Delaware Smarter Balanced Assessments

2014–2015 Technical Report

Addendum to the Smarter Balanced Technical Report

Submitted to
Delaware 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/Lit) 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 (CAT) technologies 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/Lit and mathematics.

Delaware 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/Lit and mathematics and are on track for college and career readiness.

The Delaware State Board of Education formally adopted the CCSS in ELA/Lit and mathematics on Aug 19, 2010 (State Board meeting minutes, 2010). Delaware 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 2010, the Delaware Department of Education fully implemented CCSS in all grade levels in SY 2013-2014. The Delaware statewide assessments in ELA/Lit 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. Data Recognition Corporation (DRC) and 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/Lit and mathematics. 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:** An online adaptive test that provides an individualized assessment for each student.

- **Performance Task:** A task that challenges students to apply their knowledge and skills to respond to real-world problems. Performance tasks can best be described as collections of questions and activities that are coherently connected to a single theme or scenario. They are used to better measure capacities such as depth of understanding, research skills, and complex analysis, 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 districts, 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/Lit and mathematics administered in grades 3–8 and 11 under the Delaware Smarter Balanced assessments. The report includes eight chapters covering 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 Delaware data for the summative assessment only. The number of students who took the interim assessments is provided in Appendix B. While this report includes information on all aspects of the technical quality of the Smarter Balanced test administration for Delaware, 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 Delaware System of Student Assessments (DeSSA) 2014–2015 Smarter Balanced Assessment testing window spans three months for grades 3-8 and two months for grade 11 in the online summative assessments and five months for the interim assessments. The paper-and-pencil fixed forms for summative assessments were administered for 15 days during the online summative window. Table 1 shows the testing windows for both online and paper-and-pencil assessments.

Table 1. 2014–2015 Testing Windows

Tests	Grade	Start Date	End Date	Mode
Summative Assessments	3–8	3/10/2015	6/5/2015	Online Adaptive
	11	4/13/2015	6/5/2015	Online Adaptive
	3–8, 11	5/4/2015	5/20/2015	Paper Fixed Forms
Interim Comprehensive Assessments	3–8, 11	1/5/2015	6/5/2015	Online Fixed Forms
Interim Assessment Blocks	3–8, 11	1/27/2015	6/5/2015	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-and-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 Fixed-Form	Paper
	Braille Fixed-Form	Paper
Interim Assessments	English	Online
	Braille	Online
	Spanish (math only)	Online

To ensure standardized administration conditions, Test Administrators (TAs) follow procedures outlined in the *Test Administration Manual* (TAM). TAs must review the TAM before 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. TAs follow required administration procedures and directions. 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 Test Coordinators (DTCs), School Test Coordinators (STCs), and TAs. The main responsibilities of these key personnel are described below. More detailed descriptions can be found in *Online Smarter Balanced Test Administration Manual*, provided online at the DeSSA portal, http://de.portal.airast.org.

District Test Coordinator (DTC)

The District Test Coordinator's (DTC) primary responsibility is to coordinate the administration of the Smarter Balanced assessments in the district.

DTCs are responsible for the following:

- Completing all required DeSSA training
- Reviewing scheduling and testing requirements with STCs
- Training district personnel in the use of the reporting system
- Working with schools to review DELSIS and TIDE student rolls
- Ensuring STCs and TAs understand protocols in the event that a student moves to a new district and/or school
- Ensuring that the STCs and TAs in their districts are appropriately trained regarding the state and Smarter Balancedassessment administration and security policies and procedures.
- Reviewing and submitting incidents, exemptions, security incidents, and data reviews to Delaware Department of Education (DDOE) from the Assessment Request System (ARS)
- Completing required DeSSA security forms and ensuring that all STCs and TAs have completed DeSSA security forms before administering any assessments
- General oversight responsibilities for all administration activities in their district.

School Test Coordinator (STC)

The School Test Coordinator's (STC) primary responsibilities are to coordinate the administration of the Smarter Balanced assessments and ensure that testing within his or her school is conducted in accordance with the test procedures and security policies established by the DDOE.

STCs are responsible for the following:

- Attending School Test Coordinator training
- Completing all required DeSSA training
- Completing all required security forms and ensuring that all TAs have completed all required security forms
- Ensuring that all TAs complete Smarter Balanced assessment training modules
- Working with technology personnel to ensure the DeSSA secure browser has been installed and is working on all computers be used with testing

- Completing test schedule
- Reviewing students in both DELSIS and TIDE applications before students are tested
- Ensuring that TAs understand protocols in the event that a student moves to a new district and/or school
- Ensuring all students in DSCYF, DAPI, or CDAP programs have home school record
- Ensuring accommodations have been reviewed and updated in Assessment Accommodations Database and are correct in DeSSA TIDE
- Entering any security issues, incidents, data reviews, unique accommodations, or exemptions required for any Smarter Balanced assessment testing window are entered in the ARS
- General oversight responsibilities for all administration activities in their school, and they oversee TAs

Test Administrators (TAs)

TAs administer the Smarter Balanced assessments. The assessments may only be administered by:

- Delaware-certified educators–teachers, administrators, or guidance counselors
- Paraprofessionals–if closely supervised by a Delaware-certified educator
- Translators–must be closely supervised by a Delaware-certified educator if not a Delaware-certified educator
- Substitute teachers–must be closely supervised by a Delaware-certified educator if not a Delaware-certified educator

If there is a severe shortage of staff, a test may be administered by student teachers acting as TAs – if closely supervised by a Delaware-certified educator.

Student teachers and school support staff may act as proctors

TAs are responsible for the following:

- Completing Smarter Balanced assessment administration training.
- Viewing student information beforetesting to ensure that the correct student receives the proper test with the appropriate accommodations/supports. TAs should report any potential data errors in the ARS for correction.
- Administering the Smarter Balanced assessment.
- Reporting all potential test security incidents to their STC and DTC in a manner consistent with DDOE policies and security procedures.
- Reviewing necessary manuals and user guides.
- Completing all required DeSSA training associated with assessments to be administered.
- Preparing the testing environment, ensuring that students have the necessary equipment and materials as appropriate (scratch paper, pencils, rulers, etc.).
- Reporting testing irregularities.

- Disposing of all testing materials in a secure manner including print-on-request document, scratch paper, and PT materials.

2.2.2 Online Administration

Smarter Balanced assessments allow schools to choose testing dates, allowing students to test in intervals rather than in one long period. To minimize the interruption of classroom instruction and efficiently utilize its facility, each district/school set their testing schedule within the state test window. 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 (STCs) oversee all aspects of testing at their schools and serve as the main point of contact, while TAs administer the online assessments. 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 TAs must complete the required DeSSA training courses listed on the DeSSA portal, http://de.portal.airast.org.

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 the 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 6.3 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 *Online Smarter Balanced Test Administration Manual* aloud to the students and walk 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 CAT test, students are allowed to scroll back to review and edit answers, as long as he or she is in the same test session, and the test session has not been paused for more than 20 minutes. No pause rule is implemented for the performance tasks. Students can return to the performance tasks to review and edit items they have previously completed before submitting the assessment. During an active CAT session, if a student reviews and changes the response to a previously answered item, then all following items to which the student already responded remain the same. No new items are selected due to the change of response for the previously answered item. For example, a student paued for an hour after answering item 10. After the pause, the student went back to item 5 and changed the answer. If the response change in item 5 changed the item score from wrong to right, the student's overall score would improve; however, there will be no change in items 6–10.

For the summative test, an assessment can be started in one test session and completed in a different session. 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.

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 TA to determine an appropriate stopping point; however, for ELA/Lit 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 must remain in the room at all times during a test session to monitor student testing. Once the test session ends, the TA must ensure that each student has successfully logged out of the system, collect any handouts or scratch paper that students used during the assessment and securely shred them.

2.2.3 Paper-and-Pencil Test Administration

The paper-and-pencil versions of the Smarter Balanced ELA/Lit and mathematics assessments are provided as an accommodation for students who cannot take the assessments online. For Delaware, paper-and-pencil tests were only offered in regular and Braille format.

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

For the ELA/Lit 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 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 completes the assessments, the DTC returns the test booklets to the testing vendor. The testing vendor scans the answer document and hand-scores the hand-scored items. Once all the items have been hand-scored, the testing vendor scores the overall test.

2.2.4 Braille Test Administration

In SY 2014–2015, the Online Smarter Balanced Assessment was 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 is available to students in several formats:

- The Braille interface includes a text-to-speech component for mathematics consistent with the read-aloud assessment accommodation. The Job Access with Speech (JAWS) screen reading software provided by Freedom Scientific is an essential component that students use with the Braille interface.
- Mathematics items are presented to students in Nemeth Braille through the adaptive online summative test or the performance task via a Braille embosser.
- Students taking the summative ELA/Lit assessment can emboss both reading passages and items as they progress through the assessment. If a student has a Refreshable Braille Display (RBD), a 40 cell RBD is recommended. The summative ELA/Lit is presented to the student with items in either contracted or un-contracted Literary Braille (for items containing only text) and via a Braille embosser (for items with tactile or spatial components that cannot be read by an RBD).

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

2.3 Training and Information for Test Coordinators and Administrators

All Test Coordinators (TCs), TAs, and school administrative staff who will be involved in Smarter Balanced administration must complete the Smarter Balanced Test Administrator Training Modules. Modules include security, test administration, and other information related to the administration of Smarter Balanced assessments. Successful completion of training is required before administration of Smarter Balanced assessments. More detailed information can be found in the *Online Smarter Balanced Test Administration Manual,* provided at the DeSSA portal, http://de.portal.airast.org.

Before administering a Smarter Balanced assessment, TAs must read the manuals and complete the training listed below. All individuals participating in or otherwise associated with any test administration must complete the following training requirements. Table 3 presents the training requirements.

Table 3. Smarter Balanced Summative Training Requirements

Smarter Balanced Participant Role	Required Training	Optional Training
DTC/District Administrator	• Smarter Balanced Practice Test CAT and Performance Task • Performance Task Overview • Security Module • Overview of Smarter Balanced Summative Assessment • DeSSA TIDE Training • Test Administrator Training • Administering the Classroom Activity • Administering the Performance Task	• Understanding Scoring and When Scores Will Be Received • Let's Talk Universal Tools • Student Interface for Online Testing • What Is a CAT?
STC	• Completion of Smarter Balanced Practice Test CAT and Performance Task • Performance Task Overview • Security Module • Overview of Smarter Balanced Summative Assessment • DeSSA TIDE Training • Test Administrator Training • Administering the Classroom Activity • Administering the Performance Task	• Understanding Scoring and When Scores Will Be Received • Let's Talk Universal Tools • Student Interface for Online Testing • What Is a CAT?
TA	• Completion of Smarter Balanced Practice Test CAT and Performance Task • Performance Task Overview • Security Module • Overview of Smarter Balanced Summative Assessment • Test Administrator Training • Administering the Classroom Activity • Administering the Performance Task	• Understanding Scoring and When Scores Will Be Received • Let's Talk Universal Tools • Student Interface for Online Testing • What Is a CAT?
Other (These individuals include	• Security Module • Overview of Smarter Balanced Summative Assessment	• Let's Talk Universal Tools • Student Interface for Online Testing

Smarter Balanced Participant Role	Required Training	Optional Training
but are not limited to principals, paraprofessionals, translators, etc.)	• Test Administrator Training	
Special Education Staff/Coordinator	• Completion of Smarter Balanced Practice Test CAT and Performance Task • Performance Task Overview • Security Module • Accessibility Guidelines • Accessibility and Accommodations	• Let's Talk Universal Tools • Student Interface for Online Testing
English Language Learners Staff/ Coordinator	• Completion of Smarter Balanced Practice Test CAT and Performance Task • Performance Task Overview • Security Module • Accessibility Guidelines • Accessibility and Accommodations	• Let's Talk Universal Tools • Student Interface for Online Testing
Students	• Let's Talk Universal Tools • What Is a CAT (Computer Adaptive Test)? • Student Interface for Online Testing	

2.3.1 Practice and Training Test Site

In January 2015, separate training sites were opened for TAs and students. TAs can practice administering assessments and starting and ending test sessions on the TA training site, and students can practice taking an online assessment on the student practice and training site. The Smarter Balanced assessment practice tests mirror the Smarter Balanced summative assessments for ELA/Lit 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/Lit), 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/Lit. Training tests are available for both mathematics and ELA/Lit and are organized by grade bands (grades 3–5, 6–8, and 11), with each test containing 5–10 questions.

A student can log in directly to the practice and training test site as a "Guest" without a TA-generated test session ID, or the student can log in through a training test session created by the TA in the TA training site. Items in the student training test include all item types that are included in the operational item pool, including multiple-choice, grid, and natural language items.

2.3.2 Manuals and User Guides

The manuals and user guides shown in Table 4 are available on the DeSSA portal, http://de.portal.airast.org.

Table 4. Manuals and User Guides

Resource	Description
DeSSA TIDE User Guide	DeSSA TIDE is the system used to manage student information and user accounts for online testing. The DeSSA TIDE User Guide provides a step-by-step approach to using the enhanced user management system.

Resource	Description
Test Administrator User Guide	The Test Administrator User Guide supports individuals using the test delivery system applications to manage testing for students participating in the summative assessment. This resource provides information about the test delivery system, including the TA and student applications.
Usability, Accessibility, and Accommodations Guidelines	The Usability, Accessibility, and Accommodations Guidelines focus on universal tools, designated supports, and accommodations for the Smarter Balanced assessments. The guidelines are intended for school-level personnel and decision-making teams, particularly Individualized Educational Program (IEP) and 504 teams, as they prepare for and implement the Smarter Balanced assessments. The guidelines provide information for classroom teachers, English development educators, special education teachers, and related services personnel 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.
Accessibility Guidelines for Delaware System of Student Assessments (DeSSA)	The Accessibility Guidelines for DeSSA provide information about identifying and documenting students who are eligible to receive designated supports and accommodations on Smarter Balanced and other DeSSA assessments. It also provides information on determining which assessments are appropriate for students and lists the designated supports and accommodations permitted on each assessment and in each content area. Finally, it explains the procedures for documenting supports and accommodations, including the necessary forms and deadlines.
Smarter Balanced Test Administration Manual for Paper and Pencil	The Smarter Balanced Test Administration Manual (TAM) for Paper and Pencil will provide administration information and requirements for administering the paper-and-pencil test.
Smarter Balanced Test Administration Manual for Interim Comprehensive Assessments	The Smarter Balanced Test Administration Manual (TAM) for Interim Comprehensive Assessments will provide administration information and requirements for administering the interim comprehensive assessment.
Administering a Classroom Activity	The Administering a Classroom Activity document provides instructions, details, and information to locate, prepare, and administer the classroom activity.
Administering a Performance Task	The Administering a Performance Task document provides instructions, details, and information to locate, prepare, and administer the performance task.
Technology Specifications Manual (TSM) for Online Testing	The Technology Specifications Manual provides technology staff with the technical specifications for online testing, including information on Internet and network requirements, general hardware and software requirements, secure browser installation, and text-to-speech function.
Secure Browser Installation Manual	The Secure Browser Installation Manual provides instructions for installing the secure browser on supported operating systems and is organized by operating system. This document is a supplement to the Technical Specifications Manual for Online Testing.

Resource	Description
Braille Requirements and Testing Manual	The Braille Requirements and Testing Manual includes information about supported operating systems and required hardware and software for Braille testing. It also includes a quick guide for TAs who are testing students with a Braille accommodation. This manual consolidates information that was previously split between the Technical Specifications Manual and the Test Administrator User Guide.

2.3.3 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 PowerPoint format; two modules were also narrated. Table 5 lists the training modules.

Table 5. Smarter Balanced Developed Training Modules

Module Name	Primary Audience	Objective
Accessibility and Accommodations	• TAs • Teachers • STCs	This module describes the recommended uses of available universal tools, designated supports, and accommodations for student accessibility to Smarter Balanced assessments.
Let's Talk Universal Tools	• Students • TAs • Teachers	This module acquaints students and teachers with the online, universal tools (e.g., types of calculators, expandable text) available in the Smarter Balanced assessments. This module should be shown to students in a classroom/group setting. It is suggested that teachers be in the room to answer questions from students as they view the module.
Performance Task (PT) Overview	• DTCs and STCs • Teachers	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	• Students • DTCs and STCs • TAs • Teachers	This module explains how to navigate the Student Interface.
What Is a CAT (Computer Adaptive Test)?	• District and School Test Coordinators • Teachers	This module provides the characteristics and advantages of a CAT.

2.4 TEST SECURITY

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

Test security is critical important to protect the intellectual properties, reduce test fraud and theft, and maintain the integrity of the state assessments; therefore, to ensure the validity and reliability of test scores, and fairness in testing for all Delaware students. The Test Security Manual provided online at the DeSSA portal, de.portal.airast.org sets forth test security policies, procedures, and responsibilities for DeSSA assessments. This manual is intended to be used for training who administer the state assessments.

In 2015, each district and charter school adopted and enforced a plan setting forth procedures for test security and submitted its Test Security Plan to the state by October, 2014. All unethical or inappropriate practice and behaviors must be reported in writing in the process for test preparation, test administration, and scoring. In addition, all personnel associated with assessment administration must read and sign the Test Security and Non-Disclosure Agreement as documentations.

The Test Security Manual provides examples for appropriate practices in assessment administration. Any test security violations must be reported to the Office of Assessment at the Delaware Department of Education and documented, such as missing test materials, unauthorized access to test materials, test misadministration, and any other deviations from acceptable security requirements.

In the Test Security Manual, the test security incidents during testing are defined in three levels: Impropriety, Irregularity, and Breach. Impropriety refers to an unusual circumstance that has a low impact on individual or a group of students with low risk of potentially affecting student performance on the test, which can be corrected and contained at the local level. Irregularity refers to an unusual circumstance that may potentially affecting student performance on the test, which can be corrected and contained at the local level; but must be submitted in the online appeal system for resolution. Breach refers to an event that poses a threat to the validity of the assessment (e.g., exposure of secured test materials). These circumstances have external implications and may result in a decision to remove certain test items from the operation.

The Manual specifically indicates the test security in the administration of the Smarter Balanced assessments in ELA/Lit and mathematics. For example, scratch papers and any materials developed during the classroom activities must be securely disposed prior to the administration of Performance Task (PT). Unless needed as a print-on-demand or Braille accommodation, no copies of any test items, stimuli, reading passages, PT materials, writing prompts or any secured test materials. The electronic policy clearly signifies prohibiting usages of cell phones and other electronic devices in testing area.

2.4.2 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 only in accordance with their user rights.

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

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

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

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

Student test materials and reports should not be exposed so that student names could be identified with student results except by authorized individuals with an appropriate need to know.

All students, including home-schooled students, must be enrolled or registered at their testing schools in order to take the online, paper-and-pencil, or Braille assessments. Student enrollment information, including demographic data, is generated using a DDOE 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-and-pencil versions of the assessments, 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 DTCs, STCs or teachers can view their students' scores. TAs do not have access to student scores.

2.4.3 System Security

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

A hierarchy of control: As described in Section 2.2.1, DTCs, STCs, and TAs have well-defined roles and access to the testing system.

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 login to the system. Newly added STCs, TAs, and teachers require access to all DeSSA applications via the DeSSA Single Sign-On System.

Secure browser: A key role of the STC 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.

2.4.4 Security of the Testing Environment

The STCs, 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.

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, 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, 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 before the pause. This measure was implemented to prevent students from using the time to look up answers.

Room Preparation

The room should be prepared before 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

TAs should provide adequate spacing between students' seats. Students should be seated so that they will not be tempted to look at the answers of others. Because the online CAT is adaptive, it is unlikely that

students will see the same test questions as other students; however, students should be discouraged from communicating through appropriate seating arrangements. For the performance tasks, different forms are spiraled within a classroom so that students receive different forms of the performance tasks.

After the Test

At the end of a test session, the 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. These materials should be securely shredded or stored in a locked area immediately. The printed reading passages and questions for any content area assessment provided for a student who is allowed to use this accommodation in an individual setting must also be shredded immediately after a test session ends.

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

2.4.5 Test Security Violations

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

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

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

Breach: This is a test security incident that poses a threat to the validity of the test. Breaches require immediate attention and escalation to the 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 must to document all test security incidents. The DTC is responsible for reporting test security incidents to the state via the ARS. Throughout testing, test security incidents are reported in accordance with the guidelines in the DeSSA Test Security Manual at the DeSSA portal, http://de.portal.airast.org. The deadline for all incident submissions is one week after the testing window closes.

