Connecticut Smarter Balanced Assessments

2022–2023 Technical Report

Submitted to
Connecticut State Department of Education
by Cambium Assessment, Inc.

TABLE OF CONTENTS

List of Tables

List of Figures

List of Exhibits

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. Connecticut was 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, as well as optional interim assessments that provide meaningful feedback and actionable data that teachers and educators can use to help students succeed. SBAC, a state-led enterprise, is intended to provide leadership and resources to improve teaching and learning by creating and maintaining a suite of summative and interim assessments and tools aligned to the CCSS in ELA/L and mathematics.

The Connecticut State Board of Education formally adopted the CCSS in ELA/L and mathematics on July 7, 2010. All students in Connecticut, including students with significant cognitive disabilities who are eligible to take the Connecticut Alternate Assessment (CTAA), an alternate assessment based on alternate academic achievement standards (AA-AAAS), are taught content that aligns to the same academic standards. Connecticut CCSS define the knowledge and skills students need to succeed in college and careers after graduating from high school. These standards include rigorous content and application of knowledge through higher-order skills and align with college and workforce expectations.

The Connecticut statewide assessments in ELA/L and mathematics aligned with the CCSS were administered for the first time in spring 2015 to students in grades 3–8 and 11 in all public elementary and secondary schools. In 2015–2016, Connecticut adopted the Scholastic Aptitude Test (SAT) to replace the Smarter Balanced grade 11 assessments for high school students.

The Smarter Balanced assessments are composed 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 (CAT).** The CAT is an online adaptive test that provides an individualized online assessment for each student.
- **Performance Task (PT).** A PT is a task that challenges students to apply their knowledge and skills to respond to real-world problems. PTs can best be described as collections of items 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, none of which can be adequately assessed with selected-response or constructed-response items. Some PT items can be scored by the computer, but most are handscored.

Optional interim assessments allow teachers to monitor student progress throughout the year and provide information that teachers can use to improve their instruction and learning. These tools are used at the discretion of schools and districts, and teachers can employ them to gauge students' progress in mastering specific concepts at strategic points during the school year.

There are three types of interim assessments available as fixed-form tests:

- **Interim Comprehensive Assessments (ICAs).** ICAs test the same content and report scores on the same scale as the summative assessments.
- **Interim Assessment Blocks (IABs).** IABs focus on smaller sets of related concepts and provide more detailed information about student learning.
- The **Focused Interim Assessment Block (FIAB)** focuses on specific sets of related concepts that measure no more than three assessment targets and provide more detailed information about student learning than IAB.

Starting in the 2015–2016 summative test administration, Connecticut made four changes in the summative tests:

- Replaced the summative ELA/L and mathematics assessments in grade 11 with the SAT reading and writing, language, and mathematics tests.
- Removed the summative field-test items and off-grade items from the ELA/L and mathematics CAT item pool.
- Removed PTs in ELA/L while keeping PTs in mathematics assessment. For the paper-pencil tests, the test booklet will include both non-PT and PT components, but only the non-PT component will be scored for ELA/L.
- Reported scores for combining claim 2 (writing) and claim 4 (research/inquiry) in ELA/L.

Due to the COVID-19 pandemic, the U.S. Department of Education waived testing requirements in the 2019–2020 school year (https://www2.ed.gov/policy/gen/guid/secletter/200320.html). For the 2021–2022 school year, the U.S. Department of Education did not grant waivers for standardized testing but did waive certain accountability requirements (e.g., mandatory high participation rates) due to the impacts of the pandemic in many states, resulting in lower participation rates than in previous years. In the 2020–2021 school year, all public schools in Connecticut participated in the Smarter Balanced summative assessments in grades 3–8, with

the participation rates of 90.7–95.2% in ELA/L and 88.1–94.4% in mathematics. In the 2020–2021 test administration, the Connecticut State Department of Education (CSDE) allowed remote testing in addition to in-person testing. The percentage of the students who took the summative tests remotely ranged from 9.8% to 14.4% in ELA/L and from 9.7% to 13.8% in mathematics in grades 3–8.

In the 2022–2023 school year, the participation rates increased, ranging from 96.9% to 97.9% in ELA/L and from 96.2% to 97.7% in mathematics, and remote testing was not allowed for the summative tests.

This report provides a technical summary of the 2022–2023 summative assessments in ELA/L and mathematics administered in grades 3–8 under the Connecticut Smarter Balanced assessments. The report is divided into eight chapters: Overview; Test Administration; Summary of the 2022–2023 Operational Test Administration; Validity; Reliability; Scoring; Reporting and Interpreting Scores; and Quality Control Procedures. The data included in this report are based on Connecticut data for the summative assessment only. For the interim assessments, the number of students who took ICAs and IABs and a summary of their performance are provided in Appendix A.

While this report includes information on all aspects of the technical quality of the Smarter Balanced test administration for Connecticut, it is an addendum to the 2022–2023 Smarter Balanced technical report. The Smarter Balanced technical report contains 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.

Smarter Balanced produces a technical report on the Smarter Balanced assessments that covers all aspects of their compliance with the technical qualities described in the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014) and the requirements of the U.S. Department of Education, outlined in *Peer Review of State Assessment Systems: Non-Regulatory Guidance for States* (U.S. Department of Education, 2015). The Smarter Balanced technical report includes analysis of the data at the consortium level, combining data from the consortium members.

2. TEST ADMINISTRATION

2.1 TESTING WINDOWS

The 2022–2023 Smarter Balanced assessments testing window spanned approximately two and a half months for the summative assessments and eight months for the interim assessments. The paper-pencil fixed-form tests for summative assessments were administered concurrently during the online summative window. Table 1 shows the testing windows for the online, remote, and paper-pencil assessments.

Table 1. 2022–2023 Testing Windows

Tests	Grades	Start Date	End Date	Mode
Summative Assessments	3–8	March 27, 2023	June 2, 2023	Online Adaptive
	3–8	March 27, 2023	June 2, 2023	Paper-Pencil Fixed Forms
Interim Comprehensive Assessments	3–8, 11	August 31, 2022	June 9, 2023	Online Fixed Forms
Interim Assessment Blocks	3–8, 11	August 31, 2022	June 9, 2023	Online Fixed Forms

2.2 TEST OPTIONS AND ADMINISTRATIVE ROLES

The 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, several assessment options were available for the 2022–2023 administration to accommodate students' needs. Table 2 lists the testing options that were offered in 2022–2023. A testing option is selected by content area. Once a testing option is selected, it applies to all tests in the content area.

Table 2. 2022–2023 Testing Options

Assessments	Test Options	Test Mode
Summative Assessments	English	Online
	Braille	Online
	Braille HAT (Hybrid Adaptive Test) (mathematics only)	Online
	Spanish (mathematics only)	Online
	Paper-Pencil, Large-Print, Fixed-Form Test*	Paper-Pencil
	Paper-Pencil, Braille, Fixed-Form Test*	Paper-Pencil
Interim Assessments	English	Online
	Braille	Online
	Spanish (mathematics only)	Online

* For the paper-pencil fixed-form tests, all student responses on the paper-pencil tests were entered in the Data Entry Interface (DEI) by test administrators.

To ensure standardized administration conditions, teachers (TEs) and test administrators (TAs) follow procedures outlined in the *Smarter Balanced ELA/L and Mathematics Online, Summative*

Test Administration Manual (TAM). TEs and TAs must review the TAM prior to the beginning of testing to ensure that the testing room is prepared appropriately (e.g., removing certain classroom posters, arranging desks). Make-up procedures should be established for any students who are absent on testing days. TEs and TAs follow required administration procedures and directions and read the boxed directions verbatim to students, ensuring standardized administration conditions.

2.2.1 Administrative Roles

The key personnel involved with the test administration for the Connecticut State Department of Education (CSDE) are District Administrators (DAs), District Coordinators (DCs), School Coordinators (SCs), teachers (TEs), and test administrators (TAs). Their main responsibilities are described in the following subsections. More detailed descriptions can be found in the TAM provided online at this URL: https://ct.portal.cambiumast.com/resources/.

District Administrator

The DA may add users with DC roles in the Test Information Distribution Engine (TIDE). For example, a director of special education may need DC privileges in TIDE to access district-level data for the purposes of verifying test settings for designated supports and accommodations. DAs have the same test administration responsibilities as DCs. Their primary responsibility is to coordinate the administration of the Smarter Balanced assessment in the district.

District Coordinator

The DC is primarily responsible for coordinating the administration of the Smarter Balanced assessments at the district level.

DCs are responsible for the following:

- Reviewing all Smarter Balanced policies and test administration documents
- Reviewing scheduling and testing requirements with SCs, TEs, and TAs
- Working with SCs and technology coordinators (TCs) to ensure that all systems, including the CAI Secure Browser, are properly installed and functional
- Importing users (including SCs, TEs, and TAs) into TIDE
- Verifying all student information and eligibility in TIDE
- Scheduling and administering training sessions for all SCs, TEs, TAs, and TCs
- Ensuring that all personnel are trained on how to administer the Smarter Balanced assessments properly

- Monitoring the secure administration of the tests
- Investigating and reporting all testing improprieties, irregularities, and breaches reported by the TEs and TAs
- Attending to any secure material according to CSDE and Smarter Balanced policies

School Coordinator

The SC is primarily responsible for coordinating the administration of the Smarter Balanced assessments at the school level and ensuring that testing within his or her school is conducted in accordance with the test procedures and security policies established by the CSDE.

SC responsibilities include the following:

- Based on testing windows, establishing a testing schedule with DCs, TEs, and TAs
- Working with technology staff to ensure timely computer setup and installation
- 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
- Identifying students who may require designated supports and test accommodations, and ensuring that procedures for testing these students follow CSDE and Smarter Balanced policies
- Attending all district trainings and reviewing all Smarter Balanced policies and test administration documents
- Ensuring that all TEs and TAs attend school or district trainings and review online training modules posted on the Connecticut state portal
- Establishing secure and separate testing rooms if needed
- Downloading and planning the administration of the classroom activity with TEs and TAs
- Monitoring secure administration of the tests
- 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 CSDE and Smarter Balanced policies

Teacher

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

Test Administrator

A TA is primarily responsible for administering the 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 do not have access to student results.

TAs are responsible for the following:

- Completing Smarter Balanced test administration training
- Reviewing all Smarter Balanced policy and test administration documents before administering any Smarter Balanced assessments
- Viewing student information before testing to ensure that a student receives the proper test with the appropriate supports, and reporting any potential data errors to SCs and DCs, as appropriate
- Administering the Smarter Balanced assessments
- Reporting all potential test security incidents to the SCs and DCs in a manner consistent with Smarter Balanced, CSDE, and district policies

2.2.2 Online Test Administration

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

SCs oversee all aspects of testing at their schools and serve as the main point of contact, while TEs and TAs administer the online assessments 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 CAI's online TA Certification Course. Staff who complete this course receive a certificate of completion.

To start a test session, the TE or TA must first enter the TA Interface of the online testing system using his or her own computer. A session ID is generated when the test session is created. Students who are taking the assessment with the TE or TA must enter their State Student Identification Number (SSID), their first name, and the 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 features (see Section 2.6 for a list of accommodations). Students can begin testing only when the TA or TE confirms the settings. The TA or TE then reads the *Directions for Administration* in the *Online Smarter Balanced Test Administration Manual* aloud to the students and guides them through the login process.

Once an assessment has started, the student must answer all the test items presented on a page before proceeding to the next page. Skipping items is not permitted. For the online computer-adaptive test (CAT), students are allowed to scroll back to review and edit previously answered items, as long as these items are in the same test session and this session has not been paused for more than 20 minutes. Students may review and edit responses they have previously provided before submitting the assessment. During an active CAT session, if a student reviews and changes the response to a previously answered item, then all items that follow to which the student already responded remain the same. If a student changes the answers, no new items are assigned. For example, a student pauses for 10 minutes after completing Item 10. After the pause, the student goes back to Item 5 and changes the answer. If the response change in Item 5 changes the item score from wrong to right, the student's overall score will improve; however, there will be no change in Items 6–10.

There is no pause rule implemented for the performance tasks (PTs). The same rules that apply to the CAT for reviews and changes to responses also apply to PTs.

For the summative test, an assessment can be started in one component and completed in another. For the CAT, the assessment must be completed within 45 calendar days of the start date or the assessment opportunity will expire. For the PTs, the assessment must be completed within 20 calendar days of the start date.

During a test session, TEs or TAs may pause the test for a student or group of students to take a break. It is up to the TEs or TAs to determine an appropriate stopping point; however, to ensure the integrity of test scores or testing, the CAT cannot be paused for more than 30 minutes for English language arts/literacy (ELA/L) and mathematics. If that happens, the student must restart a new test session, which starts from where the student left off. The viewing and editing options of previous responses are no longer available.

The TAs or TEs must always remain in the room during a test session to monitor student testing. Once the test session ends, the TAs or TEs must ensure that each student has successfully logged out of the system. Then the TAs or TEs must collect and send for secure shredding any handouts or scratch paper that students used during the assessment.

2.2.3 Paper-Pencil Test Administration

The paper-pencil versions of the Smarter Balanced ELA/L and mathematics assessments are provided as an accommodation for students who do not have access to a computer and students who are visually impaired. For Connecticut, paper-pencil tests were offered only in braille and large print.

The DA must order the accommodated test materials on behalf of the students who need to take the paper-pencil test via TIDE. Based on the paper-pencil orders submitted in TIDE, the testing contractor ships the appropriate test booklets and the *Paper-Pencil Test Administration Manual* to the district.

Separate test booklets are used for 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 PTs in both content areas. The TEs and TAs are asked not to administer the ELA PT on the paper-pencil test.

After the student has completed the assessments, the TEs and TAs enter the student responses into the Data Entry Interface (DEI) and return the test booklets to the testing vendor. The tests submitted via the DEI are then scored.

The total number of students who took paper-pencil tests is presented in Table 3. Please note that students who took the paper-pencil tests took the test in the classroom.

Table 3. Number of Students Who Took Paper-Pencil Tests in the 2022–2023 Summative Test Administration

Subject	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Total
ELA/L	1	6	5	2	8	7	29
Mathematics	1	4	5	2	6	6	24

2.2.4 Braille Test Administration

The adaptive braille test was available with the same test blueprint in English in both ELA/L and mathematics. In the 2017–2018 test administration, Smarter Balanced added the Braille Hybrid Adaptive Test (Braille HAT) for mathematics. The Braille HAT consists of a fixed-form segment, a CAT segment, and a fixed-form PT. The fixed-form segment includes items with tactile graphics

which can be embossed at the testing location or received as a package of pre-embossed materials through the CSDE. All items on the Braille HAT can be presented to the students using a Refreshable Braille Display (RBD).

The braille interface is described as follows:

- 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 the Nemeth Braille Code for Mathematics via a braille embosser through the online CAT and a fixed-form PT.
- Students taking the summative ELA/L assessment can emboss both reading passages and items as they progress through the assessment. If a student has an 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 or TAs must ensure that the technical requirements are met. These requirements apply to the student's computer, the TE's or 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, DCs, and SCs oversee all aspects of testing at their schools and serve as the main points of contact, and TEs and TAs administer the online assessments. The online CAI TA Certification Course, webinars, user guides, manuals, and training sites are used to train TEs and TAs about the online testing requirements and the mechanics of starting, pausing, and ending a test session. Training materials for test administration are provided online.

2.3.1 Online Training

Multiple online training opportunities are offered to key staff.

TA Certification Course

CAI's online TA Certification Course is available as an optional course to any user in TIDE. This web-based course is about 30–45 minutes long and covers information on testing policies and 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 are required to answer multiple-choice questions about the information provided.

Office Hour Webinars

During the testing window, the CSDE and CAI held office hours every Thursday from 3:00 p.m.–4:00 p.m. During office hours, the CSDE and CAI staff provided brief, weekly assessment updates and were available for phone support to answer any questions from districts. All office hour sessions were recorded, and the recordings were posted to the portal.

Practice and Training Test Site

In January 2015, separate practice and 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. 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 item types and levels of difficulty (approximately 30 items each in ELA/L and mathematics), as well as an opportunity to practice the PT.

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

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. The student training test includes all item types in the operational item pool, including multiple-choice items and grid items. Teachers can also use these training tests to help students become familiar with the online platform and item types.

Manuals and User Guides

The following manuals and user guides are available on the Connecticut portal: https://ct.portal.cambiumast.com/.

The *Test Coordinator Manual* provides information for DCs and SCs regarding policies and procedures for the 2023 Smarter Balanced assessments in ELA/L and mathematics.

The *Smarter Balanced Summative Assessment Test Administration Manual* provides information for TEs and TAs administering the Smarter Balanced online summative assessments in ELA/L and mathematics. It includes screen captures and step-by-step instructions on how to administer the online tests.

The *Assistive Technology Manual* provides an overview of the embedded and non-embedded assistive technology tools that can be used to help students with specific accessibility needs complete online tests in the Test Delivery System (TDS). It includes lists of supported devices and applications for each type of assistive technology that students may need, as well as setup instructions for the assistive technologies that require additional configuration in order to work with TDS.

The technology resource manuals contain technology requirements and instructions that will assist technology coordinators in preparing computers and devices for online testing. A guide is created for each of the approved operating systems (Windows, Mac, iPad, Linux, ChromeOS).

The *Centralized Reporting System User Guide* provides information about the reporting system, including instructions for viewing score reports, accessing test management resources, creating and editing rosters, and searching for students for interim and summative assessments.

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

The *Assessment Viewing Application User Guide* provides an overview of how to access and use the Assessment Viewing Application (AVA). AVA allows teachers to view items on the Smarter Balanced interim assessments.

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

All manuals and user guides pertaining to the 2022–2023 online testing are available on the portal, and DAs, DCs, and SCs used the manuals and user guides to train TAs and TEs in test administration policies and procedures.

Brochures and Quick Guides

The following brochures and quick guides are available on the Connecticut portal, https://ct.portal.cambiumast.com/.

Accessing Participation Reports: This brochure provides instructions for how to extract participation reports for the Smarter Balanced assessments.

Accessing TIDE: This brochure provides a brief overview of user management in TIDE and how to log in to the system. School personnel will need to use TIDE account credentials to access all secure online systems used to administer Connecticut Comprehensive Assessment Program online assessments.

Embedded and Non-Embedded Designated Supports for English Learners: This brochure provides recommendations for students who are English learners (ELs) on what supports they may benefit from when participating in the Connecticut statewide assessments. These designated supports are intended as a language support for students who have limited English language skills, whether or not they are identified in the Public School Information System (PSIS) as EL or EL with a disability. The use of these supports may result in the student needing additional overall time to complete the assessment.

How to Access the Data Entry Interface (DEI): This brochure describes how to access the DEI to submit the Smarter Balanced paper-pencil tests.

How to Activate a Test Session for the Interim Assessments: This document provides a step-by-step guide on how to start a test session for the Smarter Balanced interim assessments, including the interim assessment blocks (IABs). It includes a complete list of all interim test labels as they appear in the TA Interface.

Managing Student Test Settings Brochure: This brochure provides a brief overview on how to manage student test settings in TIDE. Students' embedded accommodations, non-embedded accommodations, and designated supports must be set in TIDE prior to test administration for these settings to be reflected in the TDS.

Monitoring Test Progress: Test Status Code Report and Test Completion Rates: This brochure contains instructions for generating Test Status Code Reports and Test Completion Rates in TIDE. These are excellent tools that should be used to track test completion for students at both the district and school level.

User Role Permissions for Online Systems Brochure: This brochure outlines the user roles and permissions for each secure online testing system used to administer the online assessments for the Connecticut Comprehensive Assessment Program. These systems include TIDE, the Centralized Reporting System (CRS), TA Interface, DEI, and AVA.

Understanding and Creating Rosters: This document provides instructions for how to create, view, and modify rosters in TIDE and in the CRS. Rosters are groups of students associated with a teacher in a particular school. Rosters typically represent entire classrooms in lower grades, or individual classroom periods in upper grades.

2.3.2 District Coordinator Training Workshops

DC training workshops were held on January 18 and 20, 2023. Five remote training sessions were held during this period. Training was provided for the administration of the Smarter Balanced assessments for ELA/L and mathematics. During the training, DCs were provided with information to support training of the SCs, TEs, and TAs.

2.4 Test Security

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

2.4.1 Student-Level Testing Confidentiality

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

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

1. *Test eligibility* refers to the assignment of a test to a particular student.
2. *Test accommodation* refers to the assignment of a test setting to specific students based on needs.
3. *Test session* refers to the authentication process of a 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 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 email message; if information must be sent via email or fax, include only the SSID number, not the student's name
- Having students log in and test under another student's SSID number

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

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

Students log in to the online assessment using their legal first name, SSID number, and a test session ID. Only students can log in to an online test session. 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 and TAs are required to affix the student label to the student's answer document.

After a test session, only staff with the administrative roles of DA, DC, SC, or TE 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 designated 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, all DAs, DCs, SCs, TAs, and TEs have defined roles and levels of access to the testing system. When the TIDE testing window opens, the CSDE provides a verified list of DAs to the testing contractor, who uploads the information into TIDE. DAs are then responsible for selecting and entering the DCs' and SCs' information into TIDE, and the SC is responsible for entering TA and TE information into TIDE. Throughout the year, the DA, DC, and SC are also expected to delete information in TIDE for any staff members who have transferred to other schools, resigned, or no longer serve as TAs or TEs.

Password Protection: All access points by different roles at the state, district, school principal, and school staff levels require a password to log in to the system. Newly added SCs, TAs, and TEs receive separate passwords through their personal email addresses assigned by the school.

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

2.4.3 Security of the Testing Environment

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

Testing personnel are reminded in the online training and user manuals that assessments should be administered in testing rooms that do not crowd students. Good lighting, ventilation, and freedom from noise and interruption are important factors to consider 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 and where students are expected to report once they leave without disrupting others. 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 during testing, the TAs or TEs 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 items answered before the pause. This measure is implemented to prevent students from using the time outside of the testing room to look up answers.

Room Preparation

The room should be prepared prior to the start of the test session. Any information displayed on bulletin boards, chalkboards, or charts that students might use to help answer test items should be removed or covered. This rule applies to rubrics, vocabulary charts, student work, posters, graphs, content-area strategies charts, and other materials. The cell phones of both testing

personnel and students must be turned off and stored in the testing room out of sight. 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 space 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 items as other students; however, through appropriate seating arrangements, students should be discouraged from communicating with each other. For the PTs, different forms are distributed throughout a classroom so that students receive different forms of the PTs.

After the Test

At the end of the 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 items for any content-area assessment provided for a student allowed to use this accommodation in an individual setting must also be shredded immediately after a test session ends.

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

2.4.4 Test Security Violations

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

> **Impropriety:** This is a test security incident that has a low impact on the individual or group of students who are testing and has a low risk of potentially affecting student performance on the test, test security, or test validity (e.g., students leaving the testing room without authorization).
>
> **Irregularity:** This is a test security incident that impacts an individual or group of students who are testing and may potentially affect student performance on the test, test security, or test validity. These circumstances can be contained at the local level (e.g., disruption during the test session, such as a fire drill).

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

District and school personnel are required to document all test security incidents in the test security incident log. The log serves as the document of record for all test security incidents and should be maintained at the district level and submitted to the CSDE at the end of testing.

2.5 Student Participation

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

2.5.1 Home-Schooled Students

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

2.5.2 Exempt Students

Students who have a significant medical emergency are exempt from participating in the Smarter Balanced assessments.

2.6 Online Testing Features and Testing Accommodations

The Smarter Balanced Assessment Consortium's *Usability, Accessibility, and Accommodations Guidelines* (UAA Guidelines) are intended for school-level personnel and decision-making teams, including Individualized Education Program (IEP) and Section 504 Plan teams, as they prepare for and implement the Smarter Balanced assessments. The UAA 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 UAA Guidelines are also intended for assessment staff and administrators who oversee the decisions that are made in instruction and assessment.

The *Connecticut Assessment 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. They focus on universal tools, designated supports, and accommodations for the Smarter Balanced assessments of ELA/L and mathematics. At the same time, the UAA 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, DCs, and SCs 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. One or more of the pre-selected universal tools can be deactivated by a TE/TA in the TA Interface of the testing system for a student who may be distracted by the ability to access a specific tool during a test session.

For additional information about the availability of designated supports and accommodations, refer to the Connecticut's Assessment Guidelines for complete information at this URL: https://ct.portal.cambiumast.com/resources/guides/csde-assessment-guidelines.

2.6.1 Online Universal Tools for All Students

Universal tools are 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 have been pre-set in TIDE. In the 2022–2023 test administration, the following features of universal tools were available for *all* students to access. For specific information on how to access and use these features, refer to the *Test Administrator User Guide* at this URL: https://ct.portal.cambiumast.com.

Embedded Universal Tools

Breaks: The student can pause and resume the assessment. However, if an assessment is paused for more than 20 minutes, students will not be allowed to return to previous test items.

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 indicate that it would be appropriate.

Digital Notepad: This tool is used for making notes about an item. The digital notepad is item-specific and 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 Glossary: Grade- and context-appropriate definitions of specific construct-irrelevant terms are shown in English on the screen via a pop-up window. The student can access the embedded glossary by clicking on any of the pre-selected terms.

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

Highlighter: This tool is used to highlight passages or sections of passages and test items.

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

Line Reader: Students can use the line reader tool to assist in reading by raising and lowering the tool for each line of text on the screen.

Mark for Review: Students can mark an item to return to later during testing. However, for the CAT, if the assessment is paused for more than 20 minutes, students will not be allowed to return to marked test items.

Mathematics Tools: These digital tools (e.g., embedded ruler, embedded protractor) are used for measurements related to mathematics items. They are available only with the specific items for which the Smarter Balanced item specifications indicate that one or more of these tools would be appropriate.

Strikethrough: This tool allows users to cross out response options. If the response option is an image, a strikethrough line will not appear, but the image will be grayed out.

Writing Tools (for interim ELA/L performance tasks): Selected writing tools (e.g., bold, italic, bullets, undo/redo) are available for all student-generated responses.

Zoom: Students can zoom in and zoom out on test items, text, or graphics.

Non-Embedded Universal Tools

Breaks: Breaks may be given at predetermined intervals or after completion of sections of the assessment for students taking a paper-pencil test. Sometimes, students are allowed to take breaks when individually needed in order 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.

Scratch Paper/White Board with Marker: 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 and acceptable to the CSDE.

2.6.2 Designated Supports and Accommodations

Designated supports for the Smarter Balanced assessments are 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 Assessment Consortium 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 can adjust screen background or font color, based on student needs or preferences. This may include reversing the colors for the entire interface or choosing the color of font and background. Black on white, reverse contrast, black on rose, medium gray on light gray, and yellow on blue were offered for the online assessments.

Illustration Glossary: The illustration glossaries are provided for selected construct-irrelevant terms for math. Illustrations for these terms appear on the computer screen when students select them. Students with the illustration glossary setting enabled can view the illustration glossary. Students can also adjust the size of the illustration and move it around the screen.

