Delaware Smarter Balanced Assessments

2016–2017 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 (SBAC) developed a next-generation assessment system. The assessments are designed to measure the Common Core State Standards (CCSS) in English language arts/literacy (ELA/Lit) and mathematics for grades 3–8 and 11, and to provide valid, reliable, and fair test scores about student academic achievement. Delaware was among the 18 member states (plus the U.S. Virgin Islands) leading the development of assessments in ELA/Lit and mathematics.The system includes both summative assessments, for accountability purposes, and optional interim assessments that provide meaningful feedback and actionable data that teachers and educators can use to help students succeed. Smarter Balanced, a state-led enterprise, is intended to provide leadership and resources to improve teaching and learning by creating and maintaining a suite of summative and interim assessments and tools aligned to the CCSS in ELA/Lit and mathematics.

The Delaware State Board of Education formally adopted the CCSS in ELA/Lit and mathematics on August 19, 2010 (State Board meeting minutes, 2010). Delaware CCSS define the knowledge and skills that students need to succeed in college and careers after graduating from high school. These standards include rigorous content and application of knowledge through higher-order skills and align with college and workforce expectations.

Since the adoption of the CCSS in 2010, the Delaware Department of Education fully implemented the CCSS in all grade levels in SY 2013–2014. The new 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 schools. In 2015–2016, Delaware adopted the SAT to replace the Smarter Balanced grade 11 assessments for high school students. The American Institutes for Research (AIR) delivered and scored the Smarter Balanced assessments and produced score reports. Measurement Incorporated (MI) scored the handscored items.

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

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

Optional interim assessments allow teachers to check student progress throughout the year, giving them information that they can use to improve instruction and learning. 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 specific sets of related concepts and provide more detailed information about student learning.

This report provides a technical summary of the 2016–2017 summative assessments in ELA/Lit and mathematics administered in grades 3–8 under the Delaware Smarter Balanced assessments. The report includes eight chapters: overview, test administration, summary of 2016–2017 operational administration, validity, reliability, scoring, reporting and interpreting scores, and quality control process. The data included in this report are based on Delaware data for the summative assessment only. For the interim assessments, the number of students who took ICAs and IABs is provided in Appendix A.

While this report includes information on all aspects of the technical quality of the Smarter Balanced test administration for 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 Smarter Balanced technical report.

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

2. TEST ADMINISTRATION

2.1 TESTING WINDOWS

The 2016–2017 Delaware Smarter Balanced assessment testing window spanned approximately three months for grades 3–8 for the online summative assessments and the full school year for the interim assessments. The paper-pencil fixed-form summative assessments were administered for 15 days during the online summative testing window. Table 1 shows the testing windows for both online and paper-pencil assessments.

Table 1. 2016–2017 Testing Windows

Tests	Grades	Start Date	End Date	Mode
Summative Assessments	3–8	03/08/2017	06/01/2017	Online Adaptive
	3–8	04/25/2017	05/12/2017	Paper Fixed-Form
Interim Comprehensive Assessments	3–8	08/29/2016	07/14/2017	Online Fixed-Form
Interim Assessment Blocks	3–8	08/29/2016	07/14/2017	Online Fixed-Form

2.2 TEST OPTIONS AND ADMINISTRATIVE ROLES

Smarter Balanced assessments are administered primarily online. To ensure that all eligible students in tested grades were given the opportunity to take the Smarter Balanced assessments, a number of assessment options were available for the 2016–2017 administration to accommodate students' needs. Table 2 lists the testing options that were offered in 2016–2017. Testing options are selected by content area. Once an option is selected, it applies to all tests in the content area.

Table 2. Testing Options in 2016–2017

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

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

2.2.1 Administrative Roles

The key personnel involved with test administration are District Test Coordinators (DTCs), District Accommodations Managers (DAMs), School Test Coordinators (STCs), and Test Administrators (TAs). The main responsibilities of these key personnel are described below. More detailed descriptions can be found in TAM, provided online at the Delaware System of Student Assessments (DeSSA) portal, http://de.portal.airast.org.

District Test Coordinator (DTC)

DTCs are responsible for coordinating testing in their district. They ensure that STCs and TAs in their districts are appropriately trained and aware of policies and procedures. DTCs also ensure that their STCs are trained in the reporting system.

DTCs responsibilities include the following:

- Oversee all test administration-related activities in the district
- Complete all required DeSSA trainings
- Complete all required DeSSA security forms
- Finalize testing schedules and requirements with STCs
- Ensure that all STCs and TAs are trained to properly administer the Smarter Balanced assessments
- Ensure that all STCs and TAs understand the protocols in the event that a student moves to a new district and/or school
- Ensure that all STCs and TAs are appropriately trained regarding the test security policies and procedures
- Ensure that all STCs and TAs understand the procedures to submit incidents, exemptions, and data review to Delaware Department of Education (DDOE) from the KACE/DOE Help Desk
- Ensure that all STCs and TAs have completed DeSSA security forms
- Create and manage appeals through the Test Information Distribution Engine (TIDE)

District Accommodations Manager (DAM)

DAMs are responsible for ensuring that student accommodations are correctly entered into TIDE.

DAMs responsibilities include the following:

- Complete District Accommodations Manager training
- Update the accessibility features in TIDE
- Enter any security issues, data reviews, unique accommodations, and/or exemption requests during the testing window in the KACE/DOE Help Desk

School Test Coordinator (STC)

STCs help to coordinate the administration of the Smarter Balanced assessments and ensure that testing operates smoothly and properly at the school level. STCs responsibilities include the following:

- Oversee all test administration related-activities in the school
- Complete the School Test Coordinator training
- Complete required security forms for reporting incidents
- Ensure that all TAs complete Smarter Balanced assessment training modules
- Ensure that the DeSSA secure browser has been installed for test administration
- Develop the test schedule
- Review student records in the Delaware Student Information System (DELSIS) and TIDE applications prior to testing
- Ensure that all TAs understand the protocols for student relocation
- Ensure that all students in Department of Services for Children, Youth and their Families (DSCYF), Delaware Adolescent Program, Inc. (DAPI), or the Consortium Discipline Alternative Program (CDAP) have a homeschool record
- Ensure that accommodations have been reviewed and updated in TIDE
- Enter any security issues, incidents, data reviews, unique accommodations, or exemptions in the KACE/DOE Help Desk

Test Administrator (TA)

TAs are qualified personel who administer the Smarter Balanced assessments. The pool of TAs may include the following authorized personnel:

- Delaware-certified educators (teachers, administrators, or guidance counselors)
- Paraprofessionals, if closely supervised by a Delaware-certified educator
- Translators (if they are not Delaware-certified educators, they must be closely supervised by a Delaware-certified educator)
- Substitute teachers (if they are not Delaware-certified educators, they must be closely supervised by one)

If there is a severe shortage of staff, a test may be administered by the following:

- Student-teachers acting as TAs, if closely supervised by a Delaware-certified educator
- Student-teachers and school support staff acting as proctors

TAs responsibilities include the following:

- Complete Smarter Balanced training
- Review necessary manuals and user guides

- Review student information before testing for accuracy to ensure that each student receives the right testing materials and/or is tested with the appropriate accommodations and supports
- Report any errors in student information to the KACE/DOE Help Desk for corrections
- Prepare for testing activities, such as environment, and equipments and materials (e.g., scratch paper, pencils, rulers, etc.)
- Administer the Smarter Balanced assessments
- Report all potential test security incidents and irregularities to the STC and/or DTC by following the security procedures
- Securely dispose of all testing materials including print-on-demand documents, scratch paper, and performance task (PT) materials

2.2.2 Online Test Administration

Within the state's testing window, each school needs to set testing schedules to allow students to complete a test in intervals (e.g., multiple sessions) rather than in one long period. With online testing, schools do not need to handle test booklets and address the storage and potential security problems inherent in large shipments of materials to different sites.

STCs oversee all aspects of testing at their schools and serve as the main contact person, while TAs administer the online assessments only. TAs are trained in the online testing requirements and the mechanics of starting, pausing, and ending a test session. Training materials for the test administration are provided online. All school personnel who serve as TAs must complete the required DeSSA training courses listed on the DeSSA portal, http://de.portal.airast.org. Prior to testing, DAMs are responsible for ensuring that student accommodations are correctly entered into TIDE.

To start a test session, the TA must first enter the TA Interface of the online test delivery system using his or her own computer. A test session ID is generated when the test session is created. Students who are taking the assessment need to enter their statewide student identifier (SSID), first name, and the test session ID into the Student Interface using computers provided by the school. The TA then verifies that the student is taking the appropriate assessment with the appropriate accessibility feature(s) (see Section 2.6 for a list of accommodations). Students can begin testing only when the TA confirms the settings. The TA reads the *Directions for Administration* in the *Online Smarter Balanced Test Administration Manual* aloud to the student(s) and guides them through the login process.

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

The summative assessment may be started in one test session and completed in a different session. The CAT must be completed within 45 calendar days of the start date or the assessment will expire. The PT must be completed within 20 calendar days of the start date.

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

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 and collect and shred any handouts or scratch papers that students used.

2.2.3 Paper-Pencil Test Administration

The paper-pencil versions of the Smarter Balanced ELA/Lit and mathematics assessments are provided as an accommodation for students who cannot access a computer or students with blindness or visual impairement. Although the online Braille was available, only the paper-pencil Braille test was used in Delaware in 2016–2017 administration.

The nonembedded support for the paper-pencil version must be set in TIDE prior to the deadline for the student to receive paper-pencil test booklets with the initial shipment. To request Braille paper-pencil materials, the nonembedded accommodation for Braille (paper-pencil version) must also be set. The list of students with this accommodation is extracted from TIDE and submitted to DDOE for approval. If the request is approved, the testing contractor ships the appropriate test booklets to the district. Additional orders may be entered into TIDE by the DTC after the initial order is received in the district. Additional orders for paper test materials must be approved by DDOE if the request exceeds 50 test booklets or if the request is for one or more Braille test booklets.

Separate test booklets are used for ELA/Lit and for mathematics. The items from the CAT and the PT components are combined into one test booklet, including two sessions for CAT and one session for PT in both content areas. Thus, the TA can break up the assessment into separate sessions.

After the student completes the assessment, the DTC returns the test booklets to the testing contractor. The testing contractor scans the answer document and scores the test, including the handscored items.

2.2.4 Braille Test Administration

In SY 2016–2017, the online Braille test was also available. The Braille interface is described below in several formats:

- The Braille interface includes a text-to-speech component for mathematics consistent with the read-aloud assessment accommodation. The Job Access with Speech (JAWS) screen reading software provided by Freedom Scientific is an essential component that students use with the Braille interface.
- Mathematics items are presented to students in Nemeth Braille through the adaptive online summative test or in the PT 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 DTCs, DAMs, STCs, 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, all individuals participating in or otherwise associated with any test administration must complete the training requirements in Table 3 and read the applicable manuals relevant to their roles. Table 3 presents the training requirements based on roles.

Table 3. Smarter Balanced Assessment Training Requirements

Role	Required Training	Course Number	Components of the Required Training	Estimated Time to Complete
All Roles	DeSSA Entry Training	24246	• Test Security • DeSSA Overview • TA Interface • Student Interface	• 30 min • 30 min • 15 min • 30 min
All Roles	Optional Training – Introduction to TIDE	25493	• TIDE Training	• 40 min
Smarter Balanced Summative Test Administrator	Smarter Balanced Summative TA Training	24619	• Smarter Balanced Summative TA Training	• 30 min
District Test Coordinator (DTC) and School Test Coordinators (STC)	DeSSA District and School Test Coordinator Training	24248	• TIDE Training • ORS Training • Smarter Balanced Interim TA Training • THSS Training	• 30 min • 35 min • 30 min • 30 min
Smarter Balanced Interim Test Administrator	Smarter Balanced Interim TA Training	24288	• Smarter Balanced Interim TA Training • THSS Training • AVA Training	• 30 min • 30 min • 5 min
Staff Performing Accommodations Data Entry	District and School Accommodations Manager Training	24250	• District and School Accommodations Manager Training	• 25 min
Special Education Staff/ Coordinator **English Language Learners Staff/ Coordinator** **General Education With Supports Staff/ Coordinator**	Accessibility Coordinator Training	24483*	• DeSSA Overview • Accessibility (TBD)	• 30 min • 50 min
Secretaries, Administrative Support	Security Training	24621	• Security module only	• 30 min
TAs who are giving paper-pencil assessment only* **(if TA is giving online and paper-pencil, take these and the online requirements)**	DeSSA Paper-Pencil TA Training for Smarter Balanced, DCAS, and EOC	24620	• Paper-Pencil TA Training • Security Training • DeSSA Overview	• 20 min • 30 min • 30 min
Students and Educators (optional)	Student Training	24472 24473 24484	• Let’s Talk Universal Tools • What is a CAT? • Student Interface	• 30 min • 20 min • 30 min

* TAs who administer the paper-pencil version must take the corresponding training (Summative 24619).

2.3.1 Practice and Training Test Site

In August 2016, separate training sites were opened for TAs and students. TAs can practice administering an assessment, such as starting and ending a test session on the TA Training Site, and students can take an online practice test on the Student Practice and Training Site. The Smarter Balanced assessment practice tests mirror the corresponding summative assessments. Each test provides students with a grade-specific testing experience, and students are able to practice with a variety of question types and difficulty levels (approximately 30 items each in mathematics and ELA/Lit), as well as the PT.

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 Smarter Balanced assessments for ELA/Lit and mathematics. Training tests are organized by grade bands (grades 3–5 and 6–8), with each test containing five to ten questions.

A student can log in directly to the practice and training test site as a guest without a TA-generated test session ID number, 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 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
Test Information Distribution Engine User Guide	The Test Information Distribution Engine (TIDE) is the system used to manage student information and user accounts for online testing. The TIDE User Guide provides a step-by-step approach to using the enhanced user management system.
Online Reporting System User Guide	The Online Reporting System (ORS) is the system used to view student performance and participation data. The ORS User Guide provides information on how to use the ORS to create reports.
Test Administrator User Guide	The Test Administrator (TA) 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, the TA Interface, and the Student Interface.
Accessibility Guidelines for Delaware System of Student Assessments (DeSSA)	This document provides information about identifying and documenting students who are eligible to receive designated supports and accommodations on Smarter Balanced and other DeSSA assessments. The document 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 ELA/Literacy and Mathematics Online Summative Test Administration Manual	This test administration manual (TAM) provides needed information regarding policies and procedures for the Smarter Balanced English Language Arts/Literacy and Mathematics Online Summative Assessments.
Smarter Balanced Summative ELA/Literacy Assessment Paper-Pencil Test Administration Manual	This TAM provides an overview of the Smarter Balanced Summative ELA/Literacy Assessment paper-pencil test administration and supplements the Online Summative TAM.
Smarter Balanced Summative Mathematics Assessment Paper-Pencil	This TAM provides an overview of the Smarter Balanced Summative Mathematics Assessment paper-pencil test administration and supplements the Online Summative TAM.

Resource	Description
Test Administration Manual	
SmarterELA/Literacy and Mathematics Interim Comprehensive Assessment and Interim Assessment Blocks Test Administration Manual	This TAM provides needed information regarding policies and procedures for the Smarter Balanced ELA/Literacy and Mathematics Interim Comprehensive Assessment and Interim Assessment Blocks.
Technology Specifications Manual for Online Testing	The 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 supporting the text-to-speech accommodation.
DeSSA Test Security Manual	The DeSSA Test Security Manual provides information regarding test security policies for all DeSSA tests. School personnel, including TAs, should review this document carefully.
Secure Browser Installation Manual	This 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.
Smarter Braille Requirements and Testing Manual	This Smarter Balanced Braille Requirements and Testing Manual provides information about supported hardware and software requirements and how to configure JAWS. Information about administering a test to a student requiring Braille and navigating a test with JAWS is also included.

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 as PowerPoint presentations; two modules included narration. Table 5 lists the training modules.

Table 5. Smarter Balanced-Developed Training Modules

Module Name	Primary Audience	Objective
Let’s Talk Universal Tools	• Students • TAs • Teachers	This presentation provides an overview of the Embedded Universal Tools available to students when using the Test Delivery System (TDS) for the online Smarter Balanced Assessment.
Student Interface for Online Testing	• Students • DTCs and STCs • TAs • Teachers	This presentation provides information on how students log in and navigate the test delivery system, including information on layout and functionality of the test tools.
What Is a CAT (Computer Adaptive Test)?	• DTCs and STCs • Teachers	This presentation, produced by Smarter Balanced, introduces TAs and students to the concept of a computer adaptive test, or CAT.

2.4 TEST SECURITY

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

2.4.1 DeSSA Test Security Manual

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

In preparation for the 2016–2017 school year, each district, school, 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 2016. All unethical or inappropriate practices and behaviors in the process of test preparation, test administration, and scoring must be reported in writing. In addition, all personnel associated with assessment administration must read and sign the Test Security and Non-Disclosure Agreement as documentation.

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

The Test Security Manual defines the test security incidents during testing in three levels: Impropriety, Irregularity, and Breach. Impropriety refers to an unusual circumstance that has a low impact on an individual or a group of students, with a low risk of potentially affecting student performance on the test; an impropriety can be corrected and contained at the local level. Irregularity refers to an unusual circumstance that may potentially affect student performance on the test; an irregularity 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); a breach has external implications and may result in a decision to remove certain test items from field 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 of prior to the administration of a PT. Unless needed as a print-on-demand or Braille accommodation, no copies may be made of any test items, stimuli, reading passages, PT materials, writing prompts, or any secured test materials. The electronic policy clearly prohibits usages of cell phones and other electronic devices in the testing area.

2.4.2 Student-Level Testing Confidentiality

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

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

1. *Test eligibility:* the assignment of a test for a particular student
2. *Test accommodation:* the assignment of a test setting to specific students based on needs
3. *Test session:* 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 a student log in and test under another student’s SSID number

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

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

Students log in to the online assessment using their legal first name, SSID number, and the test session ID. Only students can log in to an online test session. TAs, proctors, or other personnel are not permitted to log in to the system on behalf of students, although they are permitted to assist students who need help logging in. For the paper-pencil versions of the assessments, 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 online test delivery 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 log in 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 STCs 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

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

Room Preparation: The room should be prepared 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 PTs, different forms are spiraled within a classroom so that students receive different PTs.

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

For the paper-pencil versions, specific instructions 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. (Examples: Administrators modifying student answers, or students sharing test items through social media.)

District and school personnel must document all test security incidents. DTCs are responsible for reporting test security incidents to the state via the KACE/DOE Help Desk. 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.

2.4.6 Monitoring Test Administration

The observation of the 2016–2017 test administration of Smarter Balanced assessments was intended to improve test administration and monitoring for the 2017–2018 test administration. The Office of Assessment at the Department of Education scheduled on-site visits (upon agreement with schools) during the testing window, and all observers followed the procedure for the on-site visits without interfering with test activities.

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: (1) Computer sign-on and start-up process; (2) Security; (3) Test environment and administration procedures; (4) Test atmosphere; (5) Calculator use in mathematics; (6) Accommodations; and (7) Classroom activity for PTs.

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

2.5 STUDENT PARTICIPATION

All students (including retained students) currently enrolled in grades 3–8 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 Student Exemptions

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

- Students with significant cognitive disabilities who meet the criteria for the ELA/Lit alternate assessment based on alternate achievement standards (approximately one percent or less of the student population)
- Students with significant cognitive disabilities who meet the criteria for the mathematics alternate assessment based on alternate achievement standards (approximately one percent or less of the student population)
- English language learners (ELLs) who enrolled in a U.S. school within the last 12 months before the beginning of the testing time 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 ACCOMMODATIONS

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

The Smarter Balanced *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 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) and ELLs, 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 online test environment (e.g., test-to-speech, Spanish-English stacked) and non-embedded features that must be provided by the school (e.g., translator, enhanced lighting).

The administration of Smarter Balanced assessments is 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 universal tools, designated supports, and accommodations in both embedded and non-embedded versions. Embedded resources are part of the online test delivery system whereas non-embedded resources are provided outside of that system.

State-level users, DAs and DAMS 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 test delivery 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* at https://portal.smarterbalanced.org/library/en/usability-accessibility-and-accommodations-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 delivery system. Universal tools are available to all students based on their preference and selection and have been preset in TIDE. In 2016–2017 test administration, the following features (universal tools) were available for all students to access. For specific information on how to access and use these features, refer to the *Test Administrator User Guide* at the DeSSA portal, http://de.portal.airast.org.

Embedded Universal Tools

Breaks: The number of items per session can be flexibly defined based on the student's need. Breaks of more than 20 minutes will prevent the student from returning to items already attempted by the student (exception is the PT). There is no limit on the number of breaks that a student might be given. The use of this universal tool may result in the student needing additional overall time to complete the assessment.

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 may be available for the full-write portion of an ELA/Lit PT. The use of this universal tool may result in the student needing additional overall time to complete the assessment.

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. The use of this accommodation may result in the student needing additional overall time to complete the assessment.

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 PTs in which students complete the full write portion of an ELA/Lit PT. The student clicks the notepad icon for the notepad to appear. During the ELA/Lit PTs, the notes are retained from segment to segment so that the student may go back to the notes even though he or she cannot go back to specific items in the previous segment.

Highlighter: A digital tool for marking desired text, item questions, item answers, or parts of these with a color. Highlighted text remains available throughout each test segment.

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

Mark for review: This tool allows students to flag items for future review during the assessment. Markings are not saved when the student moves on to the next segment or after a break of more than 20 minutes.

Pause: The student can pause the assessment and and then return to their last test question. However, if an assessment is paused for more than 20 minutes, students will not be allowed to return to previous test questions.

Strikethrough: This function allows the student to cross out response options.

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

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

Non-Embedded Universal Tools

Assistive listening device: FM system or other system that mitigates student hearing impairment.

Breaks: All students may take breaks as needed. The term “frequent breaks” refers to multiple, planned, short breaks during testing based on a specific student’s needs (for example, the student fatigues easily).

During each break, the testing clock is stopped. 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 may be available for the full-write portion of an ELA/Lit PT. The use of this universal tool may result in the student needing additional overall time to complete the assessment.

Familiar TA: The student knows the TA and/or interpreter.

Refocus: The student's attention can be refocused on the test with use of intermittent verbal, picture symbol, signed, cued speech, or physical prompts. Refocus should not in any way cue a student to return to a previous item or indicate that the student may have made an error. This would be considered a test security violation. The use of this universal tool may result in the student needing additional overall time to complete the assessment.

Scratch/blank/grid 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 mathematics assessments. A student can use an assistive technology device for scratch paper as long as the device is consistent with the child's IEP or Section 504 Plan and is acceptable to the state.

Small group: A small group is a subset of a larger testing group assessed in a separate location. There is no specific number defined for a small group, but a group of two to eight students is typical. A group of one also is permissible. Small groups may be appropriate for human read-aloud, translated test administration, WhisperPhone®, or to reduce distractors for some students. If selecting a small group, it is not necessary to also select a separate setting.

Thesaurus: A thesaurus provides synonyms of terms while a student interacts with text included in the assessment and may be available for the full write portion of an ELA/Lit PT. The use of this universal tool may result in the student needing additional overall time to complete the assessment.

Time of Day: Student should be tested during the time of day that is best for the student (e.g., only morning).

Additional Non-Embedded Universal Tool options include *modified lighting, specialized equipment or furniture, and specified area or seating.*

2.6.2 Designated Supports and Accommodations

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

Accommodations are modifications in testing conditions and/or presentations of the test to facilitate the access for students with special needs in order to demonstrate what they know and can do. Consortium-approved accommodations do not compromise the learning expectations, construct, grade-level standard, or intended outcome of the assessments.

Below are brief descriptions of embedded and non-embedded supports and accommodations.

Embedded Designated Supports

Color contrast (Computer): 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 mathematics stimuli items, ELA/Lit 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 mathematics): Translation of test directions is a language support available before beginning the actual test items. Students can see test directions in another language. As an embedded designated support, translated test directions are automatically a part of the stacked translation designated support.

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

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

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

Non-Embedded Designated Supports

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

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

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

Disable universal tools: Disabling any universal accessibility tools that might be distracting or which students do not need to use, or are unable to use.

Interpreter – native language: Provide a native language translator to translate test questions (including multiple-choice options) into native language.

Human read aloud passages (For ELA/PT 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*. All or portions of the content may be read aloud.

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

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

Read aloud items (for mathematics items and ELA/Lit items; but not for reading passages): Text is read aloud to the student by a trained and qualified human reader who follows the administration guidelines provided in the *Smarter Balanced Online Summative Test Administration Manual*. 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. 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 in school: The test location is altered so that the student is tested in an in-school setting different from that made available for most students.

Separate setting not in school/homebound: The test location is altered so that the student is tested in a non-school setting different from that made available for most students.

