Hawai‘i Smarter Balanced Assessments 2024–2025 Technical Report

Submitted to
Hawai‘i 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) has developed a next-generation assessment system designed to accomplish two goals: first, to measure students' mastery of the *Common Core State Standards* (CCSS) in English language arts/literacy (ELA/L) and mathematics in grades 3–8 and 11, and second, to provide valid, reliable, and fair test scores of students' academic achievement. At the time of development, Hawaiʻi was one of 18 member states (plus the U.S. Virgin Islands) leading the development of assessments in ELA/L and mathematics. The system includes summative assessments for accountability purposes and optional interim assessments that supply meaningful feedback and actionable data that teachers and educators can use to help students succeed. SBAC, a state-led collaboration, 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 Hawaiʻi State Board of Education formally adopted the CCSS in ELA/L and mathematics on June 18, 2010. All students in Hawaiʻi, including students with significant cognitive disabilities who are eligible to take the Hawaiʻi State Alternate Assessment (an alternate assessment based on Alternate Academic Achievement Standards), are taught the same academic content standards. The Hawaiʻi CCSS define the knowledge and skills that students need to succeed in college and careers after graduating from high school. These standards include rigorous content and application of knowledge through higher-order skills and align with college and workforce expectations.

Since the adoption of the CCSS in 2010, the Hawaiʻi Department of Education (HIDOE) began implementing the CCSS in the 2012–2013 school year with grades K–2 and 11–12. This transition was fully implemented in all grade levels in the 2013–2014 school year. The new Hawaiʻi statewide assessments in ELA/L and mathematics aligned with the CCSS were administered for the first time in spring 2015 to students in grades 3–8 and 11 in all public elementary and secondary schools.

The Smarter Balanced assessments comprise the end-of-year summative assessment designed for accountability purposes, and the optional interim assessments that support teaching and learning throughout the year. The summative assessments evaluate 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).

- The **Computer-Adaptive Test (CAT)** provides an individualized assessment for each student.
- The **Performance Task (PT)** challenges students to apply their knowledge and skills 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, which cannot be adequately assessed with selected- or constructed-response items. The computer can score some PT items, but most are handscored.

The optional interim assessments allow teachers to monitor student progress throughout the year and provide information that they can use to improve instruction and learning. These tools are used at the discretion of schools and complex areas, 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:

- The **Interim Comprehensive Assessment (ICA)** tests the same content and reports scores on the same scale as the summative assessments.
- The **Interim Assessment Block (IAB)** focuses on specific sets of related concepts that measure three to eight assessment targets and provide 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 the IAB alone.

In the 2019–2020 school year, the U.S. Department of Education waived testing requirements due to the COVID-19 pandemic (https://www2.ed.gov/policy/gen/guid/secletter/200320.html). For the 2020–2021 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. Starting in the 2021–2022 school year, all students were required to take ELA/L and mathematics summative assessments.

Starting with the 2020–2021 Smarter Balanced summative test administration, Hawai'i shortened the full test blueprints for ELA/L and mathematics and allowed schools to administer remote test administrations to individual students.

The American Institutes for Research (AIR) delivered the Hawai'i statewide assessments in ELA/L and mathematics through the 2018–2019 school year. Starting with SY 2020–2021, Cambium Assessment, Inc. (CAI) (formerly a segment of AIR) delivered and scored the Smarter Balanced assessments and produced the score reports. Measurement Incorporated (MI) scored the handscored items.

This report provides a technical summary of Hawai'i's 2024–2025 administration of the Smarter Balanced summative assessments in English language arts/literacy (ELA/L) and mathematics in grades 3–8 and 11. The report is divided into eight chapters: Overview; Test Administration; Summary of the 2024–2025 Operational Test Administration; Validity; Reliability; Scoring; Reporting and Interpreting Scores; and Quality Control Procedures. The data included in this report are based on Hawai'i 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 in Hawai'i, it is an addendum to the 2024–2025 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.

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

2 TEST ADMINISTRATION

2.1 TESTING WINDOWS

The 2024–2025 Smarter Balanced Assessment (SBA) testing window spanned approximately three months for the summative assessments for most schools and spanned the entire school year for the interim assessments. The paper-pencil fixed forms for the summative assessments were administered concurrently during the three-month online summative window. Table 1 shows the testing windows for both online and paper-pencil assessments.

Table 1. 2024–2025 Testing Windows

Tests	Grade	Start Date	End Date	Mode
Summative Assessments	3–8	2/18/2025	5/30/2025	Online Adaptive
		3/10/2025 (Multi-track)	6/13/2025 (Multi-track)	
	11	2/18/2025	5/30/2025	Online Adaptive
		11/18/2024 (Block Scheduled)	5/30/2025 (Block Scheduled)	
	3–8, 11	2/18/2025	5/16/2025	Paper Fixed-Form
	3–8, 11	2/18/2025	6/13/2025	Remote Online Adaptive
	3–8, 11	2/18/2025	5/16/2025	Braille Paper Fixed-Form
Interim Comprehensive Assessments	3–8, 11	8/13/2024	7/18/2025	Online Fixed-Form
Interim Assessment Blocks	3–8, 11	8/13/2024	7/18/2025	Online Fixed-Form
Focused Interim Assessment Blocks	3–8, 11	8/13/2024	7/18/2025	Online Fixed-Form

2.2 TEST OPTIONS AND ADMINISTRATIVE ROLES

The Smarter Balanced Assessment (SBA) is administered primarily online. To ensure that all eligible students in the tested grades were given the opportunity to take the SBA, several assessment options were available to accommodate students' needs. Table 2 lists the testing options offered in 2024–2025. A testing option is selected by content area. Once an option is selected, it is applied to all tests in the content area.

Table 2. 2024–2025 Testing Options

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

To ensure that standardized administration conditions are met, test administrators (TAs) follow procedures outlined in the *Smarter Balanced ELA/L and Mathematics Online, Summative Test Administration Manual*

(TAM). TAs must review the TAM before testing to ensure that the testing room is prepared for testing (e.g., removing certain classroom posters, arranging desks). Make-up procedures should be established for students who are absent on the day(s) of testing. 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 are principals (PRs), test coordinators (TCs), and TAs. The main responsibilities of the key personnel are outlined in the following descriptions. More detailed descriptions can be found in the TAM provided online at: https://smarterbalanced.alohahsap.org/resource-list/en/smarter-balanced-summative-test-administration-manual-2024-2025.

Principals

The PR's primary responsibility is to ensure that testing in his or her school is conducted in accordance with the test procedures and security policies established by the Hawaiʻi State Department of Education (HIDOE).

PRs are responsible for performing the following functions:

- Reviewing all Smarter Balanced policies and test administration documents
- Reviewing scheduling and test requirements with TCs and TAs
- Working with TCs and technology coordinators to ensure that all systems, including the CAI Secure Browser, are properly installed and functioning
- Designating or acting as the TC
- Importing users (TCs) into the Test Information Distribution Engine (TIDE)
- Scheduling and administering training sessions for all TCs, TAs, and technology coordinators (refer to Section 2.3, Training and Information for Test Coordinators and Administrators)
- Ensuring that all personnel understand and are trained on the proper administration of the Smarter Balanced assessments
- Monitoring secure test administration
- Investigating and reporting all testing improprieties, irregularities, and breaches reported by TCs or TAs
- Attending to any secure materials according to state and Smarter Balanced policies

Test Coordinator

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

TCs are responsible for performing the following functions:

- Identifying TAs and proctors (if appropriate) and ensuring that TAs complete the TA Certification Course
- Establishing a testing schedule with PRs and TAs based on the testing windows
- Working with technology staff to ensure timely computer setups and installations
- Working with TAs to review student information in TIDE to ensure that student information and test settings for designated supports and accommodations are applied correctly
- Identifying students who may require designated supports and test accommodations and ensuring that procedures for testing these students follow state and Smarter Balanced policies
- Attending all school trainings and reviewing all Smarter Balanced policy and test administration documents
- Ensuring that all TAs attend school trainings and review online training modules posted on the portal
- Establishing secure and separate testing rooms if needed
- Monitoring secure administration of the test
- Monitoring testing progress during the testing window and ensuring that all students participate, as appropriate
- Investigating and reporting all testing improprieties, irregularities, and breaches reported by the TAs in coordination with the PRs
- Attending to any secure materials according to state and Smarter Balanced policies

Test Administrator

The TA's primary responsibility is to administer the Smarter Balanced assessments. The TA's role is designed for test administrators, such as technology staff, who administer tests but should not have access to student results.

