Connecticut Smarter Balanced Summative Assessments

2014–2015 Technical Report

Addendum to the Smarter Balanced Technical Report

CSDE

CONNECTICUT STATE
DEPARTMENT OF EDUCATION

DRAFT

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 developed a system of valid, reliable, and fair next-generation assessments aligned to the *Common Core State Standards (CCSS)* in English language arts/literacy (ELA/L) and mathematics for grades 3–8 and 11. The system—which includes both summative assessments for accountability purposes and optional interim assessments for instructional use—uses *computer adaptive testing* technologies to the greatest extent possible to provide meaningful feedback and actionable data that teachers and other educators can use to help students succeed. The Smarter Balanced Assessment Consortium (the Consortium) is a state-led enterprise intended to provide leadership and resources to improve teaching and learning by creating and maintaining a suite of summative, interim, and formative assessments and tools aligned to the CCSS in ELA/L and mathematics.

Connecticut is among 18 member states (plus the U.S. Virgin Islands) leading a Smarter Balanced Assessment Consortium that developed a new assessment system to measure whether students are meeting the CCSS for ELA/L and mathematics and are on track for college and career readiness.

The Connecticut State Board of Education formally adopted the CCSS in ELA/L and mathematics on [Date and Year] (State Board meeting minutes, 20xx). All students in Connecticut, including Students with Significant Cognitive Disabilities who are eligible to take the Connecticut State Alternate Assessment, an AA-AAAS, are taught to the same academic content standards. Connecticut CCSS define the knowledge and skills students need to succeed in college and careers when they graduate. They align with college and workforce expectations, are clear and consistent, include rigorous content and application of knowledge through higher-order skills, are evidence-based, and are informed by standards in top-performing countries.

Since the adoption of the CCSS in 20xx, the Connecticut Department of Education fully implemented CCSS in all grade levels in SY 2013–2014. 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. The 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 end-of-year summative assessment designed for accountability purposes and optional interim assessments designed to support teaching and learning throughout the year. Summative assessments determine students' progress toward college and career readiness in ELA/L and math. These are given at the end of the school year and consist of two parts: a computer adaptive test (CAT) and a performance task.

- **Computer Adaptive Test (CAT):** 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. They 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, which cannot be adequately assessed with selected- or constructed-response items. Some performance task items can be scored by the computer, but most will be manually scored.

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

- Interim Comprehensive Assessments (ICAs) that test the same content and report scores on the same scale as the summative assessments.

- Interim Assessment Blocks (IABs) that focus on smaller sets of related concepts and provide more detailed information for instructional purposes.

This report provides a technical summary of the 2014–2015 summative tests in ELA/L and mathematics administered in grades 3–8 and 11 under the Connecticut Smarter Balanced assessments. The report includes eight chapters on overview, test administration, summary of 2014–2015 operational administration, validity and reliability of the test scores, reporting and interpreting scores, and quality control process. The data included in this report are based on Connecticut data for the summative assessment only, except for the evidence on relations to other variables in the validity section. The external validity was examined using Hawaii data. For the interim assessments, the number of students who took is provided in Appendix B. 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 are included in the Consortium technical report.

The Consortium 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 2014–2015 Smarter Balanced Assessments testing window spans three months for the summative assessments and six months for the interim assessments. The paper-pencil fixed forms for summative assessments were administered concurrently during the three-month online summative window. Table 1 shows the testing windows for both online and paper-pencil assessments.

Table 1. 2014–2015 Testing Windows

Tests	Grade	Start Date	End Date	Mode
Summative Assessments	3–8	3/17/2015	6/12/2015	Online Adaptive
	11	4/27/2015	6/12/2015	Online Adaptive
	3–8	3/17/2015	6/12/2015	Paper Fixed Forms
	11	4/27/2015	6/12/2015	Paper Fixed Forms
Interim Comprehensive Assessments	3–8, 11	1/27/2015	6/12/2015	Online Fixed Forms
Interim Assessment Blocks	3–8, 11	1/27/2015	6/12/2015	Online Fixed Forms

2.2 TEST ADMINISTRATION

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 2014–2015 administration to accommodate students' needs. Table 2 lists the testing options that were offered in 2014–2015. A testing option is selected for each content area. Once the testing option is selected, it applies to all tests within that content area, whether in online or paper-pencil format.

Table 2. Summary of Tests and Testing Options in 2014–2015

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

To ensure standardized administration conditions, Teachers (TEs) and Test Administrators (TAs) follow procedures outlined in the *Test Administration Manual* (TAM). TEs and TAs must review the TAM prior to the beginning of testing, ensure that the testing room is prepared for testing (e.g., removing certain classroom posters, arranging desks), and establish make-up procedures 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 for all assessments.

2.2.1 Administrative Roles

The key personnel involved with the test administration are District Administrators (DAs), District 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 *Test Administration Manual*, provided online at this URL: http://ct.portal.airast.org/resources/.

District Administrator (DA)

The District Administrator (DA) is a District Test Coordinator (DC) 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 District Test Coordinator's (DC) primary responsibility is to coordinate the administration of the Smarter Balanced assessment in the district.

DCs are responsible for the following:

- Reviewing all Smarter Balanced policy and test administration documents
- Reviewing scheduling and test requirements with SCs and TEs/TAs
- Working with SCs and Technology Coordinators to ensure all systems, including the secure browser, are properly installed and functioning
- Importing users (SCs, TEs, TAs) into TIDE
- Verifying all student information and eligibility in TIDE
- Scheduling and administering training sessions for all SCs, TEs, TAs, and Technology Coordinators
- Ensuring that all personnel are trained on how to properly administer the Smarter Balanced assessments
- Monitoring secure administration of the test
- Investigating and reporting all testing improprieties, irregularities, and breaches reported by the TEs/TAs
- Attending to any secure material according to state and Smarter Balanced policy

School Test Coordinator (SC)

The School Test Coordinator's (SC) primary responsibilities are to coordinate the administration of the Smarter Balanced assessment and ensure that testing within his or her school is conducted in accordance with the test procedures and security policies established by the Connecticut State Department of Education (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 correct student information and test settings for designated supports and accommodations are applied
- Identifying students who may require designated supports and test accommodations and ensuring that procedures for testing these students follow state and Smarter Balanced policy
- Attending all district trainings and reviewing all Smarter Balanced policy 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 o any secure material according to state and Smarter Balanced policy

Teachers (TEs)

A Teacher responsible for administering the Smarter Balanced assessments must have the same qualifications as a Test Administrator (TA). This role has the same test administration responsibilities as a TA. The TE role also allows users to view student 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 Administrators (TAs)

A Test Administrator’s primary responsibility is to administer the Smarter Balanced assessments. The Test Administrator (TA) role does not allow for 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 (see section 1.4 Training and reviewing all Smarter Balanced policy and test administration documents prior to administering any Smarter Balanced assessments
- Viewing student information prior to testing to ensure that the correct student receives the proper test with the appropriate supports. TAs should report any potential data errors to SCs and DCs as appropriate
- Administering the Smarter Balanced assessments

- Reporting all potential test security incidents to the SC/DC in a manner consistent with Smarter Balanced, state, and district policies

2.2.2 Online Administration

Smarter Balanced Assessments allow schools to choose testing dates to test students in intervals rather than in one long period. 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.

School Test Coordinators (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. 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 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 certification course receive a certificate of completion and appear in the online testing system.

To start a test session, the TA must first enter the TA Interface of the online testing system using his or her own computer. A session ID is generated when the test session is created. Students who are taking the assessment with the TA need to enter their State Student Identification Number (SSID), first name, and session ID into the student interface using computers provided by the school. The TA then verifies that the students are taking the appropriate content area assessment(s), using the correct test opportunity, and are provided with the appropriate assessment accommodations, such as testing in a small group (see Section 2.6 for a list of accommodations). Students can begin testing only after the TA confirms that the students are taking the appropriate assessments(s) and approves them to be tested. The TA needs to read the *Directions for Administration* in the *Test Administration Manual* aloud to the students and guide them through the login process.

Once an assessment is started, the student must answer all test questions presented on a page before proceeding to the next page; students are not allowed to skip questions. For the online computer adaptive test (CAT), students are allowed to scroll back to review and edit answers as long as the student is in the same test session and only if the test session has not been paused for more than 20 minutes. There is no pause rule implemented for the performance tasks. Students can return to the performance tasks to review and edit items they have previously completed.

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 or the assessment opportunity will expire. For the performance tasks, the assessment must be completed within 10 calendar days of the start date or the assessment opportunity will expire.

TEs/TAs can also pause a single student’s assessment or all of the assessments during a test session (for example, to give students a break). It is up to the TE/TA to determine an appropriate stopping point; however, for ELA/L and math CAT, the assessments cannot be paused for more than 20 minutes to ensure the integrity of the assessments. If an assessment is paused for more than 20 minutes, the student can continue the same assessment opportunity but must do so in a new test session. In the new test session, answers provided in the previous session are not available for review or editing.

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

2.2.3 Paper-Pencil Test Administration

The paper-pencil versions of the Smarter Balanced ELA/L and mathematics assessments are provided as an accommodation for students who cannot take the assessments online. For Connecticut, paper/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/pencil version needs to submit a request on behalf of the student to the Department. If the request is approved, the testing contractor will ship the appropriate test booklets and paper/pencil Test Administration Manual to the district.

For the ELA/L and mathematics assessments, each content area has a separate test booklet. The CAT and the performance task are combined into one test book. In both content areas, three sessions (two for CAT and one for performance task) are included in each test booklet so that the TE/TA can break up the assessment into separate sessions.

The student enters his or her answers into the test booklet using a pencil. 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 handscores the handscored items. Once all the items have been handscored, the testing vendor will score the overall test.

2.2.4 Braille Test Administration

In SY 2014–2015, the Online Smarter Balanced Assessment is made available to students who use Braille as a mode of instruction, allowing these students to have access to the adaptive online summative assessments and the online performance task.

The Braille interface of the online Smarter Balanced Assessments to students in the following 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 adaptive online summative test or the performance task via a Braille embosser.
- Students taking the summative ELA/L are able to 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 is presented to the student with items in either contracted or un-contracted Literary Braille (for items containing only text) and via a Braille embosser (for items with tactile or spatial components that cannot be read by a Refreshable Braille Display).

Prior to administering the online summative assessments using the Braille interface, TAs must ensure that the technical requirements are met. These requirements apply to the student’s computer, Test Administrator’s computer, and any supporting Braille technologies used in conjunction with the Braille interface.

2.3 Training and Information for Test Coordinators and Administrators

DAs, DCs, and 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. 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.

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 the Test Delivery System (TDS), including how to set student settings in TIDE and how to start and monitor a test session using the Test Administrator (TA) Interface.

Online Reporting System (ORS): The webinar provides an overview of the Online Reporting System, including how to retrieve student results for the Smarter Balanced Spring 2015 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. 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 practice tests mirror the Smarter Balanced summative assessments for English language arts/literacy (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 mathematics and ELA/L), 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 mathematics and ELA/L. Training tests are available for both mathematics and ELA/L and are organized by grade bands (grades 3–5, grades 6–8, and grades 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, i.e., multiple-choice items, grid items, and natural language items.

Manuals and User Guides

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

The *Test Coordinator Manual* provides information for District/School Test Coordinators regarding policies and procedures for the 2015 Smarter Balanced assessments in mathematics and English language arts/literacy.

The *Summative Assessment Test Administration Manual* provides information for Test Examiners administering the Smarter Balanced online summative assessments in English language arts/literacy 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, navigating an online test with JAWS, and how to administer a test to a student requiring Braille.

The *System Requirements for Online Testing* 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 (TIDE) User Guide* is designed to help users navigate TIDE. Users can find information on managing user account information, managing student account information, student test settings and accommodations, appeals, and voice packs

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

The *Test Administrator (TA) User Guide* is designed to help users navigate TDS including the Student Interface and the Test Administrator Interface, and help support Test Administrators manage and administer online testing for students.

The *Assessment Viewing Application (AVA) 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 (THSS) User Guide* provides information on THSS for Scorers and Score Managers responsible for human-scored item responses on the Smarter Balanced interim assessments.

All manuals and user guides pertaining to the 2014–2015 online testing were available on the portal, and DAs, DCs, and SCs can use to train TAs regarding test administration policies and procedures.

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 PPT format; two modules were also narrated.

Assessment Viewing Application (AVA) 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 (ORS) Module: This module explains how to navigate ORS, including participation reports and score reports.

Performance Task Overview Module: This module provides an overview of what a performance task is 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. This module includes how students log into the testing system and select a test, the layout of the test and the functionality of the test tools, and how students navigate through the test.

Teacher Hand Scoring System (THSS): This module provides an overview of THSS. The hand scoring system is to be used by teachers for scoring 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 staff needs to know in order to prepare for online test administration.

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

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

What is a CAT? Module: This module describes what a Computer Adaptive Test is and how it works when taking English Language Arts and Mathematics online assessments.

2.3.2 District Training Workshops

District Test Coordinator Workshops were held on January 21–23, 2015, at the Institute of Technology and Business Development (ITBD) in New Britain. Training were provided for the administration of the Smarter Balanced Assessments for English language arts/literacy and mathematics for English language arts and mathematics. During the training, District Test Coordinators were provided with information to support training School Coordinators, Teachers, and Test Administrators.

2.4 TEST SECURITY

All test items, test materials, and student-level testing information are secure materials for both online and paper-pencil 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 impropriety.

2.4.1 Student-Level Testing Confidentiality

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

- Giving out 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.

Student test materials and reports should not be exposed in such a manner that student names could be identified with student results except by authorized individuals with an educational need to know.

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

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

After a test session, only staff with the administrative roles of DAs, 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 is kept protected and that it is 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, district personnel, SCs, TAs, and teachers have well-defined roles and access to the testing system. When the TIDE window opens, CSDE provides a verified list of District Administrators (DAs) to the testing contractor who uploads the information into TIDE. DAs are then responsible for selecting and entering the 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 teachers.

Password protection: All access points by different roles—at the state level, district level, school principal level, and school staff level—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 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 TA is 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 prior to the pause. This measure was 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, etc. 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 in such a way 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 with one another through appropriate seating arrangements. For the performance tasks, different forms are spiraled within a classroom so students receive different forms of the performance tasks.

After the Test

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

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

2.4.4 Test Security Violations

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

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

Irregularity: 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: A test security incident that poses a threat to the validity of the test. Breaches require immediate attention and escalation to the state agency. 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 incident in the test security incident log. The Test Security Incident Log is the document of record for all test security incidents and should be maintained at the district level and submitted to the CSDE at the end of testing**.**

2.5 STUDENT PARTICIPATION

All students (including retained students) currently enrolled in grades 3–8 and 11 at public schools in Connecticut are required to participate in the Smarter Balanced Assessment. Students must be tested in the enrolled grade assessment; out-of-grade-level testing is not allowed for the administration of Smarter Balanced.

2.5.1 Home-Schooled Students

Students who are home-schooled may participate in the Smarter Balanced Assessment 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 Assessment:

- A student who has a significant medical emergency
- An English Language Learner (ELL) 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* are intended for school-level personnel and decision-making teams, including Individualized Education Program (IEP) and Section 504 teams, as they prepare for and implement the Smarter Balanced assessments. The *Guidelines* provide information for classroom teachers, English language development educators, special education teachers, and instructional assistants to use in selecting and administering universal tools, designated supports, and accommodations for those students who need them. The *Guidelines* are also intended for assessment staff and administrators who oversee the decisions that are made in instruction and assessment.

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

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

State-level users, Test Coordinators, and Teachers 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 prior to starting a test session.

All of the 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 Test Administrator 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 *digitally-delivered* (i.e., embedded) or separately-delivered (i.e., 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 2014–2015, the following features were available for *all* students to access. These are known as universal tools. For specific information on how to access and use these features, refer to the *Test Administrator (TA) User Guide* at http://ct.portal.airast.org.

The following are **embedded universal tools**:

Zoom in on test questions, text, or graphics.

Highlight passages or sections of passages and test questions.

Pause the assessment and return to the test question 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 on 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 on 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 the student is not able to go back 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 assessment must be completed within 45 calendar days of its starting date. The performance tasks must be completed within 10 calendar days of the starting date.

The following are **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 sixth grade and can be used on all math 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 state.

Thesaurus: A thesaurus contains synonyms of terms while a student interacts with text included in the assessment. 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 those 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 be made aware of the range of designated supports available. Smarter Balanced 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 504 Plans. Consortium-approved accommodations do not compromise the learning expectations, construct, grade-level standard, or intended outcome of the assessments.

The following lists the **embedded and non-embedded designated supports**:

Embedded

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 are able to focus their attention on a specific part of a test item by masking.

Print size: The selected print size becomes the default for all passages and items in the student's test. Regardless of the default print size assigned, all students can toggle between the five levels of print size on each test page via the [**Zoom In**] and [**Zoom Out**] buttons.

Text-to-speech (for math stimuli and items, ELA/L items, and ELA/L performance task stim and items): Text is read aloud to the student via embedded text-to-speech technology. The student is able to control the speed as well as 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 prior to beginning the actual test items. Students can see test directions in another language. As an embedded designated support, translated test directions are automatically a part of the stacked translation designated support.

Translations (glossaries) for math: Translated glossaries are a language support. The translated glossaries are provided for selected construct-irrelevant terms for math. 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 math: Stacked translations are a language support. Stacked translations are 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: Any universal tools that might be distracting, that students do not need to use, or that students are unable to use may be disabled.

Non-Embedded

Bilingual dictionary: A bilingual/dual language word-to-word dictionary is a language support. A bilingual/dual language word-to-word dictionary 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: Ear mufflers, white noise, and/or other equipment to reduce environmental noises.

Read Aloud (for math items, ELA/L items/not 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 made available for most students.

Translated test directions: PDF of directions translated in each of the languages currently supported. Bilingual adult can read to student.

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

The following lists the **embedded and non-embedded accommodations**:

Embedded

American Sign Language (ASL) for ELA/L listening items and math items: Test content is translated into ASL video. 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: A raised-dot code that individuals read with the 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 math.

Closed Captioning for ELA/L listening stims: 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 is able to control the speed as well as raise or lower the volume of the voice via a volume control.

Non-Embedded

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, 11 math tests): A non-embedded calculator for students needing a special calculator, such as a Braille calculator or a talking calculator, currently unavailable within the assessment platform.

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

Print on Demand: Paper copies of either 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 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 input devices to the computer in order 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 2014–15 administration. Tables 4–9 provide the number students who were offered with the designated supports and/or accommodations.

Table 3. SY 2014–2015 Universal Tools, Designated Supports, and Accommodations

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

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

[1] For calculator-allowed items only
[2] For ELA/L performance task full writes
[3] Includes embedded ruler, embedded protractor
[4] For ELA/L items
[5] Includes bold, italic, underline, indent, cut, paste, spell check, bullets, undo/redo
[6] For ELA/L PT stimuli, ELA/L PT and CAT items (not ELA/L CAT reading passages), and math items: Must be set in TIDE by TC, TA or TE before test begins.
[7] For math items
[8] For math items
[9] For math test
[10] For ELA/L listening items and math items
[11] For ELA/L listening items
[12] For ELA/L reading passages grades 6-8 and 11: Not available for grades 3-5. Must be set in TIDE by state-level user. TCs must submit a student's Verification of Need form to the Assessment Section for review and approval or disapproval.
[13] For ELA/L performance task full writes
[14] For ELA/L performance task full writes

[15] For ELA/L performance task full writes
[16] For ELA/L non-writing items and math items
[17] For math items
[18] Includes adapted keyboards, large keyboard, StickyKeys, MouseKeys, FilterKeys, adapted mouse, touch screen, head wand, and switches
[19] For calculator-allowed items only
[20] For math items beginning in grade 4

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

Accommodations	Grade						
	3	4	5	6	7	8	11
Embedded Accommodations							
American Sign Language	2	1	6	7	9	10	11
Closed Captioning	17	20	15	24	27	38	43
Language: Braille English		2		1		2	1
Streamlined Mode	29	55	38	24	17	22	31
Text-to-Speech: Passage & Items							
Non-Embedded Accommodations							
Alternate Response Options	4	2	1	1	2		3
Print on Demand: Stimuli & Items				1		2	
Read Aloud Passages	30	33	29	4	6	6	7
Scribe Items (Writing)	5	8	4	1	4		3
Speech-to-Text	36	30	27	45	20	29	26

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

Designated Supports	Subgroup	Grade						
		3	4	5	6	7	8	11
Color Choices	Overall	91	102	107	28	17	7	1
	ELL	16	11	13	1			
	IDEA Eligible	13	17	23	4	8	4	
Masking	Overall	205	273	190	143	200	157	37
	ELL	25	55	48	38	64	55	8
	IDEA Eligible	127	148	99	80	139	93	33
Permissive Mode	Overall	104	94	106	47	61	33	27
	ELL	25	19	19	2	3	1	4
	IDEA Eligible	83	78	83	37	59	31	20
Print Size	Overall	13	8	9	5	7	0	8
	ELL			2	1			
	IDEA Eligible	10	4	5	3	2		5
Text-to-Speech: Items	Overall	4,244	4,099	4,034	2,514	2,124	1,938	531
	ELL	1,597	1,451	1,297	668	543	490	143
	IDEA Eligible	2,326	2,565	2,696	2,002	1,685	1,558	373
Text-to-Speech: Stimuli	Overall	16	9	11	11	6	14	
	ELL	6	3	3	2	2	1	
	IDEA Eligible	8	3	6	7	4	13	
Text-to-Speech: Stimuli & Items	Overall	3,904	3,846	3,794	2,769	2,370	2,099	592
	ELL	1,486	1,372	1,215	698	571	494	142
	IDEA Eligible	2,231	2,465	2,582	2,268	1,922	1,744	451