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

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

Unique accommodation (DOE approved): Support or accommodation not listed in these guidelines by Smarter Balanced. By application only.

Embedded Accommodations

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

Braille: This is a raised-dot code that individuals read with 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 mathematics.

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

Print on request: Paper copies of either passages/stimuli and/or items are printed for students. Student may request one or more test questions to be printed electronically from the online system for student to review on paper. All printed test material must be shredded at end of test session. (TA must approve each print request.)

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

Text-to-Speech (for 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. May only be activated by DOE.

Non-Embedded Accommodations

100s Number Table (grade 4 and above mathematics tests): A paper-based table listing of numbers from 1–100 will be available from Smarter Balanced for reference.

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

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

Calculator (for grades 6–8 mathematics 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.

Human 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.

Interpreter – Visual Communication: An adult with the necessary qualifications provides translation/ interpretation of the mathematics test using cued speech or signed English to a student with disabilities. Reading passages may not be translated through visual communication. This support must be approved by DOE.

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

Physical Assistance from a TA: Using physical assistance from a test administrator, such as direct assistance with turning pages, recording answers for the paper-pencil test (scribing) or navigating in electronic format.

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., open application programs, pull down menus, save 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 2016–2017 administration. Tables 7 through 12 provide the number of students who were offered the accommodations and/or designated supports.

Table 6. Universal Tools, Designated Supports, and Accommodations in 2016–2017

	Universal Tools	**Designated Supports**	**Accommodations**
Embedded	Breaks Calculator[1] Digital Notepad English Dictionary[2] English Glossary Expandable Passages Global Notes Highlighter Keyboard Navigation Mark for Review Mathematics Tools[3] Spell Check Strikethrough Writing Tools[4] Zoom	Color Contrast (Computer) Masking Text-to-Speech[5] Translated Test Directions[6] Translations (Glossary)[7] Translations (Stacked) [8] Turn Off Any Universal Tools	American Sign Language[9] Braille Closed Captioning[10] Print-on-Request Streamlined Mode Text-to-Speech[11]
Non-embedded	Assistive Listening Device Breaks English Dictionary[12] Familiar TA Modified Lighting Refocus Scratch/blank/grid Paper Small Group Specialized Equipment or Furniture Specified area or Seating Thesaurus[13] Time of Day	Bilingual Dictionary[14] Color Contrast (Printed) Color Overlay Disable universal tools Interpreter – Native Language[22] Human Read Aloud Passages for PT[23] Magnification Noise Buffers Read Aloud Items[15] Scribe[16] Separate Setting in school Separate Setting not in school/homebound Translated Test Directions Translations (Glossary)[17] Unique Accommodation[22] WhisperPhone®	100s Number Table[20] Abacus Alternate Response Options[18] Calculator[19] Human Read Aloud Passages[21] Interpreter – Visual Communication[22] Multiplication Table[20] Physical Assistance from a TA Scribe Speech-to-Text

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

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

[2] For ELA/Lit performance task full-writes

[3] Includes embedded ruler, embedded protractor

[4] Includes bold, italic, underline, indent, cut, paste, spell check, bullets, undo/redo

[5] For ELA/Lit PT stimuli, ELA/Lit PT and CAT items (not ELA/Lit CAT reading passages), and mathematics stimuli and items: Must be set in TIDE before test begins.

[6] For mathematics items
[7] For mathematics items
[8] For mathematics test
[9] For ELA/Lit listening items and mathematics items
[10] For ELA/Lit listening items
[11] For ELA/Lit CAT reading passages. Must be set in TIDE by state-level user.
[12] For ELA/Lit performance task full writes
[13] For ELA/Lit performance task full writes
[14] For ELA/Lit performance task full writes
[15] For ELA/Lit items (not ELA/Lit reading passages) and mathematics items
[16] For ELA/Lit non-writing items and mathematics items
[17] For mathematics items on the paper-pencil test
[18] Includes adapted keyboards, large keyboard, StickyKeys, MouseKeys, FilterKeys, adapted mouse, touch screen, head wand, and switches
[19] For calculator-allowed items only in grades 6–8 and 11
[20] For mathematics items beginning in grade 4
[21] For ELA/Lit CAT reading passages, all grades – must be approved by DOE
[22] Must be approved by DOE
[23] For ELA/Lit performance task passages

Table 7. Students with Embedded and Non-Embedded Accommodations in ELA/Lit

Accommodations	**Grade**					
	3	**4**	**5**	**6**	**7**	**8**
Embedded Accommodations						
American Sign Language	10	20	10	6	6	10
Braille			1	1		1
Closed Captioning	10	18	9	40	39	20
Print-on-Request: Items	1	3		3	1	1
Print-on-Request: Passages	18	25	10	4	13	6
Print-on-Request: Passages & Items	310	412	401	392	336	289
Print-on-Request: Stimuli	2	1	3			
Print-on-Request: Stimuli & Items	66	70	63	7	11	7
Streamlined Mode	18	27	16	77	65	64
Text-to-Speech: Passages	7	1				
Text-to-Speech: Passages & Items	7	14	23	21	19	13
Non-Embedded Accommodations						
Alternate Response Options		1	1			1
Human Read Aloud Passage	28	14	16	6	7	
Interpreter – Visual Communication				1		1
Physical Assistance from a TA	26	20	19		4	4
Scribe Items (Writing)	103	91	75	18	12	9
Speech-to-Text	13	23	33	15	22	19

Table 8. Students with Embedded Designated Supports in ELA/Lit

Designated Supports	**Subgroup**	**Grade**					
		3	**4**	**5**	**6**	**7**	**8**
Color Contrast	Overall	14	29	37	37	26	28
	ELL	6	5	2	4	1	1
	Special Ed	12	17	30	35	22	27
Masking	Overall	144	297	235	284	237	229
	ELL	31	121	26	61	43	41
	Special Ed	65	150	142	206	199	167
Text-to-Speech: Items	Overall	2,293	2,432	2,110	1,747	1,248	1,187
	ELL	993	654	353	289	243	198
	Special Ed	976	1,091	1,105	1,156	971	934
Text-to-Speech: Stimuli & Items	Overall	2,285	2,372	2,090	1,724	1,288	1,171
	ELL	990	659	353	282	233	196
	Special Ed	974	1,071	1,132	1,158	1,024	931
Turn Off Any Universal Tools	Overall	8,712	8,653	8,837	8,906	8,789	8,851
	ELL	1,093	554	261	261	226	226
	Special Ed	348	420	427	458	391	482

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

Designated Supports	**Subgroup**	**Grade**					
		3	**4**	**5**	**6**	**7**	**8**
Bilingual Dictionary	Overall	67	43	20	59	84	72
	ELL	67	41	18	59	82	72
	Special Ed	5	2	4	8	14	13
Color Contrast	Overall				6	3	5
	ELL				2	1	1
	Special Ed				5	3	5
Color Overlay	Overall	9	9	5	3	2	5
	ELL						1
	Special Ed	9	5	3	3	2	5
Disable Universal Tools	Overall	5	11		2		
	ELL	1	3		1		
	Special Ed	5	11		2		
Magnification	Overall	10	9	5	10	10	11
	ELL	1					2
	Special Ed	5	4	4	7	8	9
Noise Buffers	Overall	85	138	142	64	29	29
	ELL	13	35	20	5		1
	Special Ed	58	96	96	54	17	19
Read Aloud Items	Overall	506	488	402	273	226	229
	ELL	180	125	57	22	14	25
	Special Ed	279	287	310	231	217	211
Read Aloud Passages	Overall	480	383	329	181	98	89
	ELL	180	111	52	12	5	10
	Special Ed	267	234	259	163	94	85
Scribe Items (Non-Writing)	Overall	44	31	26	7	4	1
	ELL	16	9	9			
	Special Ed	19	19	16	5	2	1
Separate Setting in School	Overall	597	464	441	375	319	334
	ELL	107	62	46	36	31	34
	Special Ed	504	400	388	333	286	294
Separate Setting Not in School/Homebound	Overall	2	1	2	1	1	3
	ELL	2					
	Special Ed	1	1	1	1	1	3
Simplified Test Directions	Overall	371	233	149	116	140	151
	ELL	321	198	86	46	64	61
	Special Ed	97	77	77	81	81	96
Translated Test Directions	Overall	4	6		4	6	6
	ELL	3	6		4	6	6
	Special Ed	1					
Unique Accommodation	Overall	4	3	2			
	ELL	1					
	Special Ed	4	1	2			
WhisperPhone	Overall	213	146	81	22	14	9
	ELL	110	43	21	13	10	8
	Special Ed	123	114	66	6	4	1

Table 10. Students with Embedded and Non-Embedded Accommodations in Mathematics

Accommodations	**Grade**					
	3	**4**	**5**	**6**	**7**	**8**
Embedded Accommodations						
American Sign Language	10	20	10	6	6	10
Print-on-Request: Stimuli & Items	331	406	408	380	339	288
Streamlined Mode	27	32	33	108	100	91
Non-Embedded Accommodations						
100s Number Table	199	365	215	57	47	51
Abacus				2	5	
Alternate Response Options		1		1		3
Calculator	89	134	121	302	365	337
Interpreter – Visual Communication				1		1
Multiplication Table	85	958	1,064	941	848	732
Physical Assistance from a TA	34	20	26	4	3	4
Scribe Items (Writing)	100	81	65	15	8	9
Speech-to-Text	15	18	31	12	18	18

Table 11. Students with Embedded Designated Supports in Mathematics

Designated Supports	**Subgroup**	**Grade**					
		3	**4**	**5**	**6**	**7**	**8**
Color Contrast	Overall	13	19	32	37	22	24
	ELL	4	3	2	4	1	4
	Special Ed	10	9	28	35	19	19
Masking	Overall	138	280	240	288	230	220
	ELL	33	125	34	67	45	50
	Special Ed	62	138	136	203	188	167
Translation (Glossary): Spanish	Overall	132	45	43	35	30	37
	ELL	130	41	39	32	29	37
	Special Ed	15	8	5	7	5	3
Translation (Glossary): Other Languages	Overall	10	14	11	14	17	17
	ELL	10	12	10	13	17	17
	Special Ed	1					1
Text-to-Speech: Stimuli & Items	Overall	2,333	2,429	2,126	1,712	1,291	1,182
	ELL	991	657	373	251	192	172
	Special Ed	994	1,095	1,135	1,159	1,028	926
Turn Off Any Universal Tools	Overall	8,721	8,630	8,805	8,844	8,745	8,762
	ELL	1,144	592	302	286	260	268
	Special Ed	319	390	385	407	356	424

Table 12. Students with Non-Embedded Designated Supports in Mathematics

Designated Supports	**Subgroup**	**Grade**					
		3	**4**	**5**	**6**	**7**	**8**
Color Contrast	Overall				6	2	3
	ELL				2	1	1
	Special Ed				5	2	3
Color Overlay	Overall	9	7	7	2	2	6
	ELL			1			1
	Special Ed	9	4	5	2	2	5
Disable Universal Tools	Overall	3	11				
	ELL	1	3				
	Special Ed	3	11				
Human Read Aloud Stimuli & Items	Overall	516	401	345	167	90	65
	ELL	211	132	70	16	11	9
	Special Ed	280	233	263	146	79	57
Human Read Aloud in Spanish	Overall	21	15	12	4	10	7
	ELL	21	15	12	4	8	6
	Special Ed				1	1	1
Interpreter – Native Language	Overall	11	22	13	16	11	21
	ELL	11	22	13	15	9	20
	Special Ed				1	1	
Magnification	Overall	11	7	5	7	8	10
	ELL						1
	Special Ed	5	4	4	6	5	8
Noise Buffers	Overall	85	138	142	64	31	27
	ELL	13	36	22	4		1
	Special Ed	58	97	93	54	19	17
Scribe Items	Overall	40	36	27	6	1	
	ELL	20	16	13			
	Special Ed	17	16	13	4	1	
Separate Setting in School	Overall	601	460	425	362	318	344
	ELL	118	71	51	41	38	47
	Special Ed	494	392	370	314	279	291
Separate Setting Not in School/Homebound	Overall	7	7	8	5	9	1
	ELL	1					
	Special Ed	6	7	5	4	9	
Simplify Directions in English	Overall	413	262	171	142	150	182
	ELL	360	224	110	69	74	91
	Special Ed	100	82	78	87	79	99
Translated Test Directions	Overall	13	15	6	7	8	8
	ELL	13	15	6	7	8	8
	Special Ed	1	1				
Translations (Glossaries) – Paper-Pencil	Overall		1	2	4	19	15
	ELL		1	1	4	19	15
	Special Ed			1		1	
Unique Accommodation	Overall	3	1				
	ELL						
	Special Ed	3					
WhisperPhone	Overall	179	125	73	24	13	10
	ELL	78	29	17	13	10	9
	Special Ed	115	110	58	8	3	2

2.7 DATA FORENSICS PROGRAM

2.7.1 Data Forensics Report

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

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

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

2.7.2 Changes in Student Performance

Score changes between years are examined using a regression model. For between-year comparisons, the scores between 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 between the current year (e.g., 2017) and the year before the current year (e.g., 2016).

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

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

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

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

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

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

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

2.7.3 Item Response Time

The online environment also allows item response time 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.

2.7.4 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 time index might flag such a student.

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

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

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

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

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

2.8 PREVENTION AND RECOVERY OF DISRUPTIONS IN TEST DELIVERY SYSTEM

AIR is continuously improving our ability to protect our systems from interruptions. AIR's test delivery system is designed to ensure that student responses are captured accurately and stored on more than one server in case of a failure. Our architecture, described below, is designed to recover from failure of any component with little interruption. Each system is redundant, and critical student response data is transferred to a different data center each night.

AIR has developed a unique monitoring system that is very sensitive to changes in server performance. Most monitoring systems provide warnings when something is going wrong. Ours does, too, but it also provides warnings when any given server is performing differently from its performance over the prior few hours, or differently than the other servers preforming the same jobs. Subtle changes in performance often precede actual failure by hours or days, allowing us to detect potential problems, investigate them, and mitigate them *before* a failure. On multiple occasions, this has enabled us to make adjustments and replace equipment before any problems occurred.

AIR has also implemented an escalation procedure that enables us to alert clients within minutes of any disruption. Our emergency alert system notifies our executive and technical staff by text message, who immediately join a call to understand the problem.

The section below describes AIR system architecture and how it recovers from device failures, internet interruptions, and other problems.

2.8.1 High-Level System Architecture

Our architecture provides redundancy, robustness, and reliability required by a large-scale, high stakes testing program. Our general approach, which has been adopted by Smarter Balanced as standard policy, is pragmatic and well supported by our architecture.

Any system built around an expectation of flawless performance of computers or networks within schools and districts is bound to fail. Our system is designed to ensure that the testing results and experience are able to respond robustly to such inevitable failures. Thus, AIR’s test delivery system (TDS) is designed to protect data integrity and prevent student data loss at every point in the process.

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

Student Machine

Student responses are conveyed to our servers in real time as students respond. Long responses, such as essays, are saved automatically at configurable intervals (usually set to one minute), so that student work is not at risk during testing.

Responses are saved asynchronously, with a background process on the student machine waiting for confirmation of successfully stored data on the server. If confirmation is not received within the designated time (usually set to 30–90 seconds), the system will prevent the student from doing any more work until connectivity is restored. The student is offered the choice of asking the system to try again or pausing the test and returning at a later time. For example:

- If connectivity is lost and restored within the designated time period, the student may be unaware of the momentary interruption.
- If connectivity cannot be silently restored, the student is prevented from testing and given the option of logging out or retrying the save.
- If the system fails completely, upon logging back in the system the student returns to the item at which the failure occurred.

In short, data integrity is preserved by confirmed saves to our servers and prevention of further testing if confirmation is not received.

Test Delivery Satellites

The test delivery satellites communicate with the student machines to deliver items and receive responses. Each satellite is a collection of web and database servers. Each satellite is equipped with redundant array of independent disks (RAID) systems to mitigate the risk of disk failure. Each response is stored on multiple independent disks.

One server for every four satellites serves as a backup hub. This server continually monitors and stores all changed student response data from the satellites, creating an additional copy of the real-time data. In the unlikely event of failure, data are completely protected. Satellites are automatically monitored, and upon failure, they are removed from service. Real-time student data are immediately recoverable from the satellite, backup hub, or hub (described below), with backup copies remaining on the drive arrays of the disabled satellite.

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

Hub

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

Demographic and History Servers

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

Quality Assurance System

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

Database of Record

The Database of Record (DoR) is the final storage location for the student data. These clustered database servers with RAID systems hold the completed student data.

2.8.2 Automated Backup and Recovery

Every system is backed up nightly. Industry-standard backup and recovery procedures are in place to ensure safety, security, and integrity of all data. This set of systems and processes is designed to provide complete data integrity and prevent loss of student data. Redundant systems at every point, real-time data integrity protection and checks, and well-considered real-time backup processes prevent loss of student data, even in the unlikely event of system failure.

2.8.3 Other Disruption Prevention and Recovery

We have designed our system to be extremely fault-tolerant. The system can withstand failure of any component with little or no interruption of service. One way that we achieve this robustness is through redundancy. Key redundant systems are as follows:

- Our hosting provider has redundant power generators that can continue to operate for up to 60 hours without refueling. With the multiple refueling contracts that are in place, these generators can operate indefinitely.
- Our hosting provider has multiple redundancies in the flow of information to and from our data enters by partnering with nine different network providers. Each fiber carrier must enter the data center at separate physical points, protecting the data center from a complete service failure

caused by an unlikely network cable cut.

- On the network level, we have redundant firewalls and load balancers throughout the environment.
- We use redundant power and switching within all of our server cabinets.
- Data are protected by nightly backups. We complete a full weekly backup and incremental backups nightly. Should a catastrophic event occur, AIR is able to reconstruct real time data using the data retained on the TDS satellites and hubs.
- The server backup agents send alerts to notify system administration staff in the event of a backup error, at which time they will inspect the error to determine whether the backup was successful or if they will need to rerun the backup.

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

3. SUMMARY OF 2016–2017 OPERATIONAL TEST ADMINISTRATION

3.1 STUDENT POPULATION

All students enrolled in grades 3–8 in all public elementary and secondary schools are required to participate in the Smarter Balanced ELA/Lit and mathematics assessments. Tables 13 and 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 Summative ELA/Lit Assessment

Group	G3	G4	G5	G6	G7	G8
All Students	10,600	10,386	10,461	10,189	10,070	10,069
Female	5,171	5,150	5,230	5,055	4,936	4,942
Male	5,429	5,236	5,231	5,134	5,134	5,127
African American	3,206	3,143	3,077	3,133	3,201	3,096
Asian	371	383	367	381	358	348
Hispanic/Latino	1,997	1,838	1,824	1,776	1,604	1,646
American Indian/Alaska Native	36	41	31	43	45	45
White	4,513	4,518	4,708	4,458	4,570	4,678
English Language Learner	1,635	886	440	392	339	322
Special Education	1,438	1,474	1,526	1,483	1,431	1,432
CD 504	331	411	462	456	488	492
Title I	1,035	1,046	1,247	1,336	1,567	1,714

Table 14. Number of Students in Summative Mathematics Assessment

Group	G3	G4	G5	G6	G7	G8
All Students	10,669	10,442	10,519	10,211	10,087	10,058
Female	5,203	5,183	5,255	5,072	4,943	4,944
Male	5,466	5,259	5,264	5,139	5,144	5,114
African American	3,216	3,155	3,089	3,138	3,199	3,092
Asian	394	398	378	389	362	356
Hispanic/Latino	2,031	1,871	1,861	1,794	1,636	1,669
American Indian/Alaska Native	36	41	31	43	45	45
White	4,514	4,514	4,706	4,447	4,552	4,641
English Language Learner	1,707	954	507	435	385	379
Special Education	1,441	1,479	1,543	1,478	1,420	1,415
CD 504	336	416	468	455	488	489
Title I	1,045	1,052	1,254	1,339	1,568	1,714

3.2 SUMMARY OF OVERALL STUDENT PERFORMANCE

Tables 15–18 present a summary of the 2016–2017 summative test results for all students and by subgroups, including the average and the standard deviation of scale scores, the percentage of students in each achievement level, and the percentage of proficient students. Figures 1–2 present the percentage of proficient students in 2014–2015, 2015–2016, and 2016–2017 for all students (cohort comparisons). The percentages of proficient students by subgroups across three years are provided in Appendix B.

Table 15. ELA/Lit Percentage of Students in Achievement Levels for Overall and by Subgroups (Grades 3–5)

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
Grade 3								
All Students	10,600	2433.34	87.23	23	25	25	27	52
Female	5,171	2442.07	85.68	20	25	25	30	55
Male	5,429	2425.03	87.89	27	25	24	24	48
AmeriIndian/AlaskaNat	36	2430.11	84.27	25	22	25	28	53
Asian	371	2494.18	80.20	7	15	25	54	78
African American	3,206	2401.07	81.12	34	30	21	14	36
Hispanic	1,997	2407.33	80.11	33	29	23	16	39
White	4,513	2461.57	82.79	13	21	27	39	66
ELL	1,635	2397.09	77.50	37	31	20	12	32
Special Education	1,438	2354.54	72.72	59	25	11	4	15
CD 504	331	2426.70	75.61	22	31	25	22	47
Title I	1,035	2455.52	77.98	14	23	29	34	63
Grade 4								
All Students	10,386	2477.16	92.05	26	20	26	28	54
Female	5,150	2486.93	89.62	22	20	27	31	58
Male	5,236	2467.55	93.40	30	20	25	25	50
AmeriIndian/AlaskaNat	41	2478.63	81.25	17	32	29	22	51
Asian	383	2542.79	82.16	7	10	26	57	83
African American	3,143	2442.78	88.37	39	22	24	15	39
Hispanic	1,838	2451.99	84.30	34	24	25	17	42
White	4,518	2505.05	86.58	15	18	27	39	67
ELL	886	2412.51	74.63	53	26	15	6	21
Special Education	1,474	2380.80	78.40	69	18	10	3	12
CD 504	411	2467.46	86.07	27	26	24	22	47
Title I	1,046	2483.96	82.10	21	21	31	27	58
Grade 5								
All Students	10,461	2519.67	93.28	21	19	34	26	60
Female	5,230	2532.70	92.07	17	18	34	31	65
Male	5,231	2506.65	92.67	24	21	34	21	55
AmeriIndian/AlaskaNat	31	2519.09	95.88	23	16	29	32	61
Asian	367	2591.94	86.65	4	9	29	58	87
African American	3,077	2484.14	89.06	31	24	32	13	45
Hispanic	1,824	2494.54	83.85	27	26	32	15	47
White	4,708	2546.86	88.37	12	15	36	36	72
ELL	440	2413.63	74.34	67	20	12	1	13
Special Education	1,526	2417.70	80.31	63	21	14	2	16
CD 504	462	2510.73	80.49	22	22	38	18	56
Title I	1,247	2526.28	83.35	15	21	39	25	64

Table 16. ELA/Lit Percentage of Students in Achievement Levels for Overall and by Subgroups (Grades 6–8)

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
Grade 6								
All Students	10,189	2529.67	93.36	22	26	34	18	52
Female	5,055	2542.42	91.00	17	25	36	21	57
Male	5,134	2517.11	93.96	26	27	32	15	47
AmeriIndian/AlaskaNat	43	2545.39	78.41	7	40	37	16	53
Asian	381	2602.23	87.76	6	12	36	46	82
African American	3,133	2493.67	87.49	33	32	27	8	35
Hispanic	1,776	2502.33	86.42	30	31	31	9	39
White	4,458	2558.35	87.40	13	22	39	26	65
ELL	392	2412.50	69.68	74	22	4	1	4
Special Education	1,483	2428.53	74.32	66	25	9	1	10
CD 504	456	2523.53	82.67	20	32	36	12	48
Title I	1,336	2525.97	87.97	21	29	35	14	49
Grade 7								
All Students	10,070	2553.72	97.76	22	25	37	17	54
Female	4,936	2567.96	94.16	17	24	39	20	59
Male	5,134	2540.04	99.20	26	26	35	13	48
AmeriIndian/AlaskaNat	45	2558.70	87.89	20	27	38	16	53
Asian	358	2633.95	92.81	6	10	35	48	83
African American	3,201	2514.87	94.85	34	30	29	8	36
Hispanic	1,604	2527.44	90.31	27	31	34	8	42
White	4,570	2583.77	88.59	12	20	45	23	68
ELL	339	2435.64	76.85	71	23	6	1	7
Special Education	1,431	2450.33	80.50	64	26	10	1	11
CD 504	488	2549.74	86.82	19	32	38	12	50
Title I	1,567	2550.83	92.35	21	27	38	14	53
Grade 8								
All Students	10,069	2565.99	99.66	22	26	37	15	52
Female	4,942	2585.21	95.55	16	25	40	20	60
Male	5,127	2547.46	100.04	28	28	34	11	45
AmeriIndian/AlaskaNat	45	2585.27	88.57	16	18	51	16	67
Asian	348	2646.32	98.18	7	13	37	43	80
African American	3,096	2528.00	94.46	34	30	29	7	36
Hispanic	1,646	2543.16	95.44	28	30	33	9	42
White	4,678	2592.31	93.09	13	23	42	21	64
ELL	322	2457.50	78.75	66	26	7	1	8
Special Education	1,432	2463.79	81.11	63	27	9	1	10
CD 504	492	2554.60	91.73	22	29	39	9	48
Title I	1,714	2565.50	93.43	20	28	40	13	52

