# South Dakota Smarter Balanced Summative Assessments

# 2016–2017 Technical Report

## Addendum to the Smarter Balanced Technical Report


**Submitted to**
**South Dakota 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/L) and mathematics for grades 3–8 and 11, and to provide valid, reliable, and fair test scores about student academic achievement. South Dakota is among 18 member states (plus the U.S. Virgin Islands) leading the development of assessments in ELA/L and mathematics. The system includes both summative assessments, for accountability purposes, and optional interim assessments that provide meaningful feedback and actionable data that teachers and educators can use to help students succeed. Smarter Balanced, a state-led enterprise, is intended to provide leadership and resources to improve teaching and learning by creating and maintaining a suite of summative and interim assessments and tools aligned to the CCSS in ELA/L and mathematics.

The South Dakota State Board of Education formally adopted the CCSS in ELA/L and mathematics on November 28, 2010 (State Board meeting minutes, approved January, 2011). South Dakota CCSS define the knowledge and skills students need to succeed in college and careers after graduating from high school. They align with college and workforce expectations, are clear and consistent, include rigorous content and application of knowledge through higher-order skills, are evidence-based, and are informed by standards in top-performing countries.

Since the adoption of the CCSS in 2010, the South Dakota Department of Education fully implemented CCSS in all grade levels in SY 2013–2014. The new South Dakota statewide assessments in ELA/L and mathematics aligned with the CCSS were administered for the first time in spring 2015 to students in grades 3–8 and 11 in all public elementary and secondary schools. The American Institutes for Research (AIR) delivered and scored the Smarter Balanced assessments, and produced score reports. Measurement Incorporated (MI) scored the hand-scored items.

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

- **Computer Adaptive Test:** An online adaptive test that provides an individualized assessment for each student.
- **Performance Task:** A task that challenges students to apply their knowledge and skills to respond to real-world problems. Performance tasks can best be described as collections of questions and activities that are coherently connected to a single theme or scenario. They are used to better measure capacities such as depth of understanding, research skills, and complex analysis, 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 they can use to improve instruction and learning. These tools are used at the discretion of schools and districts, and teachers can employ them to check students' progress at mastering specific concepts at strategic points during the school year. The interim assessments are available as fixed form tests and consist of the following features:

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

- **Interim Assessment Blocks (IABs)** that focus on smaller sets of related concepts and provide more detailed information about student learning.

This report provides a technical summary of the 2016–2017 summative assessments in ELA/L and mathematics administered in grades 3–8 and 11 under the South Dakota Smarter Balanced assessments. The report includes eight chapters: overview, test administration, the 2016–2017 operational administration, validity, reliability, scoring, reporting and interpreting scores, and the quality control process. The data included in this report are based on South Dakota 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 South Dakota, it is an addendum to the Smarter Balanced technical report. The information on item and test development, item content review, field-test administration, item data review, item calibrations, content alignment study, standard setting, and other validity information is included in the Smarter Balanced technical report.

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

# 2. TESTING ADMINISTRATION

## 2.1 TESTING WINDOWS

The 2016–2017 Smarter Balanced assessment testing window spanned approximately two months for the online summative assessments and approximately five or eight months for the interim assessments. The paper-pencil fixed-forms for summative assessments were administered for one month during the online summative window. Table 1 shows the testing windows for both online and paper-pencil summative and interim assessments.

Table 1. 2016–2017 Testing Windows

| Tests | Grade | Start Date | End Date | Mode |
|---|---|---|---|---|
| Summative Assessments | 3–8, 11 | 3/8/2016 | 5/5/2017 | Online Adaptive |
| | 3–8, 11 | 3/20/2016 | 4/21/2017 | Paper Fixed-Form |
| Interim Comprehensive Assessments | 3–8, 11 | 9/7/2016 | 2/24/2017 | Online Fixed-Form |
| Interim Assessment Blocks | 3–8, 11 | 9/7/2016 | 5/5/2017 | Online Fixed-Form |

## 2.2 TEST OPTIONS AND ADMINISTRATIVE ROLES

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

Table 2. Summary of Tests and Testing Options in 2016–2017

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

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

### 2.2.1 Administrative Roles

The key personnel involved with the test administration are District Administrators (DAs), District Test Coordinators (DTCs), School Test Coordinators (STCs), TAs, and TEs. The main responsibilities of these key personnel are described below. More detailed descriptions can be found in the TAM provided online at this URL: http://sd.portal.airast.org/resources/.

**District Administrator (DA)**

The DA's role is assigned by the state Department of Education. This role can add users with any other roles in the Test Information Distribution Engine (TIDE) excluding DA roles. DAs have the same test administration responsibilities as DTCs. Their primary responsibility is to coordinate the administration of the Smarter Balanced assessment in the district.

**District Test Coordinator (DTC)**

The DTC's primary responsibility is to coordinate the administration of the South Dakota Smarter Balanced assessment in their district.

DTCs are responsible for the following:

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

**School Test Coordinator (STC)**

The STC's primary responsibilities are to coordinate the administration of the South Dakota Smarter Balanced assessment and ensure that testing within his or her school is conducted in accordance with the test procedures and security policies established by the South Dakota State Department of Education (SDDOE).

STCs are responsible for the following:

- Establishing a testing schedule with DTCs, TEs, and TAs based on test administration windows
- Working with technology staff to ensure timely computer setup and installations
- Working with TEs and TAs to review student information in TIDE to ensure that student information and test settings for designated supports and accommodations are correctly applied

- Entering student test settings in TIDE
- Identifying students who may require designated supports and test accommodations and ensuring that procedures for testing these students follow state and Smarter Balanced policy
- Attending all district trainings and reviewing all state and Smarter Balanced policy and test administration documents
- Ensuring that all TEs and TAs attend school or district trainings and review online training modules posted on the portal
- Establishing secure and separate testing rooms if needed
- Monitoring secure administration of the test
- Monitoring testing progress during the testing window and ensuring that all students participate, as appropriate
- Investigating and reporting all testing improprieties, irregularities, and breaches reported by the TEs and TAs
- Attending to any secure material according to state and Smarter Balanced policies

**Teacher (TE)**

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

**Test Administrator (TA)**

TAs are primarily responsible for administering the South Dakota Smarter Balanced assessments. The TA's role does not allow access to student results and is designed for TAs, such as technology staff, who administer tests but should not have access to student results.

TAs are responsible for the following:

- Completing South Dakota Smarter Balanced assessment administration training (see Section 1.4)
- Training and reviewing all state and Smarter Balanced policies and test administration documents before administering any South Dakota Smarter Balanced assessments
- Viewing student information before testing to ensure that a student receives the proper test with the appropriate supports. TAs should report any potential data errors to STCs and DTCs as appropriate
- Administering the South Dakota Smarter Balanced assessments
- Reporting all potential test security incidents to the STCs and DTCs in a manner consistent with Smarter Balanced, state, and district policies

### 2.2.2 Online Administration

Within the state's testing window, schools can set testing schedules, 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 that are 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 TEs and TAs administer the online assessments only. TEs and TAs are trained in the online testing requirements and the mechanics of starting, pausing, and ending a test session. Training materials for the test administration are provided online. All school personnel who serve as TEs and TAs are encouraged to complete an online TA Certification Course before testing begins. Staff who complete this course receive a certificate of completion.

To start a test session, the TAs or TEs must first enter the TA Interface of the online testing system using his or her own computer. A session ID is generated when the test session is created. Students who are taking the assessment with the TE or TA need to enter their Statewide Student Identification number (SSID), first name, and session ID into the student interface using computers provided by the school. The TE or TA then verifies that the students are taking the appropriate assessments with the appropriate accessibility feature(s) (see Section 2.6 for a list of accommodations). Students can begin testing only when the TE or TA confirms the settings. The TE or TA needs to read *Section 10 Day of Test Administration* in the *Online, Summative, 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, in the same test segment, and this session has not been paused for more than 20 minutes. Students may review and edit the 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 would be no change in items 6–10. No pause rule is implemented for the performance tasks. The same rules that apply to the CAT for reviews and changes to responses also apply to performance tasks.

For the summative test, an assessment can be started in one component and completed in a different component. For the CAT, the assessment must be completed within 45 calendar days of the start date, otherwise, the assessment opportunity will expire. For the performance tasks, the assessment must be completed within 10 calendar days of the start date.

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

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

### 2.2.3 Paper-Pencil Test Administration

The paper-pencil versions of the South Dakota Smarter Balanced ELA/L and mathematics assessments are provided as an alternate test administration method for students who could not gain access to a computer

or students with blindness or visual impairments. For South Dakota, paper-pencil tests were offered in the standard, non-accommodated format and the braille format.

The DA at the district with student(s) who need to take the paper-pencil version must submit a request for appropriate materials, on behalf of the student 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 and paper-pencil *Test Administration Manual* to the district. For ELA/L, the field (i.e., schools, districts) also receives a Listening Script that contains secure information needed to administer the listening session.

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

After the student has completed the assessments, the DA returns the test booklets, answer booklets, and the listening script 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 2016–2017, 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/L assessment can emboss both reading passages and items as they progress through the assessment. If a student has a Refreshable Braille Display (RBD), a 40-cell RBD is recommended. The summative ELA/L is presented to the student with items in either contracted or un-contracted literary braille (for items containing only text) and via a braille embosser (for items with tactile or spatial components that cannot be read by an RBD).

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

## 2.3 Training and Information for Test Coordinators and Administrators

All DAs, DTCs, and STCs oversee all aspects of testing at their schools and serve as the main point of contact, while TEs and TAs administer the online assessments. The online TA Certification Course, webinars, user guides, manuals, and regional training sites are used to train district and school coordinators about the online testing requirements and the mechanics of starting, pausing, and ending a test session. Training materials for the administration are provided online at http://sd.portal.airast.org/resources. District and School Coordinators are responsible for training TEs and TAs.

### 2.3.1 Online Training

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

**TA Certification Course**

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

**Webinars**

The following six webinars were offered to the field:

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

*Test Information Distribution Engine (TIDE)*: This webinar provides an overview of how to navigate the TIDE system, including how to register users, enroll students, manage and edit users/students, and process/view test invalidations.

*Test Administrator Interface for Online Testing*: This webinar prepares DTCs, STCs, and TAs for the assessments by providing an overview of the Test Administrator Interface and the Test Delivery System (TDS), including how to start and monitor a test session using the TA Interface.

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

*Interim Teacher Hand Scoring System (THSS)*: This webinar provides an overview of the THSS for the Smarter Balanced interim assessments. This application allows scorers to score test responses that require human scoring.

*Assessment Viewing Application (AVA)*: This webinar provides an overview of the AVA for the Smarter Balanced interim assessments. This application allows district-level and school-level users to view the Interim assessments (ICAs and IABs) for administrative or instructional purposes.

Each of these interactive webinars lasts approximately one hour. Participants can ask questions during and after the presentation. The audio portion of the webinar is recorded. The PowerPoint slides and audio files of the interactive webinars are available on the portal after the live webinars are completed at http://sd.portal.airast.org/resources/.

**Practice and Training Test Site**

In September 2016, separate training sites were opened for TEs, TAs, and students. TEs and TAs can practice administering assessments and starting and ending test sessions on the TA Training Site; and students can practice taking an online assessment on the Student Practice and Training Site. The Smarter Balanced assessment practice tests mirror the corresponding summative assessments for ELA/L and

mathematics. Each test provides students with a grade-specific testing experience, including a variety of question types and difficulty levels (approximately 30 items each in ELA/L and mathematics), as well as a performance task.

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

A student can log in directly to the practice and training test site as a "Guest" without a TA-generated test session ID, or the student can log in through a training test session created by the TE or 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 items, grid items, and natural language items.

**Manuals and User Guides**

The following manuals and user guides are available on the SD portal at http://sd.portal.airast.org/.

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

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

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

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

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

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

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

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

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

The *Usability, Accessibility, and Accommodations Guidelines* provides information for school-level personnel and decision-making teams, particularly Individualized Education Program (IEP) teams, to use

in selecting and administering universal tools, designated supports, and accommodations for those students who need them.

All manuals and user guides pertaining to the 2016–2017 online testing are available on the portal, and DAs, DTCs, and STCs can use these manuals and user guides to train TEs/TAs regarding test administration policies and procedures.

**Training Modules**

The following training modules were created to help users in the field understand the overall South Dakota Smarter Balanced assessments as well as how each system works. The modules were provided in two formats: a PowerPoint version and a narrated version.

*AIR Ways Interim Reporting Module*: This module explains how to navigate the system and to view interim assessment performance reports.

*Braille Training Module*: This module provides an overview of technology and features available to Test Administrators preparing to administer a braille test.

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

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

*Teacher Hand Scoring System (THSS)*: This module provides an overview of THSS. Teachers use the handscoring system for scoring items on the interim assessments.

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

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

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

### 2.3.2 District Trainings

The SDDOE provided district-wide trainings for 2016–2017.

## 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 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 they ensure authorized data access. All aspects of the system, including item development and review, test delivery, and reporting, are password-protected. Our systems use role-based security models that ensure that users may access only the data to which they are entitled and may edit data in accordance with their user rights.

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

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

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

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

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

All students must be enrolled or registered at their testing schools in order to take the online, paper-pencil, or braille assessments. Student enrollment information, including demographic data, is generated at the district level and uploaded directly into TIDE during the testing period.

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

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

### 2.4.2 System Security

The objective of system security is to ensure that all data are protected and accessed appropriately by the right user groups. It is about protecting data and maintaining data and system integrity as intended, including ensuring that all personal information is secured, that transferred data (whether sent or received)

is not altered in any way, that the data source is known, and that any service can only be performed by a specific, designated user.

**A hierarchy of control**: As described in Section 2.2, STCs, TAs, and TEs have well-defined roles and access to the testing system. When the TIDE window opens, the SDDOE creates a verified list of DAs that is uploaded into TIDE. DAs are then responsible for selecting and entering the DTCs' and STCs' information into TIDE, and STCs are responsible for entering TAs' and TEs' information in TIDE. Throughout the year, the DAs, DTCs, and STCs are also expected to delete information in TIDE for any staff members who have transferred to other schools, resigned, or no longer serve as TAs or TEs.

**Password protection**: All access points by different roles—at the state, district, and school levels—require a password to log in to the system. Newly added users receive separate passwords through their personal e-mail addresses assigned by the school.

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

### 2.4.3 Security of the Testing Environment

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

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

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

If a student needs to leave the room for a brief time, the TEs or 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. TEs and TAs are encouraged to minimize access to the testing rooms by posting signs in halls and entrances in order to promote optimum testing conditions; they should also post "TESTING—DO NOT DISTURB" signs on the doors of testing rooms.

**Seating Arrangements**

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

**After the Test**

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

For the paper-pencil versions, the *Paper-Pencil Test Administration Manual* for mathematics or ELA/L provides specific instructions on how to package and secure the test booklets in order to be returned to the testing contractor's office.

### 2.4.4 Test Security Violations

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

**Impropriety**: 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 (for example, student[s] leaving the testing room without authorization).

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

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

District and school personnel must document all test security incidents in the test security incident log. Districts send the security logs on an as-needed basis to the SDDOE Office of Assessment. This log is the document of record for all test security incidents and should be maintained at the district level and submitted to the SDDOE at the end of testing.

## 2.5 STUDENT PARTICIPATION

All students (including retained students) currently enrolled in grades 3–8 and 11 at public schools in South Dakota are required to participate in the South Dakota Smarter Balanced assessment. Students must be tested in the enrolled grade assessment; with the exception of grade 12 students, out-of-grade-level testing is not allowed for the administration of Smarter Balanced summative assessments.

### 2.5.1 Exempt Students

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

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

## 2.6 ONLINE TESTING FEATURES AND TESTING ACCOMMODATIONS

The Smarter Balanced Assessment Consortium's *Usability, Accessibility, and Accommodations Guidelines (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/L and mathematics. At the same time, the *Guidelines* support important instructional decisions about accessibility and accommodations for students who participate in the Smarter Balanced assessments.

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

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

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

For additional information about the availability of designated supports and accommodations, refer to the Smarter Balanced *Usability, Accessibility, and Accommodations Guidelines* for complete information at http://sd.portal.airast.org/resources/.

### 2.6.1 Online Universal Tools for ALL Students

Universal tools are the access features of an assessment or exam that are embedded or non-embedded components of the test administration system. Universal tools are available to all students based on their preference and selection, and they have been preset in TIDE. In 2016–2017 test administration, the following features of universal tools are available for *all* students to access. For specific information on how to access and use these features, refer to the *Test Administrator User Guide* at http://sd.portal.airast.org.

**Embedded Universal Tools**

*Zoom in*: The student can zoom in on test questions, text, or graphics.

*Highlight*: The student can highlight passages or sections of passages and test questions.

*Pause*: A student may 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/L performance task.

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

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

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

*Cross out response options*: A student may use the strikethrough function to cross out response options.

*Mark a question for review*: A student can mark a question 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.

*Mouse pointer*: Students may be given a mouse pointer with various colors or sizes. Pointer colors include black, green, yellow, red, and white.

*Take as much time as needed to complete a Smarter Balanced assessment*: Testing may be split across multiple sessions so that the testing does not interfere with class schedules. The CAT assessment must be

completed within 45 calendar days of its starting date. The performance tasks must be completed within 10 calendar days of the starting date.

*Spell check*: This is a writing tool for checking the spelling of words in student-generated responses. Spell check only gives an indication that a word is misspelled; it does not provide the correct spelling. This tool is available only with the specific items for which the Smarter Balanced item specifications indicated that it would be appropriate. Spell check is bundled with other embedded writing tools for all performance task full writes (planning, drafting, revising, and editing). A full write is the second part of a performance task.

*Writing tools*: Selected writing tools (i.e., bold, italics, bullets, and undo and re-do) are available for all student-generated responses.

**Non-Embedded Universal Tools**

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

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

*Scratch paper*: Scratch paper to make notes, write computations, or record responses may be made available. Only plain paper or lined paper is appropriate for ELA/L. Graph paper is required beginning in grade 6 and can be used on all mathematics assessments. A student can use an assistive technology device for scratch paper as long as the device is consistent with the child's IEP 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 Section 504 Plans. Consortium-approved accommodations do not compromise the learning expectations, construct, grade-level standard, or intended outcome of the assessments.

**Embedded Designated Supports**

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

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

*Text-to-speech* (for mathematics stimuli items, ELA/L items, not for 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.

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

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

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

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

**Non-Embedded Designated Supports**

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

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

*Color overlays*: Color transparencies are 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 provided for by the zoom universal tool.

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

*Read aloud (for mathematics items and ELA/L items but not passages)*: Text is read aloud to the student by a trained and qualified human reader who follows the administration guidelines provided in the *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 *Online Summative Test*

*Administration Manual* and in *Guidelines for Read Aloud, Test Reader*. All or portions of the content may be read aloud.

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

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

*Simplified Test Directions*: The test administrator simplifies or paraphrases the test directions found in the *Test Administration Manual* according to the Simplified Test Directions guidelines.

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

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

**Embedded Accommodations**

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

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

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

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

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

**Non-Embedded Accommodations**

*100s number table (grade 4 and above mathematics tests)*: A paper-based list of all the digits from 1 to 100 in table format will be available from Smarter Balanced for reference.

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

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

*Calculator (for grades 6–8 and 11 mathematics tests)*: A non-embedded calculator for students needing a special calculator, such as a braille calculator or a talking calculator, currently unavailable in the assessment platform.

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

*Print-on-demand*: Paper copies of either passages/stimuli and/or items are printed for students. For those students needing a paper copy of a passage or stimulus, permission to request printing must first be set in TIDE, or member's comparable platform. For those students needing a paper copy of one or more items, the SDDOE must be contacted by the school or district coordinator to review the student's case before setting the accommodations for the student.

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

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

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

Table 3 presents a list of universal tools, designated supports, and accommodations that were offered in the 2016–2017 administration. Tables 4–9 provide the number of students who were offered the accommodations and designated supports.

Table 3. SY 2016–2017 Universal Tools, Designated Supports, and Accommodations

| | Universal Tools | Designated Supports | Accommodations |
|---|---|---|---|
| Embedded | Breaks<br>Calculator[1]<br>Digital Notepad<br>English Dictionary[2]<br>English Glossary<br>Expandable Passages<br>Global Notes<br>Highlighter<br>Keyboard Navigation<br>Mark for Review<br>Mathematics Tools[3]<br>Mouse Pointer<br>Spell Check<br>Strikethrough<br>Writing Tools[4]<br>Zoom | Color Contrast<br>Masking<br>Text-to-Speech[5]<br>Translated Test Directions[6]<br>Translations (Glossary)[7]<br>Translations (Stacked) [8]<br>Turn off Any Universal Tools | American Sign Language[9]<br>Braille<br>Closed Captioning[10]<br>Streamline<br>Text-to-Speech[11] |
| Non-Embedded | Breaks<br>English Dictionary[12]<br>Scratch Paper<br>Thesaurus[13] | Bilingual Dictionary[14]<br>Color Contrast<br>Color Overlay<br>Magnification<br>Read Aloud[15]<br>Noise Buffers<br>Scribe[16]<br>Separate Setting<br>Simplified Test Directions<br>Translated Test Directions<br>Translations (Glossary)[17] | Abacus<br>Alternate Response Options[18]<br>Calculator[19]<br>Multiplication Table[20]<br>Print on Demand<br>Read Aloud[21]<br>Scribe<br>Speech-to-Text<br>100s Number Table[20] |

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

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

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

| Accommodations | Grade | | | | | | |
|---|---|---|---|---|---|---|---|
| | **3** | **4** | **5** | **6** | **7** | **8** | **11** |
| **Embedded Accommodations** | | | | | | | |
| American Sign Language | 1 | 1 | 6 | 2 | 2 | | 1 |
| Closed Captioning | 2 | 1 | | 2 | 4 | | 3 |
| Streamlined Mode | 31 | 17 | 13 | 12 | 5 | 4 | 4 |
| Text-to-Speech: Passages | | | | | | | |
| Text-to-Speech: Passages and Items | 782 | 692 | 645 | 777 | 670 | 623 | 289 |
| **Non-Embedded Accommodations** | | | | | | | |
| Alternate Response Options | 1 | 1 | | 1 | | | 1 |
| Print on Demand: Items | | 1 | 1 | | | | |
| Print on Demand: Stimuli | | | | 1 | | | |
| Print on Demand: Stimuli & Items | | | | 1 | | 1 | 1 |
| Read Aloud Stimuli | 13 | 9 | 8 | 26 | 21 | 14 | 6 |
| Scribe Items (Writing) | 18 | 30 | 31 | 12 | 17 | 13 | 3 |
| Speech-to-Text | 1 | 6 | 6 | 4 | 6 | 5 | 9 |

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

| Designated Supports | Subgroup | Grade | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | **3** | **4** | **5** | **6** | **7** | **8** | **11** |
| Color Contrast | Overall | 6 | 23 | 9 | 11 | 15 | 8 | |
| | LEP | | | | | | | |
| | IDEA Eligible | 3 | 4 | 4 | 2 | 3 | 4 | |
| Masking | Overall | 15 | 17 | 12 | 6 | 6 | 11 | 2 |
| | LEP | 8 | 3 | 4 | 1 | | 1 | 2 |
| | IDEA Eligible | 9 | 13 | 10 | 5 | 6 | 10 | |
| Text-to-Speech: Items | Overall | 1,107 | 962 | 890 | 896 | 767 | 666 | 294 |
| | LEP | 248 | 154 | 113 | 97 | 88 | 117 | 66 |
| | IDEA Eligible | 776 | 739 | 708 | 608 | 542 | 417 | 218 |
| Text-to-Speech: Stimuli | Overall | 3 | 2 | 2 | 2 | 3 | 5 | 2 |
| | LEP | | | | | | | 1 |
| | IDEA Eligible | 3 | 1 | 2 | | 1 | 3 | 2 |

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

| Designated Supports | Subgroup | Grade | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | 3 | 4 | 5 | 6 | 7 | 8 | 11 |
| Bilingual Dictionary | Overall | 4 | 3 | 1 | 2 | | | 1 |
| | LEP | 4 | 3 | 1 | 2 | | | 1 |
| | IDEA Eligible | 2 | 1 | | | | | |
| Color Contrast | Overall | 1 | 3 | 1 | 2 | | | 1 |
| | LEP | 1 | | | | | | |
| | IDEA Eligible | 1 | 1 | 1 | 2 | | | |
| Color Overlay | Overall | 3 | 3 | | 2 | 2 | 1 | 1 |
| | LEP | 1 | | | | | | |
| | IDEA Eligible | 3 | 1 | | 2 | 1 | 1 | |
| Magnification | Overall | 3 | 3 | 4 | 3 | 2 | 2 | 5 |
| | LEP | 1 | | | | | | 2 |
| | IDEA Eligible | 3 | 2 | 3 | 3 | 1 | 1 | 1 |
| Noise Buffers | Overall | 16 | 13 | 13 | 6 | 7 | 4 | 4 |
| | LEP | 8 | 4 | 3 | 2 | | 2 | 3 |
| | IDEA Eligible | 9 | 9 | 9 | 5 | 7 | 4 | 2 |
| Read Aloud Items | Overall | 47 | 58 | 51 | 62 | 38 | 34 | 23 |
| | LEP | 5 | 7 | 3 | 1 | 1 | 1 | 3 |
| | IDEA Eligible | 42 | 49 | 43 | 58 | 34 | 30 | 20 |
| Read Aloud Stimuli | Overall | 23 | 24 | 22 | 24 | 14 | 9 | 10 |
| | LEP | 1 | 3 | 3 | | | | 3 |
| | IDEA Eligible | 22 | 19 | 16 | 23 | 12 | 9 | 8 |
| Scribe Items (Non-Writing) | Overall | 25 | 22 | 17 | 13 | 11 | 7 | 6 |
| | LEP | | | | | | | |
| | IDEA Eligible | 23 | 18 | 16 | 12 | 11 | 4 | 6 |
| Separate Setting | Overall | 732 | 686 | 681 | 616 | 514 | 487 | 293 |
| | LEP | 187 | 106 | 84 | 47 | 42 | 48 | 14 |
| | IDEA Eligible | 555 | 581 | 592 | 536 | 436 | 404 | 270 |
| Simplified Test Directions | Overall | 139 | 98 | 116 | 66 | 32 | 46 | 28 |
| | LEP | 88 | 52 | 36 | 2 | 1 | | 5 |
| | IDEA Eligible | 57 | 50 | 86 | 55 | 30 | 42 | 24 |

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

| Accommodations | Grade | | | | | | |
|---|---|---|---|---|---|---|---|
| | 3 | 4 | 5 | 6 | 7 | 8 | 11 |
| **Embedded Accommodations** | | | | | | | |
| American Sign Language | 1 | 1 | 6 | 2 | 2 | | 1 |
| Streamlined Mode | 28 | 16 | 11 | 9 | 4 | 3 | 3 |
| **Non-Embedded Accommodations** | | | | | | | |
| Abacus | | | | | | | |
| Alternate Response Options | 1 | 1 | 1 | 1 | | | |
| Calculator | | | | 6 | 2 | 3 | 1 |
| Multiplication Table | | | | 165 | 168 | 120 | 23 |
| Print on Demand: Items | | 1 | 1 | | | | |
| Print on Demand: Stimuli and Items | | | | 1 | | 1 | 1 |
| Speech-to-Text | 1 | 5 | 7 | 4 | 6 | 6 | 6 |
| 100s Number Table | | 27 | 11 | 6 | 4 | 4 | 1 |

Table 8. Mathematics Total Students with Allowed Embedded Designated Supports

| Designated Supports | Subgroup | Grade | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | **3** | **4** | **5** | **6** | **7** | **8** | **11** |
| Color Contrast | Overall | 6 | 12 | 9 | 11 | 15 | 8 | |
| | LEP | | | | | | | |
| | IDEA Eligible | 3 | 4 | 4 | 2 | 3 | 4 | |
| Masking | Overall | 13 | 14 | 18 | 14 | 6 | 10 | 2 |
| | LEP | 8 | 3 | 4 | 1 | | 2 | 2 |
| | IDEA Eligible | 7 | 10 | 16 | 9 | 6 | 9 | |
| Translation (Glossary): Spanish | Overall | | 3 | 3 | 15 | | 4 | 3 |
| | LEP | | 3 | 3 | 15 | | 4 | 3 |
| | IDEA Eligible | | 3 | 3 | 15 | | 4 | 3 |
| Translation (Glossary): Other Languages | Overall | | 2 | | | 2 | 2 | |
| | LEP | | 2 | | | 1 | 2 | |
| | IDEA Eligible | | | | | | | |
| Text-to-Speech: Items | Overall | 195 | 191 | 183 | 155 | 109 | 77 | 40 |
| | LEP | 28 | 23 | 21 | 11 | 11 | 8 | 6 |
| | IDEA Eligible | 151 | 156 | 142 | 131 | 92 | 70 | 30 |
| Text-to-Speech: Stimuli | Overall | 6 | 4 | 2 | 2 | 3 | 3 | |
| | LEP | 6 | 4 | 2 | 2 | 3 | 3 | |
| | IDEA Eligible | | | | | | | |

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

| Designated Supports | Subgroup | Grade | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | **3** | **4** | **5** | **6** | **7** | **8** | **11** |
| Color Contrast | Overall | 1 | 3 | 1 | 3 | | | 1 |
| | LEP | 1 | | | | | | |
| | IDEA Eligible | 1 | 1 | 1 | 2 | | | |
| Color Overlay | Overall | 2 | 3 | | 3 | 2 | 1 | 1 |
| | LEP | 1 | | | | | | |
| | IDEA Eligible | 2 | 1 | | 2 | 1 | 1 | |
| Translation (Glossary): Spanish | Overall | 2 | 4 | 3 | 21 | 11 | 17 | 1 |
| | LEP | 2 | 4 | 3 | 21 | 11 | 17 | 1 |
| | IDEA Eligible | 1 | | | | | | |
| Translation (Glossary): Other Languages | Overall | | | | 1 | 5 | | |
| | LEP | | | | 1 | 5 | | |
| | IDEA Eligible | | | | | | | |
| Magnification | Overall | 3 | 3 | 4 | 3 | 2 | 2 | 4 |
| | LEP | 1 | | | | | | 2 |
| | IDEA Eligible | 3 | 2 | 3 | 3 | 1 | 1 | 1 |
| Noise Buffers | Overall | 16 | 11 | 12 | 6 | 7 | 5 | 3 |
| | LEP | 8 | 4 | 3 | 2 | | 3 | 2 |
| | IDEA Eligible | 9 | 9 | 9 | 5 | 7 | 4 | 1 |
| Read Aloud Items | Overall | 43 | 54 | 51 | 55 | 40 | 34 | 18 |
| | LEP | 5 | 7 | 5 | 1 | 1 | 1 | 1 |
| | IDEA Eligible | 38 | 47 | 40 | 51 | 37 | 30 | 15 |

| Designated Supports | Subgroup | Grade | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | 3 | 4 | 5 | 6 | 7 | 8 | 11 |
| Read Aloud Stimuli | Overall | 35 | 34 | 32 | 51 | 32 | 21 | 14 |
| | LEP | 1 | 4 | 4 | | | 1 | 3 |
| | IDEA Eligible | 34 | 29 | 24 | 49 | 30 | 20 | 12 |
| Scribe Items | Overall | 26 | 27 | 20 | 13 | 15 | 10 | 5 |
| | LEP | 1 | | | | | | |
| | IDEA Eligible | 24 | 23 | 19 | 12 | 14 | 7 | 5 |
| Separate Setting | Overall | 725 | 695 | 683 | 645 | 547 | 533 | 299 |
| | LEP | 186 | 113 | 86 | 70 | 60 | 77 | 14 |
| | IDEA Eligible | 544 | 583 | 588 | 544 | 452 | 420 | 277 |
| Simplified Test Directions | Overall | 159 | 121 | 135 | 66 | 33 | 50 | 27 |
| | LEP | 106 | 75 | 56 | 3 | 1 | 3 | 5 |
| | IDEA Eligible | 60 | 51 | 86 | 54 | 30 | 43 | 23 |
| Translated Test Directions | Overall | 2 | 2 | 7 | 3 | 1 | 3 | 2 |
| | LEP | 1 | 1 | 4 | 1 | 1 | 2 | 2 |
| | IDEA Eligible | 2 | 1 | 3 | 2 | | 1 | |

## 2.7 DATA FORENSICS PROGRAM

### 2.7.1 Data Forensics Report

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

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

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

### 2.7.2 Changes in Student Performance

Score changes between years are examined using a regression model. For between-year comparisons, the scores between past and current years are compared, with the current-year score regressed on the test score from the previous year and the number of days between test end days between two years to control the instruction time between the two test scores. Between-year comparisons are performed between the current year (e.g., 2017) and the year before the current year (e.g., 2016).

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

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

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

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

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

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

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

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

### 2.7.3 Item Response Time

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

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

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

### 2.7.4 Inconsistent Item Response Pattern (Person Fit)

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

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

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

The person-fit index is computed using a standardized log-likelihood statistic. Following Drasgow, Levine, and Williams (1985) and 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, TA, and school).

## 2.8 PREVENTION AND RECOVERY OF DISRUPTIONS IN TEST DELIVERY SYSTEM

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

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

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

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

### 2.8.1 High-Level System Architecture

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

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

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

**Student Machine**

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

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

- If connectivity is lost and restored within the designated time period, the student may be unaware of the momentary interruption.

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

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

**Test Delivery Satellites**

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

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

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

**Hub**

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

**Demographic and History Servers**

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

**Quality Assurance System**

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

**Database of Record**

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

### 2.8.2 Automated Backup and Recovery

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

### 2.8.3 Other Disruption Prevention and Recovery

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

- Our hosting provider has redundant power generators that can continue to operate for up to 60 hours without refueling. With the multiple refueling contracts that are in place, these generators can operate indefinitely.
- Our hosting provider has multiple redundancies in the flow of information to and from our data enters by partnering with nine different network providers. Each fiber carrier must enter the data center at separate physical points, protecting the data center from a complete service failure caused by an unlikely network cable cut.
- On the network level, we have redundant firewalls and load balancers throughout the environment.
- We use redundant power and switching within all of our server cabinets.
- Data are protected by nightly backups. We complete a full weekly backup and incremental backups nightly. Should a catastrophic event occur, AIR is able to reconstruct real time data using the data retained on the TDS satellites and hubs.
- The server backup agents send alerts to notify system administration staff in the event of a backup error, at which time they will inspect the error to determine whether the backup was successful or if they will need to rerun the backup.