TAs are responsible for performing the following functions:

- Completing Smarter Balanced test administration training and reviewing all Smarter Balanced policy and test administration documents before administering any Smarter Balanced assessments
- Reviewing student information for accuracy before testing to ensure that students receive the proper test with the appropriate supports and reporting any potential data errors to TCs and PRs, as appropriate
- Administering the Smarter Balanced assessments
- Reporting all potential test security incidents to the TCs or PRs in a manner consistent with Smarter Balanced, state, and school policies

2.2.2 Online Administration

Within the state's testing window, schools can set the testing schedule and customize their testing conditions, such as allowing students to test in intervals (i.e., multiple sessions) rather than in one long period and minimizing the interruption of classroom instruction and efficiently using 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.

Starting with SY 2020–2021, a new feature was developed within the universally used Test Delivery System (TDS) that allowed tests to be administered remotely by a TA to students who remained at home. The decision to allow students to test remotely was made at the school level in cases when a parent or guardian refused to take a student to campus for testing but insisted on the student being tested. This new feature allowed TAs to pre-schedule a testing session, host online video and chat features with a group of students, and video monitor students in a testing session.

To ensure that TAs were able to use these new features, an additional *Remote Testing TA Certification Course* was developed. TAs scheduled to administer remote testing sessions were required to complete this course prior to test administration. In addition, before a student was eligible for remote test administration, a parent or guardian had to provide written consent to the school to administer a remote test that would contain video and audio components allowing the TA to view and monitor the student. The school's TC was responsible for ensuring that these students had positive consent for remote testing within the TIDE system. Additional resources were developed tor TAs to understand the requirements for remote testing and posted to the state portal at https://smarterbalanced.alohahsap.org/resource-list/en/remote-summative-test-administration-2024-2025.

TCs oversee all aspects of testing at their schools and serve as the main point of contact; TAs administer the online assessments only. TAs are trained in the online testing requirements and the mechanics of starting, pausing, and ending a test session. Training materials for the test administration are provided online. All school personnel who serve as TAs must complete an online TA Certification Course. Staff who complete this certification course receive a certificate of completion and are qualified to administer assessments.

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

Once an assessment is started, the student must answer all the test questions presented on a page before proceeding to the next page. Skipping questions is not permitted. For the CAT, students can 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. In addition, students can review and edit only previously answered items before submitting the assessment. During an active CAT session, if a student reviews and changes the response to a previously answered item, all following items to which the student already responded remain the same. No new items are assigned to this student for changing answers. For example, a student paused for 10 minutes after completing Item 10. After the pause, the student went back to Item 5 and changed the answer. If the updated response to Item 5 changed the item score from wrong to right, the student's overall score would improve; however, there would be no change in Items 6–10. For PTs, there is no pause rule; but the same rules that apply to the CAT for reviews and changes to responses also apply to PTs.

The CAT must be completed within 45 calendar days of the start date, or the assessment opportunity will expire. The ELA/L performance task must be completed within 10 calendar days of the start date.

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

The TA must remain in the room when the test is administered in person and be present continuously when using the video feature for remote test administrations to monitor student testing. When the test session ends, the TA must ensure that each student has successfully logged out of the system. The TA must also collect and shred any handouts or scratch paper that students may have used during the CAT session; if handouts or scratch paper were used for the ELA/L PT, the TA must collect and securely store them until the ELA/L PT has been submitted. After the PT's submission, the TA must securely shred all handouts and/or scratch paper.

The number of students who took summative tests remotely in 2024–2025 is presented in Table 3.

Table 3. Number of Students Who Took Tests Remotely in the 2024–2025 Summative Test Administration

Subject	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 11	Total
ELA/L	6	3	10	11	5	13	1	49
Mathematics	7	3	8	11	5	12	1	47

2.2.3 Paper-Pencil Test Administration

There are two matching versions of the paper-pencil Smarter Balanced ELA/L and mathematics assessments. One version is provided as an accommodation for students who cannot access a computer, and the other is a braille version for students with blindness or visual impairments. Both versions contain the same items and are based on the Smarter Balanced full-length blueprints for ELA/L and mathematics used in SY 2024-25. TCs from schools with any student(s) who require the paper-pencil assessment must submit a request to HIDOE for test materials on behalf of the student(s) before the testing window opens. If the request is approved by HIDOE, the testing contractor will ship the appropriate test booklets and the paper-pencil TAM to the school.

Separate test booklets are used for the ELA/L and mathematics assessments, which are based upon the Smarter Balanced full-length blueprint. The items from the CAT and the PT components are combined into one test booklet, including two sessions for the CAT and one session for the PT in both content areas. Thus, the TA can break up the assessment into separate test sessions. After the student completes the assessment, the TC will return the test booklets to the testing contractor, and the testing contractor will scan the answer document and score the test, including the handscored items.

The total number of students who took paper-pencil tests is shown in Table 4.

Table 4. Number of Students Who Took Paper-Pencil Tests in the 2024–2025 Summative Test Administration

Subject	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 11	Total
ELA/L		1	1	1	2	1		6
Math		1	1	1	1	1		5

2.2.4 Braille Test Administration

The adaptive braille test was available with the same test blueprint 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 computer-adaptive 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 HIDOE. All items on the Braille HAT can be presented to students using a Refreshable Braille Display (RBD). The blueprints for the Braille HAT follow the Smarter Balanced full-length blueprints for mathematics used in SY 2024-25. This was not an option for administration in Hawai'i in 2024–2025, and no versions of these tests were taken.

The braille interface comprises several formats as follows:

- The braille interface includes a text-to-speech (TTS) component for mathematics consistent with the read-aloud assessment accommodation. The Job Access with Speech (JAWS) screen-reading software provided by Freedom Scientific is an essential component that students use with the braille interface.
- Mathematics items are presented to students in Nemeth Braille Code via a braille embosser through the adaptive online summative test 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 uncontracted literary braille (for items containing only text) and via a braille embosser (for items with tactile or spatial components that cannot be read by an RBD).

Before administering the online summative assessments using the braille interface, TAs must ensure that technical requirements are met. These requirements apply to the student's computer, the 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

PRs and TCs oversee all aspects of testing at their schools and serve as the main points of contacts; TAs administer the online assessments. The online TA Certification Course, webinars, user guides, manuals, and training sites are used to train TAs on the online testing requirements and the mechanics of starting, pausing, and ending a test session. Training materials for administration are provided online.

2.3.1 Online Training

Multiple training opportunities are offered to key assessment staff through the state portal.

TA Certification Course

There are three TA Certification Courses that are available for TAs: an Interim Assessment TA Certification Course, a Summative Assessment TA Certification Course, and a Remote Assessment TA Certification Course. TAs must complete an online TA Certification Course every year in order to administer assessments. The Interim Assessment TA Certification Course must be completed to administer Interim Assessments, while the Summative Assessment TA Certification Course must be completed to administer Summative Assessments. For 2024-2025, TAs administering summative tests must complete both the Interim and Summative TA Certification Courses. These web-based courses are each about 30–45 minutes long and cover information on testing policies and the steps for administering Interim and Summative test sessions in the online testing system. The courses are interactive, requiring participants to start test sessions under different scenarios. Participants are required to answer multiple-choice questions about the information provided throughout the training and at the end of the Summative TA course. A third TA Certification Course of about 20 minutes is required for TAs administering tests in a remote format. For 2024–2025, TAs administering remote tests were required to take all courses.

Webinars

The following five webinars were offered to users in the field:

- *Accessibility and Accommodations.* This webinar provides an overview of the accessibility features and supports available to students during testing, including universal tools, designated supports, and accommodations.