Table 17. Mathematics Percentage of Students in Achievement Levels for Overall and by Subgroups (Grades 3–5)

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
Grade 3								
All Students	10,669	2440.96	79.38	22	25	31	23	53
Female	5,203	2441.13	76.43	21	26	32	21	53
Male	5,466	2440.79	82.09	22	24	30	24	54
AmeriIndian/AlaskaNat	36	2432.93	95.13	28	28	19	25	44
Asian	394	2503.09	77.28	6	12	30	53	82
African American	3,216	2409.25	74.09	34	30	26	10	36
Hispanic	2,031	2420.07	72.83	28	30	29	13	42
White	4,514	2467.00	73.75	12	20	35	33	68
ELL	1,707	2416.14	73.57	30	31	28	12	40
Special Education	1,441	2367.60	76.44	57	26	13	4	18
CD 504	336	2435.30	68.50	22	28	33	17	50
Title I	1,045	2462.15	71.86	13	22	37	28	65
Grade 4								
All Students	10,442	2483.34	82.58	19	31	28	22	50
Female	5,183	2481.87	79.24	18	33	29	20	49
Male	5,259	2484.79	85.72	19	30	28	24	52
AmeriIndian/AlaskaNat	41	2486.36	74.67	12	46	22	20	41
Asian	398	2557.89	79.37	3	14	26	57	83
African American	3,155	2448.62	76.67	31	37	23	10	32
Hispanic	1,871	2459.40	72.32	24	40	25	11	37
White	4,514	2510.45	77.16	9	25	34	32	65
ELL	954	2432.89	70.68	38	40	15	6	22
Special Education	1,479	2399.97	75.61	58	29	11	2	13
CD 504	416	2480.85	74.87	16	35	30	20	49
Title I	1,052	2498.48	73.06	11	30	35	24	58
Grade 5								
All Students	10,519	2511.47	89.71	27	29	20	24	44
Female	5,255	2512.50	87.22	26	30	20	24	44
Male	5,264	2510.44	92.12	27	28	20	25	44
AmeriIndian/AlaskaNat	31	2503.33	81.96	19	45	16	19	35
Asian	378	2589.12	87.92	7	17	16	60	76
African American	3,089	2472.86	81.64	41	33	17	9	26
Hispanic	1,861	2486.88	81.01	34	35	17	13	31
White	4,706	2540.34	84.49	16	26	23	35	59
ELL	507	2426.06	71.18	68	25	5	2	7
Special Education	1,543	2420.55	74.44	69	23	6	2	8
CD 504	468	2509.06	80.70	26	37	17	20	37
Title I	1,254	2521.94	82.55	20	32	23	25	48

Table 18. Mathematics Percentage of Students in Achievement Levels for Overall and by Subgroups (Grades 6–8)

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
Grade 6								
All Students	10,211	2523.83	103.51	28	30	22	19	41
Female	5,072	2527.35	98.64	26	31	23	19	42
Male	5,139	2520.35	108.01	30	30	22	19	40
AmeriIndian/AlaskaNat	43	2554.03	85.63	16	33	28	23	51
Asian	389	2610.81	106.64	9	14	26	50	76
African American	3,138	2479.59	96.23	43	34	15	7	22
Hispanic	1,794	2496.29	94.07	37	34	20	10	29
White	4,447	2557.18	95.78	16	28	28	28	56
ELL	435	2412.80	84.55	76	18	4	1	5
Special Education	1,478	2410.37	92.15	75	20	4	1	6
CD 504	455	2521.41	92.68	26	36	22	15	38
Title I	1,339	2525.23	89.73	24	35	24	16	40
Grade 7								
All Students	10,087	2538.73	109.13	30	29	22	19	41
Female	4,943	2540.92	105.67	29	30	23	18	41
Male	5,144	2536.62	112.33	31	27	22	19	41
AmeriIndian/AlaskaNat	45	2532.10	82.57	31	33	22	13	36
Asian	362	2640.43	116.98	8	15	24	53	77
African American	3,199	2492.96	101.05	45	31	16	7	23
Hispanic	1,636	2507.07	102.26	41	30	19	10	30
White	4,552	2573.75	98.88	18	27	28	27	55
ELL	385	2422.66	92.22	77	17	4	2	6
Special Education	1,420	2424.41	89.72	77	16	5	2	7
CD 504	488	2539.97	91.81	27	35	24	14	38
Title I	1,568	2540.27	103.83	28	30	25	18	42
Grade 8								
All Students	10,058	2550.50	119.68	36	26	18	20	38
Female	4,944	2560.01	114.35	32	27	19	22	41
Male	5,114	2541.30	123.94	39	26	17	18	35
AmeriIndian/AlaskaNat	45	2580.21	117.73	24	20	29	27	56
Asian	356	2668.26	135.38	11	17	15	58	72
African American	3,092	2498.60	107.54	53	26	13	8	21
Hispanic	1,669	2525.95	109.63	43	28	17	12	29
White	4,641	2584.07	112.26	24	26	22	28	50
ELL	379	2452.28	102.23	72	17	5	5	10
Special Education	1,415	2432.37	91.70	79	16	4	1	5
CD 504	489	2538.35	106.94	37	32	17	14	31
Title I	1,714	2551.92	107.96	33	29	21	17	38

Figure 1. ELA/Lit %Proficient Across Years

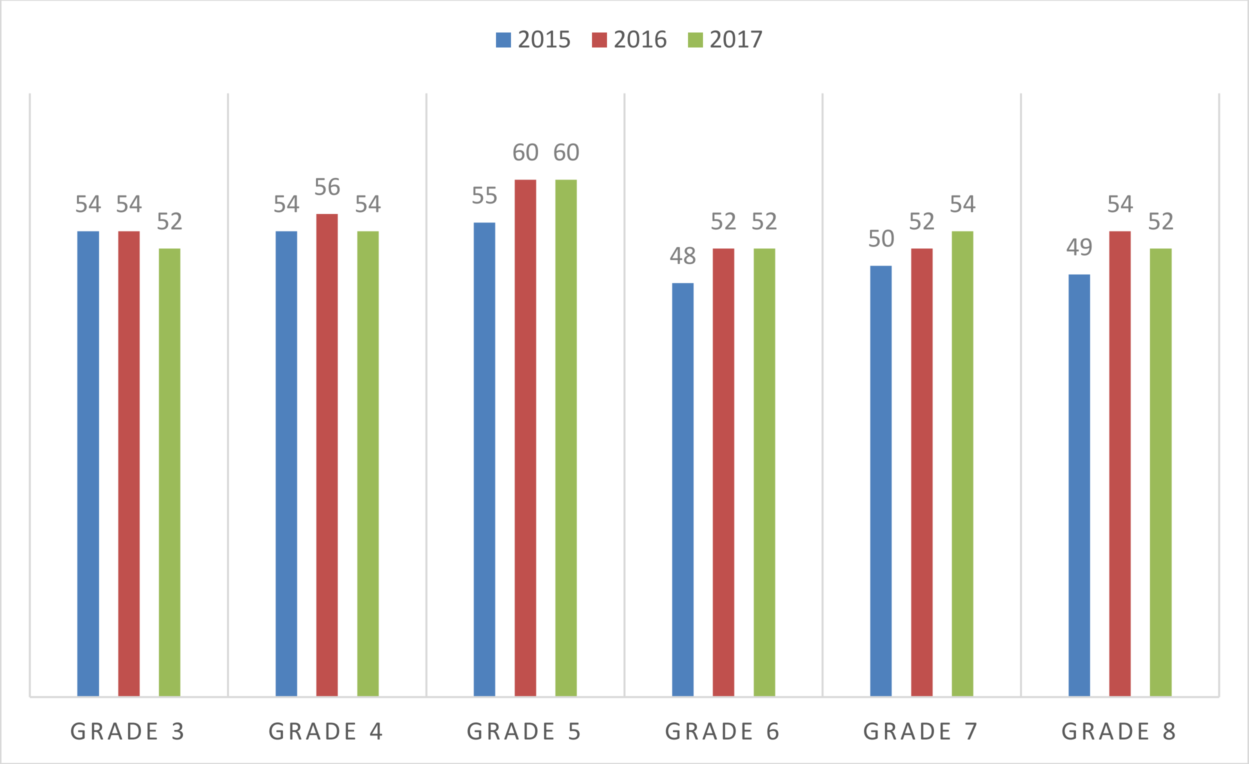
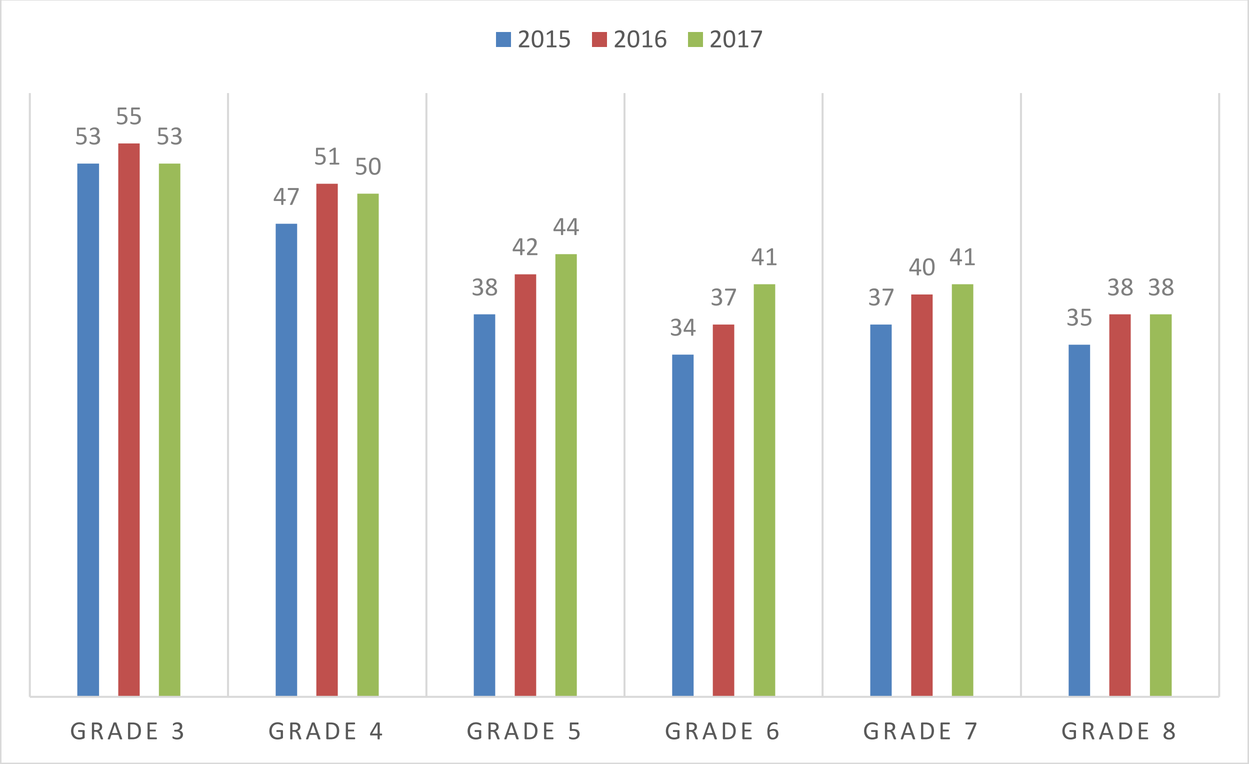


Figure 2. Mathematics %Proficient Across Years



For the reporting categories, because the precision of scores in each reporting category is not sufficient to report scores, given a small number of items, the scores on each reporting category are reported using one of the three performance categories, taking into account the SEM of the reporting category score: (1) Below standard, (2) At/Near standard, or (3) Above standard (see Section 6.5 for the rules). Tables 19 and 20 present the distribution of performance categories for each reporting category. The reporting categories are four claims in ELA/Lit, and three claims in mathematics, combining claims 2 and 4.

Table 19. ELA/Lit Percentage of Students in Performance Categories
for Reporting Categories

Grade	**Performance Category**	**Claim 1: Reading**	**Claim 2: Writing**	**Claim 3: Listening**	**Claim 4: Research**
3	Below	30	25	17	20
	At/Near	44	47	63	51
	Above	26	28	20	29
4	Below	23	25	19	18
	At/Near	49	49	58	53
	Above	28	25	23	29
5	Below	21	18	15	18
	At/Near	48	49	62	47
	Above	30	33	23	34
6	Below	27	27	17	21
	At/Near	50	48	64	51
	Above	24	24	19	28
7	Below	26	24	20	19
	At/Near	47	48	64	52
	Above	27	28	15	29
8	Below	28	26	15	21
	At/Near	45	50	66	51
	Above	27	24	18	27

Table 20. Mathematics Percentage of Students in Performance Categories for Reporting Categories

Grade	**Performance Category**	**Claim 1: Concepts and Procedures**	**Claims 2 & 4: Problem Solving & Modeling and Data Analysis**	**Claim 3: Communicating Reasoning**
3	Below	29	22	18
	At/Near	36	48	50
	Above	35	30	31
4	Below	33	26	24
	At/Near	33	48	46
	Above	34	26	29
5	Below	37	29	28
	At/Near	33	47	48
	Above	30	24	24
6	Below	37	35	32
	At/Near	37	46	47
	Above	26	19	21
7	Below	39	31	26
	At/Near	34	48	55
	Above	27	21	20
8	Below	41	36	32
	At/Near	35	41	48
	Above	24	23	20

3.3 TEST TAKING TIME

The Smarter Balanced assessments are not timed. The time spent on each item may vary among individual students, which may provide us with useful information about student testing behaviors and motivation, for example. Although the length of a test session could be monitored by TAs who are knowledgeable about his/her school and its students, additional time for students who need it would be arranged.

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, while items associated with a stimulus appear on the screen together and the page time is the total time spent on all associated items. In this case, the page time for each item is the average time for all the items associated with the stimulus. For each student, the total testing time for the test was the sum of the page time for all items.

Tables 21 and 22 present an average testing time and the testing time at percentiles for the overall test, the CAT component, and the PT component.

Table 21. ELA/Lit Test Taking Time

Grade	Average Testing Time (hh:mm)	Median Testing Time (hh:mm)	Testing Time in Percentiles (hh:mm)				
			75th	80th	85th	90th	95th
Overall Test							
3	5:03	4:23	6:06	6:40	7:30	8:40	10:52
4	5:25	4:53	6:40	7:10	7:52	8:47	10:34
5	5:22	4:53	6:34	7:04	7:42	8:37	9:57
6	4:35	4:12	5:30	5:52	6:21	7:09	8:35
7	4:13	3:55	5:04	5:24	5:52	6:31	7:40
8	3:55	3:35	4:41	5:02	5:29	6:12	7:13
CAT Component							
3	2:24	2:08	2:53	3:07	3:25	3:54	4:48
4	2:40	2:25	3:13	3:28	3:46	4:15	5:05
5	2:37	2:24	3:09	3:23	3:40	4:05	4:46
6	2:19	2:09	2:47	2:58	3:11	3:30	4:02
7	2:12	2:03	2:37	2:47	3:01	3:18	3:49
8	1:58	1:50	2:21	2:30	2:43	3:00	3:30
PT Component							
3	2:38	2:10	3:20	3:44	4:15	5:03	6:20
4	2:45	2:23	3:29	3:48	4:14	4:54	6:05
5	2:45	2:24	3:28	3:49	4:16	4:52	5:47
6	2:16	2:01	2:49	3:05	3:25	3:56	4:54
7	2:02	1:48	2:34	2:49	3:08	3:34	4:22
8	1:57	1:41	2:26	2:41	3:00	3:26	4:12

Table 22. Mathematics Test Taking Time

Grade	Average Testing Time (hh:mm)	Median Testing Time (hh:mm)	Testing Time in Percentiles (hh:mm)				
			75th	80th	85th	90th	95th
Overall Test							
3	2:46	2:23	3:22	3:42	4:09	4:46	6:01
4	2:43	2:26	3:22	3:38	3:58	4:28	5:14
5	3:25	3:02	4:13	4:37	5:05	5:43	6:45
6	2:51	2:35	3:26	3:43	4:02	4:31	5:21
7	2:17	2:06	2:45	2:57	3:12	3:32	4:11
8	2:34	2:20	3:08	3:24	3:42	4:08	4:53
CAT Component							
3	1:47	1:30	2:09	2:21	2:40	3:07	3:51
4	1:50	1:38	2:17	2:28	2:43	3:04	3:37
5	1:54	1:42	2:21	2:34	2:48	3:08	3:38
6	1:47	1:39	2:10	2:19	2:31	2:48	3:16
7	1:41	1:33	2:01	2:10	2:21	2:36	3:04
8	1:49	1:40	2:14	2:24	2:38	2:56	3:25
PT Component							
3	1:00	0:48	1:15	1:23	1:34	1:51	2:23
4	0:53	0:45	1:05	1:12	1:21	1:35	1:57
5	1:31	1:15	1:54	2:07	2:24	2:49	3:30
6	1:03	0:53	1:18	1:26	1:37	1:54	2:29
7	0:36	0:31	0:46	0:51	0:58	1:06	1:22
8	0:45	0:38	0:56	1:02	1:10	1:21	1:39

3.4 Student Ability–Item Difficulty Distribution for the 2016–2017 Operational Item Pool

Figures 3 and 4 display the empirical distribution of the Delaware student scale scores in the 2016–2017 administration and the distribution of the administered summative item difficulty parameters. 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) to better measure low performing students.

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

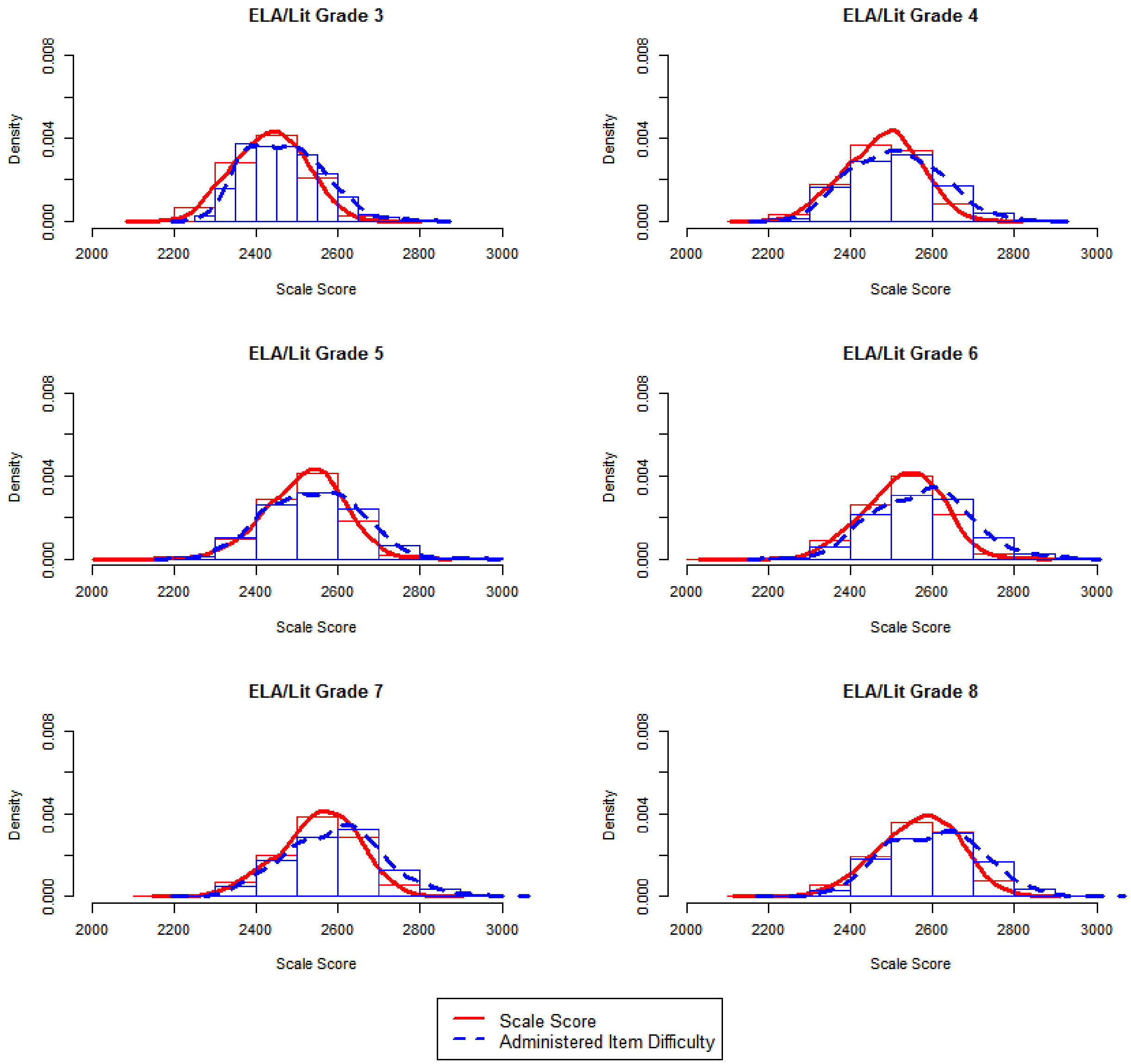
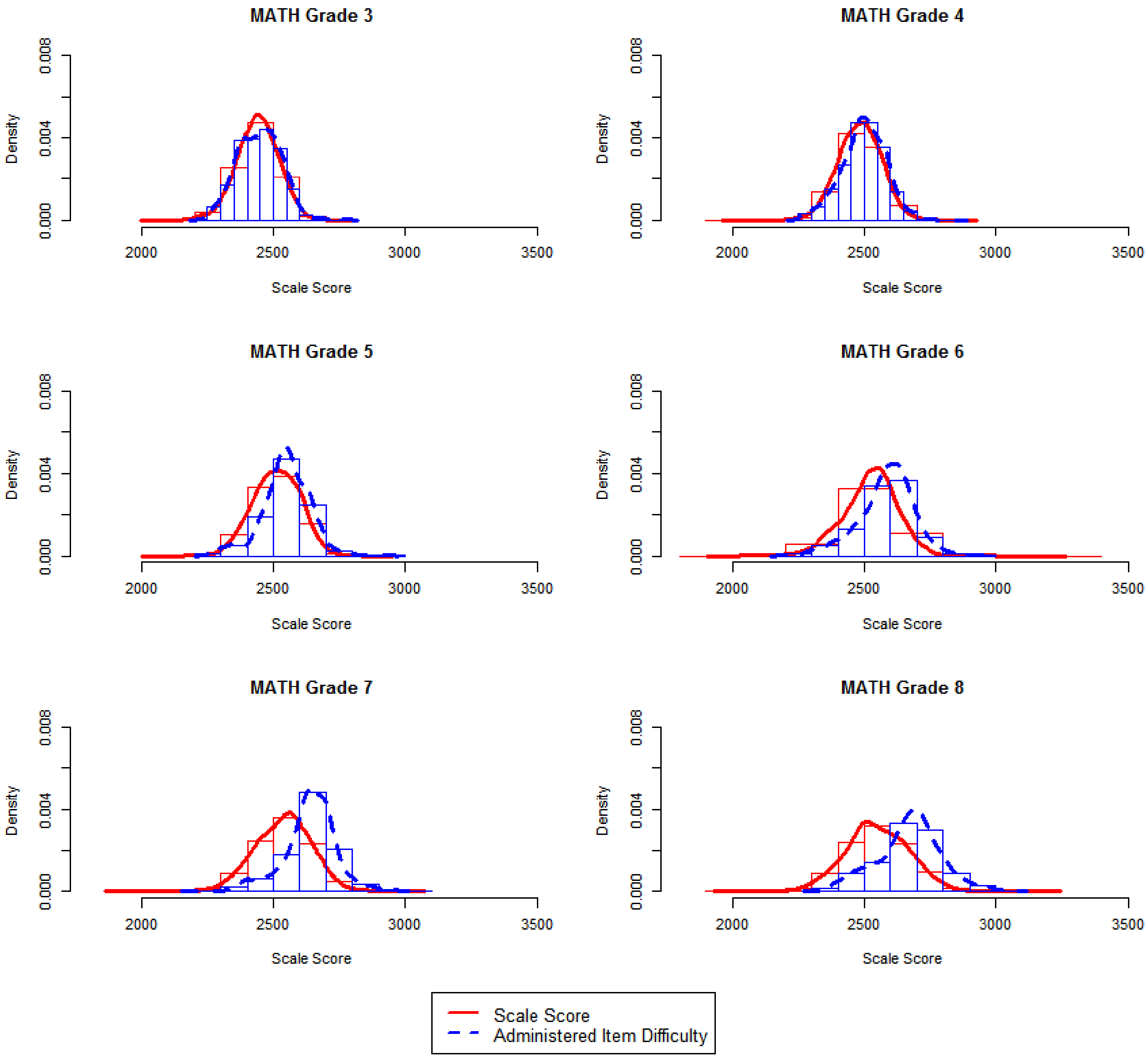


Figure 4. Student Ability–Item Difficulty Distribution for Mathematics



4. VALIDITY

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

Validity evidence provided in this chapter is as follows:

- Test Content
- Internal Structure

Evidence on test content validity is provided with the blueprint match rates for the delivered tests. Evidence on internal structure is examined in the results of intercorrelations among reporting category scores.