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.

Mouse Pointer: This embedded support allows the mouse pointer to be set to a larger size and/or for the color of the mouse pointer to be changed. A TA sets the size and color of the mouse pointer prior to testing.

Print Size Online: This tool allows the font size viewed by the student in the TDS to be pre-set for the entire test. This support is generally most beneficial for students with visual disabilities. Selections are entered in the TIDE system prior to testing.

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 (for mathematics stimuli items and ELA/L items): Text is read aloud to the student via embedded text-to-speech technology. The student can control the speed of the voice and raise or lower the volume of the voice via a volume control. This support is also available in Spanish on mathematics tests when both Spanish-stacked translations and text-to-speech for stimuli and items are selected in TIDE.

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

Translations (glossaries) (for mathematics): Translated glossaries are a language support 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, Burmese, Cantonese, Filipino, Hmong, Korean, Mandarin, Punjabi, Russian, Somali, Spanish, 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, that students do not need to use, or that students are unable to use.

Non-Embedded Designated Supports

Amplification: The student adjusts the volume control beyond the computer's built-in settings using headphones or other non-embedded devices.

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

Color Overlay: Color transparencies may be placed over a paper-pencil assessment.

Illustration Glossary: The illustration glossaries are a language support provided for selected construct-irrelevant terms for math. Illustrations for these terms appear in a supplement to the paper/pencil test and are identified by item number.

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

Native Language Reader of Teacher Script: All teacher test directions/script as published in the Smarter Balanced TAM may be read and clarified in English, or using the student's native language for English learners and multilingual learners. A non-certified or certified staff person trained in test administration and security may administer this support.

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

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

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

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

Translated Test Directions: The TA uses a PDF file of directions translated in each of the languages currently supported. A bilingual adult can read the 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. Graphic material (e.g., maps, charts, graphs, diagrams, illustrations) is presented in a raised format (paper or

thermoform). Contracted and non-contracted braille is available, and Nemeth Code is available for mathematics.

Braille Transcript: A braille transcript of the closed captioning is available for the listening passages of the ELA/L assessment in the following braille codes: English Braille, American Edition (EBAE) uncontracted; and EBAE contracted.

Closed Captioning (for ELA/L listening stimuli items): This is printed text that appears on the computer screen as audio materials are presented.

Speech-to-Text: This tool allows students to dictate their responses into an open text box.

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 of the voice and raise or lower the volume of the voice via a volume control.

Non-Embedded Accommodations

100s Number Table: This is a paper-based list of all the digits from 1 through 100 in table format.

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

Alternate Response Options: Alternate response options include but are not limited to an adapted keyboard, large keyboard, Sticky Keys, Mouse Keys, Filter Keys, adapted mouse, touch screen, head wand, and switches.

Calculator (for grades 6-8 mathematics tests): When the embedded Desmos calculator or specialized calculator is inaccessible, the provision of a hand-held calculator may be appropriate: either a basic calculator (grade 6) or a scientific calculator (grades 7-8).

Human Signer: This sign language accommodation allows a qualified test administrator to sign or provide visual language support for the test directions, test content, and/or reading passages to a student who is deaf or hard of hearing.

Math Manipulatives: These tools are available to allow eligible students to use concrete mathematical tools strategically to support their decision making. Students eligible for this accommodation typically have visual or math-related disabilities.

Multiplication Table: This is a paper-based single digit (1–9) multiplication table students use for reference.

Paper Tests (large print and braille): Paper tests are available in large print and braille for students who need these accommodations in paper format.

Print-on-Demand: Paper copies of passages, stimuli, and/or items are printed for students. For those students who need a paper copy of a passage or stimulus, permission for the students to request printing must first be set in TIDE.

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

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

Specialized Calculator (for grades 6–8 mathematics tests): A non-embedded calculator may be provided for students who need a special calculator, such as a braille calculator or a talking calculator that is currently unavailable within the assessment platform.

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 4 presents a list of universal tools, designated supports, and accommodations that were offered in the 2022–2023 administration. Tables 5–10 provide the number of students who utilized any of the offered accommodations and designated supports.

Table 4. 2022–2023 Universal Tools, Designated Supports, and Accommodations

Universal Tools	Designated Supports	Accommodations
Embedded		
Breaks Calculator[1] Digital Notepad English Glossary Expandable Passages/Stimuli/Items Highlighter Keyboard Commands Line Reader Mark for Review Mathematics Tools[2] Strikethrough Writing Tools[3] Zoom	Color Contrast Illustration Glossary[4] Masking Mouse Pointer Print Size Online Streamline Text-to-Speech[5] Translated Test Directions[4] Translations (Glossary)[4] Translations (Spanish-Stacked)[6] Turn Off Any Universal Tools	American Sign Language[7] Braille Braille Transcripts[8] Closed Captioning[8] Speech-to-Text Text-to-Speech[9]
Non-Embedded		
Breaks Scratch Paper/White Board	Amplification Color Contrast Color Overlay Illustration Glossary[4] Magnification Native Language Reader of Teacher Script Noise Buffers Read Aloud[10] Read Aloud in Spanish[6] Separate Setting Simplified Test Directions Translated Test Directions Translations (Glossary)[6]	100s Number Table Abacus Alternate Response Options[11] Calculator[1] Human Signer Math Manipulatives Multiplication Table[6] Paper Tests (Large Print and Braille) Print-on-Demand Read Aloud[12] Scribe Specialized Calculator[1] Speech-to-Text

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

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

[2] Includes embedded ruler, embedded protractor

[3] For interim ELA/L performance tasks; includes bold, italic, underline, indent, cut, paste, spell check, bullets, undo/redo

[4] For mathematics items

[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. Also available in Spanish for mathematics tests.

[6] For mathematics tests

[7] For ELA/L listening items and mathematics items

[8] For ELA/L listening items

[9] For ELA/L reading passages; must be set in TIDE by state-level user

[10] For ELA/L items (not ELA/L reading passages) and mathematics items

[11] Includes adapted keyboards, large keyboard, Sticky Keys, Mouse Keys, Filter Keys, adapted mouse, touch screen, head wand, and switches

[12] For ELA/L reading passages, all grades

Table 5. ELA/L Total Students with Allowed Embedded and Non-Embedded Accommodations

Accommodations	Grade					
	3	4	5	6	7	8
Embedded Accommodations						
American Sign Language	2	8	9	2	11	3
Braille		2	1	1	1	
Braille Transcripts		1	2	1	1	1
Closed Captioning	40	47	49	46	66	59
Speech-to-Text	569	723	755	697	507	366
Text-to-Speech: Passages and Items	1,759	1,805	1,727	1,445	1,287	1,182
Non-Embedded Accommodations						
Alternate Response Options	14	18	17	13	14	17
Customized Medical Accommodations	1	1		1		
Speech-to-Text	43	78	135	90	48	51

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

Designated Supports	Subgroup	Grade					
		3	4	5	6	7	8
Color Contrast	Overall	8	14	17	46	29	33
	LEP		1	1	12	10	14
	IDEA	5	11	7	7	14	16
Masking	Overall	216	242	250	196	153	159
	LEP	52	57	65	17	22	19
	IDEA	150	177	168	123	126	145
Mouse Pointer	Overall	8	4	3	31	6	3
	LEP		1		4		
	IDEA	7	3	3	3	3	3
Print Size Online	Overall	47	60	58	47	31	121
	LEP	11	6	7	6	1	22
	IDEA	32	35	36	28	18	32
Streamline	Overall	301	322	256	260	185	190
	LEP	124	131	115	93	29	37
	IDEA	167	176	141	127	139	125
Text-to-Speech: Items	Overall	8,357	7,840	7,710	6,572	5,994	5,851
	LEP	3,516	3,368	3,122	2,170	1,927	1,808
	IDEA	2,234	2,373	2,571	2,486	2,399	2,256

Table 7. ELA/L Total Students with Allowed Non-Embedded Designated Supports

Designated Supports	Subgroup	Grade					
		3	4	5	6	7	8
Color Contrast	Overall	8	8	1	3	4	5
	LEP		2		1		
	IDEA	5	4		3	3	3
Color Overlay	Overall	8	1	6	4	3	4
	LEP						
	IDEA	6	1	5	3	3	3
Magnification	Overall	12	18	11	21	7	9
	LEP	2	2	4	2		
	IDEA	6	8	5	11	4	3
Medical Device	Overall	3	1	2	2	6	3
	LEP				1		1
	IDEA	1		1		2	1
Noise Buffers	Overall	24	25	18	20	26	23
	LEP	3	2	3	5	7	6
	IDEA	14	13	12	12	16	13
Read-Aloud Items	Overall	360	303	295	217	156	174
	LEP	87	87	67	41	40	50
	IDEA	293	232	249	189	139	151
Separate Setting	Overall	4,826	5,043	5,223	4,725	4,449	4,548
	LEP	962	991	989	714	617	620
	IDEA	3,521	3,759	3,904	3,684	3,531	3,562
Simplified Test Directions	Overall	1,219	1,242	1,149	1,071	1,132	1,001
	LEP	457	401	384	346	373	378
	IDEA	872	947	921	822	912	769
Translated Test Directions	Overall	85	87	68	124	142	142
	LEP	84	85	66	124	138	141
	IDEA	10	13	9	13	18	17

Table 8. Mathematics Total Students with Allowed Embedded and Non-Embedded Accommodations

Accommodations	Grade					
	3	4	5	6	7	8
Embedded Accommodations						
American Sign Language	2	7	8	2	10	3
Braille		2	1	1	3	
Speech-to-Text	524	657	691	614	422	322
Non-Embedded Accommodations						
100s Number Table	2,157	2,033	1,586	1,035	755	516
Abacus	2	3	3		3	1
Alternate Response Options	16	15	12	12	13	11
Customized Medical Accommodations	1	1		1		
Multiplication Table		3,314	3,540	3,531	3,288	2,988
Specialized Calculator			12	86	117	134
Speech-to-Text	41	72	125	83	39	44

Table 9. Mathematics Total Students with Allowed Embedded Designated Supports

Designated Supports	Subgroup	Grade					
		3	**4**	**5**	**6**	**7**	**8**
Color Contrast	Overall	9	13	17	43	27	32
	LEP		1	1	11	10	14
	IDEA	6	10	7	6	13	16
Illustration Glossary	Overall	1,751	1,676	1,465	1,376	1,146	1,056
	LEP	1,644	1,656	1,425	1,290	1,117	1,027
	IDEA	237	208	233	193	172	173
Masking	Overall	173	199	218	186	145	143
	LEP	48	54	55	15	20	15
	IDEA	111	142	139	97	98	112
Mouse Pointer	Overall	7	4	2	29	7	1
	LEP		1		4		
	IDEA	6	3	2	2	4	1
Print Size Online	Overall	48	57	57	43	36	119
	LEP	11	6	7	5	1	22
	IDEA	33	32	35	23	21	31
Streamline	Overall	298	315	252	254	174	191
	LEP	123	131	115	90	30	37
	IDEA	164	174	135	126	126	125
Text-to-Speech: Stimuli and Items	Overall	10,429	10,089	9,803	8,336	7,750	7,506
	LEP	3,837	3,756	3,454	2,420	2,136	2,081
	IDEA	3,961	4,153	4,256	3,818	3,580	3,344
Translations (Glossary): Spanish	Overall	934	907	865	1,089	970	970
	LEP	924	892	844	1,020	951	931
	IDEA	97	92	110	132	137	129
Translations (Glossary): Other Languages	Overall	82	105	80	71	66	66
	LEP	80	101	79	66	64	65
	IDEA	3	8	4	4	5	3
Translations (Spanish-Stacked)	Overall	651	743	682	759	724	785
	LEP	647	732	675	727	710	769
	IDEA	30	55	45	49	51	69

Table 10. Mathematics Total Students with Allowed Non-Embedded Designated Supports

Designated Supports	Subgroup	Grade					
		3	**4**	**5**	**6**	**7**	**8**
Color Contrast	Overall	7	4	2	2	5	7
	LEP						
	IDEA	4	3	1	2	4	5
Color Overlay	Overall	9	3	6	4	4	2
	LEP		1				
	IDEA	7	3	5	3	4	1
Illustration Glossary	Overall	77	103	48	26	29	33
	LEP	73	102	47	26	29	32
	IDEA	19	20	12	5	1	5
Magnification	Overall	15	19	12	24	11	11
	LEP	2	3	4	2		
	IDEA	9	9	5	14	6	4
Medical Device	Overall	2	1	4	2	7	3
	LEP				1		1
	IDEA	1		3		3	1
Noise Buffers	Overall	21	23	18	21	25	22
	LEP	3	1	3	5	7	7
	IDEA	13	11	12	13	15	12
Read Aloud Stimuli and Items	Overall	311	260	307	198	126	139
	LEP	66	81	55	38	24	30
	IDEA	259	204	235	178	118	131
Read Aloud Stimuli and Items (Spanish)	Overall	30	46	38	12	15	19
	LEP	30	42	37	11	13	19
	IDEA	3	6	6	3	3	1
Separate Setting	Overall	4,869	5,110	5,256	4,734	4,477	4,543
	LEP	980	1,003	995	718	618	624
	IDEA	3,550	3,790	3,917	3,676	3,560	3,556
Simplified Test Directions	Overall	1,209	1,247	1,129	1,065	1,107	1,011
	LEP	446	398	380	344	355	372
	IDEA	867	971	915	818	888	776
Translated Test Directions	Overall	70	88	55	114	131	137
	LEP	70	86	53	114	128	136
	IDEA	8	14	6	11	14	17
Translations (Glossary): Spanish	Overall	82	92	69	119	124	132
	LEP	81	92	67	116	122	132
	IDEA	10	2	7	16	8	8
Translations (Glossary): Other Languages	Overall	5	9	10	13	10	10
	LEP	4	8	9	12	10	10
	IDEA	1		2	1		

2.7 Testing 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. For discrete items, each item appears on the screen one item at a time, whereas stimulus-based items appear on the screen together. The page time is the time spent on one item for discrete items and the time spent on all items associated with a stimulus for stimulus-based items. 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 (stimulus-based items).

The Smarter Balanced summative assessments are not timed, and an individual student may need more or less testing time overall. The length of a test session is determined by TEs/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/TAs must use their best professional judgment when allowing students extra time. Students should be actively engaged in responding productively to test items.

Tables 11 and 12 present an average testing time and the testing time at percentiles for the overall test, the CAT component, and the PT component.

Table 11. ELA/L Testing Time

Grade	Average Testing Time (hh:mm)	SD of Testing Time (hh:mm)	Testing Time by Percentile (hh:mm)				
			75th	80th	85th	90th	95th
Overall Test (CAT Component)							
3	1:46	0:58	2:06	2:16	2:28	2:48	3:24
4	1:51	1:01	2:11	2:21	2:34	2:53	3:31
5	1:49	0:52	2:10	2:19	2:31	2:48	3:19
6	1:51	0:52	2:14	2:24	2:37	2:55	3:28
7	1:44	0:49	2:05	2:14	2:25	2:42	3:12
8	1:38	0:46	1:58	2:06	2:16	2:31	3:02

Table 12. Mathematics Testing Times

Grade	Average Testing Time (hh:mm)	SD of Testing Time (hh:mm)	Testing Time by Percentile (hh:mm)				
			75th	80th	85th	90th	95th
Overall Test							
3	2:00	1:03	2:28	2:40	2:55	3:18	3:56
4	2:10	1:07	2:39	2:52	3:08	3:30	4:13
5	2:18	1:09	2:50	3:03	3:20	3:42	4:23
6	2:07	1:00	2:34	2:46	3:00	3:21	3:58
7	1:53	0:54	2:17	2:26	2:39	2:57	3:29
8	1:54	0:54	2:19	2:29	2:41	2:58	3:31
CAT Component							
3	1:24	0:45	1:43	1:52	2:03	2:18	2:48
4	1:34	0:50	1:56	2:05	2:18	2:35	3:07
5	1:34	0:47	1:56	2:05	2:16	2:32	2:58
6	1:28	0:41	1:46	1:55	2:05	2:19	2:45
7	1:27	0:42	1:46	1:54	2:03	2:17	2:43
8	1:26	0:41	1:46	1:53	2:03	2:16	2:41
PT Component							
3	0:37	0:24	0:47	0:51	0:57	1:06	1:22
4	0:36	0:23	0:45	0:49	0:55	1:03	1:17
5	0:44	0:31	0:56	1:01	1:09	1:20	1:39
6	0:39	0:25	0:50	0:55	1:01	1:09	1:25
7	0:26	0:19	0:33	0:36	0:41	0:47	0:59
8	0:28	0:19	0:35	0:39	0:43	0:50	1:01

2.8 Data Forensics Program

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

For online administrations, a set of quality assurance (QA) reports is generated during and after the testing window. One of the QA reports focuses on flagging possible testing anomalies. Testing anomalies are analyzed by examining changes in student performance from year to year, test taking time, item response patterns using a person-fit index, and item response change analyses.

Analyses are performed at the student level and summarized for each aggregate unit, including testing session, TA, and school. Flagging criteria used for these analyses are described below and are configurable by an authorized user. When the aggregate unit size is small, the aggregate

unit is flagged if the percentage of flagged students is greater than 50% in the analysis. The default small aggregate unit size is 5 or fewer students but this value is configurable. For each aggregate unit, small groups are identified based on the number of tests included in the aggregate unit from that analysis. Thus, a small unit identified in one analysis may not be a small unit in another analysis. The QA reports are provided to state clients to monitor testing anomalies throughout the testing window.

2.8.1 Changes in Student Performance

Changes in student scores between administration years are examined using a regression model to check for outliers. For these between-year comparisons, students' current-year scores are regressed on their test scores from the previous year and on the number of days between the two years' test-end dates (to control for the instruction time between the two test scores).

A large score gain or loss in student scores between administration years is detected by examining the residuals for outliers. The residuals are computed as the observed value minus the regression model's predicted value. To detect unusual residuals, the studentized residuals are computed. An unusual increase or decrease in student scores between administration years is flagged when the absolute value of the studentized residual is greater than 3.

The residuals of students are also 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 of the residuals in the aggregate unit (e.g., testing session, TA, school). For each aggregate unit, a *t* value is computed and flagged when $|t|$ is greater than 3,

$$t = \frac{\sum_{i=1}^{n} \hat{e}_i \,/n}{\sqrt{\frac{s^2}{n} + \frac{\sum_{i=1}^{n} \sigma^2(1-h_{ii})}{n^2}}},$$

where *s* is the standard deviation of residuals in an aggregate unit; *n* is the number of students in an aggregate unit (e.g., testing session, TA, school), σ^2 is the MSE from the regression, and $\hat{e}_i$ is the residual for the *i*th student.

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

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

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

2.8.2 Test-Taking Time

The summative assessments are not timed, and thus an individual student's test taking times may vary across students. However, unusual test-taking times such as excessively shorter or longer test-taking time may indicate irregularities in test administration. An example of an unusual test-taking time is a test record for an individual who scores very well on the test even though the average time spent is far less than that required of students statewide. If students already know the answers to the items, the test-taking time may be much shorter than the test-taking time for those who have 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 state average testing time and standard deviation are computed based on all students available when the analysis was performed. Students and aggregate units are flagged if the test-taking time is different from the state average by three standard deviations or more, although the flagging criteria can be adjusted by an authorized user.

2.8.3 Inconsistent Item Response Pattern (Person Fit)

In item response theory (IRT) models, person-fit measurement is used to identify test-takers 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. However, 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 in a test. An unlikely response to a single test item 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). 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 less than -3 are flagged. Aggregate units are flagged with *t* less than -3,

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

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.

2.8.4 Item Response Change

Students are allowed to revisit items as many times as they wish within a session and may also mark items to be revisited prior to completing the session. However, excessively high rates of response change, especially high rates of item score increases (i.e., response changes from wrong to right), may indicate irregularities in test administration. For example, test administrators (TAs) could review students' responses and either coach them to modify their responses or keep the session active and change responses themselves.

To identify irregular patterns of response change, the item score for the final response to each item and the penultimate response if one exists are examined, and the number of instances in which the item score increases are counted.

The average and standard deviation of positive item score changes are computed based on all students available when the analysis was performed. Students and aggregate units are flagged if the number of positive item score changes is larger than the state average by three standard deviations or more, although the flagging criteria can be adjusted by an authorized user.

2.9 Prevention and Recovery of Disruptions in Test Delivery System

CAI is continuously improving its ability to protect testing systems from interruptions. CAI's TDS is designed to ensure that student responses are captured accurately and stored on more than one server in case of a failure. The CAI architecture, described in the following paragraphs, is designed to recover from a failure of any component with little interruption. Each system is redundant, and critical student response data are transferred to a different data center each night.

CAI has developed a unique monitoring system that is extremely sensitive to changes in server performance. Most monitoring systems provide warnings when something is going wrong. The

CAI system does, too, but it also provides warnings when any given server performs differently from its performance over the few hours prior or differently than the other servers performing the same jobs. Subtle changes in performance often precede actual failure by hours or days, allowing CAI to detect of potential problems, investigate them, and mitigate them. This system has enabled CAI to make adjustments and replace equipment on multiple occasions before any problems occurred.

CAI has also implemented an escalation procedure to alert clients within minutes of any disruption. The emergency alert system notifies CAI's executive and technical staff by text message, who then immediately join a call to identify and address the problem.

The following subsection describes CAI system architecture and how it recovers from device failures, Internet interruptions, and other problems.

2.9.1 High-Level System Architecture

CAI's architecture provides the redundancy, robustness, and reliability required by a large-scale, high-stakes testing program. The general approach, which Smarter Balanced has adopted as standard policy, is pragmatic and well supported by the system architecture.

CAI posits that any system built around an expectation of the flawless performance of computers or networks within schools and districts is bound to fail. Therefore, the system is designed to ensure that the testing results and experience respond robustly to such inevitable failures. CAI's TDS is designed to protect data integrity and prevent student data loss at every point throughout the test administration process. Fault tolerance and automated recovery are built into every component of the system.

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

Student Machine

Student responses are conveyed to CAI's 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 CAI 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 serves as a backup hub for every four satellites. 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 malfunction, they are removed from service. Real-time student data are immediately recoverable from the satellite, backup hub, or hub (described in the next subsection), 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 earlier. This real-time backup copy remains on the hub until the hub receives a 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 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 a notification immediately goes out to CAI's 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.

2.9.2 Automated Backup and Recovery

Every system is backed up nightly. Industry-standard backup and recovery procedures are in place to ensure the 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.

2.9.3 Other Disruption Prevention and Recovery Systems

These testing systems are designed to be extremely fault-tolerant. The systems can withstand failure of any component with little or no service interruption. This robustness is archived through redundancy. Key redundant systems are as follows:

- The system's 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.
- The hosting provider has multiple redundancies in the flow of information to and from the system's data centers 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 are redundant firewalls and load balancers throughout the environment.
- The system uses redundant power and switching in all server cabinets.
- Data are protected by nightly backups. A full weekly backup and incremental nightly backups protect data. Should a catastrophic event occur, CAI 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 need to rerun it.

The system's TDS 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 is redundant at every component, and in the event of failure, the unique design ensures that data are always stored in at least two locations. The engineering that led to this system protects student responses from loss.

3. SUMMARY OF 2022–2023 OPERATIONAL TEST ADMINISTRATION

3.1 STUDENT POPULATION

All Connecticut students enrolled in grades 3–8 in all public schools are required to participate in the Smarter Balanced English language arts/literacy (ELA/L) and mathematics assessments. Before the testing window opens, the state or districts send Cambium Assessment, Inc. (CAI) a student enrollment file to load into the Test Information Distribution Engine (TIDE). Using this enrollment file, the participation rates are calculated as the percentage of students who attempted the tests. Tables 13 and 14 present the participation rates in percentages for all students and by subgroups who attempted the tests. Tables 15 and 16 present the number of Connecticut students who meet attemptedness requirements for the Smarter Balanced summative scoring and reporting.

Table 13. Participation Rates by Percentage in ELA/L Summative Assessment

Group	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
All Students	97.88	97.79	97.74	97.60	97.32	96.93
Female	98.57	98.33	98.25	98.09	97.99	97.28
Male	97.24	97.27	97.26	97.14	96.69	96.62
Black or African American	96.99	96.99	96.91	96.88	96.90	96.44
AmerIndian/Alaskan	100.00	99.04	97.56	100.00	98.08	96.94
Asian	97.59	98.33	97.30	97.83	97.36	98.39
Hispanic or Latino	97.71	97.43	97.56	97.43	97.01	96.76
Pacific Islander	100.00	100.00	100.00	97.44	100.00	100.00
White	98.28	98.13	98.14	97.89	97.65	97.03
Multi-Racial	97.61	98.09	97.58	97.50	97.18	96.87
LEP	97.90	97.35	97.42	96.86	96.09	95.88
IDEA	89.51	89.39	89.39	89.76	89.23	88.49

Note. AmerIndian/Alaskan = American Indian or Alaska Native; Pacific Islander = Native Hawaiian or Other Pacific Islander; LEP = Limited English Proficiency Status; IDEA= Individuals with Disabilities Education Act.

Table 14. Participation Rates by Percentage in Mathematics Summative Assessment

Group	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
All Students	97.70	97.66	97.54	97.25	96.77	96.17
Female	98.43	98.24	98.07	97.74	97.48	96.52
Male	97.01	97.11	97.04	96.78	96.09	95.86
Black or African American	96.67	96.65	96.65	96.61	96.03	95.43
AmerIndian/Alaskan	100.0	99.04	97.56	98.70	98.08	96.94
Asian	97.54	98.28	97.25	97.57	97.21	98.13
Hispanic or Latino	97.49	97.33	97.33	96.90	96.32	95.66
Pacific Islander	100.0	100.0	100.0	97.44	100.0	97.78
White	98.15	98.03	97.97	97.62	97.24	96.54
Multi-Racial	97.28	97.98	97.29	97.10	96.39	95.62
LEP	97.82	97.27	97.23	96.55	95.45	94.92

Group	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
IDEA	89.05	89.03	88.74	88.69	88.16	86.71

Table 15. Number of Students in ELA/L Summative Assessment

Group	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
All Students	35,822	35,921	36,384	36,506	36,914	38,070
Female	17,487	17,591	17,867	17,925	18,147	18,489
Male	18,331	18,326	18,512	18,561	18,745	19,525
Black or African American	4,288	4,258	4,455	4,603	4,721	4,967
AmerIndian/Alaskan	81	103	80	77	102	95
Asian	1,911	1,954	2,019	1,897	1,892	1,951
Hispanic or Latino	10,895	11,242	11,026	11,121	11,189	11,496
Pacific Islander	32	30	45	38	29	46
White	16,820	16,583	17,063	17,090	17,393	17,966
Multi-Racial	1,795	1,751	1,696	1,680	1,588	1,549
LEP	4,891	4,856	4,430	3,716	3,345	3,120
IDEA	5,556	5,698	6,014	5,969	6,066	6,142

Table 16. Number of Students in Mathematics Summative Assessment

Group	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
All Students	35,743	35,860	36,300	36,358	36,691	37,757
Female	17,458	17,564	17,829	17,857	18,044	18,338
Male	18,281	18,292	18,466	18,482	18,625	19,364
Black or African American	4,270	4,239	4,443	4,589	4,674	4,912
AmerIndian/Alaskan	81	103	80	76	102	95
Asian	1,910	1,952	2,019	1,890	1,887	1,947
Hispanic or Latino	10,866	11,227	10,990	11,051	11,106	11,358
Pacific Islander	32	30	45	38	29	45
White	16,794	16,562	17,031	17,041	17,318	17,872
Multi-Racial	1,790	1,747	1,692	1,673	1,575	1,528
LEP	4,892	4,851	4,415	3,700	3,321	3,088
IDEA	5,556	5,696	5,993	5,896	5,996	6,032

3.2 Summary of Student Performance

Tables 17–20 summarize overall student performance in the 2022–2023 summative test for all students and by subgroups, including the average and the standard deviation of overall scale scores, the percentage of students in each achievement level, and the percentage of proficient students.