For the 2015 Smarter administration, the investigation of test security violations was an ongoing process in the joint effort with those who administered the state assessments from schools and districts, as well as the contractors, American Institutes for Research (AIR) and Data Recognition Corporation (DRC). For instance, the Quality Assurance (QA) files provided by AIR might be used to trace potential cheating incidents; the Scoring Reports by DRC provided some evidence of potential cheating during testing and irregular behaviors (e.g., suicide attempts, family violence).

2.4.6 Monitoring Test Administration

The observation of the 2015 administration of Smarter assessments was intended to improve test administration and monitoring for the 2016 test administration. The Office of Assessment at the Department of Education scheduled on-site visits (upon agreement with schools) during the test window and all observers followed the procedure for the on-site visits without interfering with test activities. (Smarter Balanced Spring 2015 Site Visits).

The Observation and Discussion Form provides each observer with a general checklist for the appropriate test practices and standardized test conditions. The observation includes seven elements (a) Computer sign-on and start-up process; (b) Security; (c) Test environment and administration procedures; (d) Test atmosphere; (e) Calculator use in mathematics; (f) Accommodations; and (g) Classroom activity for Performance Tasks.

The Feedback Form was used to collect comments from schools and districts regarding Smarter Balanced administration, test materials, technology, service and Help Desk, and other aspects of testing. Communication with principles, test coordinators, and teachers were encouraged to collect questions, feedback, and comments prior to and/or after test sessions.

The feedback from the District Test Coordinators (DTC) and School Test Coordinators to (STC) on the Survey of the 2015 administration of Smarter Balanced assessments provided useful information for the improvement of future test administration. For example, the results suggest that the Smarter Test Administration Manual and portal resources provided necessary information (76%), which were easily accessible (67%); however, the online training should be improved. Over 70% of the DTCs and STCs agreed that their questions and concerns were solved in a timely manner by the Help Desk and 80% of them agreed that technology issues were resolved in a timely manner. In terms of scheduling test administration, over 50% of the respondents suggested to reconsider the recommended testing time in ELA/Lit.

2.5 STUDENT PARTICIPATION

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

2.5.1 Home-Schooled Students

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

- Students with the most significant cognitive disabilities who meet the criteria for the ELA/Lit alternate assessment based on alternate achievement standards (approximately 1% or fewer of the student population).

- Students with the most significant cognitive disabilities who meet the criteria for the mathematics alternate assessment based on alternate achievement standards (approximately 1% or fewer of the student population).
- ELLs who enrolled within the last 12 months before the beginning of testing in a U.S. school have a one-time exemption. These students may instead participate in their state's English language proficiency assessment consistent with state and federal policy. Students who are participating in the Interim Comprehensive Assessments or Interim Assessment Blocks may also have an exemption from completing the ELA/Lit assessment.

School personnel should follow federal and state policies regarding student participation.

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 ELA/Lit and mathematics. At the same time, the *Guidelines* support important instructional decisions about and the connection between accessibility and accommodations for students who participate in the Smarter Balanced assessments.

Following the Smarter Balanced Guidelines, the Accessibility Guidelines for Delaware System of Student Assessments on the DeSSA portal, http://de.portal.airast.org, contain the Delaware policies for governing the provision and documentation of test supports and available accommodations for students participating in the DeSSA Smarter Balanced assessments. The Delaware Guidelines clearly describe the process for the inclusion of students with disabilities (SWD), English language learners (ELL), the process for identification of those who need accommodations, and the selection and provision of the appropriate accommodation(s) and related supports. This document also provides test users with the state policy for "General Education Students Receiving Supports" who are eligible to receive supports (e.g., text-to-speech on items), not accommodations, on the Smarter Balanced ELA/Lit and mathematics assessments. The two types of accessibility features are classified as Embedded features provided directly through the on-line test environment (e.g., test-to-speech, Spanish-English staked) and Non-Embedded features that must be provided by school (e.g., translator, enhanced lighting).

In 2015, the administration of Smarter Balanced assessments can be classified into four general categories in Delaware: (a) Testing without accommodation(s) and supports; (b) Testing without accommodation(s), but with supports; (c) Testing with accommodation(s), but without supports, and (d) Testing with accommodation(s) and supports.

The summative assessments contain embedded universal tools, designated supports, and accommodations. Embedded resources 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 before 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 TA in the TA Interface of the testing system for a student who may be distracted by the ability to access a specific tool during a test session.

For additional information about the availability of designated supports and accommodations, refer to the Smarter Balanced *Usability, Accessibility, and Accommodations Guidelines* for complete information at 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 User Guide* at the DeSSA portal, http://de.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 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/Lit 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 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/Lit performance tasks in which students complete a full write. The student clicks the notepad icon for the notepad to appear. During the ELA/Lit performance tasks, the notes are retained from segment to segment so that the student may go back to the notes even though he or she cannot 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 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/Lit 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/Lit. Graph paper is required beginning in grade 6 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 or Section 504 Plan and is acceptable to the state.

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

2.6.2 Designated Supports and Accommodations

Designated Supports

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 understand 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.

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 can focus their attention on a specific part of a test item by masking.

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

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

Translations (glossaries) for math: Translated glossaries are a language support and 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 available for some students; they provide the full translation of each test item above the original item in English.

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

Non-Embedded

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/Lit 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, navigation buttons) may be adjusted by the student with an assistive technology device. Magnification allows increasing the size to a level not allowed by the Zoom universal tool.

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

Read aloud (for math items and ELA/Lit items but not for passages): Text is read aloud to the student by a trained and qualified human reader who follows the administration guidelines provided in the *Smarter Balanced Online Summative Test Administration Manual* and the *Guidelines for Read Aloud, Test Reader*. All or portions of the content may be read aloud.

Read aloud in Spanish (for mathematics tests): Spanish text is read aloud to the student by a trained and qualified human reader who follows the administration guidelines provided in the Smarter Balanced Test Administration Manual and the read aloud guidelines. All or portions of the content may be read aloud.

Scribe (for ELA/Lit 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: This is a PDF file of directions translated in each of the languages currently supported. A bilingual adult can read this file to the student.

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

Accommodations

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 accommodations.**

Embedded

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

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

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

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

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

Non-Embedded

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

Alternate response option: Alternate response options include but are not limited to adapted keyboards, large keyboards, StickyKeys, MouseKeys, FilterKeys, adapted mouse, touch screen, head wand, and switches.

Calculator (for grades 6–8, and 11 math tests): A non-embedded calculator for students needing a special calculator, such as a Braille calculator or a talking calculator, currently unavailable in 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 STC must fill out a Verification of Student Need Form and contact DDOE to have the accommodation set for the student.

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

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

Table 6 presents a list of universal tools, designated supports, and accommodations that were offered in the 2014–2015 administration. Tables 7–12 provide the number students who were offered the designated supports and/or accommodations.

Table 6. SY 2014–2015 Universal Tools, Designated Supports, and Accommodations

	Universal Tools	Designated Supports	Accommodations
Embedded	Breaks Calculator[1] Digital Notepad	Audio Glossary Color Contrast Masking	American Sign Language[10] Braille Closed Captioning[11]

	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	Text-to-Speech[6] Translated Test Directions[7] Translations (Glossary)[8] Translations (Stacked) [9] Turn off Any Universal Tools	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/Lit and math unless otherwise noted.

[1] For calculator-allowed items only
[2] For ELA/Lit performance task full writes
[3] Includes embedded ruler, embedded protractor
[4] For ELA/Lit items
[5] Includes bold, italic, underline, indent, cut, paste, spell check, bullets, undo/redo
[6] For ELA/Lit PT stimuli, ELA/Lit PT and CAT items (not ELA/Lit CAT reading passages), and math items: Must be set in TIDE before test begins.
[7] For math items
[8] For math items
[9] For math test
[10] For ELA/Lit listening items and math items
[11] For ELA/Lit listening items
[12] For ELA/Lit 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/Lit performance task full writes
[14] For ELA/Lit performance task full writes
[15] For ELA/Lit performance task full writes
[16] For ELA/Lit 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 7. ELA/Lit Total Students with Allowed Embedded and Non-Embedded Accommodations

Accommodations	Grade						
	3	4	5	6	7	8	11
Embedded Accommodations							
American Sign Language	6	4	2	3	3		4
Closed Captioning	8	17	9	12	10	7	5
Streamlined Mode	15	9	3	1		1	16
Text-to-Speech: Passage & Items				22	10	11	7
Non-Embedded Accommodations							
Alternate Response Options				1	3	2	
Print on Demand: Stimuli & Items	358	325	349	325	304	268	129
Read Aloud Passages	28	11	8	5	7	10	6
Scribe Items (Writing)	101	76	78	24	17	20	1
Speech-to-Text	2	2	10	5	5	6	2

Table 8. ELA/Lit Total Students with Allowed Embedded Designated Supports

Designated Supports	Subgroup	Grade						
		3	4	5	6	7	8	11
Color Choices	Overall	60	106	117	57	74	70	20
	ELL	8	8	8	1	3	6	10
	Special Ed	36	85	75	53	65	65	7
Masking	Overall	327	253	291	157	100	134	70
	ELL	124	75	51	22	26	22	13
	Special Ed	131	110	135	129	86	120	55
Print Size	Overall	212	229	103	43	32	33	20
	ELL	120	108	34	12	11	10	14
	Special Ed	63	86	54	32	22	25	5
Text-to-Speech: Items	Overall	1832	1619	1539	1082	981	986	327
	ELL	716	435	243	126	147	139	53
	Special Ed	839	885	973	857	834	807	272
Text-to-Speech: Stimuli	Overall	1	6	4	15	8	7	28
	ELL			1	2	1	1	4
	Special Ed	1	6	3	14	8	6	22
Text-to-Speech: Stimuli & Items	Overall	251	148	112	76	88	68	14
	ELL	95	38	8	18	16	13	1
	Special Ed	118	70	83	54	66	54	14

Table 9. ELA/Lit Total Students with Allowed Non-Embedded Designated Supports

Designated Supports	Subgroup	Grade						
		3	4	5	6	7	8	11
Bilingual Dictionary	Overall	90	31	66	57	56	58	43
	ELL	87	31	64	57	56	58	43
	Special Ed	14	3	13	9	9	9	4
Color Contrast	Overall	14	5	4	1	2	3	2
	ELL	12		1			1	
	Special Ed	4	5	4	1	2	2	2
Color Overlay	Overall	7	12	15	17	13	19	2
	ELL	3	1	3				
	Special Ed	4	10	11	13	12	14	1
Magnification	Overall	25	11	19	12	8	6	5
	ELL	11	2	2	1		1	
	Special Ed	11	6	16	11	7	4	4
Noise Buffers	Overall	112	112	123	48	47	60	5
	ELL	18	6	11	8	12	12	
	Special Ed	74	98	107	27	24	38	2
Read Aloud Items	Overall	465	453	394	254	167	107	25
	ELL	127	102	31	21	25	17	7
	Special Ed	300	309	271	196	140	88	16
Read Aloud Stimuli	Overall	28	12	8	5	8	10	6
	ELL	6	1	1	1		1	
	Special Ed	25	12	7	5	7	6	5
Scribe Items (Non-Writing)	Overall	69	57	53	8	8	7	4
	ELL	10	1	1				
	Special Ed	56	54	47	8	8	7	4
Separate Setting	Overall	1,121	1,173	1,125	1,181	1,157	1,186	378
	ELL	204	143	92	68	81	66	16
	Special Ed	819	870	872	1,027	1,026	1043	324

Table 10. Mathematics Total Students with Allowed Embedded Accommodations

Accommodations	Grade						
	3	4	5	6	7	8	11
Embedded Accommodations							
American Sign Language	6	4	2	5	3		4
Streamlined Mode	17	10	3		4	1	
Non-Embedded Accommodations							
Abacus	15	18	20	37	27	16	1
Alternate Response Options				1	3	2	
Calculator	13	41	38	145	156	159	148
Multiplication Table	132	994	1,035	891	751	629	56
Print on Demand: Stimuli & Items	342	304	298	318	298	268	129
Speech-to-Text	2	2	7	4	4	6	2

Table 11. Mathematics Total Students with Allowed Embedded Designated Supports

Designated Supports	Subgroup	Grade						
		3	4	5	6	7	8	11
Color Choices	Overall	70	102	107	73	77	67	24
	ELL	8	10	8	1	3	6	14
	Special Ed	36	80	64	62	61	61	7
Masking	Overall	316	258	271	148	101	139	76
	ELL	126	78	49	22	28	26	17
	Special Ed	129	111	125	121	85	121	58
Print Size	Overall	213	232	106	54	34	36	26
	ELL	120	109	36	14	13	14	20
	Special Ed	63	87	55	35	22	24	5
Translation (Glossary): English	Overall	10,231	9,977	9,998	10,067	9,728	9,492	7,485
	ELL	1,000	601	334	278	320	262	134
	Special Ed	1,277	1,347	1,383	1,402	1,317	1,346	759
Translation (Glossary): Spanish	Overall			1		1		1
	ELL			1		1		1
	Special Ed							
Translation (Glossary): Other Languages	Overall	34	10	9	6	9	3	4
	ELL	31	10	9	6	8	3	4
	Special Ed	2				2		
Translations: Stacked	Overall	81	66	59	62	76	57	23
	ELL	81	66	59	62	76	57	23
	SPED	11	7	6	6	3	6	
Text-to-Speech: Items	Overall	1,796	1,607	1,530	1,146	994	971	313
	ELL	689	430	231	126	150	120	42
	Special Ed	824	890	987	879	829	808	271
Text-to-Speech: Stimuli	Overall	2	5	3	7	5	6	26
	ELL				1	1		3
	Special Ed	2	5	1	5	5	6	22
Text-to-Speech: Stimuli & Items	Overall	417	275	290	164	212	148	83
	ELL	168	62	54	45	50	43	19
	Special Ed	191	146	189	98	115	92	55

Table 12. Mathematics Total Students with Allowed Non-Embedded Designated Supports

Designated Supports	Subgroup	Grade						
		3	4	5	6	7	8	11
Color Contrast	Overall	14	5	4	1	2	3	
	ELL	11		1			1	
	Special Ed	5	5	4	1	2	2	
Color Overlay	Overall	7	12	15	16	12	21	1
	ELL	3	1	3				
	Special Ed	4	10	11	12	11	17	
Translation (Glossary): All Languages	Overall	3	21	41	13	10	15	13
	ELL	2	21	34	13	10	14	12
	Special Ed		1	5	3		2	1
Magnification	Overall	16	11	20	12	7	5	5
	ELL	3	2	2	1		1	1
	Special Ed	9	6	17	11	6	4	4
Noise Buffers	Overall	109	112	124	47	46	63	6
	ELL	17	6	11	8	12	14	
	Special Ed	70	98	108	28	23	39	3
Read Aloud Items	Overall	490	479	382	283	174	113	25
	ELL	148	126	41	28	30	23	8
	Special Ed	305	308	271	199	140	88	15
Scribe Items (Non-Writing)	Overall	72	58	55	7	9	5	4
	ELL	10	1	3				
	Special Ed	59	54	48	6	9	5	4
Separate Setting	Overall	1,142	1,198	1,160	1,200	1,172	1,199	393
	ELL	213	158	100	80	92	74	27
	Special Ed	824	879	897	1,043	1,027	1,047	331
Translated Test Directions	Overall	19	21	13	11	15	2	6
	ELL	19	21	13	11	15	1	6
	Special Ed	2	1			2	1	

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 support cheating detection, 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-and-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, after the first operational administration. Following the first operational administration, the analyses used to detect the testing anomalies can be run any time within the testing window. AIR evaluated evidence including 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 an authorized 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. The system flags unusual changes in an aggregate performance between administrations and/or years 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 is flagged if the percentage of flagged students is greater than 50%. The aggregate unit size for the score change is based on the number of students included in the within-year 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 adding up the page time for all items and item groups.

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

The average and the standard deviation of test-taking time are computed across all students for each opportunity. Students and aggregate units 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 when the analysis was performed. The QA report includes a list of the flagged aggregate units with the number of flagged students in the aggregate unit.

2.7.3 Inconsistent Item Response Pattern (Person Fit)

In item response theory (IRT) models, person-fit measurement is used to identify 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), he or she will respond correctly to those items at a higher probability than indicated by his or her ability as estimated across all items. In this case, the person-fit index will be large for the student. We note, however, that if a student has prior knowledge of the entire test content, this will not be detected based on the person-fit index, although the item response 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, Pornel, 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/Lit and mathematics assessments. Tables 13–14 present the demographic composition of Delaware students who meet attemptedness requirements for scoring and reporting of the Smarter Balanced assessments.

Table 13. Number of Students in SY 2014–2015 Summative ELA/Lit Assessment

Group	G3	G4	G5	G6	G7	G8	G11
All Students	10,231	9,910	9,922	10,023	9,716	9,546	7,497
Female	5,122	4,932	4,890	4,943	4,735	4,669	3,721
Male	5,109	4,978	5,032	5,080	4,981	4,877	3,776
African American	3,016	3,060	3,115	3,097	3,068	3,109	2,315
Asian	375	385	361	352	354	328	283
Hispanic/Latino	1,763	1,702	1,533	1,601	1,453	1,267	854
American Indian/Alaska Native	38	43	41	48	52	38	36
White	4,631	4,331	4,585	4,694	4,555	4,574	3,892
English Language Learner	984	558	303	247	285	258	138
Special Education	1,279	1,349	1,381	1,389	1,328	1,350	765
CD 504	332	376	412	416	351	404	258
Title I	1,161	1,274	1,621	1,814	1,902	1,957	810

Table 14. Number of Students in SY 2014–2015 Summative Mathematics Assessment

Group	G3	G4	G5	G6	G7	G8	G11
All Students	10,268	9,995	10,017	10,084	9,754	9,512	7,521
Female	5,150	4,970	4,935	4,981	4,753	4,646	3,731
Male	5,118	5,025	5,082	5,103	5,001	4,866	3,790
African American	3,026	3,063	3,148	3,111	3,064	3,091	2,321
Asian	391	401	375	358	360	329	284
Hispanic/Latino	1,784	1,736	1,565	1,635	1,490	1,264	860
American Indian/Alaska Native	38	43	41	48	52	38	37
White	4,620	4,362	4,602	4,701	4,556	4,558	3,906
English Language Learner	1,032	613	346	291	334	267	141
Special Education	1,280	1,355	1,390	1,405	1,324	1,350	778
CD 504	333	377	409	417	350	402	256
Title I	1,163	1,279	1,628	1,826	1,912	1,943	816

3.2 SUMMARY OF OVERALL STUDENT PERFORMANCE

Table 15 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 15. 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/Lit							
3	2438.10	84.73	21	25	25	29	54
4	2477.39	88.02	25	21	25	29	54
5	2509.37	89.30	24	21	34	22	55
6	2522.78	92.41	24	28	32	16	48
7	2547.11	96.00	25	25	35	15	50
8	2559.13	97.90	24	27	35	14	49
11	2581.57	112.79	24	24	31	21	52
Mathematics							
3	2439.39	75.47	21	26	32	21	53
4	2476.86	75.44	19	35	29	17	47
5	2498.56	84.99	31	31	20	18	38
6	2510.54	96.32	33	32	19	15	34
7	2529.61	102.70	31	32	22	15	37
8	2541.72	111.97	37	28	19	17	35
11	2541.14	119.80	52	25	15	8	23

3.3 TEST TAKING TIME

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

In the Test Delivery System (TDS), item response latency is captured as the item page time (the length of time that each item page is presented) in milliseconds. Discrete items appear on the screen one item 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 associated items appear on the screen together. 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 16 and 17 present an average testing time and testing time by hourly intervals for the overall test, the CAT component, and the PT component.