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

4.1 Evidence on Test Content

The Smarter Balanced summative assessment includes two components: 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 a fixed-form test. The content coverage in all PT forms is the same.

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

Tables 23 and 24 present the percentages of tests aligned with the test blueprint constraints for ELA/Lit CAT. Table 23 provides the blueprint match rates for item and passage requirements for each claim. All tests met the requirements for item and passage. For DoK and item type constraints, the Smarter Balanced blueprint specifies the minimum number of items, not the maximum. Table 24 presents the percentages of tests that satisfied the DoK and item type constraints for each claim. All tests met the requirements, except for the claim 2 DoK2 requirement in grades 3 and 6, which each administered one DoK2 item fewer than required in claim 2.

Tables 25–26 provide the percentages of tests aligned with the test blueprint constraints for mathematics CAT, the blueprint match rates for claims, DoK, and target constraints. In mathematics, all tests met all blueprint requirements, except for grade 3, 6 and 8. In grade 3, the violation was one or two items fewer than required in claim 1 no-calculator segment for target sets of E, J and K. In grade 6, the violation was in claim 1 no-calculator segment for target sets of E and F and target B and claim 3 calculator segement for target sets of A and D, each administered some fewer or more items than required. In grade 8, the violation was in claim 1 no-calculator segment for Target B, Target C, and DoK2 or higher, which administered one item fewer or more than the item requirement.

Table 23. ELA/Lit Percentage of Delivered Tests Meeting Blueprint Requirements for Each Claim and the Number of Passages Administered

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

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

Table 24. ELA/Lit Percentage of Delivered Tests Meeting Blueprint Requirements for Depth-of-Knowledge and Item Type

DoK and Item Type Constraints	**Minimum Required Items**	**%BP Match**					
		Grade 3	**Grade 4**	**Grade 5**	**Grade 6**	**Grade 7**	**Grade 8**
Claim 1 DoK2	7	100%	100%	100%	100%	100%	100%
Claim 1 DoK3 or higher	2	100%	100%	100%	100%	100%	100%
Claim 2 DoK2	4	91%	100%	100%	70%	100%	100%
Claim 2 DoK3 or higher	1	100%	100%	100%	100%	100%	100%
Claim 2 Brief Write	1	100%	100%	100%	100%	100%	100%
Claim 3 DoK2 or higher	3	100%	100%	100%	100%	100%	100%

Table 25. Mathematics Percentage of Delivered Tests Meeting Blueprint Requirements for Each Claim and Targets (Grades 3–5)

Claim	**Target**	**Grade 3**		**Grade 4**		**Grade 5**	
		Required Items	**%BP Match**	**Required Items**	**%BP Match**	**Required Items**	**%BP Match**
Total Adaptive Test Length		34	100%	34	100%	34	100%
1	Overall	20	100%	20		20	100%
	Priority Cluster	15	100%				
	Targets B, C, G, I	6	100%				
	Targets D, F	6	100%				
	Target A	3	100%				
	Supporting Cluster	5	100%				
	Targets E, J, K	4	99%				
	Target H	1	100%				
	Priority Cluster			15	100%		
	Target A, E, F			9	100%		
	Target G			3	100%		
	Target D			2	100%		
	Target H			1	100%		
	Supporting Cluster			5	100%		
	Target I, K			3	100%		
	Target B, C, J			1	100%		
	Target L			1	100%		
	Priority Cluster					15	100%
	Target E, I					6	100%
	Target F					5	100%
	Target C, D					4	100%
	Supporting Cluster					5	100%
	Target J, K					3	100%
	Target A, B, G, H					2	100%
	DOK 2 or higher	7	100%	7	100%	7	100%
2	Overall	3	100%	3	100%	3	100%
	Target A	2	100%	2	100%	2	100%
	Targets B, C, D	1	100%	1	100%	1	100%
3	Overall	8	100%	8	100%	8	100%
	Targets A, D	3	100%	3	100%	3	100%
	Targets B, E	3	100%	3	100%	3	100%
	Targets C, F	2	100%	2	100%	2	100%
	DOK 3 or higher	2	100%	2	100%	2	100%
4	Overall	3	100%	3	100%	3	100%
	Targets A, D	1	100%	1	100%	1	100%
	Targets B, E	1	100%	1	100%	1	100%
	Targets C, F	1	100%	1	100%	1	100%
2&4	DOK 3 or higher	2	100%	2	100%	2	100%

Table 26. Mathematics Percentage of Delivered Tests Meeting Blueprint Requirements for Each Claim and Targets (Grades 6–8)

Claim	**Target**	**Grade 6**		**Grade 7**		**Grade 8**	
		Required Items	**%BP Match**	**Required Items**	**%BP Match**	**Required Items**	**%BP Match**
Total Adaptive Test Length		33	100%	34	100%	34	100%
1-Calc	Overall	6	100%	10	100%	14	100%
	Priority Cluster	3	100%	6	100%	11	100%
	Target A	2	100%				
	Target G	1	100%				
	Targets A, D			6	100%		
	Target D					4	100%
	Targets E, G					4	100%
	Targets F, H					3	100%
	Supporting Cluster	3	100%	4	100%	3	100%
	Targets H, I, J	3	100%				
	Targets E, F			2	100%		
	Targets G, H, I			2	100%		
	Targets I, J					3	100%
	DOK 2 or higher	2	100%	4	100%	5	100%
1-No Calc	Overall	13	100%	10	100%	6	100%
	Priority Cluster	11	100%	9	100%	4	100%
	Targets E, F	6	99%				
	Target A	2	100%				
	Target B	1	99%			2	87%
	Target D	2		3	100%		
	Target B, C			6	100%		
	Target C					2	87%
	Supporting Cluster	2	100%	1	100%	2	100%
	Target C	2	100%				
	Target E			1	100%		
	Target A					2	100%
	DOK 2 or higher	5	100%	4	100%	4	94%
2	Overall	3	100%	3	100%	3	100%
	Target A	2	100%	2	100%	2	100%
	Targets B, C, D	1	100%	1	100%	1	100%
3-Calc	Overall	7	100%	8	100%	8	100%
	Targets A, D	3	99%	2–3	100%	2–3	100%
	Targets B, E	2–3	100%	3	100%	3	100%
	Targets C, F, G	2	100%	1–2	100%	1–2	100%
	DOK 3 or higher	1	100%	2	100%	2	100%
3-NoCalc	Overall	1	100%				
4	Overall	3	100%	3	100%	3	100%
	Targets A, D	1	100%	1	100%	1	100%
	Targets B, E	1	100%	1	100%	1	100%
	Targets C, F	1	100%	1	100%	1	100%
2&4	DOK 3 or higher	2	100%	2	100%	2	100%

Table 27 summarizes the target coverage, the average and the range of 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 27. Average and Range of the Number of Unique Targets Assessed Within Each Claim Across all Delivered Tests

Grade	**Total Targets in BP**				**Mean**				**Range (Minimum – Maximum)**			
	C1	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**
ELA/Lit												
3	14	5	1	3	10.2	4.0	1	3	8–13	3–5	1–1	3–3
4	14	5	1	3	10.3	4.1	1	3	8–13	3–5	1–1	3–3
5	14	5	1	3	10.1	4.7	1	3	7–13	3–5	1–1	3–3
6	14	5	1	3	9.3	4.1	1	3	8–11	3–5	1–1	3–3
7	14	5	1	3	9.4	4.9	1	3	7–11	3–5	1–1	3–3
8	14	5	1	3	9.4	4.0	1	3	8–11	3–4	1–1	3–3
Mathematics												
3	11	4	6	6	10.8	2	5.5	3	9–11	2–2	4–6	2–3
4	12	4	6	6	10.0	2	5.5	3	9–10	2–2	3–6	3–3
5	11	4	6	6	9.0	2	5.3	3	9–9	2–2	3–6	3–3
6	10	4	7	6	10.0	2	4.8	3	8–10	1–2	3–7	3–3
7	9	3	7	6	8.0	2	4.8	3	8–8	2–2	3–6	3–3
8	10	4	7	6	10.0	2	4.8	3	10–10	2–2	3–6	3–4

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

4.2 Evidence on Internal Structure

The measurement and reporting model used in the Smarter Balanced 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 28 and 29. The correction for attenuation indicates what the correlation would be if reporting category scores could be measured with perfect reliability, corrected (adjusted) for measurement error estimates.

The observed correlation between two reporting category scores with measurement errors can be corrected for attenuation as $r_{x'y'} = \frac{r_{xy}}{\sqrt{r_{xx}*r_{yy}}}$, where $r_{x'y'}$ is the correlation between x and y corrected for attenuation, r_{xy} is the observed correlation between x and y, r_{xx} is the reliability coefficient for x, and r_{yy} is the reliability coefficient for y.

When corrected for attenuation (above diagonal), the correlations among reporting scores are higher than observed correlations. The disattenuated correlations are quite high, especially in mathematics. The correction for attenuation is large in mathematics because the marginal reliabilities of claim 2 and 4 and claim 3 scores are low. The low reliabilities are due to the low performance with large standard errors and a shortage of easy items in the item pool.

Because the reliabilities for reporting category scores are low, the performance of each reporting category scores is reported in three performance categories. The distribution of performance categories for each reporting category is provided in Tables 19–20, Section 3.2. Scale scores are not reported for reporting categories.

Table 28. Correlations Among Reporting Categories for ELA/Lit

Grade	Claim	Observed & Disattenuated Correlation			
		Claim 1	Claim 2	Claim 3	Claim 4
3	Claim 1: Reading		0.88	0.93	0.90
	Claim 2: Writing	0.70		0.87	0.87
	Claim 3: Listening	0.64	0.61		0.89
	Claim 4: Research	0.67	0.66	0.58	
4	Claim 1: Reading		0.90	0.90	0.90
	Claim 2: Writing	0.68		0.85	0.89
	Claim 3: Listening	0.63	0.61		0.90
	Claim 4: Research	0.65	0.66	0.61	
5	Claim 1: Reading		0.90	0.91	0.92
	Claim 2: Writing	0.69		0.86	0.90
	Claim 3: Listening	0.63	0.61		0.90
	Claim 4: Research	0.69	0.70	0.63	
6	Claim 1: Reading		0.91	0.93	0.92
	Claim 2: Writing	0.70		0.91	0.90
	Claim 3: Listening	0.63	0.64		0.89
	Claim 4: Research	0.67	0.69	0.59	
7	Claim 1: Reading		0.90	0.92	0.94
	Claim 2: Writing	0.71		0.89	0.93
	Claim 3: Listening	0.63	0.62		0.92
	Claim 4: Research	0.70	0.71	0.60	
8	Claim 1: Reading		0.91	0.91	0.94
	Claim 2: Writing	0.72		0.88	0.92
	Claim 3: Listening	0.62	0.60		0.89
	Claim 4: Research	0.70	0.69	0.58	

Table 29. Correlations Among Reporting Categories for Mathematics

Grade	**Reporting Categories**	**Observed & Disattenuated Correlation**		
		Claim 1	**Claim 2&4**	**Claim 3**
3	Claim 1		0.97	0.94
	Claim 2 & 4	0.78		1.00
	Claim 3	0.77	0.73	
4	Claim 1		0.97	0.96
	Claim 2 & 4	0.80		0.99
	Claim 3	0.79	0.76	
5	Claim 1		1.00	0.95
	Claim 2 & 4	0.79		1.00
	Claim 3	0.77	0.74	
6	Claim 1		1.00	0.96
	Claim 2 & 4	0.82		1.00
	Claim 3	0.78	0.76	
7	Claim 1		1.00	0.98
	Claim 2 & 4	0.79		1.00
	Claim 3	0.76	0.70	
8	Claim 1		1.00	0.97
	Claim 2 & 4	0.80		1.00
	Claim 3	0.77	0.73	

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

5. RELIABILITY

Reliability refers to the consistency in test scores. Reliability is evaluated in terms of the standard errors of measurement (SEM). In classical test theory, reliability is defined as the ratio of the true score variance to the observed score variance, assuming the error variance is the same for all scores. Within the 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 assessments is provided with marginal reliability, SEM, and classification 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 SEM, estimated at different points on the ability scale, for all students.

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

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

where N is the number of students; $CSEM_i$ is the conditional 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, the greater the precision of the test.

Another way to examine test reliability is with 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, the greater the accuracy of test scores.

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

Table 30. Marginal Reliability for ELA/Lit and Mathematics

Grade	N	Number of Items Specified in Test Blueprint		Marginal Reliability	Scale Score Mean	Scale Score SD	Average CSEM
		Min	**Max**				
ELA/Lit							
3	10,600	41	44	0.93	2433.34	87.23	23.71
4	10,386	40	44	0.92	2477.16	92.05	26.03
5	10,461	41	45	0.92	2519.67	93.28	25.74
6	10,189	41	45	0.92	2529.67	93.36	25.92
7	10,070	41	45	0.92	2553.72	97.76	26.83
8	10,069	43	45	0.92	2565.99	99.66	27.62
Mathematics							
3	10,669	39	40	0.94	2440.96	79.38	19.26
4	10,442	37	40	0.94	2483.34	82.58	19.40
5	10,519	38	40	0.94	2511.47	89.71	22.23
6	10,211	38	39	0.94	2523.83	103.51	25.44
7	10,087	38	40	0.94	2538.73	109.13	27.76
8	10,058	38	40	0.94	2550.50	119.68	29.85

5.2 STANDARD ERROR CURVES

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

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

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

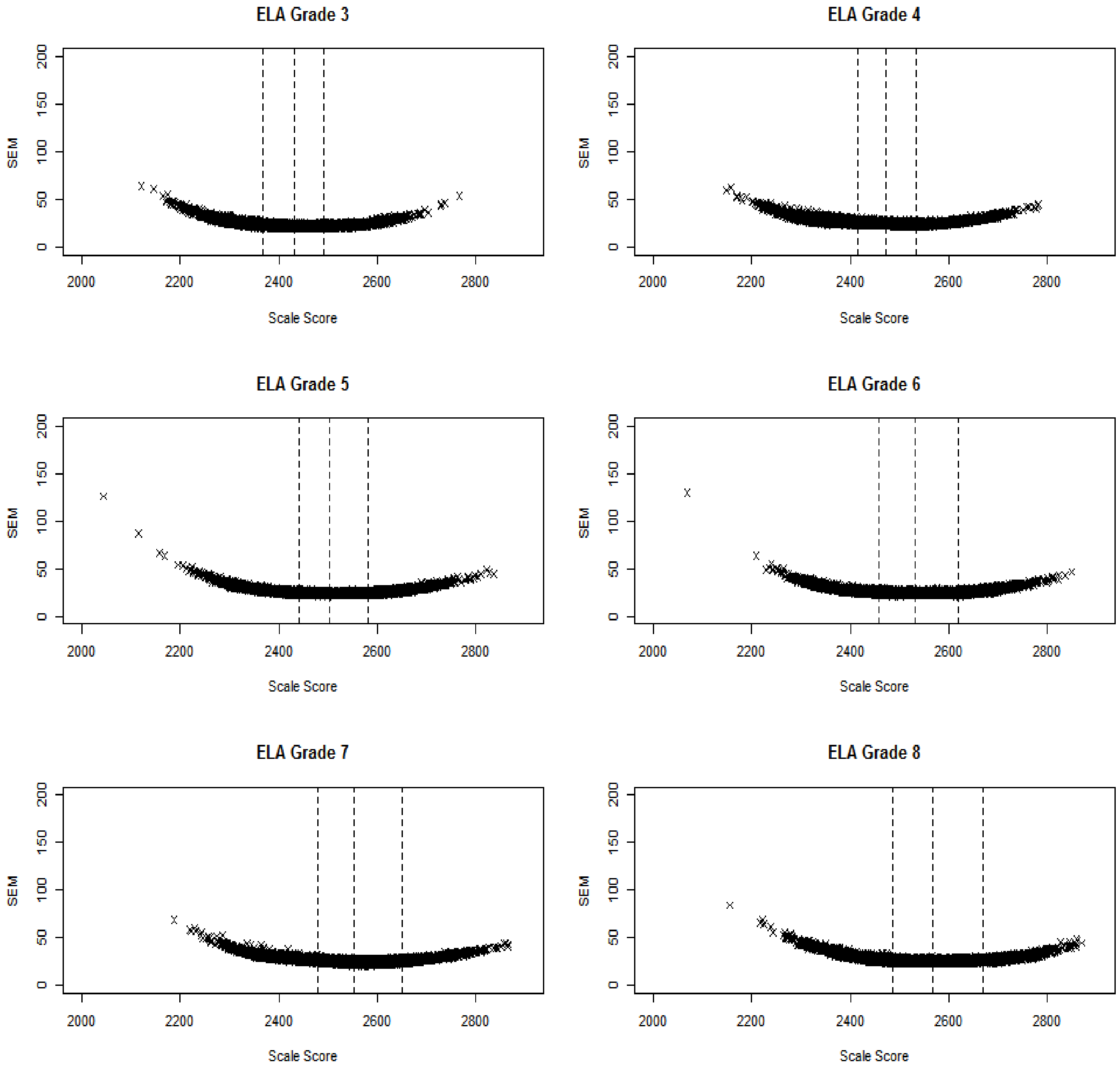
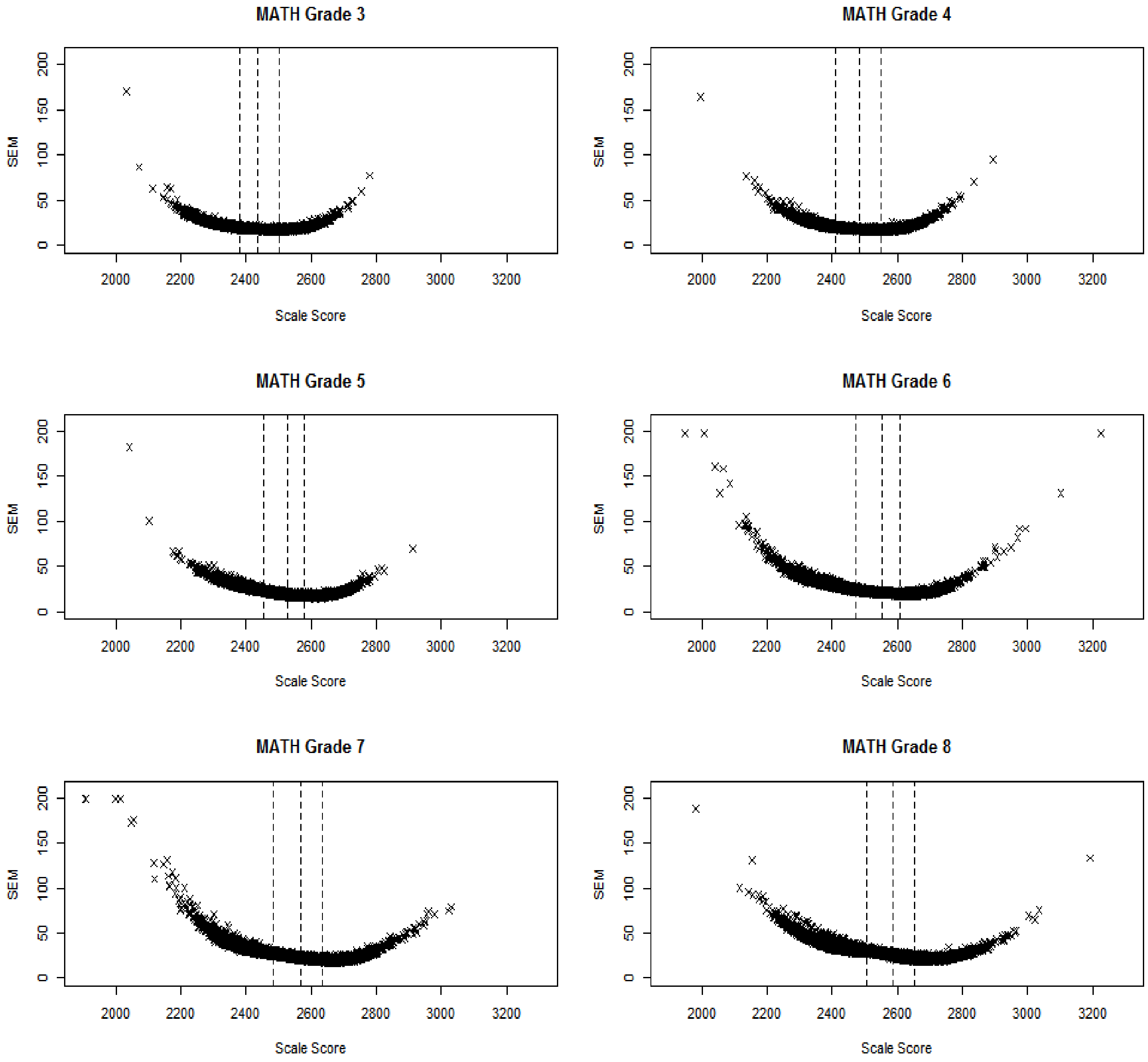


Figure 6. Conditional Standard Error of Measurement for Mathematics



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

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

Grade	Level 1	Level 2	Level 3	Level 4	Average CSEM
ELA/Lit					
3	26.83	22.40	21.99	23.52	23.71
4	28.51	25.27	24.27	25.73	26.03
5	28.08	24.22	24.16	26.89	25.74
6	28.63	24.46	24.65	26.85	25.92
7	30.40	25.54	24.85	28.03	26.83
8	31.12	26.03	25.88	28.97	27.62
Mathematics					
3	23.37	18.36	17.04	18.75	19.26
4	24.24	18.35	17.02	19.18	19.40
5	28.72	20.64	18.02	18.75	22.23
6	33.18	22.18	20.09	22.54	25.44
7	36.29	24.80	21.32	22.54	27.76
8	36.39	28.31	23.74	23.16	29.85

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

Grade	L2 Cut	L3 Cut	L4 Cut	|L2–L3|	|L3–L4|	|L2–L4|
ELA/Lit						
3	23.16	21.83	21.96	1.33	0.13	1.20
4	25.87	24.96	23.98	0.91	0.98	1.89
5	24.49	23.83	24.70	0.66	0.87	0.21
6	25.06	24.40	25.24	0.66	0.84	0.18
7	26.43	24.74	25.34	1.69	0.60	1.09
8	26.18	25.67	27.03	0.51	1.36	0.85
Mathematics						
3	19.06	17.74	16.73	1.32	1.01	2.33
4	19.57	17.27	16.67	2.30	0.60	2.90
5	23.36	18.75	17.48	4.61	1.27	5.88
6	24.20	20.97	19.62	3.23	1.35	4.58
7	27.02	22.64	20.39	4.38	2.25	6.63
8	30.13	26.04	21.39	4.09	4.65	8.74

5.3 Reliability of Achievement Classification

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

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

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

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

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

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

Instead of assuming a normal distribution of $\hat{\theta}_i \sim N\left(\theta_i, se^2(\hat{\theta}_i)\right)$, we can estimate the above probabilities directly using the likelihood function.

The likelihood function of theta given a student's item scores represents the likelihood of the student's ability at that theta value. Integrating the likelihood values over the range of theta at and above the cut point (with proper normalization) represents the probability of the student's latent ability or the true score being at or above that cut point. If a student with estimated theta is below the cut point, 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

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

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

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

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

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

where d stands for dichotomous and p stands for polytomous items; $\mathbf{b}_j = (a_j, b_j, c_j)$ if the *j*th item is a dichotomous item, and $\mathbf{b}_j = (a_j, b_{j1}, \ldots, b_{jK_i})$ if the *j*th item is a polytomous item; a_j is the item's discrimination parameter (for Rasch model, $a_j = 1$), c_j is the guessing parameter (for Rasch and 2PL models, $c_j = 0$), D is 1.7 for non-Rasch models and 1 for Rasch model.