- *Smarter Balanced Test Coordinators Training.* This webinar provides information about accessing and using the Interim Assessments, Summative Assessments, Centralized Reporting System, and Digital Library.

- *Test Information Distribution Engine.* This webinar provides an overview of how to navigate the Test Information Distribution Engine (TIDE), including managing student information and monitoring test progress.

- *Centralized Reporting System.* This webinar provides information on the Centralized Reporting System (CRS), including an overview of accessing student reports and the distribution of reports to parents and guardians.

- *Remote Interim Administration.* This webinar provides information about setting up and administering remote interim assessments using the Test Delivery System (TDS) and the CAI Secure Browser.

Each of these webinars is about one hour long. The interactive nature of these training webinars allows the participant to ask questions during and after the presentation. After the live webinar, a streaming video recording of the webinar is made available on the state portal.

Practice and Training Test Site

Starting in August 2022, separate online training sites were opened for TCs, TAs, and students. TAs could practice administering assessments and starting and ending test sessions on the TA Training Site, and students could practice taking an online assessment on the Student Practice and Training Site. The Smarter Balanced assessment practice tests mirror the corresponding summative assessments for ELA/L and mathematics. Each test provides students with a grade-specific testing experience, including a variety of question types and difficulty levels (approximately 30 items each in ELA/L and mathematics) and a performance task in ELA/L.

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

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

Manuals and User Guides

The following manuals and user guides are available on the Hawaiʻi Statewide Assessment Program Portal:

The *Smarter Balanced Online Summative Test Administration Manual* provides information for TCs 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 *Smarter Balanced Interim Assessments Test Administration Guide* provides an overview of how to prepare for and administer the Smarter Balanced Interim assessments.

The *Online Calculators in the Test Delivery System Manual* and the *Desmos User Guide* provide instructions for using the online Desmos Calculators during testing.

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

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

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

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

The *Test Information Distribution Engine User Guide* and *Quick Guide to TIDE* are designed to help users navigate TIDE. Users can find information on managing user account information, student account information, student test settings and accommodations, testing incidents, creating and editing rosters, and voice packs.

The *Centralized Reporting System User Guide* provides information about the CRS, including instructions for viewing score reports, managing test administration, and searching for students. It is also a component of the Smarter Balanced Interim Assessments that allows authorized users to view individual student responses on both the Interim Comprehensive Assessments (ICAs) and the Interim Assessment Blocks (IABs).

The *Guide to Navigating the Online HSAP Administration* is designed to help users navigate the TDS, including the Student Interface and the TA Interface, and to help 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), which allows teachers to view items on the Smarter Balanced interim assessments.

The *Usability, Accessibility, and Accommodations Guidelines* describe the current universal tools, designated supports, and accommodations adopted by the Smarter Balanced states to ensure valid assessment results for all students taking its assessments.

All manuals and user guides pertaining to the 2024–2025 online testing were available on the portal, and PRs and TCs were able to use these manuals and guides when training TAs on test administration policies and procedures.

Training Modules

The following training modules were created to help users in the field understand the overall Smarter Balanced assessments and how each system works. All modules were provided in PowerPoint presentation format; and three modules were also narrated.

The *Accessibility and Accommodations Module* outlines the designated supports and accommodations available for the online assessments, as described in the *Usability, Accessibility, and Accommodations Guidelines* available on the Smarter Balanced website.

The *Administering a Test Using Speech-to-Text (STT) Software Module* provides an overview of key features of the STT accommodation and its functionality during testing.

The *Centralized Reporting Module* provides an overview of the key features of the CRS, which provides teachers with detailed information about their students' performance on the Smarter Balanced Interim Assessments.

The *Embedded Universal Tools and Online Features Module* acquaints students and teachers with the online universal tools (e.g., types of calculators, expandable text) available in the Smarter Balanced assessments.

The *Individual Student Assessment Accessibility Profile (ISAAP) Module* offers an overview of the Smarter Balanced Usability, Accessibility, and Accommodations Guidelines, the ISAAP Process, and the ISAAP Tool. Smarter Balanced suggests a process and tool by which each student's needs can be matched with appropriate universal tools, designated supports, and/or accommodations.

The *Performance Task Overview Module* provides an introduction to the ELA/L performance task.

The *Read Aloud Module* is designed to help the read-aloud test reader understand the guidelines for the read-aloud designated support and accommodation when administering the Smarter Balanced assessments.

The *Scribing Protocol Training Module* is designed for test administrators acting as scribes to understand the guidelines for administering this designated support to students with this accommodation for the Smarter Balanced assessments.

The *Student Interface for Online Testing Module* explains how to navigate the Student Interface. The module includes information on how students log in to the testing system, select a test, understand the test layout, and use test tools.

The *Technology Requirements for Online Testing Module* provides current information about technology requirements, site readiness, supported devices, and CAI Secure Browser installation.

The *Test Administrator (TA) Interface for Online Testing Module* presents an overview of how to navigate the TA Interface.

The *Test Information Distribution Engine (TIDE) Module* provides an overview of the TIDE system. It includes information on logging in to TIDE and managing user accounts, student information, rosters, and testing incidents.

The *Testing with Braille Training Module* provides TAs with information on administering online tests to students using braille.

The *What Is a CAT? Module* describes the CAT and how it works when taking ELA/L and mathematics online assessments.

2.3.2 Statewide Trainings

Two series of virtual statewide trainings were held during SY 2024–2025. The first series of virtual statewide trainings was held September 16–17, 2024. The second series of virtual statewide trainings was held November 12–18, 2024. A set of in-person trainings were held January 21–31, 2025. These training sessions provided the information necessary for administering the Smarter Balanced assessments in ELA/L and mathematics. New TCs were provided with information on participation guidelines, test security and ethics, accessibility and accommodations, interim assessments, test administration procedures, technology requirements, the CRS, and family reports.

A separate series of trainings was held on November 7, 2024, and February 27, 2025. The training sessions held on February 27 focused specifically on accessibility and accommodations for all Hawai'i statewide assessments, including the Smarter Balanced summative and interim assessments, while the training held on November 7 focused specifically on the administration of Braille for all Hawai'i statewide assessments.

2.4 TEST SECURITY

The security of assessment instruments and the confidentiality of student information are vital to maintaining the validity, reliability, and fairness of the test results. All test items, test materials, and student-level testing information are classified as 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. Various features of the TDS also protect test security. This section describes student confidentiality, system security, testing environment security, and policies on testing incidents.

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.

Three elements are involved in assuring that students are accessing appropriate test content, including:

1. *Test eligibility,* which refers to the assignment of a test to a particular student
2. *Test accommodation,* which refers to the assignment of a test setting to specific students based on student needs
3. *Test session,* which refers to the authentication process that TAs must follow when creating a test session, including reviewing and approving a test and its settings for each student, and the student signing on to take the test

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

- Providing login information (usernames and passwords) to other authorized TIDE users or to unauthorized individuals
- Sending a student's name and SSID number together in an email message
- Having a student log in and test under another student's SSID number

Test materials and score reports should not be exposed to reveal student names with test scores except for authorized individuals with an appropriate need to know. If information about a test must be sent via email or fax, only the SSID number should be included, not the student's name.

All students, including homeschooled 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 HIDOE file and uploaded nightly via a secured file transfer site to the online TDS during the testing window.

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

After a test session, only staff with the administrative roles of PR, TC, or teacher (TE) can view their students' scores. TAs who are not also teachers 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 are accessed only by the appropriate user groups. The end goal of system security entails protecting and maintaining data and system integrity, safeguarding personal information, and ensuring accurate data transfer and appropriate levels of user access.

Hierarchy of Control

As described in Section 2.2.1, Administrative Roles, PRs, TCs, and TAs have well-defined roles and levels of access to the testing system. PRs are responsible for selecting and entering the TC's information into TIDE, and the TC is responsible for entering TAs' and TEs' information into TIDE. Throughout the year, the PR and TC 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 teachers.

Password Protection

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

Secure Browser

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

2.4.3 Security of the Testing Environment

The TCs 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 have been set up to prevent students from crowding. Good lighting, ventilation, and protection from noise and other interruptions are also essential factors to consider when selecting testing rooms.

