Delaware Smarter Balanced Assessments

2015–2016 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. 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. SBAC, a state-led enterprise, is intended to provide leadership and resources to improve teaching and learning by creating and maintaining a suite of summative and interim assessments and tools aligned to the CCSS in ELA/Lit and mathematics. Delaware is among the 18 member states (plus the U.S. Virgin Islands) leading the development of assessments in ELA/Lit and mathematics.

The Delaware State Board of Education formally adopted the CCSS in ELA/Lit and mathematics on Aug 19, 2010 (State Board meeting minutes, 2010). The 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 human-scored items.

The Smarter Balanced assessments consist of the end-of-year summative assessments for accountability purposes and the optional interim assessments to support teaching and learning throughout the year. The summative assessments are used to determine student achievement based on the CCSS and track student progress for college and career readiness in ELA/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 hand-scored.

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

This report provides a technical summary of the 2015–2016 summative assessments in ELA/Lit and mathematics administered in grades 3–8 under the Delaware Smarter Balanced assessments. The report includes eight chapters: overview, test administration, summary of 2015–2016 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.

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

2. TEST ADMINISTRATION

2.1 TESTING WINDOWS

The 2015–2016 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-and-pencil fixed forms for summative assessments were administered for 15 days during the online summative window. Table 1 shows the testing windows for both online and paper-and-pencil assessments. For grade 11, although the grade 11 summative assessment was no longer used in the 2015–2016 school year; the grade 11 interims remained available.

Table 1. 2015–2016 Testing Windows

Tests	Grade	Start Date	End Date	Mode
Summative Assessments	3–8	3/9/2016	6/2/2016	Online Adaptive
	3–8	5/2/2016	5/18/2016	Paper Fixed Forms
Interim Comprehensive Assessments	3–8, 11	8/31/2015	7/15/2016	Online Fixed Forms
Interim Assessment Blocks	3–8, 11	8/31/2015	7/15/2016	Online Fixed Forms

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 2015–2016 administration to accommodate students' needs. Table 2 lists the testing options that were offered in 2015–2016. Testing options are selected by content area. Once an option is selected, it would apply to all tests in the content area.

Table 2. Testing Options in 2015–2016

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

To ensure standardized administration conditions, Test Administrators (TAs) follow procedures outlined in the *Smarter 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). A 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 the test administration are District Test Coordinators (DTCs), District/School Accommodations Managers (DAMs/SAMs), School Test Coordinators (STCs), and 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)

The DTC's primary responsibility is to coordinate the administration of the Smarter Balanced assessments in the district.

DTCs are responsible for the following:

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

District/School Accommodations Manager (DAMs/SAMs)

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

DAMs/SAMs are responsible for the following:

- Attending District/School Accommodations Manager training
- Completing all required DeSSA training
- Ensuring accommodations have been reviewed and updated in TIDE

School Test Coordinator (STC)

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

STCs are responsible for the following:

- Attending School Test Coordinator training
- Completing all required DeSSA training
- Completing all required security forms and ensuring that all TAs have completed all required security forms
- Ensuring that all TAs complete Smarter Balanced assessment training modules
- Working with technology personnel to ensure that the DeSSA secure browser has been installed and is working on all computers to be used with testing
- Completing the test schedule
- Reviewing students in both DELSIS and TIDE applications before students are tested
- Ensuring that TAs understand protocols in the event that a student moves to a new district and/or school
- Ensuring that all students in Department of Services for Children, Youth and their Families (DSCYF), Delaware Adolescent Program, Inc. (DAPI), or Consortium Discipline Alternative Program (CDAP) programs have home school records
- Ensuring accommodations have been reviewed and updated in the Assessment Accommodations Database and are correct in TIDE
- Entering in the SysAID any security issues, incidents, data reviews, unique accommodations, or exemptions required for any Smarter Balanced assessment
- General oversight responsibilities for all administration activities in their school and overseeing TAs

Test Administrators (TA)

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

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

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

Student teachers and school support staff may act as proctors.

TAs are responsible for the following:

- Completing Smarter Balanced assessment administration training.

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

2.2.2 Online Test Administration

Within the state's testing window, schools can set testing schedule, allowing students to test in intervals (e.g., multiple sessions) rather than in one long period, minimizing the interruption of classroom instruction and efficiently utilizing its facility. With online testing, schools do not need to handle test booklets and address the storage and security problems inherent in large shipments of materials to a school site.

STCs oversee all aspects of testing at their schools and serve as the main point of contact, 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/SAMs are responsible for ensuring student accommodations are correctly entered into TIDE.

To start a test session, the TA must first enter the TA Interface of the online testing system using his or her own computer. A session ID is generated when the test session is created. Students who are taking the assessment with the TA need to enter their statewide student identifier (SSID), first name, and the 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 needs to read the *Directions for Administration* in the *Online Smarter Balanced Test Administration Manual* aloud to the student(s) and walk them through the 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 scroll back to review and edit previously answered items, as long as these items are in the same test session, and this session has not been paused for more than 20 minutes. Students may review and edit their responses they have previously completed before submitting the assessment. During an active CAT session, if a student reviews and changes the response to a previously

answered item, then all following items to which the student already responded remain the same. No new items are assigned to this student for changing the answers. For example, a student paused for 10 minutes after completing item 10. After the pause, the student went back to item 5 and changed the answer. If the response change in item 5 changed the item score from wrong to right, the student's overall score would improve; however, there will be no change in items 6–10. No pause rule is implemented for the performance tasks. The same rules that apply to the CAT for reviews and changes to responses also apply to performance tasks.

For the summative test, an assessment can be started in one test session and completed in a different session. For the CAT, the assessment must be completed within 45 calendar days of the start date, otherwise, the assessment will be expired. For the Performance Task, the assessment must be completed within 10 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 CAT, the assessments 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 starts from where the student left off. Previous responses and editing are no longer available.

The TA must remain in the room 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 any handouts or scratch papers that students used during the assessment to securely shred them.

2.2.3 Paper-and-Pencil Test Administration

The paper-and-pencil versions of the Smarter Balanced ELA/Lit and mathematics assessments are provided as an accommodation for students who could not access to a computer or students with blindness or visual impaired. Although the online Braille was available, only the paper-and-pencil Braille test was used in Delaware in 2015-2016 administration.

The DTC must submit a request to DDOE on behalf of the students who need to take the paper/pencil test for test materials. If the request is approved, the testing contractor will ship the appropriate test booklets to the district.

Separate test booklets are used for ELA/Lit and for mathematics. The items from the CAT and the Performance Task components are combined into one test booklet, including two sessions for CAT and one session for performance task in both content areas. 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 vendor. The testing vendor scans the answer document and scores the test, including the hand-scored items.

2.2.4 Braille Test Administration

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

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

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

2.3 Training and Information for Test Coordinators and Administrators

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

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

Table 3. Smarter Balanced Summative 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
Smarter Summative Test Administrator	Smarter Summative Test Administrator Training	24619	• Smarter 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 Interim Training • THSS Training	• 30 min • 35 min • 30 min • 30 min
Smarter Interim Test Administrator	Smarter Interim Test Administrator Training	24288	• Smarter Interim 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 Students Coordinator**	Accessibility Coordinator Training	24483*	• DeSSA Overview • Accessibility (TBD)	• 30 min • 50 min
Secretaries, Administrative Support	Security Training	24621	• Security module only	• 30 min
Test Administrators who are giving paper-and-pencil only* (if TA is giving online and p/p, take these and the online requirements)	DeSSA Paper/Pencil Test Administrator Training for Smarter, 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-and-pencil version must take the corresponding training (Summative 24619, Interim 24288, or DCAS-EOC 24251).

2.3.1 Practice and Training Test Site

In August 2015, 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 get familiar with a variety of question types and difficulty levels (approximately 30 items each in mathematics and ELA/Lit), as well as the performance task.

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

A student can log in directly to the practice and training test site as a "Guest" without a TA-generated test session ID, or the student can log in through a training test session created by the TA in the TA training site. Items in the student training test include all item types that are 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
TIDE User Guide	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.
Test Administrator User Guide	The Test Administrator User Guide supports individuals using the test delivery system applications to manage testing for students participating in the summative assessment. This resource provides information about the test delivery system, including the TA and student applications.
Usability, Accessibility, and Accommodations Guidelines	The Usability, Accessibility, and Accommodations Guidelines focus on universal tools, designated supports, and accommodations for the Smarter Balanced assessments. The guidelines are intended for school-level personnel and decision-making teams, particularly Individualized Educational Program (IEP) and Section 504 teams, as they prepare for and implement the Smarter Balanced assessments. The guidelines provide information for classroom teachers, English development educators, special education teachers, and related services personnel to use in selecting and administering universal tools, designated supports, and accommodations for those students who need them. The guidelines are also intended for assessment staff and administrators who oversee the decisions that are made in instruction and assessment.
Accessibility Guidelines for Delaware System of Student Assessments (DeSSA)	The Accessibility Guidelines for DeSSA provide information about identifying and documenting students who are eligible to receive designated supports and accommodations on Smarter Balanced and other DeSSA assessments. 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 ELA/Literacy and	This Test Administration Manual (TAM) provides needed information regarding policies and procedures for the Smarter English Language Arts/Literacy and

Resource	Description
Mathematics Online Summative Test Administration Manual	Mathematics Online Summative Assessments.
Smarter Balanced Test Administration Manual for Paper and Pencil	The Smarter Balanced Test Administration Manual (TAM) for Paper and Pencil will provide administration information and requirements for administering the paper-and-pencil test.
Smarter Balanced Test Administration Manual for ICAs and IABs	The Smarter Balanced Test Administration Manual (TAM) for ICAs and IABs will provide administration information and requirements for administering the interim comprehensive assessment.
Technology Specifications Manual (TSM) for Online Testing	The Technology Specifications Manual provides technology staff with the technical specifications for online testing, including information on Internet and network requirements, general hardware and software requirements, secure browser installation, and text-to-speech function.
DeSSA Test Security Manual	The DeSSA Test Security Manual provides information regarding test security policies for all DeSSA tests. School personnel, including Test Administrators, should review this document carefully.
Secure Browser Installation Manual	The Secure Browser Installation Manual provides instructions for installing the secure browser on supported operating systems and is organized by operating system. This document is a supplement to the Technical Specifications Manual for Online Testing.
Braille Requirements and Testing Manual	The Braille Requirements and Testing Manual includes information about supported operating systems and required hardware and software for braille testing. It also includes a quick guide for TAs who are testing students with a braille accommodation. This manual consolidates information that was previously split between the Technical Specifications Manual and the Test Administrator User Guide.

2.3.3 Training Modules

The following training modules were created to help users in the field understand the overall Smarter Balanced assessments as well as how each system works. All modules were provided in PowerPoint format; two modules were also narrated. Table 5 lists the training modules.

Table 5. Smarter Balanced-Developed Training Modules

Module Name	Primary Audience	Objective
Let's Talk Universal Tools	• Students • TAs • Teachers	The Smarter Universal Tools Training module 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	The Student Interface Training module provides information on how students log in and navigate the student testing system, including information on layout and functionality of the test tools.
What Is a CAT (Computer Adaptive Test)?	• DTCs and STCs • Teachers	This is a presentation produced by Smarter Balanced, introducing test administrators 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 secured materials for all assessments. The importance of maintaining test security and the integrity of test items is stressed throughout the webinar trainings and in the user guides, modules, and manuals. Features in the testing system also protect test security. This section describes system security, student confidentiality, and policies on testing impropriety.

2.4.1 DeSSA Test Security Manual

Test security is critically important to protect the intellectual properties, reduce test fraud and theft, and maintain the integrity of the state assessments; and therefore, to ensure the validity and reliability of test scores, and fairness in testing for all Delaware students. The Test Security Manual provided online at the DeSSA portal, 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 the preparation for the 2015-2016 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, 2014. All unethical or inappropriate practice 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 reported to the Office of Assessment at the Delaware Department of Education and documented.

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

The manual specifically indicates the test security in the administration of the Smarter Balanced assessments in ELA/Lit and mathematics. For example, scratch papers and any materials developed during the classroom activities must be securely disposed of prior to the administration of a performance task (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 signifies prohibiting 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 appropriate test content:

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

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

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

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

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

Students log in to the online assessment using their legal first name, SSID number, and 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-and-pencil versions of the assessments, TAs are required to affix the student label to the student's answer document.

After a test session, only staff with the administrative roles of DTCs, STCs or teachers can view their students' scores. TAs do not have access to student scores.

2.4.3 System Security

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

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

Password protection: All access points by different roles—at the state level, district level, school principal level, and school staff level—require a password to 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 the STC is to ensure that the secure browser is properly installed on the computers used for the administration of the online assessments. Developed by the testing contractor, the secure browser prevents students from accessing other computers or Internet applications and from copying test information. The secure browser suppresses access to commonly used browsers such as Internet Explorer and Firefox and prevents students from searching for answers on the Internet or communicating with other students. The assessments can be accessed only through the secure browser and not by other Internet browsers.

2.4.4 Security of the Testing Environment

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

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

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

If a student needs to leave the room for a brief time, the 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 performance tasks, different forms are spiraled within a classroom so that students receive different forms of the performance tasks.

After the Test

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

2.4.5 Test Security Violations

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

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

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

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

District and School personnel must document all test security incidents. The DTC is responsible for reporting test security incidents to the state via SysAID. 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 administration of Smarter Balanced assessments was intended to improve test administration and monitoring for the 2017 test administration. The Office of Assessment at the Department of Education scheduled on-site visits (upon agreement with schools) during the test window and all observers followed the procedure for the on-site visits without interfering with test activities (Smarter Balanced Spring 2016 Site Visits).

The Observation and Discussion Form provides each observer with a general checklist for the appropriate test practices and standardized test conditions. The observation includes seven elements: (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 Performance Tasks.

The Feedback Form was used to collect comments from schools and districts regarding Smarter Balanced administration, test materials, technology, service and Help Desk, and other aspects of testing. Communication with 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. Eligibility for the grade 8 Smarter Balanced assessments was extended in 2015–2016 to include students in the next higher grade to accommodate skippers/repeaters. The DDOE policy on skipper/repeaters may be found here: http://www.doe.k12.de.us/cms/lib09/DE01922744/Centricity/Domain/111/Repeaters-Skippers%20Policy%20rev%2001%2026%202016.pdf.

2.5.1 Home-Schooled Students

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

2.5.2 Exempt Students

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

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

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

2.6 ONLINE TESTING FEATURES AND TESTING ACCOMMODATIONS

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

The Smarter Balanced *Guidelines* apply to all students. They emphasize an individualized approach to the implementation of assessment practices for those students who have diverse needs and

participate in large-scale content assessments. The *Guidelines* focus on universal tools, designated supports, and accommodations for the Smarter Balanced assessments of ELA/Lit and mathematics. At the same time, the *Guidelines* support important instructional decisions about accessibility and accommodations for students who participate in the Smarter Balanced assessments.

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

In 2015, the administration of Smarter Balanced assessments was 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 computer administration system, whereas non-embedded resources are provided outside of that system.

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

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

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

2.6.1 Online Universal Tools for ALL students

Universal tools are access features of an assessment or exam that are *digitally delivered* (i.e., embedded) or separately delivered (i.e., non-embedded) components of the test administration system. Universal tools are available to all students based on their preference and selection and have been preset in TIDE. In 2015–2016 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

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

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

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

Calculator: An embedded on-screen digital calculator can be accessed for calculator-allowed items when students click the calculator button. This tool is available only with the specific items for which the Smarter Balanced Item Specifications indicated that it would be appropriate.

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

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

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

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

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

Cross out response options: To cross out response options, use the strikethrough function.

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

Take as much time as needed to complete a Smarter Balanced Assessment: Testing may be split across multiple sessions so that the testing does not interfere with class schedules. The CAT must be completed within 45 calendar days of its starting date. The performance tasks must be completed within 10 calendar days of the starting date.

Non-Embedded Universal Tools

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

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

Scratch paper: Scratch paper to make notes, write computations, or record responses may be made available. Only plain paper or lined paper is appropriate for ELA/Lit. Graph paper is required beginning in grade 6 and can be used on all math assessments. A student can use an assistive technology device for scratch paper as long as the device is consistent with the child's IEP or Section 504 Plan and is acceptable to the state.

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

2.6.2 Designated Supports and Accommodations

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

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

Embedded Designated Supports

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

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

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

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

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

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

Non-Embedded 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 performance task.

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

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

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

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

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

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

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

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

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

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

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

Embedded Accommodations

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

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

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

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

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

Non-Embedded Accommodations

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

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

Calculator (for grades 6–8 math tests): A non-embedded calculator for students needing a special calculator, such as a braille calculator or a talking calculator, currently unavailable in the assessment platform.

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

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

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

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

Speech-to-text: Voice recognition allows students to use their voices as devices to input information into the computer to dictate responses or give commands (e.g., 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 2015–2016 administration. Tables 7 through 12 provide the number students who were offered the accommodations and/or designated supports.