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

# 3. SUMMARY OF 2016–2017 OPERATIONAL TEST ADMINISTRATION

## 3.1 STUDENT POPULATION

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

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

| Group | G3 | G4 | G5 | G6 | G7 | G8 | G11 |
|---|---|---|---|---|---|---|---|
| All Students | 11,398 | 11,390 | 11,049 | 10,902 | 10,565 | 10,165 | 9,032 |
| Female | 5,630 | 5,561 | 5,397 | 5,296 | 5,133 | 5,004 | 4,345 |
| Male | 5,768 | 5,829 | 5,652 | 5,606 | 5,432 | 5,161 | 4,687 |
| African American | 313 | 320 | 311 | 291 | 289 | 277 | 217 |
| Asian | 175 | 175 | 162 | 160 | 181 | 171 | 177 |
| Native Hawaiian/Pacific Islander | 4 | 12 | 10 | 10 | 12 | 8 | 15 |
| Hispanic/Latino | 657 | 665 | 559 | 603 | 469 | 466 | 388 |
| American Indian/Alaska Native | 1,737 | 1,764 | 1,676 | 1,676 | 1,685 | 1,594 | 1,043 |
| White | 7,990 | 7,990 | 7,910 | 7,805 | 7,595 | 7,396 | 7,001 |
| Multiple Ethnicities | 521 | 463 | 421 | 355 | 334 | 252 | 190 |
| LEP | 605 | 305 | 198 | 206 | 203 | 252 | 215 |
| IDEA | 1,737 | 1,669 | 1,517 | 1,417 | 1,285 | 1,091 | 718 |
| Section 504 | 247 | 229 | 266 | 265 | 262 | 234 | 206 |

Table 11. Number of Students in Summative Mathematics Assessment

| Group | G3 | G4 | G5 | G6 | G7 | G8 | G11 |
|---|---|---|---|---|---|---|---|
| All Students | 11,424 | 11,416 | 11,077 | 10,930 | 10,588 | 10,177 | 9,026 |
| Female | 5,646 | 5,582 | 5,406 | 5,305 | 5,147 | 5,007 | 4,340 |
| Male | 5,778 | 5,834 | 5,671 | 5,625 | 5,441 | 5,170 | 4,686 |
| African American | 324 | 335 | 319 | 303 | 301 | 291 | 216 |
| Asian | 181 | 182 | 170 | 165 | 182 | 177 | 179 |
| Native Hawaiian/Pacific Islander | 4 | 13 | 10 | 10 | 12 | 8 | 15 |
| Hispanic/Latino | 668 | 670 | 575 | 617 | 476 | 480 | 387 |
| American Indian/Alaska Native | 1,735 | 1,763 | 1,676 | 1,674 | 1,691 | 1,574 | 1,036 |
| White | 7,990 | 7,988 | 7,907 | 7,805 | 7,592 | 7,393 | 7,001 |
| Multiple Ethnicities | 521 | 464 | 420 | 354 | 334 | 253 | 191 |
| LEP | 632 | 336 | 231 | 237 | 229 | 287 | 213 |
| IDEA | 1,740 | 1,663 | 1,513 | 1,412 | 1,292 | 1,092 | 723 |
| Section 504 | 247 | 229 | 267 | 266 | 260 | 234 | 207 |

## 3.2 SUMMARY OF OVERALL STUDENT PERFORMANCE

Tables 12–17 present a summary of the 2016–2017 summative test results for all students and by subgroups, including the average and the standard deviation of scale scores, the percentage of students in each achievement level, and the percentage of proficient students. Figures 1–2 compare the percentage of proficient students in 2014–2015, 2015–2016, and 2016–17 for all students (cohort comparisons). The average and the standard deviation of scale scores, and the percentage of proficient students in both years are provided in Appendix B.

Table 12. ELA/L Percentage of Students in Achievement Levels for Overall and by Subgroups (Grades 3–4)

| Group | Number Tested | Scale Score Mean | Scale Score SD | % Level 1 | % Level 2 | % Level 3 | % Level 4 | % Proficient |
|---|---|---|---|---|---|---|---|---|
| **Grade 3** | | | | | | | | |
| All Students | 11,398 | 2419.65 | 87.18 | 29 | 25 | 24 | 23 | 47 |
| Female | 5,630 | 2427.58 | 87.29 | 26 | 24 | 24 | 26 | 50 |
| Male | 5,768 | 2411.91 | 86.39 | 31 | 25 | 24 | 20 | 44 |
| African American | 313 | 2387.44 | 82.28 | 44 | 23 | 21 | 12 | 33 |
| Asian | 175 | 2414.18 | 97.74 | 34 | 24 | 18 | 24 | 42 |
| Native Hawaiian/Pacific Islander | 4* | | | | | | | |
| Hispanic/Latino | 657 | 2390.11 | 78.32 | 40 | 30 | 19 | 12 | 31 |
| American Indian or Alaska Native | 1,737 | 2347.05 | 73.45 | 63 | 23 | 11 | 3 | 14 |
| White | 7,990 | 2439.70 | 80.99 | 19 | 25 | 28 | 28 | 56 |
| Multiple Ethnicities | 521 | 2412.88 | 84.60 | 32 | 23 | 25 | 20 | 45 |
| LEP | 605 | 2368.64 | 75.40 | 52 | 27 | 14 | 7 | 21 |
| IDEA | 1,737 | 2368.03 | 81.30 | 53 | 25 | 14 | 8 | 22 |
| Section 504 | 247 | 2414.45 | 83.67 | 29 | 25 | 26 | 20 | 46 |
| **Grade 4** | | | | | | | | |
| All Students | 11,390 | 2461.86 | 90.30 | 31 | 21 | 25 | 23 | 48 |
| Female | 5,561 | 2470.34 | 88.90 | 27 | 21 | 26 | 25 | 52 |
| Male | 5,829 | 2453.77 | 90.89 | 34 | 21 | 24 | 21 | 45 |
| African American | 320 | 2422.89 | 81.67 | 48 | 22 | 23 | 8 | 30 |
| Asian | 175 | 2466.66 | 94.60 | 29 | 23 | 20 | 28 | 48 |
| Native Hawaiian/Pacific Islander | 12 | 2450.44 | 61.79 | 25 | 25 | 50 | 0 | 50 |
| Hispanic/Latino | 665 | 2438.02 | 82.82 | 40 | 24 | 23 | 14 | 37 |
| American Indian or Alaska Native | 1,764 | 2385.24 | 80.67 | 67 | 17 | 12 | 4 | 16 |
| White | 7,990 | 2482.51 | 83.01 | 21 | 22 | 29 | 29 | 57 |
| Multiple Ethnicities | 463 | 2457.14 | 85.95 | 32 | 26 | 22 | 20 | 42 |
| LEP | 305 | 2377.29 | 73.66 | 73 | 14 | 9 | 3 | 12 |
| IDEA | 1,669 | 2398.24 | 86.09 | 62 | 18 | 12 | 8 | 20 |
| Section 504 | 229 | 2461.79 | 92.06 | 31 | 20 | 26 | 23 | 49 |

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

***** Suppressed the data due to the small sample size, n<10.

Table 13. ELA/L Percentage of Students in Achievement Levels for Overall and by Subgroups (Grades 5–7)

| Group | Number Tested | Scale Score Mean | Scale Score SD | % Level 1 | % Level 2 | % Level 3 | % Level 4 | % Proficient |
|---|---|---|---|---|---|---|---|---|
| **Grade 5** | | | | | | | | |
| All Students | 11,049 | 2494.33 | 94.31 | 29 | 22 | 32 | 18 | 50 |
| Female | 5,397 | 2506.42 | 92.23 | 24 | 21 | 33 | 22 | 55 |
| Male | 5,652 | 2482.78 | 94.84 | 33 | 22 | 30 | 15 | 45 |
| African American | 311 | 2459.09 | 88.26 | 42 | 24 | 25 | 9 | 34 |
| Asian | 162 | 2498.4 | 102.31 | 28 | 25 | 23 | 24 | 47 |
| Native Hawaiian/Pacific Islander | 10 | 2517.71 | 93.26 | 20 | 20 | 20 | 40 | 60 |
| Hispanic/Latino | 559 | 2463.7 | 90.28 | 40 | 27 | 24 | 9 | 34 |
| American Indian or Alaska Native | 1,676 | 2409.55 | 83.51 | 66 | 19 | 13 | 2 | 15 |
| White | 7,910 | 2515.89 | 85.49 | 19 | 22 | 37 | 22 | 59 |
| Multiple Ethnicities | 421 | 2491.29 | 90.13 | 30 | 24 | 29 | 16 | 45 |
| LEP | 198 | 2375.37 | 76.79 | 79 | 15 | 6 | 1 | 6 |
| IDEA | 1,517 | 2413.42 | 85.20 | 66 | 18 | 12 | 3 | 16 |
| Section 504 | 266 | 2497.66 | 94.73 | 27 | 26 | 28 | 19 | 47 |
| **Grade 6** | | | | | | | | |
| All Students | 10,902 | 2520.59 | 89.03 | 23 | 29 | 34 | 13 | 48 |
| Female | 5,296 | 2534.47 | 85.91 | 18 | 28 | 38 | 16 | 54 |
| Male | 5,606 | 2507.48 | 89.94 | 28 | 30 | 31 | 11 | 42 |
| African American | 291 | 2479.92 | 85.52 | 38 | 33 | 25 | 4 | 29 |
| Asian | 160 | 2533.34 | 106.25 | 23 | 22 | 33 | 22 | 55 |
| Native Hawaiian/Pacific Islander | 10 | 2502.49 | 59.84 | 30 | 40 | 30 | 0 | 30 |
| Hispanic/Latino | 603 | 2499.61 | 83.48 | 30 | 35 | 27 | 7 | 35 |
| American Indian or Alaska Native | 1,676 | 2443.98 | 81.94 | 57 | 28 | 13 | 2 | 15 |
| White | 7,805 | 2539.77 | 80.91 | 15 | 28 | 40 | 16 | 56 |
| Multiple Ethnicities | 355 | 2524.42 | 82.92 | 19 | 34 | 32 | 15 | 47 |
| LEP | 206 | 2405.02 | 73.62 | 76 | 21 | 2 | 1 | 3 |
| IDEA | 1,417 | 2435.93 | 75.49 | 61 | 28 | 10 | 1 | 11 |
| Section 504 | 265 | 2521.26 | 91.17 | 24 | 32 | 29 | 15 | 44 |
| **Grade 7** | | | | | | | | |
| All Students | 10,565 | 2546.31 | 92.16 | 23 | 25 | 39 | 13 | 52 |
| Female | 5,133 | 2559.5 | 89.63 | 19 | 23 | 42 | 15 | 58 |
| Male | 5,432 | 2533.84 | 92.80 | 27 | 27 | 36 | 10 | 46 |
| African American | 289 | 2504.94 | 90.71 | 38 | 33 | 23 | 7 | 30 |
| Asian | 181 | 2569.41 | 93.23 | 20 | 18 | 40 | 21 | 61 |
| Native Hawaiian/Pacific Islander | 12 | 2537.30 | 87.81 | 25 | 33 | 33 | 8 | 42 |
| Hispanic/Latino | 469 | 2520.66 | 83.77 | 29 | 31 | 34 | 5 | 39 |
| American Indian or Alaska Native | 1,685 | 2470.60 | 85.45 | 54 | 27 | 17 | 2 | 19 |
| White | 7,595 | 2566.10 | 84.39 | 15 | 24 | 45 | 16 | 61 |
| Multiple Ethnicities | 334 | 2537.67 | 86.53 | 25 | 28 | 37 | 10 | 46 |
| LEP | 203 | 2438.41 | 70.82 | 69 | 25 | 6 | 0 | 6 |
| IDEA | 1,285 | 2458.93 | 81.47 | 61 | 27 | 11 | 1 | 12 |
| Section 504 | 262 | 2546.50 | 90.23 | 21 | 29 | 37 | 13 | 50 |

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

***** Suppressed the data due to the small sample size, n<10.

Table 14. ELA/L Percentage of Students in Achievement Levels for Overall and by Subgroups (Grades 8 and 11)

| Group | Number Tested | Scale Score Mean | Scale Score SD | % Level 1 | % Level 2 | % Level 3 | % Level 4 | % Proficient |
|---|---|---|---|---|---|---|---|---|
| **Grade 8** | | | | | | | | |
| All Students | 10,165 | 2555.8 | 95.37 | 24 | 29 | 35 | 12 | 48 |
| Female | 5,004 | 2573.78 | 92.11 | 18 | 27 | 39 | 16 | 55 |
| Male | 5,161 | 2538.36 | 95.25 | 30 | 30 | 31 | 9 | 40 |
| African American | 277 | 2514.59 | 92.76 | 40 | 30 | 25 | 5 | 30 |
| Asian | 171 | 2554.20 | 114.3 | 33 | 18 | 31 | 18 | 49 |
| Native Hawaiian/Pacific Islander | 8* | | | | | | | |
| Hispanic/Latino | 466 | 2535.34 | 92.62 | 30 | 32 | 31 | 7 | 38 |
| American Indian or Alaska Native | 1,594 | 2476.79 | 85.79 | 56 | 28 | 14 | 2 | 16 |
| White | 7,396 | 2576.02 | 87.39 | 16 | 28 | 41 | 15 | 56 |
| Multiple Ethnicities | 252 | 2545.86 | 87.76 | 25 | 35 | 29 | 12 | 40 |
| LEP | 252 | 2442.75 | 70.84 | 74 | 22 | 4 | 0 | 4 |
| IDEA | 1,091 | 2462.37 | 78.10 | 64 | 26 | 8 | 1 | 10 |
| Section 504 | 234 | 2544.32 | 86.30 | 25 | 36 | 30 | 8 | 38 |
| **Grade 11** | | | | | | | | |
| All Students | 9,032 | 2608.06 | 105.82 | 15 | 21 | 38 | 26 | 64 |
| Female | 4,345 | 2624.95 | 100.32 | 11 | 18 | 40 | 30 | 70 |
| Male | 4,687 | 2592.41 | 108.35 | 18 | 24 | 37 | 21 | 58 |
| African American | 217 | 2518.85 | 109.84 | 41 | 29 | 23 | 8 | 30 |
| Asian | 177 | 2560.76 | 132.96 | 34 | 19 | 27 | 19 | 46 |
| Native Hawaiian/Pacific Islander | 15 | 2631.24 | 82.60 | 7 | 20 | 33 | 40 | 73 |
| Hispanic/Latino | 388 | 2571.41 | 102.75 | 23 | 25 | 39 | 13 | 52 |
| American Indian or Alaska Native | 1,043 | 2524.46 | 102.01 | 39 | 32 | 23 | 6 | 29 |
| White | 7,001 | 2626.44 | 97.29 | 10 | 19 | 41 | 30 | 71 |
| Multiple Ethnicities | 190 | 2609.11 | 98.00 | 13 | 23 | 39 | 25 | 64 |
| LEP | 215 | 2435.85 | 75.00 | 81 | 14 | 4 | 0 | 5 |
| IDEA | 718 | 2485.73 | 93.47 | 55 | 28 | 15 | 2 | 17 |
| Section 504 | 206 | 2613.87 | 96.30 | 10 | 22 | 43 | 25 | 68 |

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

***** Suppressed the data due to the small sample size, n<10.

Table 15. 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 | 11,424 | 2437.04 | 79.08 | 23 | 25 | 31 | 21 | 53 |
| Female | 5,646 | 2435.38 | 77.54 | 23 | 25 | 32 | 20 | 52 |
| Male | 5,778 | 2438.67 | 80.53 | 22 | 25 | 31 | 23 | 54 |
| African American | 324 | 2395.61 | 76.90 | 38 | 33 | 20 | 9 | 28 |
| Asian | 181 | 2428.30 | 96.17 | 30 | 20 | 32 | 18 | 50 |
| Native Hawaiian/Pacific Islander | 4* | | | | | | | |
| Hispanic/Latino | 668 | 2407.03 | 75.25 | 35 | 29 | 26 | 10 | 36 |
| American Indian or Alaska Native | 1,735 | 2368.70 | 73.07 | 57 | 26 | 14 | 3 | 18 |
| White | 7,990 | 2456.97 | 70.12 | 13 | 24 | 36 | 27 | 63 |
| Multiple Ethnicities | 521 | 2426.36 | 76.00 | 26 | 29 | 27 | 17 | 44 |
| LEP | 632 | 2385.03 | 75.21 | 47 | 28 | 19 | 6 | 25 |
| IDEA | 1,740 | 2392.40 | 83.73 | 44 | 25 | 21 | 10 | 31 |
| Section 504 | 247 | 2442.33 | 81.21 | 21 | 26 | 31 | 23 | 54 |
| **Grade 4** | | | | | | | | |
| All Students | 11,416 | 2476.83 | 80.73 | 20 | 31 | 30 | 19 | 49 |
| Female | 5,582 | 2472.98 | 77.39 | 21 | 33 | 30 | 16 | 46 |
| Male | 5,834 | 2480.50 | 83.65 | 20 | 29 | 30 | 21 | 51 |
| African American | 335 | 2426.00 | 83.03 | 38 | 37 | 18 | 6 | 24 |
| Asian | 182 | 2474.46 | 90.02 | 22 | 31 | 26 | 21 | 47 |
| Native Hawaiian/Pacific Islander | 13 | 2462.08 | 71.74 | 15 | 54 | 23 | 8 | 31 |
| Hispanic/Latino | 670 | 2450.88 | 74.78 | 30 | 38 | 23 | 10 | 33 |
| American Indian or Alaska Native | 1,763 | 2405.58 | 73.34 | 55 | 30 | 12 | 3 | 15 |
| White | 7,988 | 2497.34 | 71.63 | 11 | 30 | 35 | 24 | 59 |
| Multiple Ethnicities | 464 | 2469.88 | 76.73 | 23 | 36 | 25 | 16 | 41 |
| LEP | 336 | 2390.69 | 76.69 | 60 | 28 | 11 | 1 | 12 |
| IDEA | 1,663 | 2420.61 | 83.32 | 47 | 32 | 14 | 7 | 21 |
| Section 504 | 229 | 2477.69 | 88.23 | 22 | 31 | 26 | 21 | 47 |
| **Grade 5** | | | | | | | | |
| All Students | 11,077 | 2499.16 | 86.27 | 29 | 30 | 22 | 18 | 40 |
| Female | 5,406 | 2497.38 | 82.64 | 30 | 32 | 22 | 17 | 38 |
| Male | 5,671 | 2500.85 | 89.57 | 29 | 29 | 22 | 20 | 42 |
| African American | 319 | 2450.38 | 87.64 | 50 | 30 | 14 | 6 | 20 |
| Asian | 170 | 2496.00 | 101.47 | 36 | 24 | 22 | 18 | 40 |
| Native Hawaiian/Pacific Islander | 10 | 2516.55 | 72.07 | 20 | 40 | 20 | 20 | 40 |
| Hispanic/Latino | 575 | 2464.26 | 85.99 | 44 | 32 | 14 | 10 | 24 |
| American Indian or Alaska Native | 1,676 | 2420.18 | 76.56 | 69 | 22 | 6 | 3 | 9 |
| White | 7,907 | 2520.64 | 76.42 | 19 | 32 | 26 | 23 | 49 |
| Multiple Ethnicities | 420 | 2495.53 | 75.78 | 31 | 34 | 22 | 14 | 36 |
| LEP | 231 | 2382.99 | 77.50 | 84 | 13 | 1 | 2 | 3 |
| IDEA | 1,513 | 2431.02 | 82.54 | 63 | 24 | 8 | 5 | 13 |
| Section 504 | 267 | 2508.33 | 90.20 | 26 | 33 | 19 | 21 | 41 |

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

* Suppressed the data due to the small sample size, n<10.

Table 16. 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 | 10,930 | 2522.07 | 97.63 | 27 | 32 | 24 | 17 | 41 |
| Female | 5,305 | 2527.09 | 92.80 | 24 | 33 | 26 | 17 | 43 |
| Male | 5,625 | 2517.34 | 101.76 | 29 | 31 | 22 | 17 | 39 |
| African American | 303 | 2452.86 | 103.55 | 54 | 29 | 13 | 3 | 17 |
| Asian | 165 | 2542.37 | 118.41 | 24 | 24 | 25 | 27 | 52 |
| Native Hawaiian/Pacific Islander | 10 | 2544.47 | 63.65 | 20 | 50 | 20 | 10 | 30 |
| Hispanic/Latino | 617 | 2489.61 | 93.00 | 42 | 33 | 15 | 10 | 25 |
| American Indian or Alaska Native | 1,674 | 2434.53 | 94.95 | 62 | 27 | 8 | 2 | 10 |
| White | 7,805 | 2545.43 | 84.93 | 17 | 33 | 28 | 22 | 50 |
| Multiple Ethnicities | 354 | 2526.83 | 86.10 | 23 | 38 | 23 | 16 | 39 |
| LEP | 237 | 2388.13 | 106.23 | 81 | 14 | 3 | 2 | 5 |
| IDEA | 1,412 | 2427.64 | 97.74 | 66 | 25 | 7 | 2 | 9 |
| Section 504 | 266 | 2519.88 | 106.73 | 29 | 36 | 17 | 18 | 35 |
| **Grade 7** | | | | | | | | |
| All Students | 10,588 | 2542.64 | 102.70 | 26 | 30 | 25 | 18 | 43 |
| Female | 5,147 | 2542.65 | 98.21 | 26 | 31 | 26 | 17 | 43 |
| Male | 5,441 | 2542.62 | 106.78 | 27 | 29 | 24 | 19 | 44 |
| African American | 301 | 2481.76 | 113.96 | 50 | 27 | 15 | 8 | 23 |
| Asian | 182 | 2576.51 | 104.93 | 23 | 23 | 22 | 32 | 54 |
| Native Hawaiian/Pacific Islander | 12 | 2541.99 | 86.38 | 17 | 50 | 25 | 8 | 33 |
| Hispanic/Latino | 476 | 2500.15 | 98.49 | 43 | 33 | 17 | 8 | 25 |
| American Indian or Alaska Native | 1,691 | 2454.15 | 90.77 | 62 | 26 | 9 | 2 | 12 |
| White | 7,592 | 2567.12 | 91.67 | 16 | 31 | 30 | 23 | 53 |
| Multiple Ethnicities | 334 | 2531.15 | 100.34 | 31 | 31 | 24 | 15 | 39 |
| LEP | 229 | 2417.08 | 94.32 | 75 | 19 | 5 | 1 | 6 |
| IDEA | 1,292 | 2441.70 | 98.95 | 67 | 23 | 7 | 3 | 10 |
| Section 504 | 260 | 2543.16 | 99.44 | 27 | 32 | 21 | 20 | 41 |
| **Grade 8** | | | | | | | | |
| All Students | 10,177 | 2554.05 | 111.51 | 32 | 28 | 21 | 20 | 41 |
| Female | 5,007 | 2561.20 | 107.70 | 28 | 29 | 23 | 20 | 43 |
| Male | 5,170 | 2547.14 | 114.67 | 35 | 27 | 19 | 19 | 38 |
| African American | 291 | 2497.74 | 107.31 | 50 | 28 | 15 | 7 | 22 |
| Asian | 177 | 2559.34 | 145.46 | 37 | 25 | 13 | 25 | 38 |
| Native Hawaiian/Pacific Islander | 8* | | | | | | | |
| Hispanic/Latino | 480 | 2519.90 | 105.53 | 44 | 28 | 17 | 11 | 28 |
| American Indian or Alaska Native | 1,574 | 2452.67 | 95.16 | 72 | 20 | 6 | 3 | 8 |
| White | 7,393 | 2580.39 | 99.97 | 21 | 29 | 25 | 24 | 49 |
| Multiple Ethnicities | 253 | 2541.38 | 110.11 | 36 | 28 | 21 | 16 | 37 |
| LEP | 287 | 2419.41 | 88.75 | 81 | 15 | 3 | 0 | 4 |
| IDEA | 1,092 | 2443.73 | 94.60 | 74 | 20 | 5 | 2 | 7 |
| Section 504 | 234 | 2552.77 | 100.56 | 35 | 29 | 21 | 15 | 36 |

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

***** Suppressed the data due to the small sample size, n<10.

Table 17. Mathematics Percentage of Students in Achievement Levels for Overall and by Subgroups (Grade 11)

| Group | Number Tested | Scale Score Mean | Scale Score SD | % Level 1 | % Level 2 | % Level 3 | % Level 4 | % Proficient |
|---|---|---|---|---|---|---|---|---|
| **Grade 11** | | | | | | | | |
| All Students | 9,026 | 2590.15 | 114.24 | 32 | 29 | 27 | 13 | 40 |
| Female | 4,340 | 2593.42 | 107.65 | 29 | 29 | 30 | 12 | 41 |
| Male | 4,686 | 2587.12 | 119.96 | 34 | 28 | 25 | 14 | 38 |
| African American | 216 | 2502.37 | 113.16 | 64 | 20 | 12 | 3 | 15 |
| Asian | 179 | 2570.99 | 131.29 | 42 | 23 | 22 | 13 | 35 |
| Native Hawaiian/Pacific Islander | 15 | 2592.63 | 119.96 | 33 | 27 | 20 | 20 | 40 |
| Hispanic/Latino | 387 | 2547.09 | 111.03 | 44 | 33 | 18 | 5 | 23 |
| American Indian or Alaska Native | 1,036 | 2483.10 | 106.16 | 72 | 19 | 8 | 2 | 9 |
| White | 7,001 | 2611.92 | 103.89 | 24 | 30 | 31 | 15 | 46 |
| Multiple Ethnicities | 191 | 2577.27 | 105.74 | 34 | 37 | 21 | 8 | 30 |
| LEP | 213 | 2423.59 | 90.11 | 90 | 8 | 1 | 0 | 2 |
| IDEA | 723 | 2450.15 | 101.78 | 83 | 13 | 4 | 1 | 5 |
| Section 504 | 207 | 2591.37 | 106.17 | 29 | 37 | 25 | 10 | 34 |

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

***** Suppressed the data due to the small sample size, n<10.

Figure 1. ELA/L %Proficient Across Years

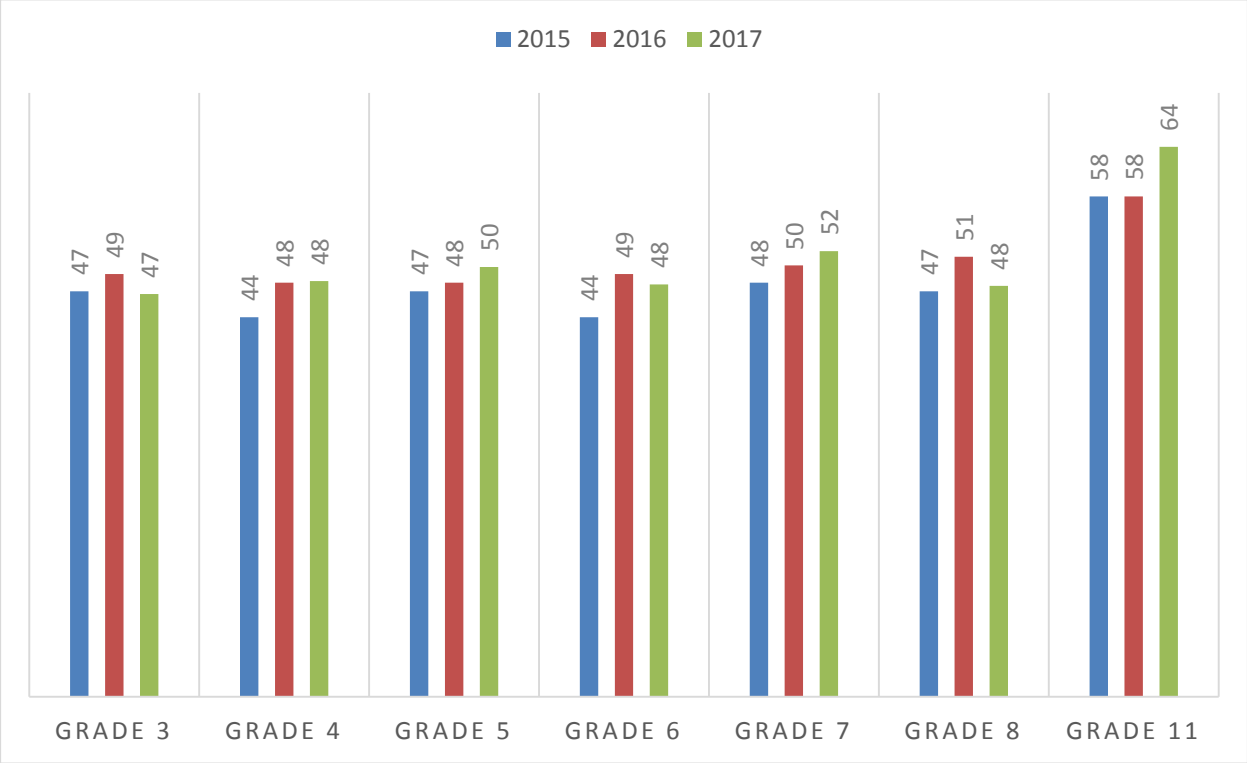
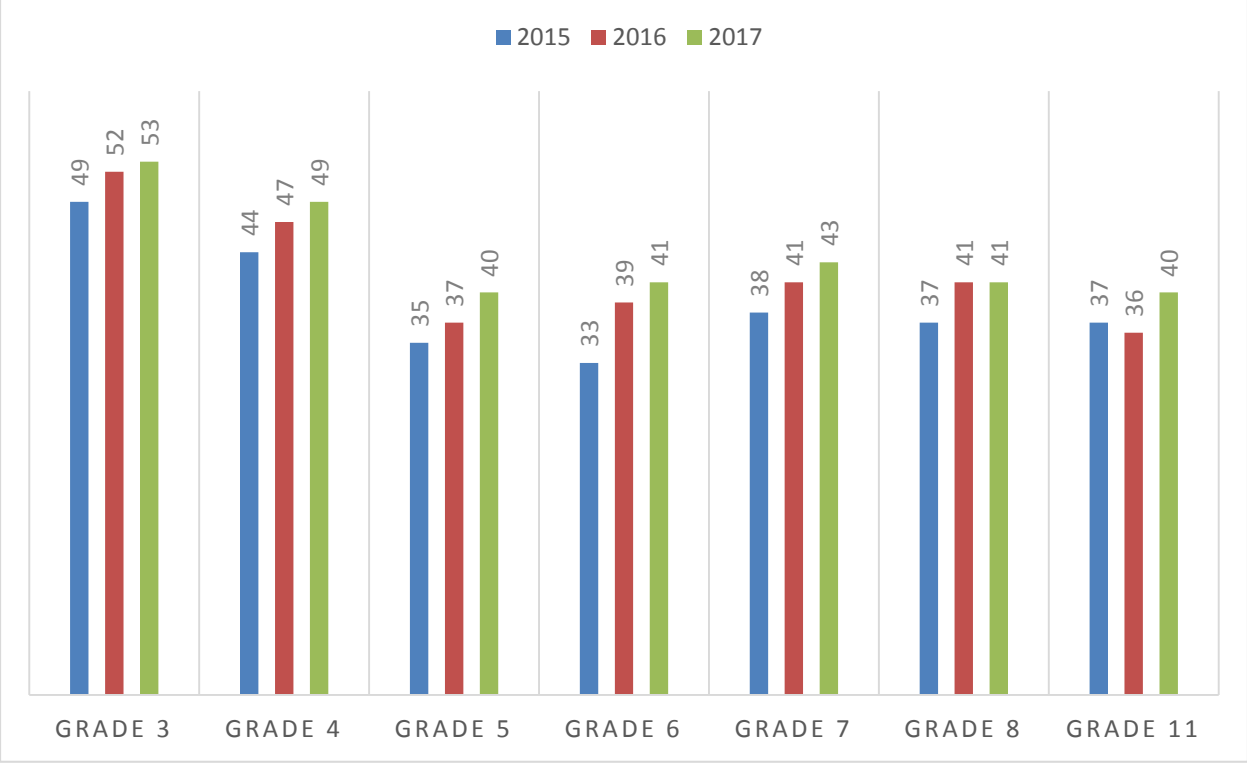



Figure 2. Mathematics %Proficient Across Years



For the reporting categories, because the precision of scores in each reporting category is not sufficient to report scores, given a small number of items, the scores on each reporting category are reported using one of the three performance categories, taking into account the SEM of the reporting category score: (1) Below standard, (2) At/Near standard, or (3) Above standard (see Section 6.5 for the rules). Tables 18 and 19 present the distribution of performance categories for each reporting category. The reporting categories are four claims in ELA/L, and three claims in mathematics, combining claims 2 and 4.