TAs must establish procedures to maintain a quiet environment during each test session, recognizing that some students may finish more quickly than others. If students are allowed to leave the testing room when they finish their assessments, TAs must 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, TAs are encouraged to have students read a book after they have completed the assessment.

If a student needs to leave the room for a brief time, the TAs must pause the student's assessment. If a pause lasts longer than 20 minutes during the CAT component, the student can continue the assessment in a new test session. However, 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 spent outside the testing room to look up answers.

Room Preparation

The testing room should be prepared before the start of the test session. Any information displayed on bulletin boards, chalkboards, or charts that students might use to answer test questions should be removed or covered. This rule applies to rubrics, vocabulary charts, student work, posters, graphs, content-area strategy charts, etc. All cell phones belonging to testing personnel and students must be turned off and stored out of sight in the testing room. TAs are encouraged to minimize access to the testing rooms by posting signs in halls and entrances to promote optimal testing conditions; they should also post "TESTING—DO NOT DISTURB" signs on the doors of testing rooms.

Seating Arrangements

TAs should provide adequate spacing between students' seats. Student seating should be arranged to prevent them from looking at other students' answers. Because the online CAT is adaptive, it is unlikely that students will see the same test questions as other students; however, students should be discouraged from communicating through appropriate seating arrangements. For the ELA/L performance task, different forms are distributed throughout the testing room so that students are less likely to receive the same forms as their neighbors.

After the Test

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

For the paper-pencil tests, specific instructions on how to package and secure the test booklets for return to the testing contractor's office are provided in the paper-pencil *Test Administration Manual*.

2.4.4 Test Security Violations

Every individual who administers or proctors the assessments is responsible for understanding the required security procedures associated with administering the assessments. The *Smarter Balanced Online Summative Test Administration Manual* outlines and categorizes prohibited testing practices into three groups, described here.

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., student[s] leaving the testing room without authorization).

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

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

Complex and school personnel are required to document all test security incidents in the test security incident log. This log is the document of record for all test security incidents and should be maintained at the complex level and submitted to HIDOE at the end of testing.

2.5 Student Participation

All students enrolled in grades 3–8 and 11 at public or public charter schools in Hawaiʻi are required to participate in the Smarter Balanced ELA/L and mathematics summative assessments, except the following:

- Students with significant cognitive disabilities who meet the criteria for a state-selected or state-developed ELA/L and mathematics alternate assessment based on the extensions of the Common Core standards (approximately 1% or fewer of the student population)
- Students in the English language learner (ELL) program whose first U.S. school in the past 12 months is a Hawaiʻi public or public charter school
- Students enrolled in the Hawaiian Language Immersion Program in grades 3–8

Only students in these three categories can be excused from taking the Smarter Balanced ELA/L assessments (all three categories) and/or the Smarter Balanced mathematics assessments (categories one and three). 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 Homeschooled Students

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

2.5.2 Exempt Students

The following categories of students are exempt from participating in the Smarter Balanced assessments based on required documentation:

- A student who has a significant medical emergency
- A student who is receiving services at an out-of-state residential program
- An ELL who has moved to the country within the year (ELA/L exemption only)
- A student who meets the requirements of Regulation 4140, Exceptions to Compulsory School Attendance

2.6 Online Testing Features and Testing Accommodations

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

The *Guidelines* apply to all students. They emphasize an individualized approach to the implementation of assessment practices for students who have diverse needs and participate in large-scale content assessments. The *Guidelines* focus on universal tools, designated supports, and accommodations for the Smarter Balanced assessments of ELA/L and mathematics. At the same time, the *Guidelines* support important instructional decisions about accessibility and accommodations for students who participate in the Smarter Balanced assessments.

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

State-level users, TCs, and teachers can set embedded and non-embedded designated supports and accommodations based on their user role in TIDE. Designated supports and accommodations must be set in TIDE prior to starting a test session.

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

For additional information about the availability of designated supports and accommodations, refer to the Smarter Balanced *Usability, Accessibility, and Accommodations Guidelines* at: https://smarterbalanced.alohahsap.org/resource-item/en/usability-accessibility-and-accommodations-guidelines-2024-2025.

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 preset in TIDE. In the 2024–2025 test administration, the following universal tools were available for all students to access. For specific information on how to access and use these features, refer to the *Smarter Balanced Online, Summative, Test Administration Manual* at: https://smarterbalanced.alohahsap.org/resource-list/en/smarter-balanced-summative-test-administration-manual-2024–2025.

Embedded Universal Tools

Breaks (Pause). A student can pause the assessment and return to the test question that he or she was working on. However, if an assessment is paused for more than 20 minutes, students will not be allowed to return to previously attempted test questions.

Calculator (for calculator-allowed mathematics items only in grades 6–8, 11). This is an embedded on-screen digital calculator for calculator-allowed items that students can access by clicking the calculator button. This tool is available only with specific items that the Smarter Balanced item specifications have indicated as appropriate.

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

English Dictionary. An English dictionary is available for the full-write portion of an ELA/L performance task. A full-write is the second component of a performance task.

English Glossary. This feature displays grade- and context-appropriate definitions of specific construct-irrelevant terms in English on the screen via a pop-up. The student can access the embedded glossary by clicking any of the pre-selected terms.

Expandable Passages and/or Stimuli. Each passage or stimulus can be expanded to take up a larger portion of the screen.

Global Notes. Global notes is a notepad that is available for the ELA/L performance task in which students complete a full-write. Students click the notepad icon for the notepad to appear. During the ELA/L performance task, the notes are retained from segment to segment and allow a student to return to the notes even though he or she cannot go back to specific items in the previous segment.

Highlighter. This tool is used to mark desired text, test questions, item answers, or parts of these with color. An enhanced highlighting feature allows multiple color options. Highlighted text remains available throughout each test segment. This tool is not available while the Line Reader tool is in use.

Keyboard Navigation. This tool allows students to navigate text using a keyboard.

Line Reader. Students use an onscreen universal tool to assist in reading by raising and lowering the tool for each line of text on the screen. If the enhanced line reader mode is enabled, all content except for the line in focus is grayed out for greater emphasis. This tool is not available while the Highlighter tool is in use.

Mark for Review. Students can mark a question for review in order to return to it later. However, for the CAT, if the assessment is paused for more than 20 minutes, students are not allowed to return to marked test questions.

Math 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 that the Smarter Balanced item specifications have indicated that one or more of these tools are appropriate.

Spellcheck. This is a writing tool for checking the spelling of words in student-generated responses. Spellcheck indicates only that a word is misspelled; it does not provide the correct spelling. This tool is available only with the specific items that the Smarter Balanced item specifications have indicated as appropriate. Spellcheck is bundled with other embedded writing tools for all performance task full-write items: planning, drafting, revising, and editing.

Strikethrough. This feature allows the student to cross out answer options. If an answer option is an image, a strikethrough line will not appear, but the image will be grayed out.

Thesaurus. A thesaurus is available for the full-write portion of an ELA/L performance task. A full-write is the second part of a performance task.

Writing Tools. Selected writing tools (e.g., bold, italic, bullets, undo, redo) are available for all student-generated responses. (Also, refer to spellcheck.)

Zoom. Students can zoom in on test questions, text, or graphics. This tool makes these features appear larger on the screen.

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 can take breaks when individually needed to reduce cognitive fatigue when they experience heavy assessment demands. The use of this universal tool may result in the student needing additional overall time to complete the assessment.

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

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

Thesaurus. A thesaurus provides synonyms of terms while a student interacts with text included in the assessment. This tool is available for the full-write portion of an ELA/L performance task. A full-write is the second part of a performance task. The use of this universal tool may result in the student needing additional time to complete the assessment.

2.6.2 Designated Supports and Accommodations

Designated supports for the Smarter Balanced assessments are features available for use by any student for whom the need has been indicated by an educator (or team of educators with the parent or 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 which supports should be designated for individual students. All educators making these decisions should be trained to use this process and should be made aware of the range of available designated supports. Smarter Balanced members have identified digitally embedded and non-embedded designated supports for students for whom an adult or team has indicated a need for the support.