Table 6. Universal Tools, Designated Supports, and Accommodations Available in 2015–2016

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

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

[1] For calculator-allowed items only in grades 6-8 and 11
[2] For ELA 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 PT stimuli, ELA PT and CAT items (not ELA/Lit CAT reading passages), and math stimuli and items: Must be set in TIDE before test begins.
[6] For math items
[7] For math items
[8] For math test
[9] For ELA listening items and math items
[10] For ELA listening items
[11] For ELA reading passages. Must be set in TIDE by state-level user. TCs must submit a student's Verification of Need form to the Assessment Section for review and approval or disapproval.
[12] For ELA performance task full writes
[13] For ELA performance task full writes
[14] For ELA performance task full writes
[15] For ELA items (not ELA reading passages) and math items
[16] For ELA non-writing items and math items
[17] For math items on the paper/pencil test
[18] Includes adapted keyboards, large keyboard, StickyKeys, MouseKeys, FilterKeys, adapted mouse, touch screen, head wand, and switches
[19] For calculator-allowed items only in grades 6-8 and 11
[20] For math items beginning in grade 4

[21] For ELA reading passages, all grades

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

Accommodations	Grade					
	3	**4**	**5**	**6**	**7**	**8**
Embedded Accommodations						
American Sign Language	8	6	3	2	5	3
Closed Captioning	14	7	12	30	12	10
Braille			1			
Streamlined Mode	10	3	4	2	1	3
Text-to-Speech: Passages & Items	4	11	12	18	18	16
Non-Embedded Accommodations						
Alternate Response Options	2		1	1	1	1
Print on Demand: Items	2		5	1	2	1
Print on Demand: Passages	13	4	4	14	2	8
Print on Demand: Passages & Items	368	376	354	310	297	281
Read Aloud Passages	60	62	103	216	175	181
Scribe Items (Writing)	55	49	42	22	12	12
Speech-to-Text	8	14	15	9	1	1

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

Designated Supports	Subgroup	Grade					
		3	**4**	**5**	**6**	**7**	**8**
Color Contrast	Overall	6	8	41	28	40	27
	ELL			3	1	1	
	Special Ed	5	7	35	24	29	25
Masking	Overall	199	183	151	185	160	100
	ELL	88	30	15	28	25	27
	Special Ed	88	84	98	158	109	90
Text-to-Speech: Items	Overall	1,997	1,862	1,773	1,106	1,068	888
	ELL	779	469	306	188	143	175
	Special Ed	900	1,004	1,030	906	876	733
Text-to-Speech: Stimuli & Items	Overall	1,942	1,833	1,753	1,093	1,046	886
	ELL	756	446	290	178	138	173
	Special Ed	887	999	1,033	906	870	739
Turn off Any Universal Tools	Overall	28	7	5	11	6	5
	ELL	4	3		1		
	Special Ed	17	2	5	10	6	5

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

Designated Supports	Subgroup	Grade					
		3	**4**	**5**	**6**	**7**	**8**
Bilingual Dictionary	Overall	1	6	8	34	43	32
	ELL	1	5	8	33	43	32
	Special Ed			3	8	12	6
Color Contrast	Overall	1	1	13	4	4	3
	ELL			3	1		
	Special Ed		1	11	4	2	2
Color Overlay	Overall	2	1	1	1	3	1
	ELL						
	Special Ed	1	1	1	1	2	
Magnification	Overall	3	3	8	33	4	5
	ELL				4		1
	Special Ed	3	3	5	31	2	3
Noise Buffers	Overall	29	44	34	19	8	4
	ELL	4	1	2	2		
	Special Ed	14	36	23	14	2	1
Read Aloud Items	Overall	208	189	218	260	226	246
	ELL	33	12	23	25	21	35
	Special Ed	124	135	159	257	216	238
Read Aloud Passages	Overall	169	165	191	247	209	226
	ELL	16	11	18	22	18	26
	Special Ed	103	112	135	244	202	223
Scribe Items (Non-Writing)	Overall	41	46	36	25	8	12
	ELL	3	4	2		1	2
	Special Ed	31	40	33	24	8	11
Separate Setting	Overall	379	406	406	618	577	570
	ELL	77	41	34	43	43	62
	Special Ed	256	308	325	564	540	528
Translated Test Directions	Overall	3	1		2	7	8
	ELL	2			1	7	8
	Special Ed	2			1	2	2

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

Accommodations	Grade					
	3	**4**	**5**	**6**	**7**	**8**
Embedded Accommodations						
American Sign Language	8	6	3	2	4	3
Streamlined Mode	9	3	4	30	19	18
Non-Embedded Accommodations						
Abacus	1	1	2	10	4	4
Alternate Response Options	2		1		1	1
Calculator	10	48	56	283	231	242
Multiplication Table		378	441	556	431	406
Print on Demand: Items	2	8	3	2	3	5
Print on Demand: Stimuli	2	1				
Scribe Items (Writing)	47	47	34	19	11	14
Speech-to-Text	8	10	12	5	1	1

Table 11. Students with Allowed Embedded Designated Supports in Mathematics

Designated Supports	Subgroup	Grade					
		3	**4**	**5**	**6**	**7**	**8**
Color Contrast	Overall	6	5	30	29	26	20
	ELL			2	1	1	
	Special Ed	5	5	28	25	23	20
Masking	Overall	201	189	157	173	135	104
	ELL	96	33	18	28	26	32
	Special Ed	87	80	98	148	106	90
Translation (Glossary): Spanish	Overall	16	19	16	12	7	4
	ELL	15	17	15	12	7	4
	Special Ed	7	5	1	2		1
Translation (Glossary): Other Languages	Overall	13	8	14	6	7	2
	ELL	12	8	14	6	7	2
	Special Ed						
Text-to-Speech: Items	Overall			1			
	ELL						
	Special Ed						
Text-to-Speech: Stimuli & Items	Overall	1964	1826	1799	1149	1114	897
	ELL	778	439	306	178	148	152
	Special Ed	878	995	1037	930	898	759
Turn off Any Universal Tools	Overall	22	6	6	12	18	7
	ELL	4	3		4	4	3
	Special Ed	17	2	6	9	5	4

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

Designated Supports	Subgroup	Grade					
		3	**4**	**5**	**6**	**7**	**8**
Color Contrast	Overall	1		14	3	3	3
	ELL			3	1		
	Special Ed			11	3	1	2
Color Overlay	Overall	1	1			3	2
	ELL						
	Special Ed		1			2	1
Translation (Glossary): Spanish	Overall	7	9	13	23	26	36
	ELL	7	8	12	23	26	36
	Special Ed	1		2	3	5	3
Translation (Glossary): Other Languages	Overall	2	2	5	1	8	2
	ELL	2	2	3	1	8	2
	Special Ed			3		1	
Magnification	Overall	2	3	6	20	4	3
	ELL				1		
	Special Ed	1	2	5	16	2	3
Noise Buffers	Overall	24	38	31	11	6	4
	ELL	2	1	1	1		
	Special Ed	9	30	21	5	2	1
Read Aloud Items	Overall	184	174	174	188	198	216
	ELL	20	11	23	20	17	25
	Special Ed	112	120	110	186	190	212
Read Aloud Items (Spanish)	Overall		1	2	2	2	2
	ELL		1	2	2	2	2
	Special Ed				2	1	2
Read Aloud Stimuli	Overall	166	164	158	183	195	219
	ELL	14	10	21	19	17	24
	Special Ed	101	110	95	181	187	213
Read Aloud Stimuli (Spanish)	Overall		1	2	2	2	2
	ELL		1	2	2	2	2
	Special Ed				2	1	2
Scribe Items	Overall	34	40	29	13	4	8
	ELL	2	2	3			1
	Special Ed	26	35	27	13	4	8
Separate Setting	Overall	336	360	357	520	517	531
	ELL	63	35	37	37	34	55
	Special Ed	223	266	274	476	487	490
Translated Test Directions	Overall	1	1	5	2	5	8
	ELL	1	1	5	2	5	8
	Special Ed			1		2	3

2.7 DATA FORENSICS PROGRAM

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

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

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

2.7.1 Changes in Student Performance

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

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

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

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

where *s* = standard deviation of residuals in an aggregate unit; *n* = number of students in an aggregate unit (e.g., testing session, test administrator, or school).

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

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

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

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.

2.7.2 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.

2.7.3 Inconsistent Item Response Pattern (Person Fit)

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

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

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

The person-fit index is computed using a standardized log-likelihood statistic. Following Drasgow, Levine, and Williams (1985), Sotaridona, 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 8 or 15 items, the "asymptotic error probabilities are quite reasonable for nominal Type I error probabilities of 0.10 and 0.05" (Snijders, 2001).

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

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

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

3. SUMMARY OF 2015–2016 OPERATIONAL TEST ADMINISTRATION

3.1 STUDENT POPULATION

All students enrolled in grades 3–8 in all public elementary and secondary schools are required to participate in the Smarter Balanced ELA/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	10,296	10,268	10,169	9,983	10,049	9,747
Female	5,122	5,132	5,053	4,923	4,957	4,761
Male	5,174	5,136	5,116	5,060	5,092	4,986
African American	3,109	3,035	3,077	3,135	3,057	3,101
Asian	363	382	386	355	347	366
Hispanic/Latino	1,789	1,781	1,761	1,549	1,642	1,508
American Indian/Alaska Native	40	38	41	43	44	50
White	4,542	4,611	4,490	4,615	4,720	4,484
English Language Learner	1,249	641	420	298	292	329
Special Education	1,334	1,452	1,451	1,418	1,440	1,364
CD 504	319	374	424	430	453	381
Title I	1,053	1,243	1,359	1,570	1,778	1,843

Table 14. Number of Students in Summative Mathematics Assessment

Group	G3	G4	G5	G6	G7	G8
All Students	10,341	10,297	10,199	10,004	10,070	9,768
Female	5,146	5,151	5,070	4,937	4,970	4,765
Male	5,195	5,146	5,129	5,067	5,100	5,003
African American	3,106	3,041	3,077	3,125	3,054	3,097
Asian	378	391	395	361	357	370
Hispanic/Latino	1,817	1,804	1,787	1,581	1,667	1,530
American Indian/Alaska Native	40	37	42	43	44	50
White	4,547	4,605	4,484	4,607	4,710	4,483
English Language Learner	1,306	683	468	339	339	367
Special Education	1,335	1,450	1,449	1,414	1,435	1,364
CD 504	319	375	423	429	450	382
Title I	1,057	1,247	1,362	1,584	1,777	1,843

3.2 SUMMARY OF OVERALL STUDENT PERFORMANCE

Tables 15–18 present a summary of the 2015–2016 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 compare the percentage of proficient students in 2014–2015 and 2015–2016 for all students and subgroups (cohort comparisons). For all students, the percentage of proficient students increased two to five percentages in ELA/Lit and mathematics, except for grade 3 ELA/Lit. The percentage of proficient students in grade 3 ELA/Lit is same in both years. The percentage of proficient students increased for subgroups in all grades and subjects as well, except for a few subgroups with small sample sizes, e.g., American Indian/Alaska Native and CD 504. The average and the standard deviation of scale scores, and the percentage of proficient students in both 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	10,296	2439.54	85.41	20	26	25	29	54
Female	5,122	2447.47	84.60	18	25	26	32	57
Male	5,174	2431.69	85.49	23	27	24	26	50
AmeriIndian/AlaskaNat	40	2438.78	81.91	13	30	38	20	58
Asian	363	2497.24	85.72	7	13	25	55	80
African American	3,109	2409.29	79.67	31	30	23	16	39
Hispanic	1,789	2414.85	77.03	27	32	24	17	41
White	4,542	2464.64	82.20	12	22	27	40	66
ELL	1,249	2390.70	67.86	36	36	21	7	28
Special Education	1,334	2357.29	69.07	58	29	10	4	14
CD 504	319	2430.37	75.76	22	27	31	21	52
Title I	1,053	2451.20	76.97	14	27	28	30	59
Grade 4								
All Students	10,268	2482.47	90.77	24	20	25	30	56
Female	5,132	2493.65	89.81	21	19	26	35	61
Male	5,136	2471.30	90.35	27	22	25	26	51
AmeriIndian/AlaskaNat	38	2482.45	85.38	18	21	37	24	61
Asian	382	2550.66	88.59	7	12	22	59	81
African American	3,035	2448.31	86.59	36	23	24	17	41
Hispanic	1,781	2455.93	83.26	33	24	24	19	43
White	4,611	2509.63	84.66	14	17	27	41	68
ELL	641	2402.07	73.87	59	25	11	5	16
Special Education	1,452	2388.69	74.74	65	21	10	3	13
CD 504	374	2469.52	84.20	26	25	28	22	49
Title I	1,243	2484.94	78.60	20	23	29	29	57
Grade 5								
All Students	10,169	2519.27	89.98	21	19	34	26	60
Female	5,053	2531.05	87.02	16	18	35	30	66
Male	5,116	2507.64	91.34	25	20	33	22	55
AmeriIndian/AlaskaNat	41	2540.14	76.22	10	22	37	32	68
Asian	386	2585.31	79.90	6	9	29	56	85
African American	3,077	2485.02	84.92	32	24	31	13	44

Hispanic	1,761	2492.85	83.95	28	23	35	14	49
White	4,490	2546.59	84.26	12	15	36	37	73
ELL	420	2418.47	75.31	65	22	11	2	13
Special Education	1,451	2420.19	76.33	62	23	13	2	15
CD 504	424	2504.35	77.90	21	26	36	17	53
Title I	1,359	2519.70	81.62	16	23	37	24	60

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	9,983	2530.18	93.45	22	27	33	18	52
Female	4,923	2544.43	89.99	17	26	35	22	57
Male	5,060	2516.32	94.68	26	28	31	15	46
AmeriIndian/AlaskaNat	43	2526.05	84.83	23	30	33	14	47
Asian	355	2602.96	90.68	6	12	32	49	81
African American	3,135	2494.48	87.39	33	32	27	8	35
Hispanic	1,549	2505.32	87.60	28	32	31	10	40
White	4,615	2556.86	87.82	13	23	39	26	65
ELL	298	2416.13	72.08	71	21	7	0	7
Special Education	1,418	2432.00	76.53	62	29	8	1	9
CD 504	430	2524.97	84.21	20	33	34	14	47
Title I	1,570	2531.80	86.74	20	27	35	17	52
Grade 7								
All Students	10,049	2552.72	98.23	23	24	36	17	52
Female	4,957	2569.41	96.42	18	23	38	21	59
Male	5,092	2536.48	97.25	28	26	33	13	46
AmeriIndian/AlaskaNat	44	2579.49	83.28	14	20	45	20	66
Asian	347	2633.06	94.34	8	10	34	48	82
African American	3,057	2514.06	90.93	36	29	29	7	35
Hispanic	1,642	2527.20	94.98	30	28	31	10	41
White	4,720	2579.83	92.37	14	21	41	23	65
ELL	292	2434.14	69.77	73	21	5	0	5
Special Education	1,440	2449.50	77.92	66	24	10	1	10
CD 504	453	2542.15	88.14	23	31	34	11	45
Title I	1,778	2550.68	93.68	22	26	37	15	52
Grade 8								
All Students	9,747	2569.63	98.07	21	25	38	16	54
Female	4,761	2588.03	94.16	15	24	41	20	61
Male	4,986	2552.06	98.50	27	27	34	12	47
AmeriIndian/AlaskaNat	50	2579.14	100.84	20	24	34	22	56
Asian	366	2642.31	98.89	7	14	35	45	80
African American	3,101	2533.27	91.18	31	31	31	7	38
Hispanic	1,508	2542.67	92.74	27	30	36	8	43
White	4,484	2597.93	92.05	13	21	43	23	66
ELL	329	2450.26	77.68	70	22	8	0	8
Special Education	1,364	2465.41	77.41	62	29	9	1	9
CD 504	381	2562.93	85.17	19	34	36	12	48
Title I	1,843	2566.68	91.75	20	27	41	13	54

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	10,341	2443.99	78.56	20	25	32	23	55
Female	5,146	2443.38	76.83	19	26	32	23	54
Male	5,195	2444.59	80.25	20	24	32	24	56
AmeriIndian/AlaskaNat	40	2442.25	76.10	18	33	28	23	50
Asian	378	2509.32	72.96	3	9	34	54	87
African American	3,106	2411.83	74.43	32	29	28	11	39
Hispanic	1,817	2423.81	68.91	26	30	33	12	44
White	4,547	2468.18	74.37	11	21	34	34	68
ELL	1,306	2410.47	66.24	32	33	28	7	35
Special Education	1,335	2364.57	78.09	57	27	13	4	17
CD 504	319	2438.58	72.05	23	29	29	19	49
Title I	1,057	2456.14	67.15	12	26	37	24	61
Grade 4								
All Students	10,297	2485.05	79.43	17	32	29	21	51
Female	5,151	2485.09	76.04	16	34	29	21	50
Male	5,146	2485.01	82.70	19	30	29	22	51
AmeriIndian/AlaskaNat	37	2488.99	61.86	11	41	32	16	49
Asian	391	2554.95	85.66	4	15	27	54	81
African American	3,041	2451.71	72.85	29	39	24	9	33
Hispanic	1,804	2462.67	70.24	23	39	27	12	38
White	4,605	2510.14	74.61	9	27	34	31	65
ELL	683	2424.93	65.84	42	40	14	4	18
Special Education	1,450	2405.51	68.66	54	34	9	3	12
CD 504	375	2478.30	78.08	21	32	32	15	47
Title I	1,247	2494.24	67.93	11	33	34	22	56
Grade 5								
All Students	10,199	2506.76	86.83	28	30	20	22	42
Female	5,070	2505.48	84.03	28	32	20	21	40
Male	5,129	2508.02	89.50	28	29	20	23	43
AmeriIndian/AlaskaNat	42	2518.53	79.06	31	26	24	19	43
Asian	395	2579.96	85.06	8	18	19	55	74
African American	3,077	2468.61	78.39	44	33	14	9	23
Hispanic	1,787	2483.33	79.52	37	34	18	12	29
White	4,484	2535.19	81.33	16	28	25	31	56
ELL	468	2426.14	74.14	69	23	5	3	8
Special Education	1,449	2416.04	72.85	72	21	4	2	6
CD 504	423	2498.39	73.39	29	36	22	13	35
Title I	1,362	2512.44	79.96	24	31	24	21	45

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	10,004	2516.33	101.78	31	32	20	17	37
Female	4,937	2519.46	98.26	30	33	20	17	37
Male	5,067	2513.28	105.02	33	31	20	17	37
AmeriIndian/AlaskaNat	43	2510.24	90.94	26	47	12	16	28
Asian	361	2606.19	114.05	9	20	25	45	70
African American	3,125	2474.12	96.10	47	32	14	6	21
Hispanic	1,581	2487.19	91.20	41	34	16	8	24
White	4,607	2547.48	92.60	19	32	25	25	50
ELL	339	2402.24	81.94	82	14	3	1	4
Special Education	1,414	2407.39	91.02	78	17	4	2	5
CD 504	429	2513.78	93.35	29	39	17	14	32
Title I	1,584	2515.50	97.35	31	32	23	14	37
Grade 7								
All Students	10,070	2534.46	106.55	30	31	22	17	40
Female	4,970	2538.62	104.81	28	31	22	18	41
Male	5,100	2530.39	108.08	31	30	22	17	38
AmeriIndian/AlaskaNat	44	2559.96	93.08	23	23	34	20	55
Asian	357	2638.93	109.73	7	16	22	54	77
African American	3,054	2487.99	97.47	46	33	15	6	21
Hispanic	1,667	2505.67	103.75	39	33	19	10	29
White	4,710	2566.19	96.62	18	30	28	25	52
ELL	339	2421.80	96.36	73	20	6	1	7
Special Education	1,435	2423.19	89.94	74	19	6	1	6
CD 504	450	2532.91	91.25	28	36	21	14	36
Title I	1,777	2534.47	96.71	27	33	24	15	39
Grade 8								
All Students	9,768	2548.93	116.98	35	28	19	19	38
Female	4,765	2557.86	111.04	31	28	21	20	41
Male	5,003	2540.42	121.77	38	27	17	18	35
AmeriIndian/AlaskaNat	50	2549.23	111.36	32	26	20	22	42
Asian	370	2658.60	138.79	11	16	21	53	74
African American	3,097	2500.30	105.42	52	28	13	7	20
Hispanic	1,530	2517.73	104.44	45	30	15	10	25
White	4,483	2583.99	108.64	21	27	24	27	51
ELL	367	2437.81	99.83	76	15	7	2	9
Special Education	1,364	2432.01	94.50	79	16	4	1	5
CD 504	382	2541.73	101.43	36	31	20	13	32
Title I	1,843	2536.93	109.59	37	30	19	14	33