Table 18. ELA/L Percentage of Students in Performance Categories for Reporting Categories

| Grade | Performance Category | Claim 1: Reading | Claim 2: Writing | Claim 3: Listening | Claim 4: Research |
|---|---|---|---|---|---|
| 3 | Below | 35 | 29 | 18 | 26 |
| | At/Near | 42 | 47 | 64 | 52 |
| | Above | 23 | 24 | 19 | 21 |
| 4 | Below | 26 | 30 | 22 | 24 |
| | At/Near | 49 | 49 | 60 | 55 |
| | Above | 25 | 21 | 18 | 21 |
| 5 | Below | 27 | 27 | 20 | 28 |
| | At/Near | 47 | 49 | 63 | 49 |
| | Above | 25 | 25 | 17 | 23 |
| 6 | Below | 27 | 32 | 16 | 22 |
| | At/Near | 51 | 48 | 68 | 57 |
| | Above | 22 | 21 | 16 | 21 |
| 7 | Below | 26 | 26 | 19 | 21 |
| | At/Near | 48 | 50 | 67 | 56 |
| | Above | 26 | 24 | 14 | 22 |
| 8 | Below | 29 | 28 | 16 | 25 |
| | At/Near | 46 | 51 | 68 | 55 |
| | Above | 25 | 21 | 16 | 20 |
| 11 | Below | 18 | 16 | 13 | 17 |
| | At/Near | 46 | 45 | 60 | 55 |
| | Above | 36 | 38 | 27 | 28 |

Table 19. Mathematics Percentage of Students in Performance Categories for Reporting Categories

| Grade | Performance Category | Claim 1: Concepts and Procedures | Claims 2 & 4: Problem Solving & Modeling and Data Analysis | Claim 3: Communicating Reasoning |
|---|---|---|---|---|
| 3 | Below | 29 | 22 | 21 |
| | At/Near | 36 | 49 | 50 |
| | Above | 35 | 29 | 29 |
| 4 | Below | 33 | 27 | 27 |
| | At/Near | 35 | 49 | 48 |
| | Above | 32 | 24 | 26 |
| 5 | Below | 41 | 30 | 31 |
| | At/Near | 35 | 49 | 51 |
| | Above | 24 | 21 | 18 |
| 6 | Below | 37 | 32 | 31 |
| | At/Near | 38 | 49 | 49 |
| | Above | 25 | 19 | 19 |
| 7 | Below | 36 | 26 | 24 |
| | At/Near | 37 | 50 | 57 |
| | Above | 27 | 24 | 20 |
| 8 | Below | 37 | 31 | 28 |
| | At/Near | 37 | 45 | 53 |
| | Above | 25 | 24 | 18 |
| 11 | Below | 43 | 27 | 21 |
| | At/Near | 33 | 53 | 61 |
| | Above | 24 | 20 | 18 |

## 3.3 TEST TAKING TIME

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

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

Tables 20 and 21 present an average testing time and the testing time at percentiles for the overall test, the CAT component, and the PT component.

Table 20. ELA/L Test Taking Time

| Grade | Average Testing Time (hh:mm) | SD of Testing Time (hh:mm) | Testing Time in Percentiles (hh:mm) | | | | |
|---|---|---|---|---|---|---|---|
| | | | 75th | 80th | 85th | 90th | 95th |
| **Overall Test** | | | | | | | |
| 3 | 3:16 | 1:41 | 4:02 | 4:23 | 4:46 | 5:21 | 6:25 |
| 4 | 3:24 | 1:39 | 4:08 | 4:27 | 4:51 | 5:27 | 6:24 |
| 5 | 3:02 | 1:16 | 3:41 | 3:55 | 4:12 | 4:36 | 5:16 |
| 6 | 2:53 | 1:12 | 3:29 | 3:41 | 3:58 | 4:21 | 5:02 |
| 7 | 2:42 | 1:09 | 3:13 | 3:25 | 3:42 | 4:04 | 4:44 |
| 8 | 2:28 | 1:04 | 2:58 | 3:09 | 3:23 | 3:46 | 4:24 |
| 11 | 2:07 | 0:50 | 2:36 | 2:44 | 2:53 | 3:09 | 3:32 |
| **CAT Component** | | | | | | | |
| 3 | 1:43 | 0:49 | 2:05 | 2:14 | 2:25 | 2:41 | 3:10 |
| 4 | 1:49 | 0:47 | 2:11 | 2:19 | 2:29 | 2:44 | 3:11 |
| 5 | 1:39 | 0:37 | 1:58 | 2:04 | 2:13 | 2:25 | 2:46 |
| 6 | 1:38 | 0:37 | 1:57 | 2:03 | 2:11 | 2:22 | 2:42 |
| 7 | 1:33 | 0:34 | 1:50 | 1:56 | 2:03 | 2:14 | 2:31 |
| 8 | 1:24 | 0:32 | 1:40 | 1:45 | 1:52 | 2:02 | 2:18 |
| 11 | 1:11 | 0:25 | 1:25 | 1:29 | 1:34 | 1:41 | 1:52 |
| **PT Component** | | | | | | | |
| 3 | 1:33 | 1:05 | 2:02 | 2:14 | 2:29 | 2:54 | 3:36 |
| 4 | 1:35 | 1:04 | 2:02 | 2:15 | 2:30 | 2:52 | 3:35 |
| 5 | 1:23 | 0:50 | 1:46 | 1:56 | 2:08 | 2:25 | 2:53 |
| 6 | 1:15 | 0:46 | 1:34 | 1:44 | 1:54 | 2:09 | 2:38 |
| 7 | 1:09 | 0:44 | 1:27 | 1:34 | 1:44 | 1:59 | 2:25 |
| 8 | 1:04 | 0:40 | 1:22 | 1:29 | 1:39 | 1:51 | 2:15 |
| 11 | 0:56 | 0:31 | 1:13 | 1:19 | 1:26 | 1:35 | 1:51 |

Table 21. Mathematics Test Taking Time

| Grade | Average Testing Time (hh:mm) | SD of Testing Time (hh:mm) | Testing Time in Percentiles (hh:mm) | | | | |
|---|---|---|---|---|---|---|---|
| | | | 75th | 80th | 85th | 90th | 95th |
| **Overall Test** | | | | | | | |
| 3 | 1:50 | 0:56 | 2:12 | 2:24 | 2:39 | 2:58 | 3:36 |
| 4 | 1:50 | 0:53 | 2:13 | 2:23 | 2:36 | 2:54 | 3:25 |
| 5 | 1:59 | 0:55 | 2:25 | 2:35 | 2:49 | 3:08 | 3:39 |
| 6 | 1:56 | 0:48 | 2:17 | 2:27 | 2:37 | 2:53 | 3:20 |
| 7 | 1:37 | 0:40 | 1:56 | 2:04 | 2:13 | 2:25 | 2:47 |
| 8 | 1:45 | 0:46 | 2:08 | 2:16 | 2:26 | 2:41 | 3:06 |
| 11 | 1:30 | 0:36 | 1:50 | 1:56 | 2:05 | 2:15 | 2:33 |
| **CAT Component** | | | | | | | |
| 3 | 1:11 | 0:37 | 1:26 | 1:33 | 1:43 | 1:57 | 2:22 |
| 4 | 1:15 | 0:37 | 1:31 | 1:39 | 1:49 | 2:02 | 2:23 |
| 5 | 1:12 | 0:32 | 1:27 | 1:33 | 1:41 | 1:52 | 2:10 |
| 6 | 1:18 | 0:32 | 1:32 | 1:38 | 1:46 | 1:57 | 2:14 |
| 7 | 1:13 | 0:30 | 1:28 | 1:34 | 1:41 | 1:49 | 2:07 |
| 8 | 1:17 | 0:34 | 1:34 | 1:40 | 1:48 | 1:59 | 2:19 |
| 11 | 1:06 | 0:27 | 1:21 | 1:25 | 1:31 | 1:39 | 1:52 |
| **PT Component** | | | | | | | |
| 3 | 0:38 | 0:26 | 0:48 | 0:53 | 1:00 | 1:10 | 1:29 |
| 4 | 0:34 | 0:23 | 0:43 | 0:48 | 0:53 | 1:01 | 1:16 |
| 5 | 0:48 | 0:31 | 1:00 | 1:06 | 1:13 | 1:24 | 1:44 |
| 6 | 0:38 | 0:23 | 0:47 | 0:51 | 0:57 | 1:04 | 1:18 |
| 7 | 0:24 | 0:16 | 0:30 | 0:32 | 0:36 | 0:42 | 0:52 |
| 8 | 0:27 | 0:17 | 0:35 | 0:38 | 0:42 | 0:48 | 0:59 |
| 11 | 0:24 | 0:14 | 0:31 | 0:33 | 0:37 | 0:42 | 0:49 |

## 3.4 STUDENT ABILITY–ITEM DIFFICULTY DISTRIBUTION FOR THE 2016–2017 OPERATIONAL ITEM POOL

Figures 3 and 4 display the empirical distribution of the South Dakota student scale scores in the 2016–2017 administration and the distribution of the summative item difficulty parameters in the operational pool. The student ability distribution is shifted to the left in all grades and subjects, more pronounced in the mathematics upper grades, indicating that the pool includes more difficult items than the ability of students in the tested population. The pool includes difficult items to measure high performing students accurately but needs additional easy items to better measure low performing students. The Smarter Balanced Assessment Consortium plans to add additional easy items to the pool, and augment the pool in, proportion to the test blueprint constraints (e.g., content, Depth-of-Knowledge (DoK), item type, item difficulties) to better measure low performing students. The Smarter Balanced plans to add more easy items to the pool, and augment the pool, proportional to the test blueprint constraints, e.g., content, Depth-of-Knowledge (DoK), item type, and item difficulties.

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

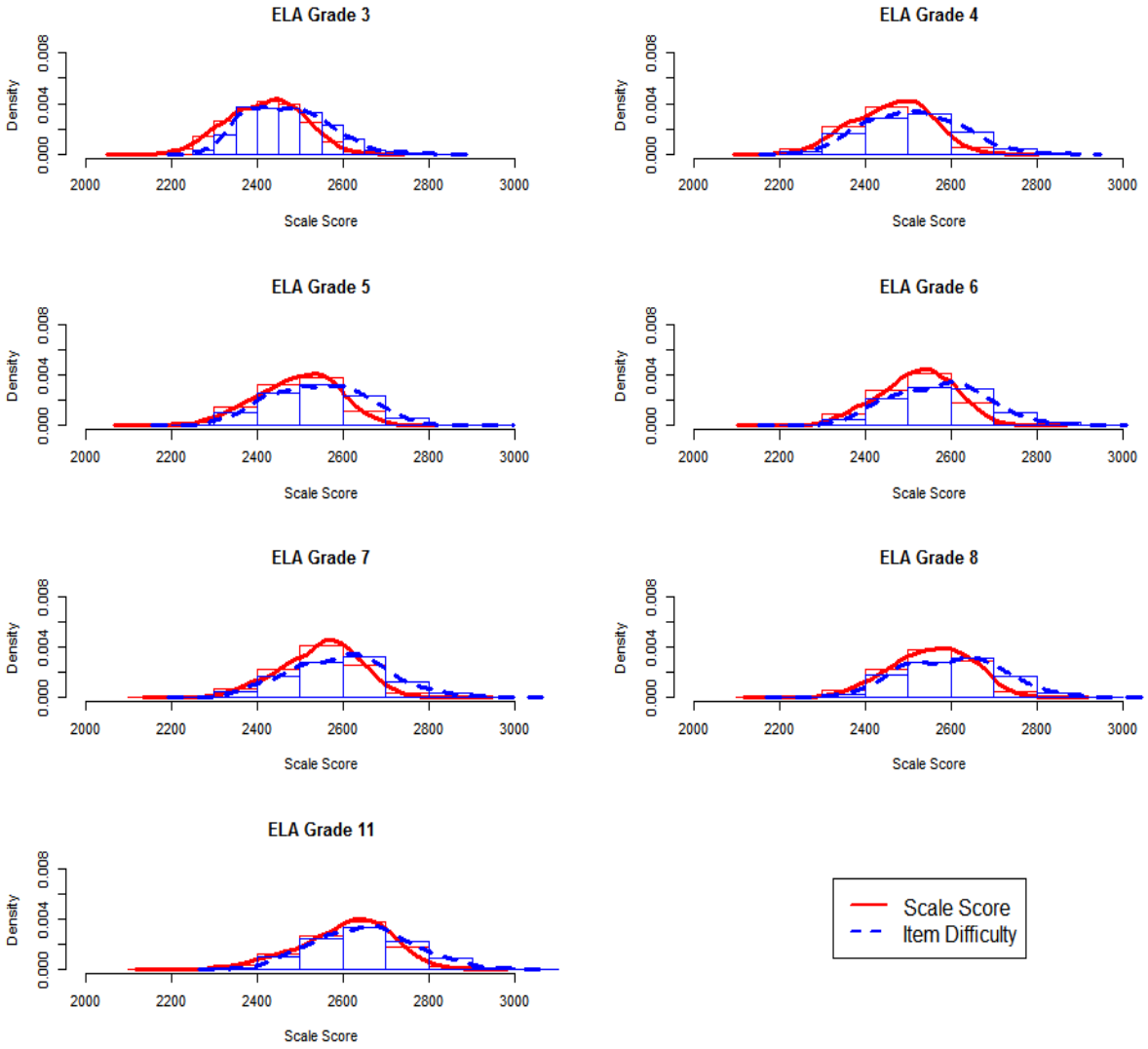
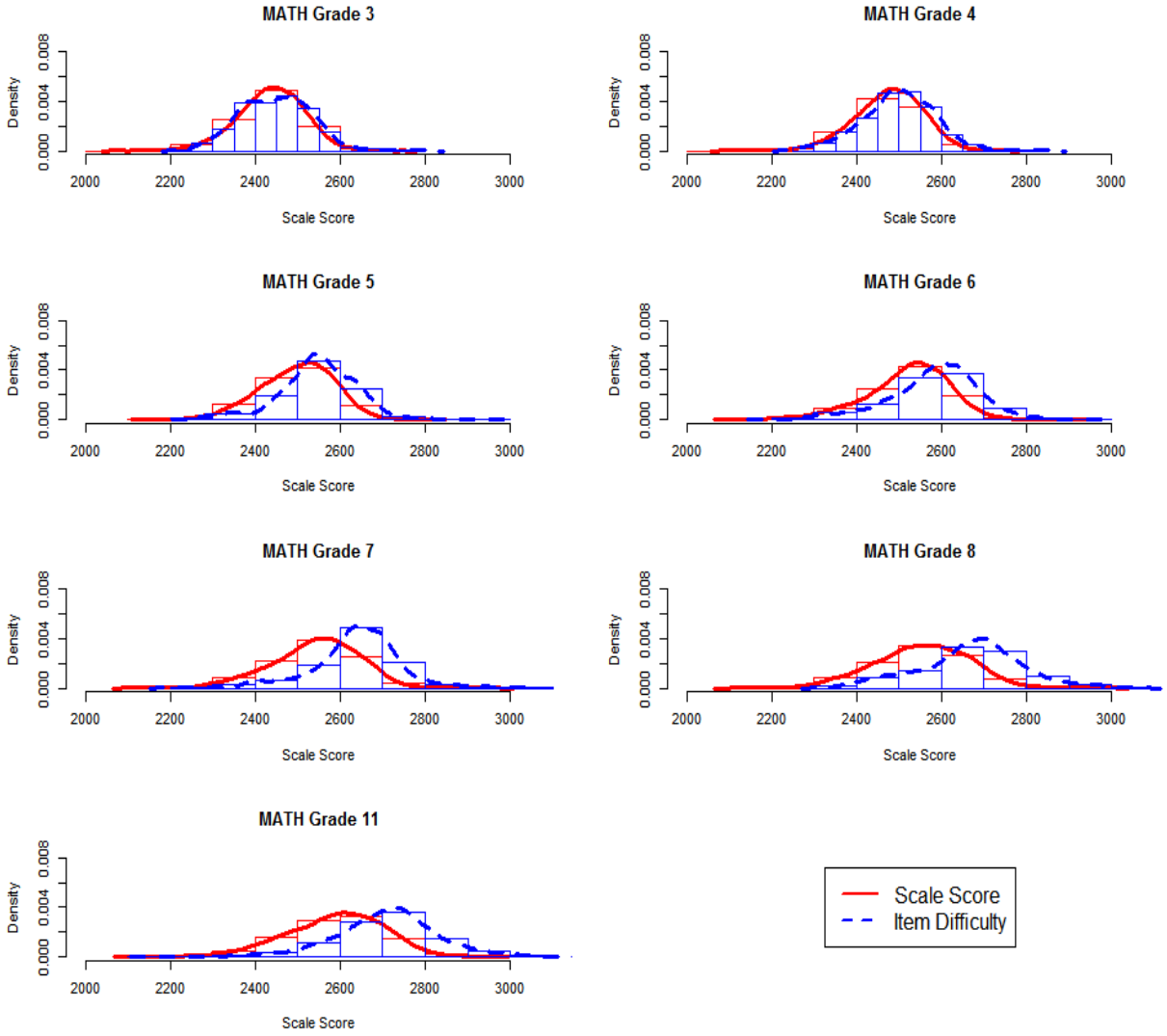


Figure 4. Student Ability–Item Difficulty Distribution for Mathematics



# 4. VALIDITY

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

Validity evidence provided in this chapter is as follows:

- Test Content
- Internal Structure

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

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

## 4.1 EVIDENCE ON TEST CONTENT

The Smarter Balanced summative assessment include 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 (South Dakota Smarter Balanced Summative Assessment Consortium, 2015) specify a range of items to be administered in each claim, content domain/standards, and targets. Moreover, blueprints constrain the DoK and item and passage types. In blueprints, all content blueprint elements are configured to obtain a strictly enforced range of items administered. The algorithm also seeks to satisfy target-level constraints, but these ranges are not strictly enforced. In ELA/L, the blueprints also specify the number of passages in reading (claim 1) and listening (claim 3) claims.

Tables 22–23 present the percentages of tests aligned with the test blueprint constraints for ELA/L CAT. Table 22 provides the blueprint match rates for item and passage requirements for each claim. All tests met the requirements, except for claim 1 literary text in grade 11. For DoK and item type constraints, the Smarter Balanced blueprint specifies the minimum number of items, not the maximum. Table 23 presents the percentages of tests that satisfied the DoK and item type constraints for each claim. All tests met the requirement, except for the claim 2 DoK2 requirement in grades 3 and 6, which each administered one DoK2 item fewer than required in claim 2.

Tables 24–26 provide the percentages of tests aligned with the test blueprint constraints for mathematics CAT, the blueprint match rates for claims, DoK, and target constraints. In mathematics, all tests met all blueprint requirements, except for grade 8. In grade 8, the violation was in claim 1 no-calculator segment for Target B, Target C, and DoK2 or higher, administered one item fewer or one item more than the item requirement.

Table 22. Percentage of ELA/L Delivered Tests Meeting Blueprint Requirements for Each Claim and the Number of Passages Administered

| Grade | Claim | Min | Max | %BP Match for | %BP Match for Passage |
|---|---|---|---|---|---|
| 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% | |
| 11 | 1-IT | 11 | 12 | 100% | 100% |
| | 1-LT | 4 | 4 | 99% | 98% |
| | 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 23. ELA/L Percentage of Delivered Tests Meeting Blueprint Requirements for Depth-of-Knowledge and Item Type

| DoK and Item Type Constraints | Minimum Required | %Blueprint Match | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Grade 3 | Grade 4 | Grade 5 | Grade 6 | Grade 7 | Grade 8 | Grade 11 |
| Claim 1 DoK2 | 7 | 100% | 100% | 100% | 100% | 100% | 100% | 100% |
| Claim 1 DoK3 or higher | 2 | 100% | 100% | 100% | 100% | 100% | 100% | 100% |
| Claim 2 DoK2 | 4 | 91% | 100% | 100% | 70% | 100% | 100% | 100% |
| Claim 2 DoK3 or higher | 1 | 100% | 100% | 100% | 100% | 100% | 100% | 100% |
| Claim 2 Brief Write | 1 | 100% | 100% | 100% | 100% | 100% | 100% | 100% |
| Claim 3 DoK2 or higher | 3 | 100% | 100% | 100% | 100% | 100% | 100% | 100% |

Table 24. Mathematics Percentage of Delivered Tests Meeting Blueprint Requirements for Claims and Targets (Grades 3–5)

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

Table 25. Mathematics Percentage of Delivered Tests Meeting Blueprint Requirements for Claims and Targets (Grades 6–8)

| Claim | Target | Grade 6 | | Grade 7 | | Grade 8 | |
|---|---|---|---|---|---|---|---|
| | | Required Items | %BP Match | Required Items | %BP Match | Required Items | %BP Match |
| Total Adaptive Test Length | | 33 | 100% | 34 | 100% | 34 | 100% |
| 1-Calc | Overall | 6 | 100% | 10 | 100% | 14 | 100% |
| | *Priority Cluster* | 3 | 100% | 6 | 100% | 11 | 100% |
| | Target A | 2 | 100% | | | | |
| | Target G | 1 | 100% | | | | |
| | Targets A, D | | | 6 | 100% | | |
| | Target D | | | | | 4 | 100% |
| | Targets E, G | | | | | 4 | 100% |
| | Targets F, H | | | | | 3 | 100% |
| | *Supporting Cluster* | 3 | 100% | 4 | 100% | 3 | 100% |
| | Targets H, I, J | 3 | 100% | | | | |
| | Targets E, F | | | 2 | 100% | | |
| | Targets G, H, I | | | 2 | 100% | | |
| | Targets I, J | | | | | 3 | 100% |
| | DOK 2 or higher | 2 | 100% | 4 | 100% | 5 | 100% |
| 1-No Calc | Overall | 13 | 100% | 10 | 100% | 6 | 100% |
| | *Priority Cluster* | 11 | 100% | 9 | 100% | 4 | 100% |
| | Targets E, F | 6 | 100% | | | | |
| | Target A | 2 | 100% | | | | |
| | Target B | 1 | 100% | | | 2 | 88% |
| | Target D | 2 | 100% | 3 | 100% | | |
| | Target B, C | | | 6 | 100% | | |
| | Target C | | | | | 2 | 88% |
| | *Supporting Cluster* | 2 | 100% | 1 | 100% | 2 | 100% |
| | Target C | 2 | 100% | | | | |
| | Target E | | | 1 | 100% | | |
| | Target A | | | | | 2 | 100% |
| | DOK 2 or higher | 5 | 100% | 4 | 100% | 4 | 94% |
| 2 | Overall | 3 | 100% | 3 | 100% | 3 | 100% |
| | Target A | 2 | 100% | 2 | 100% | 2 | 100% |
| | Targets B, C, D | 1 | 100% | 1 | 100% | 1 | 100% |
| 3-Calc | Overall | 7 | 100% | 8 | 100% | 8 | 100% |
| | Targets A, D | 3 | 100% | 2-3 | 100% | 2-3 | 100% |
| | Targets B, E | 2-3 | 100% | 3 | 100% | 3 | 100% |
| | Targets C, F, G | 2 | 100% | 1-2 | 100% | 1-2 | 100% |
| | DOK 3 or higher | 1 | 100% | 2 | 100% | 2 | 100% |
| 3-No Calc | Overall | 1 | 100% | | | | |
| 4 | Overall | 3 | 100% | 3 | 100% | 3 | 100% |
| | Targets A, D | 1 | 100% | 1 | 100% | 1 | 100% |
| | Targets B, E | 1 | 100% | 1 | 100% | 1 | 100% |
| | Targets C, F | 1 | 100% | 1 | 100% | 1 | 100% |
| 2&4 | DOK 3 or higher | 2 | 100% | 2 | 100% | 2 | 100% |

Table 26. Mathematics Percentage of Delivered Tests Meeting Blueprint Requirements for Claims and Targets (Grades 11)

| Claim | Target | Grade 11 | |
|---|---|---|---|
| | | **Required Items** | **%BP Match** |
| Total Adaptive Test Length | | 36 | 100% |
| 1-Calc | Overall | 11 | 100% |
| | *Priority Cluster* | 8 | 100% |
| | Target E | 1 | 100% |
| | Target G | 2 | 100% |
| | Target J | 1 | 100% |
| | Target K | 1 | 100% |
| | Targets L, M, N | 3 | 100% |
| | *Supporting Cluster* | 3 | 100% |
| | Target O | 1 | 100% |
| | Target P | 1 | 100% |
| | Targets C | 1 | 100% |
| | DOK 2 or higher | 4 | 100% |
| 1-No Calc | Overall | 11 | 100% |
| | *Priority Cluster* | 8 | 100% |
| | Target D, E | 1 | 100% |
| | Target F | 1 | 100% |
| | Targets H, I | 3 | 100% |
| | Target J | 1 | 100% |
| | Target K | 1 | 100% |
| | Target M | 1 | 100% |
| | *Supporting Cluster* | 3 | 100% |
| | Target O | 1 | 100% |
| | Target P | 1 | 100% |
| | Target A, B | 1 | 100% |
| | DOK 2 or higher | 4 | 100% |
| 2 | Overall | 3 | 100% |
| | Target A | 2 | 100% |
| | Targets B, C, D | 1 | 100% |
| 3-Calc | Overall | 7 | 100% |
| | Targets A, D | 2-3 | 100% |
| | Targets B, E | 3 | 100% |
| | Targets C, F, G | 1-2 | 100% |
| | DOK 3 or higher | 2 | 100% |
| 3-No Calc | Overall | 1 | 100% |
| 4 | Overall | 3 | 100% |
| | Targets A, D | 1 | 100% |
| | Targets B, E | 1 | 100% |
| | Targets C, F | 1 | 100% |
| 2&4 | DOK 3 or higher | 2 | 100% |

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

Table 27. Average and the Range of the Number of Unique Targets Assessed Within Each Claim Across all Delivered Tests

| Grade | Total Targets in BP | | | | Mean | | | | Range (Minimum–Maximum) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | C1 | C2 | C3 | C4 | C1 | C2 | C3 | C4 | C1 | C2 | C3 | C4 |
| **ELA/L** | | | | | | | | | | | | |
| 3 | 14 | 5 | 1 | 3 | 10.22 | 4.01 | 1.00 | 3.00 | 8-13 | 3-5 | 1-1 | 3-3 |
| 4 | 14 | 5 | 1 | 3 | 10.35 | 4.15 | 1.00 | 3.00 | 8-13 | 3-5 | 1-1 | 3-3 |
| 5 | 14 | 5 | 1 | 3 | 10.12 | 4.73 | 1.00 | 3.00 | 7-13 | 3-5 | 1-1 | 3-3 |
| 6 | 14 | 5 | 1 | 3 | 9.27 | 4.10 | 1.00 | 3.00 | 8-11 | 3-5 | 1-1 | 3-3 |
| 7 | 14 | 5 | 1 | 3 | 9.39 | 4.94 | 1.00 | 3.00 | 8-11 | 3-5 | 1-1 | 3-3 |
| 8 | 14 | 5 | 1 | 3 | 9.43 | 4.00 | 1.00 | 3.00 | 8-11 | 3-4 | 1-1 | 3-3 |
| 11 | 14 | 5 | 1 | 3 | 9.23 | 4.00 | 1.00 | 3.00 | 6-11 | 3-4 | 1-1 | 3-3 |
| **Mathematics** | | | | | | | | | | | | |
| 3 | 11 | 4 | 6 | 6 | 10.78 | 2.00 | 5.54 | 3.01 | 10-11 | 2-3 | 4-6 | 3-4 |
| 4 | 12 | 4 | 6 | 6 | 10.00 | 2.00 | 5.46 | 3.01 | 10-10 | 2-3 | 3-6 | 3-4 |
| 5 | 11 | 4 | 6 | 6 | 9.00 | 2.01 | 5.29 | 3.01 | 9-9 | 2-3 | 3-6 | 3-4 |
| 6 | 10 | 4 | 6 | 6 | 9.97 | 2.00 | 4.84 | 3.01 | 8-10 | 2-3 | 3-7 | 3-4 |
| 7 | 9 | 3 | 7 | 6 | 8.00 | 2.00 | 4.81 | 3.01 | 8-8 | 2-3 | 3-6 | 3-4 |
| 8 | 10 | 4 | 7 | 6 | 10.00 | 2.00 | 4.83 | 3.01 | 10-10 | 2-3 | 3-6 | 3-4 |
| 11 | 16 | 4 | 7 | 6 | 14.82 | 2.01 | 4.96 | 3.01 | 14-15 | 2-3 | 3-7 | 3-4 |

An adaptive testing algorithm constructs a test form unique to each student, targeting the student's level of ability and meeting the test blueprints. Consequently, the test forms will not be statistically parallel (e.g., equal test difficulty). However, scores from the test should be comparable, and each test form should measure the same content, albeit with a different set of test items, ensuring the comparability of assessments in content and scores. The blueprint match and target coverage results demonstrate that all 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 South Dakota Smarter Balanced Summative Assessments assumes a single underlying latent trait, with achievement reported as a total score as well as scores for each reporting category measured. The evidence on the internal structure is examined based on the correlations among reporting category scores.

The correlations among reporting category scores, both observed (below diagonal) and corrected for attenuation (above diagonal), are presented in Tables 28 and 29. The correction for attenuation indicates what the correlation would be if reporting category scores could be measured with perfect reliability, corrected (adjusted) for measurement error estimates.

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

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

Because the reliabilities for reporting category scores are low, the performance of each reporting category scores is reported in three performance categories. The distribution of performance categories for each reporting category is provided in Tables 18–19, Section 3.2. Scale scores are not reported for reporting categories.

Table 28. Correlations among Reporting Categories for ELA/L

| Grade | Reporting Categories | Observed & Disattenuated Correlation | | | |
|---|---|---|---|---|---|
| | | Claim 1 | Claim 2 | Claim 3 | Claim 4 |
| 3 | Claim 1: Reading | | 0.88 | 0.94 | 0.93 |
| | Claim 2: Writing | 0.70 | | 0.87 | 0.89 |
| | Claim 3: Listening | 0.64 | 0.60 | | 0.92 |
| | Claim 4: Research | 0.68 | 0.65 | 0.58 | |
| 4 | Claim 1: Reading | | 0.88 | 0.90 | 0.92 |
| | Claim 2: Writing | 0.66 | | 0.83 | 0.89 |
| | Claim 3: Listening | 0.61 | 0.58 | | 0.91 |
| | Claim 4: Research | 0.64 | 0.64 | 0.59 | |
| 5 | Claim 1: Reading | | 0.89 | 0.92 | 0.94 |
| | Claim 2: Writing | 0.68 | | 0.84 | 0.91 |
| | Claim 3: Listening | 0.62 | 0.59 | | 0.92 |
| | Claim 4: Research | 0.70 | 0.69 | 0.63 | |
| 6 | Claim 1: Reading | | 0.90 | 0.95 | 0.93 |
| | Claim 2: Writing | 0.69 | | 0.90 | 0.92 |
| | Claim 3: Listening | 0.61 | 0.61 | | 0.93 |
| | Claim 4: Research | 0.65 | 0.67 | 0.58 | |
| 7 | Claim 1: Reading | | 0.88 | 0.90 | 0.95 |
| | Claim 2: Writing | 0.68 | | 0.86 | 0.93 |
| | Claim 3: Listening | 0.60 | 0.58 | | 0.91 |
| | Claim 4: Research | 0.68 | 0.68 | 0.56 | |
| 8 | Claim 1: Reading | | 0.91 | 0.92 | 0.95 |
| | Claim 2: Writing | 0.71 | | 0.88 | 0.93 |
| | Claim 3: Listening | 0.60 | 0.58 | | 0.90 |
| | Claim 4: Research | 0.68 | 0.68 | 0.55 | |
| 11 | Claim 1: Reading | | 0.93 | 0.93 | 0.94 |
| | Claim 2: Writing | 0.72 | | 0.88 | 0.95 |
| | Claim 3: Listening | 0.64 | 0.62 | | 0.90 |
| | Claim 4: Research | 0.68 | 0.69 | 0.58 | |

Table 29. Correlations among Reporting Categories for Mathematics

| Grade | Reporting Categories | Observed & Disattenuated Correlation | | |
|---|---|---|---|---|
| | | **Claim 1** | **Claim 2&4** | **Claim 3** |
| 3 | Claim 1 | | 0.97 | 0.94 |
| | Claim 2 & 4 | 0.78 | | 1 |
| | Claim 3 | 0.75 | 0.72 | |
| 4 | Claim 1 | | 0.97 | 0.97 |
| | Claim 2 & 4 | 0.80 | | 1 |
| | Claim 3 | 0.79 | 0.74 | |
| 5 | Claim 1 | | 1 | 0.96 |
| | Claim 2 & 4 | 0.78 | | 1 |
| | Claim 3 | 0.75 | 0.73 | |
| 6 | Claim 1 | | 1 | 0.96 |
| | Claim 2 & 4 | 0.81 | | 1 |
| | Claim 3 | 0.76 | 0.74 | |
| 7 | Claim 1 | | 1 | 0.98 |
| | Claim 2 & 4 | 0.79 | | 1 |
| | Claim 3 | 0.75 | 0.71 | |
| 8 | Claim 1 | | 1 | 0.98 |
| | Claim 2 & 4 | 0.78 | | 1 |
| | Claim 3 | 0.76 | 0.7 | |
| 11 | Claim 1 | | 1 | 0.95 |
| | Claim 2 & 4 | 0.76 | | 1 |
| | Claim 3 | 0.71 | 0.63 | |

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

# 5. RELIABILITY

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

The reliability evidence of the Smarter Balanced summative tests is provided with marginal reliability, 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 $\sigma^2$ is the variance of the scale score. The higher reliability coefficient, the greater the precision of the test.