Figures 1 and 2 show the percentage of proficient students over the past six years for all students in ELA/L and over the past seven years for all students in mathematics (cohort comparisons). Figures 3 and 4 show the average scale scores over the past six years for all students in ELA/L and over the past seven years for all students in mathematics. In ELA/L, student performance is compared for six years because ELA/L scores in 2014–2015 were based on both computer-adaptive test (CAT) and performance task (PT) components while ELA/L

scores from 2015–2016 were based on the CAT component only. In Figures 1–4, the 2019–2020 performance is not included because the testing was cancelled due to the COVID-19 pandemic. The average and the standard deviation of scale scores, as well as the percentage of proficient students for each test administration across four years, are provided in Appendix B.

Table 17. Descriptive Statistics and Percentage of Students in Achievement Levels for Overall and by Subgroup: ELA/L (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	35,822	2417	96	32	23	20	25	45
Female	17,487	2424	96	29	22	21	27	48
Male	18,331	2411	96	35	23	19	23	43
Black or African American	4,288	2372	86	51	24	15	10	25
AmerIndian/Alaskan	81	2385	93	44	25	17	14	31
Asian	1,911	2464	93	16	18	22	43	66
Hispanic or Latino	10,895	2373	90	50	23	15	11	26
Pacific Islander	32	2417	101	28	31	16	25	41
White	16,820	2452	86	17	22	25	36	61
Multi-Racial	1,795	2426	97	29	22	20	28	49
LEP	4,891	2345	80	64	21	11	5	16
IDEA	5,556	2347	81	64	20	10	6	16
Grade 4								
All Students	35,921	2464	103	33	18	21	28	49
Female	17,591	2469	101	31	18	22	29	51
Male	18,326	2459	104	35	18	21	26	47
Black or African American	4,258	2419	92	51	19	17	12	29
AmerIndian/Alaskan	103	2430	102	48	19	16	17	33
Asian	1,954	2515	101	18	13	22	47	69
Hispanic or Latino	11,242	2416	97	52	19	16	13	29
Pacific Islander	30	2476	91	30	23	20	27	47
White	16,583	2500	91	18	18	25	39	64
Multi-Racial	1,751	2481	99	26	19	23	32	55
LEP	4,856	2380	88	67	16	12	5	17
IDEA	5,698	2385	89	67	15	11	7	17
Grade 5								
All Students	36,384	2501	109	30	18	25	26	51
Female	17,867	2509	107	27	19	25	28	54
Male	18,512	2493	109	33	18	25	24	49
Black or African American	4,455	2448	98	50	20	20	10	30
AmerIndian/Alaskan	80	2495	97	33	20	29	19	48
Asian	2,019	2560	101	13	13	27	47	74
Hispanic or Latino	11,026	2449	102	48	21	20	11	31
Pacific Islander	45	2476	104	36	20	27	18	44
White	17,063	2539	96	16	17	30	37	67
Multi-Racial	1,696	2520	106	25	17	27	32	59

LEP	4,430	2400	85	68	18	11	2	13
IDEA	6,014	2413	95	65	17	12	6	18

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

Table 18. Descriptive Statistics and Percentage of Students in Achievement Levels for Overall and by Subgroup: ELA/L (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	36,506	2520	102	28	24	29	18	48
Female	17,925	2528	101	24	25	31	20	51
Male	18,561	2512	103	31	24	28	17	45
Black or African American	4,603	2476	93	43	28	22	7	29
AmerIndian/Alaskan	77	2480	92	49	16	30	5	35
Asian	1,897	2580	97	11	17	33	39	72
Hispanic or Latino	11,121	2475	97	44	27	21	8	29
Pacific Islander	38	2492	103	37	26	29	8	37
White	17,090	2554	91	15	23	36	26	62
Multi-Racial	1,680	2532	101	23	25	30	21	52
LEP	3,716	2413	76	71	22	7	0	7
IDEA	5,969	2434	85	63	23	11	3	14
Grade 7								
All Students	36,914	2539	110	29	23	32	17	49
Female	18,147	2550	107	25	22	34	19	53
Male	18,745	2529	112	32	23	30	15	45
Black or African American	4,721	2495	101	43	28	23	7	30
AmerIndian/Alaskan	102	2513	93	32	37	24	7	30
Asian	1,892	2606	102	12	14	34	40	74
Hispanic or Latino	11,189	2490	108	46	24	23	7	30
Pacific Islander	29	2509	98	34	17	48	0	48
White	17,393	2575	96	16	21	40	23	63
Multi-Racial	1,588	2548	111	27	22	32	19	52
LEP	3,345	2410	80	79	16	4	0	5
IDEA	6,066	2447	96	64	21	12	2	15
Grade 8								
All Students	38,070	2555	111	28	24	32	17	48
Female	18,489	2569	108	23	24	34	19	53
Male	19,525	2542	113	32	24	30	14	44
Black or African American	4,967	2512	101	42	28	24	7	30
AmerIndian/Alaskan	95	2529	103	33	33	26	8	35
Asian	1,951	2627	101	10	15	36	39	75
Hispanic or Latino	11,496	2503	106	45	26	23	6	29
Pacific Islander	46	2539	104	30	33	26	11	37
White	17,966	2592	99	15	22	39	24	63
Multi-Racial	1,549	2569	110	25	21	35	19	54
LEP	3,120	2417	76	81	16	3	0	3
IDEA	6,142	2462	94	63	23	12	2	14

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

Table 19. Descriptive Statistics and Percentage of Students in Achievement Levels for Overall and by Subgroup: Mathematics (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	35,743	2432	92	29	21	26	24	50
Female	17,458	2427	89	30	22	26	22	48
Male	18,281	2436	94	27	21	26	26	52
Black or African American	4,270	2380	84	50	24	18	8	26
AmerIndian/Alaskan	81	2406	78	36	25	30	10	40
Asian	1,910	2487	89	12	15	25	48	73
Hispanic or Latino	10,866	2389	84	46	25	19	10	29
Pacific Islander	32	2414	96	41	25	9	25	34
White	16,794	2466	79	14	19	32	34	66
Multi-Racial	1,790	2438	92	27	21	26	26	53
LEP	4,892	2374	82	54	23	16	7	23
IDEA	5,556	2359	88	61	20	12	6	19
Grade 4								
All Students	35,860	2475	94	26	26	25	23	48
Female	17,564	2469	89	27	27	26	20	45
Male	18,292	2481	98	25	24	24	27	51
Black or African American	4,239	2421	85	46	30	17	7	24
AmerIndian/Alaskan	103	2439	96	40	23	26	11	37
Asian	1,952	2538	92	11	15	23	51	74
Hispanic or Latino	11,227	2431	87	42	30	18	10	28
Pacific Islander	30	2493	95	23	23	17	37	53
White	16,562	2510	81	12	24	32	33	65
Multi-Racial	1,747	2486	93	22	23	28	26	55
LEP	4,851	2412	83	51	29	14	6	19
IDEA	5,696	2396	90	59	23	12	6	17
Grade 5								
All Students	36,300	2499	100	34	24	18	24	42
Female	17,829	2495	96	36	26	17	21	39
Male	18,466	2504	103	33	23	18	26	45
Black or African American	4,443	2439	87	60	23	10	7	17
AmerIndian/Alaskan	80	2490	96	33	31	16	20	36
Asian	2,019	2572	94	11	18	18	52	71
Hispanic or Latino	10,990	2452	91	53	25	12	10	22
Pacific Islander	45	2494	91	44	13	22	20	42
White	17,031	2535	87	18	24	23	34	57
Multi-Racial	1,692	2515	100	28	24	20	29	48
LEP	4,415	2424	80	67	22	7	4	11
IDEA	5,993	2417	90	69	18	8	5	13

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

Table 20. Descriptive Statistics and Percentage of Students in Achievement Levels for Overall and by Subgroup: Mathematics (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	36,358	2512	116	35	26	19	21	40
Female	17,857	2508	113	36	27	18	19	38
Male	18,482	2516	119	34	25	19	23	42
Black or African American	4,589	2445	105	59	24	11	6	17
AmerIndian/Alaskan	76	2466	108	50	26	13	11	24
Asian	1,890	2599	111	13	16	20	51	71
Hispanic or Latino	11,051	2458	108	54	26	12	8	20
Pacific Islander	38	2483	113	39	26	24	11	34
White	17,041	2554	98	18	27	24	30	55
Multi-Racial	1,673	2522	117	32	26	18	24	42
LEP	3,700	2407	93	76	18	5	2	6
IDEA	5,896	2409	106	71	19	6	4	10
Grade 7								
All Students	36,691	2528	119	37	23	19	21	40
Female	18,044	2525	116	37	24	19	19	38
Male	18,625	2531	122	36	22	20	22	42
Black or African American	4,674	2463	101	60	24	11	5	16
AmerIndian/Alaskan	102	2504	97	45	27	18	10	27
Asian	1,887	2624	119	15	13	19	53	72
Hispanic or Latino	11,106	2469	106	58	23	12	7	19
Pacific Islander	29	2490	96	45	28	21	7	28
White	17,318	2573	104	20	25	27	29	56
Multi-Racial	1,575	2537	123	35	23	18	24	42
LEP	3,321	2410	83	82	13	3	1	4
IDEA	5,996	2426	99	75	15	7	3	10
Grade 8								
All Students	37,757	2537	127	42	22	16	20	36
Female	18,338	2538	122	41	23	17	19	36
Male	19,364	2536	131	43	21	15	21	37
Black or African American	4,912	2466	105	66	20	8	5	13
AmerIndian/Alaskan	95	2495	112	51	29	12	8	20
Asian	1,947	2642	122	15	15	19	51	70
Hispanic or Latino	11,358	2474	109	64	20	9	7	16
Pacific Islander	45	2536	101	38	33	18	11	29
White	17,872	2584	113	24	24	22	29	51
Multi-Racial	1,528	2548	129	39	21	17	23	39
LEP	3,088	2409	81	89	8	2	1	3
IDEA	6,032	2430	101	79	13	5	3	8

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

Figure 1. ELA/L Percent Proficient Across Years

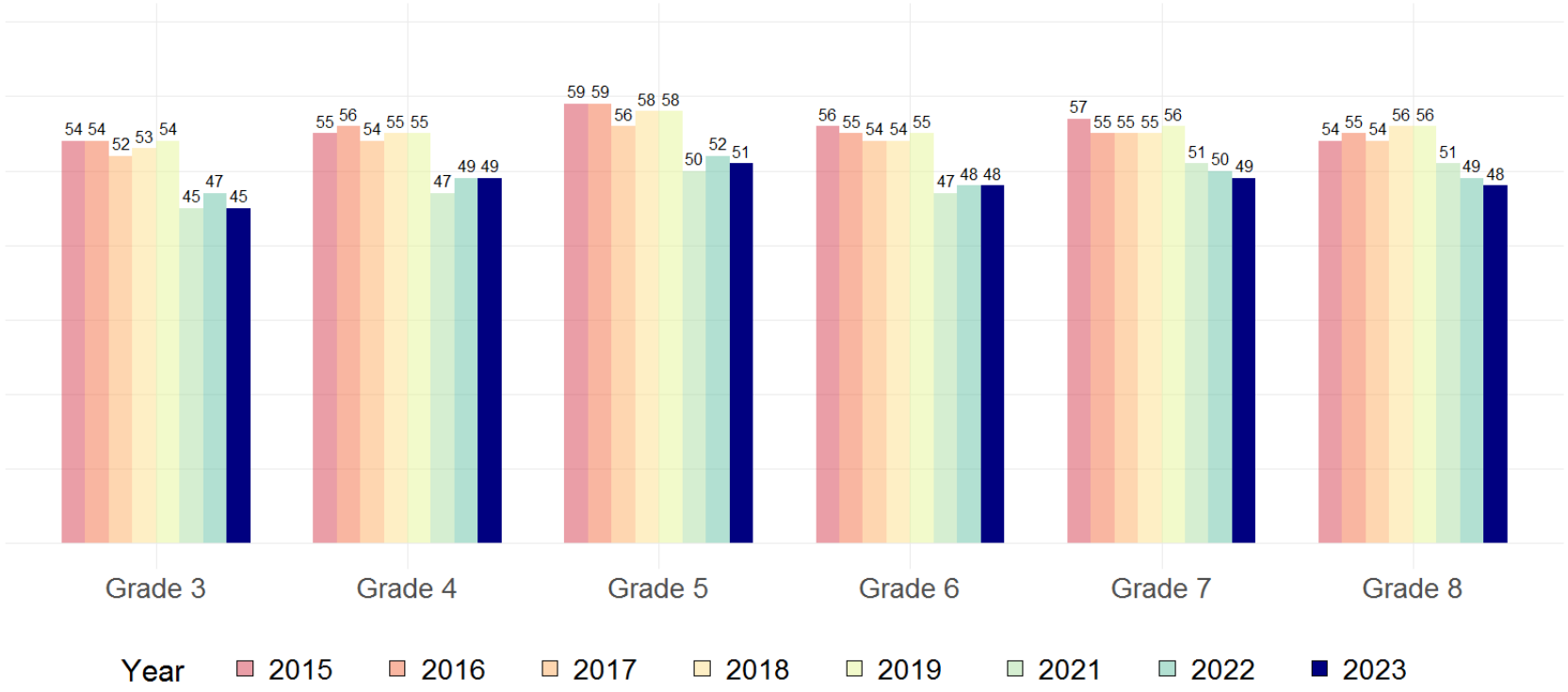


Figure 2. Mathematics Percent Proficient Across Years

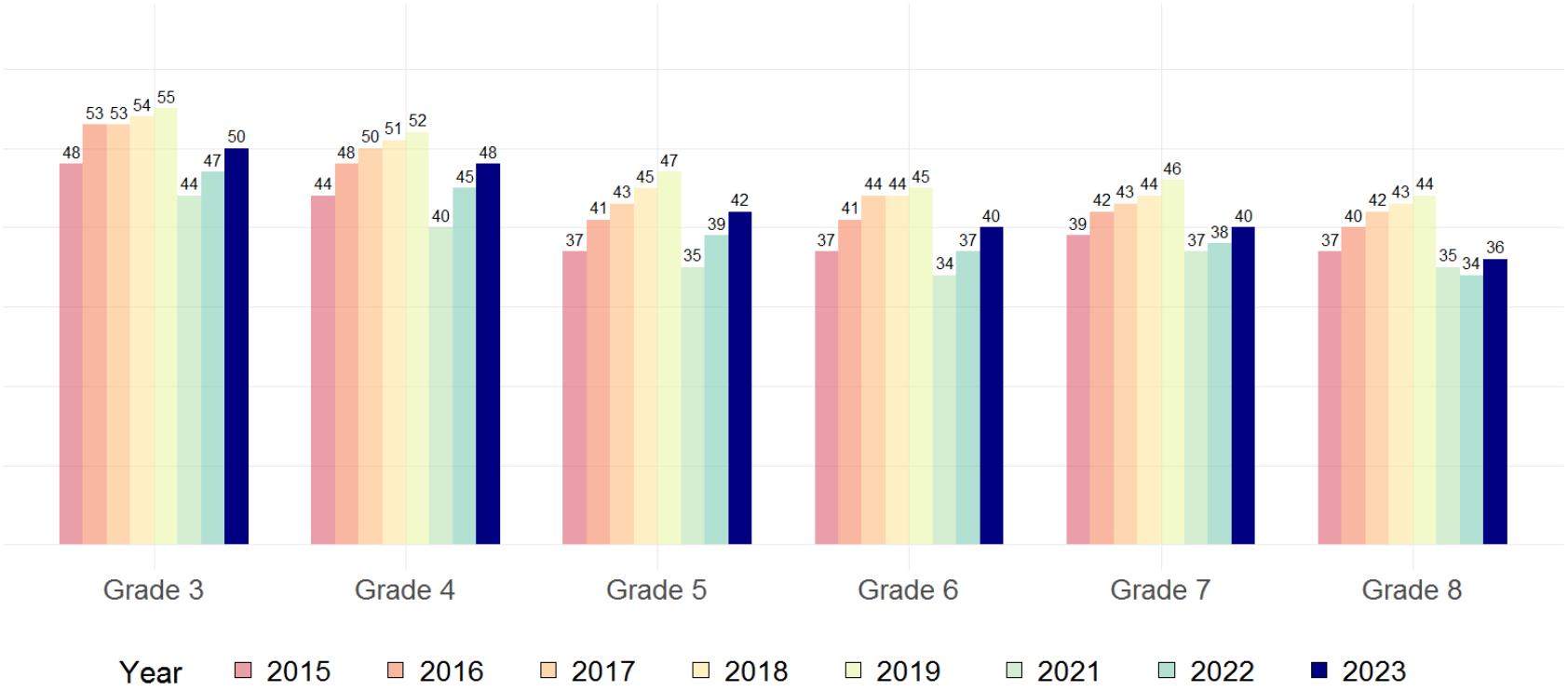


Figure 3. ELA/L Average Scale Score Across Years

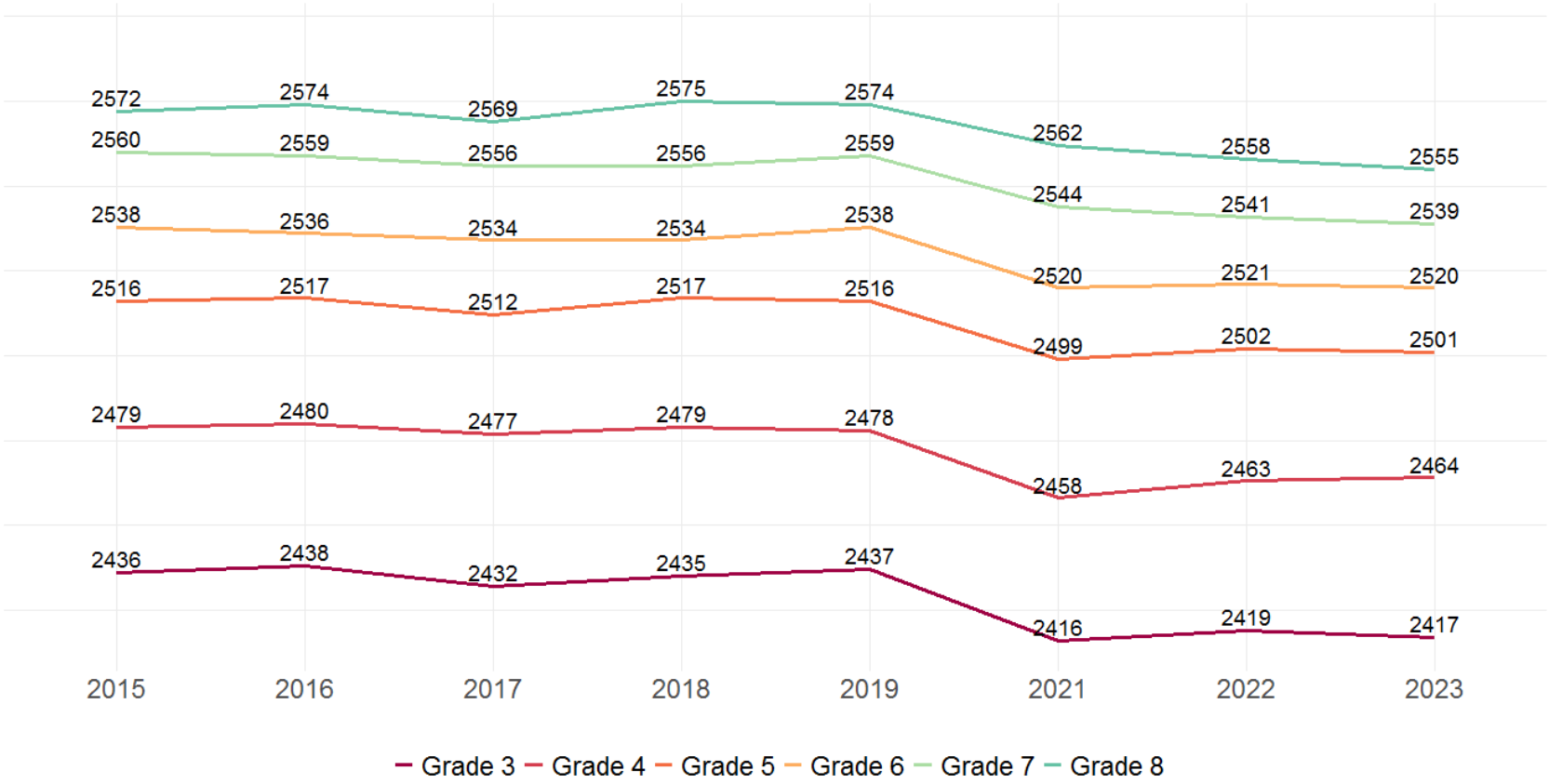
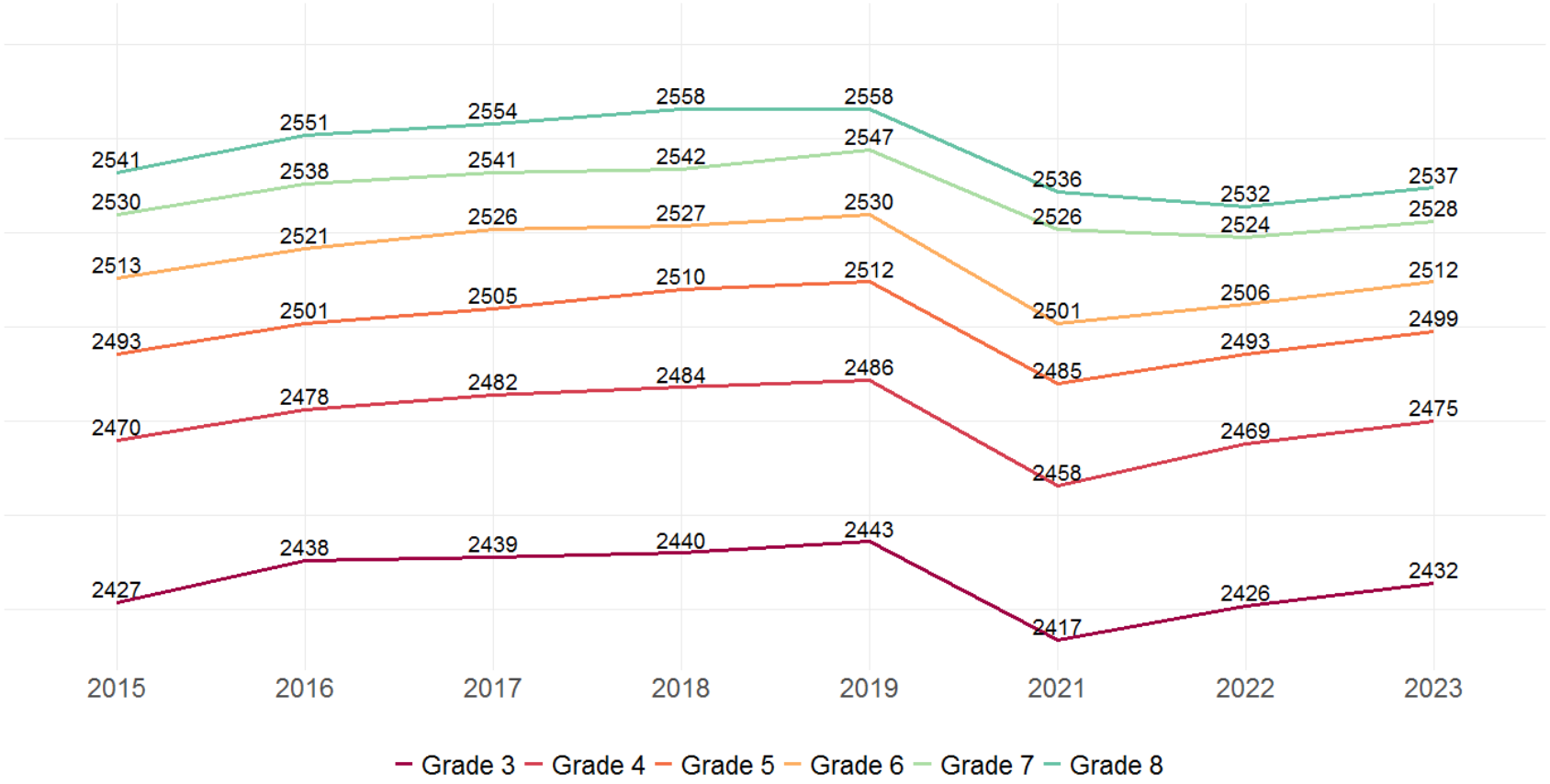


Figure 4. Mathematics Average Scale Score Across Years



Because the precision of scores in each claim 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, considering the standard error of measurement (SEM) of the claim score: (1) Below Standard, (2) At/Near Standard, or (3) Above Standard. Tables 21 and 22 present the distribution of performance categories for each claim. The number of claims is three in both ELA/L and mathematics, combining claims 2 and 4.