Accommodations are modifications in 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 only 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 the screen background or font color based on their needs or preferences. This may include reversing the colors for the entire interface or choosing the color of the font and background. Black on white, reverse contrast, black on rose, medium gray on light gray, and yellow on blue were offered for the online assessments.

Illustration Glossaries (for mathematics items). Illustration glossaries are provided for selected construct-irrelevant terms for mathematics. Illustrations for these terms appear on the computer screen when students select them. Students can also adjust the size of the illustration and move it around the screen. Only students with the illustration glossary setting enabled can use this accommodation.

Masking. Masking involves blocking off content that is not of immediate need or that may be distracting to the student. This tool allows students to focus their attention on a specific part of a test item.

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

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

Text-to-Speech (for mathematics stimuli and items and ELA/L items; not for ELA/L reading passages). Text is read aloud to the student via embedded text-to-speech technology. The student can control the speed and raise or lower the volume of the voice via a volume control. This support is also available in Spanish for mathematics tests when students have a Spanish language support selected.

Text-to-Speech in Spanish (for mathematics stimuli and items). Text is read aloud to the student via embedded text-to-speech technology in Spanish. The student can control the speed and raise or lower the volume of the voice via a volume control.

Translated Student Interface Messages (for mathematics tests in Spanish). Translation of the student interface messages is a language support available prior to beginning the actual test items. Students can see test directions in Spanish. As an embedded designated support, translated test directions are automatically a part of the Spanish language translations designated support.

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

Translations (Spanish) (for mathematics items). Dual language translations are a linguistic support available for some students; dual language translations provide the full translation of each test item above the original English language version of the item.

Turn Off Any Universal Tools. A TA may 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. Students may adjust the volume control beyond the computer's built-in settings using headphones or other non-embedded devices.

Bilingual Dictionary. The bilingual/dual-language word-to-word dictionary is a language support that can be provided for the full-write portion of an ELA/L performance task.

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

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

Illustration Glossaries (for mathematics paper-pencil tests). The illustration glossaries are a language support provided for selected construct-irrelevant terms for mathematics. 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 students to increase the size of images and text on the screen to a level not allowed by the universal Zoom tool.

Math Manipulatives. This support allows eligible students with IEPs and Section 504 Plans to represent their understanding of mathematical concepts using visual and tactile concrete materials. This list of approved mathematics manipulatives that may be provided on-site includes Algebra Tiles (recommended for grade 6 and above), Base Ten Blocks, Colored Tiles, Geoblocks Set, Geoboards and Geobands, Multi-Link Cubes, Pop Cubes, or Similar Cubes, Multi-Sensory Learning (MSL) Kit, One-Inch Blocks, Pattern Blocks, Transparent Sheets, and Two-Color Counters. Up to four manipulatives may be selected for a student; other accommodations not listed can be requested for verification.

Medical Supports. Students may have access to an electronic device for medical purposes (e.g., glucose monitor). The device may include a cell phone and should support the student for medical reasons only during testing.

Noise Buffers. Ear mufflers, white noise, and/or other equipment that reduces environmental noises may be used.

Printed Test Directions in English. Available as a supplement to the TAM, a printed copy of oral test directions in English may be provided to the student. The use of this support may result in the student needing additional overall time to complete the assessment.

Read-Aloud (for mathematics stimuli and items and ELA/L items; not for ELA/L reading passages). The 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, all grades). 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 Online Summative Test Administration Manual* and the *Guidelines for Read-Aloud, Test Reader*. All or portions of the content may be read aloud.

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

Separate Setting. The test location is altered so that the student is tested in a setting different from that made available to 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 Student Interface Messages. A bilingual adult may read aloud a PDF file of directions translated in each of the languages currently supported.

Translated Test Directions in American Sign Language (ASL). Test directions that include test administration scripts are translated into ASL video. The ASL human signer and the signed test content are viewed at the same time. Students may view portions of the ASL video as often as needed.

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). This accommodation allows test content to be translated into an 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; Nemeth Braille Code is available for mathematics.

Braille Transcript (for ELA/L listening passages). This is a braille transcript of the closed captioning created for the listening passages. The braille transcripts are available in uncontracted and contracted English Braille American Edition (EBAE).

Closed Captioning (for ELA/L listening items). Printed text may appear on the computer screen as audio materials are presented.

Speech-to-Text. Voice recognition allows students to use their voices as input devices to the computer in order to dictate responses or give commands (e.g., opening application programs, pulling down menus, saving work). Voice recognition generally can recognize speech up to 160 words per minute. Students use the testing system, along with a microphone, for this embedded accommodation.

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

Word Completion. This allows students to begin writing a word and choose from a list of words that have been predicted from word frequency and syntax rules. Word prediction is delivered via an embedded software program. The program must use only single-word prediction. Functionality such as phrase prediction, predict ahead, or next word must be deactivated. The program must have settings that allow only a basic dictionary. Expanded dictionaries, such as topic dictionaries and word banks, must be deactivated. Phonetic spelling functionality and programs with built-in speech output that reads back the information the student has written may also be used. Students who use word prediction in conjunction with speech output will need headphones unless tested individually in a separate setting.

Non-Embedded Accommodations

100s Number Table. A paper-based table listing numbers 1–100 is available for reference.

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 adapted keyboards, large keyboards, Sticky Keys, MouseKeys, FilterKeys, adapted mouse, touch screen, head wand, and switches.

Braille (paper-pencil assessment). 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). The following codes are available for the ELA/L paper-pencil assessment: EBAE uncontracted, EBAE contracted, Unified English Braille (UEB) uncontracted, and UEB contracted. The following codes are available for the mathematics paper-pencil assessment: EBAE uncontracted with Nemeth Braille Code, EBAE contracted with Nemeth, UEB uncontracted with Nemeth, UEB contracted with Nemeth, UEB uncontracted with UEB mathematics, and UEB contracted with UEB mathematics.

Calculator (for calculator-allowed items mathematics items only in grades 6–8, 11). This is a non-embedded calculator for students needing a special calculator, such as a braille calculator or a talking calculator, currently unavailable in the assessment platform.

Multiplication Table. A paper-based single digit (1–9) multiplication table is available for reference.

Print-on-Demand. This accommodation allows TAs to print paper copies of either passages/stimuli and/or items for students. For students needing a paper copy of a passage or stimulus, permission for the students to request printing must first be set in TIDE. The TC must fill out a Verification of Student Need Form and contact HIDOE to have the accommodation set for the student.

Read-Aloud (for ELA/L reading 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. Refer to the *Guidelines for Choosing the Read-Aloud Accommodation* when deciding if this accommodation is appropriate for a student.

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

Speech-to-Text. Voice recognition allows students to use their voices as input devices to the computer in order 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.

Word Completion. This allows students to begin writing a word and choose from a list of words that have been predicted from word frequency and syntax rules. Word prediction is delivered via a non-embedded software program. The program must use only single-word prediction. Functionality such as phrase prediction, predict ahead, or next word must be deactivated. The program must have settings that allow only a basic dictionary. Expanded dictionaries, such as topic dictionaries and word banks, must be deactivated. Phonetic spelling functionality and programs with built-in speech output that reads back the information the student has written may also be used. Students who use word prediction in conjunction with speech output will need headphones unless tested individually in a separate setting. Students may use their own assistive technology devices.

Table 5 presents a list of universal tools, designated supports, and accommodations that were offered in the 2024–2025 administration. Tables 6–11 present the numbers of students who were allowed to use each accommodation and/or designated support on the online ELA and mathematics assessments. Note that the overall count in the designated support tables may not match the sum of students in ELL and students with disabilities because some students are counted in both categories or because these features were approved for some students other than ELL and students with disabilities.