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

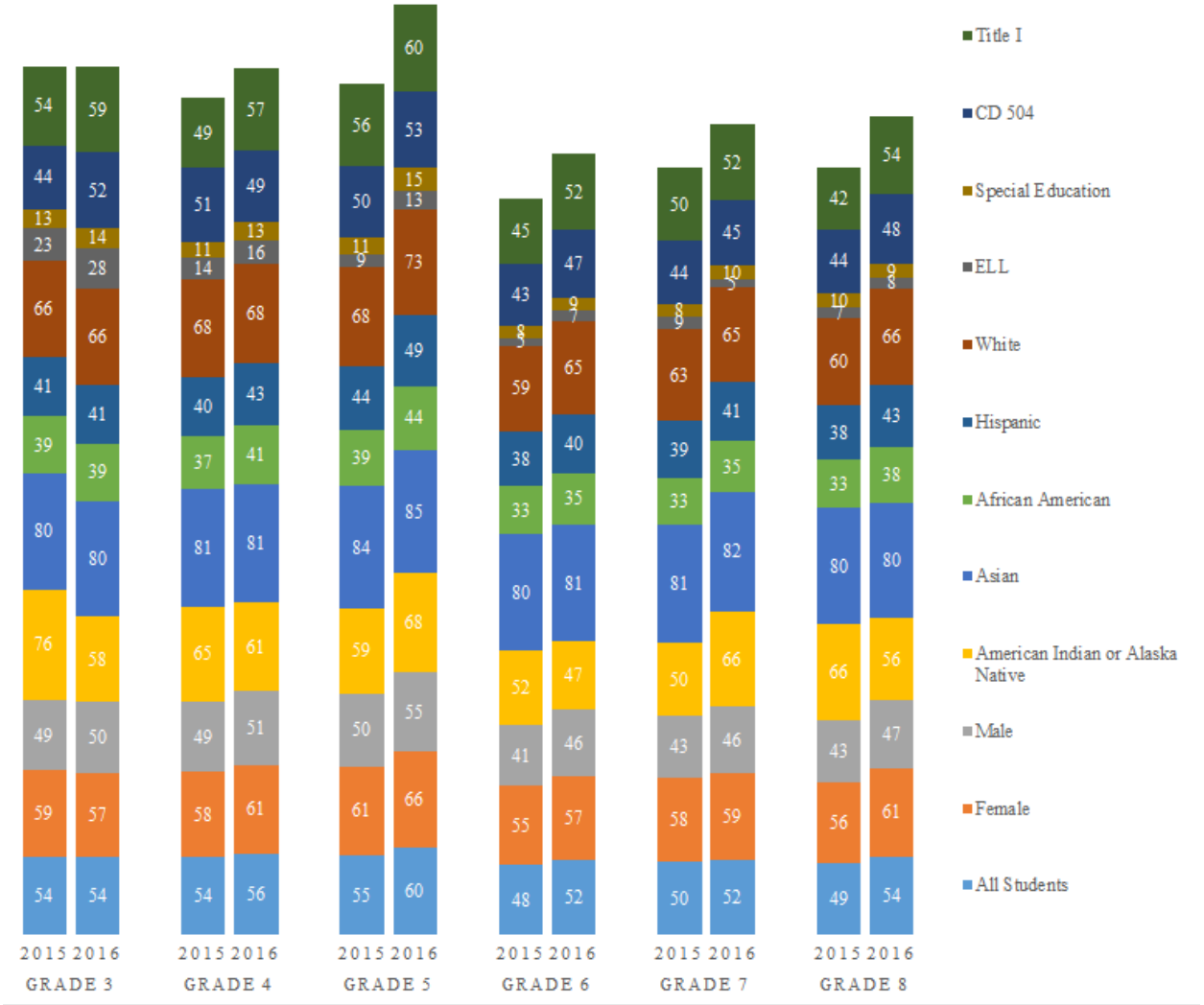
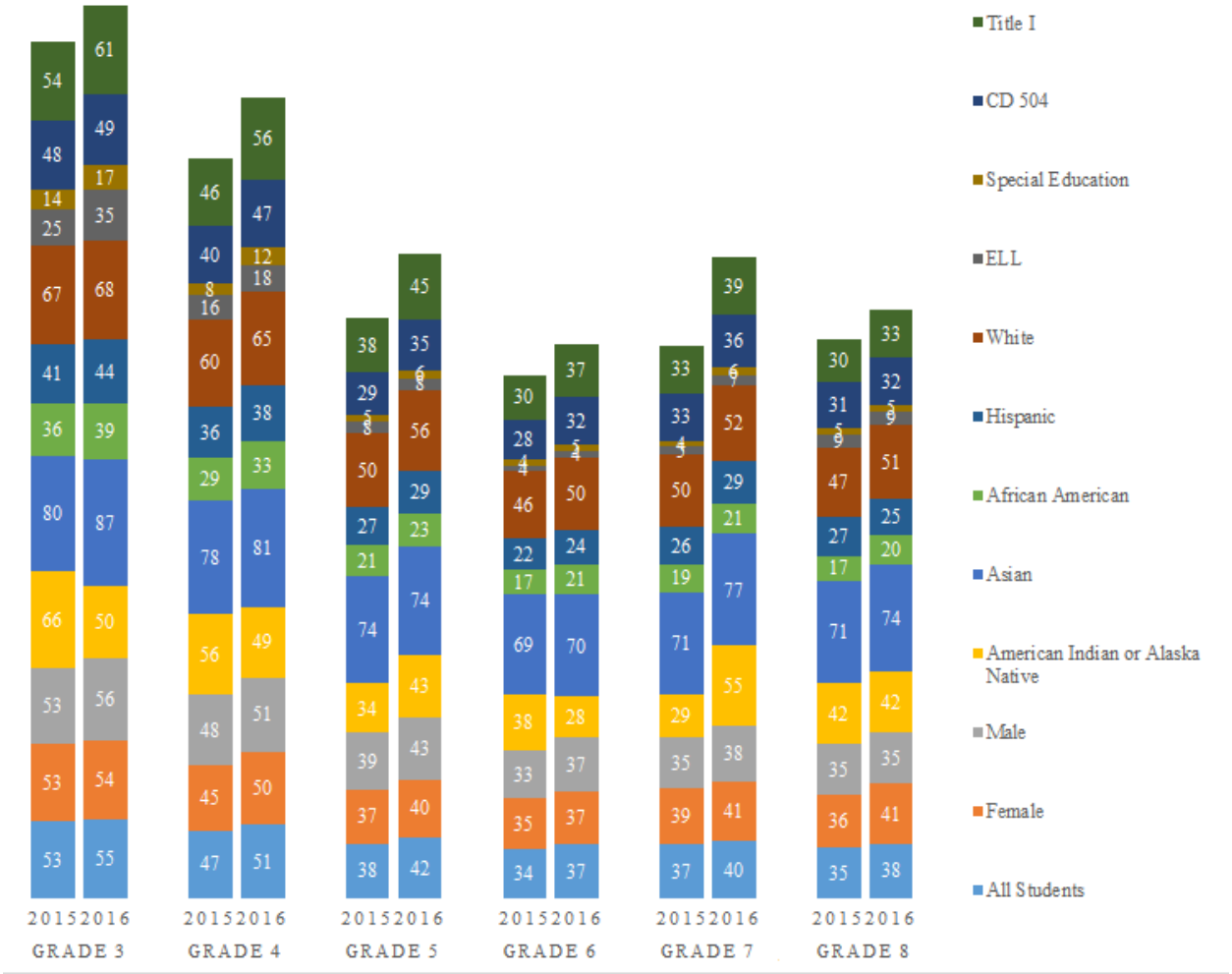


Figure 2. Mathematics %Proficient in 2014–2015 and 2015–2016



3.3 TEST TAKING TIME

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

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

Tables 19 and 20 present an average testing time and the percentage of students testing time by hourly intervals for the overall test, the CAT component, and the PT component.

Table 19. ELA/Lit Test Taking Time

Grade	Average Testing Time (hh:mm)	% Students in Each Testing Time Category							
		Less than an hour	1-2 hours	2-3 hours	3-4 hours	4-5 hours	5-6 hours	6-7 hours	More than 7 hours
Overall Test									
3	4:42	0.72	6.32	16.15	22.39	19.75	12.94	8.05	13.69
4	5:05	0.70	4.63	13.24	20.55	18.82	14.19	9.51	18.36
5	4:56	0.50	3.25	12.84	22.65	20.41	15.46	10.09	14.79
6	4:11	0.90	5.50	18.42	27.37	21.94	12.32	6.90	6.66
7	3:54	1.76	7.36	23.30	27.93	19.60	10.13	4.76	5.16
8	3:54	1.56	8.18	22.88	28.15	18.66	9.48	5.18	5.91
CAT Component									
3	2:14	5.65	44.06	32.83	11.02	3.67	1.65	0.72	0.40
4	2:24	4.66	37.76	34.98	14.23	5.06	1.98	0.74	0.60
5	2:19	3.70	40.13	36.53	13.14	4.25	1.41	0.57	0.27
6	2:06	5.56	45.40	36.81	9.39	2.06	0.50	0.14	0.13
7	2:01	7.50	49.52	32.38	7.83	1.77	0.56	0.20	0.24
8	2:00	7.48	49.96	31.66	8.13	1.96	0.43	0.18	0.19
PT Component									
3	2:28	13.66	33.60	24.42	14.48	6.75	3.19	1.59	2.30
4	2:41	10.68	29.66	26.67	15.45	8.49	4.61	2.01	2.44
5	2:37	8.29	31.05	29.71	16.76	7.31	3.37	1.70	1.81
6	2:05	14.91	39.93	27.66	10.71	4.07	1.81	0.61	0.31
7	1:53	19.96	43.34	23.60	8.05	3.04	1.17	0.42	0.44
8	1:54	20.06	42.61	23.77	8.01	3.09	1.44	0.69	0.34

Table 20. Mathematics Test Taking Time

Grade	Average Testing Time (hh:mm)	% Students in Each Testing Time Category							
		Less than an hour	1-2 hours	2-3 hours	3-4 hours	4-5 hours	5-6 hours	6-7 hours	More than 7 hours
Overall Test									
3	2:41	3.80	34.19	30.88	16.67	7.77	3.23	1.63	1.84
4	2:42	3.51	33.90	30.36	15.90	9.30	4.13	1.40	1.50
5	3:31	1.26	16.41	29.79	22.52	13.17	7.66	4.11	5.07
6	2:38	2.58	29.05	39.50	18.01	6.43	2.86	0.90	0.67
7	2:15	5.80	39.90	36.76	12.31	3.21	1.29	0.27	0.46
8	2:26	5.43	33.81	36.42	15.31	5.54	2.36	0.70	0.43
CAT Component									
3	1:45	19.74	51.01	20.24	5.77	1.93	0.77	0.22	0.33
4	1:48	18.35	49.97	20.39	7.72	2.81	0.47	0.20	0.10
5	1:56	11.87	50.98	25.06	7.69	2.89	0.91	0.37	0.24
6	1:37	14.82	62.13	19.05	3.33	0.45	0.10	0.06	0.05
7	1:36	17.97	59.84	17.87	3.12	0.79	0.20	0.11	0.10
8	1:45	15.04	54.59	22.68	5.83	1.25	0.44	0.09	0.08
PT Component									
3	0:56	66.97	26.01	5.55	1.09	0.30	0.05	0.02	0.01
4	0:55	67.86	26.33	4.53	0.92	0.19	0.12	0.05	
5	1:35	33.08	41.50	16.51	5.26	2.00	1.00	0.33	0.31
6	1:01	60.74	32.08	5.52	1.21	0.31	0.07	0.03	0.04
7	0:39	85.22	13.44	1.08	0.16	0.04	0.04	0.01	0.01
8	0:42	81.92	16.63	1.19	0.18	0.08			

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

Figures 3 and 4 display the empirical distribution of the Delaware student scale scores in the 2015–2016 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).

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

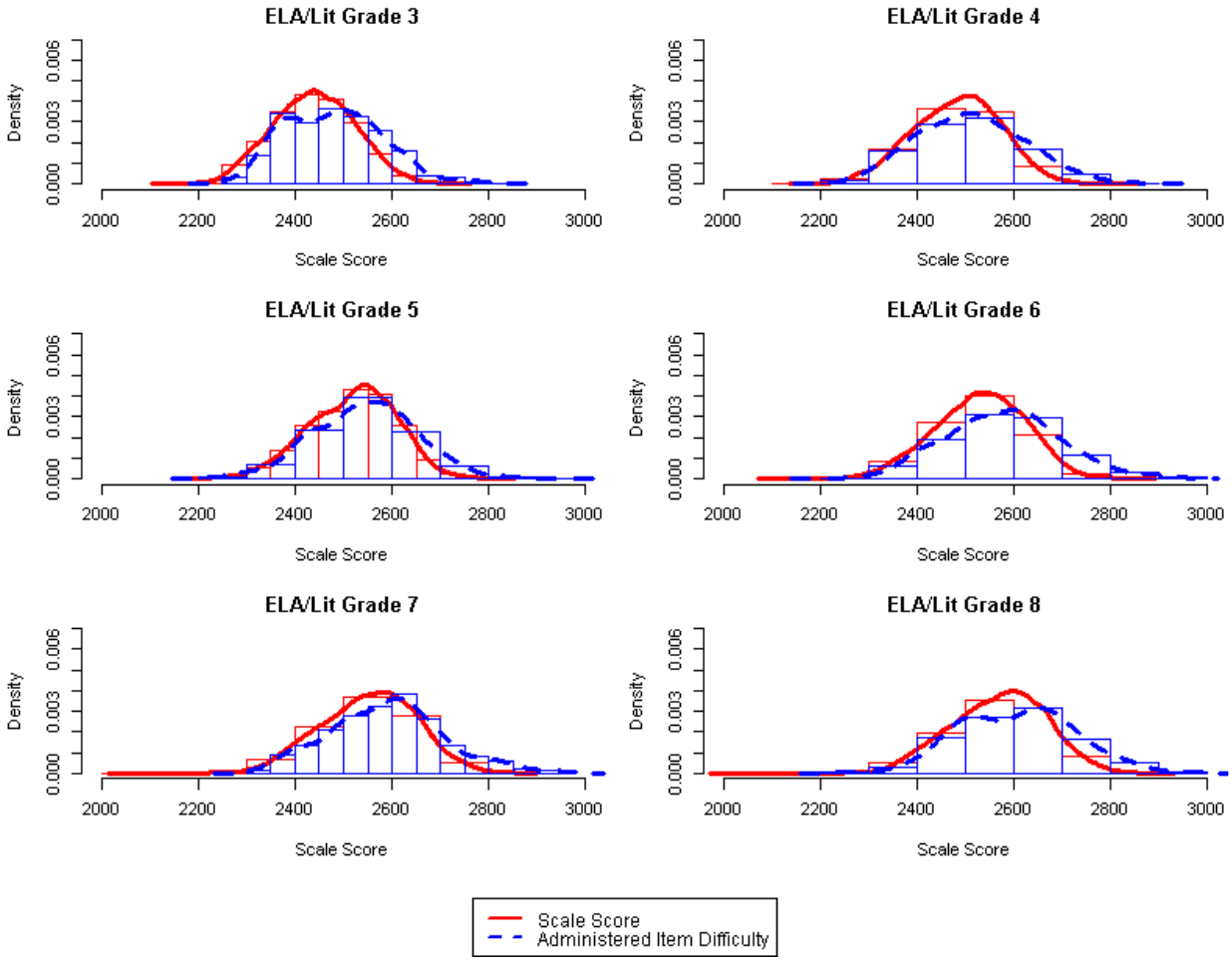
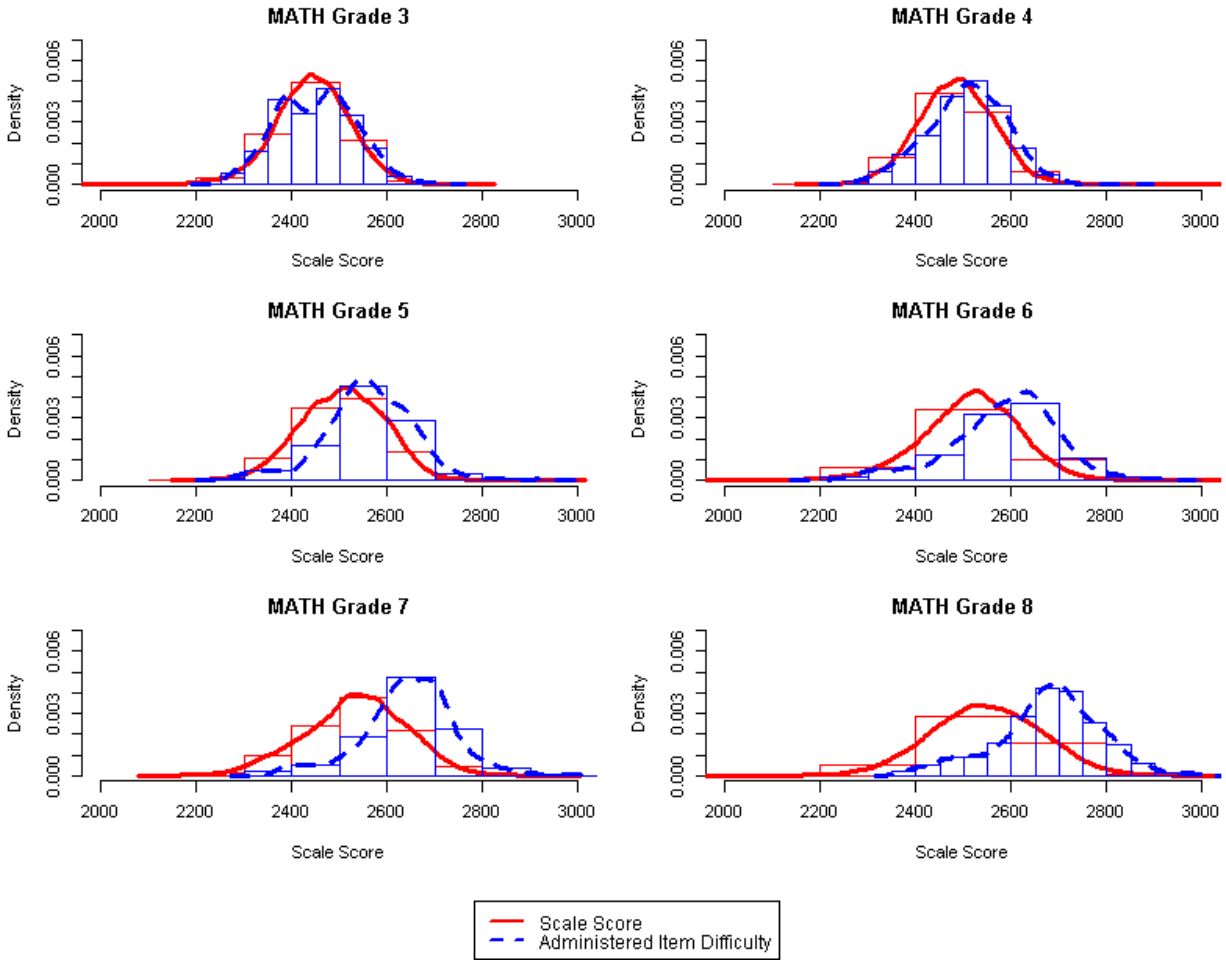


Figure 4. Student Ability–Item Difficulty Distribution for Mathematics



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
- Relations to Other Variables (External Structure)

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

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

4.1 EVIDENCE ON TEST CONTENT

The Smarter Balanced summative assessment 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 21 and 22 present the percentages of tests aligned with the test blueprint constraints for ELA/Lit CAT. Table 21 provides the blueprint match rates for item and passage requirements for each claim. For DoK and item type constraints, the Smarter Balanced blueprint specifies the minimum number of items, not the maximum. Table 22 presents the percentages of tests that satisfied the DoK and target constraints

for each claim. All tests met the requirements, except for the claim 2 DoK2 requirement in grades 3, 7, and 8, which each administered one DoK2 item fewer than required in claim 2.