Table 21. ELA/L Percentage of Students in Performance Categories by Claim

Grade	Performance Category	Claim 1 Reading	Claims 2 and 4: Writing and Research	Claim 3 Listening
3	Below	33	37	19
	At/Near	42	40	61
	Above	26	23	20
4	Below	30	34	19
	At/Near	44	42	58
	Above	26	24	23
5	Below	29	33	19
	At/Near	41	38	57
	Above	30	30	23
6	Below	32	32	20
	At/Near	44	44	61
	Above	24	23	19
7	Below	28	32	22
	At/Near	46	44	60
	Above	26	24	18
8	Below	30	32	19
	At/Near	43	44	59
	Above	26	23	22

Table 22. Mathematics Percentage of Students in Performance Categories by Claim

Grade	Performance Category	Claim 1	Claims 2 and 4	Claim 3
3	Below	35	29	27
	At/Near	30	42	45
	Above	35	29	28
4	Below	37	33	32
	At/Near	28	41	40
	Above	35	26	28
5	Below	42	36	35
	At/Near	28	41	44
	Above	30	22	21
6	Below	44	37	33
	At/Near	31	42	47
	Above	26	21	21
7	Below	45	36	32
	At/Near	28	42	47
	Above	27	22	21
8	Below	47	35	37
	At/Near	28	43	45
	Above	25	22	18

Legned.
Claim 1: Concepts and Procedures;
Claims 2 and 4: Problem Solving and Modeling and Data Analysis;
Claim 3: Communicating Reasoning

3.3 Distribution of Student Ability and Item Difficulty

Figures 5–10 display the empirical distribution of the Connecticut student scale scores in the 2022–2023 test administration and the distribution of the administered summative item difficulty parameters for each grade for overall and by claim. For overall, the student ability distribution is to the left in all grades and subjects, a pattern more pronounced in the mathematics upper grades, indicating that the pool includes more difficult items than the ability of students in the tested population. The pool includes difficult items to accurately measure high-performing students but needs additional easy items to better measure low-performing students. At the claim, the student ability distribution is shifted to the left for all claims except for claims 2 and 4 in grades 4–5 and claim 3 in grades 4–8 in ELA/L. In mathematics, the student ability distribution is shifted to the left for all claims except for claim 1 in grades 3–4. The Smarter Balanced Assessment Consortium plans to add additional easy items to the pool and to augment the pool in proportion to the test blueprint constraints (e.g., content, Depth of Knowledge [DOK], item type, and item difficulties) to better measure low-performing students.

Figure 5. Student Ability—Item Difficulty Distribution for ELA/L

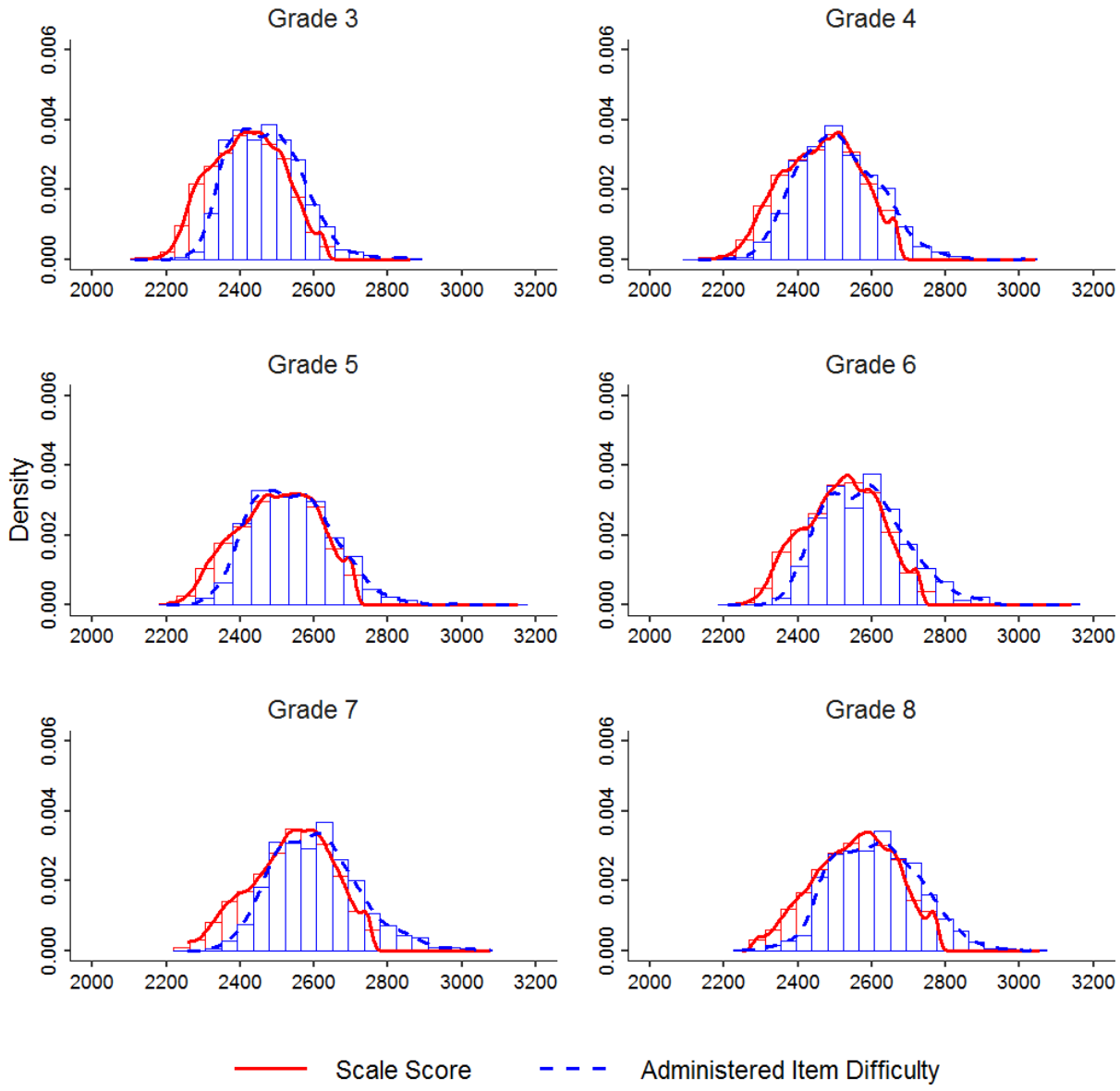


Figure 6. Student Ability—Item Difficulty Distribution by Claim: ELA/L (Grades 3–5)

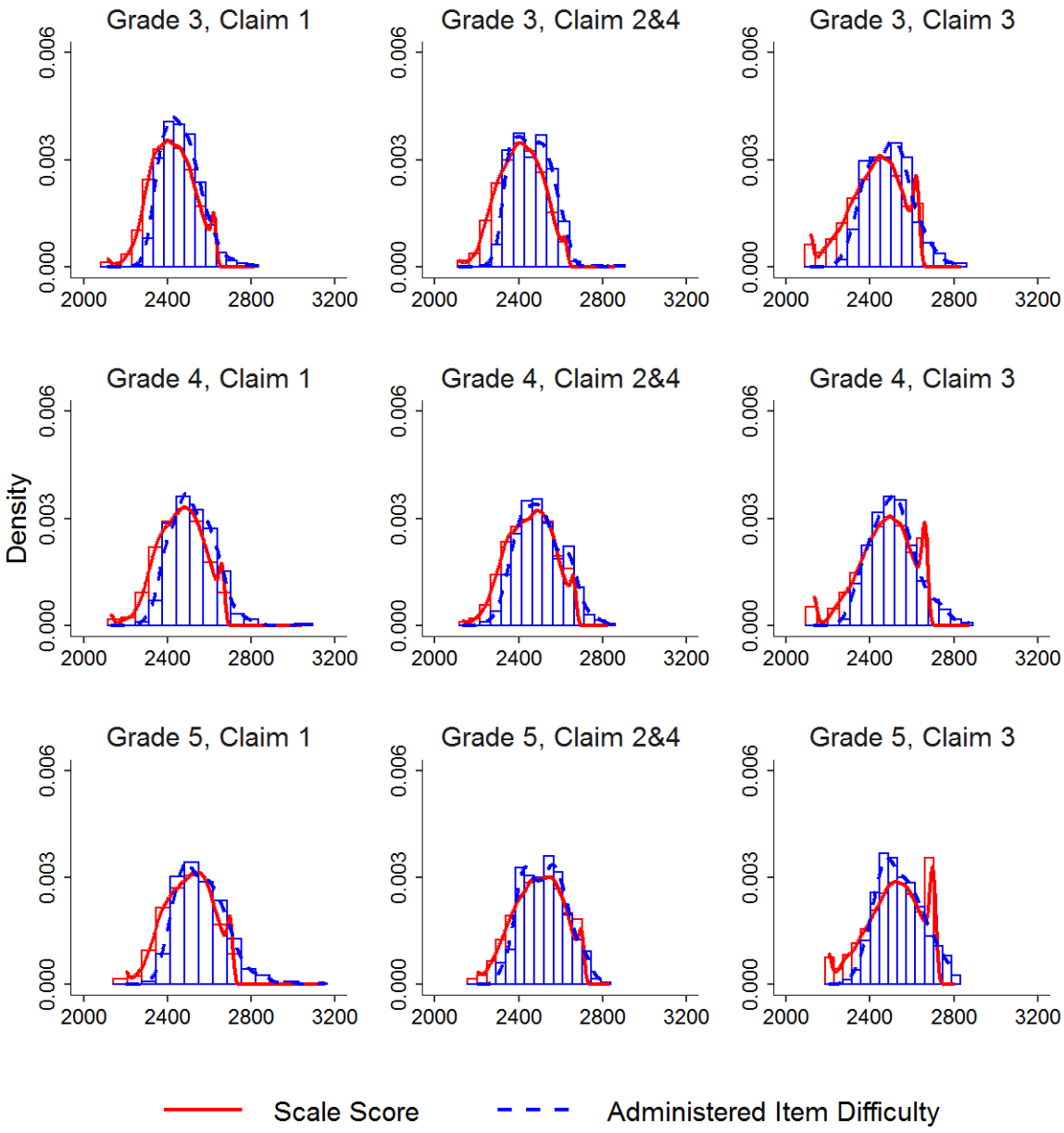


Figure 7. Student Ability—Item Difficulty Distribution by Claim: ELA/L (Grades 6–8)

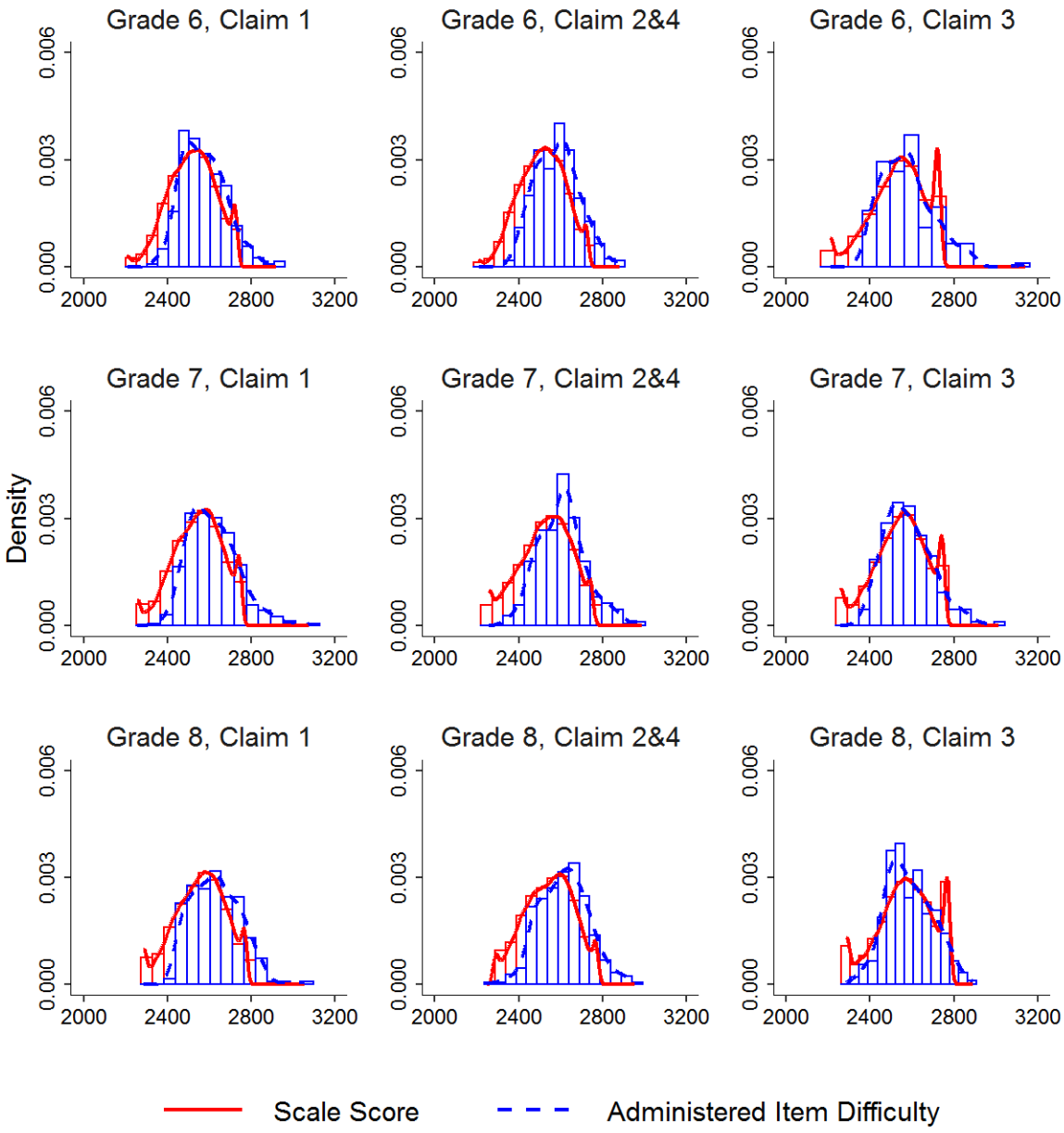


Figure 8. Student Ability—Item Difficulty Distribution for Mathematics

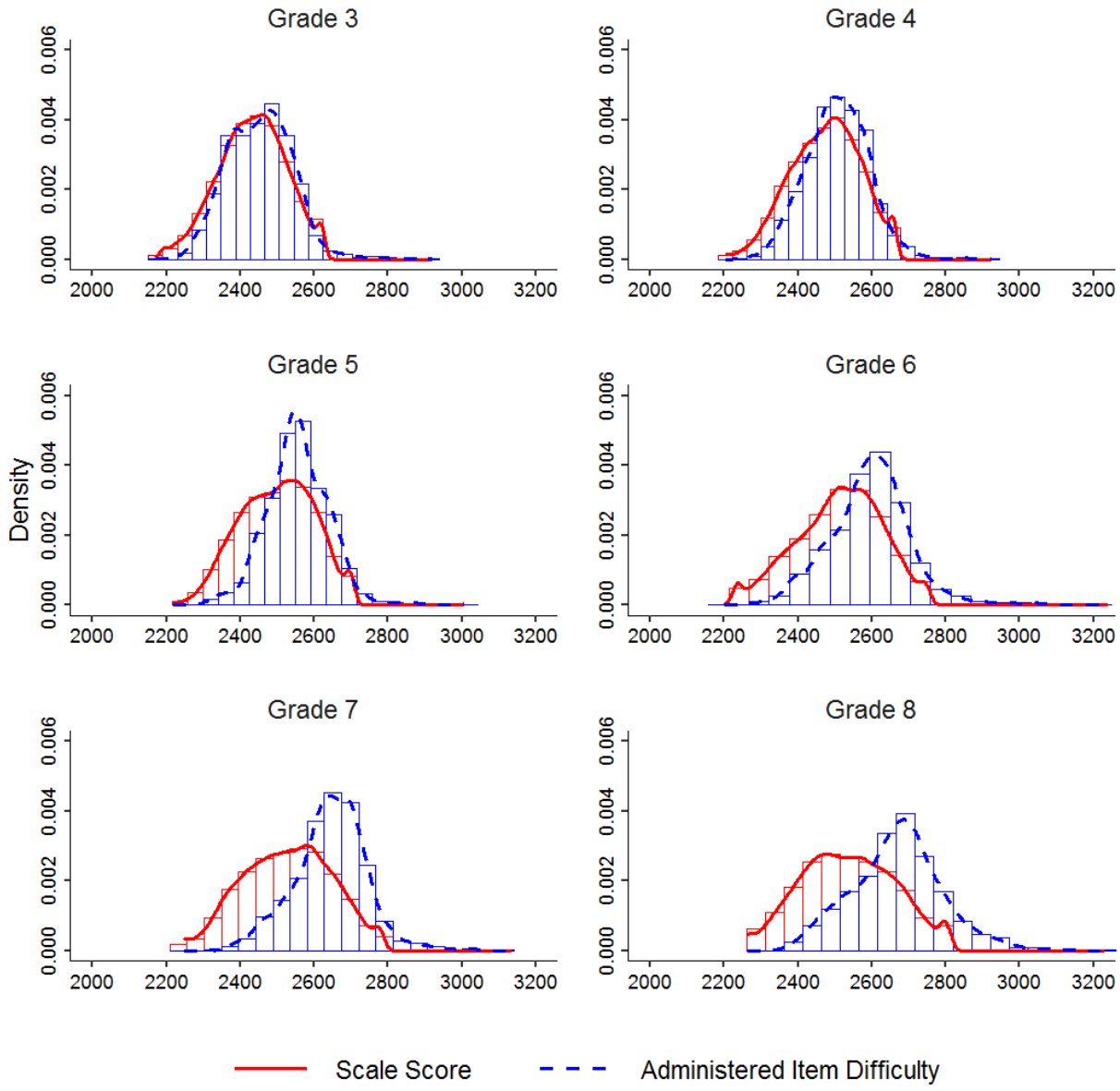


Figure 9. Student Ability—Item Difficulty Distribution by Claim: Mathematics (Grades 3–5)

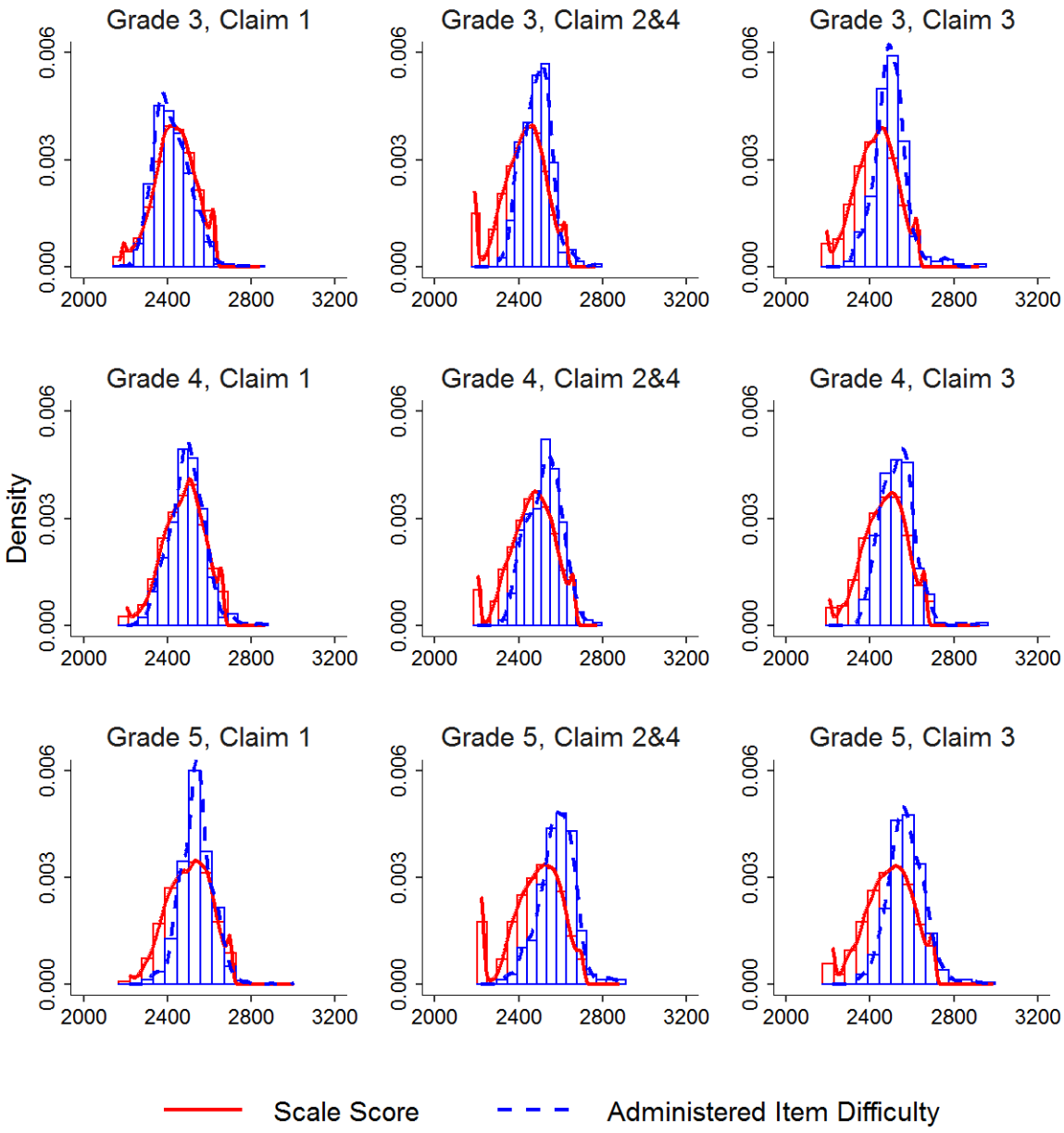
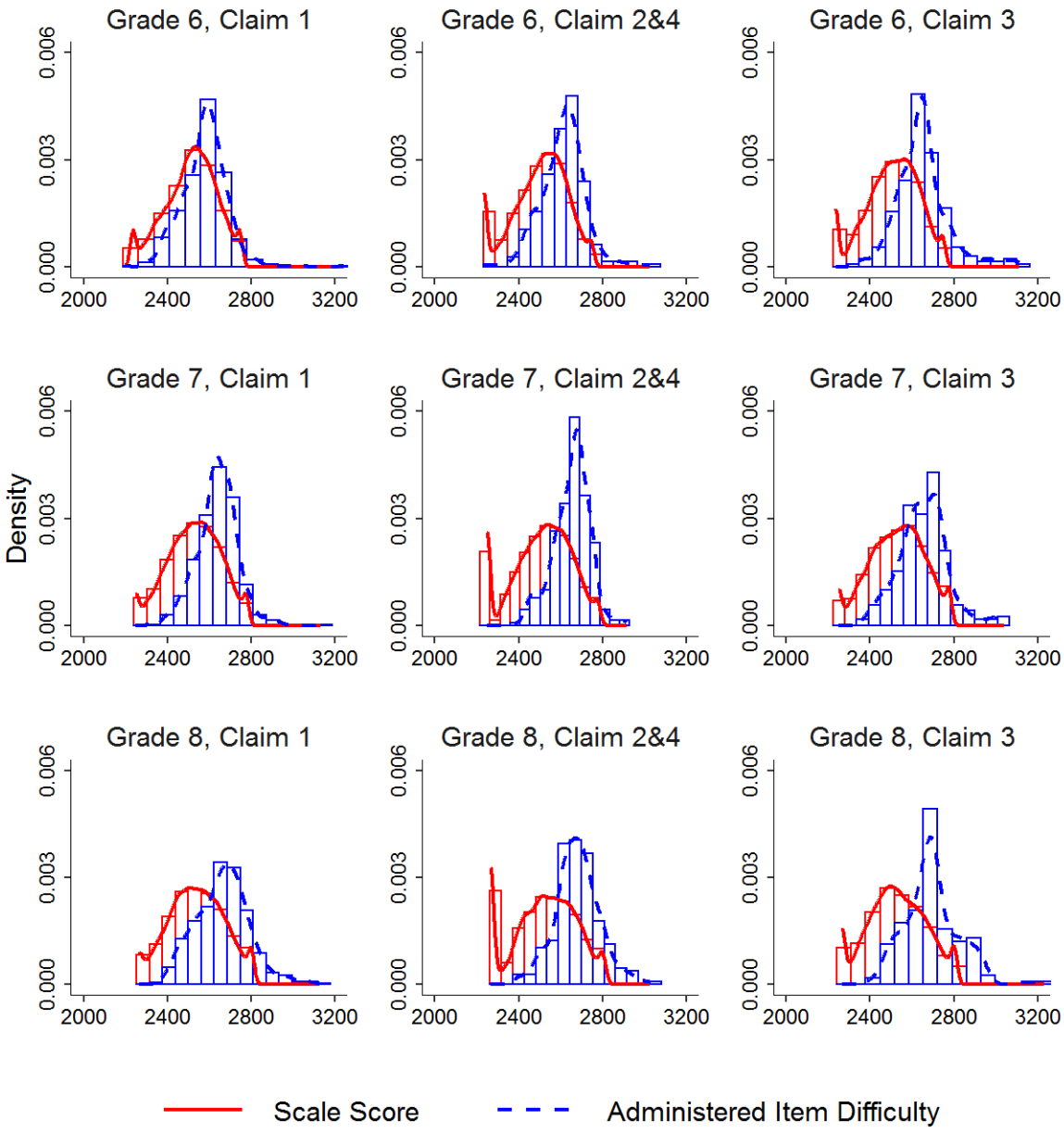


Figure 10. Student Ability—Item Difficulty Distribution by Claim: Mathematics (Grades 6–8)



4. VALIDITY

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

Validity evidence provided in this chapter is as follows:

- Test content
- Internal structure

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 inter-correlations among claim scores.

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

4.1 Evidence on Test Content

The Smarter Balanced summative assessment includes two components: the computer-adaptive test (CAT) and the performance task (PT). For the CAT, each student receives a different set of items adapted to his or her ability. For the PT, each student is administered with 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 specify a range of items to be administered in each claim, content domain/standards, and/or targets. Moreover, blueprints constrain the Depth of Knowledge (DOK) and item and passage types. For DOK constraints, the Smarter Balanced blueprint specifies either the minimum or maximum number of items, not both the minimum and maximum. 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 English language arts/literacy (ELA/L), the blueprints also specify the number of passages in reading (claim 1) and listening (claim 3) claims.

Tables 23 and 24 present the percentages of tests aligned with the ELA/L test blueprint constraints for items in claims, targets, DOK, and passages in claims 1 and 3. For the passage constraints, four passages in claim 1 reading and three to four passages in claim 3 listening are required. The composition of four reading passages in claim 1 is two literary text passages (one long and one short passage) and two informational text passages (one long and one short passage) in grades 3–5 and one literary text passage (long passage) and three informational text passages (one long and two short passages) in grades 6–8.

In ELA/L, all tests met the blueprint requirements except some targets in claim 1, which administered a few items more or less than the item requirement. The violations in claim 1 reading targets appeared in all grades due to the uneven distribution of items across targets and DOKs within and across passages.

Tables 25 and 26 provide the percentages of tests aligned with the test blueprint constraints for the mathematics CAT. In mathematics, all tests met all blueprint requirements, except for a few tests that had blueprint violations due to the application of pool filters limiting the item pool. Pool filters, such as using an alternative language like Spanish or only items with illustration or language glossaries, can result in an accommodated CAT item pool that is too limited to meet all test blueprint requirements, especially if multiple pool filters are employed on the same test.