Classification Accuracy

Using p_{il}, we can construct a $L \times L$ table as

$$\begin{pmatrix} n_{a11} & \cdots & n_{a1L} \\ \vdots & \vdots & \vdots \\ n_{aL1} & \cdots & n_{aLL} \end{pmatrix}$$

where $n_{alm} = \sum_{pl_i=l} p_{im}$. n_{alm} is the expected count of students at achievement level *lm*, pl_i is the *i*th student's achievement level, and p_{im} are the probabilities of the *i*th student being classified at achievement level *m*. In the above table, the row represents the observed level and the column represents the expected level.

The classification accuracy (*CA*) at level l ($l = 1,\cdots,L$) is estimated by

$$CA_l = \frac{n_{all}}{\sum_{m=1}^{L} n_{alm}},$$

and the overall classification accuracy is estimated by

$$CA = \frac{\sum_{l=1}^{L} n_{all}}{N},$$

where N is the total number of students.

Classification Consistency

Using p_{il}, similar to accuracy, we can construct another $L \times L$ table by assuming the test is administered twice independently to the same student group, hence we have

$$\begin{pmatrix} n_{c11} & \cdots & n_{c1L} \\ \vdots & \vdots & \vdots \\ n_{cL1} & \cdots & n_{cLL} \end{pmatrix}$$

where $n_{clm} = \sum_{i=1}^{N} p_{il} p_{im}$. p_{il} and p_{im} are the probabilities of the ith student being classified at achievement level *l* and *m*, respectively based on observed scores and hypothetical scores from equivalent test form.

The classification consistency (*CC*) at level l ($l = 1, \cdots, L$) is estimated by

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

and the overall classification consistency is

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

The analysis of the classification index is performed based on overall scale scores. Table 33 provides the percentage of classification accuracy and consistency for overall and by achievement level.

The overall classification index ranged from 79% to 83% for the accuracy and from 71% to 77% for the consistency across all grades and subjects. For achievement levels, the classification index is higher in L1 and L4 than in L2 and L3. The higher accuracy at L1 and L4 is due to the intervals used to compute the classification probability to classify students into L1 [$-\infty$, L2 cut] or L4 [L4 cut, ∞] being wider than the intervals used in L2 [L2 cut, L3 cut] and L3 [L3 cut, L4 cut]. The misclassification probability tends to be higher for narrower intervals.

The accuracy of classifications is 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 classification indexes by subgroups are provided in Appendix C.

Table 33. Classification Accuracy and Consistency by Achievement Levels

Grade	**Achievement Level**	**ELA/Lit**		**Mathematics**	
		% Accuracy	**% Consistency**	**% Accuracy**	**% Consistency**
3	Overall	80	72	82	75
	L1	89	82	89	82
	L2	72	62	73	63
	L3	70	59	79	72
	L4	89	83	90	84
4	Overall	79	71	83	77
	L1	89	83	89	82
	L2	65	53	80	73
	L3	68	58	79	71
	L4	88	82	89	84
5	Overall	80	72	83	76
	L1	88	82	89	84
	L2	68	56	77	69
	L3	76	68	72	61
	L4	87	81	90	86
6	Overall	80	72	83	76
	L1	90	83	91	86
	L2	73	64	78	70
	L3	78	70	72	63
	L4	84	76	89	83
7	Overall	81	73	83	76
	L1	90	83	91	85
	L2	72	62	76	67
	L3	80	73	74	65
	L4	84	76	89	84
8	Overall	81	73	82	75
	L1	89	82	90	85
	L2	74	64	72	63
	L3	80	73	71	61
	L4	84	75	91	86

5.4 Reliability for Subgroups

The reliability of test scores and achievement levels are also computed by subgroups. Tables 34 and 35 present the marginal reliability coefficients by subgroups. The reliability coefficients are similar across subgroups, but somewhat lower for English Language Learners (ELL) and special education subgroups, a large percentage of whom received Level 1 with large SEMs.

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

Subgroup	**Grade 3**	**Grade 4**	**Grade 5**	**Grade 6**	**Grade 7**	**Grade 8**
All Students	0.93	0.92	0.92	0.92	0.92	0.92
Female	0.92	0.92	0.92	0.92	0.92	0.92
Male	0.93	0.92	0.92	0.92	0.93	0.92
American Indian/Alaska Native	0.92	0.90	0.93	0.89	0.91	0.91
Asian	0.91	0.90	0.91	0.91	0.91	0.92
African American	0.91	0.91	0.92	0.91	0.92	0.91
Hispanic	0.91	0.90	0.91	0.91	0.91	0.92
White	0.92	0.91	0.92	0.91	0.91	0.91
English Language Learners	0.90	0.87	0.86	0.83	0.85	0.86
Special Education	0.87	0.87	0.88	0.86	0.87	0.86
CD 504	0.91	0.91	0.90	0.90	0.91	0.91
Title I	0.91	0.90	0.91	0.91	0.92	0.91

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

Subgroup	**Grade 3**	**Grade 4**	**Grade 5**	**Grade 6**	**Grade 7**	**Grade 8**
All Students	0.94	0.94	0.94	0.94	0.94	0.94
Female	0.94	0.94	0.94	0.94	0.93	0.94
Male	0.94	0.95	0.94	0.94	0.94	0.94
American Indian/Alaska Native	0.95	0.94	0.93	0.93	0.90	0.94
Asian	0.94	0.93	0.94	0.95	0.95	0.96
African American	0.93	0.93	0.91	0.92	0.91	0.91
Hispanic	0.93	0.93	0.92	0.92	0.92	0.92
White	0.94	0.94	0.94	0.94	0.94	0.94
English Language Learners	0.93	0.91	0.84	0.84	0.82	0.86
Special Education	0.91	0.91	0.86	0.86	0.83	0.84
CD 504	0.93	0.94	0.93	0.93	0.92	0.92
Title I	0.93	0.94	0.93	0.93	0.93	0.93

5.5 RELIABILITY FOR CLAIM SCORES

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

Table 36. ELA/Lit Marginal Reliability Coefficients for Claim Scores

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	0.79	2426.79	101.46	46.71
	Claim 2: Writing	11	11	0.80	2433.30	100.33	44.64
	Claim 3: Listening	8	8	0.60	2432.07	121.98	77.24
	Claim 4: Research	8	9	0.71	2435.11	118.10	63.43
4	Claim 1: Reading	14	16	0.74	2473.14	107.46	54.28
	Claim 2: Writing	11	11	0.78	2474.18	104.84	49.11
	Claim 3: Listening	8	8	0.66	2473.28	135.82	79.16
	Claim 4: Research	7	9	0.70	2480.12	120.47	65.88
5	Claim 1: Reading	14	16	0.75	2514.01	110.39	55.38
	Claim 2: Writing	11	11	0.78	2524.61	105.06	48.73
	Claim 3: Listening	8	9	0.63	2516.70	130.96	79.33
	Claim 4: Research	8	9	0.76	2519.52	120.38	58.65
6	Claim 1: Reading	14	16	0.74	2521.24	113.09	57.15
	Claim 2: Writing	11	11	0.81	2527.33	100.86	44.42
	Claim 3: Listening	8	9	0.61	2553.03	144.07	89.44
	Claim 4: Research	8	9	0.71	2522.85	127.22	68.26
7	Claim 1: Reading	14	16	0.78	2550.40	114.93	54.29
	Claim 2: Writing	11	11	0.80	2552.12	111.30	49.37
	Claim 3: Listening	8	9	0.60	2552.52	134.57	84.67
	Claim 4: Research	8	9	0.71	2552.72	128.30	68.94
8	Claim 1: Reading	16	16	0.77	2562.95	114.70	54.50
	Claim 2: Writing	11	11	0.79	2562.12	114.00	51.63
	Claim 3: Listening	8	9	0.59	2577.85	145.27	93.09
	Claim 4: Research	8	9	0.71	2561.68	130.87	69.91

Table 37. Mathematics Marginal Reliability Coefficients for Claim Scores

Grade	Reporting Categories	Number of Items Specified in Test Blueprint		Marginal Reliability	Scale Score Mean	Scale Score SD	Average CSEM
		Min	Max				
3	Claim 1	20	20	0.90	2442.41	86.59	28.05
	Claims 2 & 4	8	11	0.73	2434.79	91.25	47.34
	Claim 3	9	11	0.73	2439.17	96.12	49.61
4	Claim 1	20	20	0.90	2484.72	89.49	28.33
	Claims 2 & 4	8	10	0.76	2478.53	94.42	45.95
	Claim 3	9	10	0.76	2480.79	97.49	47.97
5	Claim 1	20	20	0.89	2513.86	97.66	31.93
	Claims 2 & 4	8	10	0.68	2504.20	99.77	56.62
	Claim 3	9	10	0.73	2505.54	111.04	57.85
6	Claim 1	19	19	0.89	2530.17	111.88	36.82
	Claims 2 & 4	9	10	0.73	2509.18	118.02	61.55
	Claim 3	10	11	0.74	2516.24	119.94	61.16
7	Claim 1	20	20	0.89	2543.69	117.91	38.50
	Claims 2 & 4	10	10	0.65	2520.93	129.67	76.94
	Claim 3	8	10	0.67	2526.80	132.22	75.71
8	Claim 1	20	20	0.89	2553.05	128.28	42.99
	Claims 2 & 4	8	10	0.72	2541.44	138.17	73.34
	Claim 3	9	10	0.72	2535.73	145.52	77.04

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

6. SCORING

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

6.1 Estimating Student Ability Using Maximum Likelihood Estimation

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

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

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

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

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

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

in the case of items with two or more points,

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

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

Standard Error of Measurement

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

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

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

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

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 subject 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 the Smarter Balanced Assessment Consortium. Table 38 presents the scaling constants for each subject for the theta-to-scale score linear transformation. Scale scores are rounded to an integer.

Table 38. Vertical Scaling Constants on the Reporting Metric

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

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

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

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

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

Table 39. Cut Scores in Scale Scores

Grade	**ELA/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

6.3 LOWEST/HIGHEST OBTAINABLE SCORES (LOSS/HOSS)

In 2014–2015 administration, Delaware applied the Smarter LOSS/HOSS to truncate extreme student ability estimates in both theta and scale score metrics. Starting in 2015–2016 administration, LOSS and HOSS truncation rule was removed.

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 or the lowest obtainable scores were assigned in the 2014–2015 administration. Since 2015–2016 administration, all incorrect and correct cases were scored by either adding 0.5 to or subtracting 0.5 from an item score with the smallest item discrimination parameter among the administered operational items (CAT and PT) for a student.

6.5 RULES FOR CALCULATING STRENGTHS AND WEAKNESSES FOR REPORTING CATEGORIES (CLAIM SCORES)

In ELA/Lit, claim scores are computed for each claim. In mathematics, claim scores are computed for claim 1, claims 2 and 4 combined, and claim 3. For each claim, three performance categories, relative strength and weakness, are produced.

If the difference between the proficiency cut score and the claim score is greater (or less) than 1.5 times standard error of the claim, a plus or minus indicator appears on the student's score report as shown in Section 7.

For summative tests, the specific rules are as follows:

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

where SS_{rc} is the student's scale score on a reporting category; SS_p is the proficiency scale score cut (Level 3 cut); and $SE(SS_{rc})$ is the standard error of the student's scale score on the reporting category.

6.6 TARGET SCORES

The target-level reports are not possible to produce for a fixed-form test because the number of items included per target (i.e., benchmark) is too few to produce a reliable score at the target level. A typical fixed-form test includes only one or two items per target. Even when aggregated, these data reflect the benchmark narrowly because they reflect only one or two ways of measuring the target. 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 in each claim (four claims) in ELA/Lit and Claim 1 only in mathematics.

Target scores are computed relative to the proficiency standard (Level 3 cut).

By defining $p_{ij} = p(z_{ij} = 1)$, representing the probability that student *j* responds correctly to item *i* (z_{ij} represents the *j*th student's score on the *i*th item). For items with one score point we use the 2PL IRT model to calculate the expected score on item *i* for student *j* with $\theta_{Level\ 3\ cut}$ as:

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

For items with two or more score points, using the generalized partial credit model, the expected score for student *j* with *Level 3 cut* on an item *i* with a maximum possible score of m_i is calculated as:

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

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

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

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

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

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

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

where n_g is the number of students who responded to any of the items that belong to the target T for an aggregate unit g. If a student did not happen to see any items on a particular target, the student is NOT included in the n_g count for the aggregate.

A statistically significant difference from zero in these aggregates may indicate that a class, teacher, school, or district is more effective (if $\bar{\delta}_{Tg}$is positive) or less effective (negative $\bar{\delta}_{Tg}$) in teaching a given target.

We do not suggest direct reporting of the statistic $\bar{\delta}_{Tg}$; instead, we recommend reporting whether, in the aggregate, a group of students performs better, worse, or as expected on this target. In some cases, insufficient information will be available and that will be indicated as well.

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

- If $\bar{\delta}_{Tg} \geq +1 * \ se(\bar{\delta}_{Tg})$, then performance is *above* the Proficiency Standard.
- If $\bar{\delta}_{Tg} \leq -1 * \ se(\bar{\delta}_{Tg})$, then performance is *below* the Proficiency Standard.
- Otherwise, performance is *near* the Proficiency Standard.

If $se(\bar{\delta}_{Tg})$ > 0.2, data are insufficient.

6.7 HANDSCORING

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

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

6.7.1 Reader Selection

MI maintains a large pool of readers at each scoring center, as well as distributive readers who work remotely from their homes. Experienced readers are defined as those who have worked on one or more previous projects and typically comprise 50–65% of all readers. 2016–2017 was the third year that MI scored operational Smarter Balanced assessments, and it is estimated that approximately twice as many experienced readers returned in comparison to 2015–2016, particularly in the distributive reader pool. MI only needs to inform experienced readers that a project is pending and invite them to return. MI routinely maintains supervisors' evaluations and performance data for each person who works on each scoring project in order to determine employment eligibility for future projects. MI employs many of these experienced readers for the Smarter Balanced project and recruits new ones as well.

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

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

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

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

6.7.2 Reader Training

All readers hired for Smarter Balanced assessment hand-scoring are trained using the rubric(s), anchor sets, and training/qualifying sets provided by Smarter Balanced. These sets were created during the original field-test scoring in 2014 and approved by Smarter Balanced. The same anchor sets are used each year. The only changes made to anchor sets across the years include occasional updates to annotations and removal of individual responses, as determined during annual meetings between the vendors and Smarter Balanced. Additionally, several of the Brief Writes anchor sets were revised between the 2014–2015 and 2015–2016 test administrations. Finally, based on challenges observed scoring the 2014–2015 and 2015–2016 administrations, in the summer of 2016 MI scoring managers developed additional item-level supplemental training materials for their respective content areas to use in conjunction with the Smarter Balanced-provided materials.

Once hired, readers are placed into a scoring group that corresponds to the subject/grade that they are deemed best suited to score (based on work history, results of the placement assessments, and performance on past scoring projects). Readers are trained on a specific item type (i.e., Brief Writes, Reading, Research, Full Writes, and/or Mathematics). Within each group, readers are divided into teams consisting of one team leader and 10–15 readers. Each team leader and reader is assigned a unique number for easy identification of their scoring work throughout the scoring session. For the 2016–2017 administration, scoring directors attempted to minimize the number of items an individual reader scored so that the reader became highly experienced in scoring responses to those items.

MI’s Virtual Scoring Center (VSC) includes an online training interface which presents rubrics, scoring guides, and training/qualifying sets. Readers are trained by a scoring director (in-person) or using scripted videos (online). The same training protocol is followed for both site-based and distributive readers.

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

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

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

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

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

Training design varies slightly depending on Smarter Balanced item type:

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

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

Multiple strategies are used to minimize rater bias. First, readers do not have access to any student identifiers. Unless the students sign their names, write about their hometowns, or in some way provide other identifying information as part of their response, the readers have no knowledge of student characteristics. Second, all readers are trained using Smarter Balanced-provided materials, which were approved as unbiased examples of responses at the various score points. Training involves constant comparisons with the rubric and anchor papers so that readers' judgments are based solely on the scoring criteria. Finally, following training, a cycle of diagnosis and feedback is used to identify any issues. Specifically, during scoring, readers are monitored and any instances of readers making scoring decisions based on anything but the criteria are discussed. Readers are further monitored, and if any continue to exhibit bias after receiving a reasonable amount of feedback they are dismissed.

6.7.3 Reader Statistics

One concern regarding the scoring of any open-response assessment is the reliability and accuracy of the scoring. MI appreciates and shares this concern and continually develops new and technically sound methods of monitoring reliability. Reliable scoring starts with detailed scoring rubrics and training materials, and thorough training sessions by experienced trainers. Quality results are achieved by daily monitoring of each reader.

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

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

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

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

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

6.7.4 Reader Monitoring and Retraining

Team leaders spot-check (i.e., read behind) each reader's scoring to ensure that he or she is on target, and conduct one-on-one retraining sessions 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 criteria developed and approved by Smarter Balanced, with a second read conducted on 15% of responses for each item for reliability purposes. Responses are selected randomly for second reading and scored by readers who are not aware of the score assigned by the first reader or even that the response has been read before. MI's QA/reliability procedures allow the hand-scoring staff to identify struggling readers very early and begin retraining at once. While retraining these readers, MI also monitors their scoring intensively to ensure that all responses are scored accurately. In fact, MI's monitoring is also used as a retraining method. MI shows readers responses that the readers have scored incorrectly, explains the correct scores, and has the readers change the scores. Between the 2014–2015 and 2015–2016 test administrations, MI developed dynamic "threshold" reports which, based on inputted criteria, immediately identify potential scoring performance issues. This enhancement allows scoring leadership to pinpoint areas of concern and take corrective action with greater efficiency than ever before.

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

6.7.5 Reader Validity Checks

Approved responses are loaded into the VSC system as validity responses. A small set of validity responses are provided by Smarter Balanced for all vendors as the common benchmark to evaluate scoring accuracy and achieve consistency in scoring across states or vendors. The "true" scores for these responses are entered into a validity database. These responses are imbedded into live scoring on an ongoing basis to be scored by the readers. A validity report is generated that includes the response identification number, the score(s) assigned by the readers, and the "true" scores. A daily and project-to-date summary of percentages of correct scores and low/high considerations at each score point is also provided. If it is determined that a validity response and/or item is performing poorly, scoring management reviews the validity responses to ensure that the true scores have been entered correctly. If so, then retraining may be conducted with the readers using the validity data as a guide for how to focus the retraining. If the true scores have been entered incorrectly, then the database is updated to show the correct true scores. Validity results are not used in isolation but as one piece of evidence along with the second read and read-behind agreement to make decisions about retraining and dismissing readers.

6.7.6 Reader Dismissal

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

6.7.7 Reader Agreements

The inter-reader reliability is computed based on scorable responses (numeric scores) scored by two independent readers only, excluding non-scorable responses (e.g., off topic, off purpose, or foreign language responses) which are scored by scoring leadership, not by two independent readers. The inter-reader reliability is based on the readers who scored student responses in Delaware.

In ELA/Lit, writing essay item response (full write) 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 handscored items range from 1–3.

Tables 40–42 provide a summary of the inter-reader reliability based on items with a sample size greater than 50. The inter-reader reliability is presented with %exact agreement, minimum and maximum %exact agreements, combined %exact and %adjacent agreement, and quadratic weighted Kappa (QWK).

Table 40. ELA/Lit Reader Agreements for Short-Answer Items

Grade	**# of Items**	**%Exact**			**%(Exact+ Adjacent)**	**QWK**
		Average	**Min**	**Max**		
3	36	77	64	89	100	0.72
4	49	76	55	93	100	0.72
5	48	75	58	91	100	0.72
6	37	75	62	94	100	0.70
7	50	74	58	94	100	0.71
8	48	74	54	99	99	0.71

Table 41. ELA/Lit Reader Agreements for Full Write Items

Grade	**Dimensions**	**# of Items**	**%Exact**			**%(Exact+ Adjacent)**	**QWK**
			Average	**Min**	**Max**		
3	Conventions	14	71	63	83	99	0.60
	Evid/Elab	14	64	54	72	98	0.67
	Org/Purp	14	63	57	71	99	0.67
4	Conventions	18	66	56	75	99	0.66
	Evid/Elab	18	66	55	82	98	0.69
	Org/Purp	18	66	55	80	99	0.69
5	Conventions	20	69	54	86	100	0.53
	Evid/Elab	20	60	45	72	98	0.68
	Org/Purp	20	61	44	72	98	0.67
6	Conventions	14	73	66	82	98	0.62
	Evid/Elab	14	63	51	75	99	0.69
	Org/Purp	14	64	53	76	99	0.69
7	Conventions	19	69	61	78	99	0.61
	Evid/Elab	19	73	58	84	99	0.75
	Org/Purp	19	73	59	83	99	0.75
8	Conventions	20	77	63	87	99	0.60
	Evid/Elab	20	69	58	78	100	0.72
	Org/Purp	20	69	58	78	100	0.72

Legend:
Evid/Elab = Evidence/Elaboration, and Org/Purp = Organization/Purpose

Table 42. Mathematics Reader Agreements

Grade	Score Points	# of Items	%Exact			%(Exact+ Adjacent)	QWK
			Average	Min	Max		
3	1	12	92	87	95	100	0.81
3	2	26	88	76	99	100	0.90
3	3	4	95	92	97	99	0.96
4	1	8	86	81	94	100	0.66
4	2	36	92	72	100	100	0.91
4	3	4	87	83	90	100	0.93
5	1	4	93	89	99	100	0.72
5	2	41	89	76	98	100	0.86
5	3	8	85	76	99	97	0.84
6	1	13	96	84	99	100	0.89
6	2	32	88	77	98	100	0.88
7	1	8	97	96	99	100	0.85
7	2	26	88	75	98	100	0.82
8	1	14	89	79	98	100	0.76
8	2	26	86	73	99	100	0.85

7. REPORTING AND INTERPRETING SCORES

The Online Reporting System (ORS) generates a set of online score reports that include the information describing student performance for students, parents, educators, and other stakeholders. The online score reports are produced immediately after students complete the test with handscored items. Because the score report on student performance are updated each time that students complete tests and these tests 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 the individual student score report, the ORS also produces aggregate score reports by class, school, district, and the state. The timely accessibility of aggregate score reports could help users monitor student testing in each subject by grade, 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. In 2016–2017, some new features are added to ORS reports.

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.

7.1 ONLINE REPORTING SYSTEM FOR STUDENTS AND EDUCATORS

7.1.1 Types of Online Score Reports

The ORS is designed to help educators and students answer questions regarding how well students have performed on ELA/Lit and mathematics assessments. The ORS is the online tool that provides educators and other stakeholders with timely, relevant score reports. The ORS for the Smarter Balanced assessment has been designed with stakeholders 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 inferences about student achievement. The ORS is also designed to present student performance in a uniform format. For example, similar colors are used for groups of similar elements, such as achievement levels, throughout the design. This design strategy allows readers to compare similar elements and to avoid comparing dissimilar elements.

Once authorized users log in to the ORS and select "Score Reports," the online score reports are presented hierarchically. The ORS starts with presenting summaries on student performance by subject and grade at a selected aggregate level. 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 43 summarizes the types of online score reports available at the aggregate level and the individual student level. Detailed information about the online score reports and instructions on how to navigate the online score reporting system can be found in the *Online Reporting System User Guide*, located in a help button on the ORS.

Table 43. Types of Online Score Reports by Level of Aggregation

Level of Aggregation	**Types of Online Score Reports**
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 school, district, and state • Student growth in scale score and achievement level over time • Writing performance descriptors and scores by dimensions

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 the subgroups. Table 44 presents the types of subgroups and subgroup category provided in ORS.