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

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

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

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

Table 30. Marginal Reliability for ELA/L and Mathematics

| Grade | N | Number of Items Specified in Test Blueprint | | Marginal Reliability | Scale Score Mean | Scale Score SD | Average CSEM |
|---|---|---|---|---|---|---|---|
| | | Min | Max | | | | |
| **ELA/L** | | | | | | | |
| 3 | 11,398 | 41 | 44 | 0.92 | 2419.65 | 87.18 | 24.13 |
| 4 | 11,390 | 40 | 44 | 0.92 | 2461.86 | 90.30 | 26.20 |
| 5 | 11,049 | 41 | 45 | 0.92 | 2494.33 | 94.31 | 25.94 |
| 6 | 10,902 | 41 | 45 | 0.92 | 2520.59 | 89.03 | 25.91 |
| 7 | 10,565 | 41 | 45 | 0.92 | 2546.31 | 92.16 | 26.69 |
| 8 | 10,165 | 43 | 45 | 0.92 | 2555.80 | 95.37 | 27.51 |
| 11 | 9,032 | 42 | 45 | 0.92 | 2608.06 | 105.82 | 30.05 |
| **Mathematics** | | | | | | | |
| 3 | 11,424 | 39 | 40 | 0.94 | 2437.04 | 79.08 | 19.28 |
| 4 | 11,416 | 37 | 40 | 0.94 | 2476.83 | 80.73 | 19.48 |
| 5 | 11,077 | 38 | 40 | 0.93 | 2499.16 | 86.27 | 22.57 |
| 6 | 10,930 | 38 | 39 | 0.94 | 2522.07 | 97.63 | 24.67 |
| 7 | 10,588 | 38 | 40 | 0.93 | 2542.64 | 102.70 | 26.76 |
| 8 | 10,177 | 38 | 40 | 0.93 | 2554.05 | 111.51 | 29.54 |
| 11 | 9,026 | 40 | 42 | 0.93 | 2590.15 | 114.24 | 30.85 |

## 5.2 STANDARD ERROR CURVES

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

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

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

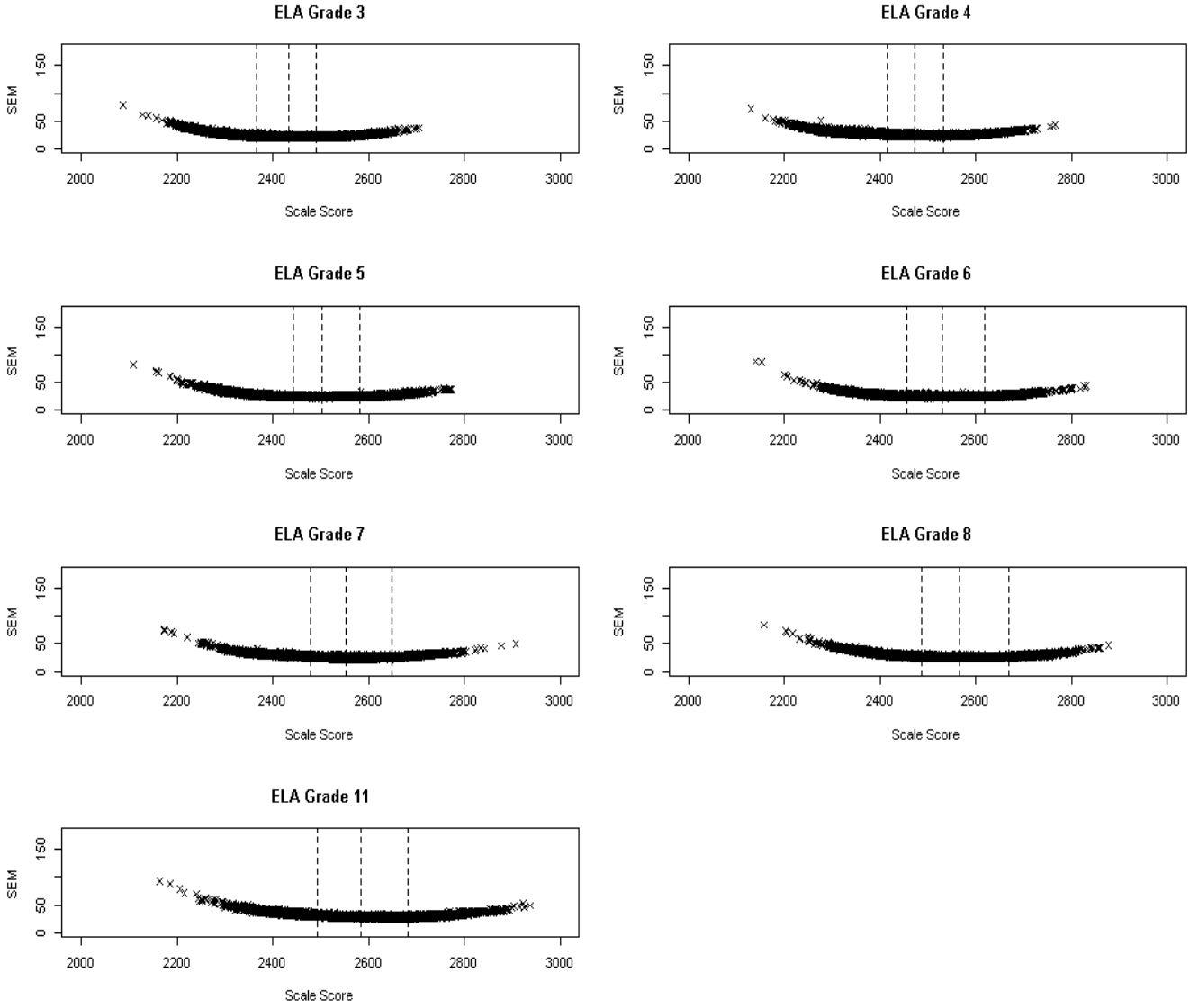
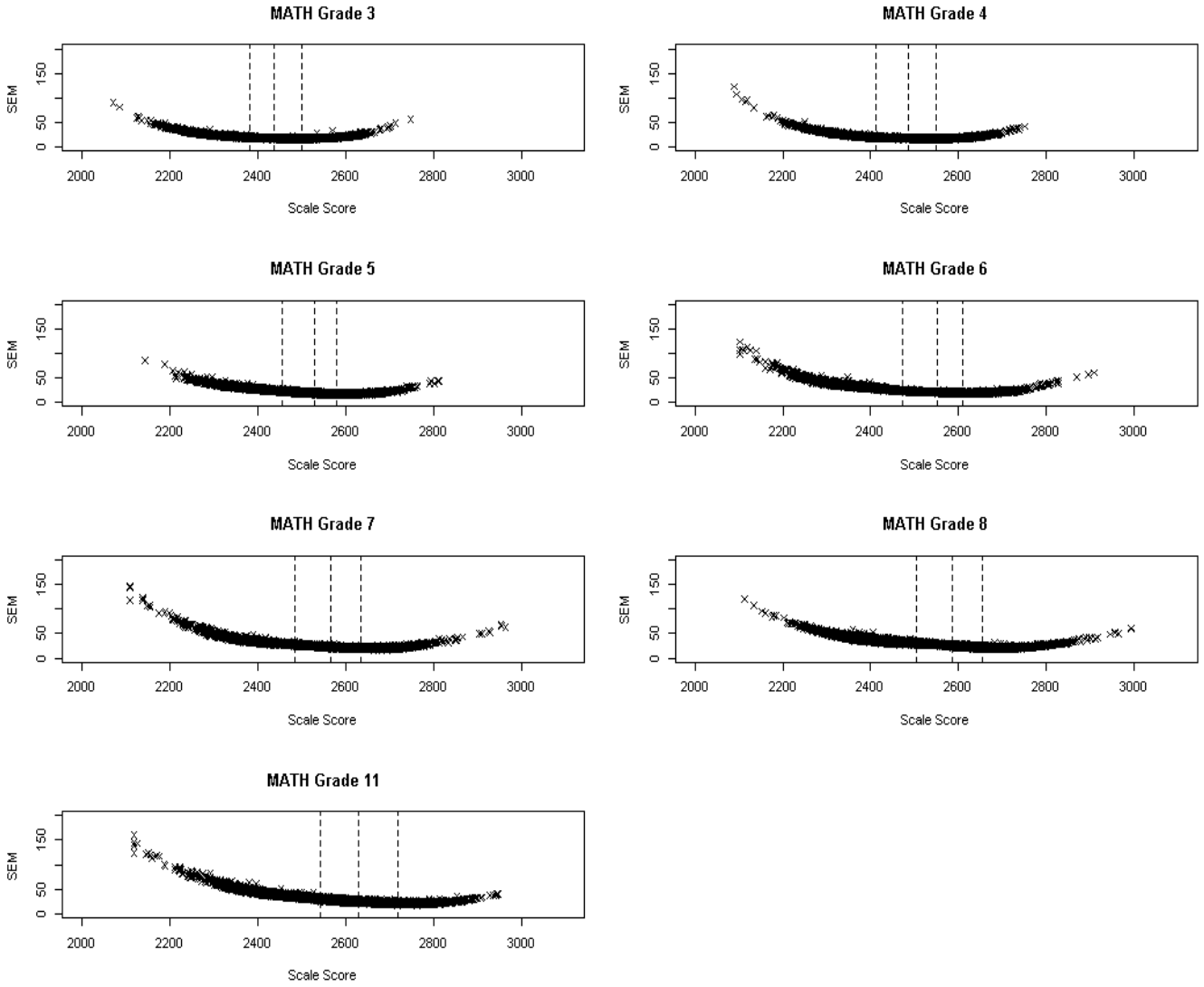


Figure 6. Conditional Standard Error of Measurement for Mathematics




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

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

| Grade | Level 1 | Level 2 | Level 3 | Level 4 | Average CSEM |
|---|---|---|---|---|---|
| **ELA/L** | | | | | |
| 3 | 27.65 | 22.52 | 22.00 | 23.24 | 24.13 |
| 4 | 28.75 | 25.45 | 24.31 | 25.31 | 26.20 |
| 5 | 28.62 | 24.27 | 24.17 | 26.43 | 25.94 |
| 6 | 28.68 | 24.58 | 24.69 | 26.68 | 25.91 |
| 7 | 30.13 | 25.63 | 24.82 | 27.69 | 26.69 |
| 8 | 30.73 | 26.18 | 25.88 | 28.44 | 27.51 |
| 11 | 35.95 | 29.27 | 27.75 | 30.18 | 30.05 |
| **Mathematics** | | | | | |
| 3 | 23.63 | 18.36 | 17.05 | 18.28 | 19.28 |
| 4 | 24.84 | 18.30 | 17.03 | 18.42 | 19.48 |
| 5 | 29.10 | 20.59 | 18.01 | 18.13 | 22.57 |
| 6 | 32.53 | 22.09 | 19.99 | 20.60 | 24.67 |
| 7 | 35.64 | 24.78 | 21.23 | 21.20 | 26.76 |
| 8 | 37.22 | 28.23 | 23.63 | 22.13 | 29.54 |
| 11 | 40.69 | 27.70 | 23.31 | 21.89 | 30.85 |

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

| Grade | L2 Cut | L3 Cut | L4 Cut | \|L2–L3\| | \|L3–L4\| | \|L2–L4\| |
|---|---|---|---|---|---|---|
| **ELA/L** | | | | | | |
| 3 | 23.24 | 21.93 | 21.83 | 1.31 | 0.10 | 1.41 |
| 4 | 26.23 | 24.87 | 23.98 | 1.36 | 0.89 | 2.25 |
| 5 | 24.52 | 23.98 | 24.89 | 0.54 | 0.91 | 0.37 |
| 6 | 25.29 | 24.28 | 24.54 | 1.01 | 0.26 | 0.75 |
| 7 | 26.22 | 24.92 | 25.67 | 1.30 | 0.75 | 0.55 |
| 8 | 26.86 | 25.85 | 26.39 | 1.01 | 0.54 | 0.47 |
| 11 | 30.61 | 28.57 | 27.99 | 2.04 | 0.58 | 2.62 |
| **Mathematics** | | | | | | |
| 3 | 19.52 | 17.61 | 16.64 | 1.91 | 0.97 | 2.88 |
| 4 | 19.53 | 17.51 | 16.97 | 2.02 | 0.54 | 2.56 |
| 5 | 23.06 | 18.68 | 17.46 | 4.38 | 1.22 | 5.60 |
| 6 | 23.91 | 20.97 | 19.47 | 2.94 | 1.50 | 4.44 |
| 7 | 27.66 | 22.51 | 20.40 | 5.15 | 2.11 | 7.26 |
| 8 | 30.37 | 25.05 | 21.71 | 5.32 | 3.34 | 8.66 |
| 11 | 30.20 | 25.19 | 21.54 | 5.01 | 3.65 | 8.66 |

## 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 $\theta_i$ is the unknown true ability of the *i*th student. The probability of the true score at achievement level *l* based on the cut scores $c_{l-1}$ and $c_l$ is estimated as

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

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 33 provides the percentages 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 $[-\infty, \text{L2 cut}]$ or L4 [L4 cut, $\infty$] is wider than the intervals used in L2 [L2 cut, L3 cut] and L3 [L3 cut, L4 cut]. The misclassification probability tends to be higher for narrow intervals. The classification indexes by subgroups are provided in Appendix C.

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

Table 33. Classification Accuracy and Consistency by Achievement Levels

| Grade | Achievement Level | ELA/L | | Mathematics | |
|---|---|---|---|---|---|
| | | % Accuracy | % Consistency | % Accuracy | % Consistency |
| 3 | Overall | 80 | 72 | 82 | 75 |
| | L1 | 90 | 84 | 89 | 83 |
| | L2 | 72 | 61 | 73 | 64 |
| | L3 | 70 | 60 | 79 | 71 |
| | L4 | 87 | 81 | 88 | 82 |
| 4 | Overall | 78 | 70 | 83 | 76 |
| | L1 | 90 | 85 | 90 | 83 |
| | L2 | 65 | 53 | 80 | 73 |
| | L3 | 68 | 58 | 79 | 71 |
| | L4 | 86 | 79 | 88 | 82 |
| 5 | Overall | 80 | 72 | 82 | 74 |
| | L1 | 91 | 85 | 90 | 85 |
| | L2 | 68 | 57 | 77 | 69 |
| | L3 | 76 | 68 | 71 | 62 |
| | L4 | 84 | 76 | 87 | 81 |
| 6 | Overall | 80 | 72 | 82 | 74 |
| | L1 | 89 | 82 | 91 | 85 |
| | L2 | 74 | 64 | 78 | 70 |
| | L3 | 77 | 70 | 72 | 62 |
| | L4 | 83 | 73 | 87 | 79 |
| 7 | Overall | 80 | 72 | 82 | 75 |
| | L1 | 89 | 82 | 90 | 84 |
| | L2 | 72 | 61 | 77 | 68 |
| | L3 | 80 | 74 | 75 | 66 |
| | L4 | 82 | 72 | 89 | 82 |
| 8 | Overall | 80 | 73 | 81 | 74 |
| | L1 | 89 | 82 | 90 | 84 |
| | L2 | 74 | 64 | 72 | 63 |
| | L3 | 80 | 73 | 71 | 61 |
| | L4 | 81 | 71 | 89 | 83 |
| 11 | Overall | 81 | 73 | 82 | 75 |
| | L1 | 88 | 81 | 91 | 86 |
| | L2 | 75 | 64 | 74 | 65 |
| | L3 | 78 | 71 | 80 | 72 |
| | L4 | 86 | 79 | 86 | 79 |

## 5.4 RELIABILITY FOR SUBGROUPS

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

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

| Subgroup | Grade 3 | Grade 4 | Grade 5 | Grade 6 | Grade 7 | Grade 8 | Grade 11 |
|---|---|---|---|---|---|---|---|
| All Students | 0.92 | 0.92 | 0.92 | 0.92 | 0.92 | 0.92 | 0.92 |
| Female | 0.93 | 0.91 | 0.92 | 0.91 | 0.91 | 0.91 | 0.91 |
| Male | 0.92 | 0.92 | 0.92 | 0.92 | 0.92 | 0.91 | 0.92 |
| African American | 0.91 | 0.89 | 0.91 | 0.90 | 0.91 | 0.91 | 0.91 |
| Asian | 0.94 | 0.92 | 0.93 | 0.94 | 0.92 | 0.94 | 0.94 |
| Native Hawaiian/ Pacific Islander | - | 0.83 | 0.93 | 0.83 | 0.91 | - | 0.88 |
| Hispanic/Latino | 0.90 | 0.90 | 0.92 | 0.90 | 0.90 | 0.91 | 0.91 |
| American Indian/ Alaska Native | 0.87 | 0.88 | 0.89 | 0.89 | 0.89 | 0.88 | 0.90 |
| White | 0.92 | 0.90 | 0.91 | 0.90 | 0.90 | 0.90 | 0.91 |
| Multiple Ethnicities | 0.92 | 0.91 | 0.92 | 0.90 | 0.91 | 0.90 | 0.91 |
| LEP | 0.89 | 0.85 | 0.84 | 0.84 | 0.83 | 0.81 | 0.77 |
| IDEA | 0.90 | 0.90 | 0.89 | 0.86 | 0.88 | 0.85 | 0.87 |
| Section 504 | 0.92 | 0.92 | 0.92 | 0.92 | 0.91 | 0.90 | 0.91 |

*Note:* cells with "–" indicate that marginal reliability is not computed due to a small sample size, $n < 10$.

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

| Subgroup | Grade 3 | Grade 4 | Grade 5 | Grade 6 | Grade 7 | Grade 8 | Grade 11 |
|---|---|---|---|---|---|---|---|
| All Students | 0.94 | 0.94 | 0.93 | 0.94 | 0.93 | 0.93 | 0.93 |
| Female | 0.94 | 0.94 | 0.93 | 0.93 | 0.93 | 0.93 | 0.92 |
| Male | 0.94 | 0.94 | 0.94 | 0.94 | 0.94 | 0.93 | 0.93 |
| African American | 0.93 | 0.93 | 0.91 | 0.92 | 0.92 | 0.91 | 0.88 |
| Asian | 0.95 | 0.95 | 0.95 | 0.95 | 0.94 | 0.95 | 0.94 |
| Native Hawaiian/ Pacific Islander | - | 0.93 | 0.93 | 0.88 | 0.91 | - | 0.94 |
| Hispanic/Latino | 0.93 | 0.93 | 0.92 | 0.92 | 0.91 | 0.91 | 0.91 |
| American Indian/ Alaska Native | 0.91 | 0.91 | 0.87 | 0.90 | 0.87 | 0.86 | 0.86 |
| White | 0.93 | 0.93 | 0.92 | 0.93 | 0.93 | 0.92 | 0.92 |
| Multiple Ethnicities | 0.94 | 0.94 | 0.91 | 0.92 | 0.92 | 0.92 | 0.91 |
| LEP | 0.92 | 0.90 | 0.83 | 0.88 | 0.85 | 0.80 | 0.72 |
| IDEA | 0.93 | 0.93 | 0.89 | 0.90 | 0.88 | 0.85 | 0.81 |
| Section 504 | 0.94 | 0.95 | 0.94 | 0.94 | 0.93 | 0.92 | 0.92 |

*Note:* cells with "–" indicate that marginal reliability is not computed due to a small sample size, $n < 10$.

## 5.5 RELIABILITY FOR CLAIM SCORES

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

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

| Grade | Reporting Categories | Number of Items Specified in Test Blueprint | | Marginal Reliability | Scale Score Mean | Scale Score SD | Average CSEM |
|---|---|---|---|---|---|---|---|
| | | Min | Max | | | | |
| 3 | Claim 1: Reading | 14 | 16 | 0.79 | 2415.08 | 102.74 | 47.64 |
| | Claim 2: Writing | 11 | 11 | 0.79 | 2419.49 | 98.96 | 45.11 |
| | Claim 3: Listening | 8 | 8 | 0.59 | 2426.00 | 121.27 | 77.55 |
| | Claim 4: Research | 8 | 9 | 0.68 | 2406.71 | 118.57 | 67.29 |
| 4 | Claim 1: Reading | 14 | 16 | 0.73 | 2461.72 | 108.42 | 56.17 |
| | Claim 2: Writing | 11 | 11 | 0.78 | 2459.15 | 103.94 | 49.15 |
| | Claim 3: Listening | 8 | 8 | 0.63 | 2458.54 | 129.74 | 78.51 |
| | Claim 4: Research | 7 | 9 | 0.67 | 2451.05 | 119.77 | 69.28 |
| 5 | Claim 1: Reading | 14 | 16 | 0.74 | 2494.40 | 113.96 | 57.80 |
| | Claim 2: Writing | 11 | 11 | 0.79 | 2496.82 | 106.13 | 48.90 |
| | Claim 3: Listening | 8 | 9 | 0.62 | 2494.21 | 129.72 | 79.58 |
| | Claim 4: Research | 8 | 9 | 0.75 | 2481.39 | 121.77 | 61.37 |
| 6 | Claim 1: Reading | 14 | 16 | 0.73 | 2517.34 | 108.99 | 56.90 |
| | Claim 2: Writing | 11 | 11 | 0.80 | 2516.15 | 100.64 | 45.03 |
| | Claim 3: Listening | 8 | 9 | 0.57 | 2546.29 | 132.92 | 86.69 |
| | Claim 4: Research | 8 | 9 | 0.66 | 2505.93 | 119.38 | 69.46 |
| 7 | Claim 1: Reading | 14 | 16 | 0.77 | 2548.38 | 110.58 | 53.44 |
| | Claim 2: Writing | 11 | 11 | 0.78 | 2542.47 | 106.27 | 49.46 |
| | Claim 3: Listening | 8 | 9 | 0.58 | 2549.61 | 128.01 | 83.29 |
| | Claim 4: Research | 8 | 9 | 0.67 | 2533.62 | 122.91 | 70.61 |
| 8 | Claim 1: Reading | 16 | 16 | 0.77 | 2556.23 | 113.07 | 54.73 |
| | Claim 2: Writing | 11 | 11 | 0.78 | 2553.85 | 109.32 | 51.18 |
| | Claim 3: Listening | 8 | 9 | 0.56 | 2570.65 | 137.94 | 91.61 |
| | Claim 4: Research | 8 | 9 | 0.67 | 2537.74 | 126.39 | 72.31 |
| 11 | Claim 1: Reading | 15 | 16 | 0.77 | 2609.15 | 121.99 | 59.06 |
| | Claim 2: Writing | 11 | 11 | 0.79 | 2614.54 | 123.91 | 56.94 |
| | Claim 3: Listening | 8 | 9 | 0.62 | 2611.97 | 148.96 | 91.36 |
| | Claim 4: Research | 8 | 9 | 0.67 | 2585.06 | 135.90 | 77.48 |

Table 37. Marginal Reliability Coefficients for Claim Scores in Mathematics

| Grade | Reporting Categories | Number of Items Specified in Test Blueprint | | Marginal Reliability | Scale Score Mean | Scale Score SD | Average CSEM |
|---|---|---|---|---|---|---|---|
| | | Min | Max | | | | |
| 3 | Claim 1 | 20 | 20 | 0.90 | 2438.93 | 86.44 | 27.71 |
| | Claim 2 & 4 | 8 | 11 | 0.73 | 2433.30 | 90.72 | 47.48 |
| | Claim 3 | 9 | 11 | 0.72 | 2430.40 | 94.05 | 49.71 |
| 4 | Claim 1 | 20 | 20 | 0.90 | 2478.28 | 86.59 | 27.79 |
| | Claim 2 & 4 | 8 | 10 | 0.76 | 2473.33 | 92.61 | 45.76 |
| | Claim 3 | 9 | 10 | 0.74 | 2470.34 | 95.36 | 48.94 |
| 5 | Claim 1 | 20 | 20 | 0.88 | 2499.34 | 91.58 | 31.40 |
| | Claim 2 & 4 | 8 | 10 | 0.64 | 2495.88 | 95.92 | 57.20 |
| | Claim 3 | 9 | 10 | 0.70 | 2490.24 | 107.32 | 59.21 |
| 6 | Claim 1 | 19 | 19 | 0.88 | 2524.87 | 105.39 | 35.80 |
| | Claim 2 & 4 | 9 | 10 | 0.72 | 2512.77 | 112.20 | 59.59 |
| | Claim 3 | 10 | 11 | 0.71 | 2513.81 | 112.17 | 60.38 |
| 7 | Claim 1 | 20 | 20 | 0.88 | 2543.36 | 108.61 | 37.10 |
| | Claim 2 & 4 | 10 | 10 | 0.66 | 2533.88 | 123.67 | 72.19 |
| | Claim 3 | 8 | 10 | 0.65 | 2529.39 | 124.72 | 73.40 |
| 8 | Claim 1 | 20 | 20 | 0.88 | 2557.64 | 119.65 | 41.91 |
| | Claim 2 & 4 | 8 | 10 | 0.65 | 2543.30 | 131.29 | 77.96 |
| | Claim 3 | 9 | 10 | 0.68 | 2536.55 | 135.06 | 76.15 |
| 11 | Claim 1 | 22 | 22 | 0.89 | 2587.38 | 118.84 | 39.49 |
| | Claim 2 & 4 | 8 | 10 | 0.65 | 2575.62 | 152.34 | 90.44 |
| | Claim 3 | 9 | 12 | 0.62 | 2581.34 | 146.41 | 90.26 |

Legend:
Claim 1 Concepts and Procedures
Claims 2 & 4 Problem Solving & Modeling and Data Analysis
Claim 3 Communicating Reasoning
Note. The low MRs are due to a restricted range in the score distribution.

# 6. SCORING

The South Dakota Smarter Balanced Summative 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 a performance category for each reporting category. This section describes the rules used in generating scores and the handscoring procedure.

## 6.1 ESTIMATING STUDENT ABILITY USING MAXIMUM LIKELIHOOD ESTIMATION

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

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

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

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

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

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

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

in the case of items with two or more points,

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

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

**Standard Error of Measurement**

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

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

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

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

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

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

## 6.2 RULES FOR TRANSFORMING THETA TO VERTICAL SCALE SCORES

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

Table 38. Vertical Scaling Constants on the Reporting Metric

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

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

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

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

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

Table 39. Cut Scores in Scale Scores

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

## 6.3 LOWEST/HIGHEST OBTAINABLE SCORES (LOSS/HOSS)

Although the observed score is measured more precisely in an adaptive test than in a fixed-form test, especially for high- and low-performing students, if the item pool does not include easy or difficult items to measure low- and high-performing students, the standard error could be large in low and high ends of the ability range. Smarter Balanced Assessment Consortium decided to truncate extreme unreliable student ability estimates. Table 40 presents the lowest obtainable score (LOT or LOSS) and the highest obtainable score (HOT or HOSS) in both theta and scale score metrics. Estimated theta's lower than LOT or higher than HOT are truncated to the LOT and HOT values, and assigned LOSS and HOSS associated with the LOT and HOT. LOT and HOT were applied to all tests and all scores (total and reporting category scores). The standard error for LOT and HOT are computed using the LOT and HOT ability estimates given the administered items.

Table 40. Lowest and Highest Obtainable Scores

| Subject | Grade | Theta Metric | | Scale Score Metric | |
|---|---|---|---|---|---|
| | | LOT | HOT | LOSS | HOSS |
| ELA/L | 3 | -5.9110 | 3.5332 | 2001 | 2811 |
| ELA/L | 4 | -5.5500 | 4.1826 | 2032 | 2867 |
| ELA/L | 5 | -5.2670 | 4.7546 | 2056 | 2916 |
| ELA/L | 6 | -5.0000 | 5.0000 | 2079 | 2937 |
| ELA/L | 7 | -4.9660 | 5.3119 | 2082 | 2964 |
| ELA/L | 8 | -4.7925 | 5.6063 | 2097 | 2989 |
| ELA/L | 11 | -4.7305 | 6.1096 | 2102 | 3032 |
| Math | 3 | -5.6030 | 3.1219 | 2071 | 2762 |
| Math | 4 | -5.3601 | 4.0264 | 2090 | 2834 |
| Math | 5 | -5.3012 | 4.7426 | 2095 | 2891 |
| Math | 6 | -5.1942 | 5.0000 | 2103 | 2911 |
| Math | 7 | -5.1311 | 5.6630 | 2108 | 2964 |
| Math | 8 | -5.0681 | 6.0272 | 2113 | 2993 |
| Math | 11 | -5.0000 | 7.1896 | 2118 | 3085 |

## 6.4 SCORING ALL CORRECT AND ALL INCORRECT CASES

In IRT maximum likelihood (ML) ability estimation methods, zero and perfect scores are assigned the ability of minus and plus infinity. For all correct and all incorrect cases, the highest obtainable scores (HOT and HOSS) or the lowest obtainable scores (LOT and LOSS) were assigned in the 2014–2015 administration. Since 2015-2016 administrations, 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/L, claim scores are computed for each claim. In mathematics, claim scores are computed for claim 1, claims 2 and 4 combined, and claim 3. For each claim, three performance categories, relative strength and weakness are produced.

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

For summative tests, the specific rules are as follows:

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

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

## 6.6 TARGET SCORES

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

Target scores are computed in two ways: (1) target scores relative to a student's overall estimated ability (θ), and (2) target scores relative to the proficiency standard (Level 3 cut).

### 6.6.1 Target Scores Relative to Student's Overall Estimated Ability

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

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

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

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

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

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

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

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

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

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

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

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

In the aggregate, a target performance is reported as a group of students performing better, worse, or as expected on this target. In some cases, insufficient information will be available and that will be indicated as well.

For target level strengths/weakness, report the following:

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

- Otherwise, performance is similar to performance on the test as a whole.
- If $se(\bar{\delta}_{Tg}) > 0.2$, data are insufficient.

### 6.6.2 Target Scores Relative to Proficiency Standard (Level 3 Cut)

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

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

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

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

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

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

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

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

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

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

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

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

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

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

- If $\bar{\delta}_{Tg} \geq +1 * se(\bar{\delta}_{Tg})$, then performance is *above* the Proficiency Standard.
- If $\bar{\delta}_{Tg} \leq -1 * se(\bar{\delta}_{Tg})$, then performance is *below* the Proficiency Standard.
- Otherwise, performance is *near* the Proficiency Standard.
- If $se(\bar{\delta}_{Tg}) > 0.2$, data are insufficient.

## 6.7 HANDSCORING

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

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

### 6.7.1 Reader Selection

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

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

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

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

MI requires all handscoring project staff (scoring directors, team leaders, readers, and clerical staff) to sign a confidentiality/nondisclosure agreement before receiving any training or secure project materials. The

employment agreement indicates that no participant in training and/or scoring may reveal information about the test, the scoring criteria, or the scoring methods to any person.

### 6.7.2 Reader Training

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

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

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

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

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

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

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

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

Training design varies slightly depending on Smarter Balanced item type:

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

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

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

### 6.7.3 Reader Statistics

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

In addition to extensive experience in the preparation of training materials and employing management and staff with unparalleled expertise in the field of 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 handscoring project.

MI has developed and operates a comprehensive system for collecting and analyzing scoring data. After the readers' scores are submitted into the VSC handscoring system, the data 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 handscoring project monitors at each MI scoring center via a secure website, and the handscoring project monitors provide updated reports to the scoring directors several times per day. MI scoring directors are experienced in examining these reports and using the information to determine the need for retraining of individual readers or the group as a whole. It can easily be determined if a reader is consistently scoring high or low, and the specific score points with which they may be having difficulty. The scoring directors share such information with the team leaders and direct all retraining efforts.

### 6.7.4 Reader Monitoring and Retraining

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

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

Each student response is scored holistically by a trained and qualified reader using the scoring criteria developed and approved by Smarter Balanced, with a second read conducted on 15% of responses for each item for reliability purposes. Responses are selected randomly for second reading and scored by readers who are not aware of the score assigned by the first reader or even that the response has been read before. MI's QA/reliability procedures allow the handscoring staff to identify struggling readers very early and

begin retraining at once. While retraining these readers, MI also monitors their scoring intensively to ensure that all responses are scored accurately. In fact, MI's monitoring is also used as a retraining method. MI shows readers responses that the readers have scored incorrectly, explains the correct scores, and has the readers change the scores. Between the 2014–2015 and 2015–2016 test administrations, MI developed dynamic "threshold" reports which, based on inputted criteria, immediately identify potential scoring performance issues. This enhancement allows scoring leadership to pinpoint areas of concern and take corrective action with greater efficiency than ever before.

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

### 6.7.5 Reader Validity Checks

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

### 6.7.6 Reader Dismissal

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

### 6.7.7 Reader Agreement

The inter-reader reliability is computed based on scorable responses (numeric scores) scored by two independent readers only, excluding non-scorable responses (e.g., off topic, off purpose, or foreign language responses) which are scored by scoring leadership, not by two independent readers. For the short-answer (SA) items, some items are scored by MI's Project Essay Grade (PEG$^{TM}$) automated scoring technology and some items are hand-scored. For the hand-scored items, the human–human agreements are computed based on the combined data across eight states (Connecticut, Delaware, Hawaii, Idaho, New Hampshire, Oregon, Washington, and West Virginia), and the U.S. Virgin Islands.

In ELA/L, 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.

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, combined %exact and %adjacent agreement, and quadratic weighted Kappa (QWK).

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

| Grade | # of Items | %Exact | | | %(Exact+ Adjacent) | QWK |
|---|---|---|---|---|---|---|
| | | Average | Min | Max | | |
| 3 | 11 | 80 | 72 | 94 | 100 | 0.67 |
| 4 | 10 | 84 | 79 | 92 | 100 | 0.72 |
| 5 | 11 | 72 | 67 | 93 | 99 | 0.64 |
| 6 | 3 | 75 | 70 | 83 | 100 | 0.70 |
| 7 | 7 | 70 | 64 | 86 | 100 | 0.53 |
| 8 | 13 | 74 | 61 | 84 | 100 | 0.65 |
| 11 | 23 | 79 | 65 | 92 | 100 | 0.71 |

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

| Grade | Dimensions | # of Items | %Exact | | | %(Exact+ Adjacent) | QWK |
|---|---|---|---|---|---|---|---|
| | | | Average | Min | Max | | |
| 3 | Conventions | 8 | 73 | 70 | 77 | 100 | 0.60 |
| | Evid/Elab | 8 | 68 | 65 | 71 | 98 | 0.69 |
| | Org/Purp | 8 | 68 | 66 | 70 | 98 | 0.69 |
| 4 | Conventions | 18 | 68 | 64 | 72 | 98 | 0.68 |
| | Evid/Elab | 18 | 70 | 66 | 73 | 99 | 0.71 |
| | Org/Purp | 18 | 70 | 66 | 74 | 99 | 0.72 |
| 5 | Conventions | 20 | 70 | 63 | 82 | 100 | 0.56 |
| | Evid/Elab | 20 | 63 | 56 | 68 | 98 | 0.71 |
| | Org/Purp | 20 | 64 | 58 | 71 | 98 | 0.72 |
| 6 | Conventions | 8 | 73 | 72 | 75 | 99 | 0.61 |
| | Evid/Elab | 8 | 66 | 61 | 72 | 98 | 0.70 |
| | Org/Purp | 8 | 65 | 60 | 70 | 98 | 0.70 |
| 7 | Conventions | 19 | 72 | 68 | 78 | 99 | 0.62 |
| | Evid/Elab | 19 | 71 | 61 | 74 | 99 | 0.74 |
| | Org/Purp | 19 | 71 | 59 | 75 | 99 | 0.74 |
| 8 | Conventions | 20 | 79 | 72 | 86 | 99 | 0.58 |
| | Evid/Elab | 20 | 69 | 63 | 74 | 99 | 0.74 |
| | Org/Purp | 20 | 69 | 63 | 75 | 99 | 0.74 |

Legend:
Evid/Elab: Evidence/Elaboration; Org/Purp: Organization/Purpose

Table 43. Mathematics Reader Agreements

| Grade | Score Points | # of Items | %Exact | | | %(Exact+ Adjacent) | QWK |
|---|---|---|---|---|---|---|---|
| | | | Average | Min | Max | | |
| 3 | 1 | 3 | 90 | 89 | 92 | 100 | 0.73 |
| 4 | 1 | 6 | 84 | 83 | 85 | 100 | 0.62 |
| 6 | 1 | 12 | 97 | 95 | 99 | 100 | 0.92 |
| 7 | 1 | 6 | 97 | 95 | 98 | 100 | 0.88 |
| 8 | 1 | 12 | 90 | 81 | 97 | 100 | 0.78 |
| 11 | 1 | 10 | 97 | 78 | 100 | 100 | 0.93 |
| 3 | 2 | 10 | 90 | 85 | 99 | 100 | 0.91 |
| 4 | 2 | 24 | 91 | 78 | 99 | 100 | 0.89 |
| 5 | 2 | 25 | 91 | 83 | 97 | 100 | 0.87 |
| 6 | 2 | 25 | 86 | 80 | 95 | 100 | 0.86 |
| 7 | 2 | 23 | 87 | 77 | 92 | 100 | 0.82 |
| 8 | 2 | 16 | 89 | 83 | 99 | 100 | 0.85 |
| 11 | 2 | 15 | 95 | 71 | 99 | 100 | 0.94 |
| 3 | 3 | 4 | 95 | 94 | 96 | 99 | 0.97 |
| 4 | 3 | 4 | 88 | 86 | 89 | 99 | 0.93 |
| 5 | 3 | 5 | 87 | 79 | 99 | 96 | 0.79 |

## 6.8 AUTOMATED SCORING

Starting with the 2015–2016 scoring, the SDDOE adopted MI's Project Essay Grade (PEG™) automated scoring technology for short-answer (SA) items in all grades in ELA/L and mathematics.