Tables 23–26 provide the percentages of tests aligned with the test blueprint constraints for mathematics CAT. Tables 23–25 provide the blueprint match rates for claims and content domains within each claim. The fidelity to the DoK and target constraints is shown in Table 26. In mathematics, all tests met the blueprint requirements for claims, but there were a few exceptions in content domains. A few tests administered one item fewer or one item more than the minimum or maximum item requirement for content domains. For the DoK and target constraints, all tests satisfied the requirements, except for grade 5. In grade 5, three percent of all delivered tests administered one DoK3 or DoK4 fewer item than required in claim 2 and 4 combined.

Table 21. Percentage of ELA/Lit Delivered Tests Meeting Blueprint Requirements for Each Claim and the Number of Passages Administered

Grade	Claim	Min	Max	%BP Match for Item Requirement	%BP Match for Passage Requirement
3	1-IT	7	8	100%	100%
	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 22. ELA/Lit Percentage of Delivered Tests Meeting Blueprint Requirements for Depth-of-Knowledge

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	97%	100%	100%	100%	90%	99%
Claim 2 DoK3 or higher	1	100%	100%	100%	100%	100%	100%
Claim 2 Brief Write	1	100%	100%	100%	100%	100%	100%
Claim 3 DoK2 or higher	3	100%	100%	100%	100%	100%	100%

Table 23. Grades 3–5 Mathematics Percentage of Delivered Tests Meeting Blueprint Requirements for Each Claim and Content Domain

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

Legend:

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

Table 24. Grades 6–7 Mathematics Percentage of Delivered Tests Meeting Blueprint Requirements for Each Claim and Content Domain

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

Legend:

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

Table 25. Grade 8 Mathematics Percentage of Delivered Tests Meeting Blueprint Requirements for Each Claim and Content Domain

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

Legend:

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

Table 26. Mathematics Percentage of Delivered Tests Meeting Blueprint Requirements for Depth-of-Knowledge and Targets

DoK and Target Constraints	Minimum Required Items				%Blueprint Match					
	G3-5	G6	G7	G8	G3	G4	G5	G6	G7	G8
Segment 1										
Claim1 DOK1	5	2	3	4	100%	100%	100%	100%	100%	100%
Claim1 DOK2 or higher	7	2	4	5	100%	100%	100%	100%	100%	100%
Claim2 Target A	2	2	2	2	100%	100%	100%	100%	100%	100%
Claim2 Target B,C,D	1	1	1	1	100%	100%	100%	100%	100%	100%
Claim2/4 DOK3 or higher	2	2	2	2	100%	100%	97%	100%	100%	100%
Claim3 DOK3 or higher	2	1	2	2	100%	100%	100%	100%	100%	100%
Claim3 Target A,D	3	3	2	2	100%	100%	100%	100%	100%	100%
Claim3 Target B,E	3	2	3	3	100%	100%	100%	100%	100%	100%
Claim3 Target C,F	2				100%	100%	100%			
Claim3 Target C,F,G		2	1	1				100%	100%	100%
Claim4 Target A,D	1	1	1	1	100%	100%	100%	100%	100%	100%
Claim4 Target B,E	1	1	1	1	100%	100%	100%	100%	100%	100%
Claim4 Target C,F	1	1	1	1	100%	100%	100%	100%	100%	100%
Segment 2										
Claim1 DOK1		3	3	2				100%	100%	100%
Claim1 DOK2 or higher		5	4	4				100%	100%	100%

Table 27 summarizes the target coverage, the number of unique targets administered in each delivered test by claim. The table includes the number of targets specified in the blueprints and the mean and range of the number of targets administered to students. Since the test blueprint is not required to cover all targets in each test, it is expected that the number of targets covered varies across tests. Although the target coverage varies somewhat across individual tests, all targets are covered at an aggregate level, across all tests combined.

Table 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	10.4	4.0	1.0	3.0	8-13	3-4	1-1	3-3
4	14	5	1	3	11.1	4.0	1.0	3.0	8-13	3-5	1-1	3-3
5	14	5	1	3	10.6	4.8	1.0	3.0	8-14	4-5	1-1	3-3
6	14	5	1	3	9.5	5.0	1.0	3.0	8-11	5-5	1-1	3-3
7	14	5	1	3	10.1	4.9	1.0	3.0	8-11	4-5	1-1	3-3
8	14	5	1	3	10.2	4.0	1.0	3.0	8-11	3-4	1-1	3-3
Mathematics												
3	11	4	6	6	10.0	2.0	5.2	3.0	9-10	2-2	3-6	3-3
4	12	4	6	6	10.0	2.0	5.6	3.0	10-10	2-2	3-6	3-3
5	11	4	6	6	9.0	2.0	5.5	3.0	8-9	2-2	3-6	3-3
6	10	4	7	6	9.9	2.0	4.3	3.0	9-10	2-2	3-6	3-3
7	9	3	7	6	8.0	2.0	4.5	3.0	8-8	2-2	3-6	3-3
8	10	4	7	6	10.0	2.0	5.3	3.0	10-10	2-2	3-6	3-3

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

4.2 EVIDENCE ON INTERNAL STRUCTURE

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

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

Table 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		0.86	0.93	0.88
	Claim 2: Writing	0.67		0.88	0.87
	Claim 3: Listening	0.62	0.60		0.89
	Claim 4: Research	0.64	0.65	0.57	
4	Claim 1: Reading		0.88	0.92	0.92
	Claim 2: Writing	0.68		0.84	0.87
	Claim 3: Listening	0.61	0.56		0.91
	Claim 4: Research	0.68	0.64	0.57	
5	Claim 1: Reading		0.85	0.93	0.91
	Claim 2: Writing	0.66		0.85	0.89
	Claim 3: Listening	0.62	0.58		0.90
	Claim 4: Research	0.66	0.66	0.57	
6	Claim 1: Reading		0.88	0.90	0.93
	Claim 2: Writing	0.67		0.85	0.92
	Claim 3: Listening	0.60	0.59		0.89
	Claim 4: Research	0.65	0.67	0.57	
7	Claim 1: Reading		0.91	0.94	0.95
	Claim 2: Writing	0.70		0.89	0.92
	Claim 3: Listening	0.61	0.59		0.93

	Claim 4: Research	0.69	0.69	0.58	
8	Claim 1: Reading		0.90	0.91	0.92
	Claim 2: Writing	0.71		0.88	0.90
	Claim 3: Listening	0.61	0.60		0.88
	Claim 4: Research	0.68	0.68	0.56	

Table 29. Correlations Among Reporting Categories for Mathematics

Grade	Reporting Categories	Observed & Disattenuated Correlation		
		Claim 1	Claim 2&4	Claim 3
3	Claim 1		0.96	0.95
	Claim 2 & 4	0.78		1
	Claim 3	0.75	0.72	
4	Claim 1		0.96	0.96
	Claim 2 & 4	0.79		0.99
	Claim 3	0.79	0.75	
5	Claim 1		1	0.95
	Claim 2 & 4	0.79		1
	Claim 3	0.76	0.74	
6	Claim 1		1	0.97
	Claim 2 & 4	0.79		1
	Claim 3	0.75	0.72	
7	Claim 1		1	0.97
	Claim 2 & 4	0.80		1
	Claim 3	0.78	0.72	
8	Claim 1		1	1
	Claim 2 & 4	0.75		1
	Claim 3	0.76	0.70	

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

4.3 EVIDENCE ON RELATIONS TO OTHER VARIABLES

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

The convergent and discriminant validities were examined based on the relationships between ELA/Lit and mathematics scores in 2014–2015 and 2015–2016. It was expected that the correlation between two tests measuring the same content (correlations between ELA/Lit scores) would be higher than the

correlation between tests measuring different contents (correlations between ELA/Lit and mathematics scores).

In Table 30, the reliability coefficients are in boldface on diagonal, the correlations between students' scores for the same subject in two years are underlined (convergent validity), and the correlations between ELA/Lit and mathematics scores within and between years are in rectangles (discriminant validity). The correlations between two grades for the same subject and between subjects for different grades are computed for grades 4–8 only because grade 3 does not have a lower grade score to correlate with.

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

The correlations for the same subject scores in two different grades were higher than the correlations between two subject scores. The correlation coefficients for the same subject scores ranged from 0.81 to 0.84 for ELA/Lit and from 0.83 to 0.87 for mathematics. The correlation between ELA/Lit and mathematics scores ranged from 0.74 to 0.81. The observed pattern of correlations within and between subjects conforms to the criteria expected for convergent and discriminant validity.

Table 30. Relationships Between ELA/Lit and Mathematics Scores

Grade	Year/Subject	N	2015 ELA/Lit	2016 ELA/Lit	2015 Math	2016 Math
3	2015 ELA/Lit	10,194	**0.92**			
	2016 ELA/Lit	10,273	n/a	**0.92**		
	2015 Math	10,194	0.80	n/a	**0.94**	
	2016 Math	10,273	n/a	0.79	n/a	**0.94**
4	2015 ELA/Lit	9,581	**0.91**			
	2016 ELA/Lit	9,581	0.81	**0.92**		
	2015 Math	9,610	0.79	0.75	**0.94**	
	2016 Math	9,610	0.74	0.80	0.84	**0.94**
5	2015 ELA/Lit	9,294	**0.92**			
	2016 ELA/Lit	9,294	0.83	**0.92**		
	2015 Math	9,362	0.80	0.76	**0.93**	
	2016 Math	9,362	0.76	0.80	0.86	**0.94**
6	2015 ELA/Lit	9,204	**0.91**			
	2016 ELA/Lit	9,204	0.82	**0.91**		
	2015 Math	9,273	0.79	0.74	**0.92**	
	2016 Math	9,273	0.77	0.80	0.83	**0.93**
7	2015 ELA/Lit	9,294	**0.92**			
	2016 ELA/Lit	9,294	0.83	**0.92**		
	2015 Math	9,402	0.80	0.77	**0.91**	
	2016 Math	9,402	0.77	0.81	0.87	**0.94**
8	2015 ELA/Lit	9,002	**0.92**			
	2016 ELA/Lit	9,002	0.84	**0.92**		
	2015 Math	9,069	0.79	0.77	**0.91**	
	2016 Math	9,069	0.75	0.79	0.85	**0.92**

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

Table 31. 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	10,296	41	44	0.92	2439.54	85.41	24.60
4	10,268	40	44	0.92	2482.47	90.77	26.31
5	10,169	41	45	0.92	2519.27	89.98	25.98
6	9,983	41	45	0.91	2530.18	93.45	27.49
7	10,049	41	45	0.92	2552.72	98.23	28.18
8	9,747	43	45	0.92	2569.63	98.07	27.72
Mathematics							
3	10,341	39	40	0.94	2443.99	78.56	19.40
4	10,297	37	40	0.94	2485.05	79.43	19.25
5	10,199	38	40	0.94	2506.76	86.83	21.89
6	10,004	38	39	0.93	2516.33	101.78	26.38
7	10,070	38	40	0.94	2534.46	106.55	26.75
8	9,768	38	40	0.92	2548.93	116.98	33.26

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

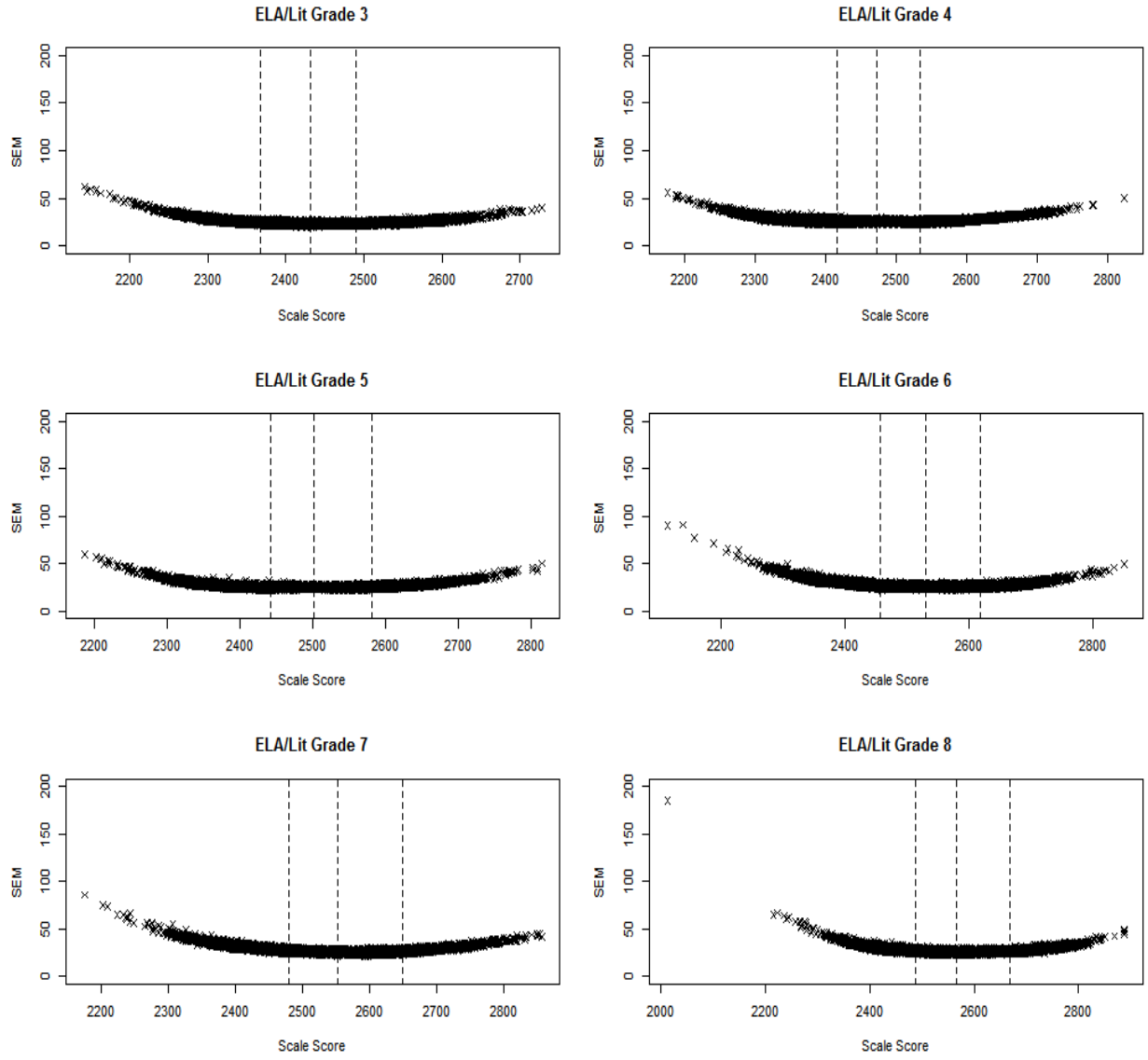
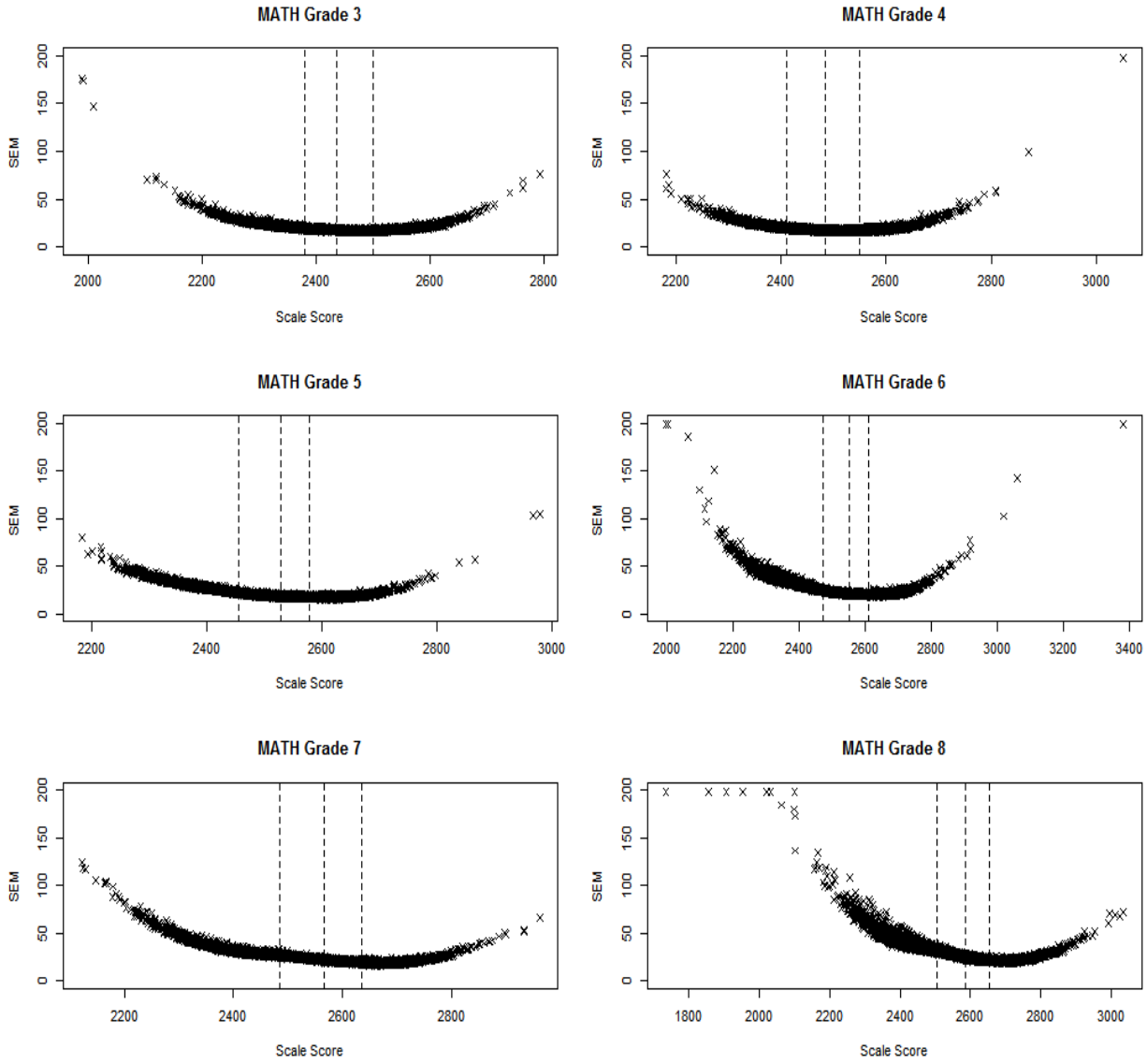


Figure 6. Conditional Standard Error of Measurement for Mathematics



The SEMs presented in the figures above are summarized in Tables 32 and 33. Table 32 provides the average conditional SEM for all scores and scores in each achievement level. Table 33 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 32. Average Conditional Standard Error of Measurement by Achievement Levels

Grade	Level 1	Level 2	Level 3	Level 4	Average CSEM
ELA/Lit					
3	27.58	23.24	23.02	24.88	24.60
4	27.74	25.23	25.14	26.80	26.31
5	27.19	24.82	24.87	27.23	25.98
6	31.48	25.97	25.80	27.57	27.49
7	32.31	26.58	25.80	29.12	28.18
8	31.95	26.11	25.90	28.48	27.72
Mathematics					
3	24.21	18.30	17.02	19.03	19.40
4	23.70	18.00	16.97	19.99	19.25
5	27.94	19.87	18.04	18.66	21.89
6	33.93	22.63	20.60	22.98	26.38
7	35.03	24.62	20.77	20.01	26.75
8	43.95	28.97	23.36	23.11	33.26

Table 33. 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	23.71	22.73	23.10	0.98	0.37	0.61
4	25.57	25.18	25.07	0.39	0.11	0.50
5	24.93	24.78	25.28	0.15	0.50	0.35
6	26.22	25.64	26.24	0.58	0.60	0.02
7	27.54	26.38	26.45	1.16	0.07	1.09
8	26.83	25.59	26.70	1.24	1.11	0.13
Mathematics						
3	19.49	17.55	16.80	1.94	0.75	2.69
4	19.60	16.93	17.43	2.67	0.50	2.17
5	22.01	18.48	17.89	3.53	0.59	4.12
6	24.52	21.10	20.52	3.42	0.58	4.00
7	27.61	22.38	19.11	5.23	3.27	8.50
8	32.01	25.41	21.83	6.60	3.58	10.18

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

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

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

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

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

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

where d stands for dichotomous and p stands for polytomous items; $\mathbf{b}_j = (a_j, b_j, c_j)$ 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 *i*th student being classified at achievement level *l* and *m*, respectively based on observed scores and hypothetical scores from equivalent test form.