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

Designated Supports	Subgroup	Grade						
		3	4	5	6	7	8	11
Bilingual Dictionary	Overall	141	158	156	162	140	126	98
	ELL	139	157	149	159	138	124	95
	IDEA Eligible	25	32	20	31	21	21	5
Color Contrast	Overall	1	3	6	1	1	1	2
	ELL							
	IDEA Eligible	1	2	2		1	1	1
Color Overlay	Overall	6	4	3	1	2	4	1
	ELL							
	IDEA Eligible	3	3	1	1	2	4	1
Magnification	Overall	8	5	7	6	9	5	4
	ELL					1		
	IDEA Eligible	6	2	6	3	4	3	3
Noise Buffers	Overall	14	15	6	1	3	2	2
	ELL		1					
	IDEA Eligible	4	2	2	1			1
Read Aloud Items	Overall	56	56	46	15	10	17	19
	ELL	8	8	6	2	4	3	5
	IDEA Eligible	39	41	42	10	7	14	17
Read Aloud Stimuli	Overall	45	43	24	9	6	13	12
	ELL	8	6	4	2	2	1	6
	IDEA Eligible	33	33	19	4	5	12	9
Scribe Items (Non-Writing)	Overall	5	5	2	1	5		3
	ELL			2				
	IDEA Eligible	2	5	1	1	5		2
Separate Setting	Overall	897	815	811	546	514	508	257
	ELL	160	143	115	68	65	47	28
	IDEA Eligible	574	565	576	389	373	334	191
Translated Test Directions	Overall	33	47	32	26	25	20	32
	ELL	33	47	30	21	21	14	24
	IDEA Eligible	6	4	7	10	5	10	9

Table 7. Mathematics Total Students with Allowed Embedded Accommodations

Accommodations	Grade						
	3	4	5	6	7	8	11
Embedded Accommodations							
American Sign Language	2	1	6	7	7	9	10
Language: Braille English		1		1			
Streamlined Mode	27	55	37	23	15	19	28
Non-Embedded Accommodations							
Abacus		1				1	
Alternate Response Options	3	3		1	2		2
Calculator	8	11	5	107	111	102	95
Multiplication Table		703	902	837	531	395	40
Print on Demand: Stimuli & Items				1		1	
Speech-to-Text	46	33	25	33	27	24	21

Table 8. Mathematics Total Students with Allowed Embedded Designated Supports

Designated Supports	Subgroup	Grade						
		3	4	5	6	7	8	11
Color Choices	Overall	88	102	104	28	32	30	1
	ELL	16	11	13	1	1		
	IDEA Eligible	12	17	23	4	10	6	
Masking	Overall	202	277	190	126	188	136	36
	ELL	26	55	49	36	64	54	8
	IDEA Eligible	123	153	99	85	137	92	32
Permissive Mode	Overall	100	89	109	49	58	29	28
	ELL	24	20	22	4	3	1	3
	IDEA Eligible	80	74	85	38	56	27	22
Print Size	Overall	13	6	7	5	5	0	8
	ELL			2	1			1
	IDEA Eligible	10	4	5	3	2		5
Translation (Glossary): English	Overall	37,563	38,200	38,429	39,370	38,403	39,227	32,014
	ELL	2,454	2,323	1,986	1,743	1,480	1,421	1,040
	IDEA Eligible	4,345	4,645	4,904	5,000	4,903	4,881	3,405
Translation (Glossary): Spanish	Overall	597	559	544	426	481	424	225
	ELL	581	554	532	419	468	415	581
	IDEA Eligible	32	44	46	33	40	35	32
Translation (Glossary): Other Languages	Overall	173	131	134	157	241	202	80
	ELL	57	38	49	47	59	54	37
	IDEA Eligible	2	2	2	1	2	0	0
Text-to-Speech: Items	Overall	241	239	233	245	230	201	25
	ELL	18	24	19	15	16	8	4
	IDEA Eligible	44	53	41	58	45	26	9
Text-to-Speech: Stimuli	Overall	3	4	1	6	3	8	2
	ELL	2	1		1		1	
	IDEA Eligible	1			6	3	8	2

Text-to-Speech: Stimuli & Items	Overall	4,988	4,839	4,586	3,315	2,923	2,553	618
	ELL	1,956	1,791	1,554	914	791	686	151
	IDEA Eligible	2,606	2,818	2,872	2,559	2,223	1,973	472

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

Designated Supports	**Subgroup**	**Grade**						
		3	**4**	**5**	**6**	**7**	**8**	**11**
Color Contrast	Overall	3	3	7	1	1	1	2
	ELL							
	IDEA Eligible	3	2	3		1	1	1
Color Overlay	Overall	6	4	4	2	2	3	1
	ELL							
	IDEA Eligible	3	3	2	2	2	3	1
Translation (Glossary): Spanish	Overall	203	197	168	129	119	120	47
	ELL	191	189	166	125	117	115	47
	IDEA Eligible	28	29	30	14	7	16	2
Translation (Glossary): Other Languages	Overall	31	25	28	31	23	21	15
	ELL	28	25	28	30	23	21	10
	IDEA Eligible	3	4	3	3			
Magnification	Overall	7	3	5	7	7	4	4
	ELL					1		
	IDEA Eligible	5	1	4	5	2	3	3
Noise Buffers	Overall	13	10	3	1	2		1
	ELL							
	IDEA Eligible	4	1	1				
Read Aloud Items	Overall	65	41	45	14	10	17	6
	ELL	12	12	13	6	7	5	1
	IDEA Eligible	48	24	30	5	4	13	4
Read Aloud Stimuli	Overall	64	34	43	11	9	15	3
	ELL	12	10	12	6	6	5	2
	IDEA Eligible	47	22	30	2	4	11	1
Scribe Items (Non-Writing)	Overall	4	5	3	1	2		3
	ELL			2				
	IDEA Eligible	2	4	1	1	2		2
Separate Setting	Overall	821	771	794	532	498	499	243
	ELL	157	136	125	65	70	53	28
	IDEA Eligible	516	522	552	373	352	328	176
Translated Test Directions	Overall	66	77	70	54	49	51	33
	ELL	61	77	68	49	45	45	26
	IDEA Eligible	8	6	10	12	5	13	9

2.7 DATA FORENSICS PROGRAM

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

For online administrations, quality assurance (QA) reports are generated during and after the test windows. These are geared toward detection of possible cheating, aggregating unusual responses at the student level to detect possible group-level testing anomalies.

Online test administration allows the testing contractor to track information that was not possible to track in the context of the paper-pencil tests. This information includes not only item responses but also item response changes, latencies between item responses and changes, number of revisits to an item or items, test start and end times, scores in each opportunity in the current year, scores in the previous year, and other selected information in the system (e.g., accommodations) as requested by the state. AIR's Test Delivery System (TDS) captures all of this information.

Unlike with paper assessments, where data analysis must await the close of the test window and processing of answer documents, AIR's TDS allows AIR psychometricians and state assessment staff to monitor testing anomalies throughout each test administration window, following the first operational administration. Following the base year, the analyses used to detect the testing anomalies can be run any time within the testing window. Evidence evaluated included changes in test scores across administrations, item response time, and item response patterns using the person-fit index. The flagging criteria used for these analyses are configurable and can be changed by the user. Analyses are performed at student level and summarized for each aggregate unit, including testing session, test administrator, and school.

2.7.1 Changes in Student Performance

Beginning in the 2015–2016 school year, for both online and paper test takers, it will be possible to examine score changes between years using a regression model. For between-year comparisons, the scores between past and current years are compared, with the current-year score regressed on the test score from the previous year and the number of days between test end days between two years to control the instruction time between the two test scores. Between-year comparisons are performed starting with the second year of the test administration.

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

The number of students with a large score gain or loss is aggregated for a testing session, test administrator, and school. Unusual changes in an aggregate performance between administrations and/or years is flagged based on the average studentized *t* residuals in an aggregate unit (e.g., a testing session or a test administrator). For each aggregate unit, a critical *t* value is computed and flagged when *t* was greater than |3|,

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

where *s* = standard deviation of residuals in an aggregate unit; *n* = number of students in an aggregate unit (e.g., testing session or test administrator); and $var(e_i) = \sigma^2(1 - h_{ii})$. 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 was flagged if the percentage of flagged students was greater than 50%. The aggregate unit size for the score change is based on the number of students included in the within- or between-year regression analyses in the aggregate unit.

2.7.2 Item Response Latency

The online environment also allows item response latency to be captured as the item page time (the time 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 summing 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 would be 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 test administrator 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 were flagged if the test-taking time was greater than |3| standard deviations of the state average. The state average and standard deviation was computed based on all students at the time 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 examinees whose response patterns are improbable given an IRT model. If a test has psychometric integrity, little irregularity will be seen in the item responses of the individual who responds to the items fairly and honestly.

If a test-taker has prior knowledge of some test items (or is provided answers during the exam), the student 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 latency index might flag such a student.

The person-fit index is based on all item responses. 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 and test administrator.

The person-fit index is computed using a standardized log-likelihood statistic. Following Drasgow, Levine, and Williams (1985) and Sotaridona, Pornell, and Vallejo (2003) define aberrant response patterns 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|.

$$t = \frac{Average\ l_z \text{ values}}{\sqrt{(s^2 + 1)/n}},$$

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, school).

3. SUMMARY OF 2014–2015 OPERATIONAL TEST ADMINISTRATION

3.1 STUDENT POPULATION

All students enrolled in grades 3–8 and 11 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 SY 2014–2015 Summative ELA/L Assessment

Group	G3	G4	G5	G6	G7	G8	G11
All Students	37,987	38,597	38,817	39,710	38,782	39,610	32,487
Female	18,577	19,065	18,884	19,307	18,838	19,223	15,869
Male	19,410	19,532	19,933	20,403	19,944	20,387	16,618
African American	4,922	4,778	4,876	4,833	5,001	5,067	4,107
Asian	1,917	1,969	1,996	1,959	1,876	1,752	1,473
Hispanic/Latino	8,995	8,770	8,382	8,454	8,082	8,059	6,008
American Indian/ Alaska Native	109	113	96	119	87	106	101
White	20,815	21,936	22,476	23,295	22,837	23,740	20,171
Multiple Ethnicities	1,197	991	962	1,009	875	850	600
Limited English Proficiency (LEP)	2,852	2,692	2,351	2,047	1,827	1,723	1,260
IDEA	4,363	4,695	4,955	5,042	4,948	4,941	3,463

Table 11. Number of Students in SY 2014–2015 Summative Mathematics Assessment

Group	G3	G4	G5	G6	G7	G8	G11
All Students	38,249	38,829	39,044	39,870	39,001	39,764	32,288
Female	18,701	19,180	18,980	19,372	18,952	19,282	15,771
Male	19,548	19,649	20,064	20,498	20,049	20,482	16,517
African American	4,943	4,783	4,889	4,841	5,026	5,073	4,074
Asian	1,961	2,002	2,019	1,979	1,901	1,791	1,473
Hispanic/Latino	9,176	8,929	8,550	8,577	8,270	8,203	6,009
American Indian/ Alaska Native	111	115	96	121	88	106	104
White	20,829	21,971	22,499	23,299	22,816	23,706	20,007
Multiple Ethnicities	1,197	988	961	1,013	875	848	596
Limited English Proficiency (LEP)	3,117	2,942	2,586	2,230	2,053	1,935	1,307
IDEA	4,384	4,695	4,958	5,042	4,957	4,921	3,429

3.2 Summary of Overall Student Performance

Table 12 presents the 2014–2015 state summary results for the average scale scores, the percentage of students in each achievement level, and the percentage of proficient students. The student performance by subgroups is included in Appendix A.

Table 12. SY 2014–2015 Percentage of Students in Achievement Levels

Grade	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
ELA/L							
3	2436.18	87.90	23	23	24	30	54
4	2478.61	92.53	26	19	24	31	55
5	2515.54	92.08	23	19	33	26	59
6	2537.81	91.55	19	25	35	21	56
7	2560.04	95.24	21	22	39	18	57
8	2572.14	95.72	20	26	37	17	54
11	2583.82	111.44	22	25	32	21	53
Mathematics							
3	2427.30	80.21	27	25	30	18	48
4	2469.93	80.10	23	33	27	17	44
5	2493.22	87.24	33	30	19	18	37
6	2513.31	99.72	32	31	21	16	37
7	2530.01	105.91	32	30	22	17	39
8	2541.01	114.32	37	26	19	18	37
11	2556.93	127.64	47	23	19	12	30

3.3 Test Taking Time

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 PRs or TCs 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 TAs must use their best professional judgment when allowing students extra time. Students should be actively engaged in responding productively to test questions.

In TDS, item response latency is captured as the item page time (the time each item page is presented) in milliseconds. For discrete items, items appear one item on the screen 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 with the stimulus appear on the screen at a time. For each student, the total time taken to finish the test was computed, by summing 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 13 and 14 present an average testing time and testing time by hourly intervals for overall test, CAT component, and PT component.

Table 13. 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						
3	3:19	2.04	13.18	29.10	29.18	26.50
4	3:24	1.57	11.60	28.54	30.22	28.07
5	3:18	1.18	11.70	31.55	30.93	24.63
6	3:12	1.74	13.62	32.91	29.85	21.88
7	2:52	2.27	18.81	39.77	25.60	13.55
8	2:45	2.98	21.45	40.30	24.14	11.12
11	1:59	13.71	39.04	35.27	9.78	2.20
CAT Component						
3	1:33	16.57	63.67	17.32	2.09	0.36
4	1:35	14.17	65.31	17.98	2.14	0.39
5	1:30	14.47	70.63	13.42	1.22	0.27
6	1:33	13.99	68.11	15.76	1.75	0.39
7	1:21	22.41	68.92	7.71	0.80	0.17
8	1:19	26.17	66.12	6.76	0.76	0.20
11	1:04	45.05	52.06	2.62	0.22	0.05
PT Component						
3	1:46	21.65	43.30	25.66	6.86	2.51
4	1:48	18.84	44.52	26.86	7.30	2.47
5	1:48	18.26	46.18	26.63	6.69	2.24
6	1:39	23.00	49.59	20.57	4.99	1.85
7	1:30	26.22	52.33	17.15	3.22	1.09
8	1:26	28.89	52.87	15.19	2.29	0.76
11	0:55	59.19	37.36	3.09	0.28	0.08

Table 14. 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	1:49	9.62	57.45	26.01	5.66	1.26
4	1:46	10.98	58.73	24.12	4.93	1.25
5	2:07	5.82	45.77	34.22	10.53	3.67
6	1:56	5.90	55.29	31.07	6.12	1.63
7	1:37	13.79	63.51	19.25	2.68	0.77
8	1:43	13.72	57.41	23.96	3.76	1.16
11	1:14	37.40	52.19	9.48	0.81	0.12
CAT Component						
3	1:07	47.43	47.76	4.37	0.38	0.06
4	1:09	46.01	47.83	5.50	0.52	0.15
5	1:12	39.24	53.90	6.14	0.57	0.14

6	1:10	39.10	56.04	4.38	0.37	0.11
7	1:11	38.90	55.58	4.81	0.55	0.16
8	1:08	42.85	52.35	4.15	0.52	0.13
11	0:52	65.97	32.76	1.18	0.08	0.02
PT Component						
3	0:42	82.59	16.20	1.06	0.11	0.04
4	0:37	88.17	11.26	0.51	0.05	0.01
5	0:55	65.37	30.50	3.35	0.59	0.19
6	0:46	80.08	18.50	1.15	0.20	0.07
7	0:26	96.28	3.55	0.14	0.02	0.01
8	0:35	90.89	8.63	0.45	0.02	0.01
11	0:23	98.06	1.90	0.05	0.00	0.00

3.4 Student Ability–Item Difficulty Distribution for the 2014–2015 Operational Item Pool

Figures 1 and 2 display the empirical distribution of the Connecticut student scale scores in the 2014–2015 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 accurately high performing students but needs easy items to better measure low performing students. The Smarter Balanced plans to add more easy items to the pool, and augment the pool, proportional to the test blueprint constraints, e.g., content, Depth-of-Knowledge (DOK), item type, and item difficulties.

Figure 1. SY 2014–2015 Student Ability–Item Difficulty Distribution for ELA/L

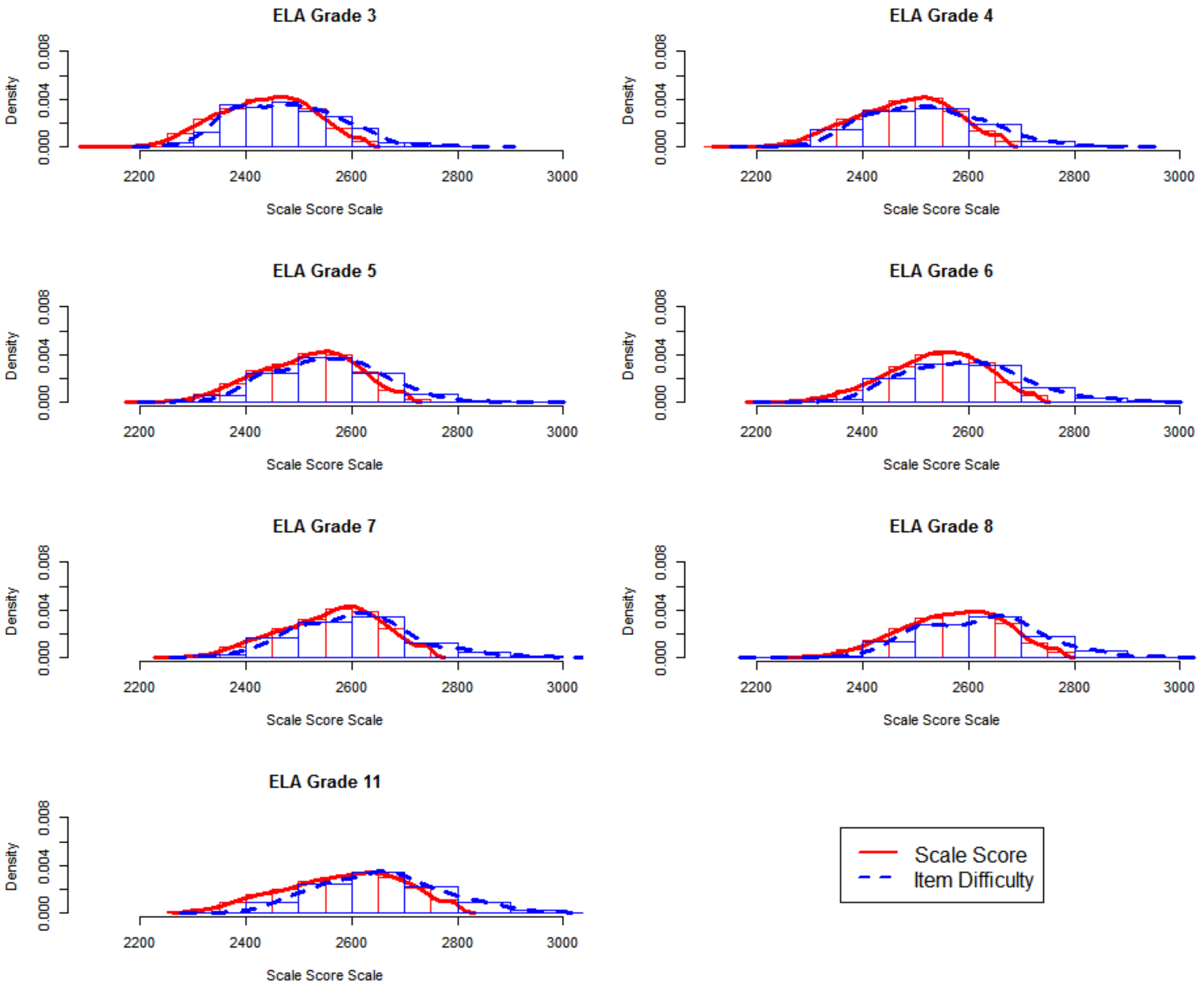
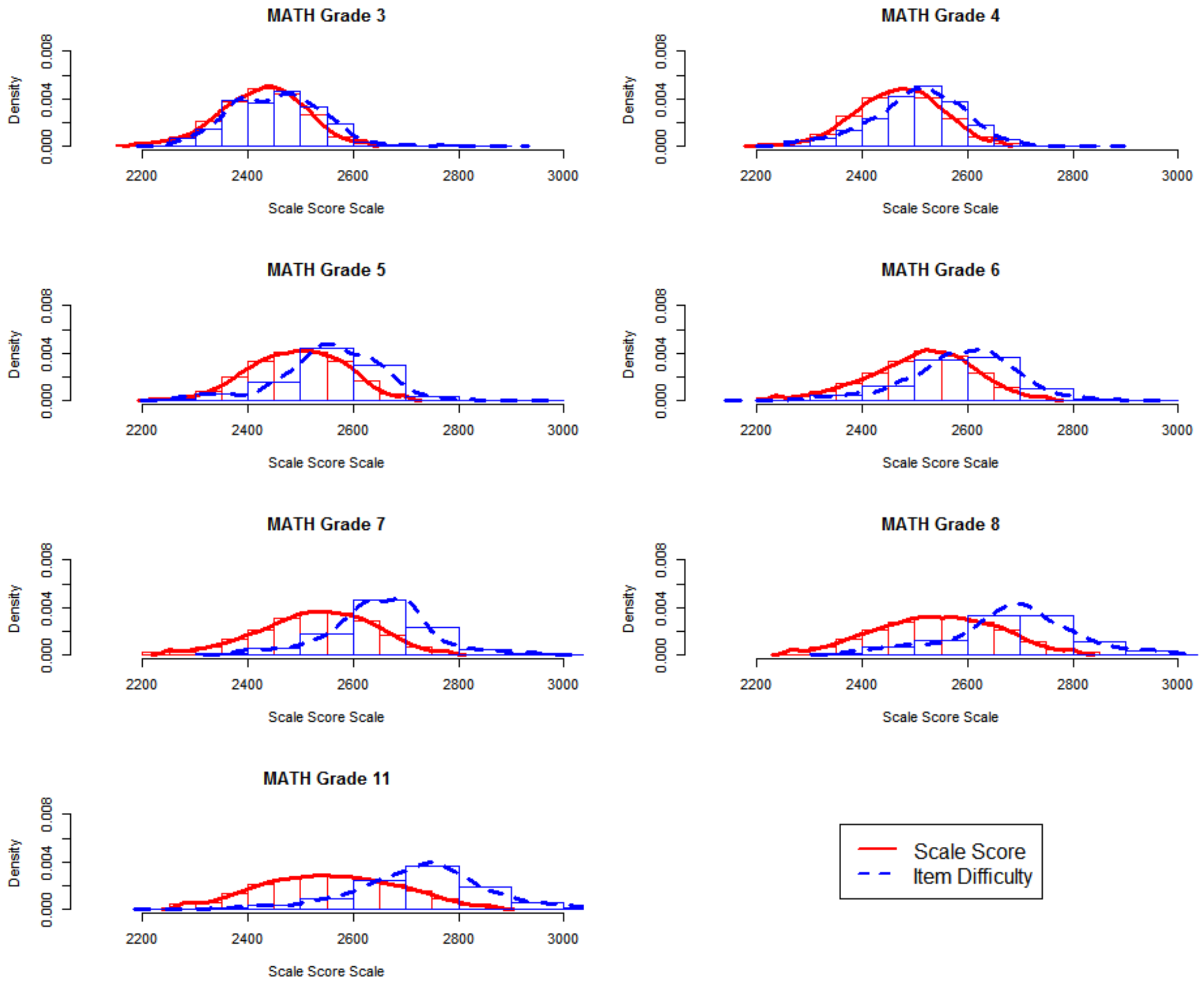


Figure 2. SY 2014–2015 Student Ability–Item Difficulty Distribution for Mathematics



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 Variables (External Structure)

Evidence on test content validity is provided with the blueprint match rates for the delivered tests. Evidence on internal structure is examined in the results of intercorrelations among reporting category scores. Evidence on external structure is examined using Hawaii data, the relationships between Smarter Balanced test scores and ACT scores.