### 6.8.1 Project Essay Grade (PEG™)

MI acquired the PEG automated scoring technology from Dr. Ellis Batten Page and his associates at Duke University in 2003. MI has re-engineered, enhanced, and extended the PEG system using the latest techniques and technologies in the field of computational linguistics, machine learning, and natural language processing.

Page's insight was that there are certain features[1] of a written text that can be measured by a computer and which can serve as indicators of the quality of that text (Page, 1966). Page's original formulation of PEG measured only 28 features (Page, 1968) and has been criticized for only dealing with surface features of the text (Ben-Simon & Bennett, 2007). Page's PEG also used standard linear regression, which cannot provide a good fit for the sort of non-linear features sometimes found in writing, such as a feature which has some optimal-scoring level which is not extreme. (An example of this might be the average number of prepositional phrases per sentence; while an average count near zero usually correlates with poor scores, it is not true that having fifty prepositional phrases per sentence correlates with high scores.) Since acquiring the technology, MI has invested a great deal of effort in transforming Page's original creation. PEG now measures millions of features, both surface and complex, and employs a whole host of linear and non-linear algorithms to correlate those features with scores.

1 Page called them "proxes" (1966), but MI prefers the term "feature" as it has become the industry standard in both machine learning and its subfield, automated writing evaluation.

The new system has been redesigned from the ground up to be a modular collection of loosely connected components, controlled by a simple set of parameters. There has been a great deal of research over the past few decades in the fields of natural language processing, text analytics, and machine learning. PEG's modular nature allows MI's researchers to quickly and easily "plug in" new algorithms and techniques, ensuring that the system stays at the cutting edge of automated scoring technology. PEG is also self-correcting; if a new component does not increase the accuracy with which PEG scores student responses, it is automatically excluded from the final model. This self-correcting feature highlights the philosophy at the heart of the system, which is that the authority on whether a particular piece of writing embodies the writing construct being measured resides with the handscoring experts and rubric creators rather than with the engineers of any particular piece of technology.

PEG's approach to measuring the writing construct begins with the assumption that the goal of the model-building process is to generate an algorithm (or *model*) that faithfully mimics the scoring done by the expert human readers who scored the training set. Rather than try to identify ahead of time the optimal features to measure and the algorithms to use in correlating scores with features, PEG's data drives the process, automatically finding those features and algorithms that best minimize the error rate on the training data. In order to jump-start the process, MI's linguists and engineers have worked together to create thousands of handcrafted features[2]. These features measure the elements of writing that expert scorers and teachers look for when scoring student responses, as well as those elements of the text that MI researchers have found commonly correlate with high- and low-scoring responses.

MI then extends this extrinsic feature set with a theoretically infinite set of intrinsic features that can be automatically extracted from the training set corpus. These are created in a number of ways, but some of the commonly used methods involve n-grams of characters, words, parts of speech, and phrases, as well as similarity measures and various matrix manipulations on the underlying intrinsic and extrinsic features. If these features correlate with the scores given by the expert human readers, then they can be encoded into the model without being directly observed by the researchers. In many cases, it would not be possible to observe them explicitly because they are too complex. MI views this model-complexity as a strength. It frees PEG from the limitations of only looking for features that engineers happen to think of and instead allows the system to leverage the collective intelligence of the humans who scored the responses in the training set.

### 6.8.2 PEG Training and Validation Samples

During the automated scoring of the eligible items on the 2016–2017 Smarter Balanced summative assessment, the PEG models employed to score the 2016–2017 Smarter Balanced operational test responses were constructed as part of the Smarter Balanced Field Test Automated Scoring Research Studies (read more: http://www.smarterapp.org/documents/FieldTest_AutomatedScoringResearchStudies.pdf). Training and validation of PEG used the responses from the 2014 Smarter Balanced field test. The field test sampling methodology is described below.

McGraw-Hill Education CTB randomly sampled 1,500 responses for each short-text, constructed-response item and 2,000 responses for each essay item from the available pool of on-grade field-test responses. Table 44 displays the item counts by score type. A random sample of 500 responses for each item were designated as validation responses; the remaining responses were designated as training responses. All responses

2 Including Page's original 28.

received two human reads. Any non-exact scores were adjudicated by scoring leadership. The score of record was the matched score (in the case of agreement between the two human reads) or the adjudicated score (in the case of disagreement). Any responses that humans had classified as non-scorable and assigned a condition code (e.g., insufficient, off-topic) were recoded as zeros. Scores of record were compared against the scores assigned by PEG and the other automated scoring engines for validation purposes.

The sets of validation responses are groups of texts that are independent of the training set, but which, like the set of training responses, have been scored by humans. After a natural language processing program, such as exists within PEG, has been trained, it is evaluated using the validation set. If the scores assigned by PEG agree to a predetermined level of accuracy with the human scores for the same responses, PEG is considered adequately trained. Otherwise, the natural language processing method is retrained using different inputs or possibly with a new algorithm. Validation responses are used to evaluate the efficacy of training. Thus, the validation set serves as a check against overtraining.

There was significant variation in the number of on-grade responses for each item. Responses were randomly sampled from the available on-grade responses in either the Standard Setting Sample or the Census Sample. Both the Standard Setting Sample and the Census Sample were representative of the student population as a whole.

### 6.8.3 Automated Scoring Processes

During the automated scoring of the eligible items on the 2016–2017 Smarter Balanced summative assessment, training and validation of PEG used responses from the 2014 Smarter Balanced field test. In addition, scoring models were generated and evaluated in the fall/winter of 2016 by Educational Testing Service (ETS), following the process described in section 6.8.4, in order to supplement the pool of eligible items.

**Response Routing**

During the 2016–2017 Smarter Balanced summative assessment administration, AIR routed all responses requiring handscoring or automated scoring to MI. Upon receipt and validation of each response, MI routed responses for those items eligible for automated scoring to PEG and the remainder of the responses to MI's handscoring system.

**Quality Assurance**

A number of strategies were in place to assure the quality of the scores assigned by PEG. First, only those items proven to be appropriate for automated scoring were scored by PEG. Second, all responses for which PEG had low confidence in assigning scores were sent to a human reader for scoring. Third, PEG only assigned numerical scores to responses. Responses suspected to be non-scorable and require condition codes were sent to a human reader for scoring. Finally, 15% of all PEG-scored responses were additionally scored by human readers for the purpose of monitoring agreement as described above. Any item with a quadratic weighted kappa (QWK) of < 0.65 will be subject to a review in order to determine whether: (1) the item has a binary or ternary score point range (i.e., 0–1 or 0–2) and thus exact agreement should be

considered[3]; (2) the item had lower than expected human-human agreement; (3) additional training data is needed to support model generalization; and/or (4) the item should be ineligible for automated scoring.

**"Alert" Procedures**

MI implements a formal process for informing clients when student responses reflect a possibly dangerous situation for the test-taker. Specifically, MI employs a set of alert procedures to notify the client of responses indicating endangerment, abuse, or psychological and/or emotional difficulties. PEG employs a rule-based detection system to flag responses that are indicative of potentially dangerous situations. Responses flagged by PEG as possible alerts were then reviewed by scoring leadership, who decide whether each response should be forwarded to the client. Once vetted, all alerts were provided to AIR, who associated the pertinent student information with the response(s) and contacted the state.

**Score Delivery**

As scores were assigned by PEG, MI verified and delivered them to AIR. MI received confirmation from AIR that each response had been received and had passed data validation.

### 6.8.4 Item Sample for Operational Scoring

Subsequent to the Field Test Automated Scoring Research Studies, ETS conducted an independent review of the automated scoring of the Smarter Balanced field-test items to determine eligibility for operational scoring. This involved an evaluation of the appropriateness of automated scoring for each item based on such outcomes as comparability to human-human metrics, QWK, standardized mean difference, and consistent performance across subgroups. Items were classified into three categories, reflected in the document *IDEAS Item Classification* (unpublished):

Category A: Meets all criteria; no reservations about operational automated scoring

Category B: Meets most criteria; no significant reservations about operational automated scoring

Category C: Does not meet criteria; not eligible for operational automated scoring

Table 44 presents the number of eligible items for automated scoring in 2016-2017. Note that South Dakota additionally excluded all ELA/L performance task items (i.e., all Research and Full Write items).

3 As QWK subtracts the possibility of chance agreement, this metric is naturally depressed when the score point range is narrow and the probability of chance agreement is therefore high.

Table 44. 2016–2017 Summative Item Pool: ETS Item Classifications for Automated Scoring

| Category | Grade | Item Type | Classification | | Total |
|---|---|---|---|---|---|
| | | | Category A | Category B | |
| **ELA/L** | | | | | |
| Brief Write | 3 | SA | 5 | 10 | 15 |
| | 4 | SA | 10 | 11 | 21 |
| | 5 | SA | 7 | 14 | 21 |
| | 6 | SA | 6 | 15 | 21 |
| | 7 | SA | 5 | 14 | 19 |
| | 8 | SA | 8 | 10 | 18 |
| | 11 | SA | 15 | 23 | 38 |
| Reading | 3 | SA | 2 | 5 | 7 |
| | 4 | SA | 21 | 5 | 26 |
| | 5 | SA | 36 | 15 | 51 |
| | 6 | SA | 7 | 7 | 14 |
| | 7 | SA | 34 | 13 | 47 |
| | 8 | SA | 33 | 13 | 46 |
| | 11 | SA | 24 | 41 | 65 |
| Research | 3 | SA | 7 | 15 | 22 |
| | 4 | SA | 3 | 16 | 19 |
| | 5 | SA | 1 | 3 | 4 |
| | 6 | SA | 10 | 15 | 25 |
| | 7 | SA | 1 | 3 | 4 |
| | 8 | SA | 1 | 4 | 5 |
| | 11 | SA | 16 | 30 | 46 |
| Full Write | 3 | WER | 3 | 3 | 6 |
| | 6 | WER | 1 | 5 | 6 |
| | 7 | WER | 1 | 0 | 1 |
| | 11 | WER | 15 | 9 | 24 |
| **Mathematics** | | | | | |
| All | 3 | SA | 11 | 22 | 33 |
| All | 4 | SA | 9 | 5 | 14 |
| All | 5 | SA | 9 | 14 | 23 |
| All | 6 | SA | 7 | 2 | 9 |
| All | 7 | SA | 3 | 11 | 14 |
| All | 8 | SA | 12 | 3 | 15 |
| All | 11 | SA | 18 | 22 | 40 |
| Total | | | 341 | 378 | 719 |

### 6.8.5 PEG-Human Agreements

During scoring of the 2016–2017 Smarter Balanced summative assessment, 15% of all PEG-scored responses were additionally scored by human readers for the purpose of determining agreement. Since PEG did not classify responses as non-scorable, agreement is based on numerical scores only consistent with the Field Test Automated Scoring Research Studies methodology.

The PEG-human statistics are computed based on the combined data for South Dakota and Vermont to increase the sample size for each item. For the PEG scored items, the human–human agreements are computed based on the combined data across eight states (Connecticut, Delaware, Hawaii, Idaho, New Hampshire, Oregon, Washington, and West Virginia), and the U.S. Virgin Islands. Table 45 summarizes

the average agreement rate statistics for 2016–2017 PEG-human and human-human agreement. This summary includes items with greater than 50 responses.

Table 45. Average Agreement Rate Statistics for Automated Scoring in ELA/L

| Grade | Item Type | Score Point Range | # of Items with n > 50 | 2016-17 PEG-Human Agreement | | | 2016-17 Human-Human Agreement | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | % Exact | % Adj. & Exact | QWK | % Exact | % Adj. & Exact | QWK |
| **ELA/L** | | | | | | | | | |
| 3 | SA | 0-2 | 31 | 77 | 99 | 0.70 | 78 | 100 | 0.74 |
| 4 | SA | 0-2 | 34 | 77 | 99 | 0.67 | 79 | 100 | 0.74 |
| 5 | SA | 0-2 | 14 | 75 | 99 | 0.67 | 78 | 100 | 0.71 |
| 6 | SA | 0-2 | 34 | 76 | 99 | 0.67 | 74 | 100 | 0.69 |
| 7 | SA | 0-2 | 15 | 76 | 100 | 0.67 | 77 | 100 | 0.71 |
| 8 | SA | 0-2 | 10 | 69 | 99 | 0.58 | 73 | 100 | 0.67 |
| 11 | SA | 0-2 | 30 | 69 | 99 | 0.62 | 76 | 100 | 0.70 |
| **Mathematics** | | | | | | | | | |
| 3 | SA | 0-1 | 8 | 92 | 100 | 0.85 | 93 | 100 | 0.85 |
| 4 | SA | 0-1 | 2 | 93 | 100 | 0.84 | 93 | 100 | 0.85 |
| 5 | SA | 0-1 | 4 | 92 | 100 | 0.65 | 93 | 100 | 0.69 |
| 6 | SA | 0-1 | 2 | 94 | 100 | 0.63 | 88 | 100 | 0.60 |
| 7 | SA | 0-1 | 2 | 96 | 100 | 0.66 | 94 | 100 | 0.62 |
| 8 | SA | 0-1 | 2 | 88 | 100 | 0.76 | 84 | 100 | 0.69 |
| 11 | SA | 0-1 | 10 | 95 | 100 | 0.69 | 93 | 100 | 0.73 |
| 3 | SA | 0-2 | 16 | 85 | 99 | 0.85 | 89 | 100 | 0.90 |
| 4 | SA | 0-2 | 12 | 87 | 98 | 0.85 | 91 | 100 | 0.91 |
| 5 | SA | 0-2 | 16 | 84 | 99 | 0.81 | 87 | 100 | 0.87 |
| 6 | SA | 0-2 | 7 | 92 | 100 | 0.88 | 93 | 100 | 0.93 |
| 7 | SA | 0-2 | 4 | 82 | 99 | 0.80 | 87 | 100 | 0.88 |
| 8 | SA | 0-2 | 10 | 77 | 100 | 0.83 | 81 | 100 | 0.87 |
| 11 | SA | 0-2 | 9 | 86 | 100 | 0.79 | 89 | 100 | 0.88 |
| 5 | SA | 0-3 | 3 | 76 | 95 | 0.76 | 81 | 95 | 0.85 |

# 7. REPORTING AND INTERPRETING SCORES

The Online Reporting System (ORS) generates a set of online score reports that include the information describing student performance for students, parents, educators, and other stakeholders. The online score reports are produced immediately after students complete tests and the tests are hand-scored. Because the score reports on students' performance are updated each time that students complete tests and these tests are hand-scored, authorized users (e.g., school principals, teachers) can view students' performance on the tests and use them to improve student learning. In addition to individual student's score report, the ORS also produces aggregate score reports by class, schools, districts, and states. The timely accessibility of aggregate score reports could help users monitor students testing in each subject by grade area, evaluate the effectiveness of instructional strategies, and inform the adoption of strategies to improve student learning and teaching during the school year. Additionally, the ORS provides participation data that helps monitor student participation rate.

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

## 7.1 ONLINE REPORTING SYSTEM FOR STUDENTS AND EDUCATORS

### 7.1.1 Types of Online Score Reports

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

Once authorized users log in to the ORS and select "Score Reports," the online score reports are presented hierarchically. The ORS starts with presenting summaries on student performance by subject and grade at a selected aggregate level. To view student performance for a specific aggregate unit, users can select the specific aggregate unit from a drop-down menu with a list of aggregate units (e.g., schools within a district, or teachers within a school). 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 46 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 46. Types of Online Score Reports by Level of Aggregation

| Level of Aggregation | Types of Online Score Reports |
|---|---|
| State<br>District<br>School<br>Teacher<br>Roster | • Number of students tested and percent of proficient (overall students and by subgroup)<br>• Average scale score and standard error of average scale score (overall students and by subgroup)<br>• Percent of students at each achievement level on overall test and by claims (overall students and by subgroup)<br>• Performance category level in each target (overall students)[1]<br>• Participation rate (overall students)[2]<br>• On-demand student roster report |
| Student | • Total scale score and standard error of measurement<br>• Achievement level on overall and claim scores with achievement level descriptors<br>• Average scale scores and standard errors of average scale scores for student's school, district, and state<br>• Student growth in scale score and achievement level over time<br>• Writing performance descriptors and scores by dimensions |

*Note*.
1: Performance category in each target is provided for all aggregate levels except for state.
2: Participation rate reports are provided at state, district, 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 47 presents the types of subgroups and subgroup category provided in ORS.

Table 47. Types of Subgroups

| Subgroup | Subgroup Category |
|---|---|
| Gender | Male |
| | Female |
| IDEA Indicator | Yes |
| | No |
| Limited English Proficiency (LEP) Status | Yes |
| | No |
| Section 504 Status | Yes |
| | No |
| Race/Ethnicity | American Indian or Alaska Native |
| | Asian |
| | Black or African American |
| | Demographic Race Two or More Races |
| | Hispanic or Latino or Other Pacific Islander |
| | Native Hawaiian or Other Pacific Islander |
| | White |

## 7.1.2 Online Reporting System

### *7.1.2.1 Home Page*

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

The home page summarizes students' performance including (1) number of students tested, and (2) percent proficient. Exhibits 1 and 2 present sampled home pages at the state level and the district level, respectively.

Exhibit 1. Home Page: State Level

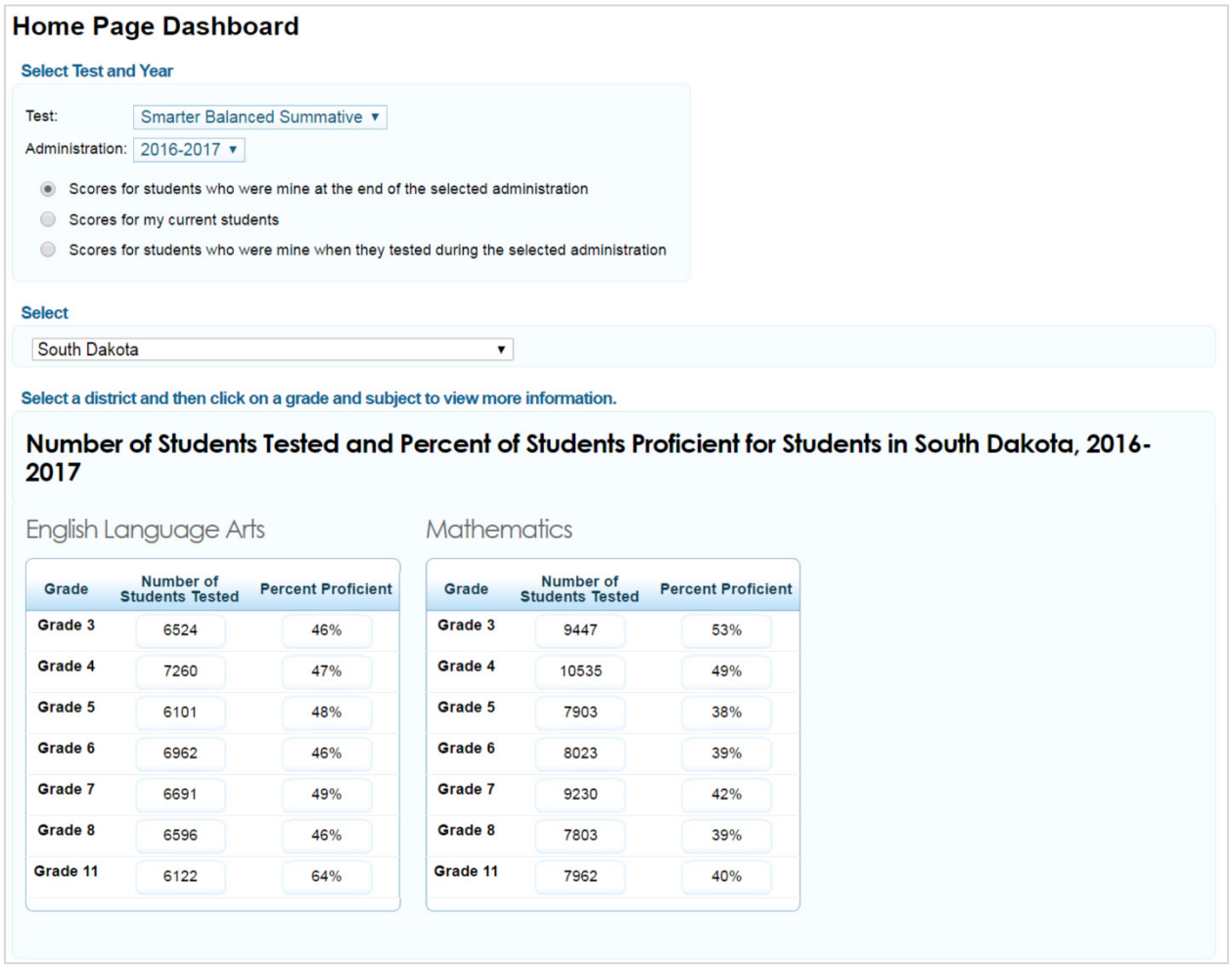
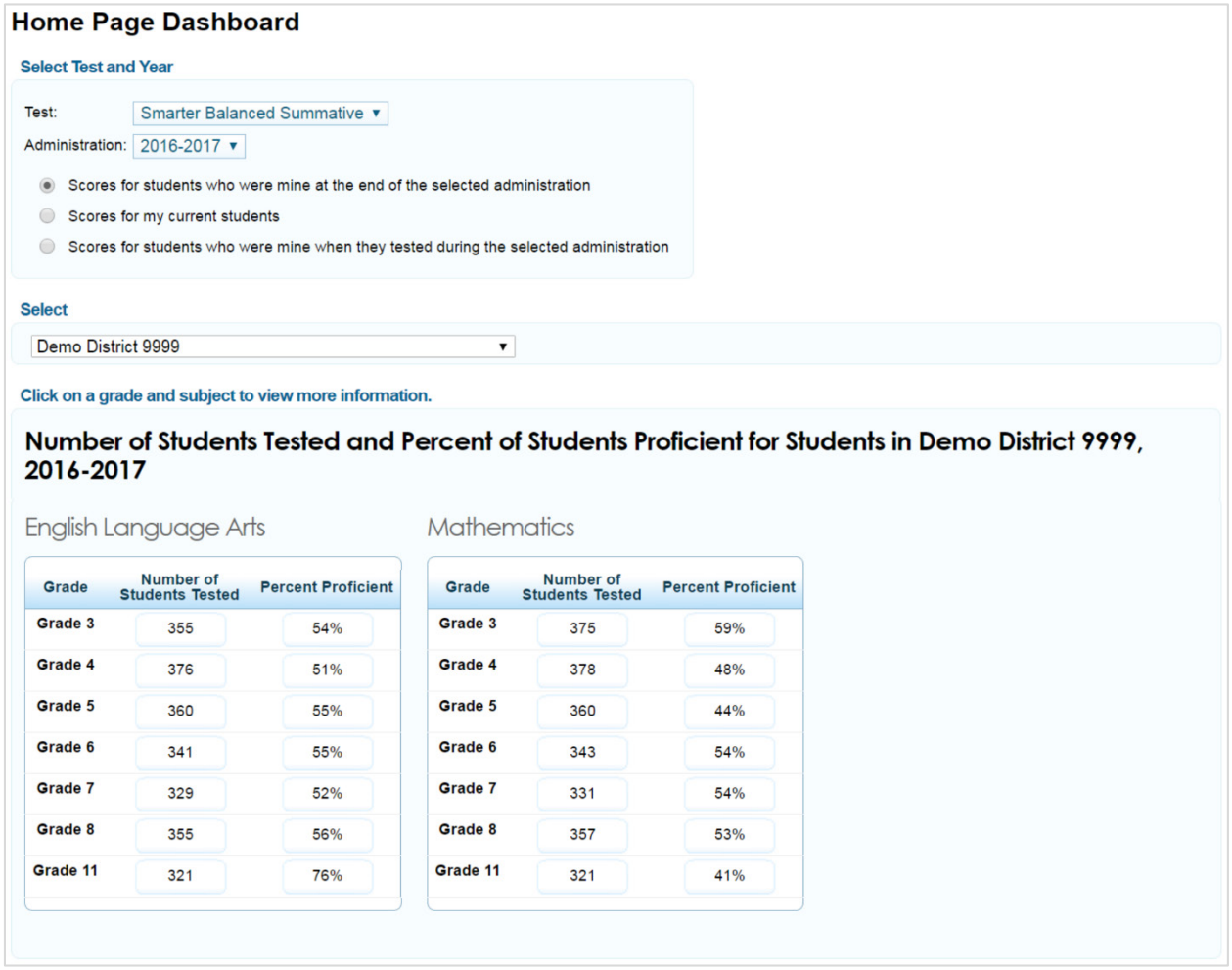


Exhibit 2. Home Page: District Level

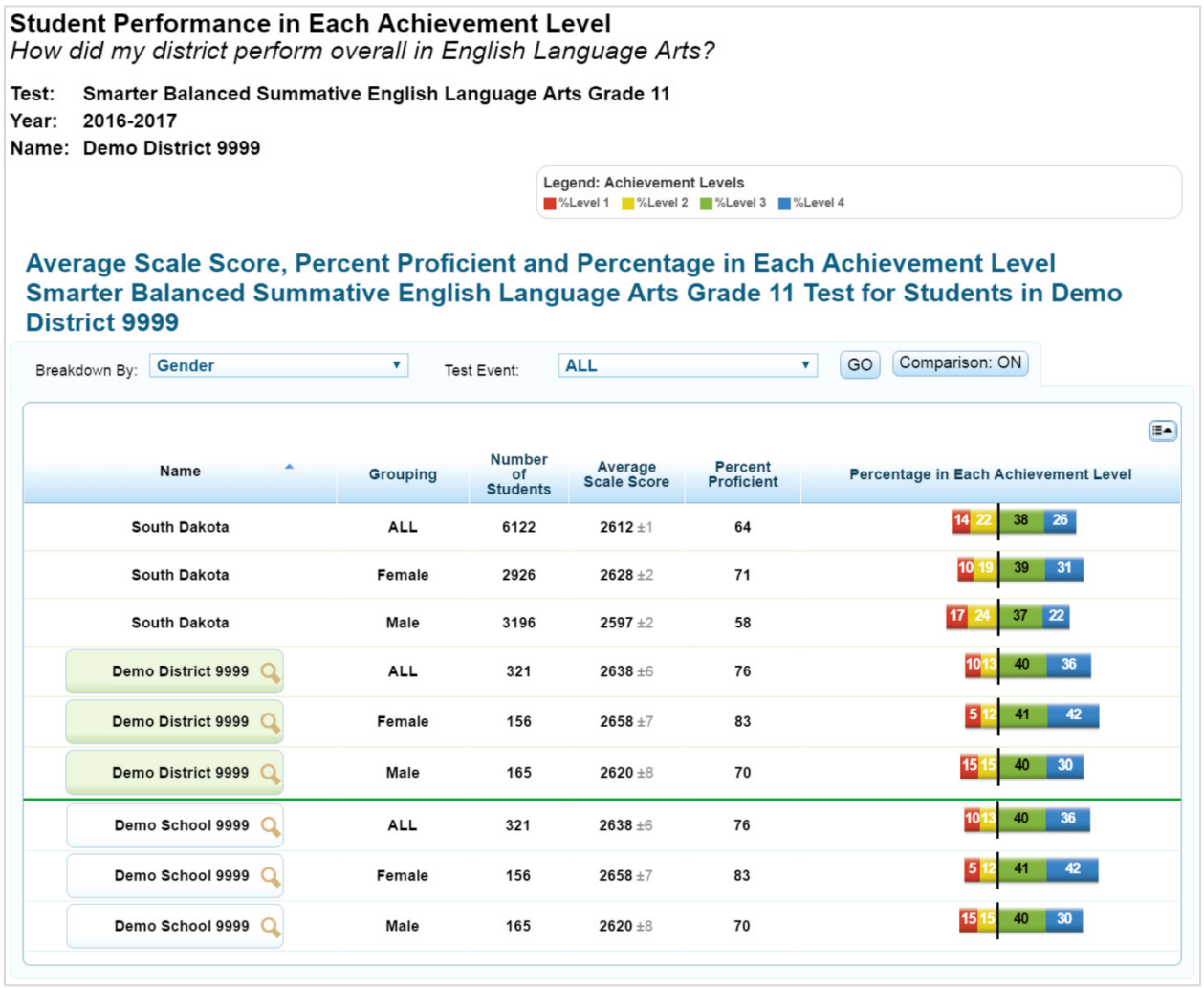



### *7.1.2.2 Subject Detail Page*

More detailed summaries of student performance in each grade on 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 and district are provided above the school summary results as well so that the school performance can be compared with the above aggregate levels.

The subject detail page provides the aggregate summaries on a specific subject area including (1) number of students, (2) average scale score and standard error of the average scale score, (3) percent 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 the subject detail page for ELA/L at a district level when a user selects a subgroup of gender.

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



**Student Performance in Each Achievement Level**

*How did my district perform overall in English Language Arts?*

**Test:** Smarter Balanced Summative English Language Arts Grade 11
**Year:** 2016-2017
**Name:** Demo District 9999

Legend: Achievement Levels
%Level 1 %Level 2 %Level 3 %Level 4

**Average Scale Score, Percent Proficient and Percentage in Each Achievement Level Smarter Balanced Summative English Language Arts Grade 11 Test for Students in Demo District 9999**

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

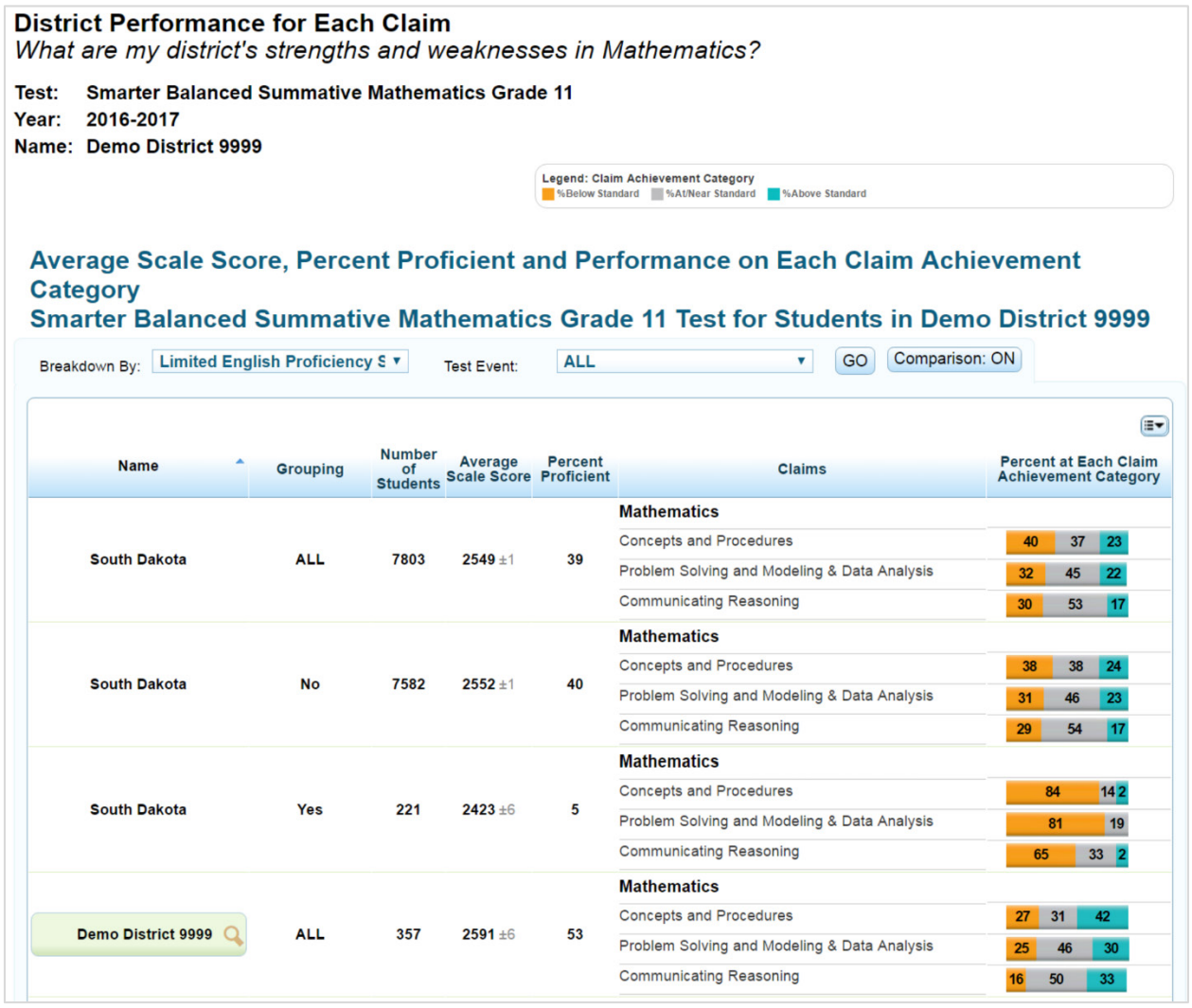
| Name | Grouping | Number of Students | Average Scale Score | Percent Proficient | Percentage in Each Achievement Level |
|---|---|---|---|---|---|
| South Dakota | ALL | 6122 | 2612 ±1 | 64 | 14 22 38 26 |
| South Dakota | Female | 2926 | 2628 ±2 | 71 | 10 19 39 31 |
| South Dakota | Male | 3196 | 2597 ±2 | 58 | 17 24 37 22 |
| Demo District 9999 | ALL | 321 | 2638 ±6 | 76 | 10 13 40 36 |
| Demo District 9999 | Female | 156 | 2658 ±7 | 83 | 5 12 41 42 |
| Demo District 9999 | Male | 165 | 2620 ±8 | 70 | 15 15 40 30 |
| Demo School 9999 | ALL | 321 | 2638 ±6 | 76 | 10 13 40 36 |
| Demo School 9999 | Female | 156 | 2658 ±7 | 83 | 5 12 41 42 |
| Demo School 9999 | Male | 165 | 2620 ±8 | 70 | 15 15 40 30 |

### *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 proficient, and (4) percent of students in each performance category for each claim.