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

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

and the overall classification consistency is

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

The analysis of the classification index is performed based on overall scale scores. Table 34 provides the percentage of classification accuracy and consistency for overall and by achievement level.

The overall classification index ranged from 78% to 83% for the accuracy and from 70% to 76% 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 slightly higher than the consistency of classifications in all achievement levels. The consistency of classification rates can be lower because the consistency is based on two tests with measurement errors while the accuracy is based on one test with a measurement error and the true score.

Table 34. Classification Accuracy and Consistency by Achievement Levels

Grade	Achievement Level	ELA/Lit		Mathematics	
		% Accuracy	% Consistency	% Accuracy	% Consistency
3	Overall	79	71	82	75
	L1	87	80	88	81
	L2	71	61	73	64
	L3	68	58	79	72
	L4	88	83	89	84
4	Overall	78	70	83	76
	L1	89	82	88	80
	L2	65	53	81	74
	L3	67	56	79	71
	L4	88	82	89	84
5	Overall	79	71	82	75
	L1	88	81	89	84
	L2	67	55	78	70
	L3	76	68	71	61
	L4	86	79	89	84
6	Overall	79	71	82	75
	L1	88	81	91	85
	L2	72	62	77	70
	L3	76	68	71	61
	L4	84	77	88	82
7	Overall	80	72	83	76
	L1	89	83	91	85
	L2	71	60	77	69
	L3	79	72	75	66
	L4	84	76	90	85
8	Overall	81	73	81	74
	L1	88	82	90	84
	L2	74	64	72	62
	L3	79	73	72	61
	L4	84	75	90	85

5.4 RELIABILITY FOR SUBGROUPS

The reliability of test scores and achievement levels are also computed by subgroups. Tables 35 and 36 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. The classification indexes by subgroups are provided in Appendix C. The smallest sample size across subgroups is 40 for American Indian/Alaska Native.

Table 35. 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	0.92	0.92	0.92	0.91	0.92	0.92
Female	0.92	0.91	0.91	0.91	0.92	0.91
Male	0.92	0.91	0.92	0.91	0.91	0.92
American Indian/Alaska Native	0.91	0.90	0.89	0.90	0.89	0.93
Asian	0.92	0.91	0.89	0.91	0.91	0.91
African American	0.90	0.91	0.91	0.90	0.90	0.91
Hispanic	0.90	0.90	0.91	0.90	0.91	0.91
White	0.91	0.90	0.90	0.91	0.91	0.91
English Language Learners	0.87	0.87	0.87	0.82	0.79	0.82
Special Education	0.85	0.86	0.87	0.84	0.83	0.84
CD 504	0.90	0.89	0.89	0.90	0.90	0.90
Title I	0.90	0.89	0.90	0.90	0.91	0.91

Table 36. 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	0.93	0.94	0.92
Female	0.94	0.94	0.93	0.93	0.94	0.92
Male	0.94	0.94	0.94	0.93	0.94	0.92
American Indian/Alaska Native	0.94	0.92	0.93	0.92	0.93	0.91
Asian	0.93	0.93	0.94	0.94	0.95	0.95
African American	0.92	0.93	0.91	0.91	0.91	0.88
Hispanic	0.92	0.93	0.92	0.91	0.92	0.89
White	0.94	0.94	0.94	0.93	0.94	0.92
English Language Learners	0.91	0.90	0.87	0.81	0.85	0.77
Special Education	0.91	0.90	0.85	0.85	0.85	0.77
CD 504	0.93	0.94	0.92	0.91	0.92	0.90
Title I	0.93	0.93	0.93	0.93	0.93	0.91

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 achievement categories, taking into account the SEM of the claim score: (1) Below standard, (2) At/Near standard, or (3) Above standard. Tables 37 and 38 present the marginal reliability coefficients for each claim score in ELA/Lit and mathematics, respectively.

Table 37. 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	0.76	2434.59	103.33	50.97
	Claim 2: Writing	11	11	0.79	2436.95	98.37	45.34
	Claim 3: Listening	8	8	0.58	2448.29	120.07	77.69
	Claim 4: Research	8	9	0.70	2437.29	118.21	65.21
4	Claim 1: Reading	14	16	0.78	2477.54	108.09	50.13
	Claim 2: Writing	11	11	0.78	2482.02	104.47	49.30
	Claim 3: Listening	8	8	0.57	2495.59	136.45	89.33
	Claim 4: Research	8	9	0.69	2476.54	121.69	68.13
5	Claim 1: Reading	14	16	0.76	2509.74	110.03	53.65
	Claim 2: Writing	11	11	0.79	2517.01	104.34	47.85
	Claim 3: Listening	8	9	0.59	2510.28	136.08	86.98
	Claim 4: Research	9	9	0.69	2539.18	110.97	62.27
6	Claim 1: Reading	14	16	0.73	2506.65	122.70	63.76
	Claim 2: Writing	11	11	0.80	2527.33	101.77	45.78
	Claim 3: Listening	8	9	0.61	2554.38	152.98	95.70
	Claim 4: Research	8	9	0.66	2542.82	117.05	68.24
7	Claim 1: Reading	14	16	0.76	2548.30	119.40	58.58
	Claim 2: Writing	11	11	0.79	2547.58	113.33	51.89
	Claim 3: Listening	8	9	0.56	2564.99	139.22	91.98
	Claim 4: Research	8	9	0.70	2551.47	128.74	69.98
8	Claim 1: Reading	16	16	0.78	2564.41	114.18	53.74
	Claim 2: Writing	11	11	0.80	2566.42	116.35	51.89
	Claim 3: Listening	8	9	0.58	2580.03	135.48	87.44
	Claim 4: Research	9	9	0.70	2569.18	128.30	70.73

Table 38. 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	0.89	2445.22	85.29	28.08
	Claims 2 & 4	8	11	0.73	2436.05	90.07	46.47
	Claim 3	9	11	0.70	2444.78	98.22	54.22
4	Claim 1	20	20	0.89	2486.04	86.30	28.85
	Claims 2 & 4	8	10	0.75	2480.13	92.78	46.53
	Claim 3	9	10	0.76	2484.04	94.38	46.23
5	Claim 1	20	20	0.89	2507.13	92.88	31.37
	Claims 2 & 4	9	10	0.71	2501.76	98.07	52.92
	Claim 3	9	10	0.71	2500.45	109.79	59.10
6	Claim 1	19	19	0.89	2515.41	110.26	37.34
	Claims 2 & 4	9	10	0.71	2512.54	110.49	59.81
	Claim 3	10	11	0.66	2511.79	122.41	70.94
7	Claim 1	20	20	0.89	2537.17	113.28	37.51
	Claims 2 & 4	10	11	0.66	2518.34	126.38	74.02
	Claim 3	8	10	0.72	2522.82	134.49	71.37
8	Claim 1	20	20	0.87	2546.72	126.47	45.93
	Claims 2 & 4	8	10	0.53	2545.40	131.28	90.01
	Claim 3	9	10	0.65	2540.00	137.18	80.72

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 hand-scoring procedure.

6.1 ESTIMATING STUDENT ABILITY USING MAXIMUM LIKELIHOOD ESTIMATION

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

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

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

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

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

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

in the case of items with two or more points,

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

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

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:

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

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 Smarter Balanced Assessment Consortium. Table 39 presents the scaling constants for each subject for the theta-to-scale score linear transformation. Scale scores are rounded to an integer.

Table 39. 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 40 provides three achievement standards for each grade and content area.

Table 40. 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 truncated extreme unreliable 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. For the 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, 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 HUMAN SCORING

AIR provides the automated electronic scoring and Measurement Incorporated (MI) provides all hand-scoring 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 "hand-scored." In mathematics, SA items and other constructed-response items are hand-scored. The procedure for scoring these items is provided by Smarter Balanced.

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

6.6.1 Reader Selection

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

MI procedures for selecting new readers are very thorough. After advertising and receiving applications, MI staff review the applications and schedule interviews for qualified applicants (i.e., those with a four-year college degree). Each qualified applicant must pass an interview by experienced MI staff, complete ELA/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.

6.6.2 Reader Training

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

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

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

Regardless of mode, the same training protocol is followed.

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

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

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

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

Training is carefully orchestrated so that readers understand how to apply the rubric in scoring the responses, reference the scoring guide, develop the flexibility needed to handle a variety of responses, and

retain the consistency needed to score all responses accurately. In addition to completing all of the initial training and qualifications, significant time is allotted for demonstrations of the VSC hand-scoring system, explanations of how to "flag" unusual responses for review by the scoring director, and instructions about other procedures necessary for the conduct of a smooth project.

Training design varies slightly depending on Smarter Balanced item type:

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

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

6.6.3 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. Unbiased scoring is ensured because the only identifying information on the student response is the identification number. Unless the students sign their names, write about their hometowns, or in some way provide other identifying information, the readers have no knowledge of them.

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

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

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

- Reader ID and team
- Number of responses scored

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

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

6.6.4 Reader Monitoring and Retraining

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

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

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

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

6.6.5 Reader Validity Checks

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

6.6.6 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.

6.6.7 Reader Agreements

The inter-reader reliability is computed based on scorable responses (numeric scores) scored by two independent readers only, excluding non-scorable responses (e.g., off topic, off purpose, or foreign language responses) which were scored by the leadership readers, not by two independent readers. The inter-reader reliability is based on the combined data across 10 states (Delaware, Hawaii, Idaho, New Hampshire, Oregon, South Dakota, Vermont, Washington, West Virginia, and Connecticut) and the U.S. Virgin Islands because the number of responses with two independent readers, after removing responses with condition codes, is too small to compute inter-reader reliability by state.

In ELA/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 hand-scored items range from 1–3.

In an adaptive test, because items are selected adapting to a student's ability while meeting the test blueprint, item usages vary across items. Tables 41–43 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, and quadratic weighted Kappa (QWK).

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

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

Table 42. ELA/Lit Reader Agreements for Full Write Items

Grade	Dimensions	# of Items	%Exact			%(Exact+ Adjacent)	QWK
			Average	Min	Max		
3	Conventions	10	60	50	65	98	0.54
	Evid/Elab	10	68	62	77	98	0.63
	Org/Purp	10	66	61	75	98	0.62
4	Conventions	14	60	45	66	97	0.62
	Evid/Elab	14	62	55	78	98	0.65
	Org/Purp	14	63	54	77	98	0.65
5	Conventions	19	63	53	72	97	0.52
	Evid/Elab	19	60	53	68	97	0.68
	Org/Purp	19	61	55	69	97	0.69
6	Conventions	13	70	64	86	97	0.64
	Evid/Elab	13	64	56	73	98	0.70
	Org/Purp	13	64	56	74	98	0.71
7	Conventions	17	70	65	74	99	0.63
	Evid/Elab	17	69	55	79	99	0.74
	Org/Purp	17	69	55	77	99	0.74
8	Conventions	18	76	66	84	99	0.58
	Evid/Elab	18	67	63	69	99	0.73
	Org/Purp	18	67	64	72	99	0.73

Table 43. Mathematics Reader Agreements

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

7. REPORTING AND INTERPRETING SCORES

The Online Reporting System (ORS) generates a set of online score reports that include the information describing student performance for students, parents, educators, and other stakeholders. The online score reports are produced immediately after students complete the test with hand-scored items. Because the score report on student performance are updated each time students complete tests and they are hand scored, authorized users (e.g., school principals, teachers) can view students' performance on the tests and use them to improve student learning. In addition to 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.

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. In order to view student performance for a specific aggregate unit, users can select the specific aggregate unit from a drop-down menu with a list of aggregate units (e.g., schools within a districts, or teachers within a school) to select. For more detailed student assessment results for a school, a teacher, or a roster, users can select the subject and grade on the online score reports.

Generally, the ORS provides two categories of online score reports: (1) aggregate score reports and (2) student score reports. Table 44 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 44. 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 performance growth over time

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

Table 45. 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 provides the summaries of students' performance including (1) number of students tested, and (2) percentage of students at Level 3 or above. Exhibits 1 and 2 present a sample of home pages at the state level and the district level, respectively.

Exhibit 1. Home Page: State Level

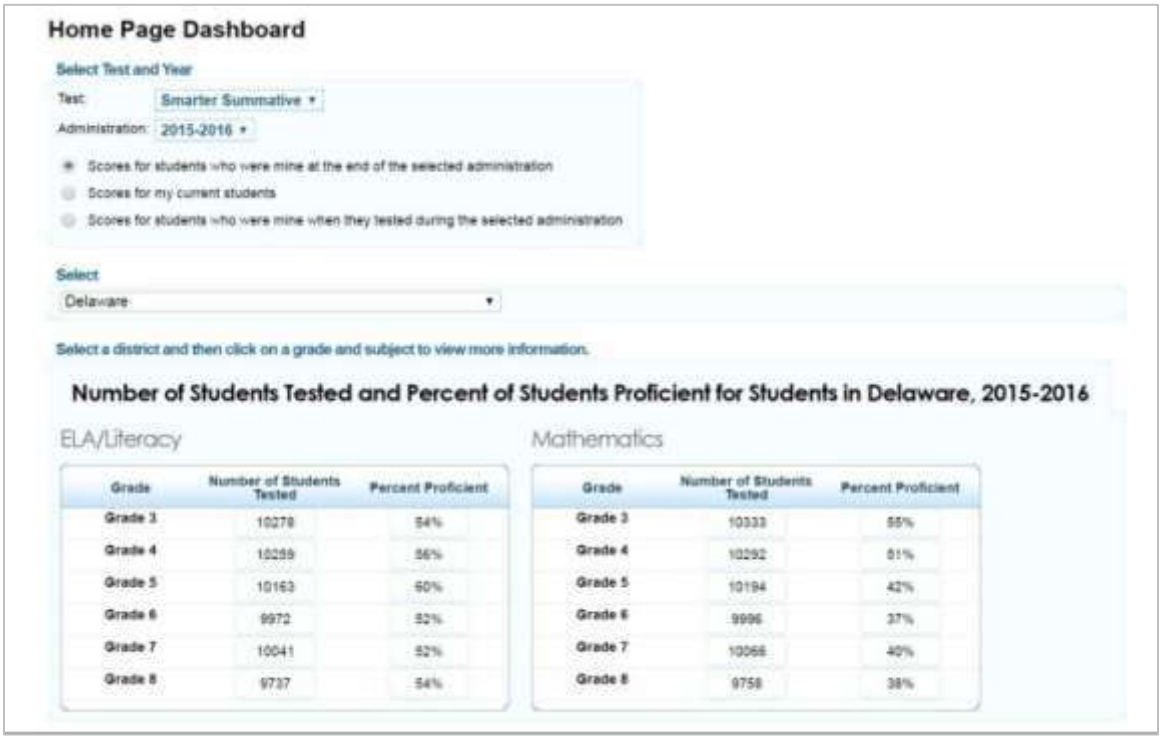
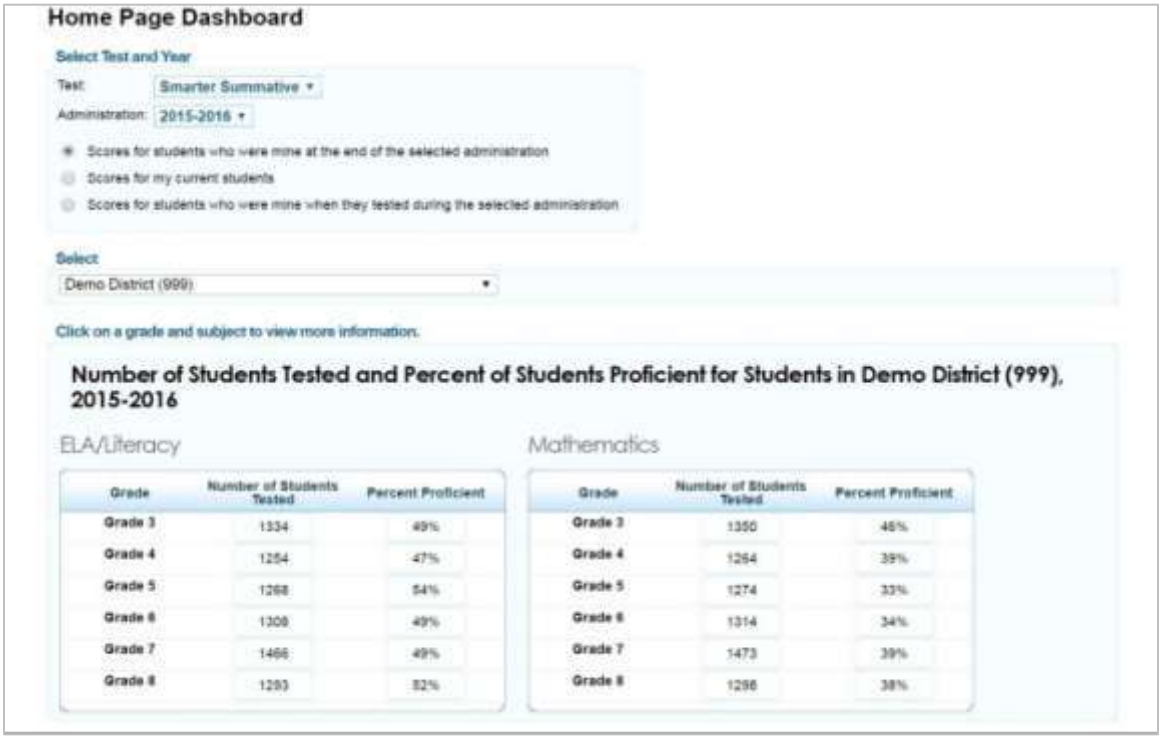


Exhibit 2. Home Page: District Level



7.1.2.2 Subject Detail Page

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

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

Exhibit 3. Subject Detail Page for ELA/Lit by Gender: District Level

Student Performance in Each Achievement Level

How did my district perform overall in ELA/Literacy?