Table 16. ELA/Lit 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	4:40	0.98	6.76	16.64	21.55	54.06
4	4:45	0.70	4.69	14.56	22.99	57.07
5	4:42	0.62	4.13	13.92	23.40	57.92
6	4:08	0.88	6.06	20.02	27.23	45.81
7	3:49	1.51	9.60	22.57	28.27	38.06
8	3:38	2.05	11.05	26.82	26.16	33.92
11	2:34	9.06	26.71	32.34	19.86	12.02
CAT Component						
3	2:11	6.26	45.47	30.56	11.68	6.03
4	2:17	4.55	40.24	36.22	13.02	5.97
5	2:15	4.09	41.84	37.05	11.90	5.12
6	2:09	5.95	43.93	35.58	11.00	3.54
7	1:55	8.75	53.09	29.64	6.05	2.48
8	1:51	10.55	55.62	26.03	5.38	2.43
11	1:22	29.41	55.54	13.08	1.61	0.37
PT Component						
3	2:29	14.26	31.76	26.14	13.37	14.47
4	2:28	11.84	33.17	27.53	14.23	13.24
5	2:28	10.30	33.47	28.26	15.45	12.52
6	1:59	17.46	42.48	25.59	8.06	6.41
7	1:53	20.89	41.17	24.23	8.86	4.85
8	1:47	24.32	41.42	22.03	7.60	4.63
11	1:12	45.65	41.01	10.48	2.12	0.74

Table 17. 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	2:38	3.52	35.06	31.29	16.32	13.81
4	2:26	4.30	37.82	33.84	14.96	9.08
5	3:04	2.10	23.69	31.70	21.68	20.82
6	2:32	2.65	32.57	38.78	17.63	8.36
7	2:11	6.58	43.49	33.73	11.46	4.74
8	2:18	6.88	37.14	35.51	14.10	6.38
11	1:31	26.22	51.03	18.62	3.25	0.88
CAT Component						
3	1:37	23.06	51.86	18.14	4.82	2.12
4	1:39	20.78	53.88	18.52	4.51	2.32
5	1:48	15.52	51.98	22.93	6.15	3.42

6	1:36	15.39	62.78	17.50	3.46	0.86
7	1:40	15.75	58.85	20.07	3.78	1.55
8	1:37	19.14	57.01	18.70	3.61	1.54
11	1:02	51.65	42.90	4.79	0.53	0.13
PT Component						
3	1:01	61.31	30.85	6.00	1.32	0.53
4	0:47	75.37	21.34	2.78	0.37	0.13
5	1:15	45.40	40.45	10.47	2.51	1.17
6	0:55	66.25	29.82	3.16	0.48	0.29
7	0:31	91.41	8.19	0.28	0.09	0.03
8	0:41	80.95	18.08	0.88	0.06	0.02
11	0:29	94.01	5.69	0.28	0.01	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 Delaware 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 accurately measure high performing students but needs additional easy items to better measure low performing students. The Smarter Balanced Assessment Consortium plans to add additional easy items to the pool and augment the pool in proportion to the test blueprint constraints (e.g., content, Depth-of-Knowledge (DOK), item type, item difficulties).

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

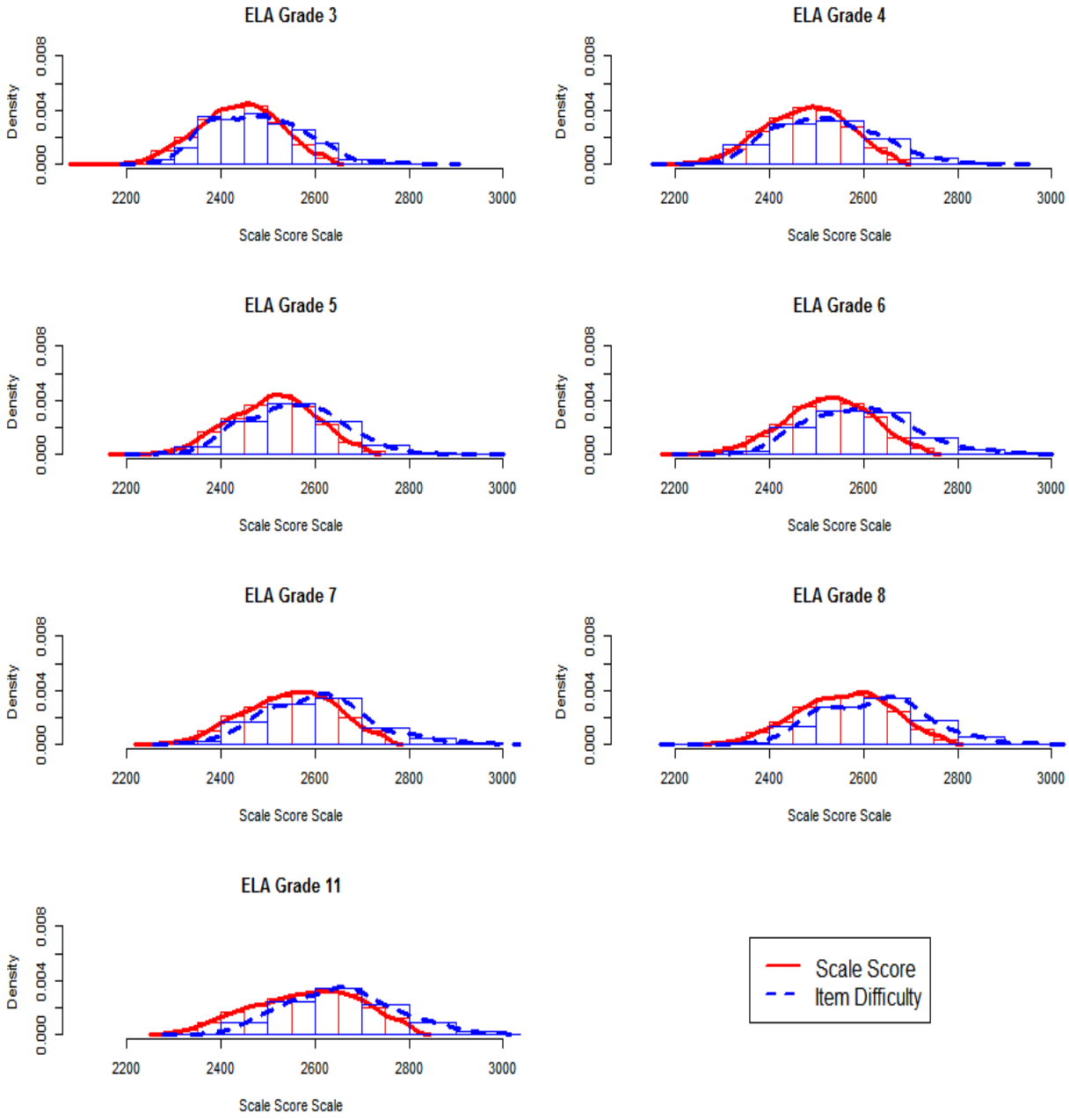
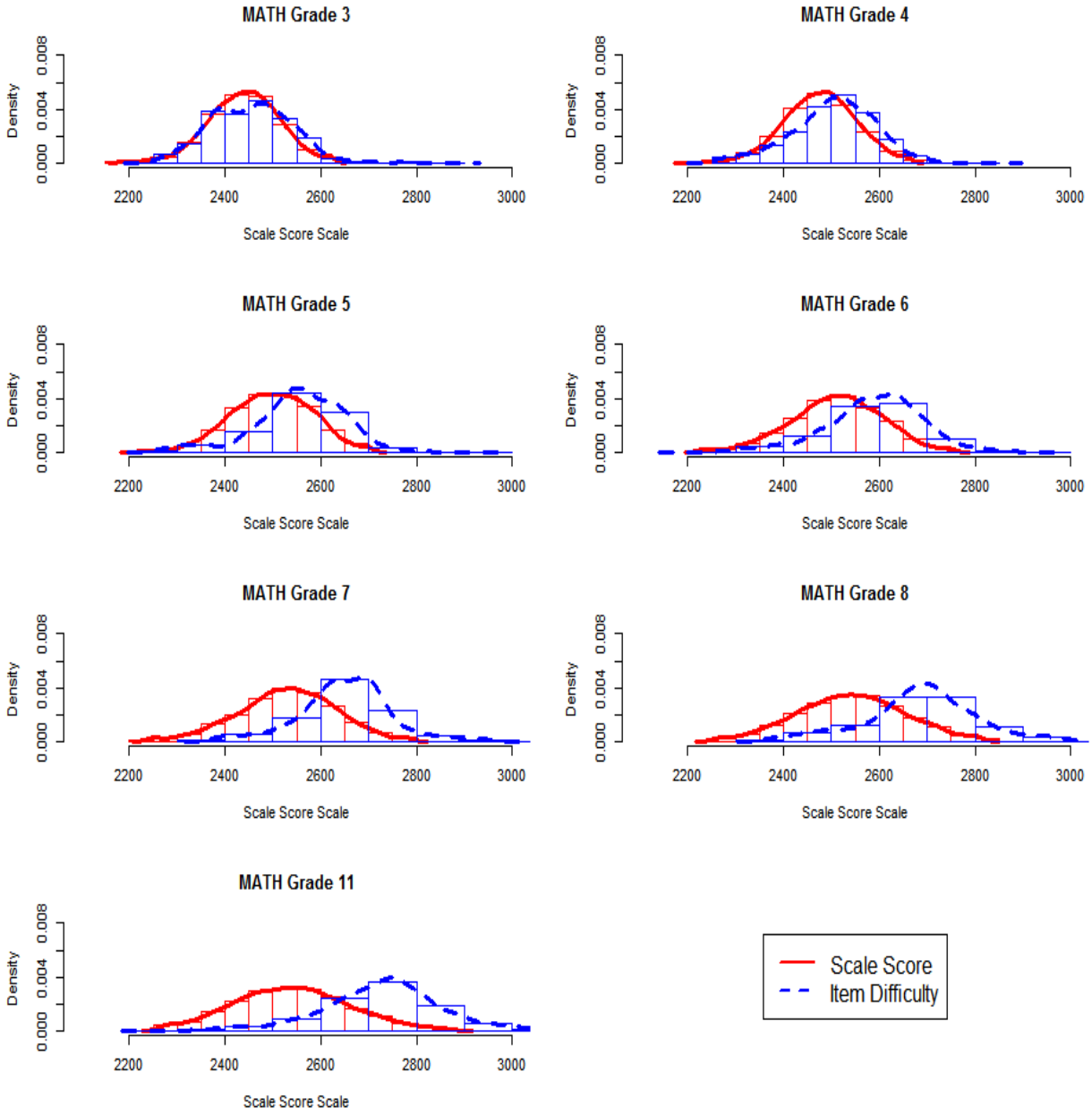


Figure 2. SY 2014–2015 Student Ability–Item Difficulty Distribution for Mathematics



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 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 by the relationships between Smarter Balanced test scores and SAT 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 or her ability. For PT, each student is administered with a fixed-form test. The content converge in all PT forms is the same.

In the adaptive item-selection algorithm, item selection takes place in two discrete stages: blueprint satisfaction and match-to-ability. The Smarter Balanced blueprints (Smarter Balanced Assessment Consortium, 2015) specify a range of items to be administered in each claim, content domain/standards, and targets. Moreover, blueprints constrain 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/Lit, the blueprints also specify the number of passages in reading (claim 1) and listening (claim 3) claims.

Tables 18–21 present the percentages of tests aligned with the test blueprint constraints for ELA/Lit and mathematics for CAT. The blueprint match rates are summarized for item and passage requirements in ELA/Lit, and for claims and content domains in mathematics, within each claim.

In ELA/Lit, 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/Lit 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 18. Percentage of ELA/Lit 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	97%	
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%	

Legend:

1-IT: Reading with Literary Text, 1-LT: Reading with Informational Text, 2-W: Writing, 3-L: Listening, and 4-CR: Research

Table 19. 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	98%	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%

Legend:

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

Table 20. 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	99%	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%

Legend:

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

Table 21. 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	100%	1	P	NoCalc	8	8	100%
1	S	NoCalc	2	2	100%	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	SP	Calc	0	2	100%
2	OTHER	Calc	0	2	100%	2	OTHER	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	95%
4	NS	Calc	0	1	100%	4	N	Calc	0	2	100%
4	SP	Calc	0	1	100%	4	SP	Calc	0	2	100%
4	OTHER	Calc	0	1	100%	4	OTHER	Calc	0	1	100%

Legend:

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

Table 22 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 22. 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/Lit												
3	14	5	1	3	11.1	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	8-13	3-4	1-1	3-3
5	14	5	1	3	11.2	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	3.3	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	7-10	2-2	3-6	3-4
4	12	4	6	6	10.0	2.0	5.4	3.0	9-10	2-2	3-6	2-3
5	11	4	6	6	9.0	2.0	5.3	3.0	9-9	2-2	3-6	3-4
6	10	4	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	8-8	2-2	3-6	3-3
8	10	4	7	6	10.0	2.0	5.0	3.0	10-10	2-2	3-6	3-3
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 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 23–24. The correction for attenuation indicates what the correlation would be if reporting category scores could be measured with perfect reliability. The observed correlation between two reporting category scores with measurement errors can be corrected for attenuation as $r_{x'y'} = r_{xy} / SQRT(r_{xx} * r_{yy})$, where $r_{x'y'}$ is the correlation between x and y corrected for attenuation, r_{xy} is the observed correlation between x and y, r_{xx} is the reliability coefficient for x, and r_{yy} is the reliability coefficient for y.

When corrected for attenuation, the correlations among reporting scores are quite high, indicating that the assessments measure a common underlying construct.

Table 23. Correlations among Reporting Categories for ELA/Lit

Grade	Reporting Categories	Observed & Disattenuated Correlation			
		Claim 1	Claim 2	Claim 3	Claim 4
3	Claim 1: Reading	1	0.88	0.93	0.92
	Claim 2: Writing	0.69	1	0.89	0.87
	Claim 3: Listening	0.61	0.59	1	0.91
	Claim 4: Research	0.66	0.63	0.56	1
4	Claim 1: Reading	1	0.92	0.95	0.95
	Claim 2: Writing	0.70	1	0.90	0.92
	Claim 3: Listening	0.63	0.60	1	0.93
	Claim 4: Research	0.67	0.65	0.57	1
5	Claim 1: Reading	1	0.92	0.97	0.95
	Claim 2: Writing	0.71	1	0.89	0.95
	Claim 3: Listening	0.63	0.58	1	0.95
	Claim 4: Research	0.68	0.69	0.57	1
6	Claim 1: Reading	1	0.89	0.99	0.95
	Claim 2: Writing	0.67	1	0.96	0.95
	Claim 3: Listening	0.60	0.61	1	1
	Claim 4: Research	0.63	0.66	0.56	1
7	Claim 1: Reading	1	0.91	1	0.97
	Claim 2: Writing	0.71	1	0.95	0.94
	Claim 3: Listening	0.64	0.61	1	1
	Claim 4: Research	0.69	0.68	0.59	1
8	Claim 1: Reading	1	0.93	0.99	0.97
	Claim 2: Writing	0.72	1	0.94	0.95
	Claim 3: Listening	0.63	0.61	1	0.99
	Claim 4: Research	0.68	0.68	0.58	1
11	Claim 1: Reading	1	0.92	0.96	0.93
	Claim 2: Writing	0.72	1	0.92	0.99
	Claim 3: Listening	0.63	0.62	1	0.96
	Claim 4: Research	0.67	0.72	0.59	1

Table 24. Correlations among Reporting Categories for Mathematics

Grade	Reporting Categories	Observed & Disattenuated Correlation		
		Claim 1	Claim 2&4	Claim 3
3	Claim 1: Concepts and Procedures	1	0.96	0.97
	Claim 2 & 4: Problem Solving & Modeling and Data Analysis	0.79	1	1
	Claim 3: Communicating Reasoning	0.76	0.74	1
4	Claim 1: Concepts and Procedures	1	0.99	0.98
	Claim 2 & 4: Problem Solving & Modeling and Data Analysis	0.77	1	1
	Claim 3: Communicating Reasoning	0.80	0.73	1
5	Claim 1: Concepts and Procedures	1	1	1
	Claim 2 & 4: Problem Solving & Modeling and Data Analysis	0.75	1	1
	Claim 3: Communicating Reasoning	0.75	0.70	1
6	Claim 1: Concepts and Procedures	1	1	1
	Claim 2 & 4: Problem Solving & Modeling and Data Analysis	0.77	1	1
	Claim 3: Communicating Reasoning	0.75	0.69	1
7	Claim 1: Concepts and Procedures	1	1	1
	Claim 2 & 4: Problem Solving & Modeling and Data Analysis	0.76	1	1
	Claim 3: Communicating Reasoning	0.71	0.64	1
8	Claim 1: Concepts and Procedures	1	1	1
	Claim 2 & 4: Problem Solving & Modeling and Data Analysis	0.72	1	1
	Claim 3: Communicating Reasoning	0.77	0.66	1
11	Claim 1: Concepts and Procedures	1	1	1
	Claim 2 & 4: Problem Solving & Modeling and Data Analysis	0.73	1	1
	Claim 3: Communicating Reasoning	0.66	0.61	1

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 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.

The evidence for convergent validity is obtained using the SAT scores. Evidence for convergent validity is determined by examining the patterns of correlations between Smarter Balanced Summative Assessments and performance on SAT. 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 SAT, we expect test scores between the common subject-area assessments to be substantially correlated.

The relationship between the Smarter Balanced assessment scores and the SAT scores in ELA/Lit and mathematics was examined to evaluate the convergent aspect of validity using grade 11 assessment data. The SAT ELA score is a sum of SAT reading and writing scores. As expected that the correlation between

the Smarter Balanced Assessment scores and the SAT scores for the same subject (convergent validity) is moderate, correlations between scores are 0.71 in ELA/Lit and 0.79 in mathematics. The results are shown in Table 25.

Table 25. Relationship Between the Smarter Balanced and SAT Test Scores

Test/Subject	N	Average Scale Score	Correlation (SB, SAT)
SB ELA	7011	2587.65	0.71
SAT ELA	7011	429.68	
SB Math	7048	2546.42	0.79
SAT Math	7048	433.45	

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

The reliability evidence of the Smarter Balanced summative 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 IRT, 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 26 presents the marginal reliability coefficients and the average conditional SEM for the total scale scores.