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

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



**District Performance for Each Claim**

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

**Test:** Smarter Balanced Summative Mathematics Grade 11
**Year:** 2016-2017
**Name:** Demo District 9999

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

**Average Scale Score, Percent Proficient and Performance on Each Claim Achievement Category**
**Smarter Balanced Summative Mathematics Grade 11 Test for Students in Demo District 9999**

Breakdown By: Limited English Proficiency S | Test Event: ALL | GO | Comparison: ON

| Name | Grouping | Number of Students | Average Scale Score | Percent Proficient | Claims | Percent at Each Claim Achievement Category |
|---|---|---|---|---|---|---|
| South Dakota | ALL | 7803 | 2549 ±1 | 39 | **Mathematics** | |
| | | | | | Concepts and Procedures | 40 37 23 |
| | | | | | Problem Solving and Modeling & Data Analysis | 32 45 22 |
| | | | | | Communicating Reasoning | 30 53 17 |
| South Dakota | No | 7582 | 2552 ±1 | 40 | **Mathematics** | |
| | | | | | Concepts and Procedures | 38 38 24 |
| | | | | | Problem Solving and Modeling & Data Analysis | 31 46 23 |
| | | | | | Communicating Reasoning | 29 54 17 |
| South Dakota | Yes | 221 | 2423 ±6 | 5 | **Mathematics** | |
| | | | | | Concepts and Procedures | 84 14 2 |
| | | | | | Problem Solving and Modeling & Data Analysis | 81 19 |
| | | | | | Communicating Reasoning | 65 33 2 |
| Demo District 9999 | ALL | 357 | 2591 ±6 | 53 | **Mathematics** | |
| | | | | | Concepts and Procedures | 27 31 42 |
| | | | | | Problem Solving and Modeling & Data Analysis | 25 46 30 |
| | | | | | Communicating Reasoning | 16 50 33 |

### *7.1.2.4 Target Detail Page*

The target detail page provides the aggregate summaries on student performance in each target. The target detail page provides (1) strength or weakness indicators in each target that are computed in two ways (i.e., performance relative to proficiency, performance relative to the overall performance), and (2) average scale scores and standard errors of average scale scores for the selected aggregate unit and the aggregate unit above the selected aggregate. It should be noted that the summaries on target-level student performance are generated for overall students only. That is, the summaries on target-level student performance are not generated by subgroup. Exhibits 5–8 present examples of target detail pages for ELA/L and mathematics at the school and the roster levels.

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

**Performance on Each Target for the English Language Arts Test**

*What are my school's strengths and weaknesses in the English Language Arts Targets?*

**Test:** Smarter Balanced Summative English Language Arts Grade 11
**Year:** 2016-2017
**Name:** Demo School 9999

**Legend: Performance Relative to the Test as a Whole**
- Performance is better than on the rest of the test
- Performance is similar to performance on the test as a whole
- Performance is worse than on the rest of the test
- Insufficient Information

**Legend: Performance Relative to Proficiency**
- Performance is above the Proficiency Standard
- Performance is near the Proficiency Standard
- Performance is below the Proficiency Standard
- Insufficient Information

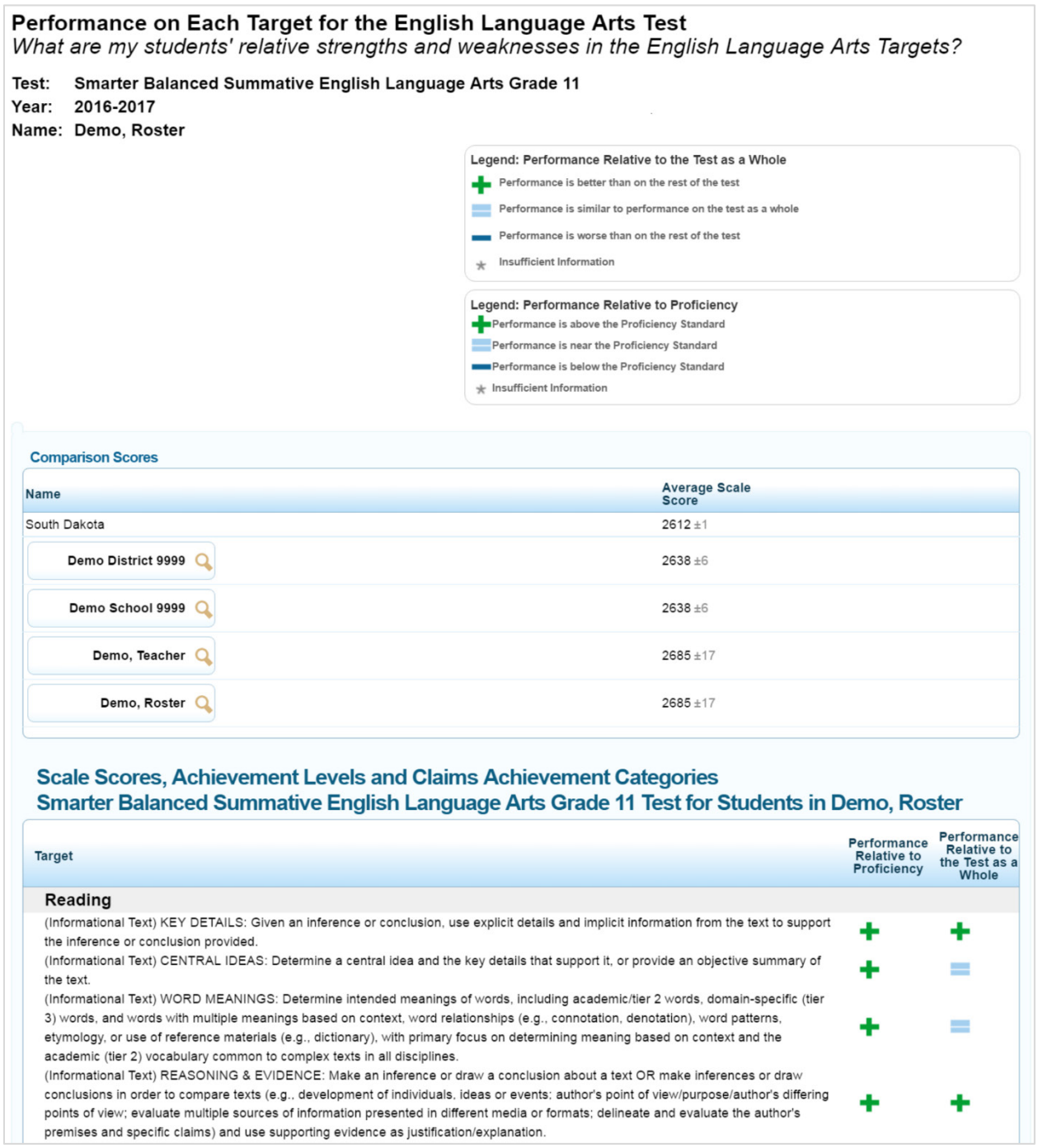
**Comparison Scores**

| Name | Average Scale Score |
|---|---|
| South Dakota | 2612 ±1 |
| Demo District 9999 | 2638 ±6 |
| Demo School 9999 | 2638 ±6 |

**Performance on Each Target**
**Smarter Balanced Summative English Language Arts Grade 11 Test for Students in Demo School 9999**

| Target | Performance Relative to Proficiency | Performance Relative to the Test as a Whole |
|---|---|---|
| **Reading** | | |
| (Informational Text) KEY DETAILS: Given an inference or conclusion, use explicit details and implicit information from the text to support the inference or conclusion provided. | + | − |
| (Informational Text) CENTRAL IDEAS: Determine a central idea and the key details that support it, or provide an objective summary of the text. | + | − |
| (Informational Text) WORD MEANINGS: Determine intended meanings of words, including academic/tier 2 words, domain-specific (tier 3) words, and words with multiple meanings based on context, word relationships (e.g., connotation, denotation), word patterns, etymology, or use of reference materials (e.g., dictionary), with primary focus on determining meaning based on context and the academic (tier 2) vocabulary common to complex texts in all disciplines. | + | = |
| (Informational Text) REASONING & EVIDENCE: Make an inference or draw a conclusion about a text OR make inferences or draw conclusions in order to compare texts (e.g., development of individuals, ideas or events; author's point of view/purpose/author's differing points of view; evaluate multiple sources of information presented in different media or formats; delineate and evaluate the author's premises and specific claims) and use supporting evidence as justification/explanation. | + | = |

Exhibit 6. Target Detail Page for ELA/L: Roster Level



**Performance on Each Target for the English Language Arts Test**

*What are my students' relative strengths and weaknesses in the English Language Arts Targets?*

**Test:** Smarter Balanced Summative English Language Arts Grade 11
**Year:** 2016-2017
**Name:** Demo, Roster

**Legend: Performance Relative to the Test as a Whole**
- Performance is better than on the rest of the test
- Performance is similar to performance on the test as a whole
- Performance is worse than on the rest of the test
- Insufficient Information

**Legend: Performance Relative to Proficiency**
- Performance is above the Proficiency Standard
- Performance is near the Proficiency Standard
- Performance is below the Proficiency Standard
- Insufficient Information

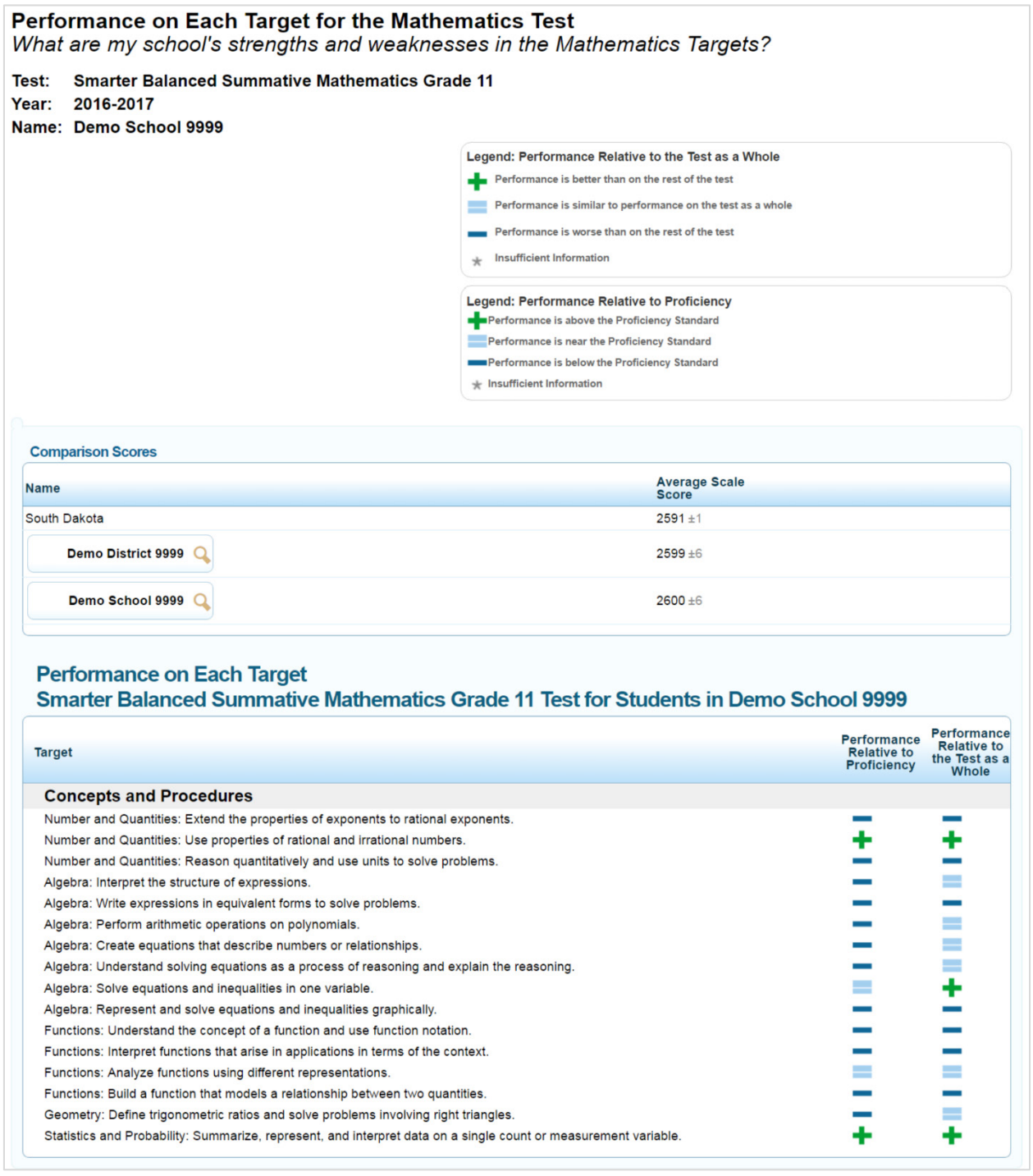
**Comparison Scores**

| Name | Average Scale Score |
|---|---|
| South Dakota | 2612 ±1 |
| Demo District 9999 | 2638 ±6 |
| Demo School 9999 | 2638 ±6 |
| Demo, Teacher | 2685 ±17 |
| Demo, Roster | 2685 ±17 |

**Scale Scores, Achievement Levels and Claims Achievement Categories**
**Smarter Balanced Summative English Language Arts Grade 11 Test for Students in Demo, Roster**

| Target | Performance Relative to Proficiency | Performance Relative to the Test as a Whole |
|---|---|---|
| **Reading** | | |
| (Informational Text) KEY DETAILS: Given an inference or conclusion, use explicit details and implicit information from the text to support the inference or conclusion provided. | + | + |
| (Informational Text) CENTRAL IDEAS: Determine a central idea and the key details that support it, or provide an objective summary of the text. | + | = |
| (Informational Text) WORD MEANINGS: Determine intended meanings of words, including academic/tier 2 words, domain-specific (tier 3) words, and words with multiple meanings based on context, word relationships (e.g., connotation, denotation), word patterns, etymology, or use of reference materials (e.g., dictionary), with primary focus on determining meaning based on context and the academic (tier 2) vocabulary common to complex texts in all disciplines. | + | = |
| (Informational Text) REASONING & EVIDENCE: Make an inference or draw a conclusion about a text OR make inferences or draw conclusions in order to compare texts (e.g., development of individuals, ideas or events; author's point of view/purpose/author's differing points of view; evaluate multiple sources of information presented in different media or formats; delineate and evaluate the author's premises and specific claims) and use supporting evidence as justification/explanation. | + | + |

Exhibit 7. Target Detail Page for Mathematics: School Level



**Performance on Each Target for the Mathematics Test**

*What are my school's strengths and weaknesses in the Mathematics Targets?*

**Test:** Smarter Balanced Summative Mathematics Grade 11
**Year:** 2016-2017
**Name:** Demo School 9999

**Legend: Performance Relative to the Test as a Whole**
- Performance is better than on the rest of the test
- Performance is similar to performance on the test as a whole
- Performance is worse than on the rest of the test
- Insufficient Information

**Legend: Performance Relative to Proficiency**
- Performance is above the Proficiency Standard
- Performance is near the Proficiency Standard
- Performance is below the Proficiency Standard
- Insufficient Information

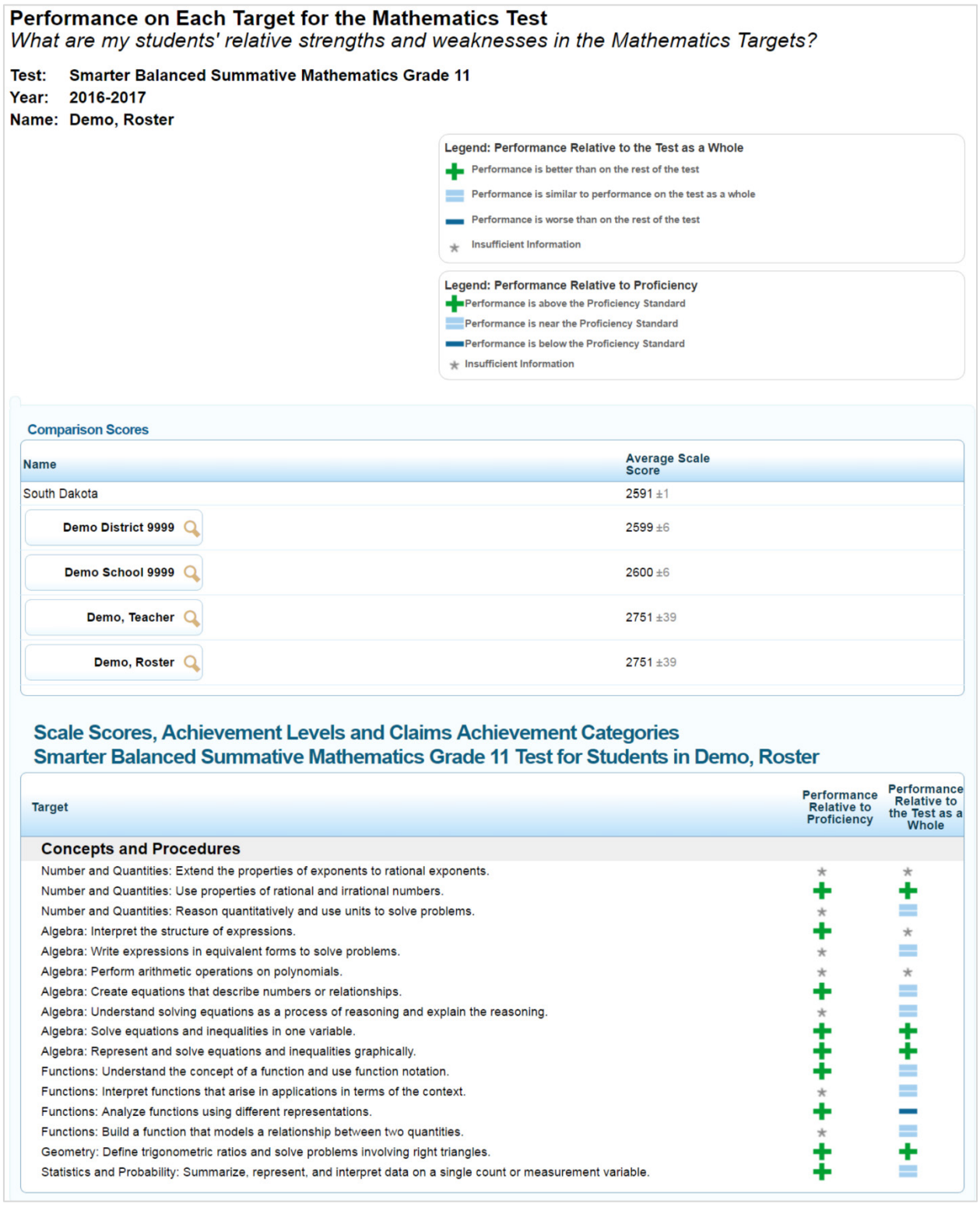
**Comparison Scores**

| Name | Average Scale Score |
|---|---|
| South Dakota | 2591 ±1 |
| Demo District 9999 | 2599 ±6 |
| Demo School 9999 | 2600 ±6 |

**Performance on Each Target**
**Smarter Balanced Summative Mathematics Grade 11 Test for Students in Demo School 9999**

| Target | Performance Relative to Proficiency | Performance Relative to the Test as a Whole |
|---|---|---|
| **Concepts and Procedures** | | |
| Number and Quantities: Extend the properties of exponents to rational exponents. | − | − |
| Number and Quantities: Use properties of rational and irrational numbers. | + | + |
| Number and Quantities: Reason quantitatively and use units to solve problems. | − | − |
| Algebra: Interpret the structure of expressions. | − | = |
| Algebra: Write expressions in equivalent forms to solve problems. | − | − |
| Algebra: Perform arithmetic operations on polynomials. | − | = |
| Algebra: Create equations that describe numbers or relationships. | − | = |
| Algebra: Understand solving equations as a process of reasoning and explain the reasoning. | − | = |
| Algebra: Solve equations and inequalities in one variable. | = | + |
| Algebra: Represent and solve equations and inequalities graphically. | − | − |
| Functions: Understand the concept of a function and use function notation. | − | − |
| Functions: Interpret functions that arise in applications in terms of the context. | − | − |
| Functions: Analyze functions using different representations. | = | = |
| Functions: Build a function that models a relationship between two quantities. | − | − |
| Geometry: Define trigonometric ratios and solve problems involving right triangles. | − | = |
| Statistics and Probability: Summarize, represent, and interpret data on a single count or measurement variable. | + | + |

Exhibit 8. Target Detail Page for Mathematics: Roster Level



**Performance on Each Target for the Mathematics Test**

*What are my students' relative strengths and weaknesses in the Mathematics Targets?*

**Test:** Smarter Balanced Summative Mathematics Grade 11
**Year:** 2016-2017
**Name:** Demo, Roster

**Legend: Performance Relative to the Test as a Whole**
- Performance is better than on the rest of the test
- Performance is similar to performance on the test as a whole
- Performance is worse than on the rest of the test
- Insufficient Information

**Legend: Performance Relative to Proficiency**
- Performance is above the Proficiency Standard
- Performance is near the Proficiency Standard
- Performance is below the Proficiency Standard
- Insufficient Information

**Comparison Scores**

| Name | Average Scale Score |
|---|---|
| South Dakota | 2591 ±1 |
| Demo District 9999 | 2599 ±6 |
| Demo School 9999 | 2600 ±6 |
| Demo, Teacher | 2751 ±39 |
| Demo, Roster | 2751 ±39 |

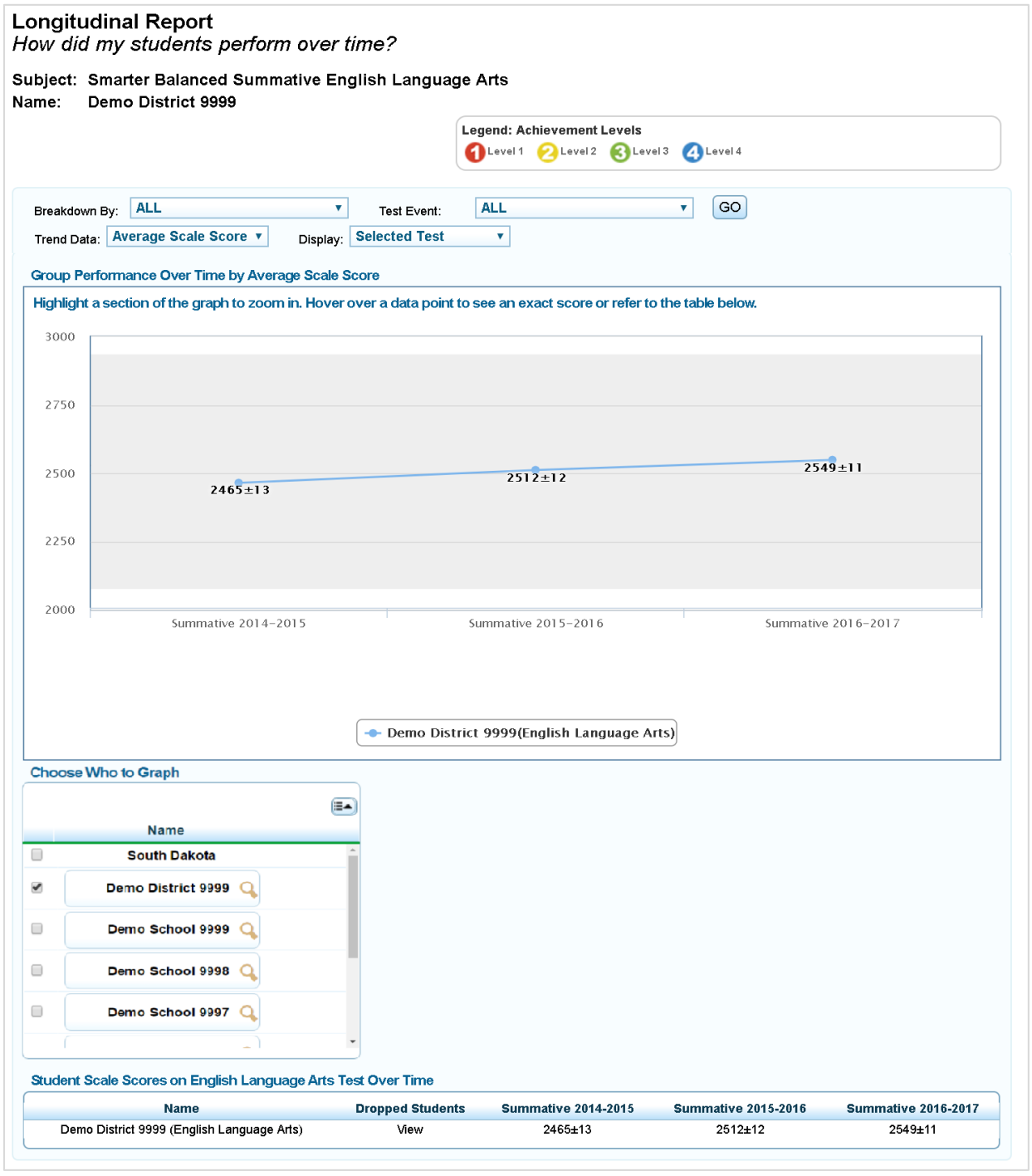
**Scale Scores, Achievement Levels and Claims Achievement Categories**
**Smarter Balanced Summative Mathematics Grade 11 Test for Students in Demo, Roster**

| Target | Performance Relative to Proficiency | Performance Relative to the Test as a Whole |
|---|---|---|
| **Concepts and Procedures** | | |
| Number and Quantities: Extend the properties of exponents to rational exponents. | ★ | ★ |
| Number and Quantities: Use properties of rational and irrational numbers. | + | + |
| Number and Quantities: Reason quantitatively and use units to solve problems. | ★ | = |
| Algebra: Interpret the structure of expressions. | + | ★ |
| Algebra: Write expressions in equivalent forms to solve problems. | ★ | = |
| Algebra: Perform arithmetic operations on polynomials. | ★ | ★ |
| Algebra: Create equations that describe numbers or relationships. | + | = |
| Algebra: Understand solving equations as a process of reasoning and explain the reasoning. | ★ | = |
| Algebra: Solve equations and inequalities in one variable. | + | + |
| Algebra: Represent and solve equations and inequalities graphically. | + | + |
| Functions: Understand the concept of a function and use function notation. | + | = |
| Functions: Interpret functions that arise in applications in terms of the context. | ★ | = |
| Functions: Analyze functions using different representations. | + | - |
| Functions: Build a function that models a relationship between two quantities. | ★ | = |
| Geometry: Define trigonometric ratios and solve problems involving right triangles. | + | + |
| Statistics and Probability: Summarize, represent, and interpret data on a single count or measurement variable. | + | = |

### *7.1.2.5 Trend Report Page*

The trend (i.e., longitudinal) page provides the trend of student performance for an aggregate, e.g., the state, district, and school, over time. The trend report can be set to plot either average scale scores or percentages of proficient students on the graph for the selected aggregate unit. In addition, the trend report can be plotted by demographic subgroups. Exhibit 9 presents an example of a trend report page for ELA/L at a district level.

Exhibit 9. Trend Report for ELA/L: District Level



#### *7.1.2.6 Student Detail Page*

When a student completes a test and the test is hand-scored, an online score report appears in the student detail page in the ORS. The student detail page provides individual student performance on the test. In each subject area, the student detail page provides (1) scale score and standard error of measurement (SEM), (2) achievement level for overall test, (3) performance category in each claim, (4) average scale scores for student's state, district, and school, and (5) writing performance descriptors in each dimension (ELA/L only).

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

Exhibits 10 and 11 present examples of student detail pages for ELA/L and mathematics, respectively.