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

4.1 Evidence on Test Content

The Smarter Balanced summative assessment includes 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. The content converge in all PT forms is 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 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 15–18 present the percentages of tests aligned with the test blueprint constraints for ELA/L and mathematics for CAT. The blueprint match rates are summarized for item and passage requirements in ELA/L, and for claims and content domains in mathematics, within each claim.

In ELA/L, all tests met the blueprint constraints for claims and passages in all delivered tests, except for very few tests. 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 more than the minimum or maximum item requirements for content domains. For the target-level constraints, most blueprint violations involved administering one item fewer or more than the minimum or maximum item requirements in both ELA/L and mathematics.

The coverage of the blueprint constraints in each test was same for *all* students indicating the validity and the comparability of all tests across all students. All tests are equivalent in the content coverage and produce comparable scores using the item parameters from the operational item pool, ensuring the comparability of assessments in content and scores.

Table 15. 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%
3	1-LT	7	8	100%	100%
3	2-W	10	10	100%	
3	3-L	8	8	100%	100%
3	4-CR	6	6	100%	
4	1-IT	7	8	100%	100%
4	1-LT	7	8	100%	100%
4	2-W	10	10	100%	
4	3-L	8	8	100%	100%
4	4-CR	6	6	100%	
5	1-IT	7	8	100%	100%
5	1-LT	7	8	100%	100%
5	2-W	10	10	100%	
5	3-L	8	9	100%	100%
5	4-CR	6	6	100%	
6	1-IT	10	12	100%	100%
6	1-LT	4	4	100%	100%
6	2-W	10	10	100%	
6	3-L	8	9	100%	100%
6	4-CR	6	6	100%	
7	1-IT	10	12	100%	100%
7	1-LT	4	4	100%	100%
7	2-W	10	10	98%	
7	3-L	8	9	100%	100%
7	4-CR	6	6	100%	
8	1-IT	12	12	100%	100%
8	1-LT	4	4	100%	100%
8	2-W	10	10	100%	
8	3-L	8	9	100%	100%
8	4-CR	6	6	100%	
11	1-IT	11	12	100%	100%
11	1-LT	4	4	100%	100%
11	2-W	10	10	100%	
11	3-L	8	9	100%	100%
11	4-CR	6	6	100%	

Table 16. Percentage of Delivered Tests Meeting Blueprint Requirements for Each Claim and Content Domain: Grade 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%
1	P	15	15	100%	15	15	100%	15	15	100%
1	S	5	5	100%	5	5	100%	5	5	100%
2	ALL	3	3	100%	3	3	100%	3	3	100%
2	G	0	2	100%	0	2	100%	0	2	100%
2	MD	0	2	100%	0	2	100%	0	2	100%
2	NBT	0	2	100%	0	2	100%	0	2	100%
2	NF	0	2	100%	1	3	100%	1	3	100%
2	OA	0	2	100%	0	2	100%	0	2	100%
3	ALL	8	8	100%	8	8	100%	8	8	100%
3	G							0	3	100%
3	MD	0	4	100%				0	4	100%
3	NBT				0	4	100%	0	4	100%
3	NF	2	6	100%	2	6	97%	2	6	100%
3	OA	0	4	100%	0	4	100%			
3	OTHER				0	2	100%			
4	ALL	3	3	100%	3	3	100%	3	3	100%
4	G	0	1	100%	0	1	100%	0	1	100%
4	MD	1	2	100%	0	2	100%	1	2	100%
4	NBT	0	1	100%	0	1	100%	0	1	100%
4	NF	0	1	100%	0	2	100%	1	2	100%
4	OA	1	2	100%	0	2	100%	0	1	100%

Table 17. Percentage of Delivered Tests Meeting Blueprint Requirements for Each Claim and Content Domain: Grade 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%
1	P	Calc	3	3	100%	6	6	100%
1	S	Calc	3	3	100%	4	4	100%
1	ALL	NoCalc	13	13	100%	10	10	100%
1	P	NoCalc	11	11	100%	9	9	100%
1	S	NoCalc	2	2	100%	1	1	100%
2	ALL	Calc	3	3	100%	3	3	100%
2	EE	Calc	0	2	100%	0	2	100%
2	G	Calc	0	2	100%	0	2	100%
2	NS	Calc	0	2	100%	0	2	100%
2	RP	Calc	0	2	100%	0	2	100%
2	SP	Calc	0	2	100%	0	2	100%
2	OTHER	Calc	0	2	100%	0	2	100%
3	ALL	Calc	7	7	100%	8	8	100%
3	EE	Calc	0	5	100%	1	5	100%
3	NS	Calc	2	6	100%	1	5	100%
3	RP	Calc	0	5	100%	1	5	100%
3	ALL	NoCalc	1	1	100%			
3	EE	NoCalc	0	1	100%			
3	NS	NoCalc	0	1	100%			
3	RP	NoCalc	0	1	100%			
4	ALL	Calc	3	3	100%	3	3	100%
4	EE	Calc	0	1	100%	0	1	99%
4	G	Calc	0	1	100%	0	1	100%
4	NS	Calc	0	1	100%	0	1	100%
4	RP	Calc	0	1	100%	0	1	99%
4	SP	Calc	0	1	100%	0	1	100%
4	OTHER	Calc	0	1	100%	0	1	100%

Table 18. Percentage of Delivered Tests Meeting Blueprint Requirements for Each Claim and Content Domain: Grade 8, 11 Mathematics

Grade 8						Grade 11					
Claim	**Content Domain**	**Segment**	**Min**	**Max**	**%BP Match**	**Claim**	**Content Domain**	**Segment**	**Min**	**Max**	**%BP Match**
1	ALL	Calc	14	14	100%	1	ALL	Calc	11	11	100%
1	P	Calc	11	11	100%	1	P	Calc	8	8	100%
1	S	Calc	3	3	100%	1	S	Calc	3	3	100%
1	ALL	NoCalc	6	6	100%	1	ALL	NoCalc	11	11	100%
1	P	NoCalc	4	4	99%	1	P	NoCalc	8	8	100%
1	S	NoCalc	2	2	99%	1	S	NoCalc	3	3	100%
2	ALL	Calc	3	3	100%	2	ALL	Calc	3	3	100%
2	EE	Calc	0	2	100%	2	A	Calc	1	2	100%
2	F	Calc	0	2	100%	2	F	Calc	0	2	100%
2	G	Calc	0	2	100%	2	G	Calc	0	2	100%
2	NS	Calc	0	2	100%	2	N	Calc	0	2	100%
2	SP	Calc	0	2	100%	2	S	Calc	0	2	100%
2	OTHER	Calc	0	2	100%	2	O	Calc	0	2	100%
3	ALL	Calc	8	8	100%	3	ALL	Calc	7	7	100%
3	EE	Calc	1	5	97%	3	A	Calc	1	4	100%
3	F	Calc	1	5	100%	3	F	Calc	0	4	100%
3	G	Calc	1	5	100%	3	G	Calc	1	4	100%
						3	N	Calc	0	4	100%
						3	ALL	NoCalc	1	1	100%
						3	A	NoCalc	0	1	100%
						3	F	NoCalc	0	1	100%
						3	G	NoCalc	0	1	100%
						3	N	NoCalc	0	1	100%
4	ALL	Calc	3	3	100%	4	ALL	Calc	3	3	100%
4	EE	Calc	1	2	99%	4	A	Calc	0	2	100%
4	F	Calc	0	1	97%	4	F	Calc	0	1	99%
4	G	Calc	0	1	100%	4	G	Calc	0	1	94%
4	NS	Calc	0	1	100%	4	N	Calc	0	2	100%
4	SP	Calc	0	1	100%	4	S	Calc	0	2	100%
4	OTHER	Calc	0	1	100%	4	O	Calc	0	1	100%

Table 19 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 19. Number of Unique Targets Assessed Within Each Claim Across all Delivered Tests

Grade	Total Targets in BP				Mean				Range (Minimum - Maximum)			
	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4
ELA/L												
3	14	5	1	3	11.0	4.0	1.0	3.0	8-14	3-5	1-1	3-3
4	14	5	1	3	10.5	4.0	1.0	3.0	8-14	3-5	1-1	3-3
5	14	5	1	3	11.1	4.7	1.0	3.0	9-13	4-5	1-1	3-3
6	14	5	1	3	9.8	5.0	1.0	3.0	8-11	4-5	1-1	3-3
7	14	5	1	3	9.6	4.0	1.0	3.2	8-11	3-5	1-1	3-4
8	14	5	1	3	10.4	4.0	1.0	3.0	8-11	3-5	1-1	3-3
11	14	5	1	3	8.7	5.0	1.0	3.0	6-11	4-5	1-1	3-3
Mathematics												
3	11	4	6	6	9.1	2.0	5.4	3.0	7-11	2-2	3-6	2-4
4	12	4	6	6	10.0	2.0	5.4	3.0	9-11	2-2	3-6	2-3
5	11	4	6	6	9.0	2.0	5.3	3.0	8-9	2-2	3-6	3-4
6	10	4	6	6	9.9	2.0	4.2	3.0	8-10	2-2	3-6	3-3
7	9	4	7	6	8.0	2.0	4.9	3.0	7-8	2-2	3-6	3-4
8	10	4	7	6	10.0	2.0	5.0	3.0	9-10	2-2	3-6	2-4
11	16	4	7	6	15.4	2.0	4.6	3.0	13-16	2-2	3-7	2-3

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. The blueprint match and target coverage results demonstrate that all test forms conform to the same content target, 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 20–21. The correlation 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 $r_{x'y'} = r_{xy} / SQRT(r_{xx} * r_{yy})$, where $r_{x},r_{y},$ 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.

Table 20. Correlations among Reporting Categories for ELA/L

Grade	Reporting Categories	Observed & Dis-attenuated Correlation			
		Claim 1	Claim 2	Claim 3	Claim 4
3	Claim 1: Reading		0.91	0.95	0.94
	Claim 2: Writing	0.72		0.92	0.91
	Claim 3: Listening	0.64	0.63		0.92
	Claim 4: Research	0.68	0.66	0.58	
4	Claim 1: Reading		0.93	0.96	0.95
	Claim 2: Writing	0.72		0.92	0.92
	Claim 3: Listening	0.65	0.63		0.94
	Claim 4: Research	0.68	0.67	0.59	
5	Claim 1: Reading		0.92	0.98	0.95
	Claim 2: Writing	0.72		0.92	0.96
	Claim 3: Listening	0.66	0.62		0.97
	Claim 4: Research	0.69	0.70	0.61	
6	Claim 1: Reading		0.91	0.99	0.95
	Claim 2: Writing	0.69		0.96	0.96
	Claim 3: Listening	0.60	0.61		1.00
	Claim 4: Research	0.64	0.67	0.58	
7	Claim 1: Reading		0.93	1.00	0.98
	Claim 2: Writing	0.72		0.96	0.95
	Claim 3: Listening	0.64	0.61		0.99
	Claim 4: Research	0.69	0.69	0.58	
8	Claim 1: Reading		0.93	0.98	0.96
	Claim 2: Writing	0.72		0.93	0.94
	Claim 3: Listening	0.63	0.60		0.96
	Claim 4: Research	0.68	0.67	0.57	
11	Claim 1: Reading		0.93	0.95	0.96
	Claim 2: Writing	0.72		0.92	1.00
	Claim 3: Listening	0.63	0.62		0.97
	Claim 4: Research	0.67	0.71	0.59	

Table 21. Correlations among Reporting Categories for Mathematics

Grade	Reporting Categories	Observed & Dis-attenuated Correlation		
		Claim 1	Claim 2&4	Claim 3
3	Claim 1		0.97	0.97
	Claim 2 & 4	0.81		1.00
	Claim 3	0.76	0.74	
4	Claim 1		1.00	0.98
	Claim 2 & 4	0.78		1.00
	Claim 3	0.80	0.74	
5	Claim 1		1.00	0.99
	Claim 2 & 4	0.76		1.00
	Claim 3	0.74	0.70	
6	Claim 1		1.00	1.00
	Claim 2 & 4	0.79		1.00
	Claim 3	0.76	0.70	
7	Claim 1		1.00	1.00
	Claim 2 & 4	0.76		1.00
	Claim 3	0.73	0.65	
8	Claim 1		1.00	1.00
	Claim 2 & 4	0.74		1.00
	Claim 3	0.77	0.68	
11	Claim 1		1.00	1.00
	Claim 2 & 4	0.75		1.00
	Claim 3	0.71	0.66	

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. 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 evidence for convergent and discriminant validity is obtained using Hawaii data. Evidence for convergent and discriminant validity is determined by examining the patterns of correlations between Smarter Balanced Summative Assessments and performance on other tests. Observed correlations should be limited only by the unreliability of the measures.

When both assessments measure student achievement in common subject areas, as with, for example, test scores based on ACT, we expect test scores between the common subject-area assessments to be substantially correlated. In addition, we expect that the magnitude of observed correlations between test scores in different subject areas will be lower than correlations between test scores in a common subject area. It is important to note, however, that test scores across subject areas and test systems are nevertheless expected to be highly correlated. This is because even though subject-area test scores measure different academic-content domains, student achievement across subject areas is influenced by factors both internal (e.g., general intelligence) and external (e.g., socioeconomic status) to the student that contribute to student achievement across all academic subject areas so that student test scores across subject areas are highly intercorrelated. So while we certainly do expect correlations between test scores across subject areas to be lower than correlations between test scores within a subject area, we nevertheless expect the correlations across subject areas to be quite high.

The relationship between the Hawaii Smarter Balanced Assessment scores and the Hawaii ACT scores in ELA/L (reading and English combined for the ACT) and mathematics was examined to evaluate the convergent and discriminant aspects of validity using grade 11 assessment data—ELA/L and mathematics for two different traits (contents) and the Smarter Balanced Assessment scores and the ACT scores for two different methods (tests).

It was expected that the correlation between the Smarter Balanced Assessment scores and the ACT scores for the same subject (convergent validity) would be moderate and higher than the correlation between Smarter Balanced ELA/L and Smarter Balanced mathematics (discriminant validity). That is, the correlation between two tests measuring the same content would be higher than the correlation between tests measuring different contents.

The results are provided in Table 22. In most scenarios, the results are as would be expected given the criteria set forth by Campbell and Fiske (1959), providing the validity evidence. First, the reliability coefficients (numbers in boldface) were higher than the convergent and discriminant coefficients for all tests. For the reliability of ACT Reading/English combined scores, the reliability of ACT English was used as a proxy since the reliability of the reading/English total score is not provided in the ACT technical report. The reliabilities of ACT English test and reading test are 0.92 and 0.88, respectively.

Second, the scores between similar traits measured by the different methods correlated more highly with each other than they did with different traits measured by the same method. This is the evidence needed for convergent validity (numbers underlined). For example, the correlation between the Smarter Balanced mathematics and the ACT mathematics scores is 0.78. This is higher than the correlation between the Smarter Balanced ELA/L and Smarter Balanced mathematics scores (r = 0.73) and between the ACT reading/English and the ACT mathematics scores (r = 0.72).

Last, the correlations of scores between different traits are lower than the correlations between similar traits. This is the evidence needed for discriminant validity (numbers in a rectangle). All correlations between the Smarter Balanced and the ACT scores in a rectangle are lower than the underlined correlations, except for the correlation between the Smarter Balanced ELA/L and mathematics scores, which is the same as the correlation between the Smarter Balanced ELA/L and ACT reading/English scores (r = 0.73).

Overall, the observed pattern of correlations within each multitrait-multimethod matrix conforms to the criteria expected for convergent and discriminant validity.

Table 22. Relationship Between the Smarter Balanced and ACT Test Scores

Test/Subject	SB ELA/L	ACT Reading/English	SB Math	ACT Math
SB ELA	**0.92**			
ACT Reading/English	0.73	**0.92***		
SB Math	0.73	0.65	**0.89**	
ACT Math	0.63	0.72	0.78	**0.91**

5. RELIABILITY

Reliability refers to the consistency in test scores. Reliability is evaluated in terms of the standard errors of measurement. In classical test theory, reliability is defined as the ratio of the true score variance to the observed score variance, assuming the error variance is the same for all scores. Within the item response theory 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 tests is provided with marginal reliability, standard error of measurement, and decision accuracy and consistency in each achievement level.

5.1 MARGINAL RELIABILITY

For the 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 standard errors of measurement, 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 standard error of measurement of the scale score for student *i;* and σ^2 is the variance of the scale score. The higher reliability coefficient indicates the greater precision of the test.

Another way to examine test reliability is with the standard error of measurement (SEM). In the item response theory, SEM is estimated as a function of test information provided by a given set of items that make up the test. In computer-adaptive testing, 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 indicates the greater accuracy of test scores.

Table 23 presents the marginal reliability coefficients and the average conditional SEM for the total scale scores.

Table 23. Marginal Reliability for ELA/L and Mathematics

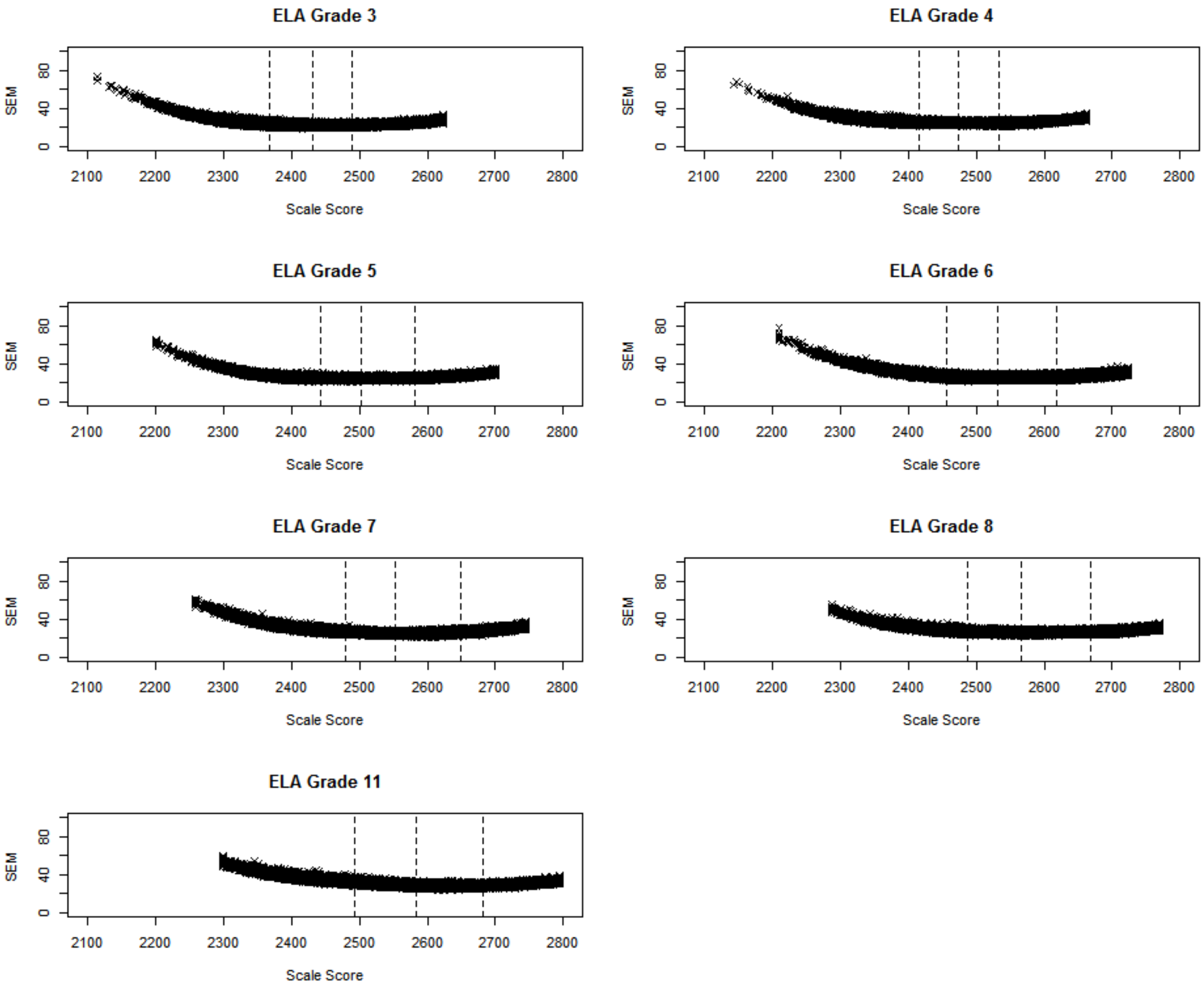
Grade	Number of Items Specified in Test Blueprint		Marginal Reliability	N	Scale Score Mean	Scale Score SD	Average CSEM
	Min	Max					
ELA/L							
3	41	44	0.92	37,987	2436.18	87.90	24.33
4	40	44	0.92	38,597	2478.61	92.53	26.15
5	41	45	0.92	38,817	2515.54	92.08	25.79
6	41	45	0.91	39,710	2537.81	91.55	27.50
7	41	45	0.92	38,782	2560.04	95.24	27.64
8	43	45	0.92	39,610	2572.14	95.72	27.56
11	42	45	0.92	32,487	2583.82	111.44	31.63
Mathematics							
3	39	40	0.94	38,249	2427.30	80.21	19.10
4	37	40	0.94	38,829	2469.93	80.10	19.37
5	38	40	0.93	39,044	2493.22	87.24	23.30
6	38	39	0.93	39,870	2513.31	99.72	26.84
7	38	40	0.91	39,001	2530.01	105.91	30.92
8	38	40	0.92	39,764	2541.01	114.32	32.95
11	40	42	0.90	32,288	2556.93	127.64	41.00

5.2 Standard Error Curves

Figures 3–4 present plots of the conditional SEM of scale scores across the range of ability. The item selection algorithm selected items efficiently, matching to each student’s ability while matching to the test blueprints, with the same precision across the range of abilities for all students

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 items that are better targeted toward these lower-achieving students, a shortage of very easy items. Content experts use this information to consider how to further target and populate item pools.