Table 26. Marginal Reliability for ELA/Lit and Mathematics

Grade	Number of Items Specified in Test Blueprint		Marginal Reliability	N	Scale Score Mean	Scale Score SD	Average CSEM
	Min	Max					
ELA/Lit							
3	41	44	0.92	10,231	2438.10	84.73	23.95
4	40	44	0.91	9,910	2477.39	88.02	25.97
5	41	45	0.92	9,922	2509.37	89.30	25.78
6	41	45	0.91	10,023	2522.78	92.41	28.20
7	41	45	0.92	9,716	2547.11	96.00	27.78
8	43	45	0.92	9,546	2559.13	97.90	27.85
11	42	45	0.92	7,497	2581.57	112.79	31.58
Mathematics							
3	39	40	0.94	10,268	2439.39	75.47	18.52
4	37	40	0.94	9,995	2476.86	75.44	18.76
5	38	40	0.93	10,017	2498.56	84.99	22.83
6	38	39	0.92	10,084	2510.54	96.32	26.87
7	38	40	0.91	9,754	2529.61	102.70	30.39
8	38	40	0.91	9,512	2541.72	111.97	32.81
11	40	42	0.88	7,521	2541.14	119.80	41.82

5.2 STANDARD ERROR CURVES

Figures 3–4 present plots of the conditional SEM of scale scores across the range of ability. The vertical lines indicate the cutscores for Level 2, Level 3, and Level 4.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/Lit

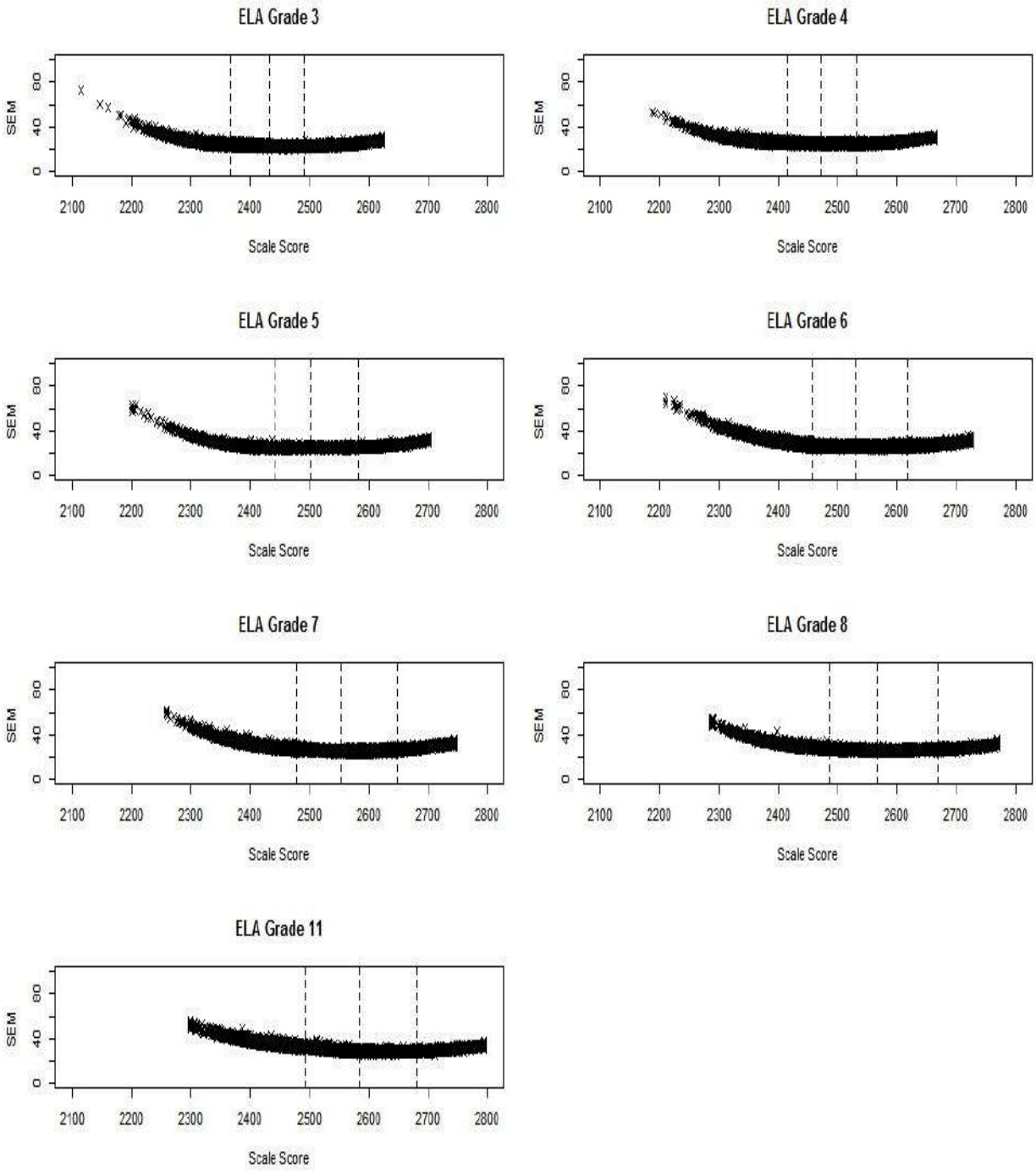
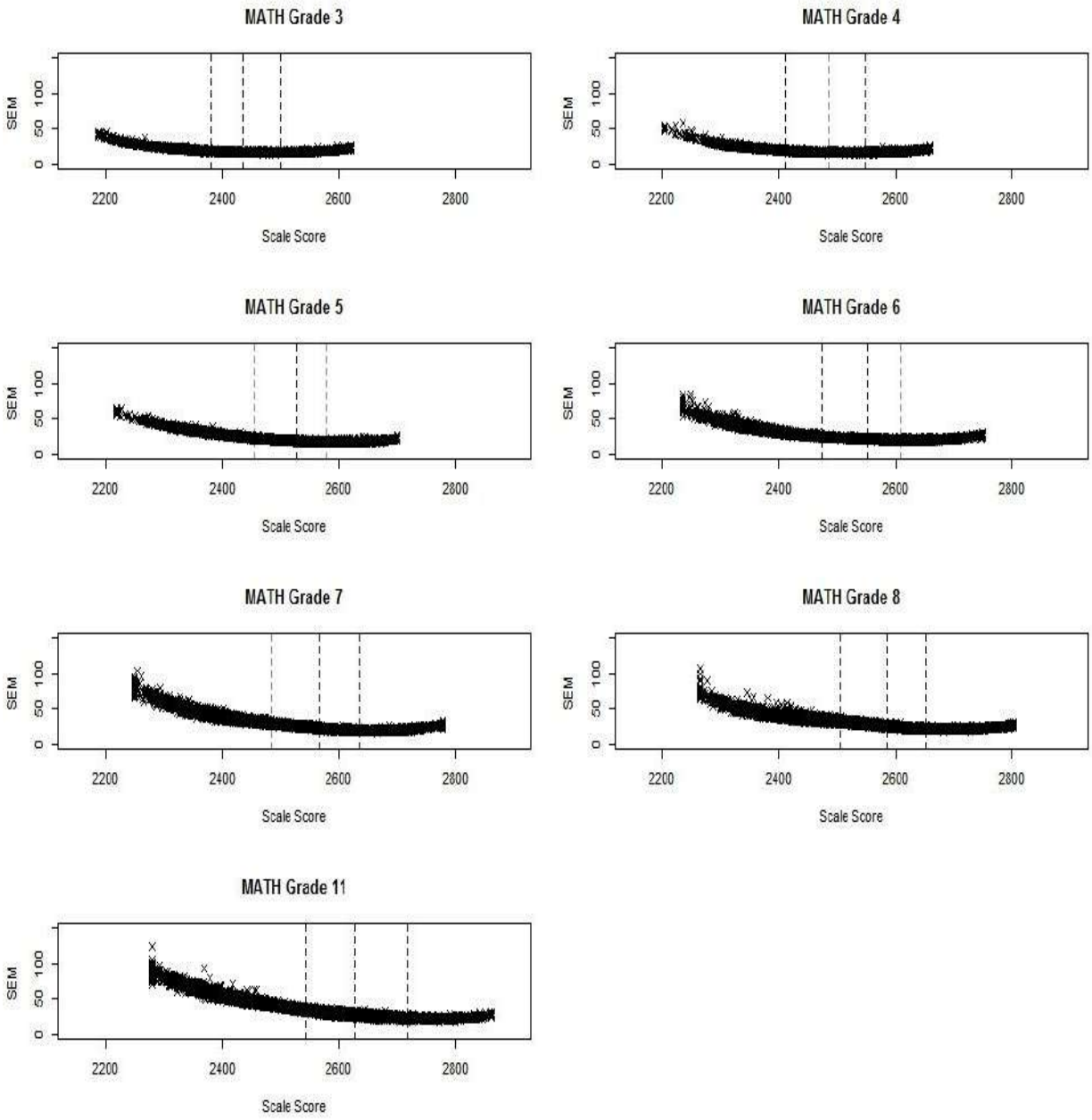


Figure 4. Conditional Standard Error of Measurement for Mathematics



The SEMs presented in the figures above are summarized in Tables 27–28. Table 27 provides the average conditional SEM for all scores and scores in each achievement level. Table 28 presents the average conditional SEMs at 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/Lit and mathematics. Average conditional SEMs at all cut scores are similar in ELA/Lit, but larger in Level 2 cut in mathematics.

Table 27. Average Conditional Standard Error of Measurement by Achievement Levels

Grade	Level 1	Level 2	Level 3	Level 4	Average CSEM
ELA/Lit					
3	27.12	22.95	22.35	23.71	23.95
4	28.09	24.94	24.56	25.96	25.97
5	27.66	24.45	24.64	26.60	25.78
6	33.80	26.01	25.79	27.51	28.20
7	31.54	26.27	25.63	28.46	27.78
8	31.85	26.27	25.97	28.01	27.85
11	37.59	30.30	28.21	30.26	31.58
Mathematics					
3	22.26	17.81	16.68	17.97	18.52
4	23.30	17.96	16.84	17.93	18.76
5	29.34	20.54	18.23	18.09	22.83
6	34.42	23.19	20.97	21.10	26.87
7	41.52	26.27	21.36	20.98	30.39
8	41.82	29.42	23.96	22.48	32.81
11	51.22	31.45	26.00	23.42	41.82

Table 28. 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/Lit						
3	23.94	21.71	22.27	2.23	0.56	1.67
4	25.18	24.94	24.73	0.24	0.21	0.45
5	24.43	24.55	24.49	0.12	0.06	0.06
6	27.02	25.74	26.04	1.28	0.30	0.98
7	26.98	25.81	26.16	1.17	0.35	0.82
8	27.34	25.86	26.59	1.48	0.73	0.75
11	31.81	28.77	28.46	3.04	0.31	3.35
Mathematics						
3	18.94	17.15	16.77	1.79	0.38	2.17
4	19.55	17.01	16.64	2.54	0.37	2.91
5	22.89	18.78	17.82	4.11	0.96	5.07
6	24.89	21.41	20.23	3.48	1.18	4.66
7	29.65	22.96	19.97	6.69	2.99	9.68
8	32.26	25.71	22.20	6.55	3.51	10.06
11	35.29	27.73	23.05	7.56	4.68	12.24

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 test, 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

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

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

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

where, d stands for dichotomous and p stands for polytomous items, $\mathbf{b}_j = (a_j, b_j, c_j)$ if the jth item is a dichotomous item, and $\mathbf{b}_j = (a_j, b_{j1}, \ldots, b_{jK_i})$ if the *j*th item is a polytomous item, a_j is the item's discrimination parameter (for Rasch model, $a_j = 1$), c_j is the guessing parameter (for Rasch and 2PL models, $c_j = 0$), 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 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 $cut_l (l = 1,\cdots,L-1)$ is estimated

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

and the false negative (*FN*) for $cut_l (l = 1,\cdots,L-1)$ 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 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, p_{ca}, is the sum of the marginal probabilities of classifications into the same level based on observed and expected classifications, hence, for cut_l $(l = 1, \cdots, L-1)$, this is estimated by

$$p_{cal} = p_{cal1} + p_{cal2},$$

where

$$p_{cal1} = \left(\frac{\sum_{k,m=1}^{l} n_{akm}}{N} + \frac{\sum_{m=1}^{l}\sum_{k=l+1}^{L} n_{akm}}{N}\right)\left(\frac{\sum_{k,m=1}^{l} n_{akm}}{N} + \frac{\sum_{k=1}^{l}\sum_{m=l+1}^{L} n_{akm}}{N}\right),$$

$$p_{cal2} = \left(\frac{\sum_{k,m=l+1}^{L} n_{akm}}{N} + \frac{\sum_{m=1}^{l}\sum_{k=l+1}^{L} n_{akm}}{N}\right)\left(\frac{\sum_{k,m=l+1}^{L} n_{akm}}{N} + \frac{\sum_{k=1}^{l}\sum_{m=l+1}^{L} n_{akm}}{N}\right).$$

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

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

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

Similarly, the same calculations can be conducted for classification consistency. Hence, for 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_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 29, 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 29. 2014–2015 Decision Accuracy and Consistency by Achievement Levels

Grade	Achievement Level	ELA/Lit				Mathematics			
		Accuracy		Consistency		Accuracy		Consistency	
		% Accuracy	Kappa	% Consistency	Kappa	% Accuracy	Kappa	% Consistency	Kappa
3	L2	94.2	0.83	91.8	0.76	94.3	0.83	91.9	0.76
	L3	92.4	0.85	89.2	0.78	92.9	0.86	90.0	0.80
	L4	92.9	0.83	89.9	0.76	94.7	0.84	92.5	0.78
4	L2	93.4	0.83	90.7	0.76	94.5	0.82	92.3	0.75
	L3	92.1	0.84	88.8	0.78	92.9	0.86	90.0	0.80
	L4	92.4	0.82	89.4	0.74	95.5	0.84	93.6	0.78
5	L2	94.2	0.84	91.7	0.77	92.9	0.83	90.0	0.77
	L3	91.9	0.84	88.7	0.77	93.7	0.87	91.0	0.81
	L4	93.3	0.81	90.5	0.73	95.4	0.85	93.5	0.78
6	L2	93.3	0.81	90.6	0.74	92.6	0.83	89.6	0.77
	L3	91.6	0.83	88.3	0.76	93.3	0.85	90.6	0.79
	L4	94.0	0.78	91.6	0.70	95.7	0.83	93.9	0.77
7	L2	93.9	0.84	91.4	0.77	92.1	0.82	89.0	0.75
	L3	92.3	0.85	89.1	0.78	93.2	0.85	90.4	0.79
	L4	94.2	0.78	91.8	0.70	96.1	0.85	94.5	0.79
8	L2	93.7	0.83	91.1	0.76	91.7	0.82	88.4	0.75
	L3	92.6	0.85	89.5	0.79	93.3	0.85	90.5	0.79
	L4	94.6	0.78	92.4	0.70	96.2	0.87	94.6	0.81
11	L2	94.2	0.84	91.8	0.78	91.1	0.82	87.6	0.75
	L3	93.1	0.86	90.2	0.80	94.6	0.85	92.4	0.79
	L4	93.6	0.81	91.0	0.74	97.8	0.85	96.8	0.79

5.4 RELIABILITY FOR SUBGROUPS

Tables 30–31 show the marginal reliability coefficients for each of the subgroups. As shown in tables, reliabilities of total scale scores are consistent across subgroups.

Table 30. Marginal Reliability Coefficients for Overall and by Subgroup for ELA/Lit

Subgroup	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 11
All Students	0.92	0.91	0.92	0.91	0.92	0.92	0.92
Female	0.92	0.91	0.91	0.90	0.91	0.91	0.92
Male	0.92	0.91	0.92	0.90	0.92	0.92	0.92
American Indian/Alaska Native	0.90	0.90	0.91	0.90	0.92	0.92	0.93
Asian	0.91	0.90	0.90	0.89	0.91	0.91	0.93
African American	0.91	0.90	0.91	0.89	0.90	0.90	0.91
Hispanic	0.90	0.89	0.90	0.90	0.90	0.90	0.91
White	0.91	0.91	0.91	0.90	0.91	0.91	0.92
ELL	0.86	0.85	0.83	0.79	0.82	0.84	0.84
Special Education	0.86	0.85	0.84	0.80	0.83	0.85	0.84
CD 504	0.90	0.88	0.90	0.89	0.90	0.90	0.92
Title I	0.90	0.90	0.91	0.90	0.91	0.91	0.90

Table 31. 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	0.92	0.91	0.91	0.88
Female	0.94	0.93	0.92	0.92	0.91	0.91	0.87
Male	0.94	0.94	0.93	0.92	0.91	0.92	0.88
American Indian/Alaska Native	0.93	0.92	0.92	0.92	0.90	0.93	0.87
Asian	0.94	0.94	0.94	0.94	0.95	0.94	0.94
African American	0.93	0.92	0.90	0.89	0.87	0.86	0.80
Hispanic	0.92	0.92	0.90	0.90	0.89	0.88	0.83
White	0.94	0.94	0.93	0.93	0.92	0.92	0.89
ELL	0.91	0.90	0.83	0.78	0.76	0.79	0.71
Special Education	0.91	0.89	0.82	0.81	0.76	0.77	0.57
CD 504	0.93	0.92	0.92	0.91	0.89	0.90	0.88
Title I	0.92	0.93	0.93	0.91	0.90	0.90	0.80

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 32–33 present the marginal reliability coefficients for each claim score in ELA/Lit and mathematics, respectively.

Table 32. Marginal Reliability Coefficients for Claim Scores in ELA/Lit

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	0.77	2432.01	98.72	47.29
	Claim 2: Writing	11	11	0.79	2434.88	98.99	45.63
	Claim 3: Listening	8	8	0.56	2441.05	109.42	72.34
	Claim 4: Research	8	9	0.67	2430.88	113.44	65.57
4	Claim 1: Reading	14	16	0.76	2472.95	104.31	51.20
	Claim 2: Writing	11	11	0.76	2475.63	97.68	47.74
	Claim 3: Listening	8	8	0.58	2479.63	115.54	74.75
	Claim 4: Research	7	9	0.65	2466.87	116.21	68.72
5	Claim 1: Reading	14	16	0.77	2505.89	103.13	48.94
	Claim 2: Writing	11	11	0.78	2499.52	99.70	46.34
	Claim 3: Listening	8	9	0.54	2500.54	123.96	84.20
	Claim 4: Research	8	9	0.67	2525.51	104.17	60.15
6	Claim 1: Reading	14	16	0.72	2501.71	114.77	60.99
	Claim 2: Writing	11	11	0.79	2520.81	102.99	47.13
	Claim 3: Listening	8	9	0.51	2538.35	125.72	88.13
	Claim 4: Research	8	9	0.61	2530.87	110.50	68.77
7	Claim 1: Reading	14	16	0.76	2537.34	111.63	54.79
	Claim 2: Writing	11	11	0.79	2548.41	106.15	48.91
	Claim 3: Listening	8	9	0.52	2548.23	123.67	85.38
	Claim 4: Research	8	9	0.66	2544.93	120.63	70.64
8	Claim 1: Reading	16	16	0.77	2558.95	110.64	52.68
	Claim 2: Writing	11	11	0.79	2553.47	111.36	50.61
	Claim 3: Listening	8	9	0.53	2560.85	121.58	82.92
	Claim 4: Research	8	9	0.65	2555.08	118.39	70.31
11	Claim 1: Reading	15	16	0.76	2586.15	119.03	58.25
	Claim 2: Writing	11	11	0.79	2577.27	129.20	59.36
	Claim 3: Listening	8	9	0.57	2558.04	140.24	91.48
	Claim 4: Research	8	9	0.67	2582.85	136.27	78.79

Table 33. 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	0.89	2439.31	79.47	26.50
	Claim 2 & 4: Problem Solving & Modeling and Data Analysis	8	11	0.76	2434.69	85.96	42.50
	Claim 3: Communicating Reasoning	9	11	0.70	2437.72	91.02	49.98
4	Claim 1: Concepts and Procedures	20	20	0.89	2476.48	78.39	26.40
	Claim 2 & 4: Problem Solving & Modeling and Data Analysis	8	10	0.67	2467.09	94.16	54.25
	Claim 3: Communicating Reasoning	9	10	0.75	2476.97	87.60	43.73
5	Claim 1: Concepts and Procedures	20	20	0.87	2497.32	88.75	31.64
	Claim 2 & 4: Problem Solving & Modeling and Data Analysis	8	10	0.60	2485.87	112.14	70.77
	Claim 3: Communicating Reasoning	9	10	0.65	2491.94	106.26	63.09
6	Claim 1: Concepts and Procedures	19	19	0.87	2508.46	101.23	36.60
	Claim 2 & 4: Problem Solving & Modeling and Data Analysis	9	10	0.60	2501.74	118.32	74.52
	Claim 3: Communicating Reasoning	10	11	0.63	2504.53	115.05	69.64
7	Claim 1: Concepts and Procedures	20	20	0.86	2528.50	107.42	40.45
	Claim 2 & 4: Problem Solving & Modeling and Data Analysis	10	10	0.57	2514.56	131.52	85.90
	Claim 3: Communicating Reasoning	8	10	0.49	2512.72	131.62	94.12
8	Claim 1: Concepts and Procedures	20	20	0.85	2539.60	117.56	45.33
	Claim 2 & 4: Problem Solving & Modeling and Data Analysis	8	10	0.56	2530.13	139.25	92.31
	Claim 3: Communicating Reasoning	9	10	0.68	2532.58	127.24	72.41
11	Claim 1: Concepts and Procedures	22	22	0.80	2534.21	126.47	57.12
	Claim 2 & 4: Problem Solving & Modeling and Data Analysis	8	10	0.47	2522.48	160.48	117.30
	Claim 3: Communicating Reasoning	9	12	0.48	2534.00	137.59	98.86

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 hand-scoring procedure.

6.1 ESTIMATING STUDENT ABILITY USING MAXIMUM LIKELIHOOD ESTIMATION

The Smarter Balanced assessments are scored using maximum likelihood estimation (MLE). The likelihood function for generating the MLEs is based on a mixture of 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, $\boldsymbol{m}_i$ is the maximum possible score of this item, $\boldsymbol{a}_i$ 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 $p_{ij}(z_{ij} \mid \theta_j, a_i, b_{i,1}, \mathsf{K}\ , b_{i,m_i})$ 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 $m_i = 1$,

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

in the case of items with two or more points,

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

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. Table 34 lists the scaling constants for each subject for the theta-to-scale score linear transformation. Scale scores will be rounded to an integer.

Table 34. 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 35 provides three achievement standards for each grade and content area.

Table 35. Theta Cut Scores and Reported Scale Scores

Grade	ELA/Lit			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 does not 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 36 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 36. 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 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/Lit 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

Data Recognition Corporation (DRC) provided all hand scoring for the online Smarter Balanced summative assessments, and Measurement Incorporated (MI) provided all hand scoring for the Smarter Balanced summative assessments in paper format. In ELA/Lit, short-answer (SA) items and full write items are scored by human raters, also identified 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 that DRC and MI follow. DRC and MI use similar procedures to score responses to all Smarter Balanced constructed response or written composition items.