Exhibit 10. Student Detail Page for ELA/L

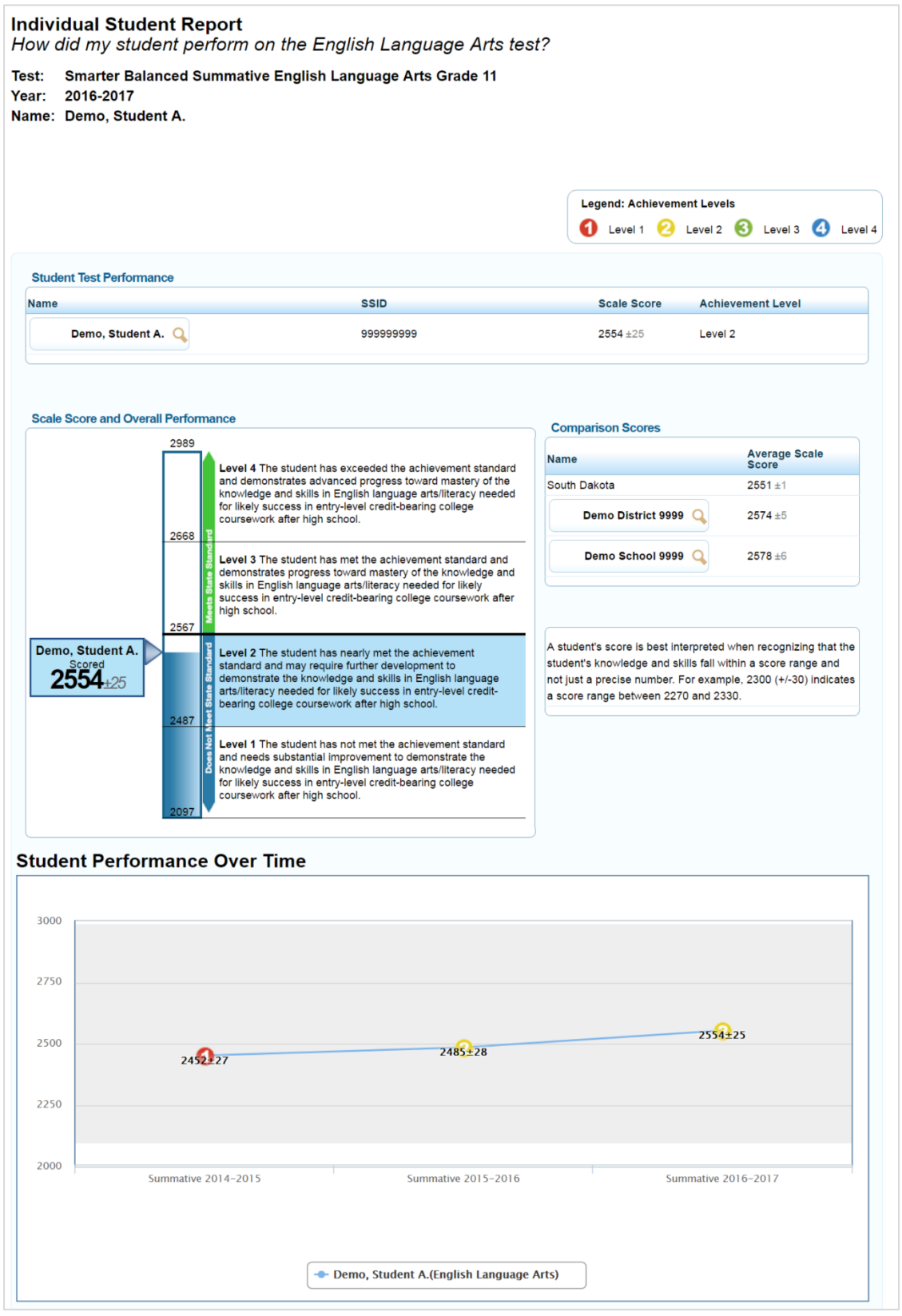


Exhibit 10. Student Detail Page for ELA/L (Continued)

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

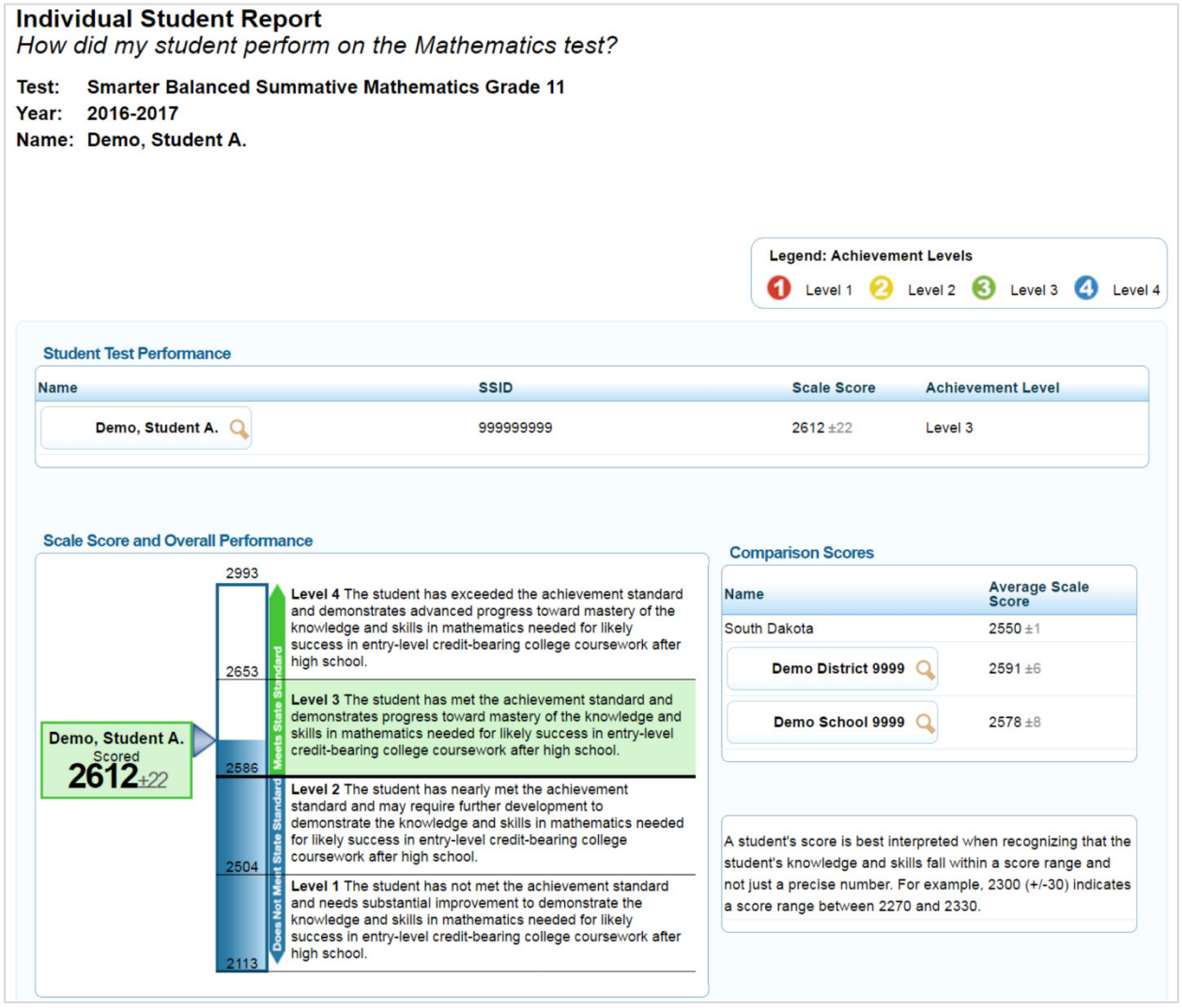
**Student Performance on Claims**

| Claim | Claim Performance | | Claim Description |
|---|---|---|---|
| Reading | Below the Standard Above the Standard | At/Near Standard | **What These Results Mean** Student may be able to read closely and analytically to comprehend a range of increasingly complex literary and informational texts. **Next Steps** Have your child study different texts that present conflicting points of view on the same topic. Compare the texts to other ideas (like myths or historical events) and point out analogies (comparing unlike ideas). |
| Writing | Below the Standard Above the Standard | At/Near Standard | **What These Results Mean** Student may be able to produce effective and well-grounded writing for a range of purposes and audiences. **Next Steps** Help your child write argumentative essays, which address opposing views and include a counterclaim, logical reasoning, and support. All essays need direct quotations and formal, subject-specific language. |
| Listening | Below the Standard Above the Standard | At/Near Standard | **What These Results Mean** Student may be able to employ effective listening skills for a range of purposes and audiences. **Next Steps** Have your child listen to a documentary and explain how presentation (graphics, tone of voice, music) relates to purpose. Point out when there is not enough evidence or when evidence is unrelated to the topic and why. |
| Research/Inquiry | Below the Standard Above the Standard | At/Near Standard | **What These Results Mean** Student may be able to engage in research and inquiry to investigate topics, and to analyze, integrate, and present information. **Next Steps** Have your child conduct a short research project based on his or her questions on a topic. He or she should include several sides of the topic, combine data, use quotations from sources, and his or her own ideas. |

**Writing Performance Based on Smarter Balanced Performance Task Writing Rubric**

| Essay | Organization/Purpose | Evidence/Elaboration | Conventions |
|---|---|---|---|
| Argumentative | The argumentative response has an inconsistent structure including an unclear claim, uneven development, few transitions, and loosely connected ideas. If present, the introduction or conclusion may be weak. The response may address the opposing argument. (2 out of 4 points) | The argumentative response provides uneven elaboration to support the claim including few facts and details cited from sources, weak elaborative techniques and ineffective language for the audience and purpose. (2 out of 4 points) | The argumentative response shows an adequate understanding of correct sentence formation, punctuation, capitalization, grammar usage, and spelling. (2 out of 2 points) |

Exhibit 11. Student Detail Page for Mathematics



**Individual Student Report**

*How did my student perform on the Mathematics test?*

**Test:** Smarter Balanced Summative Mathematics Grade 11
**Year:** 2016-2017
**Name:** Demo, Student A.

**Legend: Achievement Levels**
1 Level 1 2 Level 2 3 Level 3 4 Level 4

**Student Test Performance**

| Name | SSID | Scale Score | Achievement Level |
|---|---|---|---|
| Demo, Student A. | 999999999 | 2612 ±22 | Level 3 |

**Scale Score and Overall Performance**

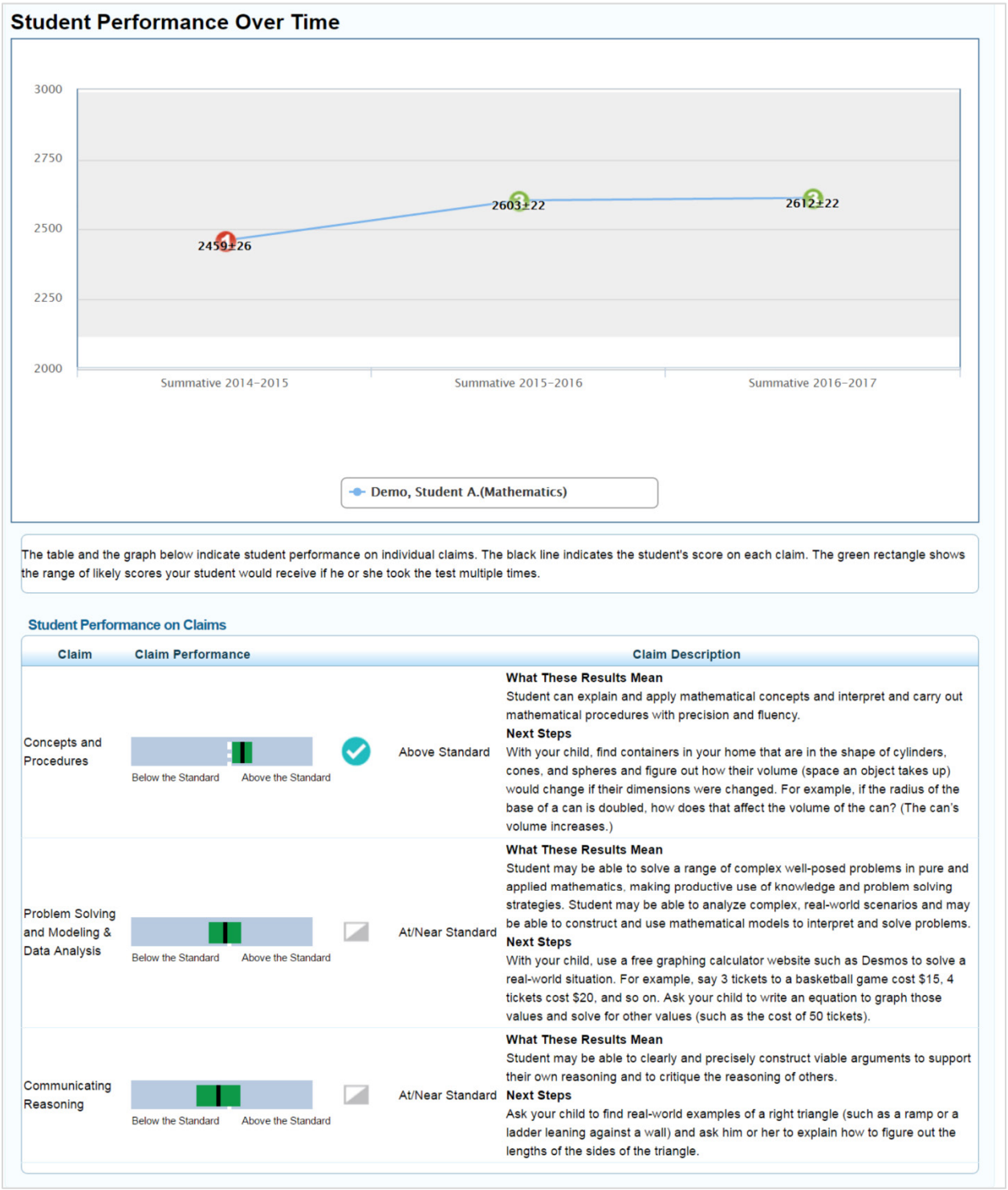


**Comparison Scores**

| Name | Average Scale Score |
|---|---|
| South Dakota | 2550 ±1 |
| Demo District 9999 | 2591 ±6 |
| Demo School 9999 | 2578 ±8 |

A student's score is best interpreted when recognizing that the student's knowledge and skills fall within a score range and not just a precise number. For example, 2300 (+/-30) indicates a score range between 2270 and 2330.

Exhibit 11. Student Detail Page for Mathematics (Continued)

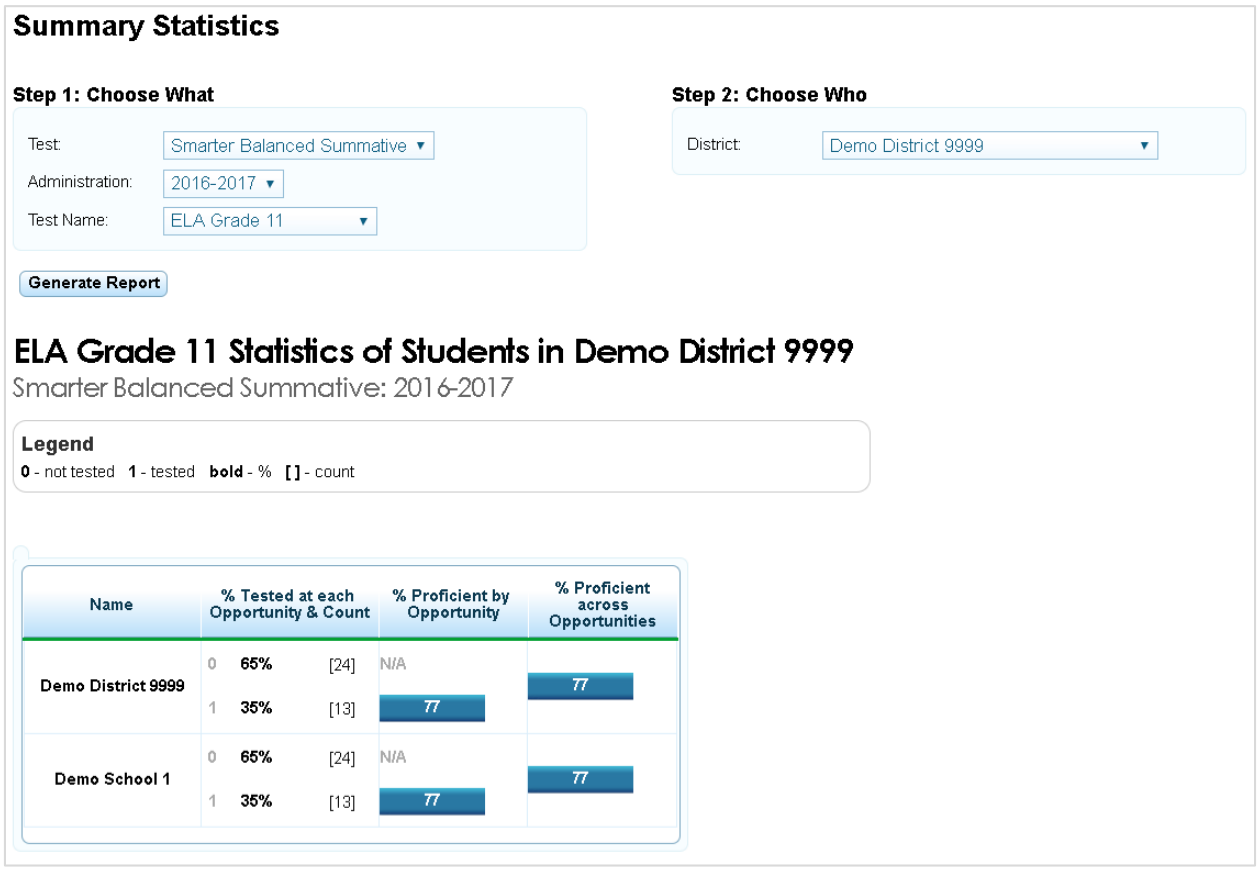



### *7.1.2.7 Participation Rate*

In addition to online score reports, the ORS provides participation rate reports for the districts and schools to help monitor the student participation rate. Participation data are updated each time students complete

tests and these tests are hand-scored. Included in the participation table are (1) number and percent of students who are tested and not tested and (2) percent proficient. Exhibit 12 presents a sample of the participation rate report at a district level.

Exhibit 12. Participation Rate Report at District Level




## 7.2 Interpretation of Reported Scores

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

### 7.2.1 Scale Score

A scale score is used to describe how well a student performed on a test and can be interpreted as an estimate of the student's knowledge and skills measured. The scale score is the transformed score from a theta score, which is estimated based on mathematical models. Low scale scores can be interpreted to mean that the student does not possess sufficient knowledge and skills measured by the test. Conversely, high scale scores can be interpreted to mean that the student has 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.

### 7.2.2 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. The SEM can be different for the same scale score, depending on how closely the administered items match the student's ability.

### 7.2.3 Achievement Level

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

### 7.2.4 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 their 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 the students' performance does not provide enough information to tell whether students reached the "Meets Standard" mark for the specific claim.

### 7.2.5 Performance Category for Targets

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

AIR reports two ways of relative strength and weakness scores for each target within a claim. The strengths and weaknesses report is generated for aggregate units of classroom, school, and district, and provides information about how a group of students in a class, school, or district performed on the reporting target, either relative to their performance on the test as a whole or relative to proficiency cut set by Smarter Balanced. Specifically, for target performance relative to the test as a whole, students' observed performance on items within the reporting element is compared with expected performance based on the overall ability estimate. At the aggregate level, when observed performance within a target is greater than expected performance, then the reporting unit (e.g., roster, teacher, school, or district) shows a relative strength in that target. Conversely, when observed performance within a target is below the level expected based on overall achievement, then the reporting unit shows a relative weakness in that target. For target performance relative to proficiency, students' observed performance on items within the reporting element is compared with proficiency cut (i.e., Achievement Level 3 cut). At the aggregate level, when observed performance within a target is greater than the proficiency cut, the reporting unit shows a relative strength in that target. Conversely, when observed performance within a target is below the proficiency cut, the reporting unit shows a relative weakness in that target.

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

### 7.2.6 Aggregated Score

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

## 7.3 Appropriate Uses for Scores and Reports

Assessment results can be used to provide information on individual students' achievement on the test. Overall, assessment results show what students know and are able to do in certain subject areas. Further, they give information on whether students are on track to demonstrate the knowledge and skills necessary for college and their careers. 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 several targets compared to their overall performance. In this case, teachers or schools can identify strengths and weaknesses of their students through the group performance by claim and target and promote instruction on specific claim or target areas that the group performance is below their overall performance. Further, by narrowing down the student performance result by subgroup, teachers and schools can determine what strategies may need to be implemented to improve teaching and student learning, particularly for students from a disadvantaged subgroup. For example, teachers can see student assessment results by LEP status and observe that LEP students are struggling with literary response and analysis in reading. Teachers can then provide additional instructions for these students to enhance their achievement of the benchmarks for literary response and analysis.

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

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

# 8. QUALITY CONTROL PROCEDURE

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

## 8.1 ADAPTIVE TEST CONFIGURATION

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

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 check the score accuracy.

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

### 8.1.1 Platform Review

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

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

Platform review is conducted by a team. The team leader projects the item as it was web approved in 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

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

## 8.3 QUALITY ASSURANCE IN DATA PREPARATION

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

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

## 8.4 QUALITY ASSURANCE IN HANDSCORING

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

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

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

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

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

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

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

### 8.4.2 Handscoring QA Monitoring Reports

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

### 8.4.3 Monitoring by State Department of Education

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

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

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

This process is also used to notify each Consortium state of possible instances of teacher or proctor interference or student collusion with others. The alert procedure is habitually explained during scorer training sessions. Within the VSC system, if a scorer identifies a response which may require an alert, he

or she flags or notes that response as a possible alert and transfers the image to the scoring manager. Scoring management then decides if the response should be forwarded to the client for any necessary action or follow-up.

## 8.5 QUALITY ASSURANCE IN TEST SCORING

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

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

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

For the computer adaptive test component, other reports such as blueprint match and item exposure reports allow psychometricians to verify that test administrations conform to the simulation results. The QA reports can be generated on any desired schedule. Item analysis and blueprint match reports are evaluated frequently at the opening of the testing window to ensure that test administrations conform to blueprint and items are performing as anticipated. Table 48 presents an overview of the QA reports.

Table 48. 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 2016–2017 Smarter Balanced summative assessment, two types of score reports were produced: online reports and printed reports (family reports only).

*8.5.1.1 Online Report Quality Assurance*

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

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

Every test undergoes a series of validation checks. Once the QA system signs off, data are passed to the DoR, which serves as the centralized location for all student scores and responses, ensuring 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; within less than a second of handscores being received by AIR and passing QA validation checks, the composite score will be available in the ORS.

#### *8.5.1.2 Paper Report Quality Assurance*

*Statistical Programming*

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

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

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

*Display Programming*

The paper report development process uses graphical programming, which takes place in a Xerox-developed programming language called VIPP and allows virtually infinite control of the visual appearance of the reports. After designers at AIR create backgrounds, our VIPP programmers write code that indicates where to place all variable information (data, graphics, and text) on the reports. The VIPP code is tested using both artificial and real data. AIR's data generation utilities can read the output layout specifications and generate artificial data for direct input into the VIPP programs. This allows the testing of these programs to begin before the statistical programming is complete. In later stages, artificial data are generated according to the input layout and run through the psychometric process and the score reporting statistical programs, and the output is formatted as VIPP input. This enables us to test the entire system. Programmed output goes through multiple stages of review and revision by graphics editors and the score reporting team to ensure that design elements are accurately reproduced and data are correctly displayed. Once we receive final data and VIPP programs, the AIR Score Reporting team reviews proofs that contain actual data based on our standard quality assurance documentation. In addition, we compare data independently calculated by AIR psychometricians with data on the reports. A large sample of reports is reviewed by several AIR staff members to make sure that all data are correctly placed on reports. This rigorous review typically is conducted over several days and takes place in a secure location in the AIR building. All reports containing actual data are stored in a locked storage area. Prior to printing the reports, AIR provides a live data file and individual student reports with sample districts for 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.

Ben-Simon, A., & Bennett, R. E. (2007). Toward more substantively meaningful automated essay scoring. *The Journal of Technology, Learning and Assessment*, *6*(1), 1–47.

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.

Leacock, C., Messineo, D., & Zhang, X. (2013). Issues in prompt selection for automated scoring of short-answer questions. Paper presented at the National Council on Measurement in Education, San Francisco, CA.

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.

Page, E. B. (1966). The imminence of... grading essays by computer. *The Phi Delta Kappan*, *47*(5), 238–243.

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*, *33*(1–2), 1–39.

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

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

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

# APPENDICES

# Appendix A: Number of Students for Interim Assessments

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

Table A–1. Number of Students Who Took ICAs

| Grade | ELA/L | | | Mathematics | | |
|---|---|---|---|---|---|---|
| | Once | Twice | Total | Once | Twice | Total |
| 3 | 97 | 0 | 97 | 98 | 0 | 98 |
| 4 | 114 | 0 | 114 | 128 | 0 | 128 |
| 5 | 102 | 0 | 102 | 160 | 0 | 160 |
| 6 | 57 | 0 | 57 | 77 | 0 | 77 |
| 7 | 39 | 0 | 39 | 79 | 0 | 79 |
| 8 | 55 | 0 | 55 | 79 | 0 | 79 |
| 11 | 79 | 0 | 79 | 487 | 1 | 488 |

For the Interim Assessment Blocks (IAB), there were seven to nine IABs for ELA/L and five to six IABs in mathematics. Students were allowed to take as many IABs as they wanted. Table A–2 presents the total number of students who took the IABs and the number of students by the number of IABs taken. For example, in grade 3 ELA/L, a total of 2,650 students took IABs, and among 2,650 students, 1,408 students took one IAB, 540 students took two IABs, and so on.

Tables A–3 and A–4 disaggregated the number of students in Table A-1 by each individual block. For example, 1,408 students in grade 3 ELA/L took one IAB only. Among 1,408 students, one 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 | 8 | 9 |
| **ELA/L** | | | | | | | | | | |
| 3 | 2,650 | 1,408 | 540 | 518 | 118 | 43 | 12 | 11 | | |
| 4 | 2,894 | 1,246 | 571 | 605 | 194 | 161 | 32 | 67 | 18 | |
| 5 | 2,310 | 1,011 | 544 | 403 | 185 | 97 | 21 | 47 | 2 | |
| 6 | 2,715 | 720 | 1,006 | 643 | 179 | 124 | 41 | | 2 | |
| 7 | 1,943 | 796 | 545 | 269 | 233 | 89 | 6 | 5 | | |
| 8 | 1,831 | 1,163 | 380 | 250 | 27 | 10 | 1 | | | |
| 11 | 608 | 327 | 108 | 61 | 45 | 67 | | | | |
| **Mathematics** | | | | | | | | | | |
| 3 | 4,100 | 1,803 | 939 | 660 | 691 | 7 | | | | |
| 4 | 4,414 | 1,700 | 885 | 664 | 571 | 593 | 1 | | | |
| 5 | 3,834 | 1,237 | 970 | 625 | 385 | 610 | 7 | | | |
| 6 | 4,238 | 1,701 | 1,258 | 700 | 287 | 209 | 83 | | | |
| 7 | 3,756 | 1,533 | 870 | 481 | 321 | 474 | 77 | | | |
| 8 | 3,428 | 1,629 | 705 | 339 | 654 | 101 | | | | |
| 11 | 2,149 | 1,086 | 385 | 312 | 363 | 3 | | | | |

Table A–3: ELA/L Number of Students Who Took IABs by Block Labels (Grades 3–6)

| Grade | Block | Number of IABs Taken | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 3 | Brief Writes | 1 | 3 | | | | | | | |
| | Editing | 207 | 275 | 353 | 110 | 43 | 12 | 11 | | |
| | Language and Vocabulary Use | 680 | 364 | 401 | 77 | 43 | 12 | 11 | | |
| | Listening and Interpretation | 234 | 194 | 397 | 71 | 43 | 12 | 11 | | |
| | Performance Task | 5 | 2 | 57 | 2 | | | | | |
| | Reading Informational Text | 10 | 21 | 13 | 39 | 1 | 7 | 11 | | |
| | Reading Literary Text | 62 | 97 | 41 | 27 | 3 | 5 | 11 | | |
| | Research | 204 | 112 | 223 | 82 | 40 | 12 | 11 | | |
| | Revision | 5 | 12 | 69 | 64 | 42 | 12 | 11 | | |
| 4 | Brief Writes | | | | | | 1 | 2 | 17 | |
| | Editing | 223 | 257 | 414 | 169 | 151 | 31 | 67 | 18 | |
| | Language and Vocabulary Use | 331 | 379 | 463 | 181 | 138 | 32 | 67 | 18 | |
| | Listening and Interpretation | 334 | 199 | 359 | 83 | 149 | 30 | 67 | 18 | |
| | Performance Task | 1 | 1 | 32 | 1 | | | | 1 | |
| | Reading Informational Text | 125 | 102 | 150 | 50 | 62 | 16 | 66 | 18 | |
| | Reading Literary Text | 62 | 131 | 140 | 32 | 36 | 20 | 66 | 18 | |
| | Research | 160 | 49 | 193 | 153 | 154 | 31 | 67 | 18 | |
| | Revision | 10 | 24 | 64 | 107 | 115 | 31 | 67 | 18 | |
| 5 | Brief Writes | 1 | 9 | 16 | 13 | 10 | 3 | 2 | 2 | |
| | Editing | 173 | 180 | 152 | 175 | 95 | 19 | 46 | 2 | |
| | Language and Vocabulary Use | 535 | 386 | 308 | 164 | 79 | 19 | 47 | 2 | |
| | Listening and Interpretation | 110 | 230 | 254 | 153 | 82 | 15 | 47 | 2 | |
| | Performance Task | 3 | 11 | 57 | 22 | 10 | 5 | | | |
| | Reading Informational Text | 50 | 33 | 64 | 26 | 36 | 19 | 47 | 2 | |
| | Reading Literary Text | | 15 | 22 | 12 | 16 | 12 | 46 | 2 | |
| | Research | 119 | 175 | 283 | 64 | 76 | 16 | 47 | 2 | |
| | Revision | 20 | 49 | 53 | 111 | 81 | 18 | 47 | 2 | |
| 6 | Brief Writes | 2 | 1 | | 7 | 3 | 1 | | 2 | |
| | Editing | 132 | 379 | 433 | 118 | 116 | 41 | | 2 | |
| | Language and Vocabulary Use | 165 | 636 | 488 | 162 | 124 | 41 | | 2 | |
| | Listening and Interpretation | 97 | 407 | 216 | 66 | 74 | 15 | | 2 | |
| | Performance Task | 20 | 8 | 63 | 9 | 4 | 1 | | 2 | |
| | Reading Informational Text | 104 | 135 | 237 | 55 | 21 | 36 | | 1 | |
| | Reading Literary Text | 145 | 311 | 92 | 45 | 58 | 34 | | 1 | |
| | Research | 43 | 86 | 176 | 114 | 108 | 41 | | 2 | |
| | Revision | 12 | 49 | 224 | 140 | 112 | 36 | | 2 | |

Table A–4: ELA/L Number of Students Who Took IABs by Block Labels (Grades 7–8, 11)

| Grade | Block | Number of IABs Taken | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 7 | Brief Writes | 6 | 22 | 2 | | | | | | |
| | Editing | 180 | 244 | 191 | 217 | 88 | 5 | 5 | | |
| | Language and Vocabulary Use | 289 | 269 | 169 | 211 | 89 | 6 | 5 | | |
| | Listening and Interpretation | 5 | 8 | 46 | 41 | 31 | 6 | 5 | | |
| | Performance Task | 31 | 20 | | | | 1 | | | |
| | Reading Informational Text | 191 | 165 | 115 | 84 | 69 | 4 | 5 | | |
| | Reading Literary Text | 67 | 149 | 33 | 87 | 13 | 5 | 5 | | |
| | Research | 1 | 15 | 65 | 98 | 78 | 6 | 5 | | |
| | Revision | 26 | 198 | 186 | 194 | 77 | 3 | 5 | | |
| 8 | Brief Writes | | 52 | 1 | | | 1 | | | |
| | Editing and Revising | 432 | 145 | 248 | 27 | 10 | | | | |
| | Listening and Interpretation | 85 | 61 | 217 | 12 | 5 | 1 | | | |
| | Performance Task | 47 | 101 | 8 | 6 | 10 | 1 | | | |
| | Reading Informational Text | 193 | 202 | 17 | 20 | 10 | 1 | | | |
| | Reading Literary Text | 178 | 106 | 19 | 21 | 10 | 1 | | | |
| | Research | 228 | 93 | 240 | 22 | 5 | 1 | | | |
| 11 | Brief Writes | | | | | | | | | |
| | Editing | 7 | 64 | 55 | 41 | 67 | | | | |
| | Language and Vocabulary Use | 223 | 43 | 42 | 45 | 67 | | | | |
| | Listening and Interpretation | 19 | 1 | 17 | 3 | 66 | | | | |
| | Performance Task | | 20 | 1 | 31 | | | | | |
| | Reading Informational Text | 47 | 17 | 6 | 4 | 1 | | | | |
| | Reading Literary Text | 28 | 26 | 33 | 12 | 3 | | | | |
| | Research | 2 | 43 | 26 | 14 | 67 | | | | |
| | Revision | 1 | 2 | 3 | 30 | 64 | | | | |

Table A–5: Mathematics Number of Students Who Took IABs by Block Labels

| Grade | Block | Number of IABs Taken | | | | | |
|---|---|---|---|---|---|---|---|
| | | **1** | **2** | **3** | **4** | **5** | **6** |
| 3 | Measurement and Data | 308 | 388 | 517 | 690 | 7 | |
| | Number and Operations in Base Ten | 711 | 536 | 472 | 690 | 7 | |
| | Number and Operations – Fractions | 539 | 444 | 474 | 690 | 7 | |
| | Operational and Algebraic Thinking | 243 | 448 | 487 | 687 | 7 | |
| | Performance Task | 2 | 62 | 30 | 7 | 7 | |
| 4 | Measurement and Data | 44 | 166 | 228 | 389 | 591 | 1 |
| | Number and Operations in Base Ten | 513 | 603 | 514 | 551 | 593 | 1 |
| | Number and Operations – Fractions | 681 | 474 | 443 | 516 | 593 | 1 |
| | Operational and Algebraic Thinking | 355 | 301 | 426 | 422 | 593 | 1 |
| | Geometry | 104 | 226 | 381 | 404 | 593 | 1 |
| | Performance Task | 3 | | | 2 | 2 | 1 |
| 5 | Measurement and Data | 41 | 186 | 258 | 221 | 610 | 7 |
| | Number and Operations in Base Ten | 322 | 737 | 482 | 368 | 610 | 7 |
| | Number and Operations – Fractions | 702 | 626 | 453 | 348 | 607 | 7 |
| | Geometry | 36 | 127 | 342 | 344 | 610 | 7 |
| | Operations and Algebraic Thinking | 136 | 233 | 338 | 259 | 610 | 7 |
| | Performance Task | | 31 | 2 | | 3 | 7 |
| 6 | Expressions and Equations | 613 | 370 | 415 | 280 | 207 | 83 |
| | Geometry | 222 | 381 | 422 | 201 | 207 | 83 |
| | Number System | 237 | 814 | 575 | 280 | 209 | 83 |
| | Statistics and Probability | 16 | 5 | 75 | 118 | 206 | 83 |
| | Performance Task | 2 | 16 | 9 | 7 | 13 | 83 |
| | Ratios and Proportional Relationships | 611 | 930 | 604 | 262 | 203 | 83 |
| 7 | Expressions and Equations | 602 | 643 | 446 | 298 | 474 | 77 |
| | Number System | 531 | 668 | 358 | 263 | 474 | 77 |
| | Geometry | 47 | 99 | 150 | 267 | 474 | 77 |
| | Statistics and Probability | 4 | 37 | 60 | 188 | 474 | 77 |
| | Performance Task | | 1 | | 1 | 5 | 77 |
| | Ratios and Proportional Relationships | 349 | 292 | 429 | 267 | 469 | 77 |
| 8 | Expressions and Equations I | 42 | 208 | 200 | 652 | 101 | |
| | Expressions and Equations II | 249 | 392 | 287 | 654 | 101 | |
| | Functions | 1,109 | 459 | 302 | 654 | 101 | |
| | Geometry | 229 | 350 | 227 | 650 | 101 | |
| | Performance Task | | 1 | 1 | 6 | 101 | |
| 11 | Algebra – Linear Functions | 351 | 313 | 296 | 363 | 3 | |
| | Algebra – Quadratic Functions | 148 | 196 | 237 | 363 | 3 | |
| | Geometry – Right Triangles and Trigonometric | 577 | 175 | 225 | 363 | 3 | |
| | Statistics and Probability | 10 | 85 | 178 | 361 | 3 | |
| | Performance Task | | 1 | | 2 | 3 | |

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

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

| Group | 2014–2015 | 2015–2016 | 2016–2017 |
|---|---|---|---|
| **Grade 3** | | | |
| All Students | 47 | 49 | 47 |
| Female | 51 | 52 | 50 |
| Male | 43 | 45 | 44 |
| American Indian/Alaska Native | 14 | 16 | 14 |
| Asian | 48 | 42 | 42 |
| African American | 30 | 31 | 33 |
| Hispanic/Latino | 31 | 39 | 31 |
| Native Hawaiian/Pacific Islander | - | 58 | 50 |
| White | 56 | 58 | 56 |
| Multiple Ethnicities | 40 | 46 | 45 |
| LEP | 21 | 22 | 21 |
| IDEA | 22 | 23 | 22 |
| Section 504 | 52 | 50 | 46 |
| **Grade 4** | | | |
| All Students | 44 | 48 | 48 |
| Female | 49 | 52 | 52 |
| Male | 39 | 45 | 45 |
| American Indian/Alaska Native | 14 | 14 | 16 |
| Asian | 45 | 53 | 48 |
| African American | 25 | 33 | 30 |
| Hispanic/Latino | 31 | 30 | 37 |
| Native Hawaiian/Pacific Islander | 33 | - | 50 |
| White | 52 | 58 | 57 |
| Multiple Ethnicities | 49 | 44 | 42 |
| LEP | 10 | 8 | 12 |
| IDEA | 17 | 19 | 20 |
| Section 504 | 45 | 51 | 49 |
| **Grade 5** | | | |
| All Students | 47 | 48 | 50 |
| Female | 53 | 55 | 55 |
| Male | 42 | 42 | 45 |
| American Indian/Alaska Native | 17 | 17 | 15 |
| Asian | 51 | 51 | 47 |
| African American | 34 | 29 | 34 |
| Hispanic/Latino | 33 | 34 | 34 |
| Native Hawaiian/Pacific Islander | - | 45 | 60 |
| White | 55 | 57 | 59 |
| Multiple Ethnicities | 43 | 50 | 45 |
| LEP | 5 | 5 | 6 |
| IDEA | 16 | 15 | 16 |
| Section 504 | 48 | 47 | 47 |