Table 23. Percentage of ELA/L Delivered Tests Meeting Blueprint Requirements for Each Claim and Target (Grades 3–5)

Claim	Content Category/Target	Required Items/Passages	% BP Match		
			Grade 3	Grade 4	Grade 5
1	**Literary Text**	7–8	100	100	100
	Target 2: Central Ideas	1–2	100	100	100
	Target 4: Reasoning and Evidence	1–2	100	100	100
	Targets 1, 3, 5, 6, and 7	3–6	100	100	100
	Target 2 or 4 Short Text	0–1	100	100	100
	Long Literary Text Passage	≥ 1	100	100	100
	Short Literary Text Passage	≤ 2	100	100	100
	Informational Text	7–8	100	100	100
	Target 9: Central Ideas	1–2	100	99	100
	Target 11: Reasoning and Evidence	1–2	100	100	100
	Targets 8, 10, 12, 13, and 14	3–6	100	100	100
	Target 9 or 11 Short Text	0–1	100	100	100
	Long Informational Text Passage	≥ 1	100	100	100
	Short Informational Text Passage	≤ 2	100	100	100
	DOK 2	≥ 7	100	100	100
	DOK 3 or Higher	≥ 2	100	100	100
2	**Writing**	10	100	100	100
	Target 1, 3, or 6: Organization/Purpose [a]	3	100	100	100
	Target 1, 3, or 6: Evidence/Elaboration [a]				
	Target 8: Language and Vocabulary Use	2	100	100	100

	Target 9: Edit/Clarify	5	100	100	100
	DOK 2	≥ 4	100	100	100
	DOK 3 or 4	1	100	100	100
	Brief-Write	1	100	100	100
3	**Listening**	8–9	100	100	100
	Target 4: Listen/Interpret	8–9	100	100	100
	DOK 2 or Higher	≥ 3	100	100	100
	Listening Passage	3–4	100	100	100
4	**Research** Target 2: Interpret and Integrate Information Target 3: Analyze Information/Sources Target 4: Use Evidence	6	100	100	100

[a] Each student will receive a total of three items, with at least one item in Organization/Purpose and at least one item in Evidence/Elaboration.

Table 24. Percentage of ELA/L Delivered Tests Meeting Blueprint Requirements for Each Claim and Target (Grades 6–8)

Claim	Content Category/Target	Required Items/Passages	% BP Match		
			Grade 6	Grade 7	Grade 8
1	**Literary Text**	4–7	100	100	100
	Target 2: Central Ideas	1	99	100	98
	Target 4: Reasoning and Evidence	1	100	100	100
	Targets 1, 3, 5, 6, and 7	2–5	100	100	100
	Target 2 or 4 Short Text	0–1	100	100	100
	Long Literary Text Passage	≥ 1	100	100	100
	Informational Text	10–12[a]	100	100	100
	Targets 9 and 11	2–5	100	100	99
	Targets 8, 10, 12, 13, and 14	7–10	100	100	99
	Target 9 or 11 Short Text	0–1	100	100	100
	Long Informational Text Passage	≥ 1	100	100	100
	Short Informational Text Passage	≤ 2	100	100	100
	DOK 1	≤ 5	100	100	100
	DOK 3 or Higher	≥ 2	100	100	100
2	**Writing**	10	100	100	100
	Target 1, 3, or 6: Organization/Purpose[b]	3	100	100	100
	Target 1, 3, or 6: Evidence/Elaboration[b]		100	100	100
	Target 8: Language and Vocabulary Use	2	100	100	100
	Target 9: Edit/Clarify	5	100	100	100
	DOK 2	≥ 4	100	100	100
	DOK 3 or 4	1	100	100	100
	Brief-Write	1	100	100	100
3	**Listening**	8–9	100	100	100
	Target 4: Listen/Interpret	8–9	100	100	100
	DOK 2 or Higher	≥ 3	100	100	100
	Listening Passage	3–4	100	100	100
4	**Research**	6	100	100	100
	Target 2: Analyze/Integrate Information Target 3: Evaluate Information/Sources Target 4: Use Evidence	6	100	100	100

[a] Required items for Informational Text are 10–12 in grades 6 and 7, and 12 in grade 8.
[b] Each student will receive a total of three items, with at least one item in Organization/Purpose and at least one item in Evidence/Elaboration.

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

Claim	Content Domain	Grade 3		Grade 4		Grade 5	
		Required Items	% BP Match	Required Items	% BP Match	Required Items	% BP Match
1	Overall	17–20	100	17–20	100	17–20	100
	DOK 2 or Higher	≥ 7	100	≥ 7	100	≥ 7	100
	Priority Cluster	13–15	100				
	Targets B, C, G, I	5–6	100				
	Targets D, F	5–6	100				
	Target A	2–3	100				
	Supporting Cluster	4–5	100				
	Targets E, J, K	3–4	100				
	Target H	1	100				
	Priority Cluster			13–15	100		
	Targets A, E, F			8–9	100		
	Target G			2–3	100		
	Target D			1–2	100		
	Target H			1	100		
	Supporting Cluster			4–5	100		
	Targets I, K			2–3	100		
	Targets B, C, J			1	100		
	Target L			1	100		
	Priority Cluster					13–15	100
	Targets E, I					5–6	100
	Target F					4–5	100
	Targets C, D					3–4	100
	Supporting Cluster					4–5	100
	Targets J, K					2–3	100
	Targets A, B, G, H					2	100
2 and 4	Overall	6	100	6	100	6	100
	DOK 3 or Higher	≥ 2	100	≥ 2	100	≥ 2	100
	2. Target A	2	100	2	100	2	100
	2. Targets B, C, D	1	100	1	100	1	100
	4. Targets A, D	1	100	1	100	1	100
	4. Targets B, E	1	100	1	100	1	100
	4. Targets C, F	1	100	1	100	1	100
3	Overall	8	100	8	100	8	100
	DOK 3 or Higher	≥ 2	100	≥ 2	100	≥ 2	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

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

Claim	Content Domain	Grade 6		Grade 7		Grade 8	
		Required Items	% BP Match	Required Items	% BP Match	Required Items	% BP Match
1	Overall	16–20	100	16–20	100	16–20	100
	DOK 2 or Higher	≥ 7	100	≥ 7	100	≥ 7	
	Priority Cluster	12–15	100				
	Targets E, F	5–6	100				
	Target A	3–4	100				
	Targets B, G	2	100				
	Target D	2	100				
	Supporting Cluster	4–5	100				
	Targets C, H, I, J	4–5	100				
	Priority Cluster			12–15	100		
	Targets A, D			8–9	100		
	Targets B, C			5–6	98		
	Supporting Cluster			4–5	100		
	Targets E, F			2–3	100		
	Targets G, H, I			1–2	100		
	Priority Cluster					12–15	100
	Targets C, D					5–6	99
	Targets B, E, G					5–6	100
	Targets F, H					2–3	100
	Supporting Cluster					4–5	100
	Targets A, I, J					4–5	100
2 and 4	Overall	6	100	6	100	6	100
	DOK 3 or Higher	≥ 2	100	≥ 2	100	≥ 2	100
	2. Target A	2	100	2	100	2	100
	2. Targets B, C, D	1	100	1	100	1	100
	4. Targets A, D	1	100	1	100	1	100
	4. Targets B, E	1	100	1	100	1	100
	4. Targets C, F	1	100	1	100	1	100
3	Overall	8	100	8	100	8	100
	DOK 3 or Higher	≥ 2	100	≥ 2	100	≥ 2	100
	Targets A, D	3	100	3	100	3	100
	Targets B, E	3	100	3	100	3	100
	Targets C, F, G	2	100	2	100	2	100

Table 27 summarizes the target coverage, the average, and the range of the numbers of unique targets administered in each delivered CAT test by claim. Because 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 CAT Components

Grade	Total Targets in Blueprint				Mean				Range (Minimum – Maximum)			
	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4
ELA/L												
3	14	5	1	3	11	5	1	3	8–14	4–5	1–1	3–3
4	14	5	1	3	12	5	1	3	8–14	4–5	1–1	3–3
5	14	5	1	3	12	5	1	3	7–14	4–5	1–1	3–3
6	14	5	1	3	11	5	1	3	9–11	4–5	1–1	3–3
7	14	5	1	3	11	5	1	3	9–11	4–5	1–1	3–3
8	14	5	1	3	10	5	1	3	8–11	4–5	1–1	3–3
Mathematics												
3	11	4	6	6	11	2	5	3	9–11	2–2	3–6	3–4
4	12	4	6	6	10	2	5	3	9–10	2–2	3–6	3–3
5	11	4	6	6	9	2	5	3	8–9	2–2	3–6	2–3
6	10	4	7	6	10	2	5	3	9–10	2–2	3–6	3–3
7	9	3	7	6	8	2	5	3	5–8	1–2	3–6	2–4
8	10	4	7	6	10	2	5	3	9–10	2–2	3–6	3–3

An adaptive testing algorithm constructs a test form unique to each student, targeting the student's level of ability and meeting the test blueprints. Consequently, the test forms will not be statistically parallel (e.g., equal test difficulty) across individual students, but test scores from the individual tests are comparable since all test forms measure the same content, albeit with a different set of test items. Although each form is unique with respect to its items, all forms align with the same curricular expectations outlined in the test blueprints.

4.2 Evidence on Internal Structure

The measurement model used in the Smarter Balanced assessments assumes a single underlying latent trait in student ability estimates, which supports the reporting of a single total ability score. During the test construction phase, the test blueprint was designed to cover multiple claims under each subject. The item selection algorithm prioritizes blueprint matching to ensure that each test contains an appropriate combination of items from each claim. Assessing the relationship between these different claim scores is a measure of internal validity according to the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014). The presence of high correlations among claim scores is evidence that the Smarter Balanced assessment measures a single underlying ability and the claim scores are related to each other.

The correlations among claim 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 claim scores could be measured with perfect

reliability, corrected (adjusted) for measurement error estimates. The observed correlation between two claim scores with measurement errors can be corrected for attenuation as $r_{x'y'} = \frac{r_{xy}}{\sqrt{r_{xx} \times r_{yy}}}$, 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 claim 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 claims 2 and 4 and claim 3 scores are low. The low reliabilities are due to large standard errors among lower scores because of a shortage of easy items in the item pool.

Because the reliability for claim scores is low, the performance of all the claim scores is reported in three performance categories. The distribution of performance categories for each claim is provided in Tables 21 and 22, Section 3.2. Scale scores are not reported for claims.

Table 28. Correlations among Claim Scores for ELA/L

Grade	Claim	Observed and Disattenuated Correlations		
		Claim 1	Claims 2 & 4	Claim 3
3	Claim 1: Reading		0.97	0.97
	Claims 2 & 4: Writing & Research	0.79		0.98
	Claim 3: Listening	0.66	0.69	
4	Claim 1: Reading		0.97	0.99
	Claims 2 & 4: Writing & Research	0.77		0.99
	Claim 3: Listening	0.68	0.71	
5	Claim 1: Reading		0.98	1
	Claims 2 & 4: Writing & Research	0.81		1
	Claim 3: Listening	0.71	0.73	
6	Claim 1: Reading		0.97	1
	Claims 2 & 4: Writing & Research	0.77		1
	Claim 3: Listening	0.69	0.69	
7	Claim 1: Reading		1	1
	Claims 2 & 4: Writing & Research	0.78		1
	Claim 3: Listening	0.68	0.71	
8	Claim 1: Reading		1	1
	Claims 2 & 4: Writing & Research	0.79		1
	Claim 3: Listening	0.72	0.71	

Table 29. Correlations among Claim Scores for Mathematics

Grade	Claim	Observed and Disattenuated Correlations		
		Claim 1	**Claims 2 & 4**	**Claim 3**
3	Claim 1		1	0.97
	Claims 2 & 4	0.82		1
	Claim 3	0.81	0.76	
4	Claim 1		0.99	0.99
	Claims 2 & 4	0.83		1
	Claim 3	0.83	0.78	
5	Claim 1		1	0.99
	Claims 2 & 4	0.80		1
	Claim 3	0.80	0.75	
6	Claim 1		1	1
	Claims 2 & 4	0.83		1
	Claim 3	0.80	0.75	
7	Claim 1		1	1
	Claims 2 & 4	0.82		1
	Claim 3	0.81	0.75	
8	Claim 1		1	1
	Claims 2 & 4	0.80		1
	Claim 3	0.80	0.74	

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

5. RELIABILITY

According to the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014), reliability refers to the consistency of test scores across replications of a testing procedure. Reliability is related to the precision of measurement for a test and is evaluated, in part, in terms of the scores' standard error 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, and reliability coefficients are the correlation between scores on two equivalent forms of the test. Within the item response theory (IRT) framework, measurement error is conditional on ability and varies across the ability scale. The amount of precision in estimating achievement can be determined by the test information function, which describes the amount of information provided by the test at each score point along the ability continuum. Test information is the inverse of measurement error; the larger the measurement error, the less test information is being provided. In computer-adaptive tests (CATs), items administered vary among students, so the amount of measurement error differs from one test to another, which yields conditional standard error of measurement (CSEM).

The reliability evidence of the Smarter Balanced summative assessments is provided with marginal reliability, CSEM, and classification accuracy and consistency in each achievement level.

5.1 MARGINAL RELIABILITY

The marginal reliability was computed for the scale scores, considering the varying measurement errors across the ability range. Marginal reliability is a measure of the overall reliability of an assessment based on the average CSEM, 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 CSEM of the scale score for student *i,* and σ^2 is the variance of the scale score. The higher the reliability coefficient, the greater the precision of the test.

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

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

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

Table 30 presents the marginal reliability coefficients and the average CSEM 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
			ELA/L			
3	35,822	38–41	0.92	2417	96	27
4	35,921	38–41	0.92	2464	103	30
5	36,384	38–41	0.93	2501	109	29
6	36,506	38–42	0.91	2520	102	31
7	36,914	38–42	0.91	2539	110	33
8	38,070	38–42	0.91	2555	111	34
			Mathematics			
3	35,743	35–40	0.95	2432	92	20
4	35,860	35–40	0.95	2475	94	20
5	36,300	35–40	0.94	2499	100	24
6	36,358	34–39	0.94	2512	116	28
7	36,691	34–40	0.94	2528	119	30
8	37,757	34–40	0.93	2537	127	32

5.2 Standard Error Curves

Figures 11 and 12 present plots of the CSEM of scale scores across the range of abilities. The vertical lines indicate the three cut scores for the four achievement levels. For most of the ability range, the selection algorithm matched items to each student's ability and to the test blueprints with similar precision. Because the item pool is finite and has fewer items located at the extremes of the ability scale, the selection algorithm had to prioritize meeting blueprint requirements over matching items to ability level for those students with very high or very low abilities. This results in higher standard errors for students with very high or very low abilities compared to students with abilities around and between the three cut scores.

Given that classifying students into achievement levels, especially into proficient or not proficient levels based on the Level 3 cut, is a high-stakes decision for schools, it is important that ability levels near and between the cut scores are measured with as much precision as possible. This increased precision near and between the cut scores is achieved by having more items in the item pool for abilities across the middle of the scale, where the cut scores are located.

A consequence of the selection algorithm's prioritization of meeting blueprint requirements is that student ability near the low and high extremes of the scale is measured with relatively less precision. This produces the expected u-curve shape for the CSEM plots in Figures 11 and 12. An adaptive test with an infinitely large item pool and a selection algorithm that focused on maximizing information over blueprint requirements would produce flatter CSEM curves. The

Smarter Balanced assessments focus on increasing precision where it is most needed, ability scores near and in between the cut scores. It is worth noting that larger standard errors are observed at the lower ends of the score distribution, relative to the higher ends. This occurs because the item pools currently have a shortage of very easy items that are better targeted toward these lower-achieving students. Content experts use this information to consider how to further target and populate item pools.

Figure 11. Conditional Standard Errors of Measurement for ELA/L

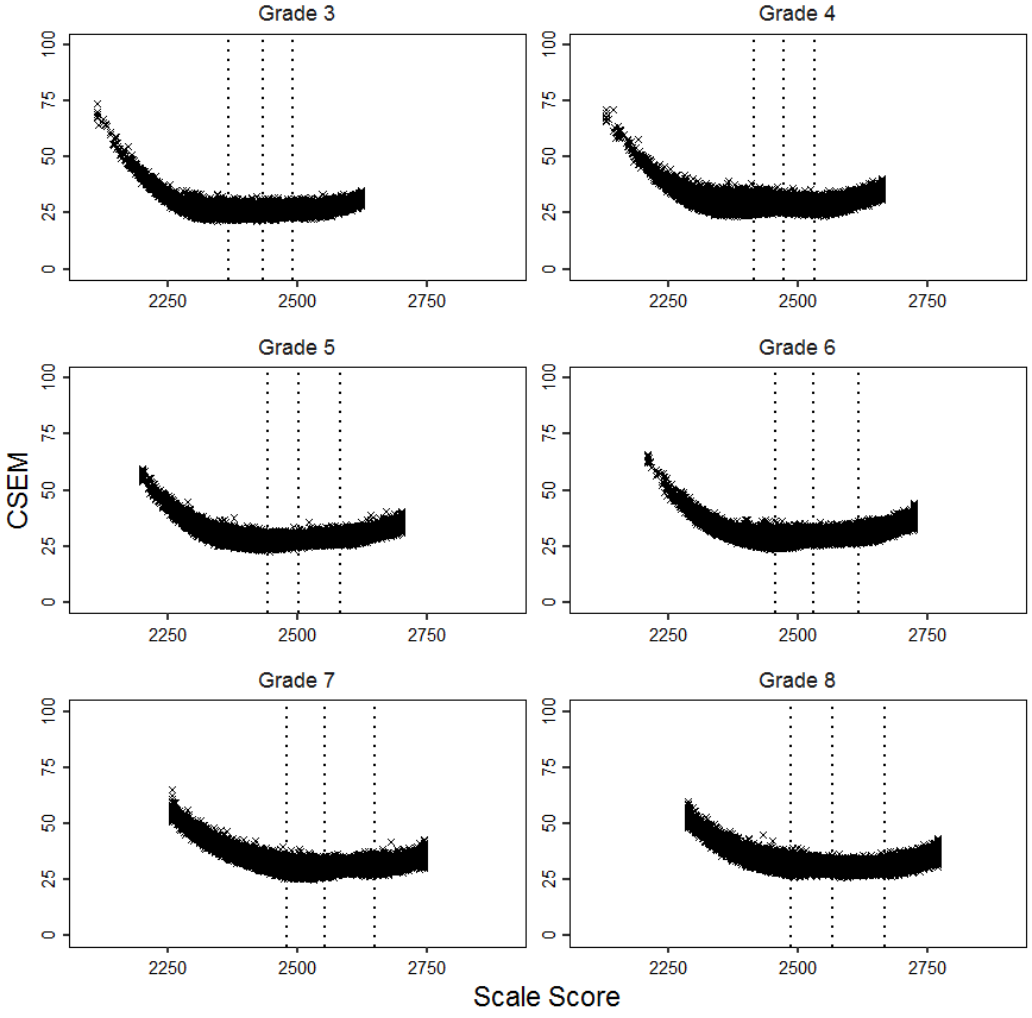
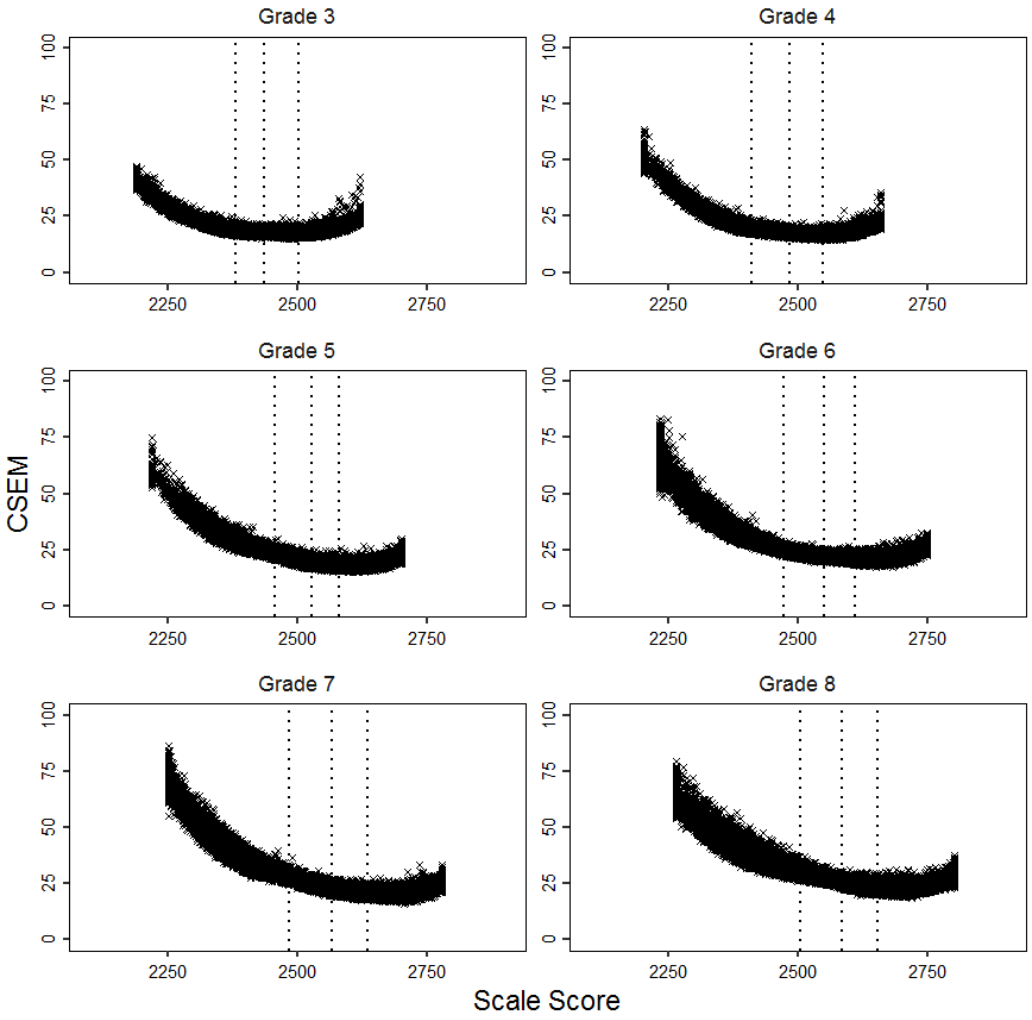


Figure 12. Conditional Standard Errors of Measurement for Mathematics



The CSEMs presented in Figures 11 and 12 are summarized in Tables 31 and 32. Table 31 provides the average CSEM for all scores and by achievement level. Table 32 presents the average CSEMs at each cut score and the difference in average CSEMs between two cut scores. As shown in Figures 11 and 12, the greatest average CSEM is in Level 1 in both English language arts/literacy (ELA/L) and mathematics. Average CSEMs at all cut scores are similar in ELA/L, but they are larger in Level 2 cut scores in mathematics.

Table 31. Average Conditional Standard Errors of Measurement by Achievement Level

Grade	Level 1	Level 2	Level 3	Level 4	Average CSEM
ELA/L					
3	29	25	26	27	27
4	31	29	29	30	30
5	30	27	28	31	29
6	32	29	30	33	31
7	38	30	31	32	33
8	39	31	30	33	34
Mathematics					
3	24	18	18	19	20
4	25	18	17	18	20
5	30	21	19	19	24
6	37	23	22	22	28
7	39	25	22	21	30
8	40	28	24	24	32

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

Grade	L2 Cut	L3 Cut	L4 Cut	|L2-L3|	|L3-L4|	|L2-L4|
ELA/L						
3	25	25	26	0	1	1
4	28	29	28	0	0	0
5	27	28	29	1	1	2
6	28	29	30	2	1	3
7	30	30	31	0	1	1
8	31	30	31	1	1	1
Mathematics						
3	19	18	17	1	0	1
4	19	17	17	2	1	3
5	23	19	18	4	1	5
6	25	22	21	3	1	5
7	28	23	21	5	2	7
8	30	26	23	5	3	8

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, & NCME, 2014). The indices 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 computer-adaptive test (CAT), because the adaptive testing algorithm constructs a test form unique to each student, the classification indices 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 taken and the classifications that would be made based on 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, 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 a SEM of $se(\hat{\theta}_i)$, and the estimated ability is distributed, as $\hat{\theta}_i \sim N\left(\theta_i, se^2(\hat{\theta}_i)\right)$, assuming a normal distribution, where θ_i is the unknown true ability of the *i*th student and Φ the cumulative distribution function of the standard normal distribution. 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 $\hat{\theta}_i \sim N\left(\theta_i, se^2(\hat{\theta}_i)\right)$, the above probabilities can be estimated 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, a probability of being at or above the cut point is an estimate of the chance that this student is misclassified as below the cut, and one minus that probability is the estimate of the chance that the student is correctly classified as below the cut score. Using this logic, the various classification probabilities can be defined.

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 $\mathbf{z}_i = \left(z_{i1},\cdots,z_{iJ}\right)$ and item parameters $\mathbf{b} = \left(\mathbf{b}_1,\cdots,\mathbf{b}_J\right)$ and 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; $\mathbf{b}_j = \left(a_j,b_j,c_j\right)$ 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 two-parameter logistic [2PL] models, $c_j = 0$), and D is 1.7 for non-Rasch models and 1 for Rasch model.

Classification Accuracy

Using p_{il}, a $L \times L$ table can be constructed 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. Because classifying students as proficient or not proficient is such a high stakes decision, classification accuracy is also considered at the proficiency level by repeating the process for overall classification accuracy of achievement levels but with the four achievement levels collapsed into two proficiency categories: proficient (achievement levels 3 and 4) and not proficient (achievement levels 1 and 2).

Classification Consistency

Using p_{il}, which is similar to accuracy, another $L \times L$ table can be constructed by assuming the test is administered twice independently to the same student group

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

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 the 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

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

As with classification accuracy, classification consistency is also considered at the proficiency level by repeating the process for overall classification consistency of achievement levels but with the four achievement levels collapsed into two proficiency categories: proficient (achievement levels 3 and 4) and not proficient (achievement levels 1 and 2).

The analysis of the classification index is performed based on overall scale scores. Table 33 provides the proportion of classification accuracy and consistency for overall, by achievement level, and at proficiency cut score.

The overall classification index ranged from 78% to 85% for the accuracy and from 71% to 79% 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 fact that the intervals used to compute the classification probabilities for students in L1 and L4 $[-\infty, \text{L2 cut; L4 cut}, \infty]$ are wider than the intervals used to compute the classification probabilities for students in L2 and L3 [L2 cut, L3 cut; L3 cut, L4 cut]. The misclassification probability tends to be higher for narrower intervals. Classification accuracy and classification consistency at the proficiency cut scores were high, ranging from 91% to 94% for accuracy and from 88% to 92% for consistency.

Accuracy of classifications is higher than the consistency of classifications in all achievement levels. The accuracy is higher than the consistency because the accuracy is based on one test with a measurement error and the true score while the consistency is based on two tests with measurement errors. The classification indices by subgroup are provided in Appendix C.