Table 44. 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 Native Hawaiian/Pacific Islander

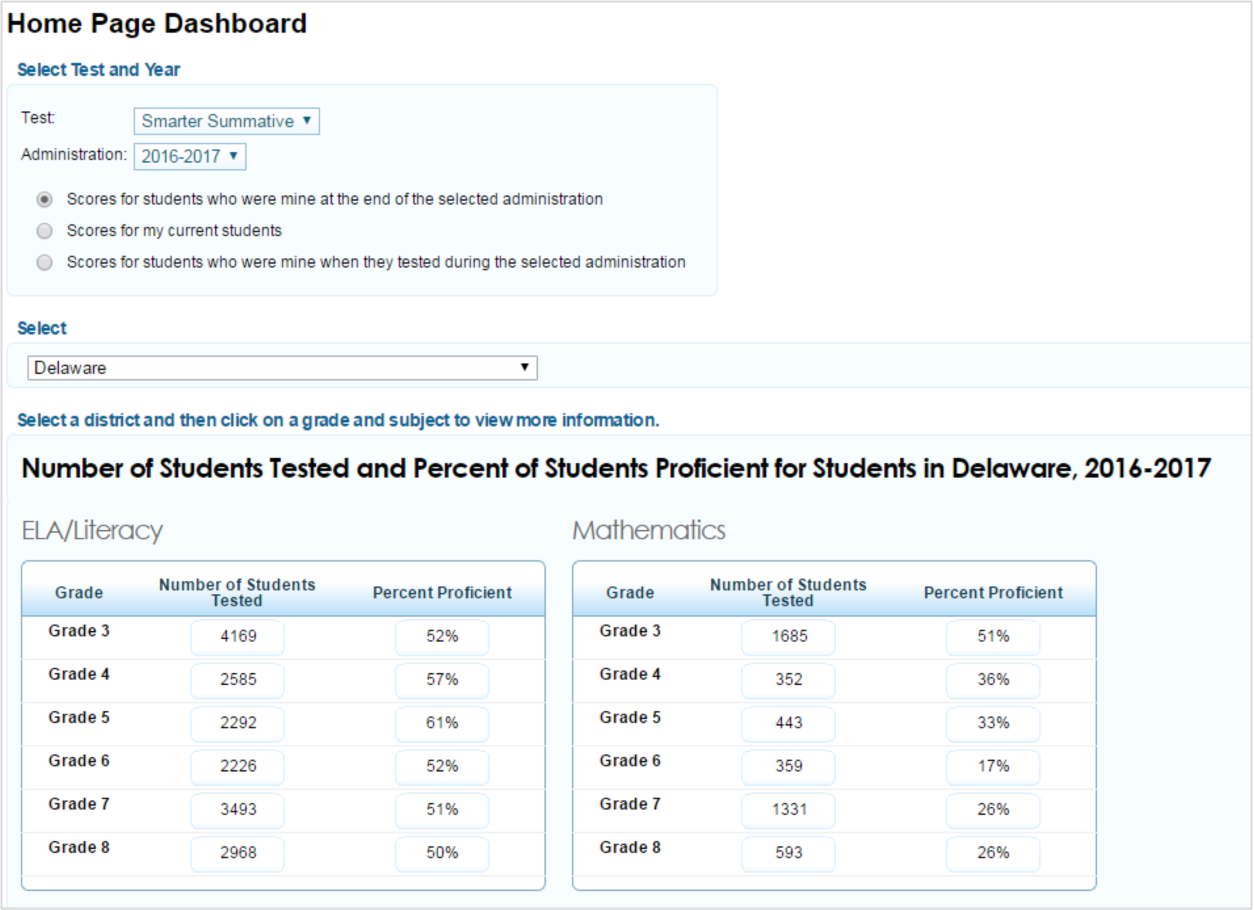
7.1.2 Online Reporting System

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 summarizes students’ performance including (1) number of students tested, and (2) percentage of students at Level 3 or above. Exhibits 1 and 2 present a sample of home pages at the state level and the district level, respectively.

Exhibit 1. Home Page: State Level



Home Page Dashboard

Select Test and Year

Test: Smarter Summative ▼

Administration: 2016-2017 ▼

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

○ Scores for my current students

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

Select

Delaware ▼

Select a district and then click on a grade and subject to view more information.

Number of Students Tested and Percent of Students Proficient for Students in Delaware, 2016-2017

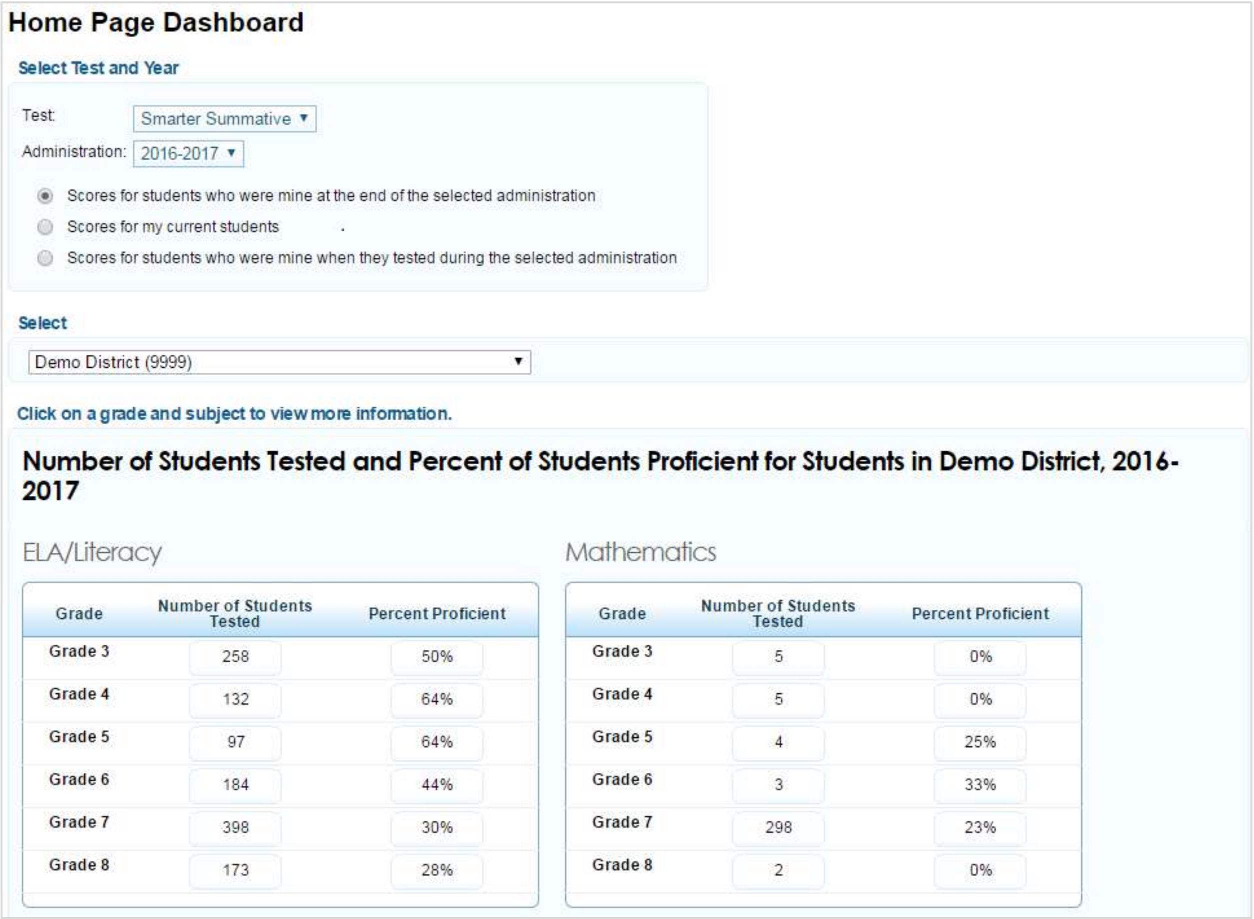
ELA/Literacy

Grade	Number of Students Tested	Percent Proficient
Grade 3	4169	52%
Grade 4	2585	57%
Grade 5	2292	61%
Grade 6	2226	52%
Grade 7	3493	51%
Grade 8	2968	50%

Mathematics

Grade	Number of Students Tested	Percent Proficient
Grade 3	1685	51%
Grade 4	352	36%
Grade 5	443	33%
Grade 6	359	17%
Grade 7	1331	26%
Grade 8	593	26%

Exhibit 2. Home Page: District Level



Home Page Dashboard

Select Test and Year

Test: Smarter Summative ▾

Administration: 2016-2017 ▾

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

○ Scores for my current students

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

Select

Demo District (9999) ▾

Click on a grade and subject to view more information.

Number of Students Tested and Percent of Students Proficient for Students in Demo District, 2016-2017

ELA/Literacy

Grade	Number of Students Tested	Percent Proficient
Grade 3	258	50%
Grade 4	132	64%
Grade 5	97	64%
Grade 6	184	44%
Grade 7	398	30%
Grade 8	173	28%

Mathematics

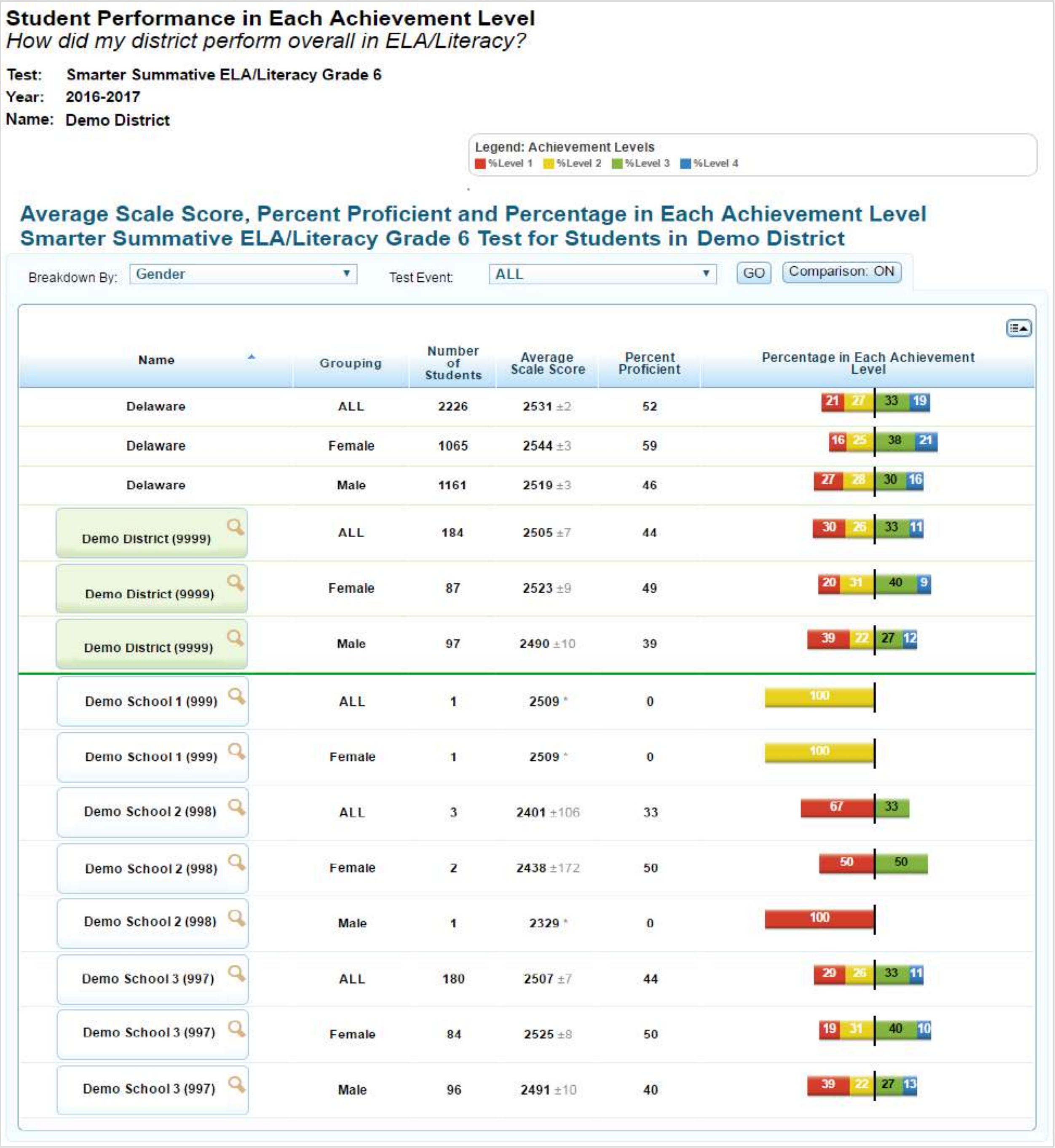
Grade	Number of Students Tested	Percent Proficient
Grade 3	5	0%
Grade 4	5	0%
Grade 5	4	25%
Grade 6	3	33%
Grade 7	298	23%
Grade 8	2	0%

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: 2016-2017
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

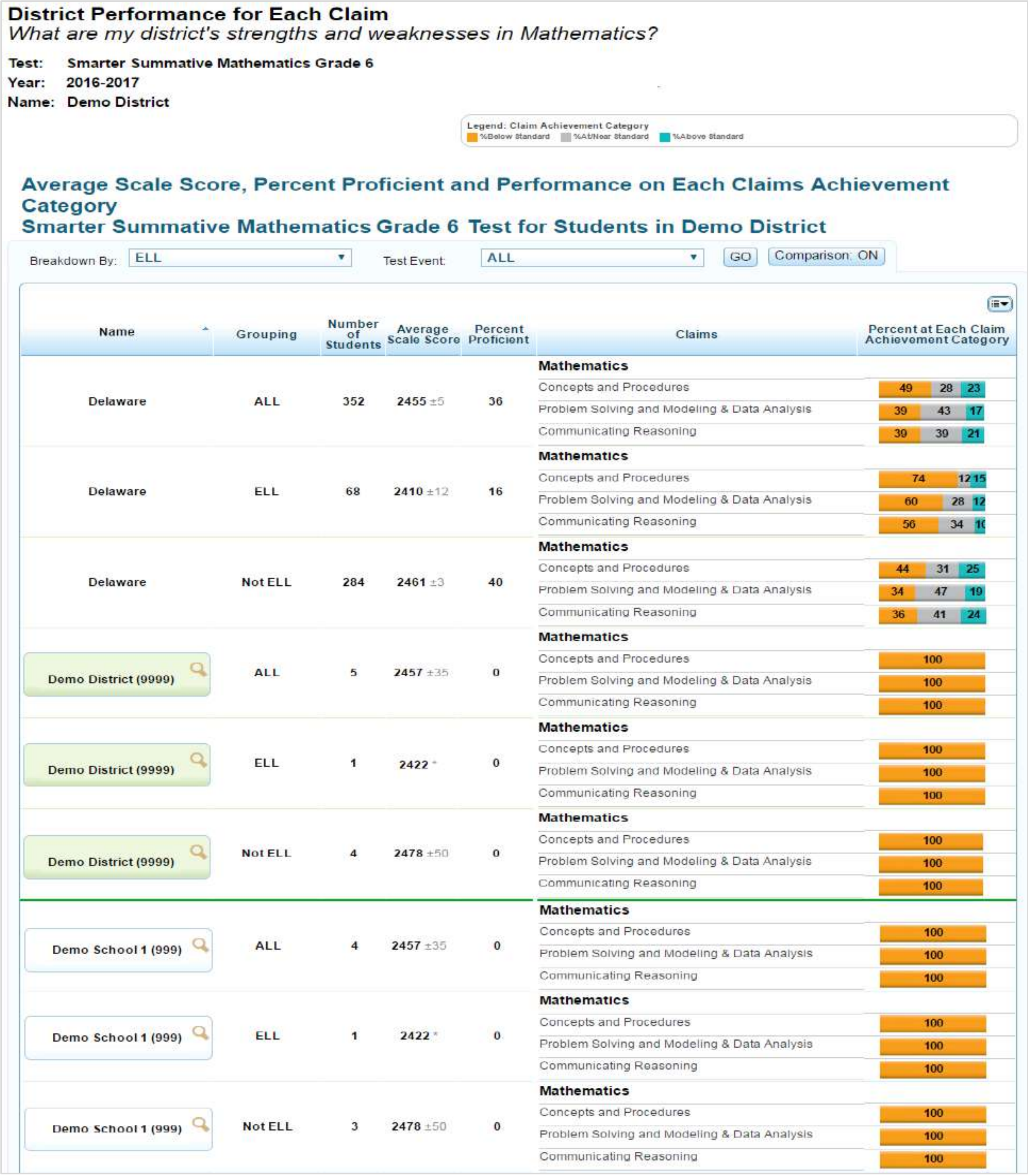
Name	Grouping	Number of Students	Average Scale Score	Percent Proficient	Percentage in Each Achievement Level
Delaware	ALL	2226	2531 ±2	52	21 27 33 19
Delaware	Female	1065	2544 ±3	59	16 25 38 21
Delaware	Male	1161	2519 ±3	46	27 28 30 16
Demo District (9999)	ALL	184	2505 ±7	44	30 26 33 11
Demo District (9999)	Female	87	2523 ±9	49	20 31 40 9
Demo District (9999)	Male	97	2490 ±10	39	39 22 27 12
Demo School 1 (999)	ALL	1	2509 *	0	100
Demo School 1 (999)	Female	1	2509 *	0	100
Demo School 2 (998)	ALL	3	2401 ±106	33	67 33
Demo School 2 (998)	Female	2	2438 ±172	50	50 50
Demo School 2 (998)	Male	1	2329 *	0	100
Demo School 3 (997)	ALL	180	2507 ±7	44	29 26 33 11
Demo School 3 (997)	Female	84	2525 ±8	50	19 31 40 10
Demo School 3 (997)	Male	96	2491 ±10	40	39 22 27 13

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 claim performance category.

Similar to the subject detail page, the summary report presents the summary results for the selected aggregate unit as well as the summary results for 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 a claim detail pages for mathematics at a 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: 2016-2017
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 ▾ Test Event: ALL ▾ GO Comparison: ON

Name	Grouping	Number of Students	Average Scale Score	Percent Proficient	Claims	Percent at Each Claim Achievement Category
Delaware	ALL	352	2455 ±5	36	**Mathematics**	
					Concepts and Procedures	49 28 23
					Problem Solving and Modeling & Data Analysis	39 43 17
					Communicating Reasoning	39 39 21
Delaware	ELL	68	2410 ±12	16	**Mathematics**	
					Concepts and Procedures	74 12 15
					Problem Solving and Modeling & Data Analysis	60 28 12
					Communicating Reasoning	56 34 10
Delaware	Not ELL	284	2461 ±3	40	**Mathematics**	
					Concepts and Procedures	44 31 25
					Problem Solving and Modeling & Data Analysis	34 47 19
					Communicating Reasoning	36 41 24
Demo District (9999)	ALL	5	2457 ±35	0	**Mathematics**	
					Concepts and Procedures	100
					Problem Solving and Modeling & Data Analysis	100
					Communicating Reasoning	100
Demo District (9999)	ELL	1	2422 *	0	**Mathematics**	
					Concepts and Procedures	100
					Problem Solving and Modeling & Data Analysis	100
					Communicating Reasoning	100
Demo District (9999)	Not ELL	4	2478 ±50	0	**Mathematics**	
					Concepts and Procedures	100
					Problem Solving and Modeling & Data Analysis	100
					Communicating Reasoning	100
Demo School 1 (999)	ALL	4	2457 ±35	0	**Mathematics**	
					Concepts and Procedures	100
					Problem Solving and Modeling & Data Analysis	100
					Communicating Reasoning	100
Demo School 1 (999)	ELL	1	2422 *	0	**Mathematics**	
					Concepts and Procedures	100
					Problem Solving and Modeling & Data Analysis	100
					Communicating Reasoning	100
Demo School 1 (999)	Not ELL	3	2478 ±50	0	**Mathematics**	
					Concepts and Procedures	100
					Problem Solving and Modeling & Data Analysis	100
					Communicating Reasoning	100

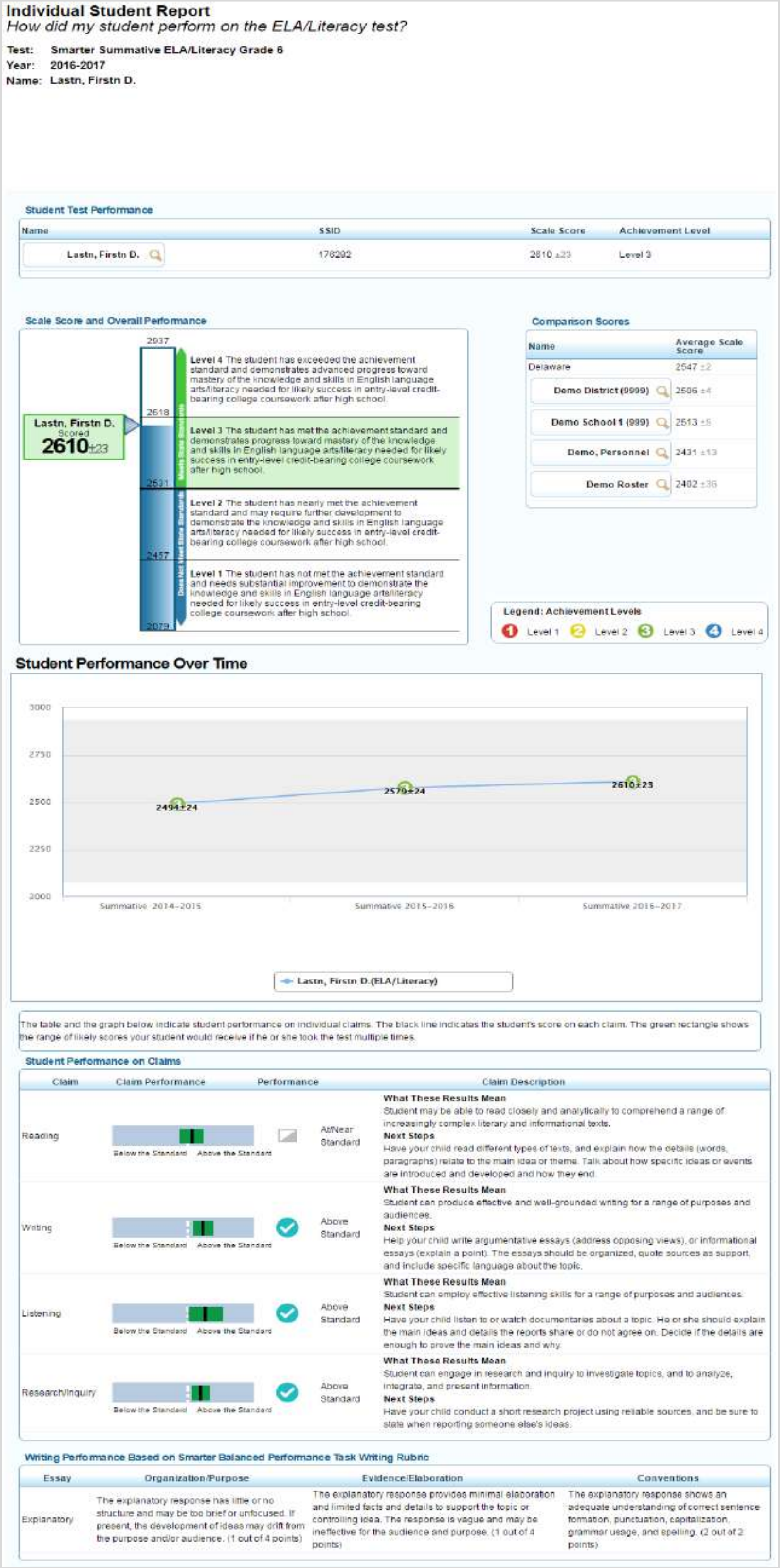
7.1.2.4 Student Detail Page

When a student completes a test and the test is handscored, an online score report appears in the student detail page in the ORS. The student detail page 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) performance category in each claim, (4) average scale scores for student's state, district, school, teacher, and associated standard errors of the average scale scores, and (5) student performance growth over time.

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 is 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. 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 define 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. Under the barrel chart, the trend of student performance over time is displayed. On the bottom of the page, student performance on each reporting category and writing dimension scores (ELA/Lit only) is displayed along with a description of his or her performance on each claim and each writing dimension.

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 6
Year: 2016-2017
Name: Lastn, Firstn D.

Student Test Performance

Name	SSID	Scale Score	Achievement Level
Lastn, Firstn D.	176282	2610 ±23	Level 3

Scale Score and Overall Performance

Lastn, Firstn D.
Scored
2610±23

Level 4 The student has exceeded the achievement standard and demonstrates advanced 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 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 high school.

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.