Figure 3. Conditional Standard Error of Measurement for ELA/L


ELA Grade 3

ELA Grade 4

ELA Grade 5

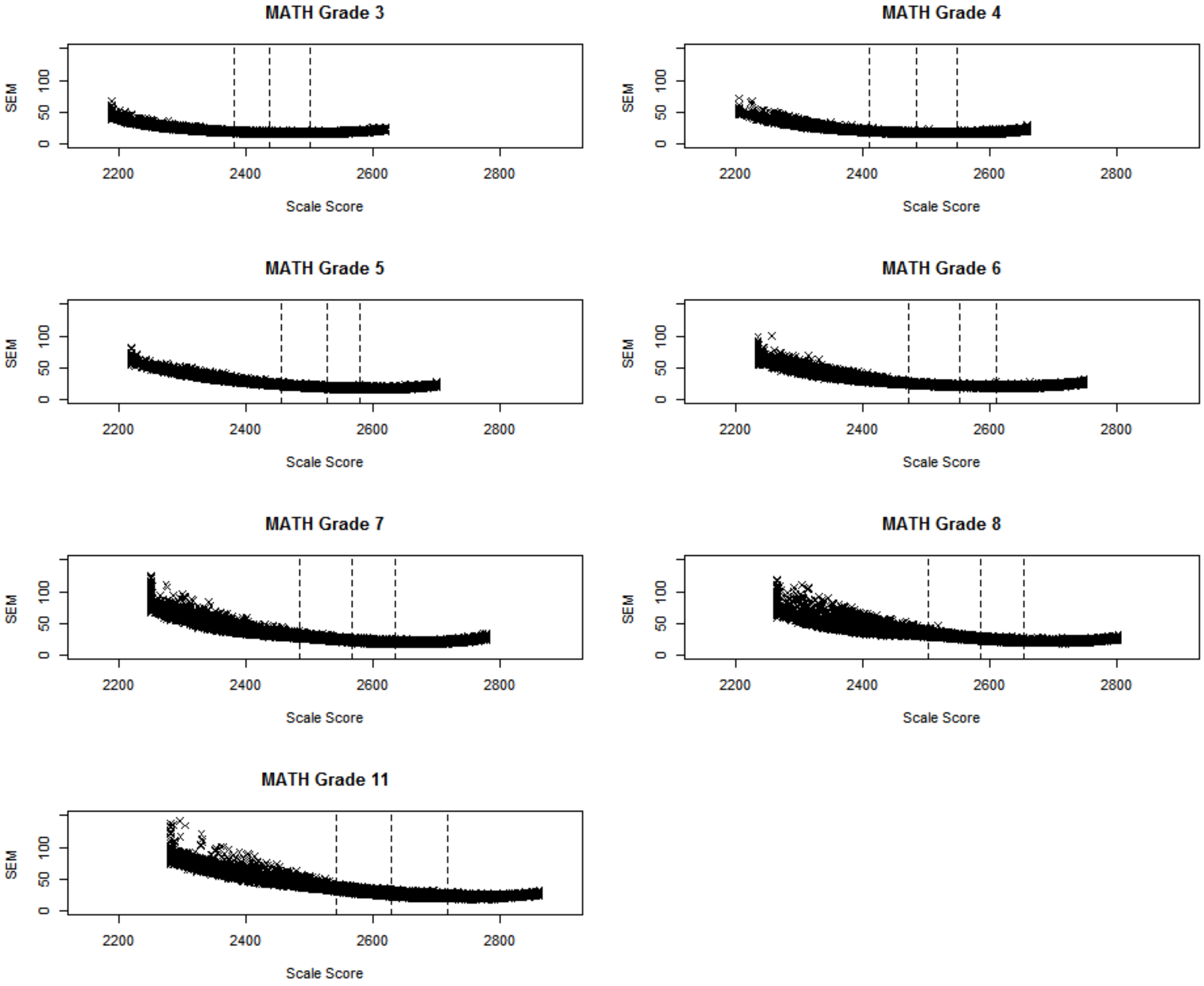
ELA Grade 6

ELA Grade 7

ELA Grade 8

ELA Grade 11

Figure 4. Conditional Standard Error of Measurement for Mathematics



The SEMs presented in Figures above are summarized in Tables 24–25. Table 24 provides the average conditional SEM for all scores and scores in each achievement level. Table 25 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 3–4, 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 in mathematics.

Table 24. Average Conditional Standard Error of Measurement by Achievement Levels

Grade	Level 1	Level 2	Level 3	Level 4	Average CSEM
ELA/L					
3	28.18	22.93	22.31	23.72	24.33
4	28.67	24.87	24.52	25.97	26.15
5	27.42	24.37	24.67	26.71	25.79
6	32.09	25.91	25.81	27.47	27.50
7	31.49	26.26	25.63	28.59	27.64
8	31.11	26.28	26.09	28.13	27.56
11	38.57	30.11	28.13	30.25	31.63
Mathematics					
3	23.04	17.81	16.69	17.88	19.10
4	24.35	18.02	16.84	17.93	19.37
5	29.81	20.52	18.24	17.95	23.30
6	35.29	22.97	20.73	21.10	26.84
7	42.66	26.38	21.37	20.62	30.92
8	42.19	29.40	23.84	22.19	32.95
11	52.43	31.20	25.03	22.82	41.00

Table 25. 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	23.77	22.60	22.42	1.17	0.18	1.35
4	25.40	24.54	24.52	0.86	0.02	0.88
5	24.32	24.51	24.95	0.19	0.44	0.63
6	26.45	25.80	26.17	0.65	0.37	0.28
7	27.09	25.75	26.17	1.34	0.42	0.92
8	27.04	25.79	26.73	1.25	0.94	0.31
11	31.75	28.50	28.36	3.25	0.14	3.39
Mathematics						
3	18.91	17.12	16.62	1.79	0.50	2.29
4	19.38	16.99	17.03	2.39	0.04	2.35
5	22.83	18.79	17.68	4.04	1.11	5.15
6	24.94	21.41	20.39	3.53	1.02	4.55
7	29.81	22.98	19.97	6.83	3.01	9.84
8	32.40	25.55	22.03	6.85	3.52	10.37
11	34.94	27.06	22.92	7.88	4.14	12.02

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 consistent classification of students as specified in standard 2.16 in the *Standards for Educational and Psychological Testing* (AERA, APA, and NCME, 2014). This index considers the consistency of classifications for the percentage of examinees that would, hypothetically, be classified in the same category on an alternate, equivalent form.

For a fixed-form test, the consistency of classifications are estimated on a single-form test scores from a single test administration based on the true-score distribution estimated by fitting a bivariate beta-binomial model or a four-parameter beta model (Huynh, 1976; Livingston & Wingersky, 1979; Subkoviak, 1976; Livingston & Lewis, 1995). For the computer-adaptive testing, because the adaptive testing algorithm constructs a test form unique to each student, targeting the student's level of ability while meeting test blueprint requirements, the consistency of classifications is based on all sets of items administered across students.

The classification index can be examined for the decision accuracy and the decision consistency. Decision 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. Decision 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 consistency is 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 $\hat{\theta}_i \sim N\left(\theta_i, se(\hat{\theta}_i)\right)$, assuming a normal distribution, where $\theta_i$is the unknown true ability of the *i*th student. The probability of the true score at achievement level *l* based on the cut scores c_{l-1} and c_l is estimated as

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

Instead of assuming a normal distribution of $\hat{\theta}_i \sim N\left(\theta_i, se(\hat{\theta}_i)\right)$, 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 $\mathbf{z}_i = (z_{i1},\cdots,z_{iJ})$ and item parameters $\mathbf{b} = (\mathbf{b}_1,\cdots,\mathbf{b}_J)$, using the *J* administered items, can be estimated as

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

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

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

where, d stands for dichotomous and p stands for polytomous items, $\mathbf{b}_j = (a_j,b_j,c_j)$ if the *j*th item is a dichotomous item, and $\mathbf{b}_j = (a_j,b_{j1},\dots,b_{jK_i})$ 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. For level 1, $cut_0 = -\infty$, and for level *L*, $cut_L = \infty$.

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 $n_{alm} = \sum_{pl_i=l} p_{im}$, pl_i is the *i*th student's achievement level. In the above table, the row represents the observed level and the column represents the expected level.

Based on the above table, the classification accuracy (*CA*) for the cut cut_l ($l = 1,\cdots,L-1$) is estimated by

$$CA_{cut_l} = \frac{\sum_{k,m=1}^{l} n_{akm} + \sum_{k,m=l+1}^{L} n_{akm}}{N},$$

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.

For classification accuracy, the false positive (*FP*) for the cut $cut_l(l = 1,\cdots,L-1)$ is estimated

$$FP_{cut_l} = \frac{\sum_{m=1}^{l}\sum_{k=l+1}^{L} n_{akm}}{N},$$

and the false negative (*FN*) for the cut $cut_l(l = 1,\cdots,L-1)$ is estimated

$$FN_{cut_l} = \frac{\sum_{k=1}^{l} \sum_{m=l+1}^{L} n_{akm}}{N}.$$

The overall false positive is estimated by

$$FP = \frac{\sum_{m=1}^{L} \sum_{k=m+1}^{L} n_{akm}}{N}.$$

The overall false negative is estimated by

$$FN = \frac{\sum_{k=1}^{L} \sum_{m=k+1}^{L} n_{akm}}{N}.$$

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 $n_{clm} = \sum_{i=1}^{N} p_{il} p_{im}$.

Based on the above table, the classification consistency (*CC*) for the cut cut_l $(l = 1, \cdots, L-1)$ is estimated by

$$CC_{cut_l} = \frac{\sum_{k,m=1}^{l} n_{ckm} + \sum_{k,m=l+1}^{L} n_{ckm}}{N}.$$

The overall classification consistency is

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

Cohen’s Coefficient Kappa Index

The probability of classification accuracy by chance, is the sum of the marginal probabilities of classified into the same level based on observed and expected classifications, hence, for the cut cut_l $(l = 1, \cdots, L - 1)$, this is estimated by

$$p_{acl} = p_{acl1} + p_{acl2},$$

where

$$p_{acl1} = \left(\frac{\sum_{k,m=1}^{l} n_{akm}}{N} + \frac{\sum_{m=1}^{l} \sum_{k=l+1}^{L} n_{akm}}{N} \right) \left(\frac{\sum_{k,m=1}^{l} n_{akm}}{N} + \frac{\sum_{k=1}^{l} \sum_{m=l+1}^{L} n_{akm}}{N} \right),$$

$$p_{acl2} = \left(\frac{\sum_{k,m=l+1}^{L} n_{akm}}{N} + \frac{\sum_{m=1}^{l} \sum_{k=l+1}^{L} n_{akm}}{N} \right) \left(\frac{\sum_{k,m=l+1}^{L} n_{akm}}{N} + \frac{\sum_{k=1}^{l} \sum_{m=l+1}^{L} n_{akm}}{N} \right).$$

For the overall classification accuracy, the chance probability is estimated by

$$p_{ac} = \sum_{l=1}^{L} \left(\frac{\sum_{m=1}^{L} n_{alm}}{N}\right) \left(\frac{\sum_{m=1}^{L} n_{aml}}{N}\right),$$

and Cohen's coefficient kappa (Cohen, 1960) is estimated by $\frac{CA_{cut_l} - p_{acl}}{1 - p_{acl}}$ for the classification accuracy at cut cut_l, and $\frac{CA - p_{ac}}{1 - p_{ac}}$ for the overall classification accuracy.

Similarly, the same calculations can be conducted for classification consistency. Hence, for cut cut_l $(l = 1, \cdots, L-1)$, the chance probability is estimated by

$$p_{ccl} = p_{ccl1} + p_{ccl2},$$

where

$$p_{ccl1} = \left(\frac{\sum_{k,m=1}^{l} n_{ckm}}{N} + \frac{\sum_{m=1}^{l} \sum_{k=l+1}^{L} n_{ckm}}{N}\right) \left(\frac{\sum_{k,m=1}^{l} n_{ckm}}{N} + \frac{\sum_{k=1}^{l} \sum_{m=l+1}^{L} n_{ckm}}{N}\right),$$

$$p_{ccl2} = \left(\frac{\sum_{k,m=l+1}^{L} n_{ckm}}{N} + \frac{\sum_{m=1}^{l} \sum_{k=l+1}^{L} n_{ckm}}{N}\right) \left(\frac{\sum_{k,m=l+1}^{L} n_{ckm}}{N} + \frac{\sum_{k=1}^{l} \sum_{m=l+1}^{L} n_{ckm}}{N}\right).$$

For the overall classification consistency, the chance probability is estimated by

$$p_{cc} = \sum_{l=1}^{L} \left(\frac{\sum_{m=1}^{L} n_{clm}}{N}\right) \left(\frac{\sum_{m=1}^{L} n_{cml}}{N}\right),$$

and Cohen's coefficient kappa is estimated by $\frac{CC_{cut_l} - p_{ccl}}{1 - p_{ccl}}$ for the classification consistency at cut cut_l, and $\frac{CC - p_{cc}}{1 - p_{cc}}$ for the overall classification consistency.

The analysis of the classification index is performed based on overall scale scores in the 2014–2015 administration. In Table 26, the decision accuracy and consistency are provided with the percentage of classification accuracy and consistency and Cohen's coefficient Kappa. 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. The accuracy and consistency indexes for each achievement level are higher for the levels with smaller standard error. Also Cohen's coefficient Kappa provides high agreement ranges across all grades and subjects. The better the test is targeted to the student's ability, the higher the reliability of classification index is.

Table 26. 2014–2015 Decision Accuracy and Consistency by Achievement Levels

Grade	Achievement Level	ELA/L				Mathematics			
		Accuracy		Consistency		Accuracy		Consistency	
		% Accuracy	Kappa	% Consistency	Kappa	% Accuracy	Kappa	% Consistency	Kappa
3	L2	94.3	0.84	91.9	0.78	94.1	0.85	91.7	0.79
	L3	92.9	0.86	90.0	0.80	93.2	0.86	90.5	0.81
	L4	93.0	0.83	90.1	0.76	95.3	0.84	93.3	0.78
4	L2	94.0	0.85	91.6	0.78	94.3	0.84	92.0	0.78
	L3	92.6	0.85	89.6	0.79	93.5	0.87	90.8	0.81
	L4	92.4	0.82	89.3	0.75	95.7	0.85	93.9	0.78
5	L2	94.5	0.85	92.3	0.78	93.1	0.85	90.3	0.78
	L3	92.9	0.85	89.9	0.79	93.9	0.87	91.4	0.81
	L4	92.5	0.80	89.4	0.73	95.4	0.84	93.5	0.78
6	L2	94.4	0.82	92.1	0.76	93.5	0.85	90.8	0.79
	L3	91.8	0.83	88.5	0.77	93.1	0.85	90.3	0.79
	L4	92.9	0.79	90.0	0.70	95.5	0.84	93.5	0.77
7	L2	94.7	0.84	92.5	0.78	92.5	0.83	89.6	0.76
	L3	92.6	0.85	89.5	0.79	93.5	0.86	90.8	0.81
	L4	93.3	0.78	90.5	0.70	95.8	0.85	94.0	0.79
8	L2	94.3	0.82	92.0	0.76	92.1	0.83	89.0	0.77
	L3	92.5	0.85	89.4	0.79	93.6	0.86	91.0	0.81
	L4	93.6	0.78	91.1	0.70	95.7	0.85	93.9	0.79
11	L2	94.4	0.84	92.0	0.77	92.2	0.84	89.1	0.78
	L3	93.0	0.86	90.0	0.80	94.5	0.87	92.3	0.82
	L4	93.4	0.80	90.7	0.73	97.0	0.86	95.8	0.80

5.4 RELIABILITY FOR SUBGROUPS

Tables 27–28 show the marginal reliability coefficients for each of the subgroups. As shown in tables, reliabilities of total scale scores are consistent across subgroups.

Table 27. Marginal Reliability Coefficients for Overall and by Subgroup for ELA/L

Subgroup	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 11
All Students	0.92	0.92	0.92	0.91	0.92	0.92	0.92
Female	0.92	0.92	0.92	0.90	0.91	0.91	0.91
Male	0.92	0.92	0.92	0.91	0.92	0.92	0.92
American Indian/ Alaska Native	0.91	0.91	0.90	0.89	0.89	0.91	0.91
Asian	0.92	0.90	0.91	0.89	0.90	0.90	0.92
African American	0.90	0.90	0.90	0.88	0.90	0.89	0.89
Hispanic/Latino	0.90	0.90	0.91	0.90	0.90	0.90	0.90
White	0.91	0.90	0.90	0.89	0.90	0.90	0.92
Multiple Ethnicities	0.92	0.92	0.92	0.91	0.91	0.92	0.92
Limited English Proficiency	0.85	0.86	0.85	0.82	0.80	0.79	0.79
IDEA	0.87	0.87	0.88	0.85	0.85	0.86	0.86

Table 28. Marginal Reliability Coefficients for Overall and by Subgroup for Mathematics

Subgroup	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 11
All Students	0.94	0.94	0.93	0.93	0.91	0.92	0.90
Female	0.94	0.94	0.92	0.92	0.91	0.91	0.89
Male	0.95	0.95	0.93	0.93	0.92	0.92	0.90
American Indian/ Alaska Native	0.94	0.93	0.88	0.9	0.87	0.88	0.84
Asian	0.95	0.95	0.94	0.94	0.94	0.94	0.93
African American	0.92	0.91	0.87	0.88	0.85	0.84	0.78
Hispanic/Latino	0.92	0.92	0.88	0.89	0.85	0.85	0.80
White	0.93	0.93	0.92	0.92	0.92	0.92	0.91
Multiple Ethnicities	0.94	0.95	0.93	0.93	0.92	0.92	0.89
Limited English Proficiency	0.89	0.89	0.82	0.81	0.71	0.68	0.65
IDEA	0.92	0.90	0.85	0.85	0.78	0.79	0.70

5.5 Reliability for Claim Scores

The marginal reliability coefficients and the measurement errors are also computed for the claim scores. 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 29–30 present the marginal reliability coefficients for each claim score in ELA/L and mathematics, respectively.

Table 29. 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	0.78	2427.59	101.72	47.92
	Claim 2: Writing	11	11	0.79	2440.32	98.20	44.94
	Claim 3: Listening	8	8	0.59	2435.66	115.70	74.26
	Claim 4: Research	8	9	0.67	2424.41	115.88	66.61
4	Claim 1: Reading	14	16	0.77	2471.58	109.84	52.66
	Claim 2: Writing	11	11	0.78	2482.94	101.34	48.05
	Claim 3: Listening	8	8	0.60	2474.30	119.28	75.65
	Claim 4: Research	7	9	0.67	2466.01	120.77	69.31
5	Claim 1: Reading	14	16	0.78	2500.72	106.67	49.71
	Claim 2: Writing	11	11	0.78	2522.51	99.47	46.49
	Claim 3: Listening	8	9	0.58	2494.93	131.49	85.28
	Claim 4: Research	8	9	0.68	2527.13	109.25	62.12
6	Claim 1: Reading	14	16	0.73	2508.12	115.42	60.52
	Claim 2: Writing	11	11	0.79	2545.27	100.03	46.22
	Claim 3: Listening	8	9	0.51	2541.41	126.33	88.50
	Claim 4: Research	8	9	0.63	2543.54	111.01	67.34
7	Claim 1: Reading	14	16	0.75	2543.88	109.39	54.36
	Claim 2: Writing	11	11	0.79	2572.58	106.55	49.27
	Claim 3: Listening	8	9	0.52	2550.32	123.01	85.51
	Claim 4: Research	8	9	0.66	2555.94	117.99	68.38
8	Claim 1: Reading	16	16	0.77	2559.62	109.28	52.66
	Claim 2: Writing	11	11	0.78	2586.46	108.04	50.33
	Claim 3: Listening	8	9	0.54	2561.65	122.40	83.27
	Claim 4: Research	8	9	0.65	2563.02	119.35	70.24
11	Claim 1: Reading	15	16	0.76	2589.09	118.71	58.47
	Claim 2: Writing	11	11	0.79	2585.73	128.25	58.61
	Claim 3: Listening	8	9	0.57	2564.82	139.27	91.36
	Claim 4: Research	8	9	0.64	2570.15	135.87	81.07

Table 30. 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: Concepts and Procedures	20	20	0.90	2426.04	83.39	26.83
	Claim 2 & 4: Problem Solving & Modeling and Data Analysis	8	11	0.76	2424.39	93.29	45.73
	Claim 3: Communicating Reasoning	9	11	0.68	2423.65	95.43	54.16
4	Claim 1: Concepts and Procedures	20	20	0.89	2469.57	83.03	26.93
	Claim 2 & 4: Problem Solving & Modeling and Data Analysis	8	10	0.67	2462.55	100.31	57.90
	Claim 3: Communicating Reasoning	9	10	0.74	2465.88	92.87	47.25
5	Claim 1: Concepts and Procedures	20	20	0.87	2489.77	90.63	32.30
	Claim 2 & 4: Problem Solving & Modeling and Data Analysis	8	10	0.60	2482.71	116.02	73.49
	Claim 3: Communicating Reasoning	9	10	0.64	2487.01	109.14	65.14
6	Claim 1: Concepts and Procedures	19	19	0.88	2510.80	105.15	36.94
	Claim 2 & 4: Problem Solving & Modeling and Data Analysis	9	10	0.61	2501.96	122.32	76.67
	Claim 3: Communicating Reasoning	10	11	0.65	2511.73	115.22	68.05
7	Claim 1: Concepts and Procedures	20	20	0.86	2528.34	110.76	41.08
	Claim 2 & 4: Problem Solving & Modeling and Data Analysis	10	10	0.56	2512.47	135.41	89.65
	Claim 3: Communicating Reasoning	8	10	0.51	2517.68	132.62	92.82
8	Claim 1: Concepts and Procedures	20	20	0.85	2536.20	118.91	46.04
	Claim 2 & 4: Problem Solving & Modeling and Data Analysis	8	10	0.58	2530.73	144.01	93.71
	Claim 3: Communicating Reasoning	9	10	0.68	2534.43	129.12	73.23
11	Claim 1: Concepts and Procedures	22	22	0.83	2552.85	134.60	55.57
	Claim 2 & 4: Problem Solving & Modeling and Data Analysis	8	10	0.50	2530.01	164.61	116.42
	Claim 3: Communicating Reasoning	9	12	0.57	2553.58	144.27	95.04

6. SCORES

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 an achievement category for each claim. This section describes the rules used in generating scores and the handscoring procedure.