Table 33. Classification Accuracy and Consistency

Grade	Achievement Level	ELA/L		Mathematics	
		% Accuracy	% Consistency	% Accuracy	% Consistency
3	Overall	80	72	83	76
	L1	91	86	89	85
	L2	69	58	73	62
	L3	65	53	78	71
	L4	87	82	90	86
	Proficiency Cut	92	89	94	92
4	Overall	78	71	85	79
	L1	91	86	91	87
	L2	60	48	80	72
	L3	62	52	79	71
	L4	88	82	91	86
	Proficiency Cut	92	89	94	92
5	Overall	80	73	84	77
	L1	91	86	91	88
	L2	64	52	76	67
	L3	72	62	71	60
	L4	87	81	90	86
	Proficiency Cut	93	90	93	91
6	Overall	79	71	84	77
	L1	91	85	93	89
	L2	69	59	77	68
	L3	72	63	70	60
	L4	84	75	90	84
	Proficiency Cut	91	88	93	89
7	Overall	79	71	85	78
	L1	91	85	92	89
	L2	67	56	76	66
	L3	75	66	74	65
	L4	84	74	91	86
	Proficiency Cut	91	88	93	89
8	Overall	79	71	84	77
	L1	89	84	92	88
	L2	68	58	72	61
	L3	76	68	71	60
	L4	84	75	91	86
	Proficiency Cut	92	88	93	90

5.4 Reliability for Subgroups

Tables 34 through 39 present the marginal reliability coefficients by the subgroups. The reliability coefficients are similar across subgroups, but somewhat lower for limited English proficiency (LEP) and the Individuals with Disabilities Education Act (IDEA) subgroups. A large percentage of students in these subgroups received Level 1 with large SEMs.

Table 34. Marginal Reliability Coefficients Overall and by Subgroup: ELA/L (Grades 3–4)

Subgroup	Grade 3				Grade 4			
	MR	SS	SD	CSEM	MR	SS	SD	CSEM
All Students	0.92	2417	96	27	0.92	2464	103	30
Female	0.92	2424	96	27	0.91	2469	101	30
Male	0.92	2411	96	27	0.92	2459	104	30
Black or African American	0.90	2372	86	28	0.90	2419	92	30
AmerIndian/Alaskan	0.92	2385	93	27	0.92	2430	102	29
Asian	0.92	2464	93	27	0.91	2515	101	30
Hispanic or Latino	0.90	2373	90	28	0.90	2416	97	30
Pacific Islander	0.93	2417	101	27	0.90	2476	91	29
White	0.90	2452	86	27	0.89	2500	91	30
Multi-Racial	0.92	2426	97	28	0.91	2481	99	30
LEP	0.88	2345	80	28	0.88	2380	88	31
IDEA	0.87	2347	81	29	0.88	2385	89	31

Table 35. Marginal Reliability Coefficients Overall and by Subgroup: ELA/L (Grades 5–6)

Subgroup	Grade 5				Grade 6			
	MR	SS	SD	CSEM	MR	SS	SD	CSEM
All Students	0.93	2501	109	29	0.91	2520	102	31
Female	0.92	2509	107	29	0.91	2528	101	31
Male	0.93	2493	109	29	0.91	2512	103	31
Black or African American	0.91	2448	98	29	0.89	2476	93	31
AmerIndian/Alaskan	0.91	2495	97	29	0.90	2480	92	30
Asian	0.91	2560	101	30	0.90	2580	97	31
Hispanic or Latino	0.92	2449	102	29	0.90	2475	97	31
Pacific Islander	0.92	2476	104	29	0.88	2492	103	35
White	0.91	2539	96	29	0.89	2554	91	31
Multi-Racial	0.92	2520	106	30	0.91	2532	101	30
LEP	0.87	2400	85	30	0.82	2413	76	32
IDEA	0.90	2413	95	30	0.86	2434	85	32

Table 36. Marginal Reliability Coefficients Overall and by Subgroup: ELA/L (Grades 7–8)

Subgroup	Grade 7				Grade 8			
	MR	SS	SD	CSEM	MR	SS	SD	CSEM
All Students	0.91	2539	110	33	0.91	2555	111	34
Female	0.91	2550	107	33	0.91	2569	108	33
Male	0.91	2529	112	33	0.91	2542	113	34
Black or African American	0.89	2495	101	34	0.88	2512	101	35
AmerIndian/Alaskan	0.88	2513	93	32	0.90	2529	103	33
Asian	0.90	2606	102	32	0.90	2627	101	32
Hispanic or Latino	0.90	2490	108	35	0.89	2503	106	35
Pacific Islander	0.89	2509	98	33	0.90	2539	104	33
White	0.89	2575	96	32	0.89	2592	99	32
Multi-Racial	0.91	2548	111	33	0.90	2569	110	35
LEP	0.78	2410	80	38	0.75	2417	76	38
IDEA	0.86	2447	96	37	0.84	2462	94	37

Table 37. Marginal Reliability Coefficients Overall and by Subgroup: Mathematics (Grades 3–4)

Subgroup	Grade 3				Grade 4			
	MR	SS	SD	CSEM	MR	SS	SD	CSEM
All Students	0.95	2432	92	20	0.95	2475	94	20
Female	0.95	2427	89	20	0.95	2469	89	20
Male	0.95	2436	94	20	0.96	2481	98	20
Black or African American	0.93	2380	84	22	0.93	2421	85	22
AmerIndian/Alaskan	0.93	2406	78	20	0.95	2439	96	22
Asian	0.95	2487	89	20	0.96	2538	92	19
Hispanic or Latino	0.94	2389	84	21	0.94	2431	87	21
Pacific Islander	0.96	2414	96	20	0.96	2493	95	19
White	0.94	2466	79	19	0.95	2510	81	19
Multi-Racial	0.95	2438	92	20	0.96	2486	93	20
LEP	0.93	2374	82	21	0.93	2412	83	22
IDEA	0.93	2359	88	23	0.93	2396	90	24

Table 38. Marginal Reliability Coefficients Overall and by Subgroup: Mathematics (Grades 5–6)

Subgroup	Grade 5				Grade 6			
	MR	SS	SD	CSEM	MR	SS	SD	CSEM
All Students	0.94	2499	100	24	0.94	2512	116	28
Female	0.94	2495	96	24	0.94	2508	113	28
Male	0.95	2504	103	24	0.94	2516	119	28
Black or African American	0.90	2439	87	27	0.91	2445	105	32
AmerIndian/Alaskan	0.94	2490	96	24	0.92	2466	108	31
Asian	0.95	2572	94	21	0.95	2599	111	25
Hispanic or Latino	0.92	2452	91	26	0.91	2458	108	32
Pacific Islander	0.93	2494	91	23	0.93	2483	113	31
White	0.94	2535	87	22	0.94	2554	98	25
Multi-Racial	0.95	2515	100	23	0.95	2522	117	27
LEP	0.87	2424	80	28	0.85	2407	93	37
IDEA	0.89	2417	90	29	0.88	2409	106	37

Table 39. Marginal Reliability Coefficients Overall and by Subgroup: Mathematics (Grades 7–8)

Subgroup	Grade 7				Grade 8			
	MR	SS	SD	CSEM	MR	SS	SD	CSEM
All Students	0.94	2528	119	30	0.93	2537	127	32
Female	0.94	2525	116	30	0.93	2538	122	32
Male	0.94	2531	122	30	0.94	2536	131	33
Black or African American	0.89	2463	101	34	0.88	2466	105	36
AmerIndian/Alaskan	0.91	2504	97	29	0.90	2495	112	35
Asian	0.96	2624	119	25	0.95	2642	122	27
Hispanic or Latino	0.90	2469	106	34	0.89	2474	109	37
Pacific Islander	0.90	2490	96	30	0.91	2536	101	31
White	0.94	2573	104	26	0.94	2584	113	29
Multi-Racial	0.94	2537	123	29	0.94	2548	129	32
LEP	0.79	2410	83	38	0.72	2409	81	43
IDEA	0.85	2426	99	38	0.84	2430	101	40

5.5 Reliability for Claim Scores

The marginal reliability coefficients and the measurement errors are also computed for the claim scores. In both ELA/L and mathematics, claims 2 and 4 are combined to generate a score. Because the precision of scores in claims is insufficient to report scores given a small number of items, the scores on each claim are reported using one of the three achievement categories, considering the SEM of the claim score: (1) Below Standard, (2) At/Near Standard, or (3) Above Standard. Tables 40 and 41 present the marginal reliability coefficients for each claim score in ELA/L and mathematics, respectively.

Table 40. Marginal Reliability Coefficients for Claim Scores in ELA/L

Grade	Claim	Number of Items Specified in Test Blueprint	Marginal Reliability	Scale Score Mean	Scale Score SD	Average CSEM
3	Claim 1: Reading	14–16	0.79	2421	103	47
	Claims 2 & 4: Writing & Research	16	0.84	2407	104	41
	Claim 3: Listening	8–9	0.59	2426	128	82
4	Claim 1: Reading	14–16	0.77	2463	112	53
	Claims 2 & 4: Writing & Research	16	0.82	2454	112	47
	Claim 3: Listening	8–9	0.61	2471	128	80
5	Claim 1: Reading	14–16	0.80	2502	115	52
	Claims 2 & 4: Writing & Research	16	0.85	2493	117	45
	Claim 3: Listening	8–9	0.62	2511	131	81
6	Claim 1: Reading	14–17	0.78	2519	113	53
	Claims 2 & 4: Writing & Research	16	0.80	2512	109	49
	Claim 3: Listening	8–9	0.58	2534	131	85
7	Claim 1: Reading	14–17	0.75	2545	117	58
	Claims 2 & 4: Writing & Research	16	0.81	2527	123	53
	Claim 3: Listening	8–9	0.62	2545	125	77
8	Claim 1: Reading	14–17	0.78	2556	120	56
	Claims 2 & 4: Writing & Research	16	0.80	2545	119	53
	Claim 3: Listening	8–9	0.63	2569	130	79

Table 41. Marginal Reliability Coefficients for Claim Scores in Mathematics

Grade	Claim	Number of Items Specified in Test Blueprint	Marginal Reliability	Scale Score Mean	Scale Score SD	Average CSEM
3	Claim 1	17–20	0.92	2434	98	28
	Claims 2 & 4	8–10	0.71	2424	105	56
	Claim 3	8–10	0.75	2426	101	50
4	Claim 1	17–20	0.92	2477	100	28
	Claims 2 & 4	8–10	0.76	2464	107	52
	Claim 3	8–10	0.77	2469	103	49
5	Claim 1	17–20	0.90	2503	105	33
	Claims 2 & 4	8–10	0.67	2483	121	70
	Claim 3	8–10	0.73	2490	115	60
6	Claim 1	16–20	0.90	2513	124	39
	Claims 2 & 4	8–10	0.72	2502	129	68
	Claim 3	8–10	0.66	2507	127	73
7	Claim 1	16–20	0.89	2527	127	42
	Claims 2 & 4	8–10	0.66	2513	138	80
	Claim 3	8–10	0.72	2528	131	70
8	Claim 1	16–20	0.89	2539	132	43
	Claims 2 & 4	8–10	0.61	2521	151	94
	Claim 3	8–10	0.68	2528	139	78

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

6. SCORING

The Smarter Balanced Assessment Consortium provided the vertically scaled item parameters by linking across all 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 claim. This section describes the rules used in generating scores, as well as the handscoring procedure.

6.1 Estimating Student Ability Using Maximum Likelihood Estimation

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

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

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

where the vector $\boldsymbol{b_i} = (b_{i,1}, \ldots, b_{i,m_i})$ 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*, and *k* indices the step of the item *i.*

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

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

in the case of items with two or more points,

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

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

Standard Error of Measurement

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

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

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

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

where m_i is the maximum possible score point (starting from 0) for the *i*th item, and D is the scale factor, 1.7. The SE is calculated based only on the answered items 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 Smarter Balanced Assessment Consortium. Table 42 presents the scaling constants for each subject for the theta-to-scale score linear transformation. Scale scores are rounded to an integer.

Table 42. Vertical Scaling Constants on the Reporting Metric

Subject	Grade	Slope (a)	Intercept (b)
ELA/L	3–8	85.8	2508.2

Mathematics	3–8	79.3	2514.9

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

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

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

The scale scores are mapped into four achievement levels using three achievement standards (i.e., cut scores). Table 43 provides three achievement standards for each grade and content area.

Table 43. 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

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 at the low and high ends of the ability range. The Smarter Balanced Assessment Consortium decided to truncate extreme unreliable student ability estimates. Table 44 presents the lowest obtainable theta score (LOT) or LOSS and the highest obtainable theta score (HOT) or HOSS. Estimated thetas lower than LOT or higher than HOT are truncated to the LOT and HOT values and are assigned LOSS and HOSS associated with the LOT and HOT. LOT and HOT were applied to all tests and all scores (total and claim scores). The standard errors for LOT and HOT are computed using the LOT and HOT ability estimates given the administered items.

Table 44. Lowest and Highest Obtainable Scores

Subject	Grade	Theta Metric		Scale Score Metric	
		LOT	HOT	LOSS	HOSS
ELA/L	3	−4.5941	1.3374	2114	2623
	4	−4.3962	1.8014	2131	2663
	5	−3.5763	2.2498	2201	2701
	6	−3.4785	2.5140	2210	2724
	7	−2.9114	2.7547	2258	2745
	8	−2.5677	3.0430	2288	2769
Mathematics	3	−4.1132	1.3335	2189	2621
	4	−3.9204	1.8191	2204	2659
	5	−3.7276	2.3290	2219	2700
	6	−3.5348	2.9455	2235	2748
	7	−3.3420	3.3238	2250	2778
	8	−3.1492	3.6254	2265	2802

6.4 Scoring All Correct and All Incorrect Cases

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

6.5 Rules for Calculating Strengths and Weaknesses for Claim Scores

In both English language arts/literacy (ELA/L) and mathematics, claim scores are computed for claim 1, claims 2 and 4 combined, and claim 3. For each claim score, three performance categories' relative strengths and weaknesses are produced. The difference between the proficiency cut score and the claim score plus or minus 1.5 times standard error of the claim is used to determine the relative strengths and weaknesses.

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 claim; 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 claim. HOSS and LOSS are automatically assigned to *Above Standard* and *Below Standard*, respectively.

6.6 Target Scores

The target-level reports are impossible to produce for a fixed-form test because the number of items included per target is too small 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 narrowly reflect the benchmark because they reflect only one or two ways of measuring the target. However, an adaptive test offers a tremendous opportunity for target-level data at the class-, school-, and district-area levels. With an adequate item pool, a class of 20 students might respond to 10 or 15 different items measuring any given target. Target scores are computed for attempted tests based on the responded items. Target scores are computed in each of the four claims in ELA/L and claim 1 for 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, the 2PL IRT model is used 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 GPCM, 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:

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

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 the individual student target scores for that target across all students in the aggregate unit.

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

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.

Direct reporting of the statistic $\bar{\delta}_{Tg}$ is not suggested. Instead reporting whether, in the aggregate, a group of students performs better, worse, or as expected on this target is

recommended. In some cases, insufficient information will be available, and that will be indicated, as well.

For target-level strengths and weaknesses, the following are reported:

- If $\bar{\delta}_{Tg} - se\left(\bar{\delta}_{Tg}\right) \geq 0.07$, then performance is better than on the overall test.
- If $\bar{\delta}_{Tg} + se\left(\bar{\delta}_{Tg}\right) \leq -0.07$, then performance is worse than on the overall test.
- Otherwise, performance is similar to performance on the overall test.
- If $se\left(\bar{\delta}_{Tg}\right) > 0.2$, data are insufficient.

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

By defining $p_{ij} = p(z_{ij} = 1)$, indicating 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 the 2PL IRT model is used to calculate the expected score on item *i* for student *j* with $\theta_{Level\ 3\ cut}$ as:

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

For items with two or more score points, using the GPCM, 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:

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

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

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

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 the individual student target scores for that target across all students in the aggregate unit.

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

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

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

Direct reporting of the statistic $\bar{\delta}_{Tg}$ is not suggested. Instead reporting whether, in the aggregate, a group of students performs better, worse, or as expected on this target is recommended. In some cases, insufficient information will be available, and that will be indicated, as well.

For target-level strengths and weaknesses, the following are reported:

- If $\bar{\delta}_{Tg} - se\left(\bar{\delta}_{Tg}\right) \geq 0.07$, then performance is *above* the Proficiency Standard.
- If $\bar{\delta}_{Tg} + se\left(\bar{\delta}_{Tg}\right) \leq -0.07$, then performance is *below* the Proficiency Standard.
- Otherwise, performance is *near* the Proficiency Standard.
- If $se\left(\bar{\delta}_{Tg}\right) > 0.2$, data are insufficient.

6.7 HANDSCORING

Constructed-response short-answer (SA) items in both English language arts/literacy (ELA/L) and mathematics for the summative assessments administered by Cambium Assessment, Inc. (CAI) are routed to Measurement Incorporated (MI) for scoring. MI provides handscoring using human raters. The methods used for handscoring and the results are described in the following sections.

For the 2022–2023 summative operational item pool, there were a total of 329 SA items in ELA/L and 312 SA items in mathematics. Table 45 shows the number of SA items by grade and subject.

Table 45. Number of Handscored Items in 2022–2023 Smarter Balanced Summative Item Pool, by Grade and Subject

Grade	ELA/L	Mathematics
3	52	51
4	58	52
5	66	76
6	52	51
7	52	34

Grade	ELA/L	Mathematics
8	49	48
Total	**329**	**312**

All guidelines for handscoring responses were specified by Smarter Balanced. Outlined in the next section is the handscoring process MI followed in spring 2023 in accordance with the Smarter Balanced guidelines. This process applied to the scoring of all constructed responses SA items for ELA/L and mathematics.

6.7.1 Rater Selection

MI has developed a pool of more than 3,000 raters experienced in scoring the Smarter Balanced assessments. MI first recruited qualified raters who had experience scoring these assessments. Recent advancements in rater evaluation practices have allowed MI to estimate rater accuracy parameters for experienced Smarter Balanced raters; these data were used to recruit the most historically accurate raters. Once recruited, experienced raters were assigned to the content area and grade band(s) with which they were most experienced.

To supplement this pool, MI also recruited raters with experience successfully scoring other large-scale assessments. MI assigned those raters to the grade level, subject area, and item type for which they were most qualified based on their performance on similar projects. Returning raters were selected based on experience and performance, as well as attendance, punctuality, and cooperation with work procedures and MI policies. MI maintains evaluations and performance data for all staff who work on each scoring project in order to determine employment eligibility for future projects. Finally, MI targeted recruitment of new raters, as needed, in an effort to continue to identify talent across the country that will best fulfill the hand-scoring requirements.

All raters possessed, at a minimum, a four-year college degree. MI collected proof of degree for all raters as a condition of employment. All raters resided in the United States and properly completed Form I-9 to verify their identity and employment authorization. Raters' I-9 forms are retained on file as required by law and made available for inspection by authorized government officers as needed. MI is an equal-opportunity employer, and believes that a diverse work force is of the utmost importance. When hiring, MI strives to ensure the work force is diverse across age, ethnicity, gender, and other demographic groups.

In selecting team leaders who monitored the raters, MI scoring leadership reviewed records of all returning staff. They looked for people who were experienced team leaders with a history of good performance on previous projects, and they also considered raters who were recommended for promotion to the team leader position.

MI requires all handscoring project staff (scoring directors, team leaders, raters, and clerical staff) to sign a confidentiality/nondisclosure agreement before receiving any training or viewing any 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 Rater Training and Scoring

All raters hired to score the Smarter Balanced assessments were 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. Additionally, MI conducts an annual review of the rater agreement and scoring materials in order to inform the development of item-specific, supplemental training materials. Supplemental materials are developed each summer and implemented in the subsequent operational administration. These additional materials are developed with a focus on challenging areas identified during the previous operational administration, such as low validity or low inter-rater reliability (IRR) statistics for a specific item (or in some cases, for specific types of responses that scorers found difficult).

Supplemental materials are also created for newly operational items for which MI identifies a need for additional examples. For instance, MI may find an approach to a mathematics item that was not encountered during field testing but appears frequently during operational scoring, or an unusual but valid way to address a research prompt that is not reflected in the existing rubric. In these cases, MI provides examples of these specific approaches along with guidance on how to score them correctly. MI also supplements materials to provide raters with additional guidance for content-wide challenging spots or to help them more accurately identify responses that should be flagged as non-scorable.

Once hired, raters were assigned to a scoring group that corresponded to the subject/grade that they were deemed best suited to score (based on work history, results of the placement assessments, and performance on past scoring projects). Raters were trained to score a specific item group of either SA (research, brief write, reading, and mathematics) items. Within each item group, raters were divided into teams supervised by team leaders and a scoring director. Each scoring director, team leader, and rater was assigned a unique number for easy identification of their scoring work throughout the scoring session. The number of items an individual rater scores was minimized to allow the rater to quickly develop experience scoring responses to a given set of items.

All raters, regardless of experience, were required to train on all anchor and training sets. Following training, all raters were required to pass the qualification sets in order to prove that

they understood and could apply the criteria accurately. Until a rater had trained and qualified successfully, he or she was not permitted to score any student responses. MI carefully orchestrated training so that raters understood that all scoring decisions must be grounded in the training materials. In addition, raters learned how to navigate the anchor set, developed the knowledge and flexibility needed to evaluate or escalate a variety of responses, and retained the necessary consistency to score all responses accurately.

In order to begin working, all scoring personnel logged in to MI's secure Scoring Resource Center (SRC). SRC includes all online training modules, serves as the portal to MI's Virtual Scoring Center (VSC) interface, and maintains the data repository of all scoring reports used for rater monitoring. MI's training system (VSC Train) provides a remote, secure application for training both team leaders and raters. VSC Train provided each trainee with a training lesson for each item that allowed the trainee to complete the following steps:

1) Review the anchor set(s)
2) Score the practice set(s)
3) Review an annotated version of the practice set(s) after submitting scores
4) Score the qualification sets

Training design varied slightly depending on Smarter Balanced item type:

- ELA/L brief write, reading, and research SA: Raters trained and qualified on a baseline lesson within a specific grade band and target. Qualification on the baseline lesson qualified the rater to score all items in that grade band and target.
- Mathematics SA: Raters trained and qualified on baseline lessons within a specific grade band. Qualification on a baseline lesson qualified the rater to score that item and all items associated with it; for items with no associated items, training was for the specific item.

Rater training time varied by grade and content area. Training for SA brief write, reading, research, and mathematics items could typically be accomplished in one day. Raters generally worked 6.5 hours per day, excluding breaks. Evening shift raters worked 3.75 hours, excluding breaks.

In addition to item-specific information, a variety of substantive procedural and policy information was provided to each trainee during training. This included information about "alert" responses and non-scorable responses, as well as instructions for how to communicate with leadership during handscoring. This ensured that raters were fully prepared to handscore responses and were also aware of all responsibilities and scoring requirements before they were allowed to begin scoring.

Each trainee's practice and qualification results were reported to the team leaders and scoring director. Scoring leadership reviewed each trainee's results, paying particular attention to frequently mis-scored responses.

Following training, all training materials remained available to raters throughout scoring via the VSC Score Resource Library. This library included the item and rubric, the annotated anchor and practice sets, and any supplemental materials that were required to ensure accurate completion of the scoring effort.

When scoring, raters had access only to those items for which they had successfully trained and qualified. The handscoring system sorts individual student responses into small sets of 5–10, grouped by item. When a rater is qualified to score multiple items, this approach eases cognitive load by presenting the rater with a scoring set in which all responses relate to the same item.

Raters were trained to recognize non-scorable responses, and these responses were systematically routed to scoring supervisors for final condition-code assignment per Smarter Balanced requirements. For some item types, condition-code responses were scored by scoring leaders trained to specialize in the scoring of these types of responses.

An "alerts" procedure was explained to raters during training sessions, where raters are trained to recognize "alerts" in their various forms, including those for suicide, criminal activity, alcohol or drug use, extreme depression, violence, rape, sexual or physical abuse, self-harm, intent to harm others, and neglect.

Multiple strategies were employed to minimize rater bias during scoring. First, raters did not have access to any student identifiers. Unless the students signed their names, wrote about their hometowns, or in some way provided other identifying information as part of their response, the raters had no knowledge of student characteristics. Second, all raters were trained using Smarter Balanced-provided materials, which were approved as unbiased examples of responses at the various score points. Training involved constant comparisons with the rubric and anchor papers so that raters' judgments were based solely on the scoring criteria. Finally, following training, a cycle of diagnosis and feedback was maintained to identify any issues. Specifically, raters were closely monitored during scoring, and any instances of raters making scoring decisions based on anything except the criteria were discussed with the raters. After this feedback had been provided, raters were further monitored, and if any continued to exhibit bias after receiving a reasonable amount of feedback, they were dismissed.

Finally, a series of automated score verifications were implemented to further ensure the accuracy of scores. For example, a blank check was conducted, which reset scores when a condition code of "blank" was assigned to a response that had one or more characters in the

response string (e.g., a response comprised of spaces or tabs). In this case, only after three independent raters had assigned a condition code of "blank" to a response that appeared blank, but which included characters in the response string, was the score recorded. A similar check was run when a score or condition code other than "blank" was assigned to a response that included no characters in the response string. Automatic resetting of double-scored responses when two raters assign non-adjacent scores, mismatched condition codes, or a combination of a condition code and a numeric score provided an additional score verification. In addition to automatically resetting and rescoring these responses, the raters' information was captured in a report and reviewed by scoring directors, one of many tools used to determine retraining needs.

6.7.3 Rater Statistics and Monitoring

At a minimum, 15% of the handscored responses received blind double reads. Additionally, 5% of the responses scored comprised pre-approved validity responses. MI's VSC system automatically and randomly routed the requisite number of responses to raters for second reads and validity in an inconspicuous manner. Raters had no means of discerning whether they were scoring a first read, a second read, or a validity response. This system also prohibited raters from being eligible to score second reads for responses they had already scored.

MI's VSC scoring system randomly seeds validity responses among operational responses during scoring. A small set of validity responses is 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. Validity responses are indistinguishable from operational responses.

VSC reports provided real-time reports throughout the scoring effort. These reports were available for access by handscoring management. Inter-rater reliability reports provide the percentage of exact, adjacent, and non-adjacent agreement for scorable responses. Score point frequency distribution reports provide the percentage per score point and include the mean and standard deviation for each item. Validity performance reports provide the percentage of exact, adjacent, and non-adjacent agreement for validity responses and were used to monitor drift. Validity performance reports are typically used to monitor and correct drift at the group level. If the data indicate that raters as a group are scoring validity responses either consistently high or consistently low, leadership will recalibrate the group by having raters review key training responses that reflect the types of responses being missed in validity. Leadership may also provide raters with a supplemental set of responses that helps reinforce the lines for the various score-points and re-anchor the raters to the proper position, arresting groupwide drift.