Comparison Scores

Name	Average Scale Score
Delaware	2547 ±2
Demo District (9999)	2506 ±4
Demo School 1 (999)	2513 ±5
Demo, Personnel	2431 ±13
Demo Roster	2402 ±36

Legend: Achievement Levels
❶ Level 1 ❷ Level 2 ❸ Level 3 ❹ Level 4

Student Performance Over Time

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

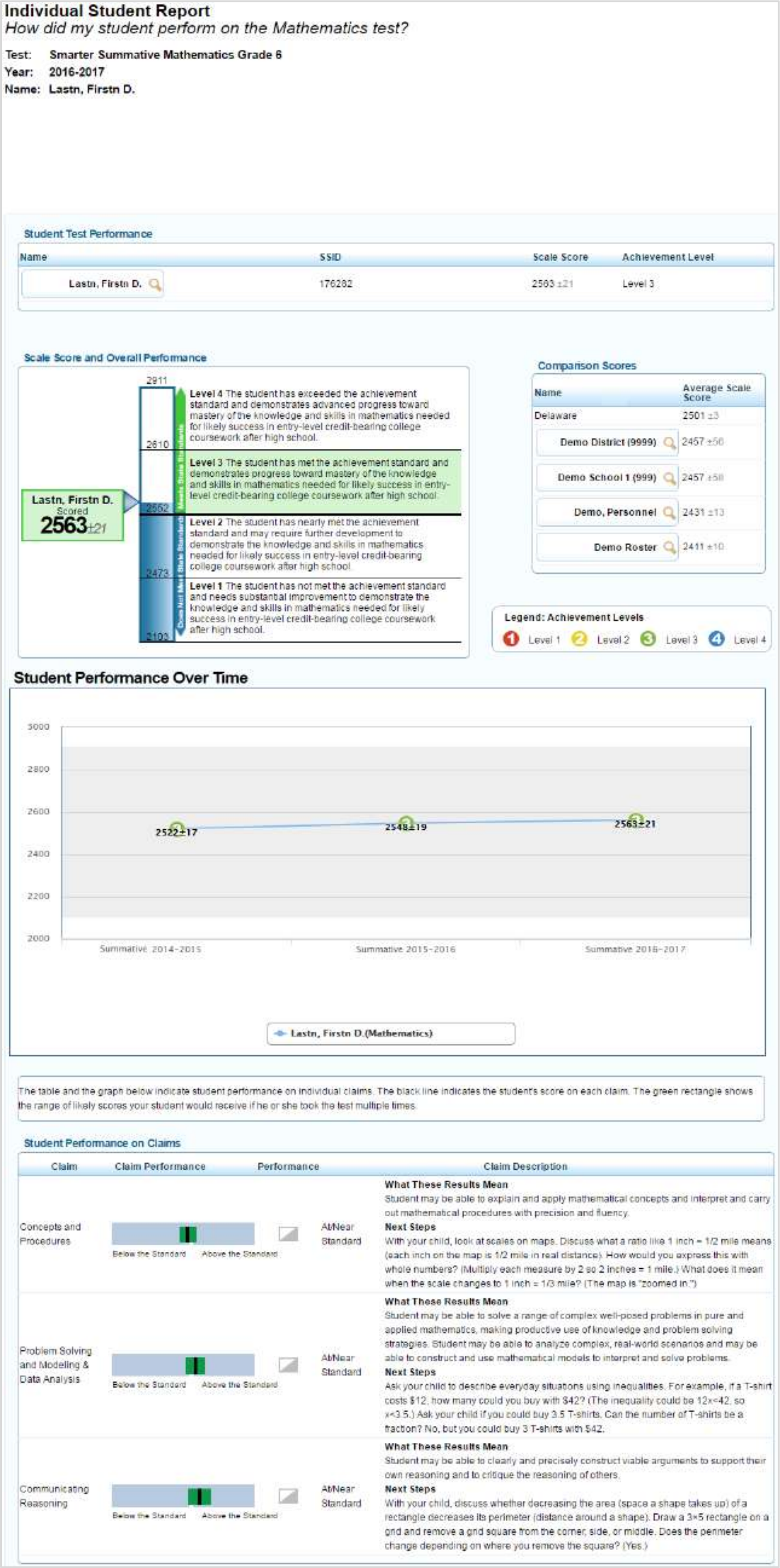
Student Performance on Claims

Claim	Claim Performance	Performance		Claim Description
Reading	Below the Standard Above the Standard	▨	At/Near Standard	**What These Results Mean** Student may be able to read closely and analytically to comprehend a range of increasingly complex literary and informational texts. **Next Steps** Have your child read different types of texts, and explain how the details (words, paragraphs) relate to the main idea or theme. Talk about how specific ideas or events are introduced and developed and how they end.
Writing	Below the Standard Above the Standard	✓	Above Standard	**What These Results Mean** Student can produce effective and well-grounded writing for a range of purposes and audiences. **Next Steps** Help your child write argumentative essays (address opposing views), or informational essays (explain a point). The essays should be organized, quote sources as support, and include specific language about the topic.
Listening	Below the Standard Above the Standard	✓	Above Standard	**What These Results Mean** Student can employ effective listening skills for a range of purposes and audiences. **Next Steps** Have your child listen to or watch documentaries about a topic. He or she should explain the main ideas and details the reports share or do not agree on. Decide if the details are enough to prove the main ideas and why.
Research/Inquiry	Below the Standard Above the Standard	✓	Above Standard	**What These Results Mean** Student can engage in research and inquiry to investigate topics, and to analyze, integrate, and present information. **Next Steps** Have your child conduct a short research project using reliable sources, and be sure to state when reporting someone else's ideas.

Writing Performance Based on Smarter Balanced Performance Task Writing Rubric

Essay	Organization/Purpose	Evidence/Elaboration	Conventions
Explanatory	The explanatory response has little or no structure and may be too brief or unfocused. If present, the development of ideas may drift from the purpose and/or audience. (1 out of 4 points)	The explanatory response provides minimal elaboration and limited facts and details to support the topic or controlling idea. The response is vague and may be ineffective for the audience and purpose. (1 out of 4 points)	The explanatory response shows an adequate understanding of correct sentence formation, punctuation, capitalization, grammar usage, and spelling. (2 out of 2 points)

Exhibit 6. Student Detail Page for Mathematics



Individual Student Report

How did my student perform on the Mathematics test?

Test: Smarter Summative Mathematics Grade 6
Year: 2016-2017
Name: Lastn, Firstn D.

Student Test Performance

Name	SSID	Scale Score	Achievement Level
Lastn, Firstn D.	176282	2563 ±21	Level 3

Scale Score and Overall Performance

Lastn, Firstn D. Scored 2563±21

Level 4 The student has exceeded the achievement standard and demonstrates advanced progress toward mastery of the knowledge and skills in mathematics 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 mathematics needed for likely success in entry-level credit-bearing college coursework after high school.

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

Comparison Scores

Name	Average Scale Score
Delaware	2501 ±3
Demo District (9999)	2457 ±50
Demo School 1 (999)	2457 ±50
Demo, Personnel	2431 ±13
Demo Roster	2411 ±10

Legend: Achievement Levels
❶ Level 1 ❷ Level 2 ❸ Level 3 ❹ Level 4

Student Performance Over Time

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

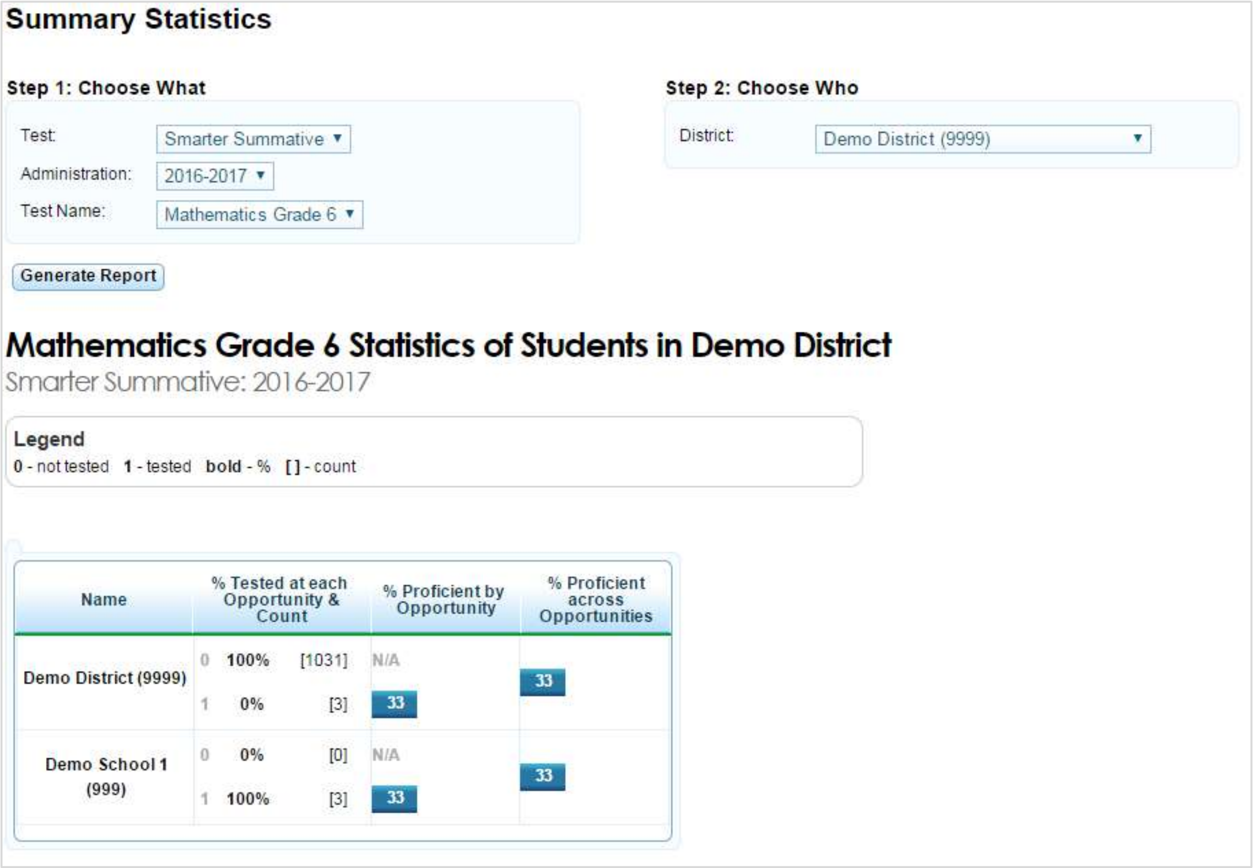
Student Performance on Claims

Claim	Claim Performance	Performance	Claim Description
Concepts and Procedures	Below the Standard Above the Standard	At/Near Standard	**What These Results Mean** Student may be able to explain and apply mathematical concepts and interpret and carry out mathematical procedures with precision and fluency. **Next Steps** With your child, look at scales on maps. Discuss what a ratio like 1 inch = 1/2 mile means (each inch on the map is 1/2 mile in real distance). How would you express this with whole numbers? (Multiply each measure by 2 so 2 inches = 1 mile.) What does it mean when the scale changes to 1 inch = 1/3 mile? (The map is "zoomed in.")
Problem Solving and Modeling & Data Analysis	Below the Standard Above the Standard	At/Near Standard	**What These Results Mean** Student may be able to solve a range of complex well-posed problems in pure and applied mathematics, making productive use of knowledge and problem solving strategies. Student may be able to analyze complex, real-world scenarios and may be able to construct and use mathematical models to interpret and solve problems. **Next Steps** Ask your child to describe everyday situations using inequalities. For example, if a T-shirt costs $12, how many could you buy with $42? (The inequality could be 12x<42, so x<3.5.) Ask your child if you could buy 3.5 T-shirts. Can the number of T-shirts be a fraction? No, but you could buy 3 T-shirts with $42.
Communicating Reasoning	Below the Standard Above the Standard	At/Near Standard	**What These Results Mean** Student may be able to clearly and precisely construct viable arguments to support their own reasoning and to critique the reasoning of others. **Next Steps** With your child, discuss whether decreasing the area (space a shape takes up) of a rectangle decreases its perimeter (distance around a shape). Draw a 3×5 rectangle on a grid and remove a grid square from the corner, side, or middle. Does the perimeter change depending on where you remove the square? (Yes.)

7.1.2.5 Participation Rate

In addition to online score reports, the ORS provides participation rate reports for districts and schools 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 proficient. Exhibit 7 presents a sampled participation rate report at the district level.

Exhibit 7. Participation Rate Report at District Level



Summary Statistics

Step 1: Choose What

Test: Smarter Summative ▾

Administration: 2016-2017 ▾

Test Name: Mathematics Grade 6 ▾

Step 2: Choose Who

District: Demo District (9999) ▾

Mathematics Grade 6 Statistics of Students in Demo District

Smarter Summative: 2016-2017

Legend

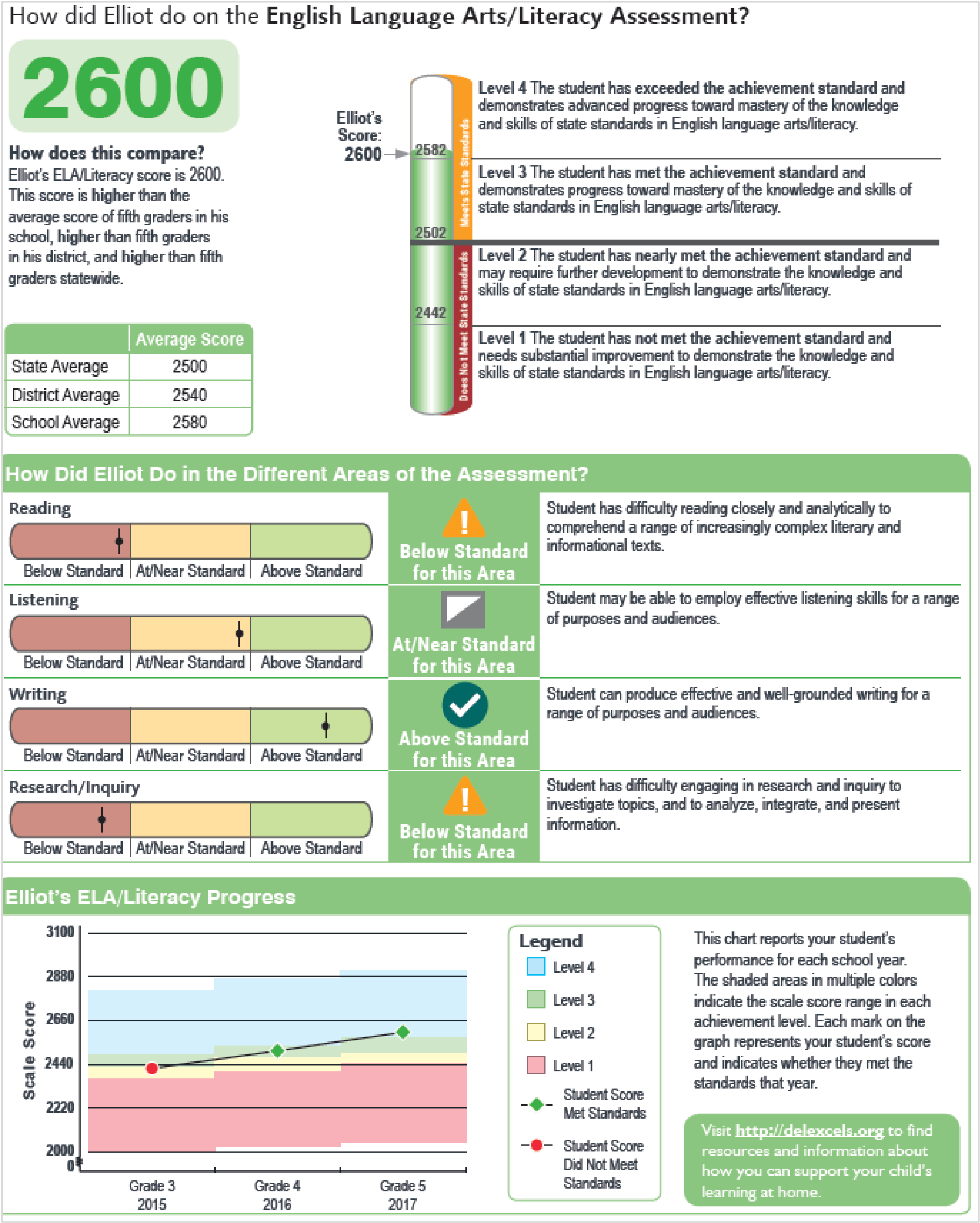
0 - not tested **1** - tested **bold** - % **[ ]** - count

Name	% Tested at each Opportunity & Count	% Proficient by Opportunity	% Proficient across Opportunities
Demo District (9999)	0 **100%** [1031]	N/A	33
	1 **0%** [3]	33	
Demo School 1 (999)	0 **0%** [0]	N/A	33
	1 **100%** [3]	33	

7.2 PAPER FAMILY SCORE REPORTS

After the testing window is closed, parents whose children participate in a test receive a full-color paper score report (hereinafter family report) that includes their children’s performance on ELA/Lit and mathematics. The family report include information on student performance that is provided on the student detailed page from the ORS with additional information on student performance. For example, the family report includes a progress chart that displays student’s performance for each school year. The progress chart shows whether student’s performance meet the standards in each year and how much student’s performance increases. Exhibits 8 and 9 present examples of paper family score reports for grade 5 ELA/Lit and mathematics.

Exhibit 8. Sample Paper Family Score Report for Grade 5 ELA/Lit



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

2600

How does this compare?

Elliot's ELA/Literacy score is 2600. This score is **higher** than the average score of fifth graders in his school, **higher** than fifth graders in his district, and **higher** than fifth graders statewide.

	Average Score
State Average	2500
District Average	2540
School Average	2580

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.

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

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

Elliot's ELA/Literacy Progress

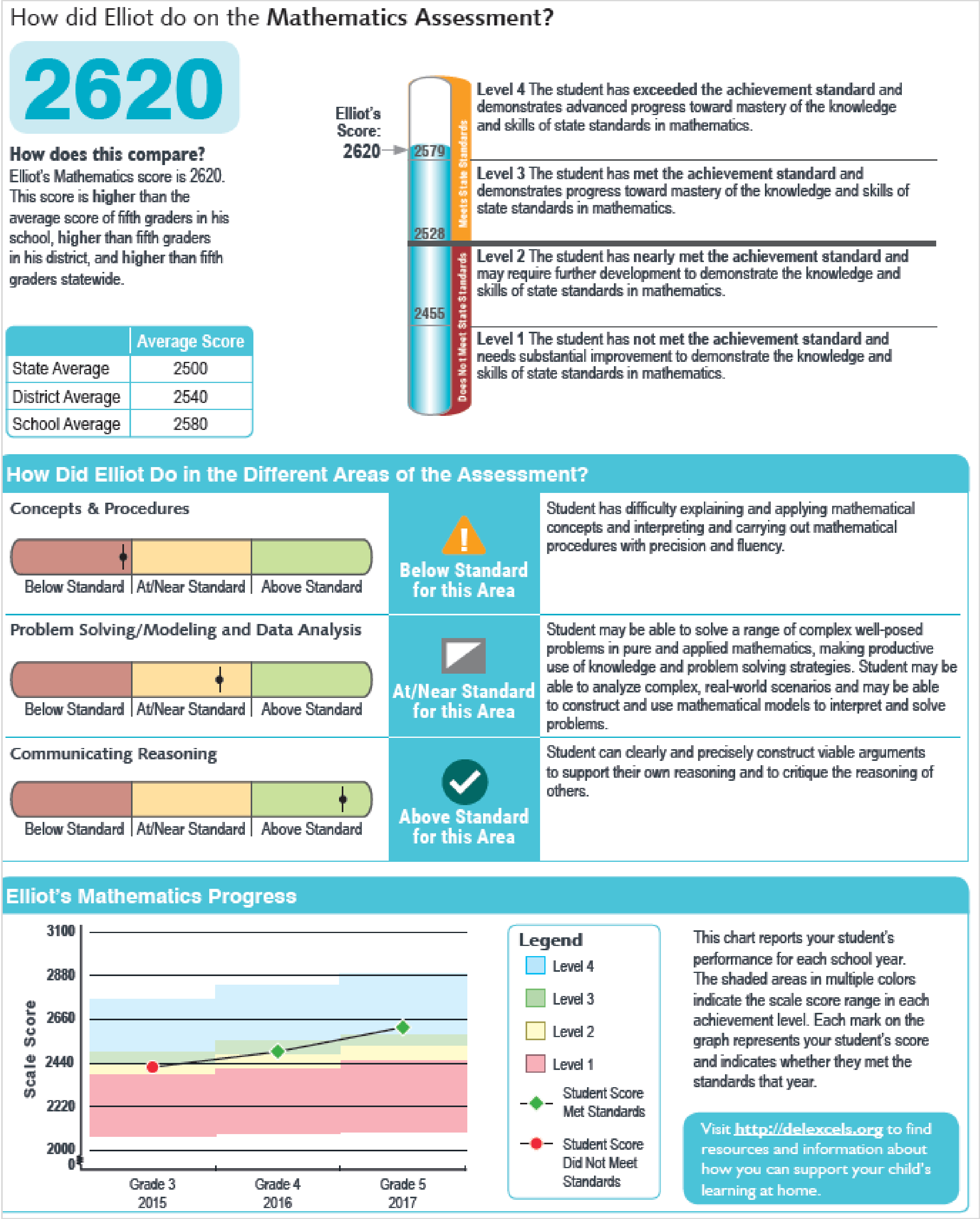
Legend

- 🟦 Level 4
- 🟩 Level 3
- 🟨 Level 2
- 🟥 Level 1
- -◆- Student Score Met Standards
- -●- Student Score Did Not Meet Standards

This chart reports your student's performance for each school year. The shaded areas in multiple colors indicate the scale score range in each achievement level. Each mark on the graph represents your student's score and indicates whether they met the standards that year.

Visit **http://delexcels.org** to find resources and information about how you can support your child's learning at home.

Exhibit 9. Sample Paper Family Score Report for Grade 5 Mathematics



How did Elliot do on the **Mathematics Assessment?**

2620

How does this compare?

Elliot's Mathematics score is 2620. This score is **higher** than the average score of fifth graders in his school, **higher** than fifth graders in his district, and **higher** than fifth graders statewide.

	Average Score
State Average	2500
District Average	2540
School Average	2580

Elliot's Score:
2620
2579
2528
2455
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 mathematics.

Level 3 The student has met the achievement standard and demonstrates progress toward mastery of the knowledge and skills of state standards in mathematics.

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 mathematics.

Level 1 The student has not met the achievement standard and needs substantial improvement to demonstrate the knowledge and skills of state standards in mathematics.

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

Concepts & Procedures

Below Standard for this Area

Student has difficulty explaining and applying mathematical concepts and interpreting and carrying out mathematical procedures with precision and fluency.

Problem Solving/Modeling and Data Analysis

At/Near Standard for this Area

Student may be able to solve a range of complex well-posed problems in pure and applied mathematics, making productive use of knowledge and problem solving strategies. Student may be able to analyze complex, real-world scenarios and may be able to construct and use mathematical models to interpret and solve problems.

Communicating Reasoning

Above Standard for this Area

Student can clearly and precisely construct viable arguments to support their own reasoning and to critique the reasoning of others.

Elliot's Mathematics Progress

Legend

- 🟦 Level 4
- 🟩 Level 3
- 🟨 Level 2
- 🟥 Level 1
- -◆- Student Score Met Standards
- -●- Student Score Did Not Meet Standards

This chart reports your student's performance for each school year. The shaded areas in multiple colors indicate the scale score range in each achievement level. Each mark on the graph represents your student's score and indicates whether they met the standards that year.

Visit http://delexcels.org to find resources and information about how you can support your child's learning at home.

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 indicate that the student does not possess sufficient knowledge and skills measured by the test. Conversely, high scale scores indicate 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, and 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 in grade 6 ELA/Lit, for instance, achievement-level descriptors are described for Level 3 as "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 high school." 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 Performance Category for Claims

Students' performance on each claim is reported in three categories: (1) *Below Standard*, (2) *At/Near Standard*, and (3) *Above Standard*. Unlike the achievement level for overall test, student performance on each of claims is evaluated with respect to the "Meets Standard achievement" standard. For students performing at either "Below Standard" or "Above Standard," this can be interpreted to mean that students' performance is clearly below or above the "Meets Standard" cut score for a specific claim. For students performing at "At/Near Standard," this can be interpreted to mean that students' performance does not provide enough information to tell whether students 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 students' scale scores are aggregated, the aggregated scale scores can be interpreted as an estimate of knowledge and skills that a group of students possess. Given that student scale scores are estimates, the aggregated scale scores are also estimates and are subject to measures of uncertainty. In addition to the aggregated scale scores, the percent of students in each achievement level 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.

7.3.6 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, performance categories for claims can be used to identify an individual student's relative strengths and weaknesses among claims within a content area.

Assessment results on student achievement on the test can be used to help teachers or schools make decisions on how to support students' learning. Aggregate score reports at the teacher and school level provide information regarding the strengths and weaknesses of their students and can be 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 (QA) 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 formats. The quality of the information produced in the Test Delivery System (TDS) is tested thoroughly before, during, and after the testing window opens.