6.7.1 Rater Selection

Measurement Incorporated

MI maintains a large pool of qualified, experienced readers at each scoring center as well as distributed readers who work remotely from their homes. MI simply informs the readers that a project is pending and invites 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/Lit and mathematics placement tests, take a grammar exercise, write an acceptable essay, and receive good recommendations from references. MI then reviews all the information about an applicant before offering employment.

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

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

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

Data Recognition Corporation

DRC retains a number of raters from year to year. This pool of experienced raters was used to staff the scoring of the 2015 Smarter Balanced assessments. To complete the rater staffing for this project, DRC placed advertisements in local newspapers and utilized a variety of web sites. Open houses were held and applications for rater positions were screened by DRC's recruiting staff. Candidates were personally interviewed by DRC staff. In addition, each candidate was required to provide an on-demand writing sample, an on-demand math sample, references, and proof of a four-year college degree. In this screening

process, preference was given to candidates with previous experience scoring large-scale assessments and degrees emphasizing expertise in mathematics or ELA/Lit. Thus, the rater pool consisted of educators and other professionals with content-specific backgrounds. These individuals were valued for their content-specific knowledge, but they were required to set aside their own biases about student performance and accept the scoring standards outlined for the Smarter Balanced assessments.

Scoring directors and team leaders were selected from a pool of employees who displayed expertise as raters and leaders on previous DRC projects. These individuals had strong backgrounds in mathematics or ELA/Lit, and demonstrated organizational, leadership, and management skills. A majority of scoring directors and team leaders had at least five years of leadership experience working on large-scale assessments. All scoring directors, team leaders, and raters signed Confidentiality/Nondisclosure Agreements before handling secure materials.

Each grade/content group of raters was assigned a scoring director. This individual led all handscoring activities for the duration of the project. Scoring directors worked with supervisors to format Smarter Balanced training materials, conducted team leader training, and were responsible for training the raters. The scoring director made sure that rater reports were available and interpreted those reports for the raters. The scoring director also supervised the team leaders. All scoring directors were monitored by the project director, the project manager, and the content specialists.

Team leaders assisted the scoring director with rater training by leading their teams in small group discussions and answering individual questions that raters may not have felt comfortable asking in a large group. Once raters were qualified, team leaders were responsible for maintaining the accuracy and workload of each team member. Ongoing monitoring identified those individuals having difficulty scoring accurately. These raters received one-on-one retraining from the team leader. Any rater who could not be successfully retrained had his/her scores purged and was released from the project.

6.7.2 Rater Training

Measurement Incorporated

All readers hired for Smarter Balanced assessment hand scoring 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 (e.g., brief writes, reading, research, full writes, mathematics). Within each group, readers are divided into teams consisting of one team leader and 10–15 readers. Each team leader and reader is assigned a unique number for easy identification of their scoring work throughout the scoring session.

MI's Virtual Scoring Center (VSC) online training interface presents rubrics, scoring guides, and training/qualifying sets in three modes (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, the 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 ensure 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 the Smarter Balanced Assessment Consortium 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 or her time 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 hand-scoring system, explanations of how to "flag" unusual responses for review by the scoring director, and instructions about other procedures which are necessary to conduct 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.
- Mathematics: Readers train on baseline items, which qualify the readers for that item as well as any items associated with it; for items with no associated items, training is for the specific item.

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

Data Recognition Corporation

As part of preparation for the 2015 Smarter Balanced Assessments, DRC's Performance Assessment Scoring (PAS) staff formatted and reviewed Smarter Balanced-provided scoring training sets. The scoring guides and associated training materials served as the raters' constant reference.

Raters were instructed on how to apply the scoring guidelines and were required to demonstrate a clear comprehension of each anchor set by performing well on the associated training materials.

The scoring director conducted a team leader training session before training the raters. This session followed the same procedures as rater training, but additional time was spent with team leaders to ensure that all team leaders would impart the same scoring rationale to their readers. During team leader training, all Smarter Balanced assessment materials were reviewed and discussed. Once the team leaders were qualified, leadership responsibilities were reviewed and team assignments were given. A ratio of one team leader per 10–12 raters ensured sufficient monitoring rates for team members.

Rater training began with the scoring director providing an intensive review of the scoring guidelines and anchor papers. Next, raters practiced by independently scoring the responses in the training sets. After each training set, the scoring director or team leaders led a thorough discussion of the responses, either in a large-group or small-group setting.

Once the scoring guidelines, anchor sets, and training sets were thoroughly discussed, each rater was required to demonstrate understanding of the scoring criteria by qualifying (i.e., scoring with acceptable agreement to the true scores) on at least one of the qualifying sets. Raters who failed to achieve the appropriate qualification percentages on the first qualifying set were given additional, individual training. Raters who did not perform at the required level of agreement (0–1 point items – 90% exact; 0–2 and 0–3 point items – 80% exact; 0–4 point items – 70% exact) by the end of the qualifying process were not allowed to score any student responses. These individuals were removed from the pool of potential raters in DRC's imaging system and released from the project.

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

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

MI has developed and operates a comprehensive system for collecting and analyzing scoring data. After the readers' scores are submitted into the VSC hand-scoring system, the data 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

DRC also operates a comprehensive system for collecting and analyzing scoring data.

1. The Reader Monitor Report monitored how often raters were in exact agreement with one another and ensured that an acceptable agreement rate was maintained. This report provided daily and cumulative exact and adjacent inter-rater agreement on the ten percent of scores that were double read.

2. The Score Point Distribution Report monitored the percentage of responses given each of the score points. For example, the mathematics daily and cumulative reports showed what percentage of 0s, 1s, 2s, 3s, and 4s a rater had given to all the responses scored at the time the report was produced. It also indicated the number of responses read by each rater so that production rates could be monitored.

The Item Status Report monitored the progress of handscoring. This report tracked each response and indicated the status (e.g., not read, complete, awaiting supervisor review, etc.). This report ensured that all responses were scored by the end of the project.

The aforementioned validity reports tracked how raters performed by comparing pre-scored responses to raters' scores for the same responses. If a rater's scoring fell below the determined agreement rate, remediation occurred.

In both DRC and MI hand scoring systems, 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 hand-scoring project monitors at each scoring center, and they provide updated reports to the scoring directors several times a day. 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

Measurement Incorporated

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

Retraining is an ongoing process once scoring is underway. Daily analysis of the Reader Status Reports enables management personnel to identify individual or group retraining needs. If it becomes apparent that a whole team or a whole group is having difficulty with a particular type of response, large group training sessions are conducted. Standard retraining procedures include room-wide discussions led by the scoring director, team discussions conducted by team leaders, and one-on-one discussions with individual readers.

It is standard practice to conduct morning room-wide retraining at MI each day, with a more extensive retraining on Monday mornings in order to re-anchor the readers after a weekend away from scoring.

Each student response is scored holistically by a trained and qualified reader using the scoring scales developed and approved by Smarter Balanced, with 10%–15% second read for reliability purposes. Item responses for the second read were selected randomly and were scored blindly. The second reader was unaware of the first reader's score. MI's quality assurance/reliability procedures allow their hand-scoring 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.

Data Recognition Corporation

Rater accuracy was monitored throughout the scoring session by means of daily and on-demand reports. These reports ensured that an acceptable level of scoring accuracy was maintained throughout the project. Interrater reliability was tracked and monitored with multiple quality control reports that were reviewed by quality assurance analysts. These reports and other quality control documents were generated at the scoring centers, where they were reviewed by the scoring directors, team leaders, content specialists, and project directors.

6.7.5 Rater Validity Checks

Measurement Incorporated

Scoring directors select responses which are loaded into the VSC system as validity responses. The "true" or rangefinding scores for these responses are entered into a validity database. These responses are embedded 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.

Data Recognition Corporation

One of the training tools that PAS utilized to ensure rater accuracy was the validity process. The goal of the validity process is to ensure that scoring standards are maintained. Specifically, the objective is to make sure that raters score student responses in a manner consistent within and across state and consortia-wide standards both within a single administration of the Smarter Balanced assessments and across consecutive administrations. During the scoring of the 2015 Smarter Balanced assessments, scoring consistency was maintained, in part, through the validity process.

The Smarter Balanced Assessment Consortium provided DRC with validity papers for each item or item type. The responses were imported into the imaging system and dispersed intermittently to the raters. By the end of the project, raters had scored validity papers for each item type they were qualified to score. Raters were unaware that they were being dealt pre-scored validity responses and assumed that they were scoring live student responses. This helped bolster the internal validity of the process. It is important to note that all raters who received validity papers had already successfully completed the training/qualifying process.

Next, the scores that the raters assigned to the validity papers were compared to the true scores in order to determine the validity of the raters' scores. For each item, the percentage of exact agreement as well as the percentage of high and low scores was computed. This data was accessed through the Validity Item Detail Report. The same sort of data was also computed for each specific rater. This data was accessed through the Validity Reader Detail Report. Both of these could be run as daily or cumulative reports.

The Validity Reader Detail Report was used to identify particular raters for retraining. If a rater on a certain day generated a lower rate of agreement on a group of validity papers, it was immediately apparent in the Validity Reader Detail Report. A lower rate of agreement was defined as anything below 70 percent exact agreement with the true scores. Any time a rater's validity agreement rate fell below 70 percent, the scoring director examined that rater's scoring. First, the scoring director attempted to ascertain what kind of validity papers the rater was scoring incorrectly. This was done to determine whether there was any sort of a trend (e.g., trending low on the 1–2 line). Once the source of the low agreement rate was determined, the rater was retrained. If it was determined that the rater had been scoring live papers inaccurately, then his/her scores were purged for that day, and the responses were re-circulated and scored by other raters.

The cumulative Validity Item Detail Report was utilized to identify potential room-wide trends in need of correction. For instance, if a particular validity response with a true score of 3 was given a score of 2 by a significant number of raters within the room, that trend would be revealed in the Validity Item Detail Report. To correct a trend of this sort, the scoring director would look for student responses similar to the validity paper being scored incorrectly. Once located, these responses would be used in room-wide re-training, usually in the form of an annotated handout or a short set of papers without printed scores given to raters as a recalibration test.

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 37–38 provide a summary of the inter-rater reliability for the Delaware 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/Lit, 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 hand-scored items range from 1–4.

Table 37. Reader Agreements for ELA/Lit

Grade	Item Type	# of Items	% Exact	Min (%Exact)	Max (%Exact)	% items w/ %Exact ≥ 80%	% items w/ %Exact ≥ 70%
3	Short Answer	32	86.10	74	98	91	100
3	WR: Conv	14	94.90	88	99	100	100
3	WR: Evid/Elab	14	95.15	88	99	100	100

Grade	Item Type	# of Items	% Exact	Min (%Exact)	Max (%Exact)	% items w/ %Exact ≥ 80%	% items w/ %Exact ≥ 70%
3	WR: Org/Purp	14	95.00	89	99	100	100
4	Short Answer	28	85.50	76	93	86	100
4	WR: Conv	19	91.53	82	98	100	100
4	WR: Evid/Elab	19	89.21	76	95	89	100
4	WR: Org/Purp	19	90.65	79	95	95	100
5	Short Answer	32	83.60	73	91	91	100
5	WR: Conv	20	87.73	81	94	100	100
5	WR: Evid/Elab	20	88.44	76	96	95	100
5	WR: Org/Purp	20	88.85	81	96	100	100
6	Short Answer	42	85.59	72	96	93	100
6	WR: Conv	14	91.67	85	95	100	100
6	WR: Evid/Elab	14	93.71	89	97	100	100
6	WR: Org/Purp	14	93.71	89	97	100	100
7	Short Answer	32	84.66	61	98	91	97
7	WR: Conv	19	91.88	84	98	100	100
7	WR: Evid/Elab	19	93.18	88	99	100	100
7	WR: Org/Purp	19	93.01	88	98	100	100
8	Short Answer	40	88.58	77	100	98	100
8	WR: Conv	21	92.87	85	100	100	100
8	WR: Evid/Elab	21	92.49	86	97	100	100
8	WR: Org/Purp	21	92.75	88	97	100	100
11	Short Answer	43	90.41	79	98	95	100
11	WR: Conv	24	93.52	88	100	100	100
11	WR: Evid/Elab	24	93.62	87	97	100	100
11	WR: Org/Purp	24	94.25	87	100	100	100

Table 38. Reader Agreements for Mathematics

Grade	Score Points	# of Items	% Exact	Min (%Exact)	Max (%Exact)	% items w/ %Exact ≥ 80%	% items w/ %Exact ≥ 70%
3	1	12	96.20	94	100	100	100
3	2	19	92.93	81	100	100	100
3	3	4	92.59	88	98	100	100
4	1	8	95.05	93	98	100	100
4	2	23	97.87	95	100	100	100
4	3	3	98.04	97	100	100	100
5	1	40	96.21	87	100	100	100
5	2	7	91.67	82	99	100	100
5	3	12	98.66	96	100	100	100
6	1	30	94.48	84	100	100	100
6	2	6	99.68	99	100	100	100
6	3	15	95.64	91	100	100	100
6	4	11	97.76	91	100	100	100
7	1	26	96.08	85	100	100	100
7	2	11	97.01	88	100	100	100
7	3	13	95.43	90	100	100	100
8	1	6	97.21	94	100	100	100
8	2	12	96.20	94	100	100	100
11	1	19	92.93	81	100	100	100
11	2	4	92.59	88	98	100	100
11	3	8	95.05	93	98	100	100

7. REPORTING AND INTERPRETING SCORES

The Online Reporting System (ORS) 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 hand-scored. Because the score report on students' performance are updated each time students complete tests and they are hand scored, authorized users (e.g., school principals, teachers) can view students' performance on the tests and use them to improve student learning. In addition to individual student's score report, the ORS produces aggregate score reports for teachers, schools, districts, and states. 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 ORS provides participation data that helps monitor student participation rate.

This section contains a description of the types of scores reported in the ORS 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 ORS is designed to help educators, students, and parents answer questions regarding how well students have achieved on ELA/Lit 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 who are not technical measurement experts in mind, ensuring that test results are 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, the 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 to the ORS and select Score Reports, the online score reports are presented hierarchically. The ORS starts with presenting summaries on student performance by subject and grade at a selected aggregate level. In order to view student performance for a specific aggregate unit, users can select the specific aggregate unit from a drop-down menu with a list of aggregate units (e.g., schools within a districts, or teachers within a school) to select. For more detailed student assessment results for a school, a teacher, or a roster, users can select the subject and grade on the online score reports.

Generally, the ORS provides two categories of online score reports: (1) aggregate score reports and (2) student score reports. Table 39 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 ORS User Guide, located in a help button on the ORS.

Table 39. Types of Online Score Reports by Level of Aggregation

Level of Aggregation	Types of Online Score Reports
State 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) • Participation rate (overall students)[1] • 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 teacher, school, district, and state

Note.
1: Participation rate reports are provided at state, 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 40 presents the types of subgroups and subgroup category provided in ORS.

Table 40. Types of Subgroups

Subgroup	Subgroup Category
Gender	Male
	Female
CD504	CD504
	Not CD504
ELL	ELL
	Not ELL
Special Education	Special Education
	Not Special Education
Title I	Title I
	Not Title I
Ethnicity	African American
	American Indian or Alaska Native
	Asian
	Hispanic
	White

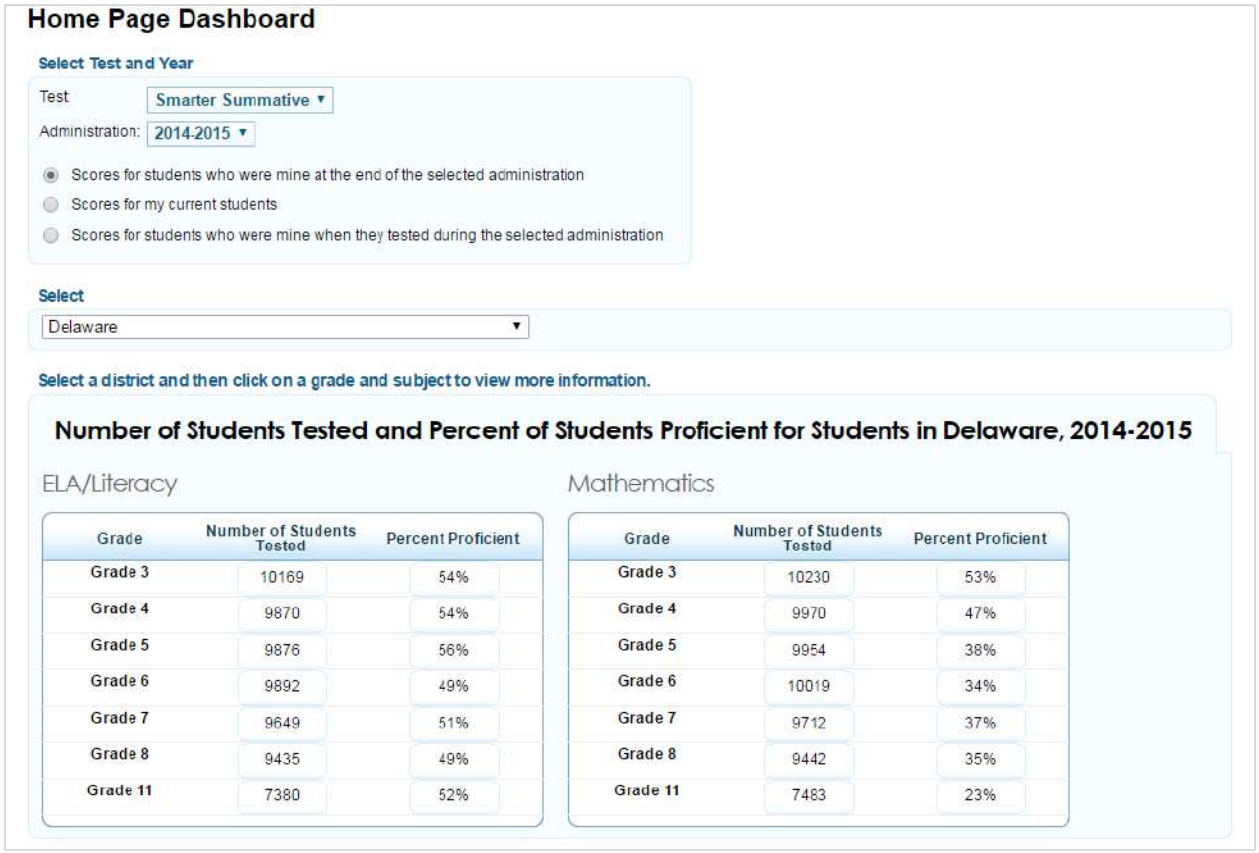
7.1.2 Online Reporting System

7.1.2.1 Home Page

When users log in to the ORS and select Score Reports, the first page displays summaries of students' performance across grades and subjects. State personnel see state summaries, 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 state personnel can see a summary of students' performance for district as well as state.

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. Exhibits 1 and 2 present sampled Home Pages at the state level and the district level.