In some cases, validity performance reports can be used to focus individualized feedback to raters who are struggling. When leadership notices a rater with low IRR, they will review the rater's mis-scored validity responses and associated data and look for a trend that suggests the scorer has drifted from the anchored responses. If such a trend is present, leadership can tailor feedback specific to that rater, typically by presenting them with live responses they have mis-scored in a way that is reflective of their overall drift from the anchor set criteria and providing targeted, thoughtful rationales for the "correct" scores.

Years of Smarter Balanced handscoring has allowed MI to amass a longitudinal dataset of rater performance data. MI's rater monitoring system uses validity responses calibrated to fit a unidimensional Item Response Theory (IRT) model for each content area/item type. This approach involves transforming raters' validity response scores into accuracy scores. Specifically, if the rater's score matches the "true" score of the validity response, an accuracy score of 2 is assigned. If the rater's score is adjacent to the score of the validity response, an accuracy score of 1 is assigned. Otherwise, for scores that are non-adjacent, an accuracy score of 0 is assigned. All accuracy score data for validity responses and readers are then fitted to a Generalized Partial Credit Model (GPCM) IRT model. Utilizing the resulting IRT parameters, MI then calculates accuracy values for each rater based on a given set of validity responses.

Extensive metrics (inter-rater reliability, calibrated validity, and sub-pools for monitoring drift) calculated by the monitoring system were used to ensure accuracy and productivity throughout the handscoring of a project. The system generated automated measures of rater performance drawing on validity, IRR, and other performance data. Raters and scoring managers received daily, automated messages summarizing raters' performance, ensuring all handscoring staff were aware of current performance and any issues that required attention. Additional outputs were also provided in manager-level reports and used to identify raters who required retraining and/or removal due to issues with accuracy and/or production. These data allowed scoring management to direct scoring leaders in review of specific VSC reports in order to determine the specific areas of attention required for any raters.

The monitoring system afforded the objective, dynamic identification of the most accurate raters, which we referred to as "expert raters." Specifically, expert raters are those with a demonstrated ability to score validity responses, including anchor validity responses originating from the field test administration,[1] highly accurately. Rater status changed daily based on current rater performance to ensure that any rater drift did not negatively impact scoring accuracy. Expert rater status was a precondition for conducting second readings.

[1] Responses and results of the 2014–2015 Smarter Balanced field-test administration were used to derive the base scale to which subsequent item parameters are aligned.

Team leaders spot-checked (i.e., read behind) raters' scoring to ensure that the raters were on target, and conducted one-on-one retraining sessions to address any problems found. At the beginning of the project, team leaders read behind every rater every day; they became more selective about the frequency and number of read-behinds as raters became more proficient at scoring.

6.7.4 Rater Retraining and Dismissal

Retraining was an ongoing process once scoring is underway. Daily analysis of the rater status reports enabled management personnel to identify individual or group retraining needs. When it became apparent that a whole team or group was having difficulty with a particular type of response, large group training sessions were conducted.

When read-behinds or daily statistics identified a rater who could not maintain acceptable agreement rates, the rater was retrained and monitored by scoring leadership personnel. Raters are released from the project if retraining is unsuccessful. In these situations, all items scored by a rater during the timeframe in question were identified, reset, and released back into the scoring pool. The aberrant rater's scores were deleted, and the responses were redistributed to other qualified raters for rescoring.

In addition to the processes described in Sections 6.7.3 and 6.7.4, several monitoring and retraining processes were added to the VSC system in spring 2023, including:

1) An additional validation stage was added to supplement Brief Writes and Research rater qualification. Immediately following the training and qualification steps described, all prospective Brief Write and Research raters were required to score, for each item, a 20-response set of pre-scored student responses sourced from the prior test administration. Like the qualification step, raters were required to meet accuracy standards during this validation to score operational responses for a given item. Any raters who failed to meet validation accuracy standards were automatically disqualified from scoring the item despite having passed qualification. This additional validation adds an additional level of quality assurance to those content areas and items that have been the most challenging to score accurately historically.
2) An automated feedback system was added to enhance the retraining methodology and augment the monitoring and feedback system used by scoring leadership. The automated feedback system identifies raters who require additional feedback—based on daily accuracy metrics—and automatically generates a custom set of responses for the rater to review. The system functions at the item level, thus providing feedback even to those raters with relatively high accuracy when the data identifies there are one or more items on which they can improve.

3) Reports using item-level accuracy expectations were implemented to identify items not meeting the expected levels of agreement. Specifically, these reports indicated the difference between expected accuracy and current accuracy for each item. Expected accuracy was defined based on historical data; in some cases (e.g., most mathematics items), expected accuracy exceeded Smarter Balanced's minimum accuracy thresholds. In this way, reports informed improvements to the scoring accuracy of all items.
4) Automated removal of raters and score resets were performed when item and rater performance failed to meet accuracy expectations. By limiting raters to scoring relatively fewer items, this approach also maximized accuracy across items.

6.7.5 Rater Agreement

Rater inter-rater reliability (IRR) was computed based only on scorable responses (numeric scores) scored by two independent raters. Non-scorable responses (e.g., off-topic, off-purpose, foreign-language) were scored by scoring leadership per the handscoring rules—and not by one expert and one random rater—and were thus excluded from IRR computations. For the handscored items, the human-human agreement was computed based on the 2022–2023 Connecticut summative assessment.

All ELA/L SA items were scored using a 0–2 rubric. Mathematics SA items were scored using 0–1, 0–2, or 0–3 rubrics. Condition codes were scored as zero.

Table 46 and Table 47 provide a summary of the human-human IRR based on items with a sample size greater than 50; as a result, only a subset of the administered items is presented in the tables. The IRR is presented with mean of percent exact agreement, minimum and maximum percent exact agreements, combined percent exact and adjacent agreement, and the mean, minimum and maximum quadratic weighted kappa (QWK). The average number of responses, as well as minimum and maximum number of responses to a given item are presented. The difference between the minimum and maximum number of responses is large because the number second scores and the number of condition codes varied widely across items.

Table 46. Inter-Rater Agreement for ELA/L Short-Answer Items

Grade	Number of Items	Number of Responses			%Exact			%(Exact+ Adjacent)	QWK		
		Mean	Min	Max	Mean	Min	Max		Mean	Min	Max
3	15	320.7	58	819	78.8	63.7	94.8	100.0	0.68	0.49	0.81
4	21	240.1	62	657	75.5	65.8	94.4	100.0	0.69	0.57	0.78
5	22	330.1	85	1159	76.0	58.8	86.6	100.0	0.72	0.53	0.84
6	19	287.3	74	732	72.8	62.1	87.6	100.0	0.61	0.48	0.86
7	23	301.9	61	709	74.2	65.9	86.9	100.0	0.69	0.53	0.85

Grade	Number of Items	Number of Responses			%Exact			%(Exact+ Adjacent)	QWK		
		Mean	Min	Max	Mean	Min	Max		Mean	Min	Max
8	28	231.0	71	838	73.2	56.5	88.5	100.0	0.67	0.46	0.88

Table 47. Inter-Rater Agreement for Mathematics Items

Grade	Score Point Range	Number of Items	Number of Responses			%Exact			%(Exact+ Adjacent)	QWK		
			Mean	Min	Max	Mean	Min	Max		Mean	Min	Max
3	0–1	11	247.3	186	298	91.1	84.4	95.6	100.0	0.79	0.65	0.90
4	0–1	10	249.5	227	310	87.0	78.0	98.0	100.0	0.68	0.42	0.96
5	0–1	9	217.7	207	237	93.5	87.3	99.2	100.0	0.74	0.28	0.96
6	0–1	10	246.7	136	395	94.2	84.8	100.0	100.0	0.73	0.06	1.00
7	0–1	10	288.8	177	350	93.7	86.2	98.9	100.0	0.74	0.41	0.91
8	0–1	14	310.5	289	351	91.2	84.4	97.0	100.0	0.76	0.53	0.93
3	0–2	34	258.0	96	322	89.9	80.8	100.0	100.0	0.92	0.86	1.00
4	0–2	38	241.6	62	317	89.8	78.4	100.0	100.0	0.89	0.57	1.00
5	0–2	58	221.5	80	267	88.6	74.0	97.3	100.0	0.88	0.60	0.98
6	0–2	41	332.6	303	383	86.6	72.8	99.1	100.0	0.84	0.72	0.99
7	0–2	23	304.8	270	345	88.0	76.7	94.9	100.0	0.83	0.65	0.96
8	0–2	30	304.0	284	349	87.6	74.4	95.5	100.0	0.85	0.65	0.92
3	0-3	6	183.2	120	268	91.4	88.2	94.3	100.0	0.96	0.94	0.98
4	0-3	4	253.5	234	308	88.1	86.0	90.3	100.0	0.94	0.93	0.95
5	0-3	9	206.1	132	257	88.8	82.0	96.4	100.0	0.92	0.87	0.97
7	0-3	1	311.0	311	311	93.6	93.6	93.6	100.0	0.92	0.92	0.92
11	0-3	4	300.3	296	304	84.8	81.5	90.9	100.0	0.95	0.94	0.97

7. REPORTING AND INTERPRETING SCORES

The Centralized Reporting System (CRS) generates a set of online score reports that includes the information describing student performance for students, parents, educators, and other stakeholders. The online score reports are produced immediately after students complete tests and any handscored items are scored. Because the score reports on students' performance are updated each time that students complete tests and handscored items are scored, authorized users (e.g., school principals, teachers) can quickly access information on students' performance and use it to improve student learning. In addition to individual students' score reports, the CRS also produces aggregate score reports by class, school, and district. The timely accessibility of aggregate score reports helps users monitor students' performance 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.

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

7.1 CENTRALIZED REPORTING SYSTEM

The CRS is designed to help educators and students answer questions about how well students have performed on English language arts/literacy (ELA/L) and mathematics assessments. The CRS is the online tool that provides all stakeholders with timely, relevant score reports. The CRS for the Smarter Balanced assessments has been designed such that score reports are easy to read and understand for all stakeholders. This is achieved by using plain, non-technical language to facilitate review by parents and the general public. The CRS 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.

Generally, the CRS provides two categories of online score reports: (1) aggregate score reports and (2) student score reports. Table 48 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 *Centralized Reporting System User Guide*, located via a Help button on the CRS.

Table 48. Types of Online Score Reports by Level of Aggregation

Level of Aggregation	Types of Online Score Reports
District School Teacher Roster	• Number of students tested and percentage of students with Level 3 or 4 (for overall students and by subgroup) • Average scale score and standard error of average scale score (for overall students and by subgroup) • Percentage of students at each achievement level on the overall test and by claims (for overall students and by subgroup) • Performance category in each target (overall students) • On-demand student roster report
Student	• Total scale score and standard error of measurement (SEM) • Achievement level on overall and claim scores with achievement-level descriptors • Average scale scores and standard errors of average scale scores for student's school, and district

Aggregate score reports at a selected aggregate level are provided for students overall and by subgroup. Users can see student assessment results in any of the subgroups. Table 49 presents the types of subgroups and subgroup category provided in the CRS.

Table 49. Types of Subgroups

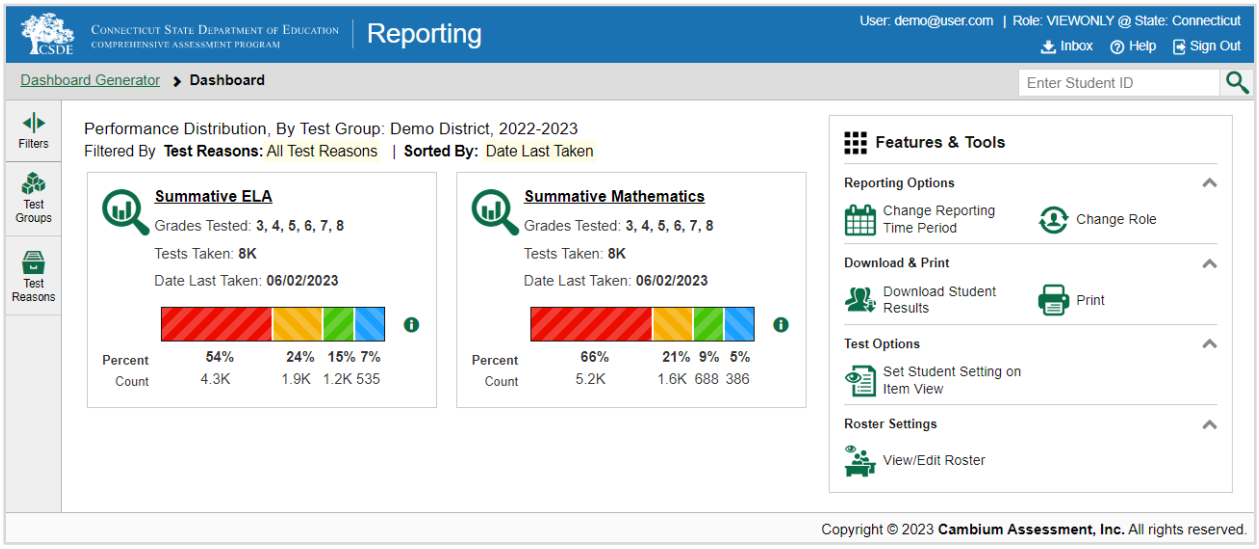
Subgroup	Subgroup Category
Gender	Male Female
IDEA Indicator	Yes Blank
Limited English Proficiency (LEP) Status	Yes Blank
Ethnicity/Race	American Indian or Alaskan Native Asian Asian/Pacific Islander Black or African American Hispanic or Latino Native Hawaiian or Other Pacific Islander White Multi-Racial

7.1.1 Dashboard

Once authorized users in the district, school, and teacher level log in to the CRS, the dashboard page shows overall test results for all tests that the students have taken grouped by test family (e.g., Smarter Balanced Summative ELA/L). The dashboard summarizes students' performance by test family for both ELA/L and mathematics across all grades, including (1) the grades of the students who have tested, (2) the number of tests taken, (3) the test date last taken, and (4) the percentage and counts of students at each achievement level. District personnel see district summaries, school personnel see school summaries, and teachers see summaries of their students.

Exhibit 1 presents an example dashboard page at the district level.

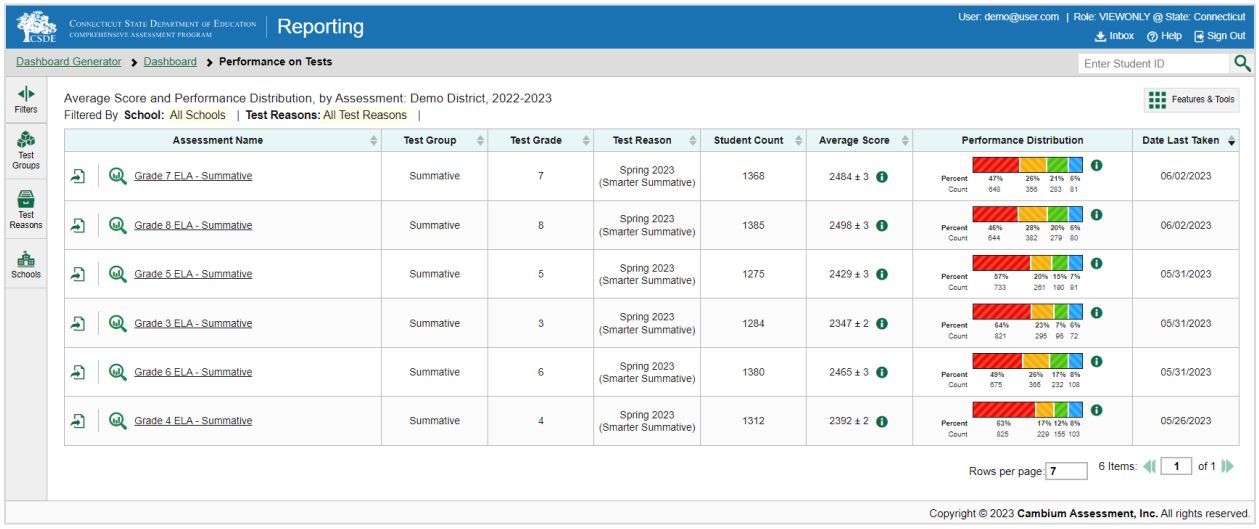
Exhibit 1. Dashboard: District Level



Once the user clicks the test family that he or she wants to explore further, it will take the user to the detailed dashboard, where the results are shown by test (e.g., Grade 3 ELA/L). The detailed dashboard page will appear by test in each grade. The detailed dashboard summarizes students' performance by test in each grade, including (1) student count, (2) average scale score and standard error of the average scale score, (3) the percentage and counts of students at each achievement level, and (4) test date last taken.

Exhibit 2 presents an example detailed dashboard page for summative ELA/L at the district level.

Exhibit 2. Detailed Dashboard: District Level



7.1.2 Aggregate Score Reports: Overall Performance

Student performance for each grade in a subject area for a selected aggregate level is presented when users select a specific assessment name. On each aggregate report, the summary report presents the summary results for the selected aggregate unit and the summary results for the aggregate unit both above and below the selected aggregate. For example, if a school is selected, the summary results of the district that the school belongs to are provided as well as the school summary results so that school performance can be compared with the other aggregate levels.

The aggregated summary report provides the summaries on a specific grade in a subject, including (1) student count, (2) the average scale score and standard error of the average scale score, (3) the percentage and counts of students in each achievement level, and (4) the percentage of proficient students. The summaries are also presented for students overall and by subgroup.

Exhibit 3 presents an example overall performance summary result for grade 3 ELA/L at the district level, and Exhibit 4 presents an example summary by gender.

Exhibit 3. Overall Performance Summary Results for Grade 3 ELA/L: District Level

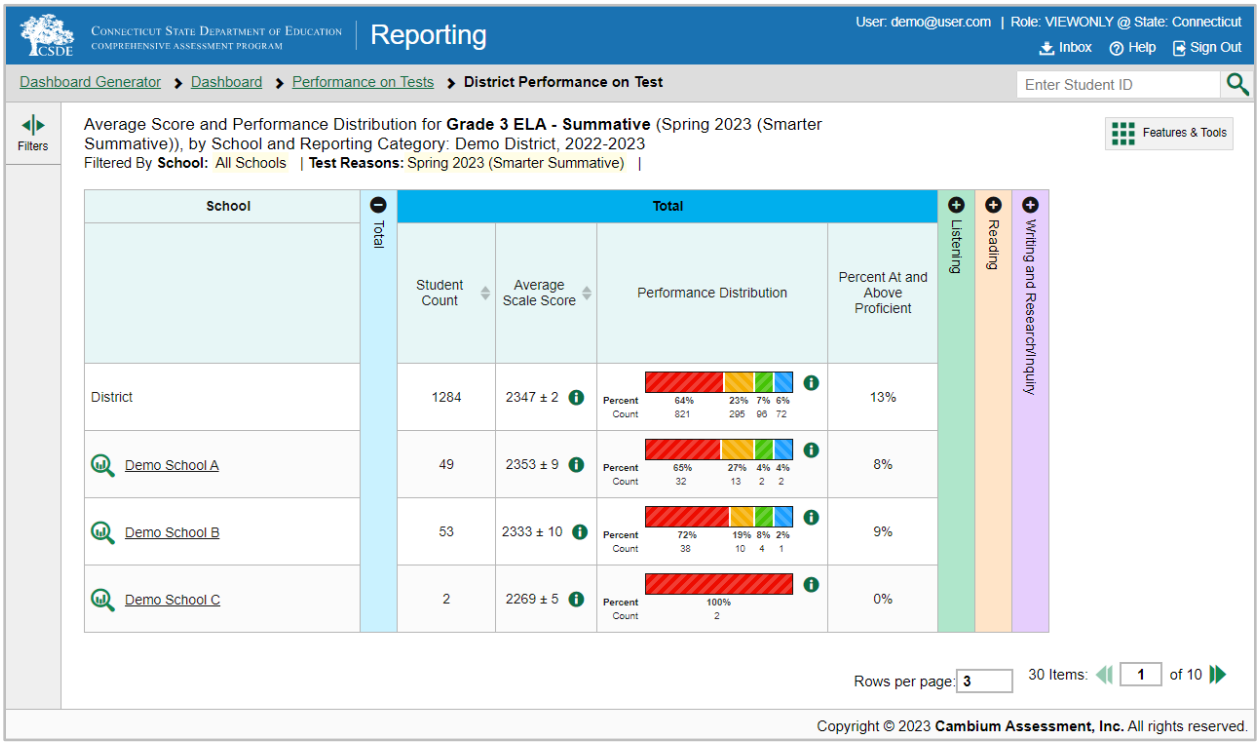
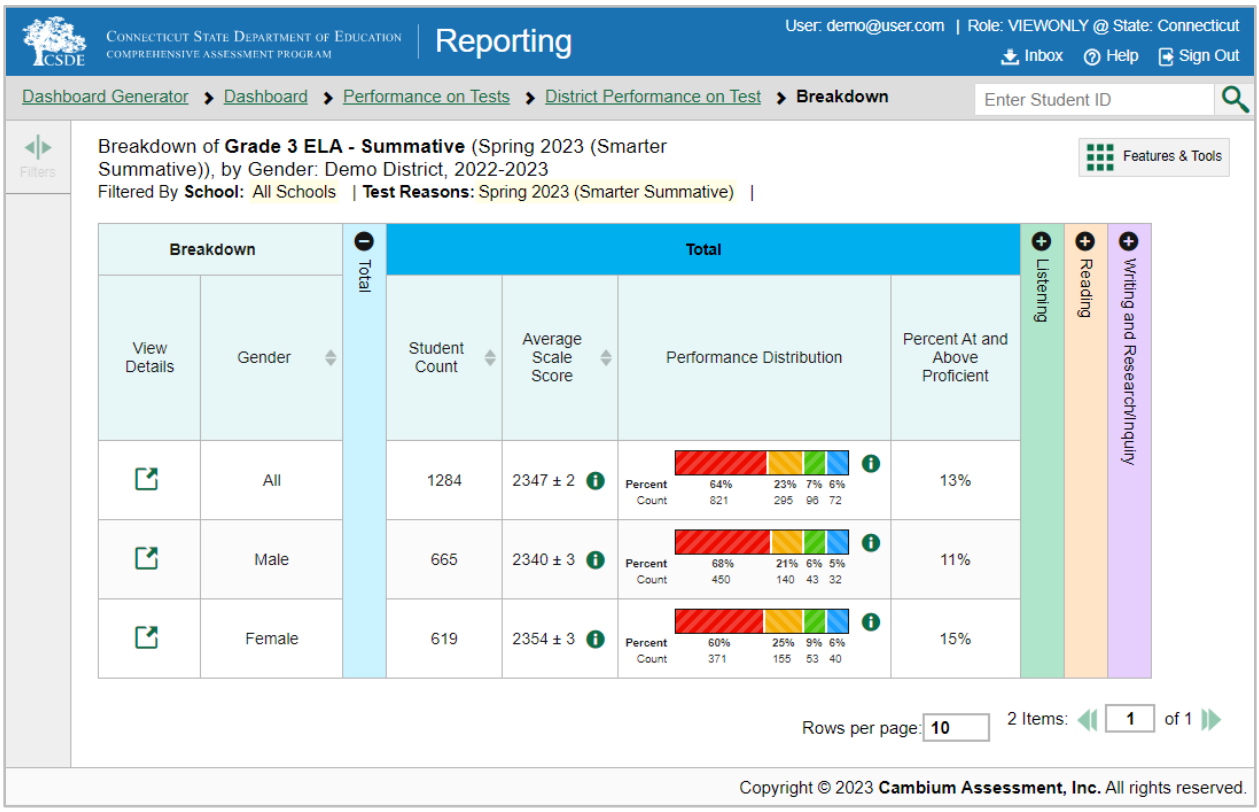


Exhibit 4. Overall Performance Summary Results for Grade 3 ELA/L by Gender: District Level



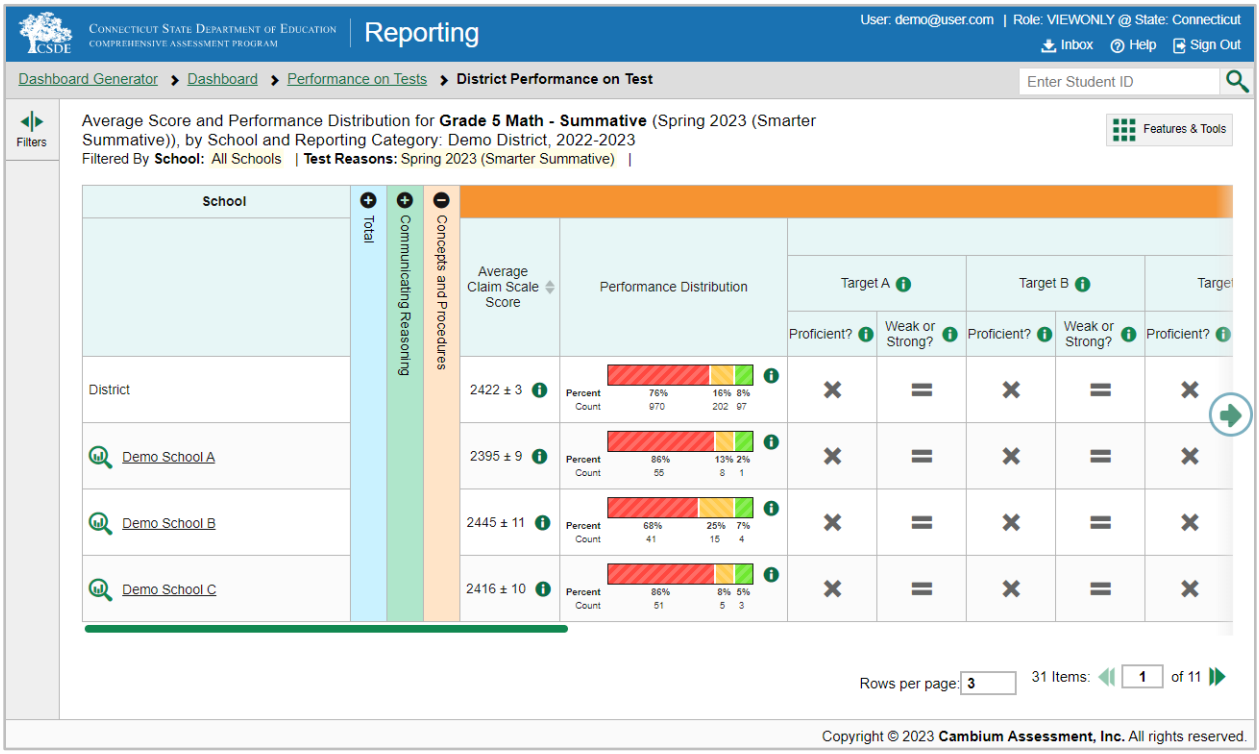
7.1.3 Aggregate Score Reports: Claim and Target Performance

Detailed summaries on aggregated claim and target results are also available on the same report page when a claim on the right side of the page is selected. For the claim result, (1) the average scale score and standard error of the average scale score and (2) performance distribution are presented. For the target result, the strength or weakness indicators on each target within a claim are presented. These strength or weakness indicators are presented in two ways. The "Proficient?" measure indicates whether the group's performance on each target is better than (check mark), less than (x mark), or not different from (half-filled circle) the proficiency standard for the selected test. The "Weak or Strong?" measure presents whether the group's performance on each target is lower than (minus sign), higher than (plus sign), or not different from (equal sign) the group's overall performance. If there is insufficient information in the "Proficient?" measure or "Weak or Strong?" measure, this is indicated with a star sign (*).