8.1 ADAPTIVE TEST CONFIGURATION

For the CAT, a test configuration file is the key file that contains all specifications for the item selection algorithm and the scoring algorithm, such as the test blueprint specification, slopes and intercepts for theta-to-scale score transformation, cut scores, and the item information (i.e., answer keys, item attributes, item parameters, and passage information). The accuracy of the information in the configuration file is 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 wide range of student response patterns. The results of simulated test administrations are used to configure and evaluate the adequacy of the item selection algorithm used to administer the Smarter Balanced summative assessments. The purpose of the simulations is to configure the adaptive algorithm to optimize item selection to meet blueprint specifications while targeting test information to student ability as well as checking the score accuracy.

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

8.1.1 Platform Review

AIR’s TDS supports a variety of item layouts. Each item goes through an extensive platform review on different operating systems 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

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

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

8.4.3 Monitoring by State Department of Education

The DDOE also directly observes MI activities, virtually. MI provides virtual access to the training activities through the online training interface. The 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 TDS during the test administration window, AIR statisticians examine the delivery demands, including the number of tests to be delivered, the length of the window, and the historic state-specific behaviors to model the likely peak loads. Using data from the load tests, these calculations indicate the number of each type of server necessary to provide continuous, responsive service, and AIR contracts for service in excess of this amount. Once deployed, our servers are monitored at the hardware, operating system, and software platform levels with monitoring software that alerts our engineers at the first signs that trouble may be ahead. The applications log not only errors and exceptions, but also 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 testing window. The item statistics analysis report is used to monitor the performance of test items throughout the testing window and serves as a key check for the early detection of potential problems with item scoring, including incorrect designation of a keyed response or other scoring errors, as well as potential breaches of test security that may be indicated by changes in the difficulty of test items. This report generates classical item analysis indicators of difficulty and discrimination, including proportion correct and biserial/polyserial correlation. The report is configurable and can be produced so that only items with statistics falling outside 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 QA reports can be generated on any desired schedule. Item analysis and blueprint match reports are evaluated frequently at the opening of the testing window to ensure that test administrations conform to blueprint and items are performing as anticipated.

Table 45 presents an overview of the QA reports.

Table 45. Overview of Quality Assurance Reports

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

8.5.1 Score Report Quality Check

In the 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 handscoring processes include rigorous training, validity and reliability monitoring, and back-reading to ensure accurate scoring. Handscored 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 handscores being received by AIR and passing QA validation checks, the composite score will be available in the ORS.

8.5.1.2 Paper Report Quality Assurance

Statistical Programming

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

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

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

Display Programming

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

REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

APPENDICES

Appendix A: 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 A–1 presents the number of students who took the ICA.

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

Grade	**ELA/Lit**			**Mathematics**		
	Once	**Twice**	**Total**	**Once**	**Twice**	**Total**
3	378	0	378	414	0	414
4	285	2	287	288	0	288
5	253	0	253	257	0	257
6	180	1	181	334	0	334
7	200	0	200	226	0	226
8	170	0	170	198	0	198

For the Interim Assessment Blocks (IAB), there were seven to nine IABs for ELA/Lit and five to six IABs in mathematics. Students were allowed to take as many IABs as they wanted. Table A–2 presents the total number of students who took the IABs and the number of students by the number of IABs taken. For example, in grade 3 ELA/Lit, a total of 1,379 students took IABs, and among these students, 795 students took one IAB, 286 students took two IABs, and so on.

Tables A–3 and A–4 disaggregated the number of students in Table A-2 by each individual block. For example, 795 students in grade 3 ELA/Lit took one IAB only. Among these students, 42 students took the Brief Writes IAB.

Table A-2. Number of Students Who Took IABs

Grade	**Total**	**Number of IABs Taken**								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
					ELA/Lit					
3	1,379	795	286	145	96	44	11	2		
4	1,234	498	375	181	85	42	40	13		
5	1,953	901	364	355	191	64	40	35		3
6	1,858	573	313	359	143	312	53	28	19	58
7	974	276	470	223	3	1		1		
8	703	424	168	107	4					
					Mathematics					
3	2,069	938	637	355	139					
4	1,747	625	421	285	125	291				
5	2,792	1,167	654	544	224	200	3			
6	2,576	918	867	473	205	111	2			
7	2,081	870	859	187	41	122	2			
8	1,570	690	193	193	451	43				

Table A-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	8	9
3	Brief Writes	42								
	Editing	77	132	82	85	43	11	2		
	Language and Vocabulary Use	331	108	85	82	41	11	2		
	Listening and Interpretation	42	167	85	69	44	11	2		
	Performance Task	18	3	1	6	1	1			
	Reading Informational Text	105	33	27	16	6	5	2		
	Reading Literary Text	102	51	40	12	9	10	2		
	Research	60	38	68	73	43	11	2		
	Revision	18	40	47	41	33	6	2		
4	Brief Writes	3	49							
	Editing	68	65	110	73	40	40	13		
	Language and Vocabulary Use	135	209	98	64	37	39	13		
	Listening and Interpretation	25	37	55	64	39	40	13		
	Performance Task		1	2						
	Reading Informational Text	145	141	68	51	7	31	13		
	Reading Literary Text	90	71	62	22	10	11	13		
	Research	28	151	70	44	40	39	13		
	Revision	4	26	78	22	37	40	13		
5	Brief Writes	1	6	93	7	6	16	35		3
	Editing	86	118	184	181	59	39	35		3
	Language and Vocabulary Use	487	169	138	43	26	40	35		3
	Listening and Interpretation	79	73	169	156	56	23			3
	Performance Task	1	6	91			1			3
	Reading Informational Text	139	71	156	152	59	39	35		3
	Reading Literary Text	18	55	62	51	47	35	35		3
	Research	73	152	101	33	9	15	35		3
	Revision	17	78	71	141	58	32	35		3
6	Brief Writes	4	9	42				6	13	58
	Editing	59	172	224	98	294	23	23	13	58
	Language and Vocabulary Use	182	134	191	125	309	53	26	19	58
	Listening and Interpretation	6	6	132	73	194	51	26	15	58
	Performance Task		4	42		1	1	11	18	58
	Reading Informational Text	229	95	76	13	132	45	28	19	58
	Reading Literary Text	34	115	112	6	18	43	25	19	58
	Research	15	48	27	119	305	50	26	18	58
	Revision	44	43	231	138	307	52	25	18	58
7	Brief Writes	1	6	40						
	Editing	32	45	113	3			1		
	Language and Vocabulary Use	2	92	172	3			1		
	Listening and Interpretation	136	129	59	1	1		1		
	Performance Task	3	4	40	2					
	Reading Informational Text	77	328	69		1		1		
	Reading Literary Text	25	324	52		1		1		
	Research					1		1		
	Revision		12	124	3	1		1		
8	Brief Writes	2	4	38						
	Editing and Revising									
	Listening and Interpretation	130	103	32						
	Performance Task	3	20	44	4					
	Reading Informational Text	37	62	104	4					
	Reading Literary Text	240	58	40	4					
	Research	12	89	63	4					

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

Grade	**Block**	**Number of IABs Taken**					
		1	**2**	**3**	**4**	**5**	**6**
3	Measurement and Data	84	379	280	139		
	Number and Operations in Base Ten	299	265	234	138		
	Number and Operations – Fractions	244	386	312	138		
	Operational and Algebraic Thinking	305	243	238	137		
	Performance Task	6	1	1	4		
4	Measurement and Data	3	88	173	81	291	
	Number and Operations in Base Ten	285	379	262	122	291	
	Number and Operations – Fractions	303	77	220	116	291	
	Operational and Algebraic Thinking	29	277	143	105	291	
	Geometry	3	21	55	75	290	
	Performance Task	2		2	1	1	
5	Measurement and Data	24	85	42	113	200	3
	Number and Operations in Base Ten	454	327	503	222	200	3
	Number and Operations – Fractions	475	514	472	198	200	3
	Geometry	178	184	254	177	200	3
	Operations and Algebraic Thinking	32	183	361	182	200	3
	Performance Task	4	15		4		3
6	Expressions and Equations	130	240	441	205	111	2
	Geometry	13	170	72	92	111	2
	Number System	165	672	390	145	104	2
	Statistics and Probability	2	9	77	170	111	2
	Performance Task	3	42	18	6	7	2
	Ratios and Proportional Relationships	605	601	421	202	111	2
7	Expressions and Equations	258	322	90	34	122	2
	Number System	388	611	185	38	122	2
	Geometry	1	46	7	35	122	2
	Statistics and Probability	2	20	97	28	121	2
	Performance Task	25	1	5	1	2	2
	Ratios and Proportional Relationships	196	718	177	28	121	2
8	Expressions and Equations I	9	49	120	441	43	
	Expressions and Equations II	47	155	159	450	43	
	Functions	597	124	137	448	43	
	Geometry	34	56	162	451	43	
	Performance Task	3	2	1	14	43	

Appendix B: Percentage of Proficient Students in 2014–2015, 2015–2016, and 2016–2017 for All Students and by Subgroups

Table B-1. ELA/Lit Percentages of Proficient Students Across Years (Grades 3–5)

Group	2014–2015	2015–2016	2016–2017
Grade 3			
All Students	54	54	52
Female	59	57	55
Male	49	50	48
American Indian/Alaska Native	76	58	53
Asian	80	80	78
African American	39	39	36
Hispanic/Latino	41	41	39
White	66	66	66
ELL	23	28	32
SPED	13	14	15
CD 504	44	52	47
Title I	54	59	63
Grade 4			
All Students	54	56	54
Female	58	61	58
Male	49	51	50
American Indian/Alaska Native	65	61	51
Asian	81	81	83
African American	37	41	39
Hispanic/Latino	40	43	42
White	68	68	67
ELL	14	16	21
SPED	11	13	12
CD 504	51	49	47
Title I	49	57	58
Grade 5			
All Students	55	60	60
Female	61	66	65
Male	50	55	55
American Indian/Alaska Native	59	68	61
Asian	84	85	87
African American	39	44	45
Hispanic/Latino	44	49	47
White	68	73	72
ELL	9	13	13
SPED	11	15	16
CD 504	50	53	56
Title I	56	60	64

Table B-2. ELA/Lit Percentages of Proficient Students Across Years (Grades 6–8)

Group	2014–2015	2015–2016	2016–2017
Grade 6			
All Students	48	52	52
Female	55	57	57
Male	41	46	47
American Indian/Alaska Native	52	47	53
Asian	80	81	82
African American	33	35	35
Hispanic/Latino	38	40	39
White	59	65	65
ELL	5	7	4
SPED	8	9	10
CD 504	43	47	48
Title I	45	52	49
Grade 7			
All Students	50	52	54
Female	58	59	59
Male	43	46	48
American Indian/Alaska Native	50	66	53
Asian	81	82	83
African American	33	35	36
Hispanic/Latino	39	41	42
White	63	65	68
ELL	9	5	7
SPED	8	10	11
CD 504	44	45	50
Title I	50	52	53
Grade 8			
All Students	49	54	52
Female	56	61	60
Male	43	47	45
American Indian/Alaska Native	66	56	67
Asian	80	80	80
African American	33	38	36
Hispanic/Latino	38	43	42
White	60	66	64
ELL	7	8	8
SPED	10	9	10
CD 504	44	48	48
Title I	42	54	52

Table B-3. Mathematics Percentages of Proficient Students Across Years (Grades 3–5)

Group	**2014–2015**	**2015–2016**	**2016–2017**
Grade 3			
All Students	53	55	53
Female	53	54	53
Male	53	56	54
AmeriIndian/AlaskaNat	66	50	44
Asian	80	87	82
African American	36	39	36
Hispanic	41	44	42
White	67	68	68
ELL	25	35	40
SPED	14	17	18
CD 504	48	49	50
Title I	54	61	65
Grade 4			
All Students	47	51	50
Female	45	50	49
Male	48	51	52
AmeriIndian/AlaskaNat	56	49	41
Asian	78	81	83
African American	29	33	32
Hispanic	36	38	37
White	60	65	65
ELL	16	18	22
SPED	8	12	13
CD 504	40	47	49
Title I	46	56	58
Grade 5			
All Students	38	42	44
Female	37	40	44
Male	39	43	44
AmeriIndian/AlaskaNat	34	43	35
Asian	74	74	76
African American	21	23	26
Hispanic	27	29	31
White	50	56	59
ELL	8	8	7
SPED	5	6	8
CD 504	29	35	37
Title I	38	45	48

Table B-4. Mathematics Percentages of Proficient Students Across Years (Grades 6–8)

Group	**2014–2015**	**2015–2016**	**2016–2017**
Grade 6			
All Students	34	37	41
Female	35	37	42
Male	33	37	40
AmeriIndian/AlaskaNat	38	28	51
Asian	69	70	76
African American	17	21	22
Hispanic	22	24	29
White	46	50	56
ELL	4	4	5
SPED	4	5	6
CD 504	28	32	38
Title I	30	37	40
Grade 7			
All Students	37	40	41
Female	39	41	41
Male	35	38	41
AmeriIndian/AlaskaNat	29	55	36
Asian	71	77	77
African American	19	21	23
Hispanic	26	29	30
White	50	52	55
ELL	5	7	6
SPED	4	6	7
CD 504	33	36	38
Title I	33	39	42
Grade 8			
All Students	35	38	38
Female	36	41	41
Male	35	35	35
AmeriIndian/AlaskaNat	42	42	56
Asian	71	74	72
African American	17	20	21
Hispanic	27	25	29
White	47	51	50
ELL	9	9	10
SPED	5	5	5
CD 504	31	32	31
Title I	30	33	38

Appendix C: Classification Accuracy and Consistency Indexes by Subgroups

Table C-1. ELA/Lit Classification Accuracy and Consistency by Achievement Levels (Grades 3–5)

Group	N	%Accuracy					%Consistency				
		All	L1	L2	L3	L4	All	L1	L2	L3	L4
Grade 3											
All Students	10,600	80	89	72	70	89	72	82	62	59	83
Female	5,171	80	88	72	70	89	72	81	63	59	84
Male	5,429	80	89	73	70	88	73	84	62	60	82
American Indian/Alaska	36	80	90*	71*	70*	89	72	86*	62*	58*	81
Asian	371	84	87	74	70	93	77	78	63	59	89
African American	3,206	80	90	72	70	85	72	84	63	59	77
Hispanic	1,997	79	88	73	70	87	71	83	62	60	78
White	4,513	80	88	73	70	89	73	79	62	59	85
ELL	1,635	79	88	72	69	85	71	83	63	59	76
Special Education	1,438	84	92	73	70	81	78	90	62	57	67
CD 504	331	78	87	74	69	85	69	78	64	58	79
Title I	1,035	79	85	74	70	88	71	76	63	60	83
Grade 4											
All Students	10,386	79	89	65	68	88	71	83	53	58	82
Female	5,150	78	88	65	68	89	70	81	53	58	83
Male	5,236	79	90	65	68	88	71	85	53	58	81
American Indian/Alaska	41	77	89*	67	67	95*	68	74*	58	58	83*
Asian	383	81	84	62	66	92	75	74	49	57	88
African American	3,143	79	90	65	68	86	71	86	53	59	77
Hispanic	1,838	77	89	65	68	86	69	84	54	58	76
White	4,518	79	87	65	68	89	71	79	53	58	84
ELL	886	80	90	65	69	82	72	86	53	56	66
Special Education	1,474	85	93	64	68	87	80	91	52	56	67
CD 504	413	77	86	65	69	87	68	80	55	57	79
Title I	1,046	77	88	65	68	88	68	80	53	59	80
Grade 5											
All Students	10,461	80	88	68	76	87	72	82	56	68	81
Female	5,230	80	88	68	76	88	72	81	56	68	82
Male	5,231	80	89	68	76	87	72	83	56	69	79
American Indian/Alaska	31	81	85*	74*	80*	82	73	80*	59*	70*	80
Asian	367	85	88	66	76	92	79	75	52	68	90
African American	3,077	79	90	67	76	83	71	85	56	68	72
Hispanic	1,824	78	86	68	76	84	69	79	57	68	74
White	4,708	81	87	68	76	89	73	80	55	68	83
ELL	440	83	90	67	75	64*	77	88	54	65	44*
Special Education	1,526	84	92	67	74	80	78	90	55	65	62
CD 504	462	78	84	68	77	86	70	78	55	70	77
Title I	1,247	79	86	69	76	87	70	77	57	69	79

*The classification index is based on n < 10.

Table C-2. ELA/Lit Classification Accuracy and Consistency by Achievement Levels (Grades 6–8)

Group	N	%Accuracy					%Consistency				
		All	**L1**	**L2**	**L3**	**L4**	**All**	**L1**	**L2**	**L3**	**L4**
Grade 6											
All Students	10,189	80	90	73	78	84	72	83	64	70	76
Female	5,055	80	89	73	78	84	72	81	64	70	77
Male	5,134	81	90	74	77	84	73	84	64	70	75
American Indian/Alaska	43	77	97*	71	75	87*	68	73*	67	65	75*
Asian	381	84	91	74	78	90	77	79	62	71	86
African American	3,133	80	90	73	78	81	73	84	64	69	68
Hispanic	1,776	80	89	73	78	79	72	84	64	70	67
White	4,458	80	89	74	77	84	72	80	64	70	78
ELL	392	89	94	75	71	98*	84	92	65	54	79*
Special Education	1,483	86	93	73	75	83	81	90	63	64	60
CD 504	459	79	87	73	78	82	70	79	65	70	69
Title I	1,336	79	88	74	78	83	71	81	64	70	73
Grade 7											
All Students	10,070	81	90	72	80	84	73	83	62	73	76
Female	4,936	80	89	72	80	85	73	81	62	73	77
Male	5,134	81	90	72	80	84	74	85	63	73	74
American Indian/Alaska	45	78	85*	71	76	87*	70	75	60	71	75
Asian	358	85	90	72	80	90	78	80	58	72	87
African American	3,201	81	91	72	79	82	74	86	63	71	70
Hispanic	1,604	80	90	72	80	78	72	83	63	73	66
White	4,570	80	87	72	80	85	73	79	61	74	76
ELL	339	87	94	71	81	74*	83	91	62	63	63*
Special Education	1,431	86	93	72	77	83	80	90	62	64	73
CD 504	489	79	88	70	78	86	71	78	63	72	74
Title I	1,567	80	89	72	79	83	72	82	62	73	73
Grade 8											
All Students	10,069	81	89	74	80	84	73	82	64	73	75
Female	4,942	80	88	73	80	85	73	80	64	73	76
Male	5,127	81	89	74	79	84	74	84	64	73	73
American Indian/Alaska	45	81	88*	79*	80	83*	73	80	63	77	69
Asian	348	83	85	75	77	90	76	77	63	71	84
African American	3,096	81	90	74	80	81	74	85	64	72	68
Hispanic	1,646	81	89	74	80	82	73	83	64	73	70
White	4,678	80	87	74	80	84	72	79	64	73	76
ELL	322	86	91	73	79	95*	80	89	62	64	67
Special Education	1,432	86	92	74	79	82	80	89	64	66	62
CD 504	495	80	89	76	78	85	72	81	66	73	70
Title I	1,714	80	88	74	79	82	72	81	65	73	71

*The classification index is based on n < 10.

Table C-3. Mathematics Classification Accuracy and Consistency by Achievement Levels (Grades 3–5)

Group	**N**	**%Accuracy**					**%Consistency**				
		All	**L1**	**L2**	**L3**	**L4**	**All**	**L1**	**L2**	**L3**	**L4**
Grade 3											
All Students	10,669	82	89	73	79	90	75	82	63	72	84
Female	5,203	82	88	73	79	90	74	81	63	72	84
Male	5,466	82	90	73	79	89	75	83	64	71	85
American Indian/Alaska Native	36	83	90	75	77*	89*	77	83	68	66*	86*
Asian	394	86	82	74	78	93	80	76	61	70	90
African American	3,216	82	90	73	79	87	74	84	64	71	79
Hispanic	2,031	81	89	73	79	87	73	82	64	71	80
White	4,514	83	87	73	79	90	76	78	63	72	86
ELL	1,707	81	90	73	79	88	74	83	64	71	80
Special Education	1,441	85	93	72	78	85	79	90	63	67	75
CD 504	336	81	85	76	78	88	73	78	65	72	81
Title I	1,045	82	86	73	79	90	74	77	63	73	84
Grade 4											
All Students	10,442	83	89	80	79	89	77	82	73	71	84
Female	5,183	83	88	80	78	89	76	80	73	71	83
Male	5,259	84	90	80	79	90	77	83	73	71	85
American Indian/Alaska Native	41	87	91*	84	81*	99*	81	77	83	72	90
Asian	398	87	82	83	77	92	81	74	73	68	90
African American	3,155	83	90	79	78	85	76	84	73	70	77
Hispanic	1,871	82	89	80	79	85	75	81	74	70	78
White	4,514	84	88	81	79	90	77	78	73	72	86
ELL	954	83	90	80	78	86	77	84	74	67	77
Special Education	1,479	87	92	80	76	90	82	90	71	68	71
CD 504	416	83	88	83	80	86	76	78	76	72	82
Title I	1,052	83	86	80	79	91	76	76	73	72	84
Grade 5											
All Students	10,519	83	89	77	72	90	76	84	69	61	86
Female	5,255	82	88	78	71	90	75	82	69	61	85
Male	5,264	83	90	77	72	90	76	85	69	62	86
American Indian/Alaska Native	31	78	96*	72	60*	89*	72	79*	72	49*	82*
Asian	378	87	88	77	73	93	82	75	70	59	93
African American	3,089	83	90	77	72	87	76	86	69	62	77
Hispanic	1,861	82	89	76	71	89	74	83	69	61	81
White	4,706	83	88	78	71	91	76	80	69	61	87
ELL	507	87	93	75	68	89	82	90	65	53	82
Special Education	1,543	88	93	77	71	79	83	91	66	57	71
CD 504	468	81	87	78	71	90	74	81	71	58	85
Title I	1,254	81	87	77	72	90	74	79	69	62	84

*The classification index is based on n < 10.

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

Group	N	%Accuracy					%Consistency				
		All	L1	L2	L3	L4	All	L1	L2	L3	L4
Grade 6											
All Students	10,211	83	91	78	72	89	76	86	70	63	83
Female	5,072	82	91	79	72	89	75	85	70	63	83
Male	5,139	83	91	77	72	90	76	86	69	63	84
American Indian/Alaska Native	43	81	83*	75	77	92	74	74*	68	68	86
Asian	389	85	89	76	70	94	79	83	65	63	90
African American	3,138	83	91	78	71	85	77	87	71	60	76
Hispanic	1,794	83	91	79	72	85	76	87	71	62	76
White	4,447	82	90	78	73	90	75	83	69	64	84
ELL	435	91	95	76	69	91*	87	94	66	57	79*
Special Education	1,478	90	95	76	71	84	86	93	67	57	75
CD 504	455	82	91	79	73	88	75	84	72	63	80
Title I	1,339	81	90	78	71	87	73	83	71	62	79
Grade 7											
All Students	10,087	83	91	76	74	89	76	85	67	65	84
Female	4,943	82	90	76	74	89	75	85	68	65	83
Male	5,144	83	91	76	75	90	76	86	67	65	84
American Indian/Alaska Native	45	82	87	83	76	77*	74	83	72	66	77*
Asian	362	87	96	74	77	94	82	84	66	68	92
African American	3,199	83	92	76	74	87	77	87	67	64	78
Hispanic	1,636	82	91	75	73	86	76	87	66	63	79
White	4,552	82	88	77	75	90	74	81	68	66	85
ELL	385	90	94	74	74	83*	86	93	62	57	78*
Special Education	1,420	90	95	73	72	80	86	93	61	61	74
CD 504	488	80	88	75	74	91	73	80	67	66	84
Title I	1,568	82	90	76	75	88	75	85	68	66	82
Grade 8											
All Students	10,058	82	90	72	71	91	75	85	63	61	86
Female	4,944	81	89	72	71	90	74	84	63	60	86
Male	5,114	83	90	72	72	91	76	86	62	62	86
American Indian/Alaska Native	45	83	92	76*	69	93	76	85	61*	65	88
Asian	356	87	90	71	73	96	83	81	64	60	94
African American	3,092	83	91	72	71	87	77	88	62	60	79
Hispanic	1,669	82	90	72	71	90	75	86	62	60	81
White	4,641	81	87	73	71	91	74	81	63	61	87
ELL	379	88	94	72	64	83	84	92	60	51	83
Special Education	1,415	90	94	71	73	78	85	93	58	58	63
CD 504	489	82	90	72	72	93	74	83	64	61	86
Title I	1,714	81	90	73	70	91	74	84	64	61	83

*The classification index is based on n < 10.