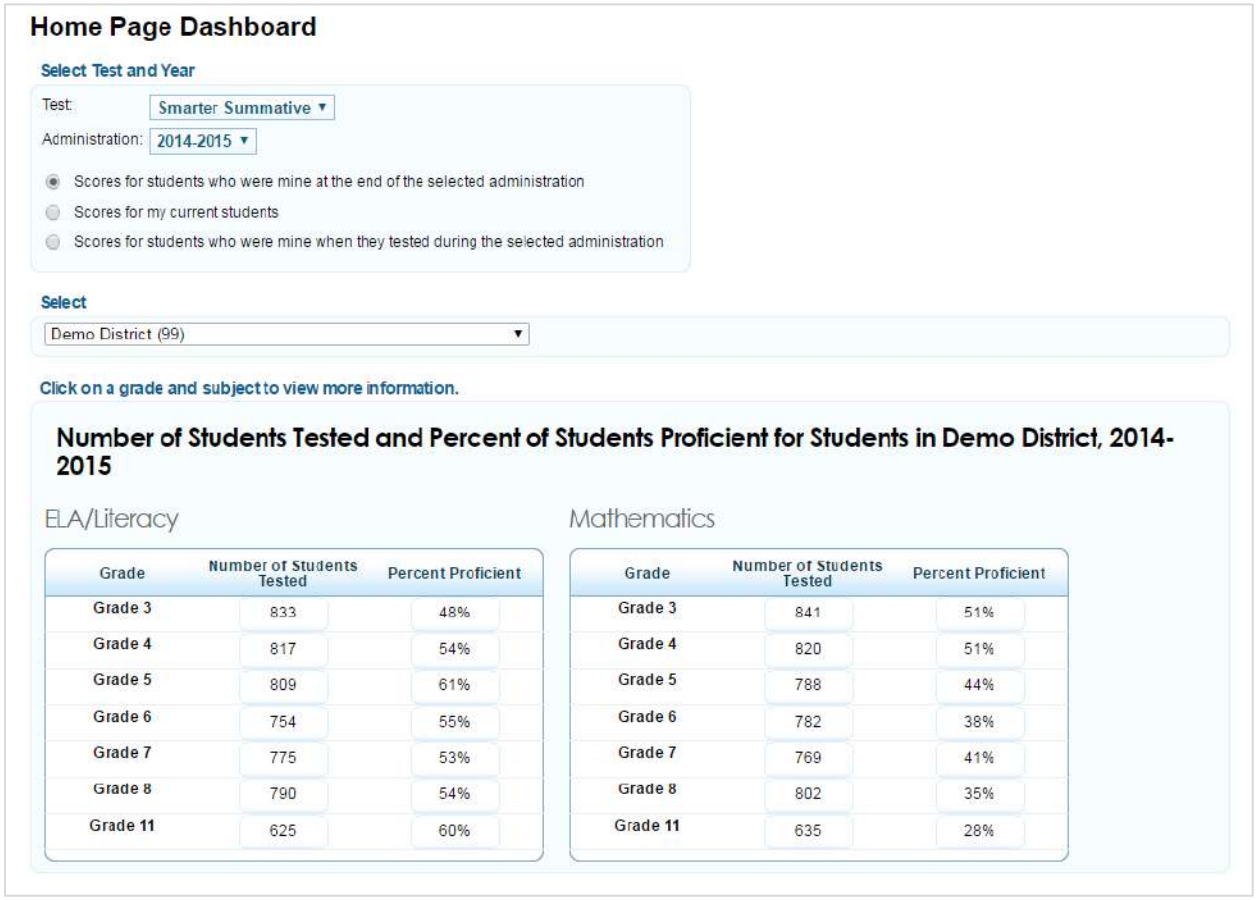
Exhibit 1. Home Page: State Level



Grade	Number of Students Tested	Percent Proficient
Grade 3	10169	54%
Grade 4	9870	54%
Grade 5	9876	56%
Grade 6	9892	49%
Grade 7	9649	51%
Grade 8	9435	49%
Grade 11	7380	52%

Grade	Number of Students Tested	Percent Proficient
Grade 3	10230	53%
Grade 4	9970	47%
Grade 5	9954	38%
Grade 6	10019	34%
Grade 7	9712	37%
Grade 8	9442	35%
Grade 11	7483	23%

Exhibit 2. Home Page: District Level

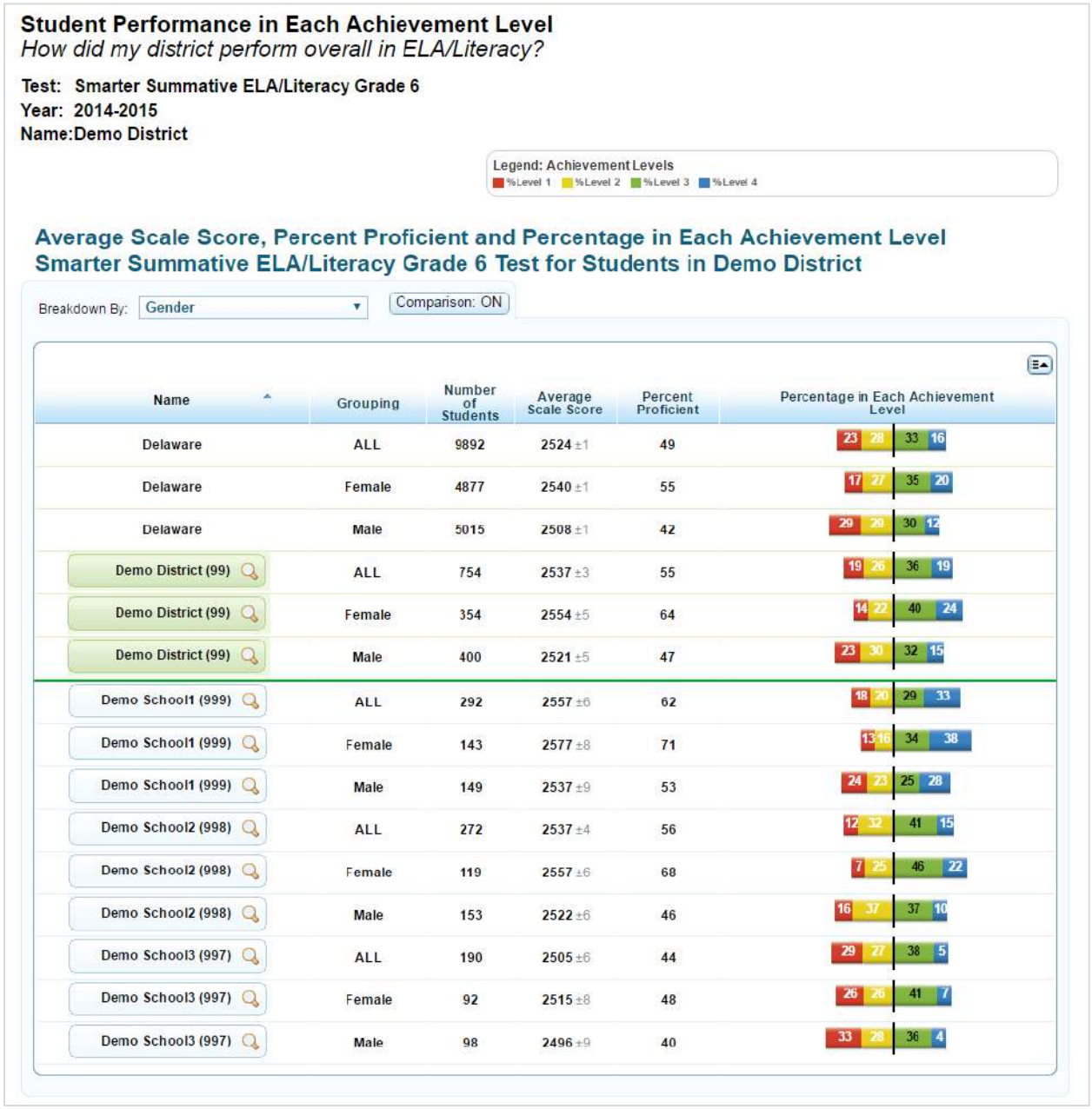


7.1.2.2 Subject Detail Page

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 state and the aggregate unit above the selected aggregate. For example, if a school is selected on the Subject Detail Page, the summary results of the state, the district, and the school 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 proficient, and (4) percent of students in each achievement level. The summaries are also presented for overall students and by subgroups. Exhibit 3 presents an example of Subject Detail Pages for ELA/Lit at the district level when a user select a subgroup of gender.

Exhibit 3. Subject Detail Page for ELA/Lit 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 Proficient and Percentage in Each Achievement Level
Smarter Summative ELA/Literacy Grade 6 Test for Students in Demo District

Breakdown By: Gender | Comparison: ON

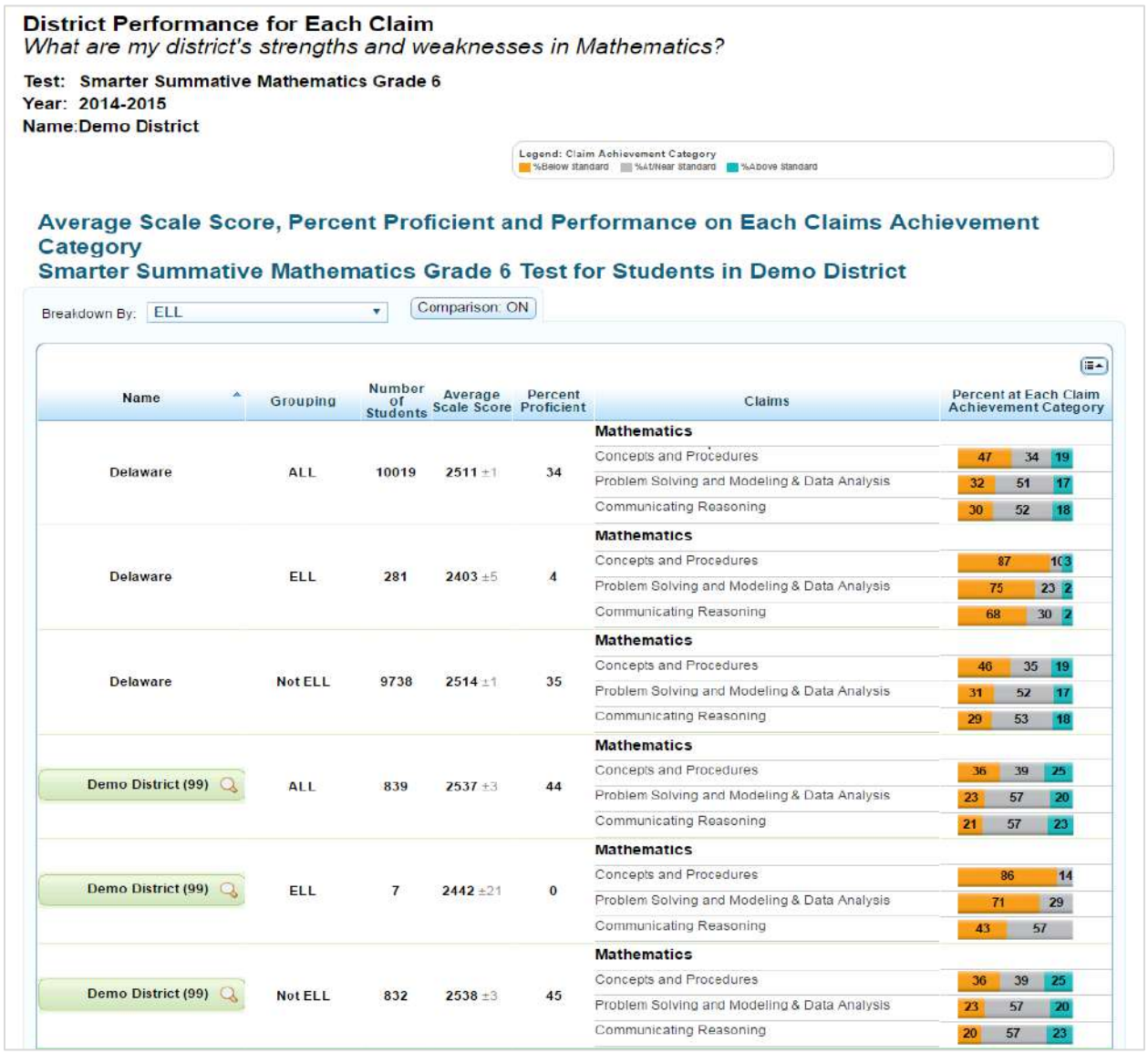
Name	Grouping	Number of Students	Average Scale Score	Percent Proficient	Percentage in Each Achievement Level
Delaware	ALL	9892	2524 ±1	49	23 28 33 16
Delaware	Female	4877	2540 ±1	55	17 27 35 20
Delaware	Male	5015	2508 ±1	42	29 29 30 12
Demo District (99)	ALL	754	2537 ±3	55	19 26 36 19
Demo District (99)	Female	354	2554 ±5	64	14 22 40 24
Demo District (99)	Male	400	2521 ±5	47	23 30 32 15
Demo School1 (999)	ALL	292	2557 ±6	62	18 20 29 33
Demo School1 (999)	Female	143	2577 ±8	71	13 16 34 38
Demo School1 (999)	Male	149	2537 ±9	53	24 23 25 28
Demo School2 (998)	ALL	272	2537 ±4	56	12 32 41 15
Demo School2 (998)	Female	119	2557 ±6	68	7 25 46 22
Demo School2 (998)	Male	153	2522 ±6	46	16 37 37 10
Demo School3 (997)	ALL	190	2505 ±6	44	29 27 38 5
Demo School3 (997)	Female	92	2515 ±8	48	26 26 41 7
Demo School3 (997)	Male	98	2496 ±9	40	33 28 36 4

7.1.2.3 Claim Detail Page

The Claim Detail Page provides the aggregate summaries on student performance in each claim for a particular grade and subject. The aggregate summaries on the Claim Detail Page include (1) number of students, (2) average scale score and standard error of the average scale score, (3) percent of proficient, and (4) percent of students in each achievement level.

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 state and the aggregate unit above the selected aggregate. Also, the summaries on claim-level performance can be presented for overall students and by subgroup. Exhibit 4 presents an example of Claim Detail Pages for mathematics at the district level when users select a subgroup of ELL.

Exhibit 4. Claim Detail Page for Mathematics by ELL: 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 Proficient and Performance on Each Claims Achievement Category
Smarter Summative Mathematics Grade 6 Test for Students in Demo District

Breakdown By: ELL Comparison: ON

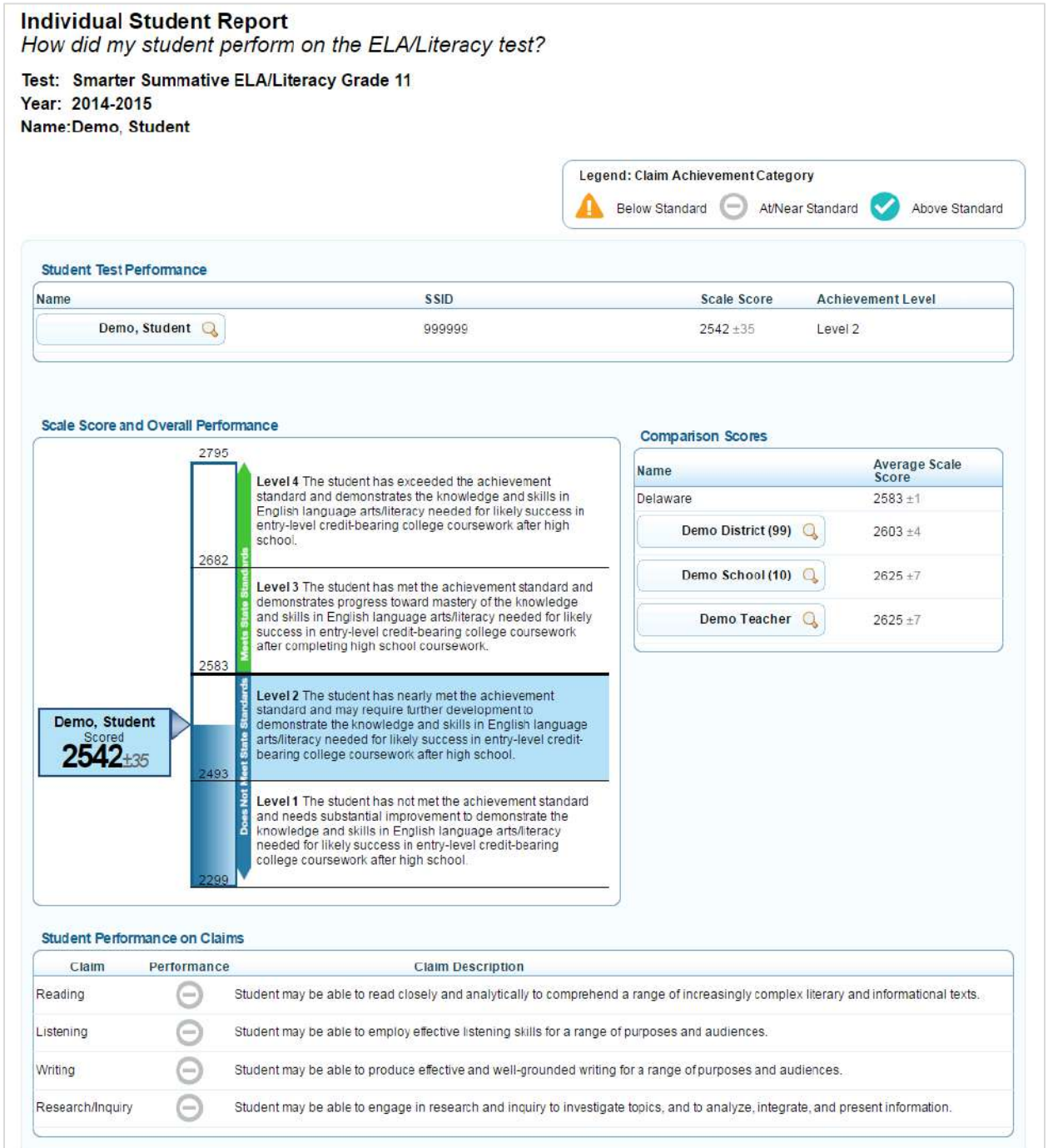
Name	Grouping	Number of Students	Average Scale Score	Percent Proficient	Claims	Percent at Each Claim Achievement Category
Delaware	ALL	10019	2511 ±1	34	**Mathematics**	
					Concepts and Procedures	47 34 19
					Problem Solving and Modeling & Data Analysis	32 51 17
					Communicating Reasoning	30 52 18
Delaware	ELL	281	2403 ±5	4	**Mathematics**	
					Concepts and Procedures	87 10 3
					Problem Solving and Modeling & Data Analysis	75 23 2
					Communicating Reasoning	68 30 2
Delaware	Not ELL	9738	2514 ±1	35	**Mathematics**	
					Concepts and Procedures	46 35 19
					Problem Solving and Modeling & Data Analysis	31 52 17
					Communicating Reasoning	29 53 18
Demo District (99)	ALL	839	2537 ±3	44	**Mathematics**	
					Concepts and Procedures	36 39 25
					Problem Solving and Modeling & Data Analysis	23 57 20
					Communicating Reasoning	21 57 23
Demo District (99)	ELL	7	2442 ±21	0	**Mathematics**	
					Concepts and Procedures	86 14
					Problem Solving and Modeling & Data Analysis	71 29
					Communicating Reasoning	43 57
Demo District (99)	Not ELL	832	2538 ±3	45	**Mathematics**	
					Concepts and Procedures	36 39 25
					Problem Solving and Modeling & Data Analysis	23 57 20
					Communicating Reasoning	20 57 23

7.1.2.4 Student Detail Page

When a student submits a completed test, an online score report appears in the Student Detail Page in the ORS. The Student Detail Page provides individual student performance on the test. In each subject area, the Student Detail Page provides (1) scale score and standard error of measurement, (2) achievement level for overall test, (3) achievement category in each claim, and (4) average scale scores for student's state, district, school, and teacher.

Specifically, on the top of the page, the student's name, scale score with standard error of measurement, and achievement level are presented. On the left middle section, the student's performance are described in detail using a barrel chart. In the barrel chart, the 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, the average scale scores and standard errors of the average scale scores for state, 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 or her performance on each of claims. Exhibits 5 and 6 present examples of Student Detail Pages for ELA/Lit and mathematics.

Exhibit 5. Student Detail Page for ELA/Lit



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

Legend: Claim Achievement Category
Below Standard | At/Near Standard | Above Standard

Student Test Performance

Name	SSID	Scale Score	Achievement Level
Demo, Student	999999	2542 ±35	Level 2

Scale Score and Overall Performance

2795
2682
2583
2493
2299

Meets State Standards
Does Not Meet State Standards

Level 4 The student has exceeded the achievement standard and demonstrates the knowledge and skills in English language arts/literacy needed for likely success in entry-level credit-bearing college coursework after high school.

Level 3 The student has met the achievement standard and demonstrates progress toward mastery of the knowledge and skills in English language arts/literacy needed for likely success in entry-level credit-bearing college coursework after completing high school coursework.

Level 2 The student has nearly met the achievement standard and may require further development to demonstrate the knowledge and skills in English language arts/literacy needed for likely success in entry-level credit-bearing college coursework after high school.

Level 1 The student has not met the achievement standard and needs substantial improvement to demonstrate the knowledge and skills in English language arts/literacy needed for likely success in entry-level credit-bearing college coursework after high school.