Table 5. 2024–2025 Universal Tools, Designated Supports, and Accommodations

Universal Tools	Designated Supports	Accommodations
	Embedded	
Breaks (Pause) Calculator[1] Digital Notepad English Dictionary[2] English Glossary Expandable Passages and/or Items Global Notes[3] Highlighter Keyboard Navigation Line Reader Mark for Review Math Tools[4] Spellcheck Strikethrough Thesaurus[2] Writing Tools[5] Zoom	Color Contrast Illustration Glossaries[6] Masking Mouse Pointer Streamline Text-to-Speech[7] Translated Student Interface Messages (in Spanish)[6] Translations (Glossaries)[6] Translations (Spanish)[6] Turn Off Any Universal Tools	American Sign Language[8] Braille Braille Transcript[9] Closed Captioning[9] Speech-to-Text Text-to-Speech[10] Word Completion
	Non-Embedded	
Breaks English Dictionary[2] Scratch Paper Thesaurus[2]	Amplification Bilingual Dictionary[2] Color Contrast Color Overlay Illustration Glossaries[11] Magnification Math Manipulatives[12] Medical Supports Noise Buffers Printed Test Directions in English Read-Aloud[13] Read-Aloud in Spanish[6] Scribe[14] Separate Setting Simplified Test Directions Translated Student Interface Messages Translated Test Directions in ASL Translations (Glossaries)[11]	100s Number Table Abacus Alternate Response Options[15] Braille[16] Calculator[1] Multiplication Table Print-on-Demand Read-Aloud[17] Scribe[2] Speech-to-Text Word Completion

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

[1] For calculator-allowed mathematics items only in grades 6–8 and 11

[2] For ELA/L performance task full-write items

[3] For ELA/L performance tasks

[4] Includes embedded ruler, embedded protractor

[5] Includes bold, italic, underline, indent, cut, paste, spellcheck, bullets, undo, redo

[6] For mathematics items

[7] For mathematics stimuli and items and ELA/L items (not for ELA/L reading passages): must be set in TIDE before test begins. Available in both English and Spanish for the mathematics tests.

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

[9] For ELA/L listening items

[10] For ELA/L reading passages. Must be set in TIDE by state-level user. TCs must submit a student's Verification of Need form to the Assessment Section for review and approval or disapproval.

[11] For mathematics paper-pencil tests

[12] Includes Algebra Tiles (recommended for grade 6 and above), Base Ten Blocks, Colored Tiles, Geoblocks Set, Geoboards and Geobands, Multi-Link Cubes, Pop Cubes, or Similar Cubes, Multi-Sensory Learning (MSL) Kit, One-Inch Blocks, Pattern Blocks, Transparent Sheets, and Two-Color Counters

[13] For mathematics stimuli and items and ELA/L items (not for ELA/L reading passages)

[14] For all items except for ELA/L performance task full-writes

[15] Includes adapted keyboards, large keyboard, Sticky Keys, MouseKeys, FilterKeys, adapted mouse, touch screen, head wand, and switches

[16] For paper-pencil assessments

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

Table 6. ELA/L Total Students with Allowed Embedded and Non-Embedded Accommodations

Accommodations	**Grade**						
	3	**4**	**5**	**6**	**7**	**8**	**11**
Embedded Accommodations							
American Sign Language	3	7	5	2	5	1	3
Braille				1	1		
Braille Transcript		4	1	1	1		
Closed Captioning	10	12	11	18	14	11	11
Speech-to-Text	1		1		2		1
Text-to-Speech: Reading Passages and Items					3	1	
Word Completion			1	1			
Non-Embedded Accommodations							
Alternate Response Options	1		3	1			
Read-Aloud Passages		1					
Scribe (Full-Write)	2	1		2	1	1	

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

Designated Supports	**Subgroup**	**Grade**						
		3	**4**	**5**	**6**	**7**	**8**	**11**
Color Contrast	Overall	1		8	2			
	ELL							
	Disability			7				
Masking	Overall	21	47	41	27	3	18	
	ELL	2	6	4	5	1	4	
	Disability	16	35	32	18	3	16	
Mouse Pointer	Overall	3	1	35		1	6	
	ELL						3	
	Disability	3	1	7		1	4	
Streamline	Overall	39	58	33	32	15	28	1
	ELL	4	5	6	4	5	5	
	Disability	21	38	15	28	12	26	1
Text-to-Speech: CAT Items	Overall	3,460	2,909	3,135	2,290	1,204	1,250	155
	ELL	784	649	644	459	322	319	41
	Disability	997	856	1,017	745	431	400	74
Text-to-Speech: PT Items	Overall	88	124	76	34	1	6	30
	ELL	17	16	7	1		2	30
	Disability	32	31	17	16	1	4	2
Text-to-Speech: PT Stimuli	Overall	4	1	6	1	2	1	
	ELL	1	1			1		
	Disability			5	1			
Text-to-Speech: PT Stimuli and Items	Overall	3,393	2,834	3,084	2,298	1,204	1,239	113
	ELL	762	638	641	495	355	352	10
	Disability	969	839	1,009	751	432	401	70

Table 8. ELA/L Total Students with Allowed Non-Embedded Designated Supports

Designated Supports	**Subgroup**	**Grade**						
		3	**4**	**5**	**6**	**7**	**8**	**11**
Amplification	Overall	3	1		1	1	1	
	ELL				1		1	
	Disability	3			1	1	1	
Bilingual Dictionary	Overall	6	3	7	3	8	16	24
	ELL	6	2	5	2	7	16	24
	Disability		2	1			3	3
Magnification	Overall			2	1	2	1	2
	ELL							
	Disability			2	1	1	1	1
Medical Supports	Overall	2	1	2	1		1	
	ELL							
	Disability	1	1	1	1			
Noise Buffers	Overall	2	3	3	1	3	4	1
	ELL			2				
	Disability	1	1	3		1	1	
Printed Test Directions in English	Overall	2	1		2			
	ELL				2			
	Disability	2	1		1			
Read-Aloud Items	Overall	58	57	58	9	6	3	2
	ELL	5	7	8		2	1	2
	Disability	39	36	38	9	5	3	1
Read-Aloud Stimuli	Overall	50	51	42	7	2	2	2
	ELL	5	5	6		1		2
	Disability	34	31	29	7	2	2	1
Scribe (Not Full-Write)	Overall	3	5	5	4	1	3	
	ELL		1	2				
	Disability	2	4	5	4	1	3	
Separate Setting	Overall	421	360	376	238	111	106	10
	ELL	65	45	45	30	12	12	
	Disability	320	289	312	193	84	76	2
Simplified Test Directions	Overall	116	103	110	64	11	15	3
	ELL	24	20	16	17	2	6	2
	Disability	78	72	73	43	7	11	2
Translated Student Interface Messages	Overall	3	3	4	1	1	3	
	ELL	3	2	4	1	1	3	
	Disability	1	2				1	
Translated Test Directions in ASL	Overall	3	4	2		2		2
	ELL	2	1					1
	Disability	3	4	2		2		2

Table 9. Mathematics Total Students with Allowed Embedded and Non-Embedded Accommodations

Accommodations	**Grade**						
	3	**4**	**5**	**6**	**7**	**8**	**11**
Embedded Accommodations							
American Sign Language	3	3	5	2	5	1	3
Braille				1	1		
Speech-to-Text					2		1
Non-Embedded Accommodations							
100s Number Table	24	11	6	12	3	4	
Abacus				1			
Alternate Response Options	2		3	1			
Calculator				3		1	
Multiplication Table		2		2	1		

Table 10. Mathematics Total Students with Allowed Embedded Designated Supports

Designated Supports	**Subgroup**	**Grade**						
		3	**4**	**5**	**6**	**7**	**8**	**11**
Color Contrast	Overall	1	1	7	1			
	ELL							
	Disability		1	7				
Illustration Glossaries	Overall	113	122	114	195	123	194	6
	ELL	74	80	61	157	120	174	6
	Disability	23	20	30	45	19	28	
Masking	Overall	21	38	40	27	3	19	
	ELL	2	6	4	5	1	4	
	Disability	16	26	31	18	3	17	
Mouse Pointer	Overall	2	1	35		1	6	
	ELL						3	
	Disability	2	1	7		1	4	
Streamline	Overall	40	58	34	30	15	28	1
	ELL	4	5	7	4	5	4	
	Disability	22	38	15	28	13	26	1
Text-to-Speech: Items	Overall	23	15	10	6	2	3	23
	ELL	9	2	1				23
	Disability	3	5	6	6	2	2	
Text-to-Speech: Stimuli	Overall	4	1	2	1		1	
	ELL	1						
	Disability	2						
Text-to-Speech: Stimuli and Items	Overall	3,558	3,040	3,207	2,430	1,293	1,335	160
	ELL	809	697	664	535	373	376	19
	Disability	1,020	887	1,045	786	453	419	78
Translations (Glossaries): Spanish	Overall	6	8	2	9	4	18	1
	ELL	6	8	2	8	4	17	1
	Disability		1		1		2	
Translations (Glossaries): Other Languages	Overall	4	26	19	9	18	39	5
	ELL	3	23	16	9	17	38	5
	Disability	1	1	2	1	2	4	
Translations (Dual Language): Spanish	Overall	3	8	5	11	10	10	3
	ELL	3	8	5	10	10	10	3
	Disability			1		1		