6.1 Estimating Student Ability Using Maximum Likelihood Estimation

The Smarter Balanced Tests are scored using maximum likelihood estimation (MLE). The likelihood function for generating the MLEs is based on a mixture of items 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 $\boldsymbol{b}_i' = (b_{i,1}, \dots, b_{i,m_i})$ for the *i*th item's step parameters, m_i is the maximum possible score of this item, a_i is the discrimination parameter for item *i*, z_{ij} is the observed item score for the person *j*, *k* indexes step of the item *i*.

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

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

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

in the case of items with two or more points,

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

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

Standard Error of Measurement

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

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

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

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

where $m_i$is the maximum possible score point (starting from 0) for the *i*th item, 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 content area test is summarized in an overall test score referred to as a *scale score*. The number of items a student answers correctly and the difficulty of the items presented are used to statistically transform theta scores to scale scores so that scores from different sets of items can be meaningfully compared. 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 Smarter Balanced Assessment consortium. **Error! Reference source not found.**31 lists the scaling constants for each subject for the theta-to-scale score linear transformation. Scale scores will be rounded to an integer.

Table 31. Vertical Scaling Constants on the Reporting Metric

Subject	Grade	Slope (a)	Intercept (b)
ELA	3–8, HS	85.8	2508.2
Math	3–8, HS	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 32 provides three achievement standards for each grade and content area.

Table 32. ELA/L Theta Cut Scores and Reported Scale Scores

Grade	**ELA/L**			**Mathematics**		
	Level 2	**Level 3**	**Level 4**	**Level 2**	**Level 3**	**Level 4**
3	2367	2432	2490	2381	2436	2501
4	2416	2473	2533	2411	2485	2549
5	2442	2502	2582	2455	2528	2579
6	2457	2531	2618	2473	2552	2610
7	2479	2552	2649	2484	2567	2635
8	2487	2567	2668	2504	2586	2653
11	2493	2583	2682	2543	2628	2718

6.3 LOWEST/HIGHEST OBTAINABLE SCORES

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 doesn't include easy or difficult items to measure low- and high-performing students, the standard error could be large in low and high ends of the ability range. Smarter Balanced decided to truncate extreme unreliable student ability estimates. Table 33 presents the lowest obtainable score (LOT) and the highest obtainable score (HOT) 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 33. Lowest and Highest Obtainable Scores

Subject	**Grade**	**Theta Metric**		**Scale Score Metric**	
		LOT	**HOT**	**LOSS**	**HOSS**
ELA	3	-4.5941	1.3374	2114	2623
ELA	4	-4.3962	1.8014	2131	2663
ELA	5	-3.5763	2.2498	2201	2701
ELA	6	-3.4785	2.5140	2210	2724
ELA	7	-2.9114	2.7547	2258	2745
ELA	8	-2.5677	3.0430	2288	2769
ELA	11	-2.4375	3.3392	2299	2795
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
Math	11	-2.9564	4.3804	2280	2862

6.4 Scoring All Correct and All Incorrect Cases

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

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

In addition to the overall scale score, relative strength and weakness at the reporting category (claim) level is produced. In ELA, claim scores are computed for each claim. In mathematics, claim scores are computed for Claim 1, Claims 2 and 4 combined, and Claim 3.

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 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 only narrowly because they reflect only one or two ways of measuring the target. An adaptive test, however, 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 (four 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 θ as:

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

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

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

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

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

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

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

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 is evidence 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, report the following:

- If $\bar{\delta}_{Tg} \geq +1 * \ se(\bar{\delta}_{Tg})$, then performance is better than on the rest of the test.
- If $\bar{\delta}_{Tg} \leq -1 * \ se(\bar{\delta}_{Tg})$, then performance is worse than on the rest of the test.

- Otherwise, performance is similar to performance on the test as a whole.
- 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 handscoring for the Smarter Balanced summative tests. In ELA/L, short-answer (SA) items and full write items are scored by human raters, also called as handscored. In mathematics, SA items and other constructed-response items are handscored. The procedure for scoring these items is provided by Smarter Balanced.

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

6.7.1 Rater Selection

MI maintains a large pool of qualified, experienced readers at each scoring center as well as distributive readers who work remotely from their home. They only need 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. They employ many of these experienced readers for this 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. Qualified applicants are those with a four-year college degree. Each qualified applicant must pass an interview by experienced MI staff, complete ELA/L and mathematics placement tests, take a grammar exercise, write an acceptable essay, and receive good recommendations from references. MI then review all the information about an applicant before offering employment.

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

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

MI requires all handscoring project staff (scoring directors, team leaders, readers, and clerical staff) to sign a Confidentiality/Nondisclosure Agreement before receiving any training or other 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 scoring methods to any person.

6.7.2 Rater Training

All readers hired for Smarter Balanced Assessment handscoring are trained using the rubric(s) and training/qualifying sets provided by Smarter Balanced. 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 (eg., Brief Writes, Reading, Research, Full Writes, Math). 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. In scoring

Connecticut students' responses, in addition to the readers hired by MI, 31 Connecticut teachers also participated in scoring.

MI's VSC online training interface presents rubrics, scoring guides, and training/qualifying sets in three modes (regardless of mode, the same training protocol is followed):

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

After the contracts and nondisclosure forms are signed, and the introductory remarks are given by the scoring director, training begins. Reader training and team leader training follow the same format, except that team leaders are required to annotate each response in the training sets, while readers are encouraged to take notes. The scoring director presents the writing or constructed-response task and introduces the scoring guide (anchor set), then discusses, room-wide, each score point. 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 assure consistency in scoring the training/qualifying responses.

All scoring personnel log in to MI's secure Scoring Resource Center (SRC). 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 Smarter Balanced before they may read actual student responses. Any readers unable to meet the qualifying standards are dismissed. All readers understand this stipulation when they are hired. MI is always sensitive to the need for accurate and consistent scoring, and any team leader or reader who is not able to demonstrate both accurate and consistent results during training is paid for his/her time spent and then dismissed.

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 qualifying, a significant amount of time is allotted for demonstrations of the VSC handscoring system, explanations of how to "flag" unusual responses for review by the scoring director, and instructions about other procedures which are 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 Narrate, Grade 6 Argumentative, etc.), then take qualifying gate sets for each item in that grade and purpose.
- Brief Writes, Reading, and Research: readers train/qualify on a baseline set within a specific grade band and target.
- Math: 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 Rater 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 handscored educational assessment, MI constantly monitors the quality of each reader's work throughout every project. Reader Status Reports are used to monitor readers' scoring habits during the Smarter Balanced handscoring project.

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

There are currently more than 20 reports available that 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 via a secure website to the handscoring project monitors at each MI scoring center, and they provide updated reports to the scoring directors several times a 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," as well as 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 Rater Monitoring and Retraining

Team leaders spot-check (read behind) each reader's scoring to ensure that he/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 Smarter Balanced, with 10%-15% second read for reliability purposes. Items responses for second read were selected randomly and were scored blindly. The second reader were unaware of the first reader's score. MI's quality assurance/reliability procedures allow their handscoring staff to identify struggling readers very early and begin retraining immediately. During the time when they retrain these readers, MI also monitors their scoring intensively to ensure that all responses are scored accurately. In fact, the monitoring MI does is also used as a retraining method (they show readers responses that they have scored incorrectly, explain the correct scores, and have them change the scores). MI's retraining methods help readers to become accurate scorers.

6.7.5 Rater Validity Checks

Scoring directors select responses which are loaded into the VSC system as validity responses. 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.

6.7.6 Rater Dismissal

When read-behinds or daily statistics identify a reader who is unable to 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

Tables 34–35 provide a summary of the inter-rater reliability for the Connecticut data. In an adaptive test, because items are selected adapting to a student's ability while meeting the test blueprint, item usages vary across items. In this summary, items with a sample size greater than 50 are used.

In ELA/L, writing essay item response is scored in three dimensions, convention (0–2 rubric), evidence/elaboration (0–4 rubric), and organization/purpose (0–4 rubric). The short answer items are scored in 0–2. In mathematics, the maximum score points of the human-scored items range from 1-4.

Table 34. Reader Agreements for ELA/L

Grade	**Item Type**	**# of Items**	**% Exact**	**Min (%Exact)**	**Max (%Exact)**	**% items w/ %Exact ≥ 80%**	**% items w/ %Exact ≥ 70%**
3	Short Answer	39	79.73	59	93	56	92
3	WR: Conv	13	93.99	78	100	92	100
3	WR: Evid/Elab	13	94.05	77	100	92	100
3	WR: Org/Purp	13	94.18	77	100	92	100
4	Short Answer	54	75.58	59	90	30	78
4	WR: Conv	19	91.41	72	100	74	100
4	WR: Evid/Elab	19	92.13	74	100	84	100
4	WR: Org/Purp	19	91.93	76	100	84	100
5	Short Answer	55	75.72	63	89	31	76
5	WR: Conv	20	84.63	75	98	65	100
5	WR: Evid/Elab	20	83.35	68	98	45	95
5	WR: Org/Purp	20	82.69	71	98	55	100
6	Short Answer	42	75.06	65	96	19	74
6	WR: Conv	13	80.77	69	97	54	92
6	WR: Evid/Elab	13	76.54	64	94	31	69
6	WR: Org/Purp	13	76.88	67	95	31	69
7	Short Answer	51	75.59	54	91	33	78
7	WR: Conv	19	90.21	82	100	100	100
7	WR: Evid/Elab	19	89.23	76	99	84	100
7	WR: Org/Purp	19	89.05	77	99	89	100
8	Short Answer	62	77.56	64	98	27	85
8	WR: Conv	21	86.63	77	100	90	100
8	WR: Evid/Elab	21	82.27	70	100	57	100
8	WR: Org/Purp	21	82.50	69	100	57	95
11	Short Answer	72	84.88	70	100	79	100
11	WR: Conv	24	95.18	91	99	100	100
11	WR: Evid/Elab	24	95.01	87	100	100	100
11	WR: Org/Purp	24	95.44	88	100	100	100

Table 35. Reader Agreements for Mathematics

Grade	Score Points	# of Items	% Exact	Min (%Exact)	Max (%Exact)	% items w/ %Exact ≥ 80%	% items w/ %Exact ≥ 70%
3	1	14	92.42	87	99	100	100
3	2	34	90.57	74	100	94	100
3	3	4	95.22	94	96	100	100
4	1	8	89.43	87	97	100	100
4	2	36	90.16	71	97	94	100
4	3	4	85.78	83	88	100	100
5	1	4	93.41	91	98	100	100
5	2	41	88.68	74	98	88	100
5	3	8	88.16	75	99	50	100
6	1	12	97.74	92	100	100	100
6	2	44	91.24	81	99	100	100
6	3	2	73.61	67	85	50	50
6	4	4	61.51	58	66	-	-
7	1	9	97.73	95	100	100	100
7	2	29	91.26	80	99	100	100
7	3	4	95.58	85	100	100	100
8	1	15	94.87	81	100	100	100
8	2	32	91.65	85	100	100	100
11	1	22	97.94	93	100	100	100
11	2	32	95.05	83	100	100	100
11	3	6	96.45	96	99	100	100

7. REPORTING AND INTERPRETING SCORES

Online Reporting System generates a set of online score reports including reliable and valid information which describe student performance for students, parents, educators, and other stakeholders. The online score reports are produced immediately after students complete tests and the tests are handscored. Because the score report on students' performance are up to date each time students complete tests and they are handscored, authorized users (e.g., school principals, teachers) can view students' performance on the tests and use them to improve student learning. In addition to individual student's score report, the Online Reporting System produces aggregate score reports for teachers, schools, and districts. It should be noted that the Online Reporting System does not produce aggregate score reports for state. The timely accessibility of aggregate score reports helps users monitor student performance in each subject and 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 Online Reporting System provides participation data that helps monitor student participation rate.

This section contains a description of the types of scores reported in Online Reporting System and a description on the ways 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 Online Reporting System (ORS) is designed to help educators, students, and parents answer questions regarding how well students have achieved on ELA/L and mathematics. The ORS is the online tool to provide educators and other stakeholders with timely, relevant score reports and guide stakeholders to make valid, actionable interpretations of student assessment results. The ORS for the Smarter Balanced Assessment has been designed with stakeholders, such as teachers, parents, and students and, who are not technical measurement experts, in mind and ensures 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 valid inferences about student achievement. Also ORS is 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 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 choose from. For more detailed student assessment results for a school, a teacher, and a rosters, 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 36 summarizes the types of online score reports available at the aggregate level and the individual student level. Detailed information about the online score reports and instructions on how to navigate the online score reporting system can be found in the Online Reporting System User Guide, located in a help button on the ORS.

Table 36. Types of Online Score Reports by Level of Aggregation

Level of Aggregation	Types of Online Score Reports
District School Teacher Roster	• 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) • Achievement level 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: Achievement 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 subgroups. Table 37 presents the types of subgroups and subgroup category provided in ORS.

Table 37. 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
	Demographic Race Two or More Races
	Hispanic or Latino Ethnicity
	White

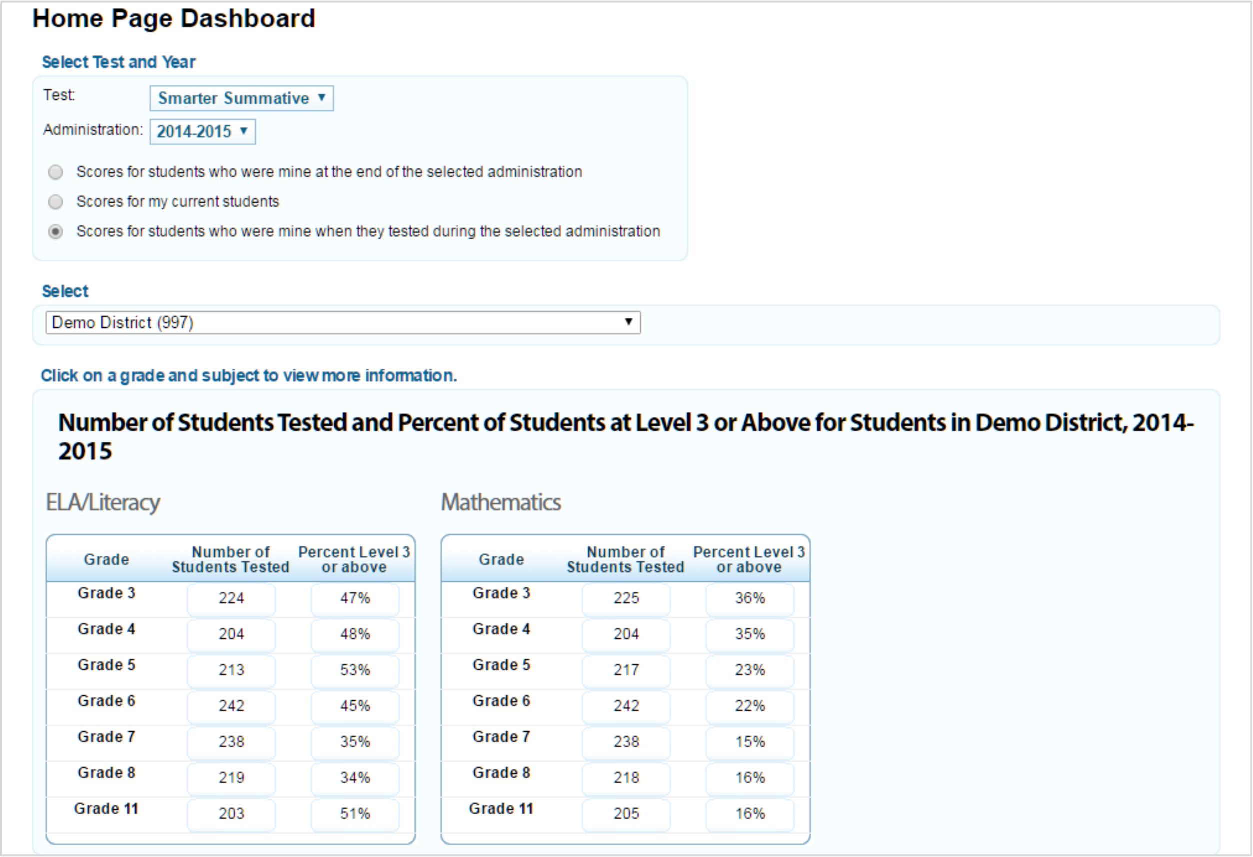
7.1.2 Online Reporting System

7.1.2.1 Home Page

The first page users see when they log onto the ORS and select Score Reports is summaries of students' performance across grades and subjects. District personnel see district summaries, school personnel see school summaries, and teachers see 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 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 sampled Home Page at a district level.

Exhibit 1. Home Page: District Level



Home Page Dashboard

Select Test and Year

Test: Smarter Summative ▾

Administration: 2014-2015 ▾

○ 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 (997) ▾

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, 2014-2015

ELA/Literacy

Grade	Number of Students Tested	Percent Level 3 or above
Grade 3	224	47%
Grade 4	204	48%
Grade 5	213	53%
Grade 6	242	45%
Grade 7	238	35%
Grade 8	219	34%
Grade 11	203	51%

Mathematics

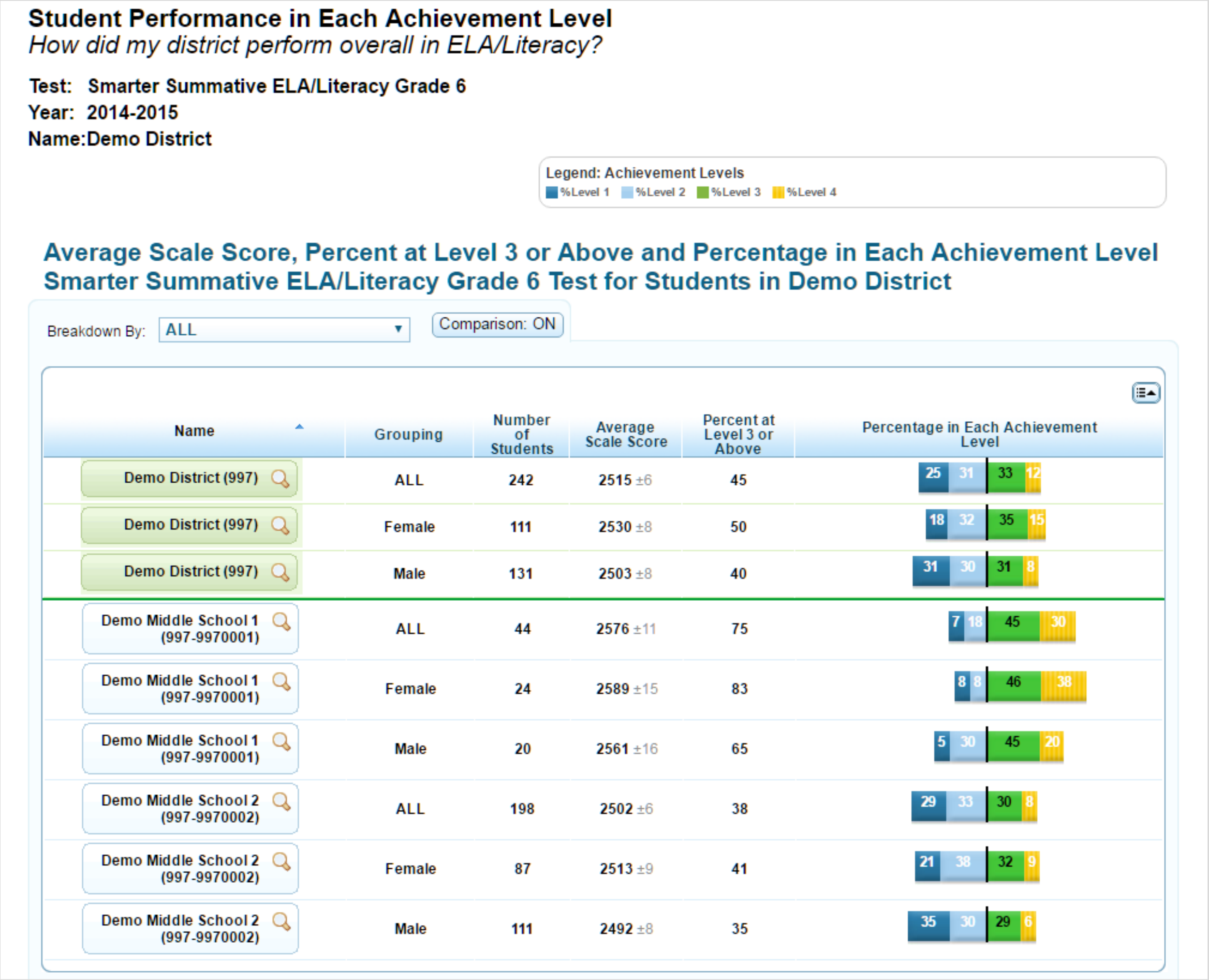
Grade	Number of Students Tested	Percent Level 3 or above
Grade 3	225	36%
Grade 4	204	35%
Grade 5	217	23%
Grade 6	242	22%
Grade 7	238	15%
Grade 8	218	16%
Grade 11	205	16%

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 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: 2014-2015
Name: Demo District