Demo, Student Scored 2542±35

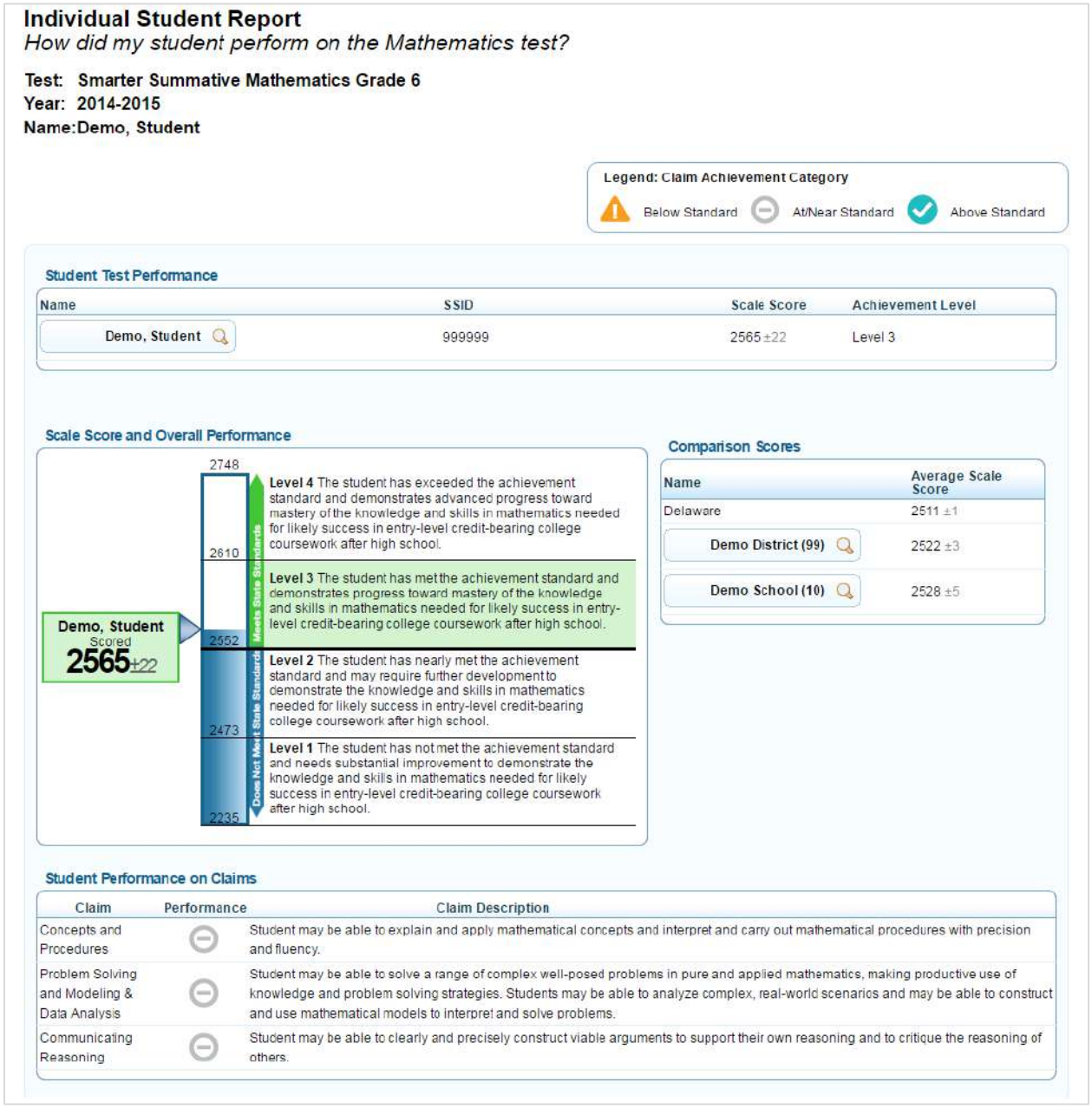
Comparison Scores

Name	Average Scale Score
Delaware	2583 ±1
Demo District (99)	2603 ±4
Demo School (10)	2625 ±7
Demo Teacher	2625 ±7

Student Performance on Claims

Claim	Performance	Claim Description
Reading		Student may be able to read closely and analytically to comprehend a range of increasingly complex literary and informational texts.
Listening		Student may be able to employ effective listening skills for a range of purposes and audiences.
Writing		Student may be able to produce effective and well-grounded writing 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.

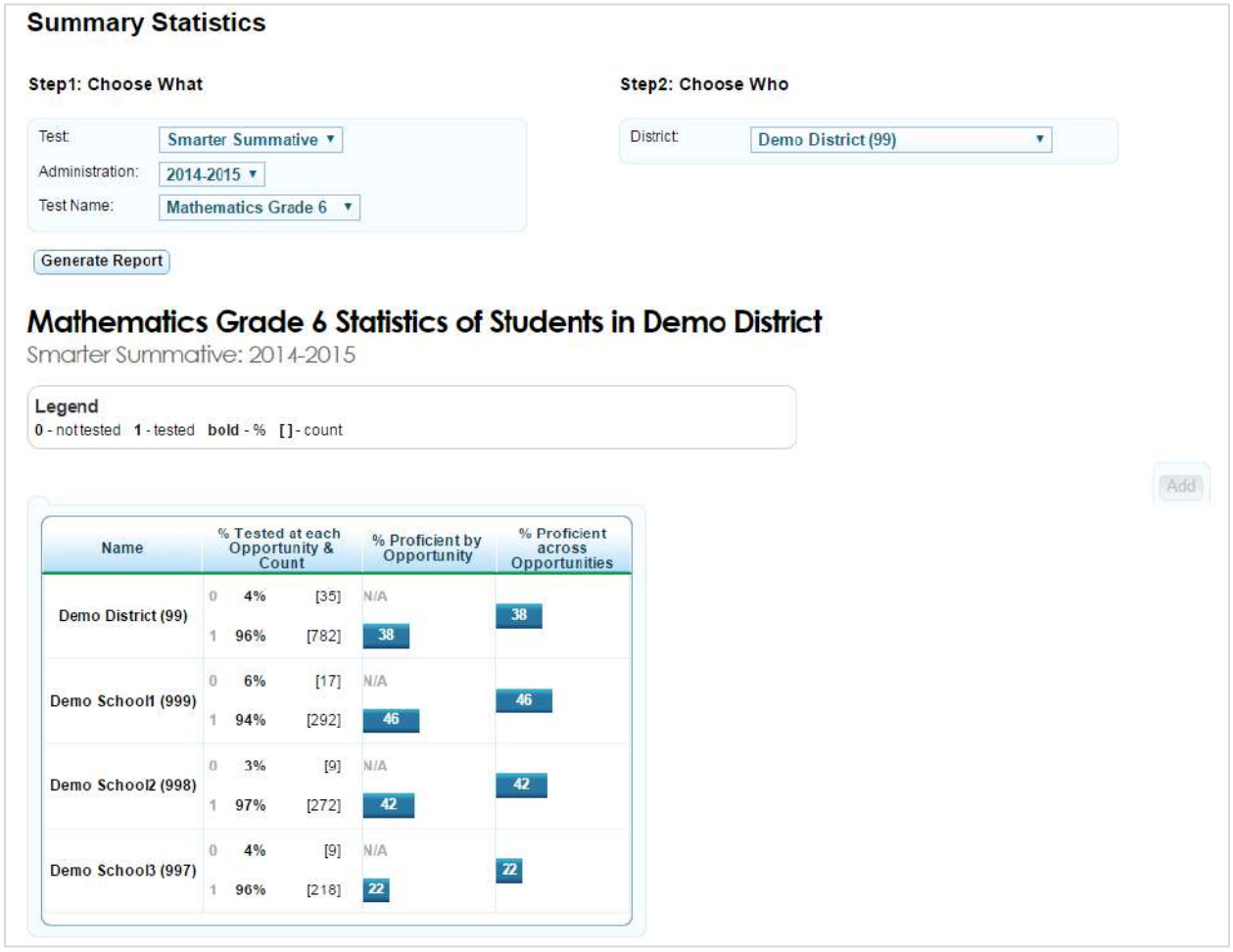
Exhibit 6. Student Detail Page for Mathematics



7.1.2.5 Participation Rate

In addition to online score reports, the ORS provides participation rate reports for the state, district, and school to help monitor student participation rate. Participation data are updated each time students complete tests and they are hand scored. Included in the participation table are (1) number and percent of students who are tested and not tested and (2) percent of students with achievement levels = 3 or 4. Exhibit 7 presents a sampled participation rate report at the district level.

Exhibit 7. Participation Rate Report at District Level

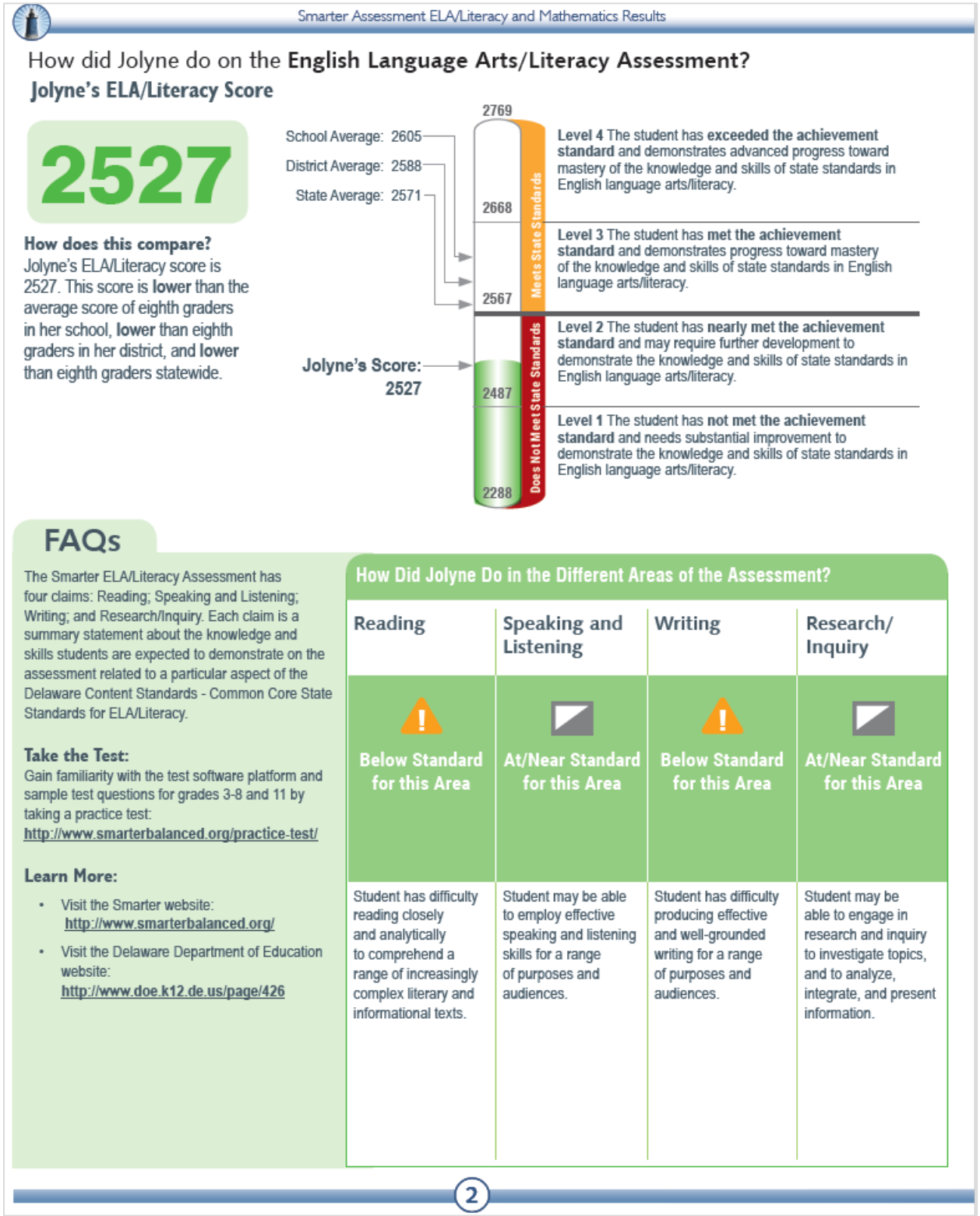


7.2 PAPER FAMILY SCORE REPORTS

After the testing window is closed, parents whose children participate in a test receive a full-color paper score report (hereafter referred to as family report) that includes their children's performance on ELA/Lit and mathematics. The family report includes information on student performance that is provided on the Student Detail Page from the ORS with additional guidance on how to interpret student achievement results in the family report.

An example of a family report is shown in Exhibit 8 and can be found online at http://de.portal.airast.org/wp-content/uploads/DE_SBAC_Family_Guide.pdf

Exhibit 8. Sample Paper Family Score Report



Smarter Assessment ELA/Literacy and Mathematics Results

How did Jolyne do on the **English Language Arts/Literacy Assessment?**

Jolyne's ELA/Literacy Score

2527

How does this compare?
Jolyne's ELA/Literacy score is 2527. This score is **lower** than the average score of eighth graders in her school, **lower** than eighth graders in her district, and **lower** than eighth graders statewide.

School Average: 2605
District Average: 2588
State Average: 2571
Jolyne's Score: 2527

2769
2668
2567
2487
2288
Meets State Standards
Does Not Meet State Standards

Level 4 The student has **exceeded the achievement standard** and demonstrates advanced progress toward mastery of the knowledge and skills of state standards in English language arts/literacy.

Level 3 The student has **met the achievement standard** and demonstrates progress toward mastery of the knowledge and skills of state standards in English language arts/literacy.

Level 2 The student has **nearly met the achievement standard** and may require further development to demonstrate the knowledge and skills of state standards in English language arts/literacy.

Level 1 The student has **not met the achievement standard** and needs substantial improvement to demonstrate the knowledge and skills of state standards in English language arts/literacy.

FAQs

The Smarter ELA/Literacy Assessment has four claims: Reading; Speaking and Listening; Writing; and Research/Inquiry. Each claim is a summary statement about the knowledge and skills students are expected to demonstrate on the assessment related to a particular aspect of the Delaware Content Standards - Common Core State Standards for ELA/Literacy.

Take the Test:
Gain familiarity with the test software platform and sample test questions for grades 3-8 and 11 by taking a practice test:
http://www.smarterbalanced.org/practice-test/

Learn More:

- Visit the Smarter website: http://www.smarterbalanced.org/
- Visit the Delaware Department of Education website: http://www.doe.k12.de.us/page/426

How Did Jolyne Do in the Different Areas of the Assessment?

Reading	Speaking and Listening	Writing	Research/ Inquiry
Below Standard for this Area	At/Near Standard for this Area	Below Standard for this Area	At/Near Standard for this Area
Student has difficulty reading closely and analytically to comprehend a range of increasingly complex literary and informational texts.	Student may be able to employ effective speaking and listening skills for a range of purposes and audiences.	Student has difficulty producing effective and well-grounded writing for a range of purposes and audiences.	Student may be able to engage in research and inquiry to investigate topics, and to analyze, integrate, and present information.

2

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 the 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 the 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 are a description of content area knowledge and skills that examinees at 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/Lit, for instance, achievement-level descriptors are described for Level 3 as "students demonstrate progress toward mastery of the knowledge and skills ELA/Lit needed for likely success in future coursework." Generally, students performing Smarter Balanced assessments 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 Aggregated Score

Students' scale scores are aggregated at roster, teacher, school, district, and state 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 at the teacher and school level provide information regarding 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 some claims. In this case, teachers or schools can identify strengths and weaknesses of their students through the group performance by claim and promote instruction on specific claim areas. 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 a disadvantaged subgroup. For example, teachers can see student assessment results by ELL status and observe that ELL students are struggling with literary response and analysis in reading. Teachers can then provide additional instructions for these students to enhance their achievement in a specific claim.

In addition, assessment results can be used to compare students' performance among different students and among different groups. Teachers can evaluate how their students perform compared with other students in schools and districts states 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 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 assessment development, administration, and scoring and reporting of results. AIR implements a series of quality control steps to ensure error-free production of score reports in both online and paper format. The quality of the information produced in the Test Delivery System (TDS) is tested thoroughly before, during, and after the testing window.

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 before the testing window.

To verify the accuracy of the scoring engine, we use simulated test administrations. The simulator generates a sample of students with an ability distribution that matches that of the population (Smarter Balanced 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 (computer adaptive test component plus a fixed-form performance task component) are performed to check scores. The simulated data are used to check whether the scoring specifications were applied accurately. The scores in the simulated data file are checked independently, following the scoring rules specified in the scoring specifications.

8.1.1 Platform Review

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

Platform review is a process in which each item is checked to ensure that it is displayed appropriately on each tested platform. A platform is a combination of a hardware device and an operating system. In recent

years, the number of platforms has proliferated, and platform review now takes place on various platforms that are significantly different from one another.

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

8.1.2 User Acceptance Testing and Final Review

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

8.2 Quality Assurance in Document Processing

Scanning Accuracy

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

8.3 Quality Assurance in Data Preparation

AIR's TDS has a real-time quality-monitoring component built in. After a test is administered to a student, the TDS passes the resulting data to our Quality Assurance (QA) system. QA conducts a series of data integrity checks, ensuring, for example, that the record for each test contains information for each item, keys for multiple-choice items, score points in each item, total number of field-test items and operation items, and ensuring 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 DDOE. AIR staff ensure that data in the extract files match the DoR before delivering to DDOE.

8.4 Quality Assurance in Hand-scoring

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

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

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

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

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

In addition to using validity responses as a qualification threshold, other validity responses are presented throughout scoring as ongoing checks for quality. Validity responses can be culled from approved existing anchor or validity responses, but they also may be generated from live scoring and included in the pool following review and approval by the Smarter Balanced Assessment Consortium. MI periodically administers validity sets to each of MI's scorers supporting the scoring effort. 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 Hand-scoring QA Monitoring Reports

MI generates detailed scorer status reports for each scoring project using a comprehensive system for collecting and analyzing score data. The scores are validated and processed according to the specifications set out by Smarter Balanced. This allows MI to manage the quality of the scorers and take any corrective actions immediately. Updated real-time reports are available that show both daily and cumulative (project-to-date) data. These reports are available to 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

DDOE also directly observes MI activities, virtually. MI provides virtual access to the training activities through the online training interface. DDOE 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 Test Delivery System during the test administration window, AIR statisticians examine the delivery demands, including the number of tests to be delivered, the length of the window, and the historic state-specific behaviors to model the likely peak loads. Using data from the load tests, these calculations indicate the number of each type of server necessary to provide continuous, responsive service, and AIR contracts for service in excess of this amount. Once deployed, our servers are monitored at the hardware, operating system, and software platform levels with monitoring software that alerts our engineers at the first signs that trouble may be ahead. The applications log not only errors and exceptions, but also 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 are captured for each assessed student, such as data about how long it takes to load, view, or respond to an item. All of this information is logged, enabling us to automatically identify schools or districts experiencing unusual slowdowns, often before they even notice.

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

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

For the computer 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 41 presents an overview of the quality assurance (QA) reports.

Table 41. 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 hand-scoring processes include rigorous training, validity and reliability monitoring, and back-reading to ensure accurate scoring. Hand-scored items are paired to the machine-scored items by our Test Integration System (TIS). The integration is based on identifiers that are never separated from their data and are checked by our quality assurance (QA) system. The integrated scores are sent to our test-scoring system, a mature, well-tested real-time system that applies client-specific scoring rules and assigns scores from the calibrated items, including calculating achievement-level indicators, subscale scores and other features, which then pass automatically to the reporting system and Database of Record (DoR). The scoring system is tested extensively before deployment, including hand checks of scored tests and large-scale simulations to ensure that point estimates and standard errors are correct.

Every test undergoes a series of validation checks. Once the QA system signs off, data are passed to the DoR, which serves as the centralized location for all student scores and responses, ensuring that there is only one place where the "official" record is stored. Only after scores have passed the QA checks and are uploaded to the DoR are they passed to the Online Reporting System (ORS), which is responsible for

presenting individual-level results and calculating and presenting aggregate results. Absolutely no score is reported in the ORS until it passes all of the QA system's validation checks. All of the above processes take milliseconds to complete so that within less than a second of hand-scores being received by AIR and passing QA validation checks, the composite score will be available in the ORS.

8.5.1.2 Paper Report Quality Assurance

Statistical Programming

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

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

Each change is followed by a complete retesting with the entire collection of scenarios on which the macro was originally tested. The main statistical program is mostly made up of calls to various macros, including macros that verify the data and conversion tables and the macros that do the many complex 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.

REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

APPENDICES

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

Table A-1. School Year 2014-2015 Grade 3 ELA/Lit 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	10,231	2438.10	84.73	21	25	25	29	54
Gender								
Female	5,122	2448.12	83.93	17	24	26	33	59
Male	5,109	2428.05	84.34	24	27	24	25	49
Ethnicity								
American Indian or Alaska Native	38	2460.64	77.44	11	13	45	32	76
Asian	375	2496.58	79.24	6	14	26	54	80
African American	3,016	2405.67	81.55	33	28	23	16	39
Hispanic	1,763	2415.31	75.73	27	32	24	17	41
White	4,631	2462.79	80.60	12	22	27	40	66
ELL Program								
ELL	984	2382.53	64.47	40	37	19	4	23
Not ELL	9,247	2444.01	84.48	19	24	26	31	58
Special Education								
Special Education	1,279	2351.30	70.03	59	28	11	3	13
Not Special Education	8,952	2450.50	79.21	15	25	27	33	60
504 Plan								
504 Plan	332	2424.21	73.41	20	36	26	18	44
No 504 Plan	9,899	2438.56	85.05	21	25	25	29	55
Title I								
Title I	1,161	2438.58	76.12	18	28	29	25	54
Not Title I	9,070	2438.04	85.77	21	25	25	29	54

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/Lit 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	9,910	2477.39	88.02	25	21	25	29	54
Gender								
Female	4,932	2486.57	86.63	22	21	26	32	58
Male	4,978	2468.30	88.44	29	22	24	25	49
Ethnicity								
American Indian or Alaska Native	43	2494.06	80.09	16	19	33	33	65
Asian	385	2541.10	83.52	10	8	21	60	81
African American	3,060	2444.39	82.78	37	25	23	15	37
Hispanic	1,702	2452.79	78.70	32	28	23	16	40
White	4,331	2503.85	83.66	16	17	28	40	68
ELL Program								
ELL	558	2399.59	69.61	61	25	10	4	14
Not ELL	9,352	2482.03	86.82	23	21	26	30	56
Special Education								
Special Education	1,349	2380.08	71.93	70	18	8	3	11
Not Special Education	8,561	2492.73	80.16	18	22	28	33	60
504 Plan								
504 Plan	376	2471.67	75.37	24	25	30	21	51
No 504 Plan	9,534	2477.62	88.47	25	21	25	29	54
Title I								
Title I	1,274	2467.85	80.05	26	25	27	22	49
Not Title I	8,636	2478.80	89.05	25	21	25	30	54

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/Lit 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	9,922	2509.37	89.30	24	21	34	22	55
Gender								
Female	4,890	2522.73	86.73	19	20	35	26	61
Male	5,032	2496.39	89.85	28	22	32	18	50
Ethnicity								
American Indian or Alaska Native	41	2518.44	86.60	15	27	37	22	59
Asian	361	2579.08	83.62	6	10	31	52	84
African American	3,115	2473.80	85.01	37	25	29	10	39
Hispanic	1,533	2486.31	79.41	31	25	32	12	44
White	4,585	2534.93	84.24	14	18	37	31	68
ELL Program								
ELL	303	2409.17	65.35	71	21	8	1	9
Not ELL	9,619	2512.53	88.12	22	21	34	23	57
Special Education								
Special Education	1,381	2408.15	70.63	71	19	10	1	11
Not Special Education	8,541	2525.74	80.83	16	21	37	25	63
504 Plan								
504 Plan	412	2502.12	82.61	23	27	33	17	50
No 504 Plan	9,510	2509.68	89.57	24	21	34	22	56
Title I								
Title I	1,621	2510.48	84.70	22	21	36	21	56
Not Title I	8,301	2509.15	90.17	24	21	33	22	55

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/Lit 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	10,023	2522.78	92.41	24	28	32	16	48
Gender								
Female	4,943	2538.94	89.09	18	27	35	20	55
Male	5,080	2507.05	92.87	29	29	30	12	41
Ethnicity								
American Indian or Alaska Native	48	2536.05	81.71	15	33	33	19	52
Asian	352	2597.36	82.96	7	13	35	45	80
African American	3,097	2490.39	87.34	34	33	26	7	33
Hispanic	1,601	2498.67	87.25	30	32	29	8	38
White	4,694	2546.31	88.39	16	26	37	22	59
ELL Program								
ELL	247	2409.07	71.95	74	21	4	0	5
Not ELL	9,776	2525.65	91.05	22	29	33	16	49
Special Education								
Special Education	1,389	2422.51	75.51	68	24	7	1	8
Not Special Education	8,634	2538.91	84.37	16	29	36	18	55
504 Plan								
504 Plan	416	2513.45	84.14	23	34	34	9	43
No 504 Plan	9,607	2523.18	92.73	24	28	32	16	48
Title I								
Title I	1,814	2515.82	86.12	24	30	34	12	45
Not Title I	8,209	2524.32	93.67	23	28	32	17	49

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/Lit 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	9,716	2547.11	96.00	25	25	35	15	50
Gender								
Female	4,735	2564.35	92.48	19	24	38	19	58
Male	4,981	2530.71	96.42	30	27	32	11	43
Ethnicity								
American Indian or Alaska Native	52	2553.58	92.61	23	27	31	19	50
Asian	354	2621.67	90.92	8	11	36	45	81
African American	3,068	2509.28	89.28	37	30	27	6	33
Hispanic	1,453	2521.82	89.96	31	30	31	8	39
White	4,555	2574.73	90.53	15	21	42	22	63
ELL Program								
ELL	285	2433.30	74.05	74	17	8	1	9
Not ELL	9,431	2550.54	94.48	23	25	36	16	51
Special Education								
Special Education	1,328	2445.75	74.50	69	23	7	1	8
Not Special Education	8,388	2563.15	88.95	18	25	40	17	57
504 Plan								
504 Plan	351	2535.62	85.44	26	30	34	10	44
No 504 Plan	9,365	2547.54	96.35	25	25	35	15	50
Title I								
Title I	1,902	2542.82	92.06	24	26	38	12	50
Not Title I	7,814	2548.15	96.91	25	25	34	16	50

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/Lit 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	9,546	2559.13	97.90	24	27	35	14	49
Gender								
Female	4,669	2576.06	93.72	18	26	38	18	56
Male	4,877	2542.93	99.07	30	28	32	11	43
Ethnicity								
American Indian or Alaska Native	38	2600.05	92.83	13	21	39	26	66
Asian	328	2634.72	92.04	7	13	41	39	80
African American	3,109	2521.52	91.16	36	32	27	5	33
Hispanic	1,267	2533.88	89.66	31	31	31	7	38
White	4,574	2585.19	93.47	16	24	40	20	60
ELL Program								
ELL	258	2454.18	76.42	68	25	6	1	7
Not ELL	9,288	2562.05	96.82	23	27	36	14	50
Special Education								
Special Education	1,350	2459.65	77.48	66	25	9	1	10
Not Special Education	8,196	2575.52	90.97	17	28	39	16	55
504 Plan								
504 Plan	404	2551.31	88.22	25	31	34	10	44
No 504 Plan	9,142	2559.48	98.30	24	27	35	14	49
Title I								
Title I	1,957	2545.16	94.44	28	30	32	10	42
Not Title I	7,589	2562.74	98.46	23	26	36	15	51

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/Lit 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	7,497	2581.57	112.79	24	24	31	21	52
Gender								
Female	3,721	2601.67	106.02	17	23	34	25	60
Male	3,776	2561.76	115.73	30	26	27	17	44
Ethnicity								
American Indian or Alaska Native	36	2570.91	118.38	31	17	33	19	53
Asian	283	2638.72	116.11	14	14	29	42	72
African American	2,315	2548.33	106.41	32	28	29	12	40
Hispanic	854	2554.93	104.50	29	30	29	13	42
White	3,892	2603.00	111.07	18	22	32	27	60
ELL Program								
ELL	138	2465.96	89.36	65	22	12	0	12
Not ELL	7,359	2583.73	112.06	23	25	31	22	53
Special Education								
Special Education	765	2475.32	86.71	59	29	10	2	12
Not Special Education	6,732	2593.64	109.02	20	24	33	23	56
504 Plan								
504 Plan	258	2572.81	115.88	27	26	27	20	47
No 504 Plan	7,239	2581.88	112.67	24	24	31	21	52
Title I								
Title I	810	2542.05	103.93	34	30	26	10	36
Not Title I	6,687	2586.35	112.89	22	24	31	23	54

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	10,268	2439.39	75.47	21	26	32	21	53
Gender								
Female	5,150	2439.88	73.28	20	27	33	20	53
Male	5,118	2438.90	77.60	22	25	32	21	53
Ethnicity								
American Indian or Alaska Native	38	2460.11	68.46	16	18	34	32	66
Asian	391	2499.64	75.27	5	15	28	52	80
African American	3,026	2408.36	70.77	33	31	27	9	36
Hispanic	1,784	2420.21	67.68	26	33	30	11	41
White	4,620	2462.02	71.35	12	21	37	30	67
ELL Program								
ELL	1,032	2395.36	63.49	40	36	21	4	25
Not ELL	9,236	2444.31	75.10	19	25	34	23	56
Special Education								
Special Education	1,280	2360.01	72.88	61	25	12	2	14
Not Special Education	8,988	2450.70	68.74	15	26	35	23	59
504 Plan								
504 Plan	333	2432.72	67.90	23	29	34	14	48
No 504 Plan	9,935	2439.62	75.70	21	26	32	21	53
Title I								
Title I	1,163	2440.77	62.47	18	28	36	18	54
Not Title I	9,105	2439.22	76.97	21	26	32	21	53

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	9,995	2476.86	75.44	19	35	29	17	47
Gender								
Female	4,970	2475.63	71.94	18	36	29	16	45
Male	5,025	2478.07	78.74	19	33	29	18	48
Ethnicity								
American Indian or Alaska Native	43	2495.29	64.66	9	35	35	21	56
Asian	401	2539.89	73.18	6	16	28	50	78
African American	3,063	2446.50	69.82	29	42	22	7	29
Hispanic	1,736	2457.00	68.13	24	40	27	9	36
White	4,362	2499.42	71.46	11	29	36	25	60
ELL Program								
ELL	613	2419.94	67.74	46	38	13	3	16
Not ELL	9,382	2480.58	74.42	17	34	31	18	49
Special Education								
Special Education	1,355	2393.13	66.91	63	29	7	2	8
Not Special Education	8,640	2489.99	67.90	12	36	33	20	53
504 Plan								
504 Plan	377	2470.55	66.12	18	42	29	11	40
No 504 Plan	9,618	2477.11	75.78	19	34	29	17	47
Title I								
Title I	1,279	2477.80	67.21	15	39	31	15	46
Not Title I	8,716	2476.72	76.58	19	34	29	18	47

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	10,017	2498.56	84.99	31	31	20	18	38
Gender								
Female	4,935	2498.81	82.06	30	33	19	18	37
Male	5,082	2498.32	87.74	31	30	20	19	39
Ethnicity								
American Indian or Alaska Native	41	2499.18	79.64	24	41	17	17	34
Asian	375	2573.75	82.16	9	17	22	52	74
African American	3,148	2461.03	79.88	48	31	14	7	21
Hispanic	1,565	2477.04	75.08	39	34	16	10	27
White	4,602	2524.87	78.65	18	32	24	26	50
ELL Program								
ELL	346	2416.48	70.57	74	18	5	3	8
Not ELL	9,671	2501.50	83.99	29	32	20	19	39
Special Education								
Special Education	1,390	2409.40	69.80	76	19	4	2	5
Not Special Education	8,627	2512.93	78.19	23	33	22	21	43
504 Plan								
504 Plan	409	2493.91	77.20	29	42	15	13	29
No 504 Plan	9,608	2498.76	85.30	31	31	20	19	38
Title I								
Title I	1,628	2500.69	83.30	30	32	20	18	38
Not Title I	8,389	2498.15	85.31	31	31	20	18	38

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	10,084	2510.54	96.32	33	32	19	15	34
Gender								
Female	4,981	2515.38	92.53	32	33	19	16	35
Male	5,103	2505.81	99.66	35	32	18	14	33
Ethnicity								
American Indian or Alaska Native	48	2518.76	89.90	33	29	21	17	38
Asian	358	2598.65	94.57	10	21	21	48	69
African American	3,111	2470.56	87.71	50	33	12	6	17
Hispanic	1,635	2486.02	90.22	42	36	14	8	22
White	4,701	2538.01	90.70	22	32	25	21	46
ELL Program								
ELL	291	2402.40	84.44	80	15	2	2	4
Not ELL	9,793	2513.75	94.78	32	33	19	16	35
Special Education								
Special Education	1,405	2404.90	82.56	80	16	3	1	4
Not Special Education	8,679	2527.64	87.04	26	35	22	17	39
504 Plan								
504 Plan	417	2506.64	83.83	34	38	16	12	28
No 504 Plan	9,667	2510.70	96.82	33	32	19	15	34
Title I								
Title I	1,826	2505.41	87.06	34	36	19	11	30
Not Title I	8,258	2511.67	98.21	33	32	19	16	35

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	9,754	2529.61	102.70	31	32	22	15	37
Gender								
Female	4,753	2535.14	99.28	28	33	23	15	39
Male	5,001	2524.36	105.59	34	31	20	15	35
Ethnicity								
American Indian or Alaska Native	52	2529.68	94.41	27	44	13	15	29
Asian	360	2622.93	107.92	12	17	21	50	71
African American	3,064	2486.74	93.77	47	34	14	5	19
Hispanic	1,490	2501.08	97.83	40	34	18	8	26
White	4,556	2560.23	94.38	19	31	28	21	50
ELL Program								
ELL	334	2416.24	90.76	78	17	4	1	5
Not ELL	9,420	2533.63	100.79	29	33	22	16	38
Special Education								
Special Education	1,324	2419.07	86.62	78	18	3	1	4
Not Special Education	8,430	2546.97	93.84	24	34	25	17	42
504 Plan								
504 Plan	350	2528.18	90.55	31	37	21	12	33
No 504 Plan	9,404	2529.67	103.13	31	32	22	15	37
Title I								
Title I	1,912	2521.75	94.27	31	36	23	10	33
Not Title I	7,842	2531.53	104.57	31	31	22	16	38

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	9,512	2541.72	111.97	37	28	19	17	35
Gender								
Female	4,646	2547.30	106.60	34	30	19	17	36
Male	4,866	2536.40	116.63	39	26	18	16	35
Ethnicity								
American Indian or Alaska Native	38	2560.00	120.30	29	29	24	18	42
Asian	329	2647.56	116.08	12	17	21	50	71
African American	3,091	2491.38	97.34	54	29	12	5	17
Hispanic	1,264	2516.39	101.03	45	28	18	9	27
White	4,558	2574.54	106.87	25	28	23	24	47
ELL Program								
ELL	267	2442.12	101.95	78	13	6	3	9
Not ELL	9,245	2544.60	110.93	36	29	19	17	36
Special Education								
Special Education	1,350	2435.10	86.27	79	17	3	1	5
Not Special Education	8,162	2559.36	105.78	30	30	21	19	40
504 Plan								
504 Plan	402	2540.62	98.95	36	33	16	14	31
No 504 Plan	9,110	2541.77	112.51	37	28	19	17	35
Title I								
Title I	1,943	2531.00	104.35	39	30	19	12	30
Not Title I	7,569	2544.48	113.69	36	28	19	18	36

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	7,521	2541.14	119.80	52	25	15	8	23
Gender								
Female	3,731	2547.73	114.08	49	27	16	7	24
Male	3,790	2534.66	124.85	54	23	14	8	23
Ethnicity								
American Indian or Alaska Native	37	2532.00	119.40	49	27	22	3	24
Asian	284	2647.91	126.77	22	21	24	33	57
African American	2,321	2496.83	103.41	67	22	8	2	10
Hispanic	860	2512.95	106.49	62	24	11	3	14
White	3,906	2565.92	119.70	42	27	20	11	31
ELL Program								
ELL	141	2456.13	103.11	85	9	4	2	6
Not ELL	7,380	2542.77	119.52	51	25	16	8	24
Special Education								
Special Education	778	2438.08	82.48	90	8	2	0	2
Not Special Education	6,743	2553.03	117.72	47	27	17	9	26
504 Plan								
504 Plan	256	2541.84	121.36	52	25	14	9	23
No 504 Plan	7,265	2541.12	119.75	52	25	15	8	23
Title I								
Title I	816	2503.62	99.59	67	22	9	2	11
Not Title I	6,705	2545.71	121.25	50	26	16	9	25

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	616	3	619	605	10	615
4	482	2	484	516	0	516
5	518	0	518	528	0	528
6	273	0	273	450	0	450
7	287	0	287	382	0	382
8	269	0	269	328	0	328
11	235	0	235	531	0	531

For the Interim Assessment Blocks (IAB), there were seven IABs for ELA/Lit 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/Lit, a total of 897 students took IABs, and among 897 students, 505 students took one IAB, 301 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/Lit and four IABs in mathematics. For example, 505 students in grade 3 ELA/Lit took one IAB only. Among 505 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	897	505	301	82	9			
4	679	392	258	23	6			
5	831	469	133	147	82			
6	366	186	122	47	11			
7	176	153	14	9				
8	269	196	48	24	1			
11	698	678	20					
Mathematics								
3	1,380	841	337	200	2			
4	1,400	931	217	250	2			
5	1,399	1014	157	223	5			
6	1,006	635	231	140				
7	1,269	1088	52	129				
8	1,168	874	131	163				
11	1,199	984	152	63				

Table B–3: ELA/Lit 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						
	Editing and Revising	58	139	81	9			
	Listening and Interpretation	117	134	60	9			
	Performance Task	137	93					
	Reading Informational Text	111	99	25	3			
	Reading Literary Text	60	3	11	8			
	Research	20	134	69	7			
4	Brief Writes							
	Editing and Revising	29	111	21	6			
	Listening and Interpretation	41	179	21	6			
	Performance Task	136	58					
	Reading Informational Text	105	72	3				
	Reading Literary Text	60	3	2	6			
	Research	21	93	22	6			
5	Brief Writes							
	Editing and Revising	37	68	134	82			
	Listening and Interpretation	39	70	146	82			
	Performance Task	242						
	Reading Informational Text	63	23	6	42			
	Reading Literary Text		4	12	40			
	Research	88	101	143	82			
6	Brief Writes	1	7	2	1			
	Editing and Revising	20	91	45	11			
	Listening and Interpretation	77	119	47	11			
	Performance Task	59		1	1			
	Reading Informational Text	27	18	1	8			
	Reading Literary Text	2		2	3			
	Research		9	43	9			
7	Brief Writes							
	Editing and Revising	18	3	9				
	Listening and Interpretation	31	13	9				
	Performance Task	69						
	Reading Informational Text	23	4					
	Reading Literary Text	5	1					
	Research	7	7	9				
8	Brief Writes							
	Editing and Revising	35	2	23	1			
	Listening and Interpretation	41	47	24	1			
	Performance Task	59			1			
	Reading Informational Text	55	25	1				
	Reading Literary Text	1						
	Research	5	22	24	1			
11	Brief Writes							
	Editing and Revising	6	19					
	Listening and Interpretation	6	4					
	Performance Task	233						
	Reading Informational Text	16						
	Reading Literary Text							
	Research	417	17					

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	56	122	194	2
	Number and Operations – Fractions	109	122	199	2
	Operational and Algebraic Thinking	498	300	198	2
	Performance Task	178	130	9	2
4	Number and Operations in Base Ten	80	75	250	2
	Number and Operations – Fractions	53	120	250	2
	Operational and Algebraic Thinking	588	169	248	2
	Performance Task	210	70	2	2
5	Measurement and Data	84	87	223	5
	Number and Operations in Base Ten	675	77	222	5
	Number and Operations – Fractions	89	129	223	5
	Performance Task	166	21	1	5
6	Expressions and Equations	67	189	140	
	Geometry	50	75	140	
	Performance Task	101	5		
	Ratios and Proportional Relationships	417	193	140	
7	Expressions and Equations	234	14	129	
	The Number System	642	47	129	
	Performance Task	142	9		
	Ratios and Proportional Relationships	70	34	129	
8	Expressions and Equations	469	109	163	
	Functions	84	62	163	
	Geometry	244	43	163	
	Performance Task	77	48		
11	Algebra – Linear Functions	301	87	63	
	Algebra – Quadratic Functions	406	113	62	
	Geometry – Right Triangles and Trigonometric Ratios	263	100	63	
	Performance Task	14	4	1	