**"-"** Suppressed data due to the small sample size, n<10.

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

| Group | 2014–2015 | 2015–2016 | 2016–2017 |
|---|---|---|---|
| **Grade 6** | | | |
| All Students | 44 | 49 | 48 |
| Female | 50 | 55 | 54 |
| Male | 37 | 44 | 42 |
| American Indian/Alaska Native | 13 | 18 | 15 |
| Asian | 43 | 60 | 55 |
| African American | 30 | 33 | 29 |
| Hispanic/Latino | 34 | 38 | 35 |
| Native Hawaiian/Pacific Islander | - | - | 30 |
| White | 51 | 58 | 56 |
| Multiple Ethnicities | 40 | 44 | 47 |
| LEP | 7 | 8 | 3 |
| IDEA | 10 | 12 | 11 |
| Section 504 | 39 | 48 | 44 |
| **Grade 7** | | | |
| All Students | 48 | 50 | 52 |
| Female | 55 | 57 | 58 |
| Male | 41 | 44 | 46 |
| American Indian/Alaska Native | 17 | 18 | 19 |
| Asian | 39 | 48 | 61 |
| African American | 25 | 35 | 30 |
| Hispanic/Latino | 36 | 41 | 39 |
| Native Hawaiian/Pacific Islander | - | 50 | 42 |
| White | 56 | 59 | 61 |
| Multiple Ethnicities | 41 | 47 | 46 |
| LEP | 4 | 6 | 6 |
| IDEA | 11 | 12 | 12 |
| Section 504 | 45 | 47 | 50 |
| **Grade 8** | | | |
| All Students | 47 | 51 | 48 |
| Female | 54 | 59 | 55 |
| Male | 39 | 43 | 40 |
| American Indian/Alaska Native | 18 | 20 | 16 |
| Asian | 48 | 44 | 49 |
| African American | 33 | 29 | 30 |
| Hispanic/Latino | 36 | 37 | 38 |
| Native Hawaiian/Pacific Islander | 46 | 50 | 63 |
| White | 54 | 59 | 56 |
| Multiple Ethnicities | 46 | 48 | 40 |
| LEP | 6 | 8 | 4 |
| IDEA | 9 | 11 | 10 |
| Section 504 | 35 | 47 | 38 |

**"-"** Suppressed data due to the small sample size, n<10.

Table B–3. ELA/L Percentages of Proficient Students Across Years (Grade 11)

| Group | 2014–2015 | 2015–2016 | 2016–2017 |
|---|---|---|---|
| **Grade 11** | | | |
| All Students | 58 | 58 | 64 |
| Female | 66 | 65 | 70 |
| Male | 50 | 52 | 58 |
| American Indian/Alaska Native | 29 | 27 | 29 |
| Asian | 39 | 40 | 46 |
| African American | 41 | 35 | 30 |
| Hispanic/Latino | 44 | 45 | 52 |
| Native Hawaiian/Pacific Islander | 60 | - | 73 |
| White | 64 | 65 | 71 |
| Multiple Ethnicities | 54 | 59 | 64 |
| LEP | 4 | 4 | 5 |
| IDEA | 12 | 13 | 17 |
| Section 504 | 54 | 53 | 68 |

**"-"** Suppressed data due to the small sample size, n<10.

Table B–4. Mathematics Percentages of Proficient Students Across Years (Grades 3–5)

| Group | 2014–2015 | 2015–2016 | 2016–2017 |
|---|---|---|---|
| **Grade 3** | | | |
| All Students | 49 | 52 | 53 |
| Female | 48 | 51 | 52 |
| Male | 50 | 53 | 54 |
| American Indian/Alaska Native | 17 | 18 | 18 |
| Asian | 40 | 43 | 50 |
| African American | 29 | 27 | 28 |
| Hispanic/Latino | 31 | 39 | 36 |
| Native Hawaiian/Pacific Islander | - | 33 | - |
| White | 59 | 63 | 63 |
| Multiple Ethnicities | 43 | 47 | 44 |
| LEP | 18 | 21 | 25 |
| IDEA | 27 | 27 | 31 |
| Section 504 | 57 | 51 | 54 |
| **Grade 4** | | | |
| All Students | 44 | 47 | 49 |
| Female | 43 | 45 | 46 |
| Male | 45 | 48 | 51 |
| American Indian/Alaska Native | 13 | 13 | 15 |
| Asian | 46 | 43 | 47 |
| African American | 19 | 24 | 24 |
| Hispanic/Latino | 28 | 30 | 33 |
| Native Hawaiian/Pacific Islander | 42 | - | 31 |
| White | 53 | 56 | 59 |
| Multiple Ethnicities | 42 | 40 | 41 |
| LEP | 9 | 7 | 12 |
| IDEA | 18 | 21 | 21 |
| Section 504 | 49 | 51 | 47 |
| **Grade 5** | | | |
| All Students | 35 | 37 | 40 |
| Female | 33 | 35 | 38 |
| Male | 38 | 38 | 42 |
| American Indian/Alaska Native | 8 | 10 | 9 |
| Asian | 39 | 44 | 40 |
| African American | 20 | 11 | 20 |
| Hispanic/Latino | 20 | 22 | 24 |
| Native Hawaiian/Pacific Islander | - | 36 | 40 |
| White | 43 | 45 | 49 |
| Multiple Ethnicities | 31 | 36 | 36 |
| LEP | 3 | 6 | 3 |
| IDEA | 11 | 11 | 13 |
| Section 504 | 39 | 38 | 41 |

**"-"** Suppressed data due to the small sample size, n<10.

Table B–5. Mathematics Percentages of Proficient Students Across Years (Grades 6–8)

| Group | 2014–2015 | 2015–2016 | 2016–2017 |
|---|---|---|---|
| **Grade 6** | | | |
| All Students | 33 | 39 | 41 |
| Female | 33 | 39 | 43 |
| Male | 33 | 40 | 39 |
| American Indian/Alaska Native | 6 | 9 | 10 |
| Asian | 35 | 45 | 52 |
| African American | 14 | 20 | 17 |
| Hispanic/Latino | 21 | 22 | 25 |
| Native Hawaiian/Pacific Islander | - | - | 30 |
| White | 41 | 48 | 50 |
| Multiple Ethnicities | 26 | 34 | 39 |
| LEP | 5 | 5 | 5 |
| IDEA | 7 | 9 | 9 |
| Section 504 | 30 | 42 | 35 |
| **Grade 7** | | | |
| All Students | 38 | 41 | 43 |
| Female | 38 | 41 | 43 |
| Male | 38 | 41 | 44 |
| American Indian/Alaska Native | 9 | 10 | 12 |
| Asian | 37 | 41 | 54 |
| African American | 16 | 23 | 23 |
| Hispanic/Latino | 22 | 29 | 25 |
| Native Hawaiian/Pacific Islander | - | 50 | 33 |
| White | 46 | 49 | 53 |
| Multiple Ethnicities | 33 | 40 | 39 |
| LEP | 7 | 3 | 6 |
| IDEA | 8 | 8 | 10 |
| Section 504 | 35 | 39 | 41 |
| **Grade 8** | | | |
| All Students | 37 | 41 | 41 |
| Female | 38 | 43 | 43 |
| Male | 36 | 38 | 38 |
| American Indian/Alaska Native | 9 | 10 | 8 |
| Asian | 48 | 38 | 38 |
| African American | 23 | 19 | 22 |
| Hispanic/Latino | 24 | 25 | 28 |
| Native Hawaiian/Pacific Islander | 15 | 40 | - |
| White | 44 | 49 | 49 |
| Multiple Ethnicities | 31 | 27 | 37 |
| LEP | 7 | 6 | 4 |
| IDEA | 6 | 7 | 7 |
| Section 504 | 24 | 36 | 36 |

**"-"** Suppressed data due to the small sample size, n<10.

Table B–6. Mathematics Percentages of Proficient Students Across Years (Grade 11)

| Group | 2014–2015 | 2015–2016 | 2016–2017 |
|---|---|---|---|
| **Grade 11** | | | |
| All Students | 37 | 36 | 40 |
| Female | 39 | 38 | 41 |
| Male | 35 | 35 | 38 |
| American Indian/Alaska Native | 10 | 7 | 9 |
| Asian | 26 | 29 | 35 |
| African American | 20 | 21 | 15 |
| Hispanic/Latino | 19 | 22 | 23 |
| Native Hawaiian/Pacific Islander | 30 | - | 40 |
| White | 43 | 42 | 46 |
| Multiple Ethnicities | 35 | 32 | 30 |
| LEP | 2 | 2 | 2 |
| IDEA | 4 | 4 | 5 |
| Section 504 | 31 | 32 | 34 |

**"-"** Suppressed data due to the small sample size, n<10.

# Appendix C: Classification Accuracy and Consistency Index by Subgroups

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

| Group | N | %Accuracy | | | | | %Consistency | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | All | L1 | L2 | L3 | L4 | All | L1 | L2 | L3 | L4 |
| **Grade 3** | | | | | | | | | | | |
| All | 11,398 | 80 | 90 | 72 | 70 | 87 | 72 | 84 | 61 | 60 | 81 |
| Female | 5,630 | 80 | 89 | 72 | 70 | 87 | 72 | 83 | 61 | 59 | 82 |
| Male | 5,768 | 80 | 90 | 72 | 70 | 86 | 72 | 85 | 61 | 60 | 79 |
| American Indian/Alaska Native | 1,737 | 85 | 93 | 71 | 70 | 81 | 79 | 91 | 60 | 59 | 65 |
| Asian | 175 | 81 | 88 | 74 | 71 | 86 | 74 | 85 | 63 | 58 | 82 |
| African American | 313 | 82 | 91 | 73 | 72 | 84 | 74 | 86 | 60 | 63 | 75 |
| Hispanic/Latino | 657 | 79 | 90 | 72 | 69 | 81 | 72 | 84 | 62 | 58 | 74 |
| Native Hawaiian/Pacific Islander | 4** | | | | | | | | | | |
| White | 7,990 | 79 | 87 | 72 | 70 | 87 | 70 | 80 | 62 | 60 | 81 |
| Multiple | 521 | 79 | 88 | 72 | 70 | 86 | 72 | 84 | 59 | 62 | 80 |
| LEP | 605 | 82 | 90 | 72 | 70 | 84 | 75 | 86 | 61 | 60 | 74 |
| IDEA | 1,737 | 83 | 92 | 72 | 69 | 85 | 76 | 89 | 62 | 57 | 75 |
| Section 504 | 247 | 80 | 90 | 73 | 71 | 84 | 72 | 85 | 62 | 61 | 76 |
| **Grade 4** | | | | | | | | | | | |
| All | 11,390 | 78 | 90 | 65 | 68 | 86 | 70 | 85 | 53 | 58 | 79 |
| Female | 5,561 | 78 | 90 | 65 | 68 | 87 | 70 | 83 | 53 | 58 | 80 |
| Male | 5,829 | 78 | 91 | 64 | 68 | 85 | 71 | 86 | 53 | 57 | 78 |
| American Indian/Alaska Native | 1,764 | 85 | 93 | 64 | 67 | 83 | 79 | 92 | 52 | 56 | 67 |
| Asian | 175 | 79 | 91 | 65 | 67 | 89 | 72 | 83 | 54 | 55 | 85 |
| African American | 320 | 79 | 90 | 64 | 69 | 83 | 71 | 86 | 52 | 60 | 66 |
| Hispanic/Latino | 665 | 78 | 90 | 65 | 67 | 81 | 70 | 85 | 54 | 57 | 73 |
| Native Hawaiian/Pacific Islander | 12 | 68 | 89* | 60* | 62* | 0* | 60 | 78* | 49* | 60* | 27* |
| White | 7,990 | 77 | 88 | 65 | 68 | 86 | 68 | 80 | 53 | 58 | 80 |
| Multiple | 463 | 78 | 89 | 65 | 68 | 86 | 69 | 83 | 55 | 56 | 78 |
| LEP | 305 | 86 | 93 | 65 | 70 | 70 | 81 | 92 | 50 | 58 | 58 |
| IDEA | 1,669 | 84 | 93 | 64 | 68 | 85 | 78 | 91 | 53 | 55 | 76 |
| Section 504 | 229 | 79 | 92 | 61 | 68 | 87 | 71 | 86 | 50 | 59 | 78 |
| **Grade 5** | | | | | | | | | | | |
| All | 11,049 | 80 | 91 | 68 | 76 | 84 | 72 | 85 | 57 | 68 | 76 |
| Female | 5,397 | 79 | 90 | 68 | 76 | 85 | 71 | 84 | 57 | 68 | 77 |
| Male | 5,652 | 80 | 91 | 68 | 77 | 83 | 73 | 86 | 57 | 68 | 74 |
| American Indian/Alaska Native | 1,676 | 86 | 94 | 68 | 76 | 81 | 81 | 92 | 56 | 66 | 63 |
| Asian | 162 | 80 | 91 | 67 | 74 | 87 | 73 | 85 | 59 | 62 | 83 |
| African American | 311 | 80 | 90 | 67 | 77 | 75 | 73 | 87 | 56 | 68 | 67 |
| Hispanic/Latino | 559 | 81 | 91 | 68 | 75 | 84 | 73 | 86 | 59 | 66 | 73 |
| Native Hawaiian/Pacific Islander | 10 | 81 | 90* | 66* | 82* | 83* | 71 | 80* | 53* | 61* | 83* |
| White | 7,910 | 78 | 88 | 68 | 76 | 84 | 70 | 80 | 56 | 68 | 77 |
| Multiple | 421 | 80 | 89 | 69 | 76 | 85 | 72 | 84 | 59 | 67 | 74 |
| LEP | 198 | 91 | 97 | 67 | 75 | 92* | 88 | 95 | 57 | 61 | 78* |
| IDEA | 1,517 | 86 | 93 | 67 | 75 | 86 | 81 | 91 | 55 | 65 | 69 |
| Section 504 | 266 | 80 | 89 | 66 | 77 | 89 | 72 | 82 | 57 | 69 | 82 |

* The classification index is based on n<10.
** Suppressed data due to the small overall sample size, n<10.

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

| Group | N | %Accuracy | | | | | %Consistency | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | All | L1 | L2 | L3 | L4 | All | L1 | L2 | L3 | L4 |
| **Grade 6** | | | | | | | | | | | |
| All | 10,902 | 80 | 89 | 74 | 77 | 83 | 72 | 82 | 64 | 70 | 73 |
| Female | 5,296 | 80 | 88 | 74 | 77 | 84 | 71 | 81 | 64 | 70 | 74 |
| Male | 5,606 | 80 | 90 | 73 | 77 | 83 | 72 | 84 | 64 | 70 | 71 |
| American Indian/Alaska Native | 1,676 | 85 | 93 | 73 | 77 | 85 | 79 | 90 | 64 | 67 | 66 |
| Asian | 160 | 81 | 94 | 73 | 74 | 86 | 74 | 87 | 62 | 68 | 79 |
| African American | 291 | 81 | 90 | 73 | 77 | 83 | 73 | 84 | 63 | 68 | 65 |
| Hispanic/Latino | 603 | 80 | 89 | 74 | 77 | 85 | 72 | 82 | 66 | 69 | 70 |
| Native Hawaiian/Pacific Islander | 10 | 74 | 70* | 72* | 82* | 0* | 64 | 63* | 61* | 74* | 27* |
| White | 7,805 | 79 | 86 | 74 | 77 | 83 | 70 | 77 | 64 | 71 | 73 |
| Multiple | 355 | 80 | 89 | 75 | 77 | 83 | 71 | 80 | 66 | 67 | 77 |
| LEP | 206 | 90 | 95 | 73 | 80* | 80* | 85 | 93 | 65 | 49* | 77* |
| IDEA | 1,417 | 85 | 92 | 74 | 74 | 86 | 79 | 89 | 64 | 62 | 67 |
| Section 504 | 265 | 79 | 86 | 74 | 76 | 84 | 71 | 78 | 65 | 67 | 77 |
| **Grade 7** | | | | | | | | | | | |
| All | 10,565 | 80 | 89 | 72 | 80 | 82 | 72 | 82 | 61 | 74 | 72 |
| Female | 5,133 | 80 | 88 | 71 | 80 | 82 | 72 | 81 | 60 | 74 | 72 |
| Male | 5,432 | 80 | 89 | 72 | 80 | 82 | 73 | 83 | 61 | 73 | 71 |
| American Indian/Alaska Native | 1,685 | 84 | 92 | 71 | 77 | 80 | 77 | 88 | 61 | 69 | 60 |
| Asian | 181 | 79 | 88 | 69 | 78 | 79 | 71 | 85 | 57 | 71 | 71 |
| African American | 289 | 81 | 89 | 73 | 80 | 78 | 73 | 83 | 64 | 69 | 67 |
| Hispanic/Latino | 469 | 80 | 89 | 72 | 80 | 77 | 72 | 82 | 62 | 74 | 61 |
| Native Hawaiian/Pacific Islander | 12 | 75 | 72* | 68* | 78* | 99* | 68 | 72* | 58* | 69* | 93* |
| White | 7,595 | 79 | 87 | 72 | 80 | 82 | 71 | 78 | 61 | 74 | 72 |
| Multiple | 334 | 80 | 88 | 73 | 79 | 77 | 71 | 83 | 62 | 72 | 68 |
| LEP | 203 | 87 | 95 | 69 | 76 | 90* | 83 | 91 | 61 | 62 | 67* |
| IDEA | 1,285 | 85 | 92 | 71 | 78 | 83 | 79 | 89 | 61 | 66 | 64 |
| Section 504 | 262 | 81 | 88 | 73 | 83 | 81 | 72 | 80 | 63 | 74 | 74 |
| **Grade 8** | | | | | | | | | | | |
| All | 10,165 | 80 | 89 | 74 | 80 | 81 | 73 | 82 | 64 | 73 | 71 |
| Female | 5,004 | 80 | 87 | 74 | 80 | 81 | 72 | 79 | 64 | 73 | 72 |
| Male | 5,161 | 81 | 90 | 74 | 80 | 80 | 73 | 83 | 64 | 73 | 70 |
| American Indian/Alaska Native | 1,594 | 85 | 92 | 73 | 79 | 73 | 78 | 89 | 64 | 70 | 53 |
| Asian | 171 | 83 | 91 | 75 | 78 | 87 | 76 | 87 | 60 | 73 | 81 |
| African American | 277 | 81 | 90 | 73 | 77 | 77 | 73 | 85 | 63 | 69 | 63 |
| Hispanic/Latino | 466 | 80 | 89 | 73 | 79 | 78 | 73 | 83 | 65 | 72 | 64 |
| Native Hawaiian/Pacific Islander | 8** | | | | | | | | | | |
| White | 7,396 | 79 | 86 | 74 | 80 | 81 | 71 | 77 | 65 | 73 | 71 |
| Multiple | 252 | 78 | 85 | 73 | 79 | 74 | 70 | 77 | 64 | 70 | 71 |
| LEP | 252 | 88 | 93 | 75 | 78 | 0* | 83 | 91 | 63 | 59 | 5* |
| IDEA | 1,091 | 86 | 92 | 74 | 78 | 77 | 80 | 90 | 64 | 64 | 67 |
| Section 504 | 234 | 79 | 86 | 74 | 78 | 85 | 71 | 76 | 66 | 72 | 71 |

*The classification index is based on n<10.
** Suppressed data due to the small overall sample size, n<10.

Table C–3. ELA/L Classification Accuracy and Consistency by Achievement Levels (Grade 11)

| Group | N | %Accuracy | | | | | %Consistency | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | All | L1 | L2 | L3 | L4 | All | L1 | L2 | L3 | L4 |
| **Grade 11** | | | | | | | | | | | |
| All | 9,032 | 81 | 88 | 75 | 78 | 86 | 73 | 81 | 64 | 71 | 79 |
| Female | 4,345 | 81 | 87 | 75 | 78 | 86 | 73 | 79 | 63 | 71 | 80 |
| Male | 4,687 | 80 | 89 | 74 | 78 | 85 | 73 | 82 | 64 | 71 | 77 |
| American Indian/Alaska Native | 1,043 | 82 | 90 | 74 | 78 | 82 | 74 | 85 | 64 | 69 | 70 |
| Asian | 177 | 85 | 93 | 76 | 79 | 89 | 78 | 89 | 63 | 72 | 82 |
| African American | 217 | 83 | 92 | 77 | 75 | 85 | 76 | 88 | 67 | 67 | 76 |
| Hispanic/Latino | 388 | 80 | 89 | 75 | 77 | 84 | 72 | 83 | 62 | 72 | 73 |
| Native Hawaiian/Pacific Islander | 15 | 77 | 56* | 79* | 80* | 77* | 69 | 43* | 63* | 67* | 79* |
| White | 7,001 | 80 | 87 | 75 | 78 | 86 | 72 | 77 | 63 | 71 | 79 |
| Multiple | 190 | 81 | 86 | 75 | 79 | 84 | 73 | 79 | 65 | 72 | 78 |
| LEP | 215 | 90 | 94 | 73 | 70* | 63* | 86 | 94 | 57 | 56* | 47* |
| IDEA | 718 | 83 | 91 | 73 | 73 | 86 | 77 | 88 | 63 | 64 | 67 |
| Section 504 | 206 | 80 | 88 | 74 | 78 | 84 | 72 | 77 | 63 | 72 | 77 |

*The classification index is based on n<10.
** Suppressed data due to the small overall sample size, n<10.

Table C–4. 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 | 11,424 | 82 | 89 | 73 | 79 | 88 | 75 | 83 | 64 | 71 | 82 |
| Female | 5,646 | 82 | 89 | 74 | 79 | 88 | 74 | 83 | 64 | 72 | 82 |
| Male | 5,778 | 82 | 90 | 73 | 79 | 88 | 75 | 83 | 64 | 71 | 83 |
| American Indian/Alaska Native | 1,735 | 85 | 92 | 73 | 77 | 85 | 79 | 89 | 62 | 68 | 74 |
| Asian | 181 | 83 | 91 | 72 | 79 | 90 | 77 | 89 | 61 | 73 | 84 |
| African American | 324 | 82 | 91 | 73 | 79 | 86 | 75 | 85 | 65 | 68 | 82 |
| Hispanic/Latino | 668 | 82 | 89 | 74 | 79 | 83 | 74 | 85 | 64 | 71 | 75 |
| Native Hawaiian/Pacific Islander | 4** | | | | | | | | | | |
| White | 7,990 | 81 | 87 | 74 | 79 | 88 | 74 | 77 | 64 | 72 | 83 |
| Multiple | 521 | 81 | 90 | 73 | 78 | 88 | 74 | 83 | 64 | 69 | 82 |
| LEP | 632 | 83 | 91 | 73 | 78 | 82 | 76 | 88 | 63 | 68 | 74 |
| IDEA | 1,740 | 84 | 92 | 73 | 79 | 87 | 78 | 89 | 63 | 70 | 79 |
| Section 504 | 247 | 81 | 89 | 75 | 77 | 89 | 74 | 82 | 65 | 70 | 83 |
| **Grade 4** | | | | | | | | | | | |
| All | 11,416 | 83 | 90 | 80 | 79 | 88 | 76 | 83 | 73 | 71 | 82 |
| Female | 5,582 | 83 | 89 | 80 | 78 | 87 | 76 | 83 | 73 | 71 | 80 |
| Male | 5,834 | 84 | 90 | 81 | 79 | 88 | 77 | 84 | 73 | 72 | 83 |
| American Indian/Alaska Native | 1,763 | 87 | 92 | 80 | 78 | 87 | 81 | 89 | 71 | 69 | 76 |
| Asian | 182 | 86 | 93 | 82 | 80 | 92 | 80 | 87 | 76 | 71 | 87 |
| African American | 335 | 85 | 91 | 82 | 77 | 84 | 78 | 86 | 75 | 68 | 76 |
| Hispanic/Latino | 670 | 83 | 90 | 80 | 78 | 85 | 76 | 84 | 74 | 69 | 78 |
| Native Hawaiian/Pacific Islander | 13 | 84 | 95* | 83* | 73* | 100* | 75 | 78* | 78* | 62* | 83* |
| White | 7,988 | 82 | 87 | 81 | 79 | 88 | 75 | 77 | 73 | 71 | 82 |
| Multiple | 464 | 82 | 85 | 80 | 80 | 88 | 75 | 80 | 73 | 71 | 81 |
| LEP | 336 | 87 | 93 | 80 | 71 | 87* | 81 | 90 | 70 | 63 | 68* |
| IDEA | 1,663 | 86 | 92 | 80 | 79 | 86 | 80 | 89 | 73 | 68 | 80 |
| Section 504 | 230 | 86 | 90 | 83 | 82 | 90 | 79 | 86 | 76 | 74 | 86 |
| **Grade 5** | | | | | | | | | | | |
| All | 11,077 | 82 | 90 | 77 | 71 | 87 | 74 | 85 | 69 | 62 | 81 |
| Female | 5,406 | 81 | 90 | 78 | 72 | 87 | 74 | 84 | 69 | 62 | 80 |
| Male | 5,671 | 82 | 91 | 77 | 71 | 88 | 75 | 86 | 68 | 62 | 82 |
| American Indian/Alaska Native | 1,676 | 88 | 93 | 76 | 70 | 86 | 83 | 91 | 65 | 58 | 75 |
| Asian | 170 | 84 | 92 | 78 | 70 | 90 | 77 | 88 | 68 | 63 | 81 |
| African American | 319 | 83 | 92 | 75 | 71 | 84 | 77 | 88 | 67 | 61 | 71 |
| Hispanic/Latino | 575 | 84 | 93 | 78 | 74 | 83 | 78 | 89 | 71 | 62 | 77 |
| Native Hawaiian/Pacific Islander | 10 | 80 | 77* | 75* | 78* | 96* | 72 | 69* | 70* | 65* | 85* |
| White | 7,907 | 80 | 87 | 78 | 71 | 88 | 72 | 79 | 70 | 62 | 82 |
| Multiple | 420 | 80 | 87 | 77 | 73 | 85 | 73 | 82 | 69 | 63 | 75 |
| LEP | 231 | 92 | 96 | 75 | 62* | 67* | 89 | 95 | 63 | 42* | 67* |
| IDEA | 1,513 | 88 | 94 | 77 | 71 | 86 | 82 | 91 | 68 | 58 | 77 |
| Section 504 | 267 | 82 | 92 | 75 | 73 | 91 | 75 | 84 | 69 | 61 | 84 |

*The classification index is based on n<10.
** Suppressed data due to the small overall sample size, n<10.

Table C–5. 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 | 10,930 | 82 | 91 | 78 | 72 | 87 | 74 | 85 | 70 | 62 | 79 |
| Female | 5,305 | 81 | 91 | 77 | 72 | 87 | 73 | 84 | 70 | 63 | 79 |
| Male | 5,625 | 82 | 92 | 78 | 73 | 87 | 75 | 86 | 70 | 62 | 80 |
| American Indian/Alaska Native | 1,674 | 88 | 95 | 78 | 72 | 80 | 83 | 92 | 70 | 59 | 67 |
| Asian | 165 | 85 | 90 | 81 | 75 | 95 | 79 | 87 | 72 | 68 | 88 |
| African American | 303 | 84 | 92 | 75 | 69 | 87 | 78 | 90 | 67 | 59 | 67 |
| Hispanic/Latino | 617 | 83 | 90 | 79 | 72 | 82 | 76 | 87 | 71 | 59 | 74 |
| Native Hawaiian/Pacific Islander | 10 | 77 | 69* | 72* | 83* | 100* | 68 | 65* | 67* | 63* | 83* |
| White | 7,805 | 80 | 89 | 77 | 72 | 87 | 72 | 80 | 70 | 63 | 80 |
| Multiple | 354 | 81 | 88 | 78 | 73 | 86 | 73 | 80 | 71 | 62 | 79 |
| LEP | 237 | 92 | 95 | 79 | 78* | 91* | 89 | 95 | 67 | 59* | 90* |
| IDEA | 1,412 | 88 | 94 | 76 | 73 | 87 | 83 | 92 | 69 | 59 | 72 |
| Section 504 | 266 | 84 | 92 | 77 | 73 | 93 | 77 | 85 | 71 | 61 | 89 |
| **Grade 7** | | | | | | | | | | | |
| All | 10,588 | 82 | 90 | 77 | 75 | 89 | 75 | 84 | 68 | 66 | 82 |
| Female | 5,147 | 82 | 90 | 77 | 75 | 88 | 74 | 83 | 68 | 66 | 81 |
| Male | 5,441 | 82 | 91 | 77 | 75 | 89 | 75 | 85 | 68 | 66 | 83 |
| American Indian/Alaska Native | 1,691 | 87 | 93 | 75 | 75 | 85 | 81 | 90 | 66 | 63 | 73 |
| Asian | 182 | 81 | 89 | 73 | 72 | 89 | 75 | 83 | 65 | 61 | 85 |
| African American | 301 | 84 | 90 | 76 | 75 | 85 | 77 | 88 | 66 | 65 | 78 |
| Hispanic/Latino | 476 | 83 | 90 | 77 | 76 | 88 | 76 | 87 | 69 | 65 | 77 |
| Native Hawaiian/Pacific Islander | 12 | 81 | 98* | 77* | 70* | 100* | 73 | 84* | 73* | 59* | 84* |
| White | 7,592 | 81 | 88 | 77 | 75 | 89 | 73 | 79 | 69 | 66 | 82 |
| Multiple | 334 | 82 | 89 | 78 | 72 | 88 | 74 | 85 | 68 | 64 | 81 |
| LEP | 229 | 90 | 95 | 74 | 73 | 81* | 86 | 93 | 64 | 60 | 63* |
| IDEA | 1,292 | 89 | 95 | 76 | 72 | 86 | 84 | 92 | 67 | 58 | 80 |
| Section 504 | 260 | 81 | 85 | 77 | 75 | 86 | 73 | 79 | 68 | 64 | 82 |
| **Grade 8** | | | | | | | | | | | |
| All | 10,177 | 81 | 90 | 72 | 71 | 89 | 74 | 84 | 63 | 61 | 83 |
| Female | 5,007 | 80 | 89 | 73 | 71 | 88 | 73 | 83 | 64 | 61 | 82 |
| Male | 5,170 | 82 | 91 | 72 | 71 | 89 | 75 | 86 | 63 | 61 | 84 |
| American Indian/Alaska Native | 1,574 | 88 | 94 | 71 | 72 | 83 | 83 | 92 | 61 | 56 | 76 |
| Asian | 177 | 84 | 93 | 70 | 67 | 95 | 79 | 89 | 65 | 54 | 89 |
| African American | 291 | 82 | 92 | 70 | 70 | 88 | 76 | 87 | 61 | 60 | 79 |
| Hispanic/Latino | 480 | 81 | 90 | 71 | 73 | 83 | 74 | 86 | 62 | 61 | 79 |
| Native Hawaiian/Pacific Islander | 8** | | | | | | | | | | |
| White | 7,393 | 79 | 87 | 73 | 71 | 89 | 71 | 79 | 64 | 61 | 83 |
| Multiple | 253 | 81 | 91 | 72 | 73 | 86 | 74 | 83 | 63 | 65 | 80 |
| LEP | 287 | 91 | 96 | 70 | 68 | 98* | 88 | 95 | 60 | 53 | 70* |
| IDEA | 1,092 | 89 | 95 | 72 | 70 | 84 | 84 | 93 | 61 | 55 | 74 |
| Section 504 | 234 | 80 | 86 | 73 | 72 | 90 | 72 | 81 | 62 | 62 | 83 |

*The classification index is based on n<10.
** Suppressed data due to the small overall sample size, n<10.

Table C–6. Mathematics Classification Accuracy and Consistency by Achievement Levels (Grade 11)

| Group | N | %Accuracy | | | | | %Consistency | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | All | L1 | L2 | L3 | L4 | All | L1 | L2 | L3 | L4 |
| **Grade 11** | | | | | | | | | | | |
| All | 9,026 | 82 | 91 | 74 | 80 | 86 | 75 | 86 | 65 | 72 | 79 |
| Female | 4,340 | 82 | 91 | 74 | 80 | 84 | 75 | 85 | 65 | 72 | 77 |
| Male | 4,686 | 83 | 91 | 74 | 79 | 88 | 76 | 86 | 65 | 71 | 81 |
| American Indian/Alaska Native | 1,036 | 89 | 94 | 72 | 79 | 88 | 85 | 92 | 62 | 69 | 76 |
| Asian | 179 | 86 | 95 | 71 | 79 | 94 | 81 | 91 | 64 | 72 | 85 |
| African American | 216 | 88 | 94 | 75 | 80 | 79* | 83 | 92 | 65 | 73 | 67* |
| Hispanic/Latino | 387 | 84 | 93 | 75 | 78 | 86 | 77 | 87 | 68 | 69 | 75 |
| Native Hawaiian/Pacific Islander | 15 | 82 | 92* | 68* | 80* | 85* | 75 | 82* | 60* | 73* | 83 |
| White | 7,001 | 81 | 89 | 74 | 80 | 86 | 74 | 82 | 65 | 72 | 79 |
| Multiple | 191 | 81 | 89 | 74 | 80 | 81 | 73 | 83 | 67 | 68 | 74 |
| LEP | 213 | 96 | 98 | 72 | 70* | 61* | 93 | 97 | 61 | 55* | 60* |
| IDEA | 723 | 93 | 97 | 73 | 74 | 86* | 89 | 96 | 61 | 61 | 68* |
| Section 504 | 207 | 82 | 92 | 75 | 79 | 90 | 75 | 83 | 69 | 70 | 79 |

*The classification index is based on n<10.
** Suppressed data due to the small overall sample size, n<10.