Test: Smarter Summative ELA/Literacy Grade 6
Year: 2015-2016
Name:Demo District 99

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 99

Breakdown By: Gender | Test Event: ALL | GO | Comparison: ON

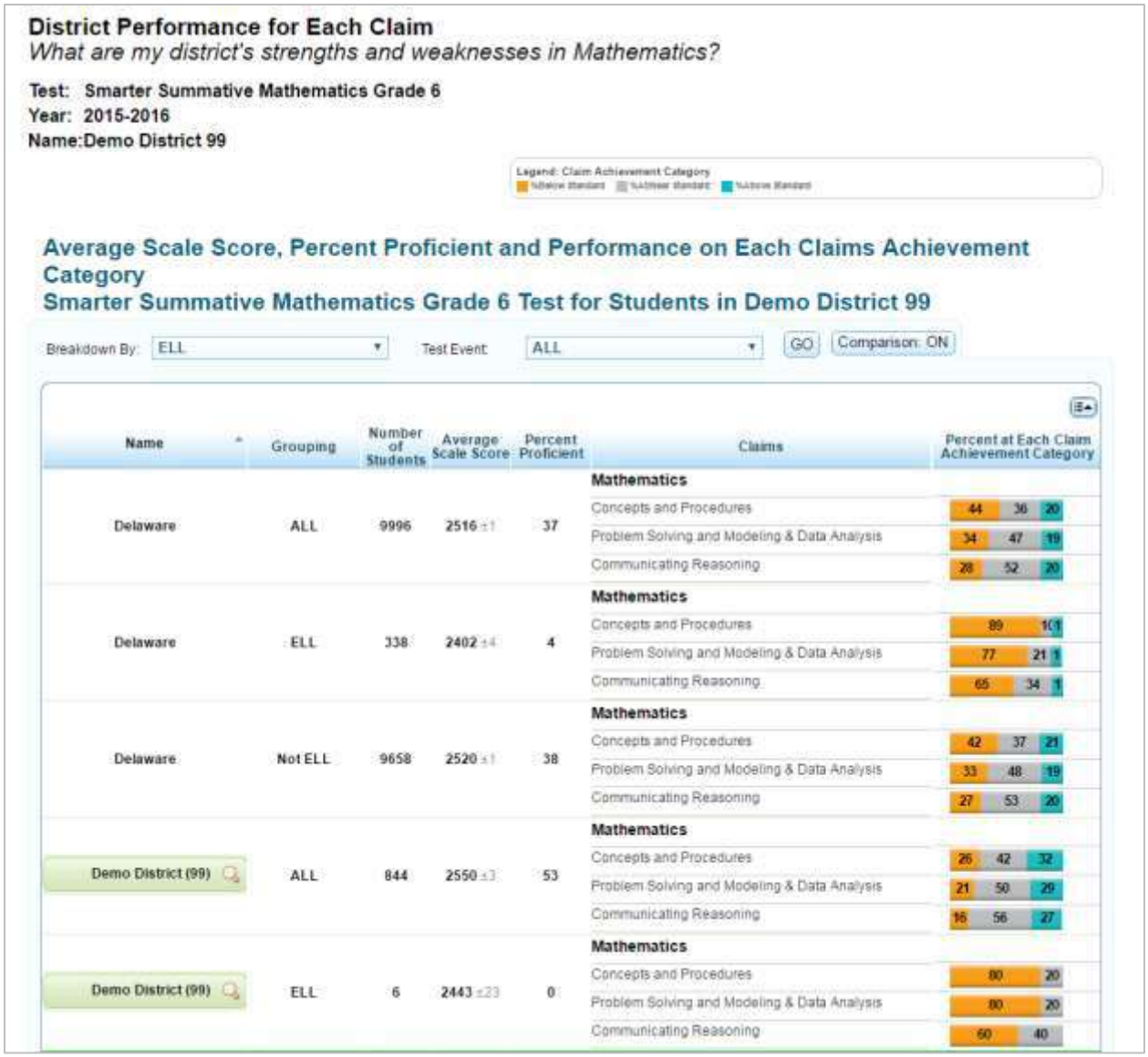
Name	Grouping	Number of Students	Average Scale Score	Percent Proficient	Percentage in Each Achievement Level
Delaware	ALL	9972	2530 ±1	52	22 27 33 18
Delaware	Female	4919	2544 ±1	57	17 26 35 22
Delaware	Male	5053	2516 ±1	46	26 28 31 15
Demo District (99)	ALL	851	2557 ±3	66	12 22 42 24
Demo District (99)	Female	404	2566 ±4	71	9 20 44 27
Demo District (99)	Male	447	2549 ±4	62	14 24 40 21
Demo Institution (01)	ALL	273	2541 ±5	59	17 24 42 17
Demo Institution (01)	Female	136	2557 ±6	69	10 21 51 18
Demo Institution (01)	Male	137	2524 ±8	49	23 28 33 16
Demo Institution (02)	ALL	256	2554 ±5	63	12 25 38 25
Demo Institution (02)	Female	137	2557 ±7	65	10 25 40 25
Demo Institution (02)	Male	119	2550 ±8	61	14 24 36 25
Demo Institution (03)	ALL	322	2574 ±4	74	7 19 45 29
Demo Institution (03)	Female	131	2584 ±7	78	8 14 40 38
Demo Institution (03)	Male	191	2566 ±6	71	7 22 48 23

7.1.2.3 Claim Detail Page

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

Similar to the subject detail page, the summary report presents the summary results for the selected aggregate unit as well as the summary results for the state and the aggregate unit above the selected aggregate. Also, the summaries on claim-level performance can be presented for overall students and by subgroup. Exhibit 4 presents an example of claim detail pages for mathematics at the district level when users select a subgroup of ELL.

Exhibit 4. Claim Detail Page for Mathematics by ELL: District Level



District Performance for Each Claim

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

Test: Smarter Summative Mathematics Grade 6
Year: 2015-2016
Name:Demo District 99

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 99

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	9996	2516 ±1	37	**Mathematics**	
					Concepts and Procedures	44 36 20
					Problem Solving and Modeling & Data Analysis	34 47 19
					Communicating Reasoning	28 52 20
Delaware	ELL	338	2402 ±4	4	**Mathematics**	
					Concepts and Procedures	89 10 1
					Problem Solving and Modeling & Data Analysis	77 21 1
					Communicating Reasoning	65 34 1
Delaware	Not ELL	9658	2520 ±1	38	**Mathematics**	
					Concepts and Procedures	42 37 21
					Problem Solving and Modeling & Data Analysis	33 48 19
					Communicating Reasoning	27 53 20
Demo District (99)	ALL	844	2550 ±3	53	**Mathematics**	
					Concepts and Procedures	26 42 32
					Problem Solving and Modeling & Data Analysis	21 50 29
					Communicating Reasoning	16 56 27
Demo District (99)	ELL	6	2443 ±23	0	**Mathematics**	
					Concepts and Procedures	80 20
					Problem Solving and Modeling & Data Analysis	80 20
					Communicating Reasoning	60 40

7.1.2.4 Student Detail Page

When a student completes a test and the test is hand-scored, an online score report appears in the student detail page in the ORS. The student detail page provides individual student performance on the test. In each subject area, the student detail page provides (1) scale score and standard error of measurement, (2) achievement level for overall test, (3) achievement 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. In addition, student performance on each reporting category is displayed along with a description of his or her performance on each claim. On the bottom of the page, student performance growth over time (i.e., year) is presented to show student performance change across school year.

Exhibits 5 and 6 present examples of student detail pages for ELA/Lit and mathematics.

Exhibit 5. Student Detail Page for ELA/Lit

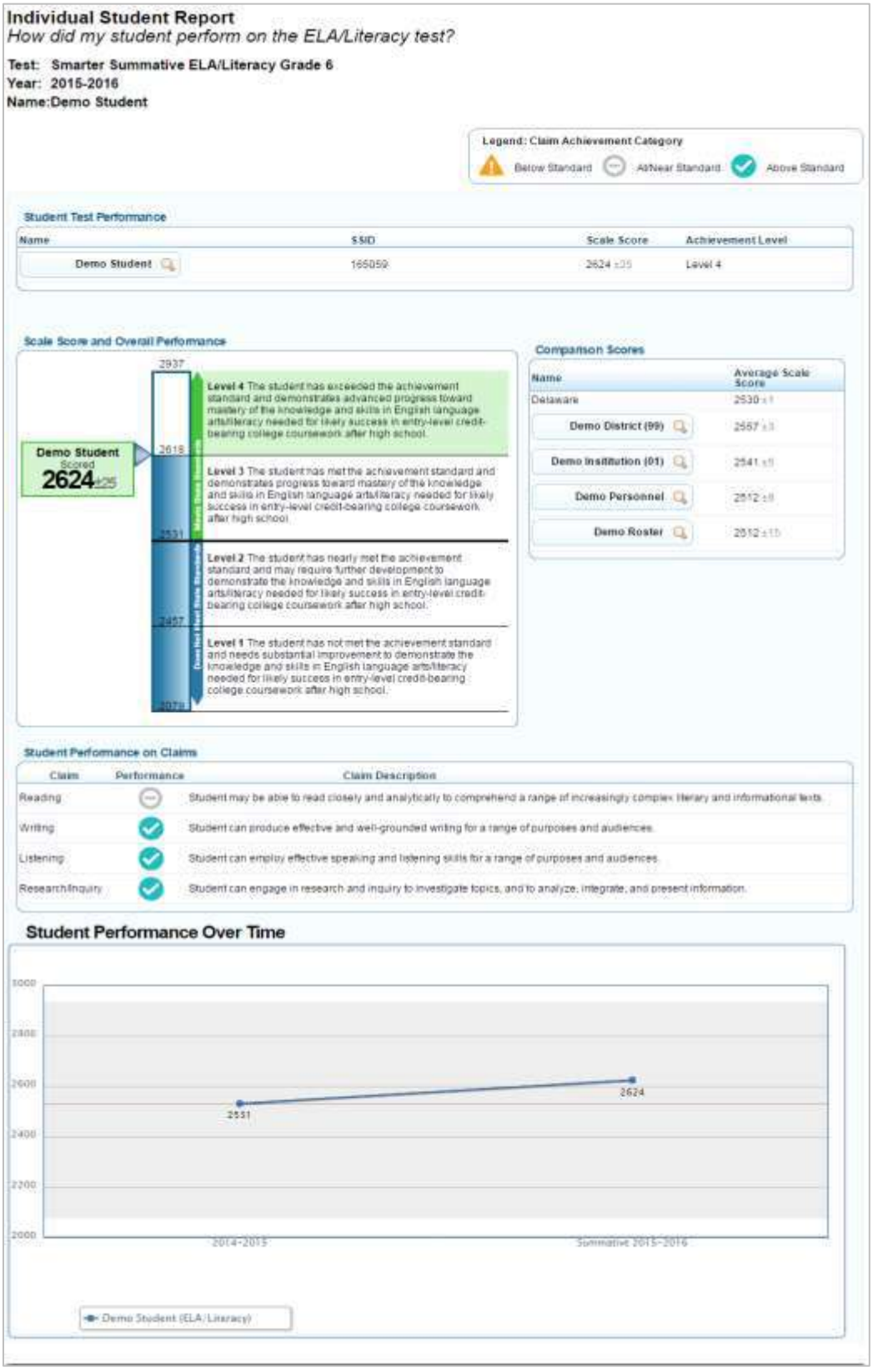
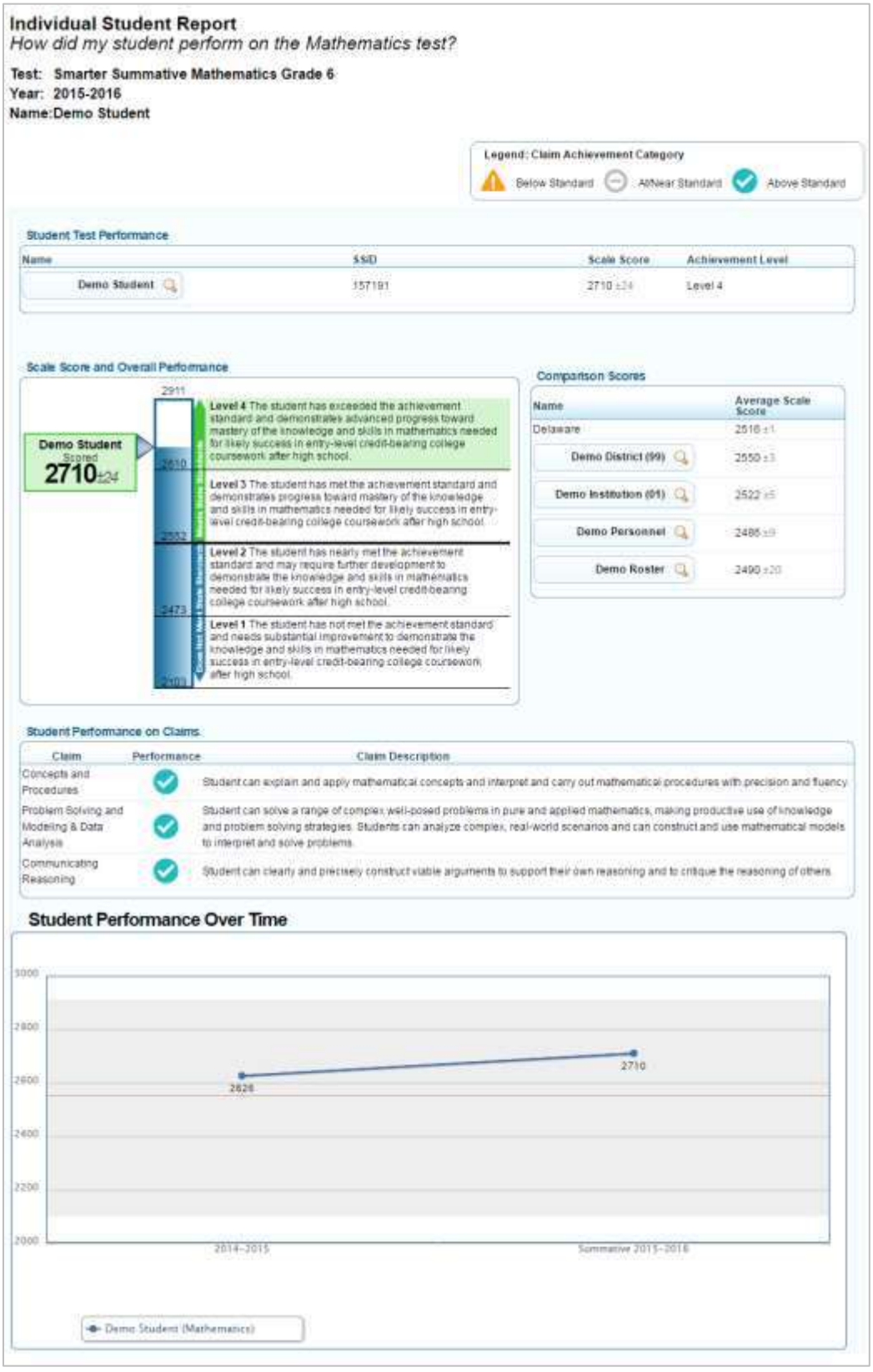


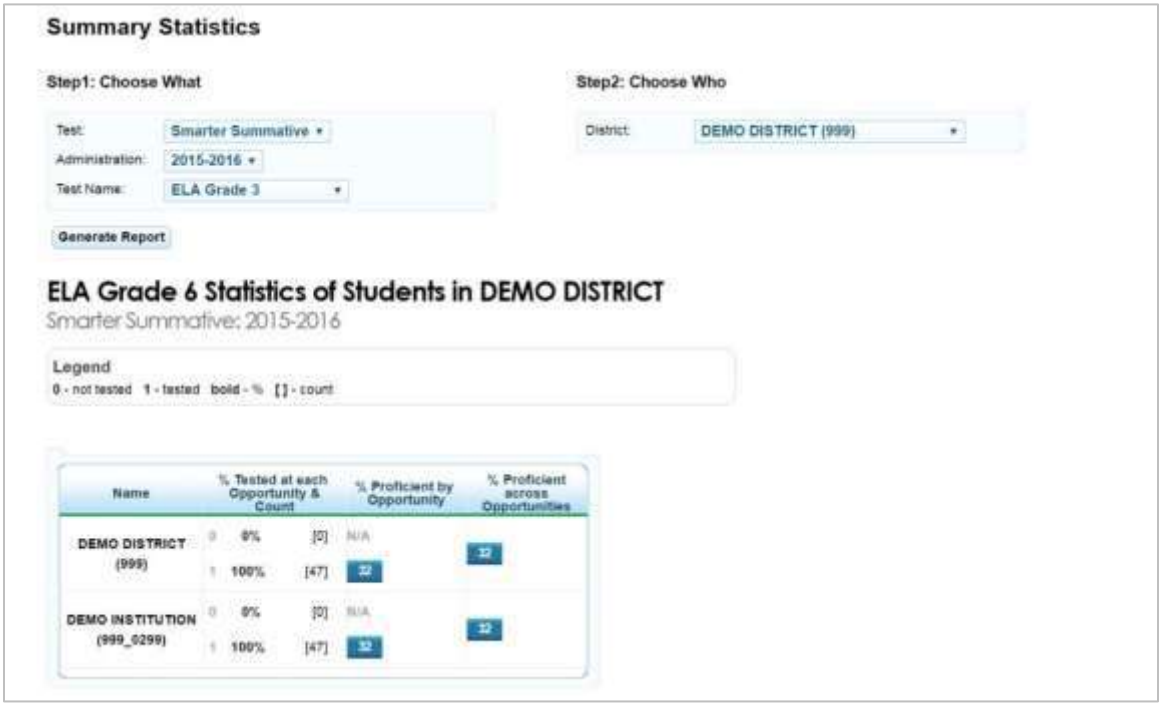
Exhibit 6. Student Detail Page for Mathematics



7.1.2.5 Participation Rate

In addition to online score reports, the ORS provides participation rate reports for 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