Like the overall performance summary results, the summary report presents results for the selected aggregate unit and the aggregate unit both above and below the selected aggregate unit. Also, the summaries on claim- and target- level performance can be presented for overall students and by subgroup.

Exhibit 5 present an example of claim- and target-level results for grade 5 mathematics at the district level.

Exhibit 5. Claim and Target Level Results for Grade 5 Mathematics: District Level

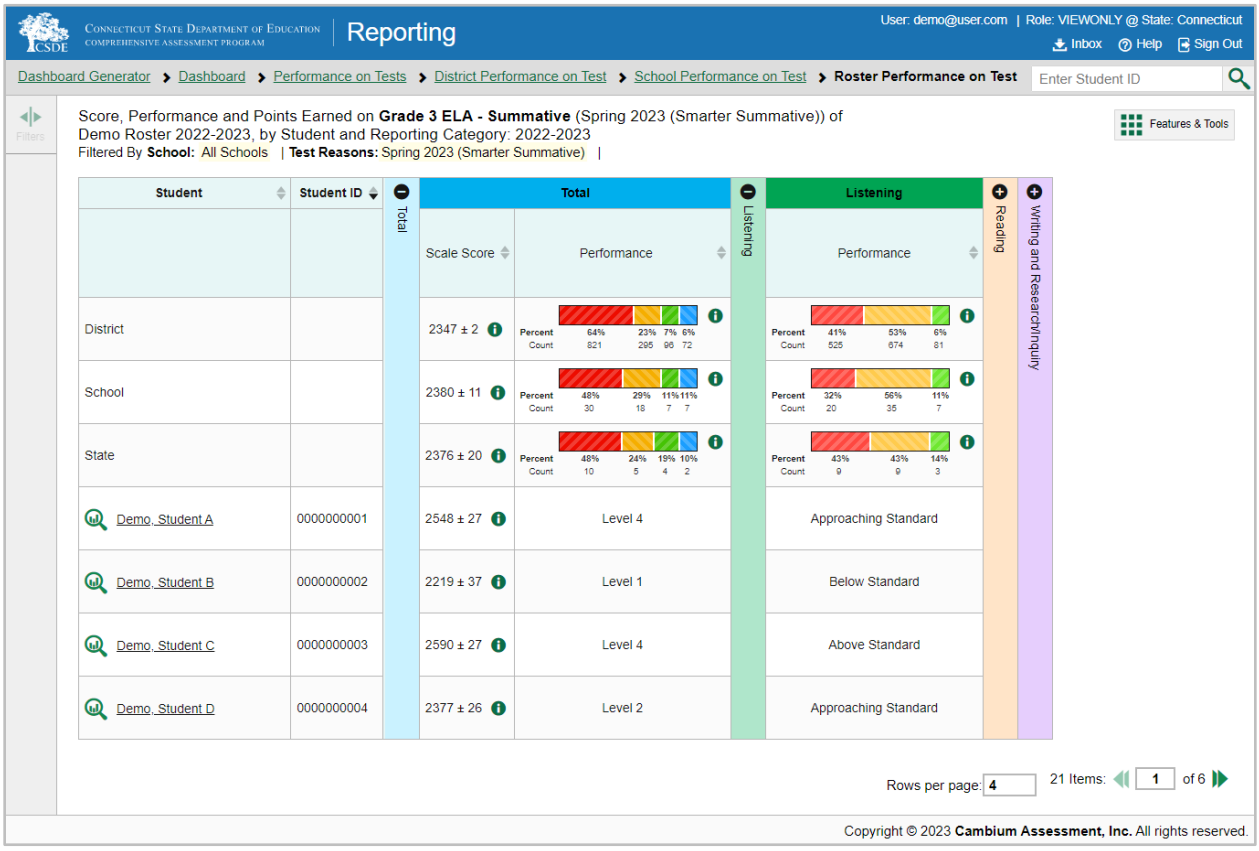


7.1.4 Roster Performance Report

Class, teacher, and school performance rosters provide users with performance data for a group of students belonging to a system-defined or user-defined class. The report includes (1) the student's overall subject scale scores with standard error of measurement, (2) the performance level, and (3) performance category for each claim.

Exhibit 6 shows a sample roster performance report for grade 3 ELA/L.

Exhibit 6. Roster Performance Report for Grade 3 ELA/L

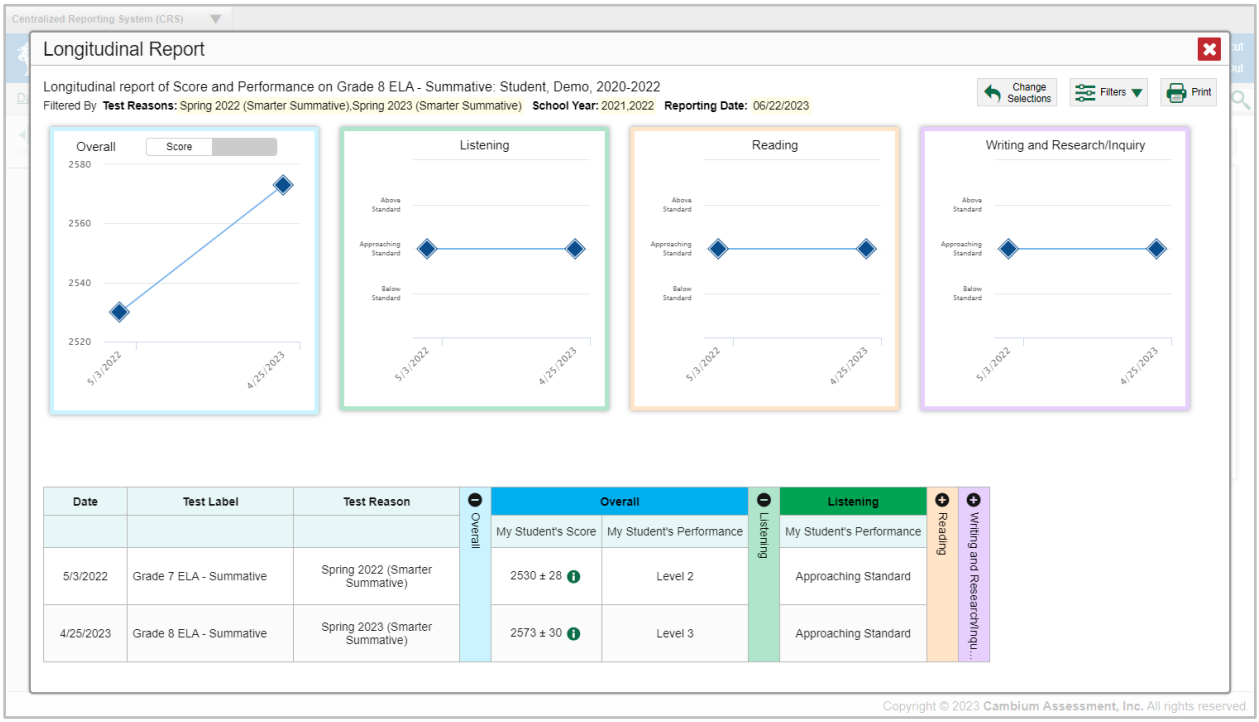


7.1.5 Trend Report

The trend (i.e., longitudinal) page provides the trend of student performance for an aggregate unit over time. The trend report can be set to plot either average scale scores or percentage of students in each achievement level on the graph for the selected aggregate unit. The trend report is also available at the individual student level.

Exhibit 7 presents an example trend report page for ELA/L at the individual student level.

Exhibit 7. Trend Report for ELA/L: Student Level



7.1.6 Individual Student Report

An individual student report can be generated and exported as a PDF file. The individual student report shows the student's overall performance on the test with detailed information on multiple pages. In each subject area, the individual student report provides (1) the scale score and SEM; (2) achievement level for overall test; (3) performance category in each claim; and (4) average scale scores for the student's district and school.

On the first page of the individual student report, the student's name, scale score with the SEM, and achievement are shown at the top of the page. In the 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 the SEM 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 were administered multiple times. Furthermore, in the barrel chart, achievement-level descriptors with cut scores at each achievement level are provided. This defines the content area knowledge, skills, and processes that test takers at the achievement level are expected to possess.

On the right side of the barrel chart, average scale scores and standard errors of the average scale scores for the student's district and school are displayed so the student's achievement can be compared with the above aggregate levels. It should be noted that the "±" next to the student's scale score is the SEM of the scale score, whereas the "±" next to the average scale

scores for aggregate levels represents the standard error of the average scale scores. On the bottom of the page, the student's performance on each claim is displayed alongside a description of his or her performance on each claim.

Exhibits 8 presents an example of individual student reports for grade 3 Math.

Exhibit 8. Individual Student Report for Grade 3 Math

CSDE Connecticut State Department of Education Comprehensive Assessment Program | Reporting

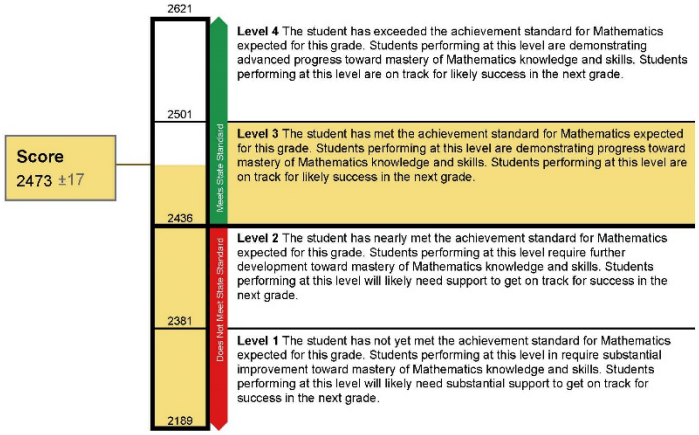
Individual Student Report

Student, Demo
Student ID: 0000000000 | Student DOB: 1/1/2014 | Enrolled Grade: 3
Date Taken: 5/9/2023

Grade 3 Math - Summative 2022-2023
Demo District
Demo School

Scale Score: 2473±17 **Performance:** Level 3

How Did Your Child Do on the Test?



Level 4 The student has exceeded the achievement standard for Mathematics expected for this grade. Students performing at this level are demonstrating advanced progress toward mastery of Mathematics knowledge and skills. Students performing at this level are on track for likely success in the next grade.

Level 3 The student has met the achievement standard for Mathematics expected for this grade. Students performing at this level are demonstrating progress toward mastery of Mathematics knowledge and skills. Students performing at this level are on track for likely success in the next grade.

Level 2 The student has nearly met the achievement standard for Mathematics expected for this grade. Students performing at this level require further development toward mastery of Mathematics knowledge and skills. Students performing at this level will likely need support to get on track for success in the next grade.

Level 1 The student has not yet met the achievement standard for Mathematics expected for this grade. Students performing at this level in require substantial improvement toward mastery of Mathematics knowledge and skills. Students performing at this level will likely need substantial support to get on track for success in the next grade.

How Does Your Child's Score Compare?

Name	Average Scale Score
Demo District	2361±2
Demo School	2352±11

Information on Standard Error of Measurement

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 (±10) indicates a score range between 2290 and 2310.

How Did Your Child Perform on Different Areas of the Test?

The table and the graph below indicate student performance on individual reporting categories. The black dot indicates the student's score on each reporting category. The lines to the left and right of the dot show the range of likely scores your student would receive if he or she took the test multiple times.

Below Standard | Approaching Standard | Above Standard

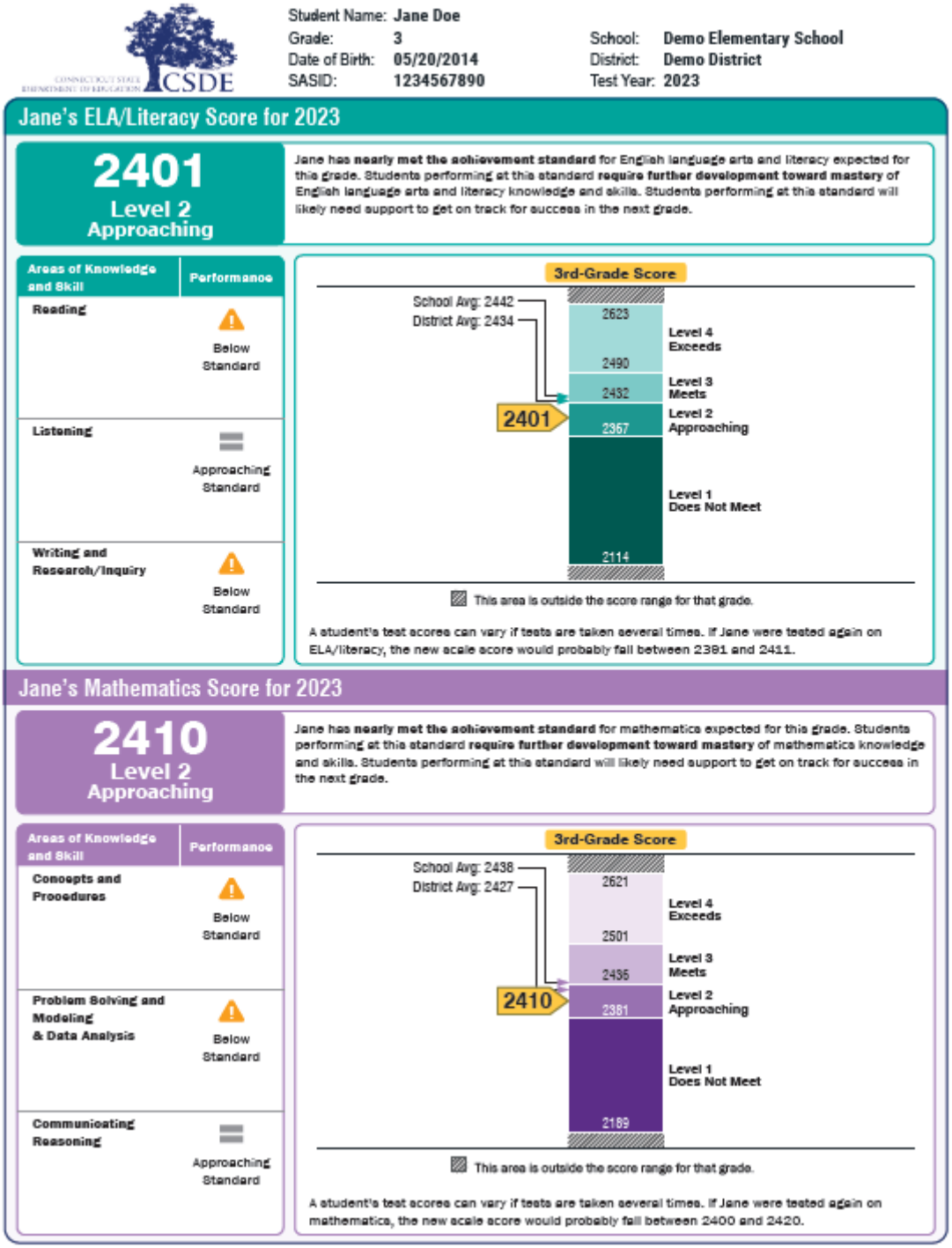
Category	Performance	Performance	Performance level Description
Communicating Reasoning	Below the Standard Above the Standard		Student may be able to clearly and precisely construct viable arguments to support their own reasoning and to critique the reasoning of others.
Concepts and Procedures	Below the Standard Above the Standard		Student can explain and apply mathematical concepts and interpret and carry out mathematical procedures with precision and fluency.
Problem Solving and Modeling & Data Analysis	Below the Standard Above the Standard		Student may be able to solve a range of complex well-posed problems in pure and applied mathematics, making productive use of knowledge and problem solving strategies. Student may be able to analyze complex, real-world scenarios and may be able to construct and use mathematical models to interpret and solve problems.

Generated on 6/22/2023 Page 1 of 1

7.1.7 Paper Family Score Reports

After the testing window is closed, parents whose children participated in a test receive a full-color paper score report (hereinafter referred to as a family report) including their child's performance on ELA/L and mathematics. The family report includes information on student performance that is similar to the student detail page from the CRS with additional guidance on how to interpret student achievement results in the family report. An example of a family report is shown in Exhibit 9.

Exhibit 9. Sample Paper Family Score Report



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 Conditional Standard Error of Measurement

A scale score (the observed score on any test) is an estimate of the true score. If a student takes a similar test multiple times, the resulting scale score will vary across administrations, sometimes being a little higher, a little lower, or the same. The SEM represents the precision of the scale score, or the range in which the student would likely score if a similar test was administered multiple 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 were tested again, it is likely that he or she 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 (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 test takers 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 ELA/L 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 demonstrating 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 the overall test, student performance on each claim is evaluated with respect to the "Meets Standard" achievement standard. For students performing at "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 their performance does not provide enough information to tell whether they reached the "Meets Standard" mark for the specific claim.

7.2.5 Performance Category for Targets

Teachers and educators sometimes need more detailed reports on student performance for instructional needs. The target report provides information on student performance about relative strength and weakness scores for each target within a claim. The strengths and weaknesses reports are generated for aggregate units of classroom, school, and district and provide information about how a group of students in a class, school, or district performed on each target, either relative to the proficiency standard (i.e., "Proficient?" target measure) or relative to their overall performance on the test ("Weak or Strong?" target measure). Target-level reports are produced for the aggregate units only, not for individual students, because each student is administered too few items in a target to produce a reliable score for each target.

For the "Proficient?" target measure, students' observed performance on items within the reporting element is compared to the expected performance on those items of someone who has an ability equal to the proficiency cut (i.e., the 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 compared to the proficiency standard. Conversely, when observed performance within a target is below the proficiency cut, the reporting unit shows a relative weakness in that target.

For the "Weak or Strong?" target measure, students' observed performance on items within the reporting element is compared with the expected performance based on the overall ability estimate. At the aggregate level, when observed performance within a target is greater than the 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.

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 some targets may be based on relatively few items, especially for a small group.

7.2.6 Aggregated Scale Score

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

7.2.7 Appropriate Uses of Test Results

Assessment results can provide information about individual students' achievements on the test. Overall, assessment results show what students know and are able to do in certain subject areas and provide further information on whether students are on track to demonstrate the knowledge and skills necessary for college and career readiness. Additionally, assessment results can be used to identify students' relative strengths and weaknesses in certain content areas. For example, performance categories for targets can be used to identify a group's relative strengths and weaknesses among targets within a claim.

Assessment results on student achievement on the test can be used to help teachers or schools make decisions on how best 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 used to improve teaching and student learning. For example, a group of students may perform very well overall on the test but potentially not perform as well in several targets compared to their overall performance. In this case, teachers and schools would be able to identify the strengths and weaknesses of their students through the group performance by claim and target. They could then promote instruction in the specific claim or target areas in which their students perform relatively lower. Furthermore, by narrowing down the student performance results by subgroup, teachers and schools can determine which strategies may be best suited to improving student learning, particularly for students from disadvantaged subgroups. For example, teachers can examine student assessment results by LEP status and may observe that LEP students need help particularly in a certain specific area, such as reading literary responses and analysis. Teachers can then provide additional focused instruction for these students to enhance their achievement in any specific target or claim in which they are struggling.

In addition, assessment results can be used to compare performance among different students and among different groups. Teachers can evaluate how their students perform compared with other students in their school and district for overall scores 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 when data are available. In the Smarter Balanced assessments, the scale scores across grades are on the same scale because the scores are vertically linked across grades. Therefore, scale scores from one grade can be compared with the next grade. i.e., measuring the growth.

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 decisions 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 consider 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. Cambium Assessment, Inc. (CAI) uses a series of quality control steps to ensure the error-free production of score reports in both online and paper-pencil formats. The quality of the information produced in the Test Delivery System (TDS) is tested thoroughly before, during, and after the testing window opens.

8.1 ADAPTIVE TEST CONFIGURATION

For the computer-adaptive test (CAT) component, 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, cut scores, the item information (i.e., answer keys, item attributes, item parameters, and passage information), and slopes and intercepts for theta-to-scale score transformation. The accuracy of the information in the configuration file is independently checked and confirmed before the testing window opens.

With the test configuration file, CAI uses simulated test administrations to configure the adaptive algorithm to optimize item selection to meet blueprint specifications while targeting test information to student ability. First, the simulator generates a sample of students with an ability distribution that matches that of the population in previous year's data. The ability of each simulated student is used to generate a sequence of item response scores while matching the blueprint and minimizing measurement error. These simulations provide a rigorous test of the adaptive algorithm. The results of these simulations are used to configure and evaluate the adequacy of the item selection algorithm used to administer the Smarter Balanced summative assessments.

After the adaptive testing simulations, another set of simulations for the combined tests (CAT and performance task [PT] components) are performed for scoring engine verification. The simulated data are generated such that verification of the scoring engine is based on a wide range of student response patterns. CAI rigorously checks whether the scoring rule specified in scoring specifications was applied accurately. The scores in the simulated data file are checked independently.

8.1.1 Platform Review

CAI's TDS supports a variety of item layouts. Each item goes through an extensive platform review on different operating systems such as 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 during which each item is checked to ensure that it is displayed appropriately on each tested platform. A platform is a combination of a hardware device and an operating system. In recent years, the number of platforms has proliferated, and platform review now takes place on various platforms that are significantly different from one another.

Platform review is conducted by a team. The team leader projects the item as it was web approved in the Item Tracking System (ITS), and team members, each using a different platform, look at the same item to confirm 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 a content approval role. The UAT period provides the Department with an opportunity to interact with the exact test that the students will use.

8.2 Quality Assurance in Document Processing

The Smarter Balanced summative assessments are administered primarily online; however, a few students took paper-pencil assessments. When test documents were scanned, a quality control sample of documents consisting of 10 test cases per document type (normally between 500 and 600 documents) was created so that all possible responses and all demographic grids were verified, including various typical errors that required editing via Measurement Incorporated's (MI) Data Inspection, Correction, and Entry (DICE) application program. This structured testing method 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 the results from the scanner, the editing process (validation and data correction), and the transfer to the CAI database were correct.

8.3 Quality Assurance in Data Preparation

CAI's TDS has a real-time, built-in quality-monitoring component. After a test is administered to a student, the TDS passes the resulting data to CAI's QA system. The QA system 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, and the total number of field-test items and operational items. The QA system ensures that the test record contains no data from items that have been invalidated.

Data pass directly from the Quality Monitor (QM) System to the Database of Record (DOR), which serves as the repository for all test information, and from which all test information for reporting is retrieved. The Data Extract Generator (DEG) is the tool that is used to retrieve data from the DOR for delivery to the Department. CAI staff ensures that data in the extract files match the DOR before delivering it to the Department.

8.4 Quality Assurance in Online Test Delivery System

To monitor the performance of the TDS during the test administration window, CAI 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 CAI contracts for service in excess of this amount. Once deployed, the servers are monitored at the hardware, operating system, and software platform levels with monitoring software that alerts CAI's 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 CAI 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—such as data about how long it takes to load, view, or respond to an item—are captured for each assessed student. All of this information is logged as well, enabling CAI to automatically identify schools or districts experiencing unusual slowdowns, often before they even notice.

A series of QA 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 2.8, Data Forensic Program.

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

For the CAT, 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 the blueprint and that items are performing as anticipated.

Table 50 presents an overview of the QA reports.

Table 50. 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 unexpectedly 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

8.4.1 Score Report Quality Check

For the Smarter Balanced summative assessments, two types of score reports are produced: online reports and printed reports (family reports only).

8.4.1.1 Online Report Quality Assurance

Scores on the online assessments are assigned automatically by the systems in real time. Every test undergoes a series of validation checks. Once the QA system signs off, data are passed to the DOR, which serves as the centralized location for all student scores and responses, ensuring that there is only one place where the official record is stored. Only after scores have passed the QA checks and are uploaded to the DOR are they passed to the Centralized Reporting System (CRS), which is responsible for presenting individual-level results and calculating and presenting aggregate results. Absolutely no score is reported in the CRS until it passes all the QA system's validation checks. All of the previously mentioned processes take milliseconds to complete so that within less than one second after CAI receives handscores and they pass QA validation checks, the composite score will be available in the CRS.

8.4.1.2 Paper Report Quality Assurance

Statistical Programming

The family reports contain custom programming and require rigorous QA processes to ensure their accuracy. All custom programming is guided by detailed and precise specifications in CAI's 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 the agreed-on procedures. Custom programming is implemented independently by two statistical programming teams working from the

specifications. The scripts are released for production when the output from both teams matches exactly.

Much of the statistical processing is repeated, and CAI has implemented a structured software development process to ensure that the repeated tasks are implemented correctly and identically each time. Small programs (called *macros*) are written to take specified data as input and produce data sets containing derived variables as output. Approximately 30 such macros reside in CAI's library for 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, the director of psychometrics, and 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 perform the many complicated calculations. This program is developed and tested using artificial data generated to test both typical and extreme cases. Additionally, 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 Variable Data Intelligent PostScript Printware (VIPP) and allows virtually infinite control of the visual appearance of the reports. After designers at CAI create backgrounds, CAI's 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. CAI'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 are run through the psychometric process and the score reporting statistical programs, and the output is formatted as VIPP input. This enables CAI to test the entire system.

Programmed output goes through multiple stages of review and revision by graphics editors and the CAI Score Reporting team to ensure that design elements are accurately reproduced, and data are correctly displayed. Once CAI receives the final data and VIPP programs, the CAI Score Reporting team reviews proofs that contain actual data based on CAI's standard quality assurance documentation. Several CAI staff members review a large sample of the reports to ensure that all data are correctly placed on reports. This rigorous review is conducted over several days and takes place in a secure location in the CAI building. All reports containing actual

data are stored in a locked storage area. Before the reports are printed, CAI provides a live data file and individual student reports with sample districts for Department staff review. CAI 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.

Billingsley, P. (1995). *Probability and Measure* (3rd ed.). New York, NY: John Wiley & Sons, Inc.

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).

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

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

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

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

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

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

U.S. Department of Education. (2015). *Peer Review of State Assessment Systems: Non-Regulatory Guidance for States*. Washington, D.C. Retrieved from https://www2.ed.gov/policy/elsec/guid/assessguid15.pdf