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

Breakdown By: ALL ▾ Comparison: ON

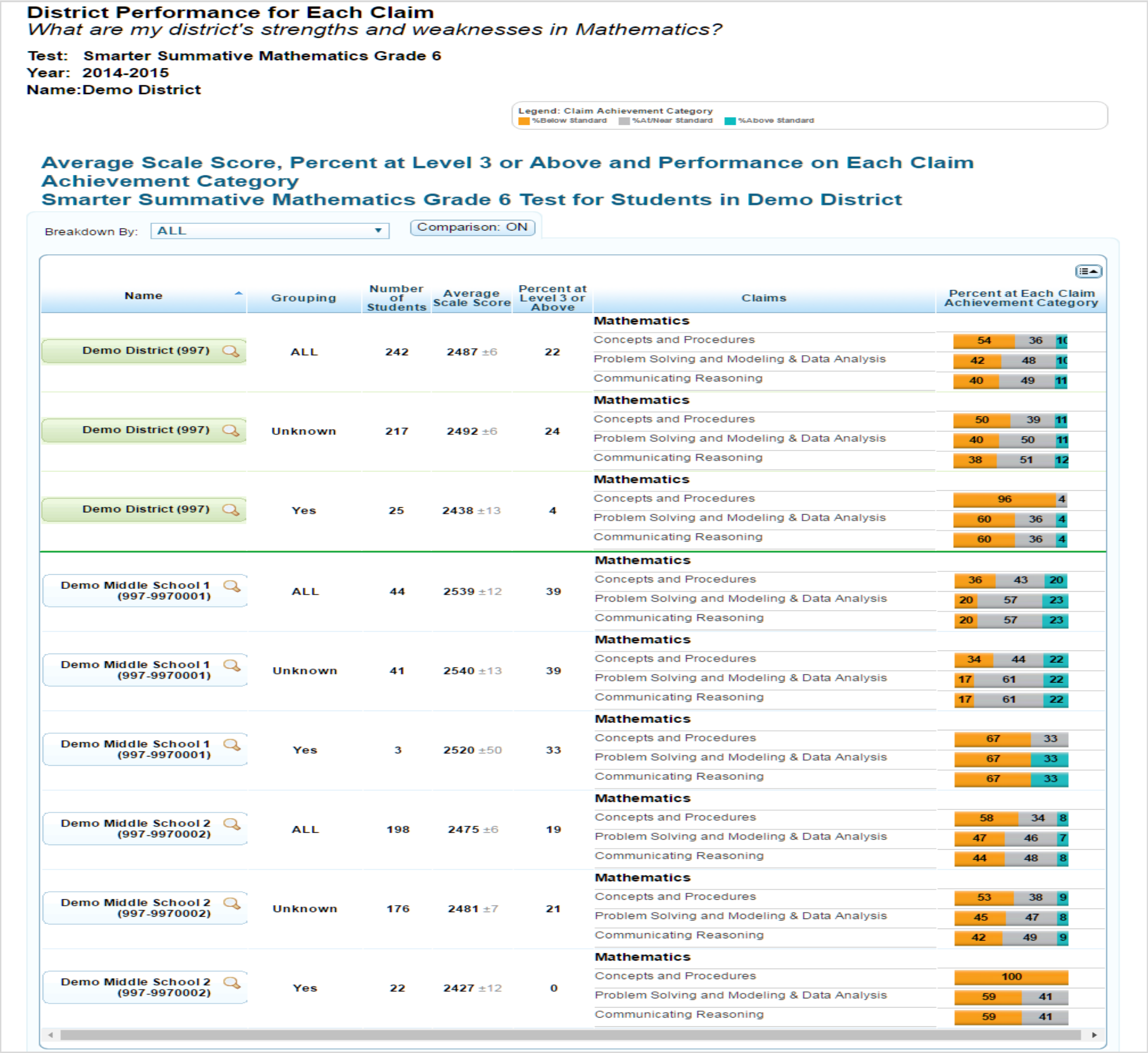
Name	Grouping	Number of Students	Average Scale Score	Percent at Level 3 or Above	Percentage in Each Achievement Level
Demo District (997)	ALL	242	2515 ±6	45	25 31 33 12
Demo District (997)	Female	111	2530 ±8	50	18 32 35 15
Demo District (997)	Male	131	2503 ±8	40	31 30 31 8
Demo Middle School 1 (997-9970001)	ALL	44	2576 ±11	75	7 18 45 30
Demo Middle School 1 (997-9970001)	Female	24	2589 ±15	83	8 8 46 38
Demo Middle School 1 (997-9970001)	Male	20	2561 ±16	65	5 30 45 20
Demo Middle School 2 (997-9970002)	ALL	198	2502 ±6	38	29 33 30 8
Demo Middle School 2 (997-9970002)	Female	87	2513 ±9	41	21 38 32 9
Demo Middle School 2 (997-9970002)	Male	111	2492 ±8	35	35 30 29 6

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, and (3) percent of students at Level 3 or above, and (4) percent of students in each achievement 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 Mathematics?

Test: Smarter Summative Mathematics Grade 6
Year: 2014-2015
Name: Demo District

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 Mathematics Grade 6 Test for Students in Demo District

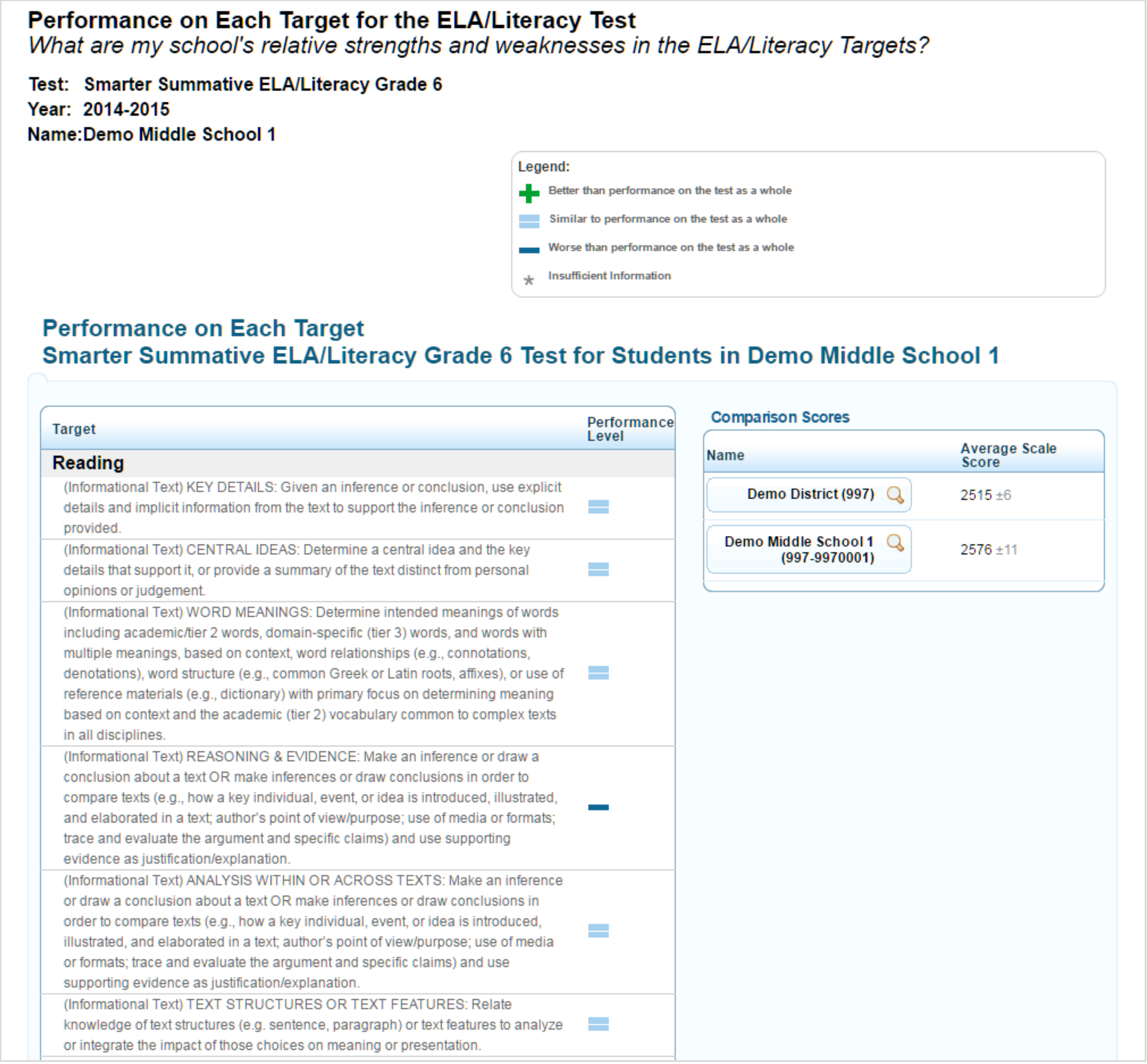
Breakdown By: ALL
Comparison: ON

Name	Grouping	Number of Students	Average Scale Score	Percent at Level 3 or Above	Claims	Percent at Each Claim Achievement Category
Demo District (997)	ALL	242	2487 ±6	22	**Mathematics**	
					Concepts and Procedures	54 36 10
					Problem Solving and Modeling & Data Analysis	42 48 10
					Communicating Reasoning	40 49 11
Demo District (997)	Unknown	217	2492 ±6	24	**Mathematics**	
					Concepts and Procedures	50 39 11
					Problem Solving and Modeling & Data Analysis	40 50 11
					Communicating Reasoning	38 51 12
Demo District (997)	Yes	25	2438 ±13	4	**Mathematics**	
					Concepts and Procedures	96 4
					Problem Solving and Modeling & Data Analysis	60 36 4
					Communicating Reasoning	60 36 4
Demo Middle School 1 (997-9970001)	ALL	44	2539 ±12	39	**Mathematics**	
					Concepts and Procedures	36 43 20
					Problem Solving and Modeling & Data Analysis	20 57 23
					Communicating Reasoning	20 57 23
Demo Middle School 1 (997-9970001)	Unknown	41	2540 ±13	39	**Mathematics**	
					Concepts and Procedures	34 44 22
					Problem Solving and Modeling & Data Analysis	17 61 22
					Communicating Reasoning	17 61 22
Demo Middle School 1 (997-9970001)	Yes	3	2520 ±50	33	**Mathematics**	
					Concepts and Procedures	67 33
					Problem Solving and Modeling & Data Analysis	67 33
					Communicating Reasoning	67 33
Demo Middle School 2 (997-9970002)	ALL	198	2475 ±6	19	**Mathematics**	
					Concepts and Procedures	58 34 8
					Problem Solving and Modeling & Data Analysis	47 46 7
					Communicating Reasoning	44 48 8
Demo Middle School 2 (997-9970002)	Unknown	176	2481 ±7	21	**Mathematics**	
					Concepts and Procedures	53 38 9
					Problem Solving and Modeling & Data Analysis	45 47 8
					Communicating Reasoning	42 49 9
Demo Middle School 2 (997-9970002)	Yes	22	2427 ±12	0	**Mathematics**	
					Concepts and Procedures	100
					Problem Solving and Modeling & Data Analysis	59 41
					Communicating Reasoning	59 41

7.1.2.4 Target Detail Page

The Target Detail Page provides the aggregate summaries on student performance in each target. The Target Detail Page provides (1) strength or weakness indicators in each target, and (2) average scale scores and standard errors of average scale scores for the district, school, and teacher levels. 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 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: 2014-2015
Name:Demo Middle School 1

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

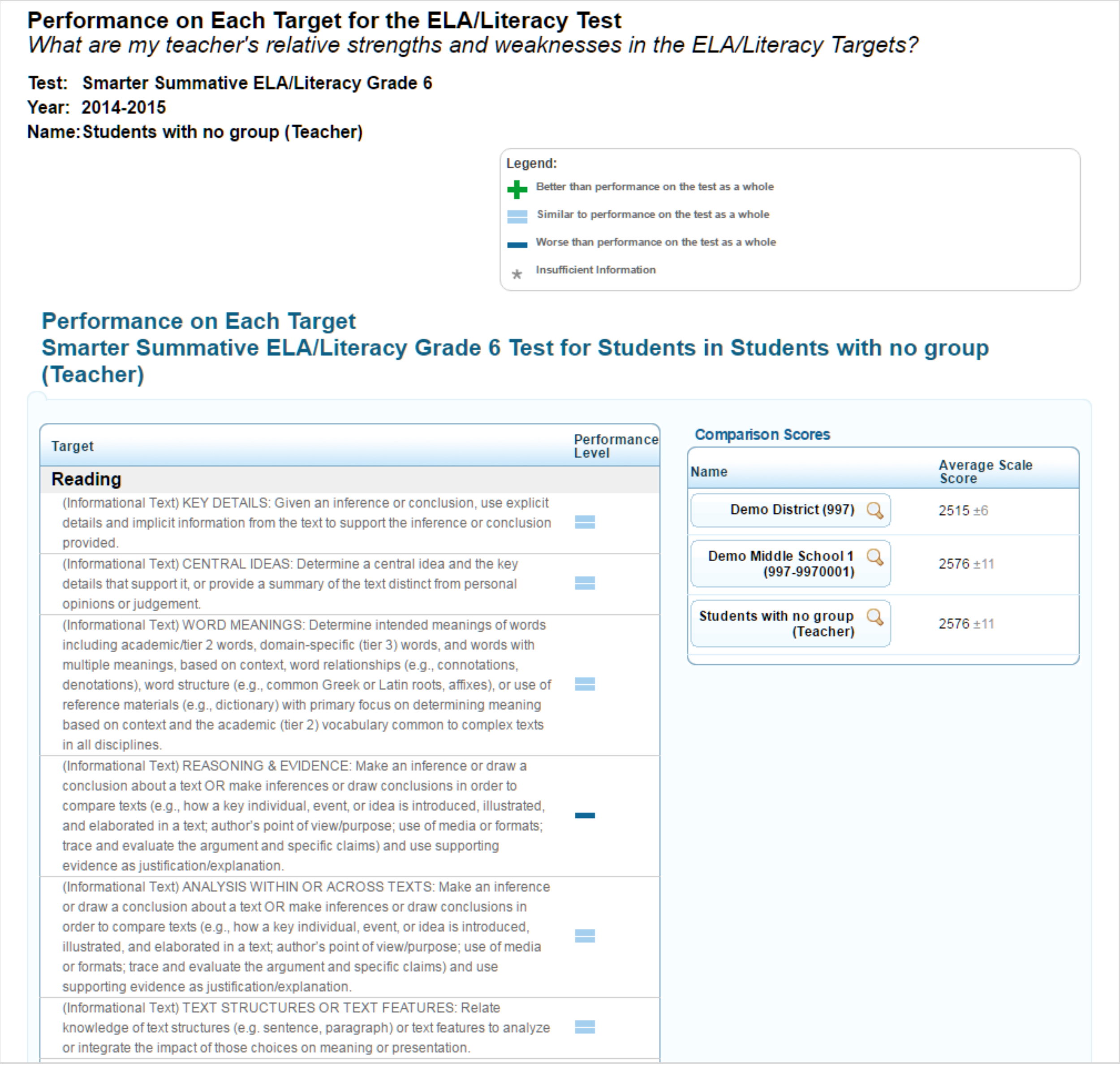
Performance on Each Target
Smarter Summative ELA/Literacy Grade 6 Test for Students in Demo Middle School 1

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.	═

Comparison Scores

Name	Average Scale Score
Demo District (997)	2515 ±6
Demo Middle School 1 (997-9970001)	2576 ±11

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: 2014-2015
Name: Students with no group (Teacher)

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

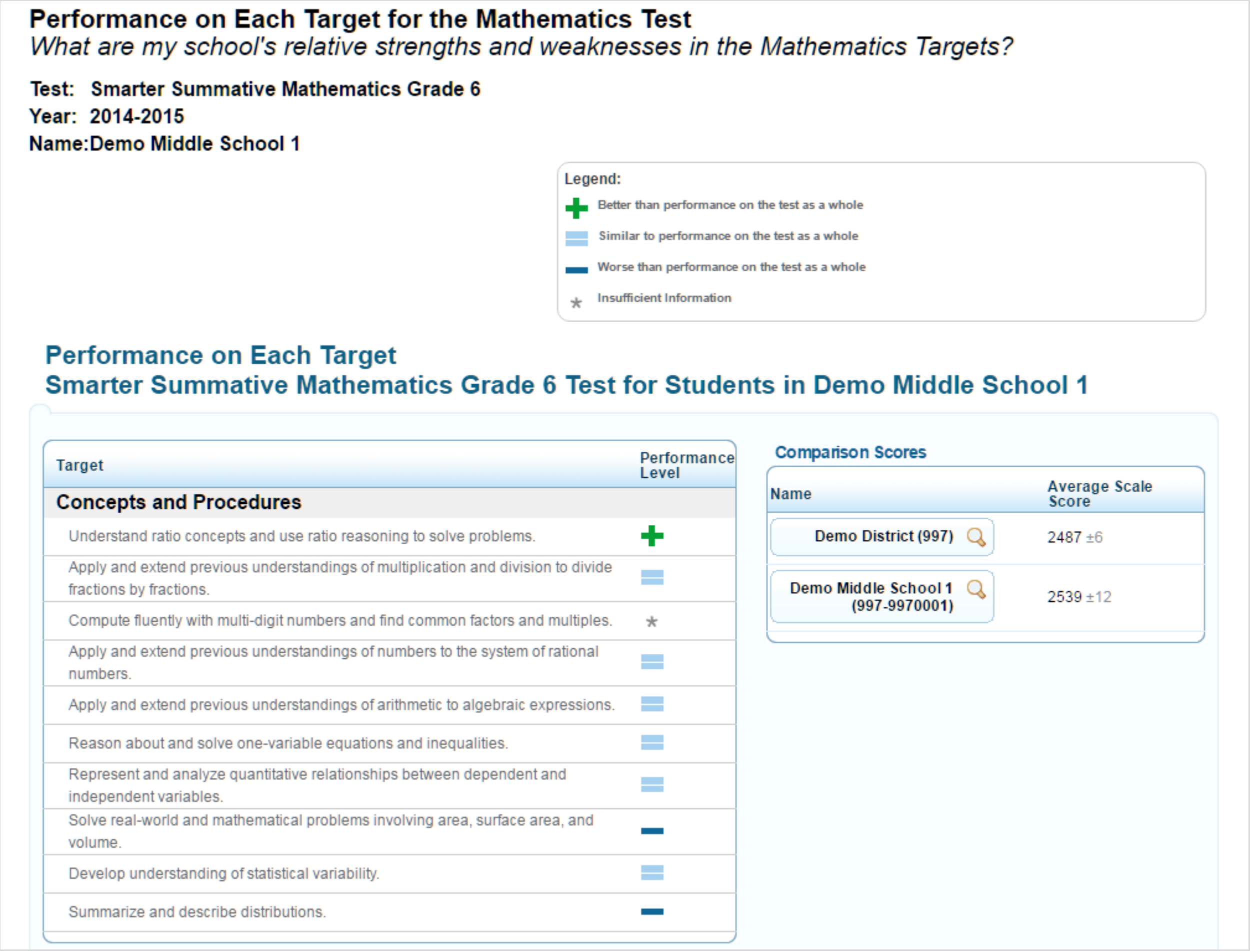
Performance on Each Target
Smarter Summative ELA/Literacy Grade 6 Test for Students in Students with no group (Teacher)

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.	=

Comparison Scores

Name	Average Scale Score
Demo District (997)	2515 ±6
Demo Middle School 1 (997-9970001)	2576 ±11
Students with no group (Teacher)	2576 ±11

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: 2014-2015
Name:Demo Middle School 1

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

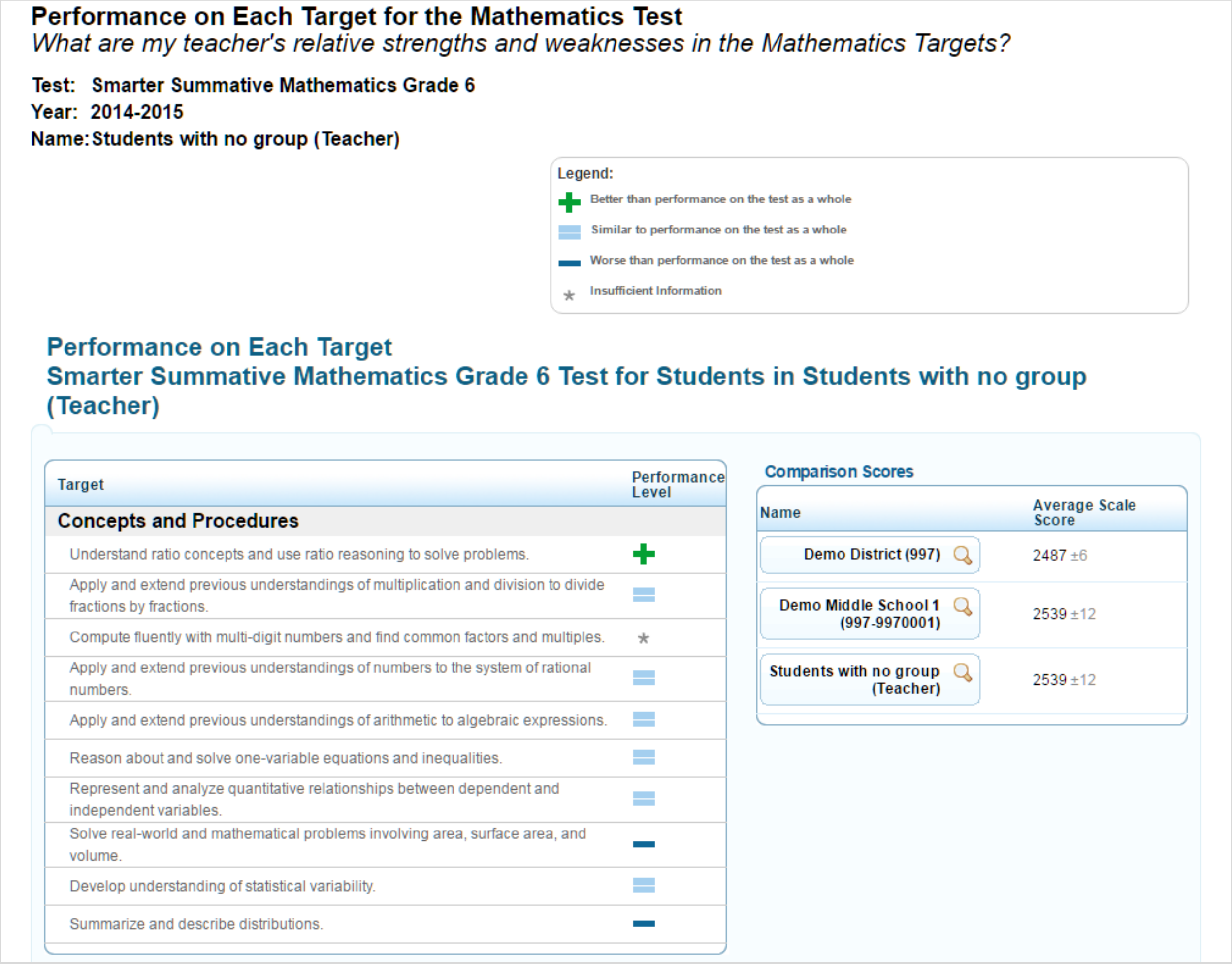
Performance on Each Target
Smarter Summative Mathematics Grade 6 Test for Students in Demo Middle School 1