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

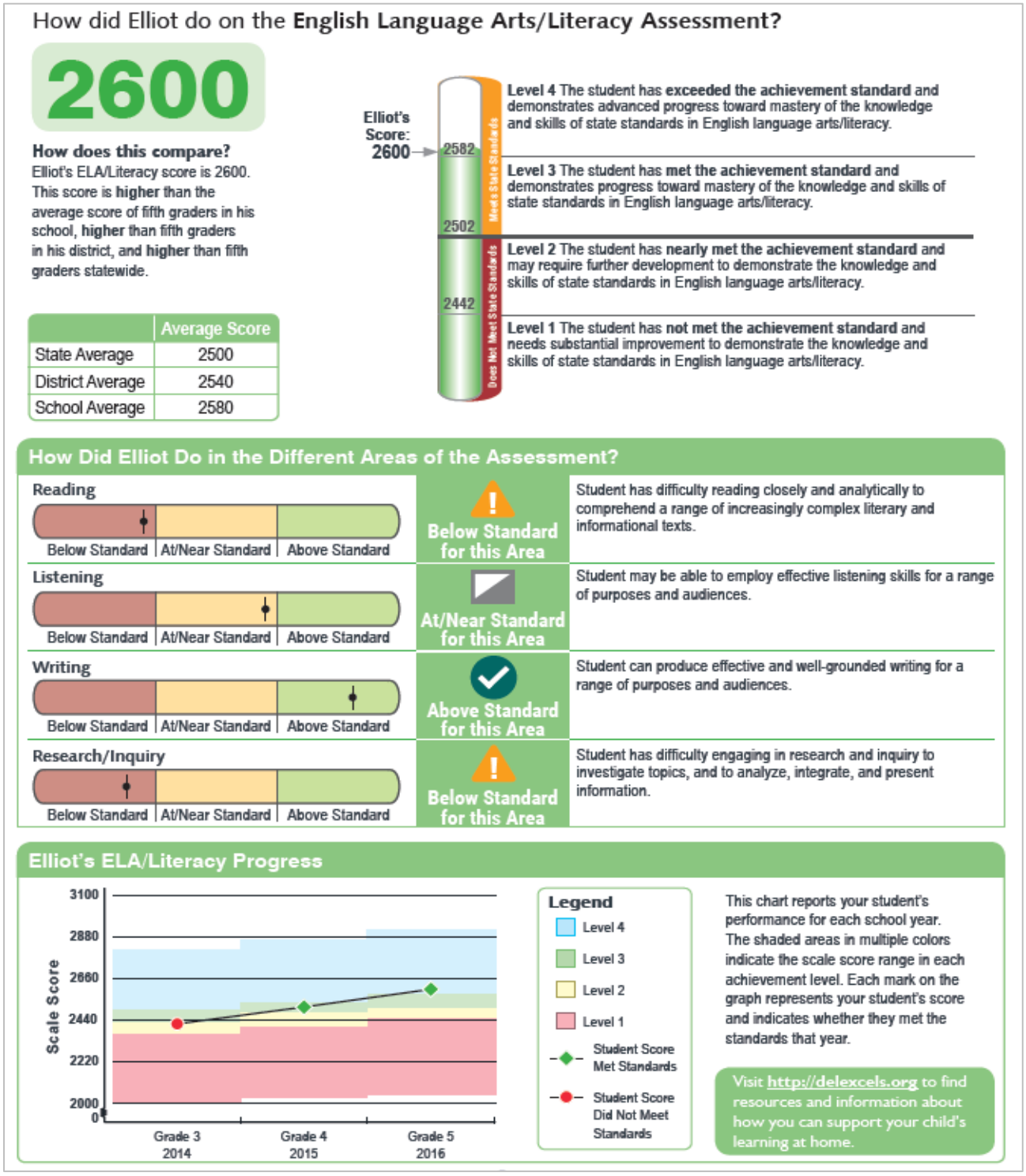
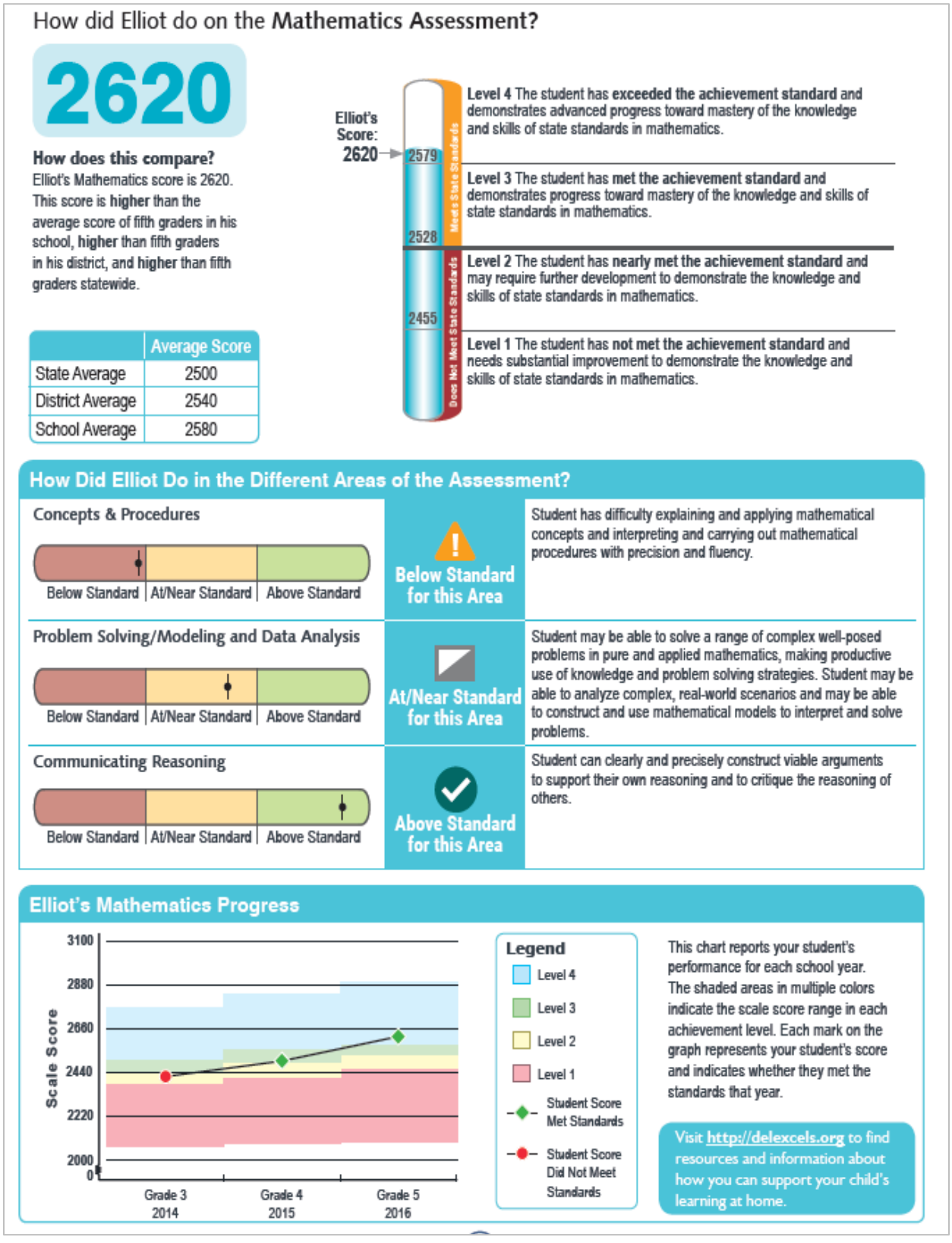


Exhibit 9. Sample Paper Family Score Report for Grade 5 Mathematics



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 at Level 3 in ELA/Lit, for instance, achievement-level descriptors are described for Level 3 as "students demonstrate progress toward mastery of the knowledge and skills ELA/Lit needed for likely success in future coursework." Generally, students performing Smarter Balanced assessments at Levels 3 and 4 are considered on track to demonstrate progress toward mastery of the knowledge and skills necessary for college and career readiness.

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

8.1 ADAPTIVE TEST CONFIGURATION

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

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

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

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

8.1.1 Platform Review

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

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

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

8.1.2 User Acceptance Testing and Final Review

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

8.2 QUALITY ASSURANCE IN DOCUMENT PROCESSING

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

8.3 QUALITY ASSURANCE IN DATA PREPARATION

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

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

8.4 QUALITY ASSURANCE IN HAND-SCORING

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

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

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

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

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

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

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

8.4.2 Human-scoring QA Monitoring Reports

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

8.4.3 Monitoring by State Department of Education

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

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

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

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

8.5 Quality Assurance in Test Scoring

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

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

For example, an item statistics analysis report allows psychometricians to ensure that items are performing as intended and serve as an empirical key check through the operational test window. The item statistics analysis report is used to monitor the performance of test items throughout the testing window and serves as a key check for the early detection of potential problems with item scoring,

including incorrect designation of a keyed response or other scoring errors, as well as potential breaches of test security that may be indicated by changes in the difficulty of test items. This report generates classical item analysis indicators of difficulty and discrimination, including proportion correct and biserial/polyserial correlation. The report is configurable and can be produced so that only items with statistics falling outside a specified range are flagged for reporting or to generate reports based on all items in the pool.

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

Table 46 presents an overview of the quality assurance (QA) reports.

Table 46. Overview of Quality Assurance Reports

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

8.5.1 Score Report Quality Check

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

8.5.1.1 Online Report Quality Assurance

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

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

Every test undergoes a series of validation checks. Once the QA system signs off, data are passed to the DoR, which serves as the centralized location for all student scores and responses, ensuring that there is only one place where the "official" record is stored. Only after scores have passed the QA checks and are uploaded to the DoR are they passed to the Online Reporting System (ORS), which is responsible for presenting individual-level results and calculating and presenting aggregate results. Absolutely no score is reported in the ORS until it passes all of the QA system's validation checks. All of the above processes

take milliseconds to complete so that within less than a second of hand-scores being received by AIR and passing QA validation checks, the composite score will be available in the ORS.

8.5.1.2 Paper Report Quality Assurance

Statistical Programming

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

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

Each change is followed by a complete retesting with the entire collection of scenarios on which the macro was originally tested. The main statistical program is mostly made up of calls to various macros, including macros that verify the data and conversion tables and the macros that do the many 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 Department staff review. AIR will work closely with the department to resolve questions and correct any problems. The reports will not be delivered unless the department approves the sample reports and data file.

REFERENCES

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

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

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

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

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 once or twice.

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

Grade	ELA/Lit			Mathematics		
	Once	Twice	Total	Once	Twice	Total
3	404	0	404	387	0	387
4	343	1	344	345	1	346
5	360	0	360	347	0	347
6				1	0	1
7						
8						
11	116	0	116			

For the Interim Assessment Blocks (IAB), there were seven IABs for ELA/Lit and four IABs in mathematics. Students were allowed to take as many IABs as they wanted. Table 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 772 students took IABs, and among 772 students, 470 students took one IAB, 209 students took two IABs, and so on.

Tables A–3 and A–4 disaggregated the number of students in Table A-2 by seven IABs in ELA/Lit and four IABs in mathematics. For example, 470 students in grade 3 ELA/Lit took one IAB only. Among 470 students, no student 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
English Language Arts/Literacy								
3	772	470	209	66	14	13		
4	863	511	213	104	35			
5	1,443	840	330	125	146	2		
6	1,616	982	312	285	34	3		
7	981	411	542	26	2			
8	1,148	571	557	20				
11	118	118						
Mathematics								
3	881	370	309	200	2			
4	769	324	238	207				
5	1,259	709	144	406				
6	1,713	1,177	368	162	6			
7	1,029	636	225	166	2			
8	1,015	720	208	86	1			
11	417	417						

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
3	Brief Writes							
	Editing and Revising	60	68	22	12	13		
	Listening and Interpretation	140	113	41	12	13		
	Performance Task	43	2	3				
	Reading Informational Text	185	36	33	11	13		
	Reading Literary Text	10	78	40	7	13		
	Research	32	121	59	14	13		
4	Brief Writes							
	Editing and Revising	22	114	88	21			
	Listening and Interpretation	164	148	102	35			
	Performance Task	1		1				
	Reading Informational Text	291	66	14	16			
	Reading Literary Text	33	2	27	34			
	Research		96	80	34			
5	Brief Writes	1	1					
	Editing and Revising	71	103	80	146	2		
	Listening and Interpretation	49	165	105	145	2		
	Performance Task	12	1					
	Reading Informational Text	618	126	20	1	2		
	Reading Literary Text	63	28	64	146	2		
	Research	26	236	106	146	2		
6	Brief Writes	8	2					
	Editing and Revising	14	181	259	34	3		
	Listening and Interpretation	251	108	255	34	3		
	Performance Task							
	Reading Informational Text	550	105	48	33	3		
	Reading Literary Text	45	73	49	3	3		
	Research	114	155	244	32	3		
7	Brief Writes							
	Editing and Revising	173	535	26	2			
	Listening and Interpretation	101	182	24	2			
	Performance Task	6	6	2	1			
	Reading Informational Text	94	10	12	2			
	Reading Literary Text	4	2		1			
	Research	33	349	14				
8	Brief Writes							
	Editing and Revising	22	521	20				
	Listening and Interpretation	302	243	20				
	Performance Task			7				
	Reading Informational Text	1	3					
	Reading Literary Text	2	32					
	Research	244	315	13				
11	Brief Writes							
	Editing and Revising	93						
	Listening and Interpretation							
	Performance Task							
	Reading Informational Text	25						
	Reading Literary Text							
	Research							

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

Grade	Block	Number of IABs Taken			
		1	**2**	**3**	**4**
3	Measurement and Data	35	192	200	2
	Number and Operations – Fractions	172	192	200	2
	Operational and Algebraic Thinking	139	234	200	2
	Performance Task	24			2
4	Number and Operations in Base Ten	111	173	207	
	Number and Operations – Fractions	179	161	206	
	Operational and Algebraic Thinking	34	142	207	
	Performance Task			1	
5	Measurement and Data	32	83	405	
	Number and Operations in Base Ten	255	105	406	
	Number and Operations – Fractions	397	99	406	
	Performance Task	25	1	1	
6	Expressions and Equations	171	276	161	6
	Geometry	570	69	157	6
	Performance Task	4	33	7	6
	Ratios and Proportional Relationships	432	358	161	6
7	Expressions and Equations	31	192	166	2
	The Number System	103	43	166	2
	Performance Task	1			2
	Ratios and Proportional Relationships	501	215	166	2
8	Expressions and Equations	218	82	78	1
	Functions	132	174	86	1
	Geometry	367	142	84	1
	Performance Task	3	18	10	1
11	Algebra – Linear Functions	148			
	Algebra – Quadratic Functions	133			
	Geometry – Right Triangles and Trigonometric	191			
	Performance Task				

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

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

Grade	2014–2015				2015–2016				Change in %Proficient
	N	Mean	SD	%Prof	N	Mean	SD	%Prof	
Grade 3									
All Students	10,231	2438.1	84.7	54	10,296	2439.5	85.4	54	0
Female	5,122	2448.1	83.9	59	5,122	2447.5	84.6	57	-2
Male	5,109	2428.1	84.3	49	5,174	2431.7	85.5	50	1
AmeriIndian/AlaskaNat	38	2460.6	77.4	76	40	2438.8	81.9	58	-18
Asian	375	2496.6	79.2	80	363	2497.2	85.7	80	0
African American	3,016	2405.7	81.6	39	3,109	2409.3	79.7	39	0
Hispanic	1,763	2415.3	75.7	41	1,789	2414.9	77.0	41	0
White	4,631	2462.8	80.6	66	4,542	2464.6	82.2	66	0
ELL	984	2382.5	64.5	23	1,249	2390.7	67.9	28	5
SPED	1,279	2351.3	70.0	13	1,334	2357.3	69.1	14	1
CD 504	332	2424.2	73.4	44	319	2430.4	75.8	52	8
Title I	1,161	2438.6	76.1	54	1,053	2451.2	77.0	59	5
Grade 4									
All Students	9,910	2477.4	88.0	54	10,268	2482.5	90.8	56	2
Female	4,932	2486.6	86.6	58	5,132	2493.7	89.8	61	3
Male	4,978	2468.3	88.4	49	5,136	2471.3	90.4	51	2
AmeriIndian/AlaskaNat	43	2494.1	80.1	65	38	2482.5	85.4	61	-4
Asian	385	2541.1	83.5	81	382	2550.7	88.6	81	0
African American	3,060	2444.4	82.8	37	3,035	2448.3	86.6	41	4
Hispanic	1,702	2452.8	78.7	40	1,781	2455.9	83.3	43	3
White	4,331	2503.9	83.7	68	4,611	2509.6	84.7	68	0
ELL	558	2399.6	69.6	14	641	2402.1	73.9	16	2
SPED	1,349	2380.1	71.9	11	1,452	2388.7	74.7	13	2
CD 504	376	2471.7	75.4	51	374	2469.5	84.2	49	-2
Title I	1,274	2467.9	80.1	49	1,243	2484.9	78.6	57	8
Grade 5									
All Students	9,922	2509.4	89.3	55	10,169	2519.3	90.0	60	5
Female	4,890	2522.7	86.7	61	5,053	2531.1	87.0	66	5
Male	5,032	2496.4	89.9	50	5,116	2507.6	91.3	55	5
AmeriIndian/AlaskaNat	41	2518.4	86.6	59	41	2540.1	76.2	68	9
Asian	361	2579.1	83.6	84	386	2585.3	79.9	85	1
African American	3,115	2473.8	85.0	39	3,077	2485.0	84.9	44	5
Hispanic	1,533	2486.3	79.4	44	1,761	2492.9	84.0	49	5
White	4,585	2534.9	84.2	68	4,490	2546.6	84.3	73	5
ELL	303	2409.2	65.4	9	420	2418.5	75.3	13	4
SPED	1,381	2408.2	70.6	11	1,451	2420.2	76.3	15	4
CD 504	412	2502.1	82.6	50	424	2504.4	77.9	53	3
Title I	1,621	2510.5	84.7	56	1,359	2519.7	81.6	60	4