Table 11. Mathematics Total Students with Allowed Non-Embedded Designated Supports

Designated Supports	**Subgroup**	**Grade**						
		3	**4**	**5**	**6**	**7**	**8**	**11**
Amplification	Overall	3	1		1		2	
	ELL				1		1	
	Disability	3			1		1	
Illustration Glossaries	Overall	1	12	9	5	1	1	
	ELL	1	9	6	5	1	1	
	Disability			4				
Magnification	Overall			2	1	2	1	2
	ELL							
	Disability			2	1	1	1	1
Math Manipulatives	Overall	887	522	315	232	72	35	
	ELL	161	129	56	27	4	3	
	Disability	231	141	153	108	31	34	
Medical Supports	Overall	2	1	2	1		1	
	ELL							
	Disability	1	1	1	1			
Noise Buffers	Overall	2	3	2	1	2	4	1
	ELL			1				
	Disability	1	1	2			1	
Printed Test Directions in English	Overall	1	1		2			
	ELL				2			
	Disability	1	1		1			
Read-Aloud Items	Overall	57	41	55	10	3	3	2
	ELL	5	6	7		1	1	2
	Disability	39	21	32	10	3	3	1
Read-Aloud Items in Spanish	Overall		1	2	1	1		
	ELL		1	2	1	1		
	Disability							
Read-Aloud Stimuli	Overall	53	41	47	7	3	3	2
	ELL	5	6	5		1	1	2
	Disability	39	21	28	7	3	3	1
Read-Aloud Stimuli in Spanish	Overall		1	2		1		
	ELL		1	2		1		
	Disability							
Scribe	Overall	2	5	5	2	2	2	
	ELL		1	1				
	Disability	1	4	5	2	1	2	
Separate Setting	Overall	414	354	371	252	118	107	11
	ELL	62	45	48	31	13	12	
	Disability	303	283	305	194	89	75	2
Simplified Test Directions	Overall	124	99	115	63	8	15	3
	ELL	24	20	17	17	1	4	2
	Disability	77	69	75	39	5	10	2

Designated Supports	Subgroup	Grade						
		3	4	5	6	7	8	11
Translated Student Interface Messages	Overall	2	3	5	1	1		
	ELL	2	2	5	1	1		
	Disability		1					
Translated Test Directions in ASL	Overall	1	3	1	1	1	1	2
	ELL							1
	Disability	1	3	1	1	1	1	2
Translations (Glossaries): Spanish	Overall	1	1	2	1	2		
	ELL	1	1	2	1	2		
	Disability							
Translations (Glossaries): Other Languages	Overall	1	4	1	3	5	8	
	ELL	1	4	1	3	4	8	
	Disability						3	

2.6.3 Usage of Designated Supports and Accommodations

The Cambium Assessment, Inc. (CAI)'s test delivery system (TDS) collects usage data for certain accessibility resources that require student interaction. Among the designated supports and accommodations, the following tools were analyzed to determine how frequently they were used: *American Sign Language, Braille Transcript, Print-on-Demand , Speech-to-Text, Text-to-Speech, and Masking.* Tables 12 through 17 present the number of students allowed to use each accommodation or designated support and the percentage of those students who used it on at least one item or passage in ELA/L and mathematics, respectively.

Table 12. ELA/L: Number of Embedded Accessibility Resource Usages (Grades 3–5)

Accessibility Resources	Subgroup	Grade 3		Grade 4		Grade 5	
		N Allowed	% Used	N Allowed	% Used	N Allowed	% Used
Accommodation							
American Sign Language	Overall	3	66.7	7	42.9	5	20.0
	ELL	2	100	1	0.0	0	0.0
	Disability	3	66.7	7	42.9	5	20.0
Braille Transcript	Overall	0	0	4	0.0	1	0.0
	ELL	0	0	1	0.0	0	0.0
	Disability	0	0	4	0.0	1	0.0
Speech-to-Text	Overall	1	100.0	0	0.0	1	100.0
	ELL	0	0	0	0.0	0	0.0
	Disability	1	100.0	0	0.0	1	100.0
Text-to-Speech: Reading Passages and Items	Overall	0	0	0	0.0	0	0.0
	ELL	0	0	0	0.0	0	0.0
	Disability	0	0	0	0.0	0	0.0
Designated Support							
Masking	Overall	21	14.3	47	36.2	41	29.3
	ELL	2	0	6	50.0	4	0.0
	Disability	16	12.5	35	34.3	32	31.3
Text-to-Speech: CAT Items	Overall	3,460	54.0	2,909	52.5	3,135	53.5
	ELL	784	47.8	649	44.2	644	48.9
	Disability	997	63.2	856	62.9	1,017	65.3
Text-to-Speech: PT Items	Overall	88	54.5	124	40.3	76	32.9
	ELL	17	41.2	16	43.8	7	71.4
	Disability	32	46.9	31	54.8	17	23.5
Text-to-Speech: PT Passages	Overall	4	75.0	1	0.0	6	16.7
	ELL	1	0	1	0.0	0	0.0
	Disability	0	0	0	0.0	5	20.0
Text-to-Speech: PT Passages and Items	Overall	3,393	66.9	2,834	66.3	3,084	63.9
	ELL	762	65.0	638	60.8	641	61.5
	Disability	969	74.0	839	72.3	1,009	76.1

Table 13. ELA/L: Number of Embedded Accessibility Resource Usages (Grades 6–8)

Accessibility Resources	**Subgroup**	**Grade 6**		**Grade 7**		**Grade 8**	
		N Allowed	**% Used**	**N Allowed**	**% Used**	**N Allowed**	**% Used**
Accommodation							
American Sign Language	Overall	2	50.0	5	20.0	1	0.0
	ELL	0	0.0	1	0.0	0	0.0
	Disability	2	50.0	5	20.0	1	0.0
Braille Transcript	Overall	1	0.0	1	0.0	0	0.0
	ELL	0	0.0	1	0.0	0	0.0
	Disability	1	0.0	1	0.0	0	0.0
Speech-to-Text	Overall	0	0.0	2	0.0	0	0.0
	ELL	0	0.0	1	0.0	0	0.0
	Disability	0	0.0	2	0.0	0	0.0
Text-to-Speech: Reading Passages and Items	Overall	0	0.0	3	66.7	1	100.0
	ELL	0	0.0	0	0.0	0	0.0
	Disability	0	0.0	3	66.7	1	100.0
Designated Support							
Masking	Overall	27	33.3	3	33.3	18	27.8
	ELL	5	40.0	1	0.0	4	0.0
	Disability	18	16.7	3	33.3	16	31.3
Text-to-Speech: CAT Items	Overall	2,290	44.5	1,204	36.3	1,250	30.1
	ELL	459	36.8	322	35.4	319	22.9
	Disability	745	57.4	431	45.0	400	46.5
Text-to-Speech: PT Items	Overall	34	41.2	1	0.0	6	33.3
	ELL	1	100.0	0	0.0	2	50.0
	Disability	16	37.5	1	0.0	4	25.0
Text-to-Speech: PT Passages	Overall	1	0.0	2	0.0	1	0.0
	ELL	0	0.0	1	0.0	0	0.0
	Disability	1	0.0	0	0.0	0	0.0
Text-to-Speech: PT Passages and Items	Overall	2,298	56.5	1,204	46.0	1,239	43.0
	ELL	495	47.1	355	40.8	352	37.5
	Disability	751	69.0	432	54.2	401	54.1