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 (997)	2487 ±6
Demo Middle School 1 (997-9970001)	2539 ±12

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: 2014-2015
Name: Students with no group (Teacher)

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 Students with no group (Teacher)

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

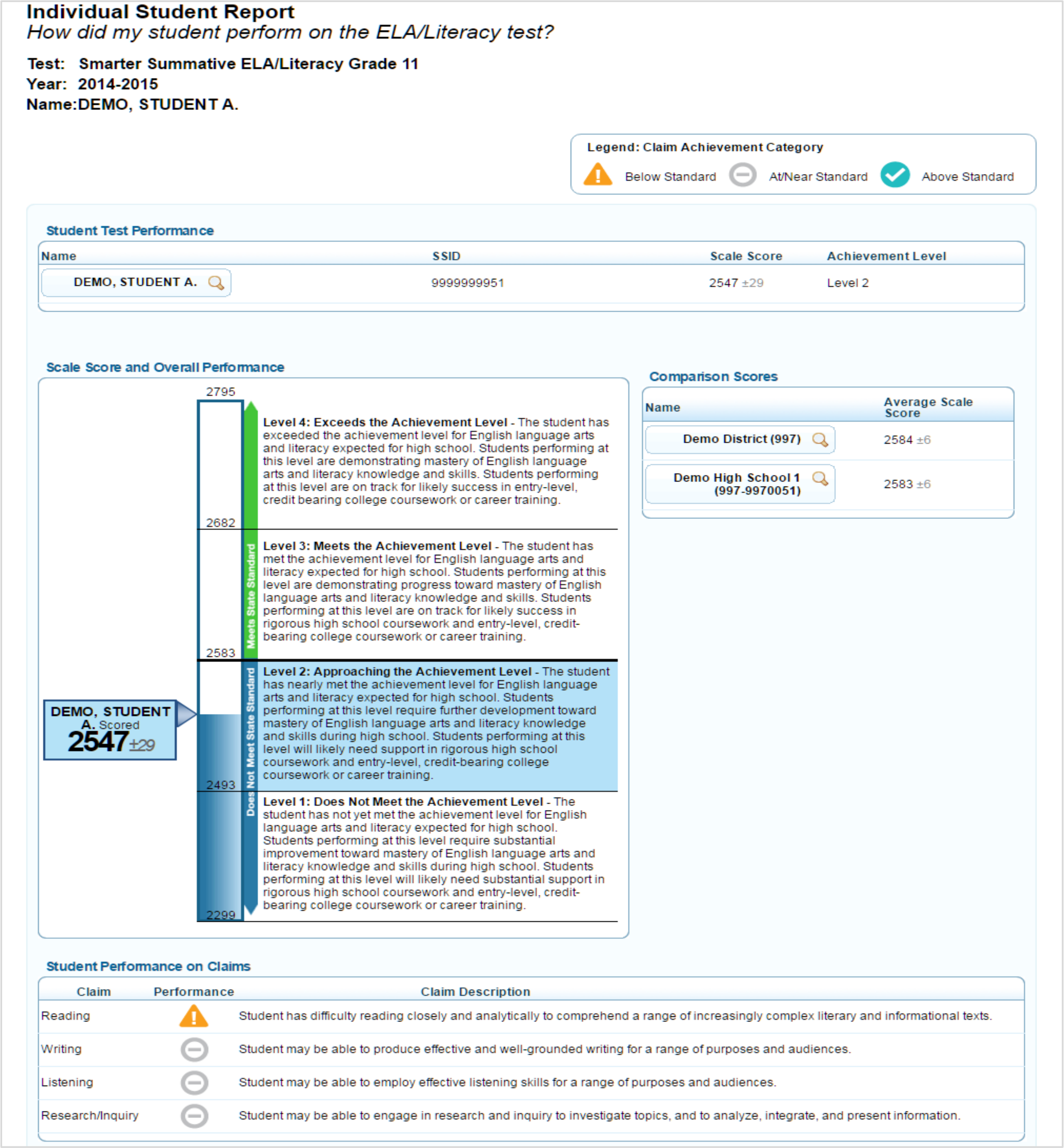
Name	Average Scale Score
Demo District (997)	2487 ±6
Demo Middle School 1 (997-9970001)	2539 ±12
Students with no group (Teacher)	2539 ±12

7.1.2.5 Student Detail Page

When a student submits a test after completing a test, an online score report appears in the Student Detail Page in 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, (2) achievement level for overall test, (3) achievement category in each claim, and (4) average scale scores for student's district, and school.

Specifically, on the top of page, student's name, scale score with standard error of measurement, and achievement level are presented. On the left middle section, student's performance are described in detail using a barrel chart. In the barrel chart, student's scale score is presented with standard error of measurement using a sign of "±." Standard error of measurement 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 examinees at the 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 standard error of measurement of the scale score whereas the ± next to the average scale scores for aggregate levels represent the standard error of the average scale scores. On the bottom of the page, student performance on claims is displayed along with a description of his/her performance on each of 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 11
Year: 2014-2015
Name:DEMO, STUDENT A.

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

Student Test Performance

Name	SSID	Scale Score	Achievement Level
DEMO, STUDENT A.	9999999951	2547 ±29	Level 2

Scale Score and Overall Performance

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

Level 3: Meets the Achievement Level - The student has met the achievement level for English language arts and literacy expected for high school. 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 rigorous high school coursework and entry-level, credit-bearing college coursework or career training.

Level 2: Approaching the Achievement Level - The student has nearly met the achievement level for English language arts and literacy expected for high school. Students performing at this level require further development toward mastery of English language arts and literacy knowledge and skills during high school. Students performing at this level will likely need support in rigorous high school coursework and entry-level, credit-bearing college coursework or career training.

Level 1: Does Not Meet the Achievement Level - The student has not yet met the achievement level for English language arts and literacy expected for high school. Students performing at this level require substantial improvement toward mastery of English language arts and literacy knowledge and skills during high school. Students performing at this level will likely need substantial support in rigorous high school coursework and entry-level, credit-bearing college coursework or career training.

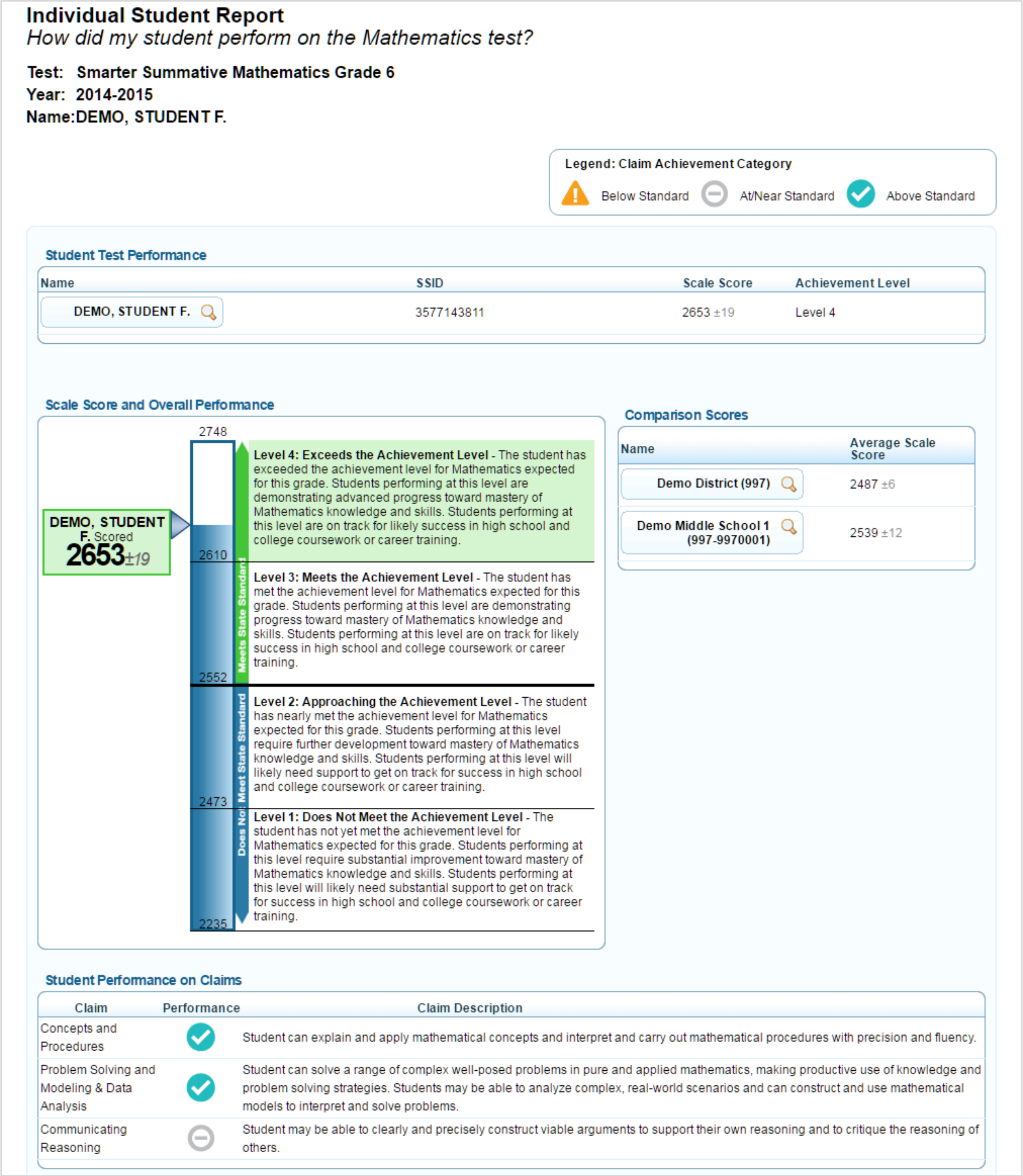
Comparison Scores

Name	Average Scale Score
Demo District (997)	2584 ±6
Demo High School 1 (997-9970051)	2583 ±6

Student Performance on Claims

Claim	Performance	Claim Description
Reading	⚠	Student has difficulty reading closely and analytically to comprehend a range of increasingly complex literary and informational texts.
Writing	⊖	Student may be able to produce effective and well-grounded writing for a range of purposes and audiences.
Listening	⊖	Student may be able to employ effective listening skills for a range of purposes and audiences.
Research/Inquiry	⊖	Student may be able to 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: 2014-2015
Name:DEMO, STUDENT F.

Legend: Claim Achievement Category

⚠ Below Standard ⊖ At/Near Standard ✓ Above Standard

Student Test Performance

Name	SSID	Scale Score	Achievement Level
DEMO, STUDENT F.	3577143811	2653 ±19	Level 4

Scale Score and Overall Performance

DEMO, STUDENT F. Scored 2653 ±19

2748

Level 4: Exceeds the Achievement Level - 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.

2610

Level 3: Meets the Achievement Level - 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.

2552

Level 2: Approaching the Achievement Level - 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.

2473

Level 1: Does Not Meet the Achievement Level - 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.

2235

Comparison Scores

Name	Average Scale Score
Demo District (997)	2487 ±6
Demo Middle School 1 (997-9970001)	2539 ±12

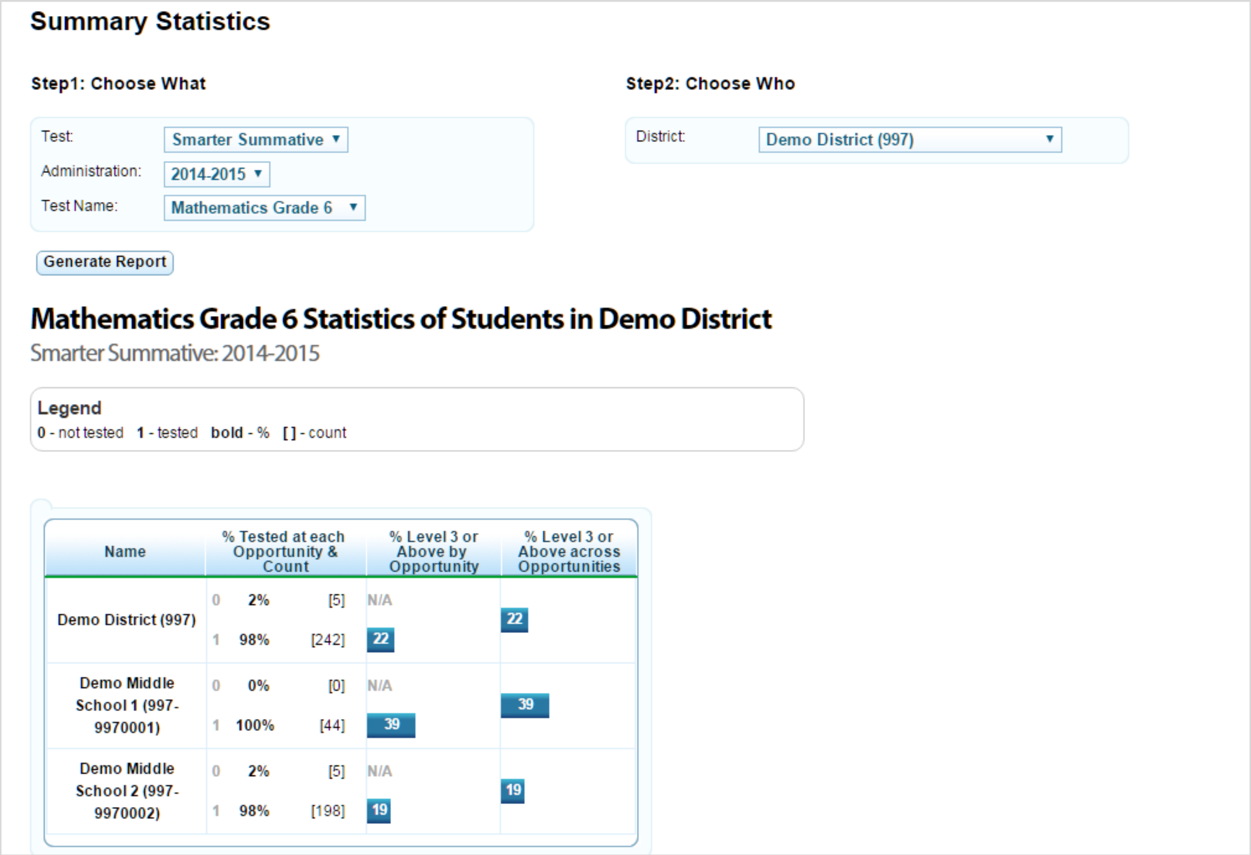
Student Performance on Claims

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. Students may be able to analyze complex, real-world scenarios and can construct and use mathematical models to interpret and solve problems.
Communicating Reasoning	⊖	Student may be able to 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 report, ORS provides participation rate reports for the district and school to help monitor student participation rate. Participation data are up to date each time students complete tests and they are handscored. Included in the participation table are (1) number and percent of students who are tested and not tested and (2) percent of students with achievement levels = 3 or 4. Exhibit 10 presents a sampled participation rate report at a district level.

Exhibit 10. Participation Rate Report at District Level



Summary Statistics

Step1: Choose What

Test: Smarter Summative ▼
Administration: 2014-2015 ▼
Test Name: Mathematics Grade 6 ▼

Generate Report

Step2: Choose Who

District: Demo District (997) ▼

Mathematics Grade 6 Statistics of Students in Demo District

Smarter Summative: 2014-2015

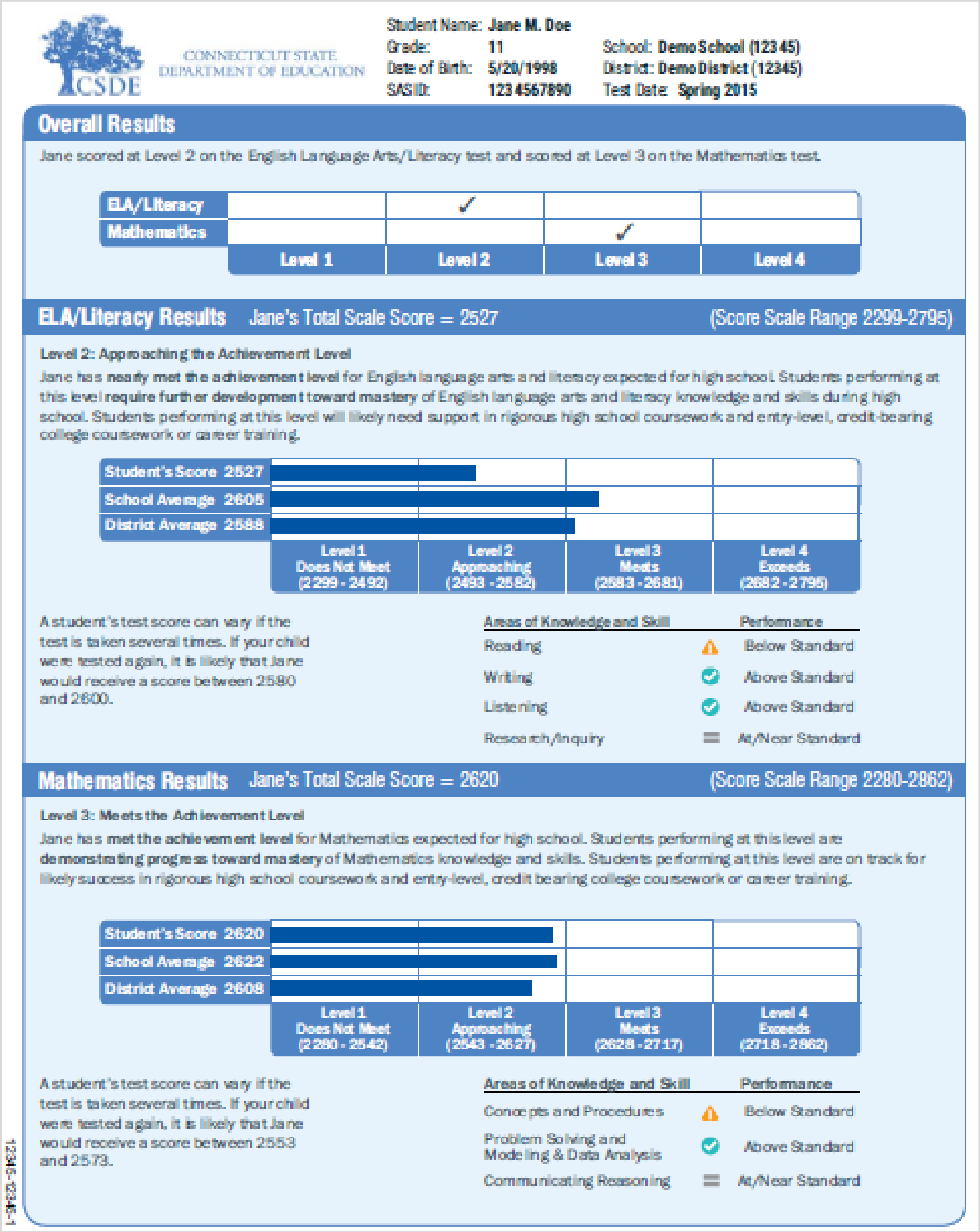
Legend
0 - not tested **1** - tested **bold** - % **[ ]** - count

Name	% Tested at each Opportunity & Count	% Level 3 or Above by Opportunity	% Level 3 or Above across Opportunities
Demo District (997)	0 **2%** [5] 1 **98%** [242]	N/A 22	22
Demo Middle School 1 (997-9970001)	0 **0%** [0] 1 **100%** [44]	N/A 39	39
Demo Middle School 2 (997-9970002)	0 **2%** [5] 1 **98%** [198]	N/A 19	19

7.2 PAPER FAMILY SCORE REPORTS

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

Exhibit 11. Sample Paper Family Score Report



CONNECTICUT STATE DEPARTMENT OF EDUCATION

CSDE

Student Name:	Jane M. Doe		
Grade:	11	School:	Demo School (12345)
Date of Birth:	5/20/1998	District:	Demo District (12345)
SAS ID:	1234567890	Test Date:	Spring 2015

Overall Results

Jane scored at Level 2 on the English Language Arts/Literacy test and scored at Level 3 on the Mathematics test.

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

ELA/Literacy Results Jane's Total Scale Score = 2527 (Score Scale Range 2299-2795)

Level 2: Approaching the Achievement Level

Jane has **nearly met the achievement level** for English language arts and literacy expected for high school. Students performing at this level **require further development toward mastery** of English language arts and literacy knowledge and skills during high school. Students performing at this level will likely need support in rigorous high school coursework and entry-level, credit-bearing college coursework or career training.

Student's Score 2527

School Average 2605

District Average 2588

Level 1 Does Not Meet (2299 - 2492)	Level 2 Approaching (2493 - 2582)	Level 3 Meets (2583 - 2681)	Level 4 Exceeds (2682 - 2795)

A student's test score can vary if the test is taken several times. If your child were tested again, it is likely that Jane would receive a score between 2580 and 2600.

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

Mathematics Results Jane's Total Scale Score = 2620 (Score Scale Range 2280-2862)

Level 3: Meets the Achievement Level

Jane has **met the achievement level** for Mathematics expected for high school. 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 rigorous high school coursework and entry-level, credit bearing college coursework or career training.

Student's Score 2620

School Average 2622

District Average 2608

Level 1 Does Not Meet (2280 - 2542)	Level 2 Approaching (2543 - 2627)	Level 3 Meets (2628 - 2717)	Level 4 Exceeds (2718 - 2862)

A student's test score can vary if the test is taken several times. If your child were tested again, it is likely that Jane would receive a score between 2553 and 2573.

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

12346-12346-1

7.3 INTERPRETATION OF REPORTED SCORES

A student's performance on a test is reported in a scale score and an achievement level for the overall test, and an achievement level for each claim. Students' scores and achievement levels are summarized at the aggregate levels. The next section provides a description about 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 students' knowledge and skills measured. The scale score is the transformed score from a theta score which is estimated based on mathematical models. Low scale scores can be interpreted that the student does not possess sufficient knowledge and skills measured by the test. Conversely, high scale scores can be interpreted that the student has proficient knowledge and skills measured by the test. Scale scores can be used to measure student growth across school years. Interpretation of scale scores is more meaningful when the scale scores are used along with achievement levels and achievement-level descriptors.