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

Grade	2014–2015				2015–2016				Change in %Proficient
	N	Mean	SD	%Prof	N	Mean	SD	%Prof	
Grade 6									
All Students	10,023	2522.8	92.4	48	9,983	2530.2	93.5	52	4
Female	4,943	2538.9	89.1	55	4,923	2544.4	90.0	57	2
Male	5,080	2507.1	92.9	41	5,060	2516.3	94.7	46	5
AmeriIndian/AlaskaNat	48	2536.1	81.7	52	43	2526.1	84.8	47	-5
Asian	352	2597.4	83.0	80	355	2603.0	90.7	81	1
African American	3,097	2490.4	87.3	33	3,135	2494.5	87.4	35	2
Hispanic	1,601	2498.7	87.3	38	1,549	2505.3	87.6	40	2
White	4,694	2546.3	88.4	59	4,615	2556.9	87.8	65	6
ELL	247	2409.1	72.0	5	298	2416.1	72.1	7	2
SPED	1,389	2422.5	75.5	8	1,418	2432.0	76.5	9	1
CD 504	416	2513.5	84.1	43	430	2525.0	84.2	47	4
Title I	1,814	2515.8	86.1	45	1,570	2531.8	86.7	52	7
Grade 7									
All Students	9,716	2547.1	96.0	50	10,049	2552.7	98.2	52	2
Female	4,735	2564.4	92.5	58	4,957	2569.4	96.4	59	1
Male	4,981	2530.7	96.4	43	5,092	2536.5	97.3	46	3
AmeriIndian/AlaskaNat	52	2553.6	92.6	50	44	2579.5	83.3	66	16
Asian	354	2621.7	90.9	81	347	2633.1	94.3	82	1
African American	3,068	2509.3	89.3	33	3,057	2514.1	90.9	35	2
Hispanic	1,453	2521.8	90.0	39	1,642	2527.2	95.0	41	2
White	4,555	2574.7	90.5	63	4,720	2579.8	92.4	65	2
ELL	285	2433.3	74.1	9	292	2434.1	69.8	5	-4
SPED	1,328	2445.8	74.5	8	1,440	2449.5	77.9	10	2
CD 504	351	2535.6	85.4	44	453	2542.2	88.1	45	1
Title I	1,902	2542.8	92.1	50	1,778	2550.7	93.7	52	2
Grade 8									
All Students	9,546	2559.1	97.9	49	9,747	2569.6	98.1	54	5
Female	4,669	2576.1	93.7	56	4,761	2588.0	94.2	61	5
Male	4,877	2542.9	99.1	43	4,986	2552.1	98.5	47	4
AmeriIndian/AlaskaNat	38	2600.1	92.8	66	50	2579.1	100.8	56	-10
Asian	328	2634.7	92.0	80	366	2642.3	98.9	80	0
African American	3,109	2521.5	91.2	33	3,101	2533.3	91.2	38	5
Hispanic	1,267	2533.9	89.7	38	1,508	2542.7	92.7	43	5
White	4,574	2585.2	93.5	60	4,484	2597.9	92.1	66	6
ELL	258	2454.2	76.4	7	329	2450.3	77.7	8	1
SPED	1,350	2459.7	77.5	10	1,364	2465.4	77.4	9	-1
CD 504	404	2551.3	88.2	44	381	2562.9	85.2	48	4
Title I	1,957	2545.2	94.4	42	1,843	2566.7	91.8	54	12

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

Grade	2014–2015				2015–2016				Change in %Proficient
	N	Mean	SD	%Prof	N	Mean	SD	%Prof	
Grade 3									
All Students	10,268	2,439.4	75.5	53	10,341	2,444.0	78.6	55	2
Female	5,150	2,439.9	73.3	53	5,146	2,443.4	76.8	54	1
Male	5,118	2,438.9	77.6	53	5,195	2,444.6	80.3	56	3
AmeriIndian/AlaskaNat	38	2,460.1	68.5	66	40	2,442.3	76.1	50	-16
Asian	391	2,499.6	75.3	80	378	2,509.3	73.0	87	7
African American	3,026	2,408.4	70.8	36	3,106	2,411.8	74.4	39	3
Hispanic	1,784	2,420.2	67.7	41	1,817	2,423.8	68.9	44	3
White	4,620	2,462.0	71.4	67	4,547	2,468.2	74.4	68	1
ELL	1,032	2,395.4	63.5	25	1,306	2,410.5	66.2	35	10
SPED	1,280	2,360.0	72.9	14	1,335	2,364.6	78.1	17	3
CD 504	333	2,432.7	67.9	48	319	2,438.6	72.1	49	1
Title I	1,163	2,440.8	62.5	54	1,057	2,456.1	67.2	61	7
Grade 4									
All Students	9,995	2476.9	75.4	47	10,297	2485.1	79.4	51	4
Female	4,970	2475.6	71.9	45	5,151	2485.1	76.0	50	5
Male	5,025	2478.1	78.7	48	5,146	2485.0	82.7	51	3
AmeriIndian/AlaskaNat	43	2495.3	64.7	56	37	2489.0	61.9	49	-7
Asian	401	2539.9	73.2	78	391	2555.0	85.7	81	3
African American	3,063	2446.5	69.8	29	3,041	2451.7	72.9	33	4
Hispanic	1,736	2457.0	68.1	36	1,804	2462.7	70.2	38	2
White	4,362	2499.4	71.5	60	4,605	2510.1	74.6	65	5
ELL	613	2419.9	67.7	16	683	2424.9	65.8	18	2
SPED	1,355	2393.1	66.9	8	1,450	2405.5	68.7	12	4
CD 504	377	2470.6	66.1	40	375	2478.3	78.1	47	7
Title I	1,279	2477.8	67.2	46	1,247	2494.2	67.9	56	10
Grade 5									
All Students	10,017	2498.6	85.0	38	10,199	2506.8	86.8	42	4
Female	4,935	2498.8	82.1	37	5,070	2505.5	84.0	40	3
Male	5,082	2498.3	87.7	39	5,129	2508.0	89.5	43	4
AmeriIndian/AlaskaNat	41	2499.2	79.6	34	42	2518.5	79.1	43	9
Asian	375	2573.8	82.2	74	395	2580.0	85.1	74	0
African American	3,148	2461.0	79.9	21	3,077	2468.6	78.4	23	2
Hispanic	1,565	2477.0	75.1	27	1,787	2483.3	79.5	29	2
White	4,602	2524.9	78.7	50	4,484	2535.2	81.3	56	6
ELL	346	2416.5	70.6	8	468	2426.1	74.1	8	0
SPED	1,390	2409.4	69.8	5	1,449	2416.0	72.9	6	1
CD 504	409	2493.9	77.2	29	423	2498.4	73.4	35	6
Title I	1,628	2500.7	83.3	38	1,362	2512.4	80.0	45	7

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

Grade	2014–2015				2015–2016				Change in %Proficient
	N	Mean	SD	%Prof	N	Mean	SD	%Prof	
Grade 6									
All Students	10,084	2510.5	96.3	34	10,004	2516.3	101.8	37	3
Female	4,981	2515.4	92.5	35	4,937	2519.5	98.3	37	2
Male	5,103	2505.8	99.7	33	5,067	2513.3	105.0	37	4
AmeriIndian/AlaskaNat	48	2518.8	89.9	38	43	2510.2	90.9	28	-10
Asian	358	2598.7	94.6	69	361	2606.2	114.1	70	1
African American	3,111	2470.6	87.7	17	3,125	2474.1	96.1	21	4
Hispanic	1,635	2486.0	90.2	22	1,581	2487.2	91.2	24	2
White	4,701	2538.0	90.7	46	4,607	2547.5	92.6	50	4
ELL	291	2402.4	84.4	4	339	2402.2	81.9	4	0
SPED	1,405	2404.9	82.6	4	1,414	2407.4	91.0	5	1
CD 504	417	2506.6	83.8	28	429	2513.8	93.4	32	4
Title I	1,826	2505.4	87.1	30	1,584	2515.5	97.4	37	7
Grade 7									
All Students	9,754	2529.6	102.7	37	10,070	2534.5	106.6	40	3
Female	4,753	2535.1	99.3	39	4,970	2538.6	104.8	41	2
Male	5,001	2524.4	105.6	35	5,100	2530.4	108.1	38	3
AmeriIndian/AlaskaNat	52	2529.7	94.4	29	44	2560.0	93.1	55	26
Asian	360	2622.9	107.9	71	357	2638.9	109.7	77	6
African American	3,064	2486.7	93.8	19	3,054	2488.0	97.5	21	2
Hispanic	1,490	2501.1	97.8	26	1,667	2505.7	103.8	29	3
White	4,556	2560.2	94.4	50	4,710	2566.2	96.6	52	2
ELL	334	2416.2	90.8	5	339	2421.8	96.4	7	2
SPED	1,324	2419.1	86.6	4	1,435	2423.2	89.9	6	2
CD 504	350	2528.2	90.6	33	450	2532.9	91.3	36	3
Title I	1,912	2521.8	94.3	33	1,777	2534.5	96.7	39	6
Grade 8									
All Students	9,512	2541.7	112.0	35	9,768	2548.9	117.0	38	3
Female	4,646	2547.3	106.6	36	4,765	2557.9	111.0	41	5
Male	4,866	2536.4	116.6	35	5,003	2540.4	121.8	35	0
AmeriIndian/AlaskaNat	38	2560.0	120.3	42	50	2549.2	111.4	42	0
Asian	329	2647.6	116.1	71	370	2658.6	138.8	74	3
African American	3,091	2491.4	97.3	17	3,097	2500.3	105.4	20	3
Hispanic	1,264	2516.4	101.0	27	1,530	2517.7	104.4	25	-2
White	4,558	2574.5	106.9	47	4,483	2584.0	108.6	51	4
ELL	267	2442.1	102.0	9	367	2437.8	99.8	9	0
SPED	1,350	2435.1	86.3	5	1,364	2432.0	94.5	5	0
CD 504	402	2540.6	99.0	31	382	2541.7	101.4	32	1
Title I	1,943	2531.0	104.4	30	1,843	2536.9	109.6	33	3

Appendix C: Classification Accuracy and Consistency Indexes by Subgroups

Table C-1. 2015–2016 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	10,296	79	87	71	68	88	71	80	61	58	83
Female	5,122	79	87	72	68	89	71	79	61	58	84
Male	5,174	78	88	71	68	87	70	81	61	57	81
American Indian/Alaska Native	40	77	94*	71	68	95*	69	78*	60	62	84*
Asian	363	83	86	72	68	92	76	75	58	58	89
African American	3,109	78	89	72	68	84	70	83	62	57	76
Hispanic	1,789	77	87	72	69	86	69	79	62	58	77
White	4,542	79	86	71	68	89	72	76	60	57	85
ELL	1,249	77	87	71	68	81	68	80	62	58	66
Special Education	1,334	82	91	71	67	82	76	87	61	54	72
CD 504	319	77	86	70	70	88	68	78	58	62	79
Title I	1,053	77	84	71	68	88	69	73	61	58	82
Grade 4											
All Students	10,268	78	89	65	67	88	70	82	53	56	82
Female	5,132	78	88	65	67	88	70	81	53	56	83
Male	5,136	78	89	65	67	88	70	83	53	56	81
American Indian/Alaska Native	38	77	92*	71*	64	90*	68	82*	57*	57	78*
Asian	382	83	87	67	67	92	77	73	54	56	90
African American	3,035	78	90	65	67	84	70	85	54	57	75
Hispanic	1,781	76	88	65	66	84	68	82	54	56	77
White	4,611	79	87	65	67	89	71	79	53	56	84
ELL	641	81	91	65	67	76	74	87	53	52	63
Special Education	1,452	84	93	65	66	83	78	90	54	53	65
CD 504	374	76	88	66	67	87	68	80	54	58	79
Title I	1,243	75	85	65	66	86	67	77	53	56	80
Grade 5											
All Students	10,169	79	88	67	76	86	71	81	55	68	79
Female	5,053	79	87	67	76	87	71	79	55	68	81
Male	5,116	79	88	67	75	85	71	83	55	68	78
American Indian/Alaska Native	41	78	84*	68*	72	89	69	72*	56*	63	84
Asian	386	83	82	65	75	90	77	73	51	65	87
African American	3,077	79	88	67	76	82	70	83	55	68	72
Hispanic	1,761	78	88	68	75	84	70	82	56	68	72
White	4,490	80	87	67	76	87	72	78	54	68	81
ELL	420	84	91	67	72	75*	78	89	56	61	59*
Special Education	1,451	83	92	67	73	73	77	89	56	62	55
CD 504	424	77	87	67	77	82	68	77	56	68	73
Title I	1,359	77	86	68	75	83	68	76	57	67	76

*The classification index is based on n<10.

Table C-2. 2015–2016 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	9,983	79	88	71	76	84	71	81	62	68	77
Female	4,923	79	87	72	77	85	71	79	62	69	78
Male	5,060	79	89	71	76	84	71	83	62	68	75
American Indian/Alaska Native	43	80	86	76	76	87*	71	79	67	68	72*
Asian	355	83	93	70	76	89	76	83	60	67	85
African American	3,135	80	89	72	77	80	72	84	63	67	69
Hispanic	1,549	79	89	71	76	80	70	82	62	68	68
White	4,615	79	86	71	76	85	70	77	60	69	78
ELL	298	88	94	70	74	64*	83	92	59	62	39*
Special Education	1,418	84	92	71	74	84	78	89	62	58	70
CD 504	430	78	87	72	78	84	70	77	63	70	76
Title I	1,570	79	87	72	77	84	70	79	62	69	75
Grade 7											
All Students	10,049	80	89	71	79	84	72	83	60	72	76
Female	4,957	80	88	71	79	85	72	81	60	72	77
Male	5,092	80	90	71	79	82	73	84	60	72	73
American Indian/Alaska Native	44	78	90*	65*	80	79*	70	81*	53*	73	72*
Asian	347	83	85	68	78	88	76	82	55	70	85
African American	3,057	81	90	71	79	81	73	85	61	72	67
Hispanic	1,642	80	90	70	78	81	72	83	61	70	71
White	4,720	80	87	71	79	85	72	79	60	73	77
ELL	292	87	94	68	76	60*	82	91	57	59	42*
Special Education	1,440	86	93	71	76	90	80	90	60	65	58
CD 504	453	79	88	71	80	83	71	81	62	71	74
Title I	1,778	79	88	71	78	82	71	81	61	71	72
Grade 8											
All Students	9,747	81	88	74	79	84	73	82	64	73	75
Female	4,761	81	87	74	80	85	73	79	65	74	76
Male	4,986	81	89	73	79	82	73	84	64	73	73
American Indian/Alaska Native	50	81	97	74	76	81	74	91	66	68	76
Asian	366	82	86	73	78	88	76	80	63	70	85
African American	3,101	81	89	74	80	80	74	84	65	73	67
Hispanic	1,508	81	89	75	80	80	73	83	65	73	65
White	4,484	80	87	73	80	84	73	78	63	74	76
ELL	329	87	92	76	76	-	82	91	63	66	12*
Special Education	1,364	85	91	74	76	79*	79	89	65	64	56*
CD 504	381	79	85	75	79	81	71	77	67	72	71
Title I	1,843	80	89	73	79	82	72	82	63	74	71

*The classification index is based on n<10.

Table C-3. 2015–2016 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	10,341	82	88	73	79	89	74	81	64	72	84
Female	5,146	81	88	73	79	89	74	80	64	71	84
Male	5,195	82	88	73	79	90	75	82	64	72	84
American Indian/Alaska Native	40	76	79*	71	67	91*	68	72*	63	58	84*
Asian	378	86	88	74	81	92	81	73	59	74	90
African American	3,106	81	89	73	79	86	74	84	64	71	78
Hispanic	1,817	80	87	74	79	86	73	81	64	72	77
White	4,547	82	87	73	79	90	75	77	64	72	86
ELL	1,306	81	88	74	79	86	73	81	65	72	74
Special Education	1,335	85	93	73	77	85	79	89	64	67	74
CD 504	319	80	85	74	77	89	73	79	65	69	82
Title I	1,057	80	83	72	79	90	73	73	63	72	84
Grade 4											
All Students	10,297	83	88	81	79	89	76	80	74	71	84
Female	5,151	83	87	81	78	88	76	78	74	70	83
Male	5,146	84	89	81	79	89	77	82	73	71	84
American Indian/Alaska Native	37	83	83*	81	80	92*	75	74*	75	72	80*
Asian	391	86	84	79	77	93	81	73	71	69	91
African American	3,041	83	89	81	79	85	76	83	75	71	75
Hispanic	1,804	83	87	81	79	87	76	80	74	72	81
White	4,605	83	86	81	78	89	76	76	73	71	85
ELL	683	84	90	81	77	85	78	85	74	67	80
Special Education	1,450	86	92	80	78	79	80	88	72	66	73
CD 504	375	82	85	80	78	92	75	80	70	72	83
Title I	1,247	83	85	82	80	88	76	74	75	73	83
Grade 5											
All Students	10,199	82	89	78	71	89	75	84	70	61	84
Female	5,070	82	89	78	71	89	75	84	70	60	83
Male	5,129	83	90	78	72	90	76	84	70	62	85
American Indian/Alaska Native	42	81	80	83	71	94*	73	79	66	63	87*
Asian	395	86	87	79	70	93	80	80	69	57	92
African American	3,077	83	90	78	71	86	76	86	69	60	78
Hispanic	1,787	82	89	78	71	87	75	84	69	60	80
White	4,484	82	89	78	72	90	75	80	71	61	85
ELL	468	87	92	76	72	87	82	91	65	56	78
Special Education	1,449	88	93	76	66	84	84	92	65	54	74
CD 504	423	81	87	78	72	89	73	81	70	63	82
Title I	1,362	81	88	78	72	88	74	82	70	62	82

*The classification index is based on n<10.

Table C-4. 2015–2016 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	10,004	82	91	77	71	88	75	85	70	61	82
Female	4,937	82	90	77	71	88	74	84	70	60	81
Male	5,067	82	91	77	71	89	76	86	69	61	82
American Indian/Alaska Native	43	82	89	78	76*	84*	74	77	75	59*	83*
Asian	361	85	93	79	70	93	79	85	71	61	89
African American	3,125	84	92	77	72	86	78	88	70	60	75
Hispanic	1,581	82	91	77	72	83	75	86	69	61	73
White	4,607	81	88	77	71	89	73	80	70	61	83
ELL	339	91	95	74	77*	86*	87	94	61	63*	76*
Special Education	1,414	91	95	78	71	87	87	94	66	57	78
CD 504	429	82	91	79	70	88	75	85	73	57	82
Title I	1,584	82	90	77	73	87	74	85	69	63	79
Grade 7											
All Students	10,070	83	91	77	75	90	76	85	69	66	85
Female	4,970	83	91	77	75	90	76	85	69	65	85
Male	5,100	83	91	77	75	90	76	86	69	66	84
American Indian/Alaska Native	44	83	90	83	74	92*	76	84	69	70	82*
Asian	357	86	87	76	73	94	81	75	67	65	92
African American	3,054	84	92	77	74	85	77	88	69	63	77
Hispanic	1,667	84	92	77	74	89	77	87	70	65	81
White	4,710	82	89	77	75	90	75	81	69	67	85
ELL	339	90	95	76	76	96*	86	93	66	67	75*
Special Education	1,435	90	95	76	72	86*	86	93	65	63	63*
CD 504	450	82	88	77	76	90	74	82	69	66	86
Title I	1,777	82	90	78	74	88	74	83	69	66	81
Grade 8											
All Students	9,768	81	90	72	72	90	74	84	62	61	85
Female	4,765	81	89	72	72	90	73	83	62	62	85
Male	5,003	82	90	72	72	90	75	85	62	61	86
American Indian/Alaska Native	50	77	86	70	65	85	71	78	59	59	85
Asian	370	85	90	70	71	94	80	80	60	61	92
African American	3,097	83	91	72	71	88	76	87	62	59	80
Hispanic	1,530	82	90	71	72	88	74	85	62	61	81
White	4,483	80	88	72	72	90	73	80	63	62	86
ELL	367	89	95	69	68	93*	85	93	56	61	73*
Special Education	1,364	89	94	71	67	89	85	93	59	54	69
CD 504	382	79	87	71	70	89	71	82	62	60	81
Title I	1,843	81	89	72	72	88	73	83	63	62	81

*The classification index is based on n<10.