7.3.2 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 standard error of measurement (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 ± next to the student's scale score provides information about the certainty, or confidence, of the score's interpretation. The boundaries of the score band are one SEM above and below the student's observed scale score, representing a range of score values that is likely to contain the true score. For example, 2680 ± 10 indicates that if a student was tested again, it is likely that the student would receive a score between 2670 and 2690. 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 students fall into based on their scale scores. For Smarter Balanced Assessments, scale scores are mapped into four achievement levels (i.e., Level 1, Level 2, Level 3, Level 4) using three achievement standards (i.e., cut scores). Achievement-level descriptors that are a description of content area knowledge and skills that examinees at the each achievement level are expected to possess. Thus achievement levels can be interpreted based on achievement-level descriptors. For the achievement level at Level 3 in ELA/L, for instance, achievement-level descriptors are described for Level 3 as "students demonstrate progress toward mastery of the knowledge and skills ELA/L needed for likely success in future coursework." Generally, students performing Smarter Balanced test at Levels 3 and 4 are considered on track to demonstrate progress toward mastery of the knowledge and skills necessary for college and career readiness.

7.3.4 Achievement Category for Claims

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

7.3.5 Achievement 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., class, 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 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 perform on a test. When student's 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 for overall and by claim are reported at the aggregate level to represent how well a group of students perform for overall and by claim.

7.4 APPROPRIATE USES FOR SCORES AND REPORTS

Assessment results can be used to provide information on individual students' 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, achievement 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 for teacher and school level provide information the strengths and weaknesses of their students and can be utilized to improve teaching and student learning. For example, a group of students performed very well in 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 subgroup. 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 district for overall and by claim. Although all students are administered different sets of items in each computer-adaptive test, scale scores are comparable across students. Furthermore, scale scores can be used to measure the growth of individual students over time 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 need to consider measurement error when using student scores to make decisions about student achievement. Moreover, although student scores may be used to help make important 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 need to take into account 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 test 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.

8.1 ADAPTIVE TEST CONFIGURATION

For the computer-adaptive testing, 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., cut scores, answer keys, item attributes, item parameters, passage information). The accuracy of the information in the configuration file is checked and confirmed numerous times independently by multiple staff members prior to 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 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 very 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 Test Delivery System 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 behind a different platform, look at the same item to see that it renders as expected.

8.1.2 User Acceptance Testing and Final Review

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

8.2 QUALITY ASSURANCE IN DOCUMENT PROCESSING

Scanning Accuracy

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

8.3 QUALITY ASSURANCE IN DATA PREPARATION

AIR's Test Delivery System 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, total number of field-test items and operation items, and that the test record contains no data from items that have been invalidated

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

8.4 QUALITY ASSURANCE IN HANDSCORING

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

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

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 Smarter Balanced. MI periodically administers validity sets to each of MI's scorers working on the scoring effort. VSC is capable of dynamically embedding calibration responses in scoring sets as individual items or in sets of whatever number of items is preferred by the state.

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

8.4.2 Handscoring QA Monitoring Reports

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

8.4.3 Monitoring by State Department of Education

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 examinee. 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. Applications log not only errors and exceptions, but latency (timing) 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, latency data is captured for each assessed student—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, 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 adaptive test component, 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 38 presents an overview of the quality assurance (QA) reports.

Table 38. Overview of Quality Assurance Reports

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

8.5.1 Score Report Quality Check

In the 2014–2015 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 item response theory [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 human-scoring processes include rigorous training, validity and reliability monitoring, and back-reading to ensure accurate scoring. Handscored items are married up 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 prior to 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 there is only one place where the “official” record is stored. Only after scores have passed the QA checks and are uploaded to the DoR are they passed to the Online Reporting System, which is responsible for presenting individual-level results and calculating and presenting aggregate results. Absolutely no score is reported in the Online Reporting System until it passes all of the QA system’s validation checks. All of the above processes take milliseconds to complete so that within less than a second of handscores being received by

AIR and passing QA validation checks, the composite score will be available in the Online Reporting System.

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 read in and verify the data and conversion tables and the macros that do the many district 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. Prior to printing the reports, AIR provides a live data file and individual student reports with sample districts for Department staff review. AIR will work closely with the department to resolve questions and correct any problems. The reports will not be delivered unless the department approves the sample reports and data file.

9. 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. *British Journal of Mathematical and Statistical Psychology, 37,* 1-21.

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.

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.

APPENDIXES

Appendix A: Percentage of Students in Achievement Levels for Overall and by Subgroups

Table A-1. School Year 2014-2015 Grade 3 ELA/L Percentage of Students in Achievement Levels for Overall and by Subgroups

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
All Students	37,987	2,436.18	87.90	23	23	24	30	54
Gender								
Female	18,577	2,446.85	85.59	19	23	25	33	58
Male	19,410	2,425.97	88.87	27	24	23	26	49
Ethnicity								
American Indian or Alaska Native	109	2,410.33	79.97	30	29	25	16	40
Asian	1,917	2,479.35	83.83	11	16	25	49	73
African American	4,922	2,386.02	79.16	43	29	18	10	28
Hispanic	8,995	2,390.32	79.93	40	29	19	12	31
White	20,815	2,463.69	79.32	12	20	27	40	68
Multiple	1,197	2,441.98	87.40	21	25	23	31	55
Limited English Proficiency								
LEP	2,852	2,353.96	68.03	58	29	11	2	13
IDEA								
IDEA Eligible	4,363	2,349.34	77.78	63	21	10	5	16

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

Table A-2. School Year 2014-2015 Grade 4 ELA/L Percentage of Students in Achievement Levels for Overall and by Subgroups

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
All Students	38,597	2,478.61	92.53	26	19	24	31	55
Gender								
Female	19,065	2,490.87	90.20	21	18	25	35	60
Male	19,532	2,466.65	93.21	30	20	23	26	50
Ethnicity								
American Indian or Alaska Native	113	2,453.53	85.62	33	25	22	20	42
Asian	1,969	2,525.29	83.61	11	14	26	49	75
African American	4,778	2,423.93	84.13	48	23	18	11	29
Hispanic	8,770	2,428.97	86.56	45	23	20	12	32
White	21,936	2,505.83	82.95	15	17	27	41	68
Multiple	991	2,489.16	92.13	21	21	24	33	57
Limited English Proficiency								
LEP	2,692	2,388.80	75.77	65	21	12	3	14
IDEA								
IDEA Eligible	4,695	2,383.59	80.42	67	18	11	4	15

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

Table A-3. School Year 2014-2015 Grade 5 ELA/L Percentage of Students in Achievement Levels for Overall and by Subgroups

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
All Students	38,817	2,515.54	92.08	23	19	33	26	59
Gender								
Female	18,884	2,529.03	89.62	18	18	34	31	64
Male	19,933	2,502.77	92.56	27	20	32	21	53
Ethnicity								
American Indian or Alaska Native	96	2,495.81	79.68	27	27	31	15	46
Asian	1,996	2,559.41	85.49	10	14	32	43	76
African American	4,876	2,460.31	84.67	43	24	25	8	33
Hispanic	8,382	2,465.44	86.45	41	24	25	10	35
White	22,476	2,542.19	82.39	13	16	37	34	71
Multiple	962	2,520.25	89.98	20	20	34	27	60
Limited English Proficiency								
LEP	2,351	2,417.50	70.25	64	24	11	1	12
IDEA								
IDEA Eligible	4,955	2,417.72	80.66	65	19	12	3	16

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

Table A-4. School Year 2014-2015 Grade 6 ELA/L Percentage of Students in Achievement Levels for Overall and by Subgroups

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
All Students	39,710	2537.81	91.55	19	25	35	21	56
Gender								
Female	19,307	2552.36	87.65	14	24	37	25	62
Male	20,403	2524.04	93.03	24	26	33	17	49
Ethnicity								
American Indian or Alaska Native	119	2514.70	83.77	25	28	35	12	47
Asian	1,959	2589.66	81.60	7	14	38	40	78
African American	4,833	2485.09	83.27	37	33	24	6	30
Hispanic	8,454	2486.67	88.41	37	31	25	7	32
White	23,295	2562.73	81.61	10	22	41	27	67
Multiple	1,009	2544.78	91.62	18	24	37	22	59
Limited English Proficiency								
LEP	2,047	2428.31	74.13	64	28	7	0	8
IDEA								
IDEA Eligible	5,042	2440.96	80.97	59	27	13	2	14

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

Table A-5. School Year 2014-2015 Grade 7 ELA/L Percentage of Students in Achievement Levels for Overall and by Subgroups

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
All Students	38,782	2,560.04	95.24	21	22	39	18	57
Gender								
Female	18,838	2,575.72	91.06	16	20	42	22	64
Male	19,944	2,545.23	96.72	26	24	36	15	50
Ethnicity								
American Indian or Alaska Native	87	2,530.74	83.43	32	29	32	7	39
Asian	1,876	2,612.76	86.76	8	12	42	37	79
African American	5,001	2,506.62	87.90	39	29	26	6	32
Hispanic	8,082	2,506.80	92.02	39	27	28	6	34
White	22,837	2,586.07	84.89	12	19	45	24	69
Multiple	875	2,567.42	91.02	18	22	42	19	60
Limited English Proficiency								
LEP	1,827	2,438.53	70.09	73	20	7	0	7
IDEA								
IDEA Eligible	4,948	2,457.15	80.03	63	24	12	1	13

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

Table A-6. School Year 2014-2015 Grade 8 ELA/L Percentage of Students in Achievement Levels for Overall and by Subgroups

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
All Students	39,610	2,572.14	95.72	20	26	37	17	54
Gender								
Female	19,223	2,589.42	91.74	14	24	40	21	62
Male	20,387	2,555.85	96.54	25	28	34	13	47
Ethnicity								
American Indian or Alaska Native	106	2,540.55	91.62	28	28	35	8	43
Asian	1,752	2,624.49	87.90	7	16	43	33	76
African American	5,067	2,518.97	85.63	36	35	24	5	29
Hispanic	8,059	2,520.09	90.80	37	31	26	6	31
White	23,740	2,597.13	87.29	12	23	43	22	65
Multiple	850	2,581.11	95.79	18	25	38	19	57
Limited English Proficiency								
LEP	1,723	2,449.55	68.14	71	24	5	0	5
IDEA								
IDEA Eligible	4,941	2,473.16	81.24	59	28	11	2	13

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

Table A-7. School Year 2014-2015 Grade 11 ELA/L Percentage of Students in Achievement Levels for Overall and by Subgroups

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
All Students	32,487	2,583.82	111.44	22	25	32	21	53
Gender								
Female	15,869	2,601.97	105.37	17	23	35	25	60
Male	16,618	2,566.50	114.29	28	26	29	18	47
Ethnicity								
American Indian or Alaska Native	101	2,569.26	109.61	27	21	40	13	52
Asian	1,473	2,627.13	110.49	14	17	34	35	70
African American	4,107	2,529.23	100.39	38	31	24	7	31
Hispanic	6,008	2,537.11	103.65	35	31	26	8	34
White	20,171	2,605.85	107.53	16	22	35	27	62
Multiple	600	2,582.18	108.56	22	27	31	20	51
Limited English Proficiency								
LEP	1,260	2,458.97	79.18	68	26	5	1	6
IDEA								
IDEA Eligible	3,463	2,487.38	96.89	56	27	13	4	17

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

Table A-8. School Year 2014-2015 Grade 3 Math Percentage of Students in Achievement Levels for Overall and by Subgroups

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
All Students	38,249	2,427.30	80.21	27	25	30	18	48
Gender								
Female	18,701	2,426.48	76.66	27	26	30	17	47
Male	19,548	2,428.09	83.46	28	24	29	20	49
Ethnicity								
American Indian or Alaska Native	111	2,406.28	85.01	40	24	24	12	36
Asian	1,961	2,476.81	80.44	12	17	31	40	71
African American	4,943	2,378.55	71.48	51	27	17	4	21
Hispanic	9,176	2,384.77	72.99	47	29	18	5	24
White	20,829	2,452.75	70.67	14	23	37	25	62
Multiple	1,197	2,432.98	79.37	27	24	28	21	49
Limited English Proficiency								
LEP	3,117	2,358.46	67.81	63	25	10	2	11
IDEA								
IDEA Eligible	4,384	2,350.19	79.67	65	20	11	3	15

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

Table A-9. School Year 2014-2015 Grade 4 Math Percentage of Students in Achievement Levels for Overall and by Subgroups

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
All Students	38,829	2,469.93	80.10	23	33	27	17	44
Gender								
Female	19,180	2,468.72	76.08	23	34	28	15	43
Male	19,649	2,471.10	83.83	23	31	27	18	45
Ethnicity								
American Indian or Alaska Native	115	2,451.55	74.18	26	40	23	11	34
Asian	2,002	2,523.28	78.81	8	21	32	39	70
African American	4,783	2,418.59	69.60	46	37	13	4	17
Hispanic	8,929	2,425.89	72.43	42	37	16	5	21
White	21,971	2,493.78	70.97	12	31	35	22	57
Multiple	988	2,480.07	82.98	20	34	24	22	46
Limited English Proficiency								
LEP	2,942	2,399.98	69.60	57	32	9	2	11
IDEA								
IDEA Eligible	4,695	2,392.17	75.78	62	26	8	3	11

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

Table A-10. School Year 2014-2015 Grade 5 Math Percentage of Students in Achievement Levels for Overall and by Subgroups

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
All Students	39,044	2,493.22	87.24	33	30	19	18	37
Gender								
Female	18,980	2,491.73	83.06	34	32	19	16	35
Male	20,064	2,494.63	91.00	33	28	19	19	38
Ethnicity								
American Indian or Alaska Native	96	2,467.95	69.25	50	30	14	6	20
Asian	2,019	2,547.04	86.53	15	24	21	39	60
African American	4,889	2,434.12	75.19	62	27	8	3	11
Hispanic	8,550	2,443.70	78.26	57	28	10	5	15
White	22,499	2,519.96	77.23	20	32	25	24	49
Multiple	961	2,497.73	86.07	32	34	15	20	35
Limited English Proficiency								
LEP	2,586	2,410.41	69.92	76	18	4	1	5
IDEA								
IDEA Eligible	4,958	2,409.31	77.10	74	18	5	2	7

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

Table A-11. School Year 2014-2015 Grade 6 Math Percentage of Students in Achievement Levels for Overall and by Subgroups

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
All Students	39,870	2,513.31	99.72	32	31	21	16	37
Gender								
Female	19,372	2,515.99	94.26	30	32	22	15	37
Male	20,498	2,510.77	104.56	33	30	20	17	37
Ethnicity								
American Indian or Alaska Native	121	2,482.65	92.22	37	42	14	7	21
Asian	1,979	2,583.50	95.48	12	23	24	41	65
African American	4,841	2,448.65	88.06	59	29	9	3	12
Hispanic	8,577	2,455.50	94.95	55	30	11	4	15
White	23,299	2,541.90	86.46	19	33	27	21	48
Multiple	1,013	2,519.61	99.55	31	30	22	17	39
Limited English Proficiency								
LEP	2,230	2,401.61	88.19	80	16	3	1	4
IDEA								
IDEA Eligible	5,042	2,407.94	95.06	75	18	5	2	7

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

Table A-12. School Year 2014-2015 Grade 7 Math Percentage of Students in Achievement Levels for Overall and by Subgroups

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
All Students	39,001	2,530.01	105.91	32	30	22	17	39
Gender								
Female	18,952	2,531.99	100.75	31	31	23	15	38
Male	20,049	2,528.15	110.54	33	28	21	18	39
Ethnicity								
American Indian or Alaska Native	88	2,490.67	91.60	49	33	13	6	18
Asian	1,901	2,604.60	101.41	12	20	26	42	68
African American	5,026	2,465.56	94.29	57	29	10	4	14
Hispanic	8,270	2,467.78	97.78	56	29	11	4	16
White	22,816	2,560.43	93.42	19	31	28	22	50
Multiple	875	2,537.32	103.40	27	32	21	19	40
Limited English Proficiency								
LEP	2,053	2,411.91	86.82	81	14	3	1	4
IDEA								
IDEA Eligible	4,957	2,421.08	93.44	75	18	5	2	7

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

Table A-13. School Year 2014-2015 Grade 8 Math Percentage of Students in Achievement Levels for Overall and by Subgroups

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
All Students	39,764	2,541.01	114.32	37	26	19	18	37
Gender								
Female	19,237	2,546.18	108.13	35	28	20	18	38
Male	20,429	2,536.44	119.53	40	24	17	18	36
Ethnicity								
American Indian or Alaska Native	105	2,505.03	101.67	54	23	15	8	23
Asian	1,788	2,621.24	112.76	16	20	23	42	64
African American	5,058	2,467.98	94.41	66	23	8	3	12
Hispanic	8,166	2,476.29	101.92	61	24	10	5	15
White	23,669	2,573.25	103.90	25	28	24	24	48
Multiple	843	2,543.55	112.24	37	28	18	18	35
Limited English Proficiency								
LEP	1,917	2,416.96	89.56	85	11	2	2	4
IDEA								
IDEA Eligible	4,848	2,429.05	93.92	80	14	4	2	6

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

Table A-14. School Year 2014-2015 Grade 11 Math Percentage of Students in Achievement Levels for Overall and by Subgroups

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
All Students	32,288	2,556.93	127.64	47	23	19	12	30
Gender								
Female	15,771	2,564.28	120.01	43	25	21	11	31
Male	16,517	2,549.92	134.16	49	21	17	12	29
Ethnicity								
American Indian or Alaska Native	104	2,524.46	109.13	58	23	17	2	19
Asian	1,473	2,639.15	125.86	22	21	28	30	58
African American	4,074	2,482.51	103.55	72	19	8	1	9
Hispanic	6,009	2,492.75	107.14	70	19	8	3	11
White	20,007	2,585.88	123.35	36	25	24	15	39
Multiple	596	2,545.47	129.62	51	24	14	11	25
Limited English Proficiency								
LEP	1,307	2,446.67	99.21	86	9	3	2	5
IDEA								
IDEA Eligible	3,429	2,447.34	100.62	84	11	4	1	5

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

Appendix B: Number of Students for Interim Assessments

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

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

Grade	English Language Arts/Literacy			Mathematics		
	Once	Twice	Total	Once	Twice	Total
3	0	0	0	0	0	0
4	0	0	0	1	0	1
5	0	0	0	7	0	7
6	0	0	0	1	0	1
7	0	0	0	1	0	1
8	1	0	1	11	0	11
11	2	0	2	0	0	0

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 B–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 846 students took IABs, and among 846 students, 645 students took one IAB, 118 students took two IABs, and so on.

Tables B–3 and B–4 disaggregated the number of students in Table B-2 by seven IABs in ELA/L and four IABs in mathematics. For example, 645 students in grade 3 ELA/L took one IAB only. Among 645 students, two students took the Brief Writes IAB.

Table B–2. Number of Students Who Took IABs

Grade	Total	Number of IABs Taken						
		1	2	3	4	5	6	7
English Language Arts/Literacy								
3	846	645	118	75	5	2	1	
4	776	450	267	40	12	6	1	
5	461	354	103	4				
6	304	131	60	111	2			
7	377	198	92	69	18			
8	641	478	151	12				
11	302	293	5	4				
Mathematics								
3	1,108	738	195	174	1			
4	1,168	663	347	158				
5	872	663	166	43				
6	310	95	209	6				
7	394	259	130	5				
8	845	570	240	35				
11	692	537	153	2				

Table B–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	2	4	2	4	1	1	
	Editing and Revising	281	107	74	3	2	1	
	Listening and Interpretation	55	77	72	4	1	1	
	Performance Task	3		1	1	1	1	
	Reading Informational Text	132	15	3	3	2	1	
	Reading Literary Text	52	4	5	1	1		
	Research	120	29	68	4	2	1	
4	Brief Writes							
	Editing and Revising	310	217	40	12	6	1	
	Listening and Interpretation	34	141	38	12	6	1	
	Performance Task	5					1	
	Reading Informational Text	33	58	13	11	6	1	
	Reading Literary Text	51	45	1	1	6	1	
	Research	17	73	28	12	6	1	
5	Brief Writes							
	Editing and Revising	172	61	4				
	Listening and Interpretation	43	59	4				
	Performance Task	6						
	Reading Informational Text	70	44					
	Reading Literary Text	10	39					
	Research	53	3	4				
6	Brief Writes							
	Editing and Revising	72	55	111	2			
	Listening and Interpretation	15	41	109	2			
	Performance Task							
	Reading Informational Text	11	8	8	2			
	Reading Literary Text	22	12	54				
	Research	11	4	51	2			
7	Brief Writes							
	Editing and Revising							
	Listening and Interpretation	6	47	67	18			
	Performance Task	2			1			
	Reading Informational Text	9	4	1	7			
	Reading Literary Text	10	38	3	10			
	Research	63	6	68	18			
8	Brief Writes							
	Editing and Revising	379	150	12				
	Listening and Interpretation	21	20	11				
	Performance Task	26						
	Reading Informational Text	11		1				
	Reading Literary Text	2	12					
	Research	39	120	12				
11	Brief Writes							
	Editing and Revising	187	2	4				
	Listening and Interpretation	36	4	4				
	Performance Task		1					
	Reading Informational Text	10	1					
	Reading Literary Text	50						
	Research	10	2	4				

Table B–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	102	108	174	1
	Number and Operations – Fractions	305	140	174	1
	Operational and Algebraic Thinking	328	141	173	1
	Performance Task	3	1	1	1
4	Number and Operations in Base Ten	365	313	158	
	Number and Operations – Fractions	183	333	158	
	Operational and Algebraic Thinking	108	48	158	
	Performance Task	7			
5	Measurement and Data	25	62	43	
	Number and Operations in Base Ten	570	117	43	
	Number and Operations – Fractions	58	139	41	
	Performance Task	10	14	2	
6	Expressions and Equations	59	193	6	
	Geometry	18	22	6	
	Performance Task	1	10		
	Ratios and Proportional Relationships	17	193	6	
7	Expressions and Equations	85	79	5	
	The Number System	45	100	5	
	Performance Task				
	Ratios and Proportional Relationships	129	81	5	
8	Expressions and Equations	390	118	35	
	Functions	147	176	35	
	Geometry	31	185	35	
	Performance Task	2	1		
11	Algebra – Linear Functions	260	104	2	
	Algebra – Quadratic Functions	134	70	2	
	Geometry – Right Triangles and Trigonometric Ratios	138	130	2	
	Performance Task	5	2		