Table 14. ELA/L: Number of Embedded Accessibility Resource Usages (Grade 11)

Accessibility Resources	**Subgroup**	**Grade 11**	
		N Allowed	**% Used**
Accommodation			
American Sign Language	Overall	3	0.0
	ELL	1	0.0
	Disability	3	0.0
Braille Transcript	Overall	0	0.0
	ELL	0	0.0
	Disability	0	0.0
Speech-to-Text	Overall	1	0.0
	ELL	0	0.0
	Disability	1	0.0
Text-to-Speech: Reading Passages and Items	Overall	0	0.0
	ELL	0	0.0
	Disability	0	0.0
Designated Support			
Masking	Overall	0	0.0
	ELL	0	0.0
	Disability	0	0.0
Text-to-Speech: CAT Items	Overall	155	26.5
	ELL	41	19.5
	Disability	74	29.7
Text-to-Speech: PT Items	Overall	30	13.3
	ELL	30	13.3
	Disability	2	0.0
Text-to-Speech: PT Passages	Overall	0	0.0
	ELL	0	0.0
	Disability	0	0.0
Text-to-Speech: PT Passages and Items	Overall	113	29.2
	ELL	10	30.0
	Disability	70	32.9

Table 15. Mathematics: Number of Embedded Accessibility Resource Usages (Grades 3–5)

Accessibility Resources	**Subgroup**	**Grade 3**		**Grade 4**		**Grade 5**	
		N Allowed	**% Used**	**N Allowed**	**% Used**	**N Allowed**	**% Used**
Accommodation							
American Sign Language	Overall	3	66.7	3	66.7	5	40.0
	ELL	2	100.0	0	0.0	0	0.0
	Disability	3	66.7	3	66.7	5	40.0
Speech-to-Text	Overall	0	0.0	0	0.0	0	0.0
	ELL	0	0.0	0	0.0	0	0.0
	Disability	0	0.0	0	0.0	0	0.0
Designated Support							
Masking	Overall	21	0.0	38	5.3	40	7.5
	ELL	2	0.0	6	0.0	4	25.0
	Disability	16	0.0	26	3.8	31	9.7
Text-to-Speech: Items	Overall	23	30.4	15	53.3	10	30.0
	ELL	9	22.2	2	50.0	1	100.0
	Disability	3	0	5	80.0	6	33.3
Text-to-Speech: Stimuli	Overall	4	0	1	0	2	0
	ELL	1	0	0	0	0	0
	Disability	2	0	0	0	0	0
Text-to-Speech: Stimuli and Items	Overall	3,558	54.6	3,040	50.9	3,207	46.7
	ELL	809	49.6	697	44.2	664	44.3
	Disability	1,020	65.5	887	62.2	1,045	60.6

Table 16. Mathematics: Number of Embedded Accessibility Resource Usages (Grades 6–8)

Accessibility Resources	**Subgroup**	**Grade 6**		**Grade 7**		**Grade 8**	
		N Allowed	**% Used**	**N Allowed**	**% Used**	**N Allowed**	**% Used**
Accommodation							
American Sign Language	Overall	2	0.0	5	20.0	1	0.0
	ELL	0	0.0	1	0	0	0.0
	Disability	2	0.0	5	20.0	1	0.0
Speech-to-Text	Overall	0	0.0	2	50.0	0	0.0
	ELL	0	0.0	1	100.0	0	0.0
	Disability	0	0.0	2	50.0	0	0.0
Designated Support							
Masking	Overall	27	14.8	3	0	19	5.3
	ELL	5	20.0	1	0	4	0.0
	Disability	18	22.2	3	0	17	5.9
Text-to-Speech: Items	Overall	6	50.0	2	50.0	3	33.3
	ELL	0	0	0	0	0	0
	Disability	6	50.0	2	50.0	2	50.0
Text-to-Speech: Stimuli	Overall	1	0	0	0	1	0
	ELL	0	0	0	0	0	0
	Disability	0	0	0	0	0	0
Text-to-Speech: Stimuli and Items	Overall	2,430	36.9	1,293	25.6	1,335	16.7
	ELL	535	33.3	373	27.9	376	12.5
	Disability	786	53.4	453	34.0	419	30.1

Table 17. Mathematics: Number of Embedded Accessibility Resource Usages (Grade 11)

Accessibility Resources	**Subgroup**	**Grade 11**	
		N Allowed	**% Used**
Accommodation			
American Sign Language	Overall	3	0
	ELL	1	0
	Disability	3	0
Speech-to-Text	Overall	1	0
	ELL	0	0
	Disability	1	0
Designated Support			
Masking	Overall	0	0
	ELL	0	0
	Disability	0	0
Text-to-Speech: Items	Overall	23	4.3
	ELL	23	4.3
	Disability	0	0
Text-to-Speech: Stimuli	Overall	0	0
	ELL	0	0
	Disability	0	0
Text-to-Speech: Stimuli and Items	Overall	160	10.0
	ELL	19	10.5
	Disability	78	14.1

2.7 TESTING TIME

The online environment allows item response time to be captured as the item page time (i.e., the time each item page is presented on the screen) 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. For discrete items, the page time is the time spent on one item; and, for stimulus-based items, it is the time spent on all items associated with a stimulus. 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 time than average overall. The length of a test session is determined by PRs or TCs who are knowledgeable about the class periods in the school's instructional schedule and the timing needs associated with the assessments. Students should be allowed extra time if they need it, but TAs must use their best professional judgment when allowing students extra time.

Tables 18 and 19 present the average testing time and the testing time at percentiles for the overall test, the computer-adaptive test (CAT) component, and the performance task (PT) component.

Table 18. ELA/L Testing Time

Grade	Average Testing Time (hh:mm)	SD of Testing Time (hh:mm)	Median Testing Time (hh:mm)	Testing Time in Percentiles (hh:mm)				
				75th	80th	85th	90th	95th
Overall Test								
3	2:43	1:49	2:18	3:29	3:50	4:18	5:03	6:16
4	3:17	2:18	2:42	4:09	4:37	5:10	5:58	7:31
5	3:28	2:13	2:58	4:25	4:51	5:26	6:12	7:34
6	3:10	1:58	2:42	3:52	4:14	4:45	5:31	6:56
7	2:54	1:41	2:30	3:34	3:57	4:25	5:07	6:08
8	2:53	1:34	2:35	3:36	3:55	4:18	4:51	5:49
11	1:54	0:59	1:45	2:21	2:32	2:45	3:03	3:37
CAT Component								
3	0:54	0:34	0:46	1:04	1:10	1:19	1:32	1:55
4	1:02	0:46	0:51	1:12	1:19	1:28	1:42	2:13
5	1:05	0:40	0:56	1:20	1:28	1:38	1:51	2:18
6	1:07	0:38	0:59	1:20	1:27	1:35	1:47	2:14
7	1:02	0:31	0:56	1:16	1:22	1:29	1:40	1:58
8	1:01	0:31	0:56	1:14	1:20	1:27	1:37	1:58
11	0:45	0:21	0:43	0:55	0:58	1:02	1:09	1:20
PT Component								
3	1:50	1:28	1:28	2:25	2:44	3:08	3:42	4:39
4	2:16	1:47	1:48	2:59	3:21	3:50	4:29	5:37
5	2:23	1:45	1:59	3:06	3:27	3:55	4:32	5:42
6	2:03	1:31	1:39	2:36	2:54	3:17	3:52	4:57
7	1:52	1:20	1:31	2:22	2:41	3:03	3:35	4:29
8	1:52	1:13	1:36	2:25	2:40	2:58	3:26	4:09
11	1:09	0:46	1:00	1:28	1:37	1:48	2:02	2:29

Note. SD: standard deviation

Table 19. Mathematics Testing Time

Grade	Average Testing Time (hh:mm)	SD of Testing Time (hh:mm)	Median Testing Time (hh:mm)	Testing Time in Percentiles (hh:mm)				
				75th	80th	85th	90th	95th
Overall Test (CAT Component)								
3	0:51	0:34	0:43	1:02	1:09	1:18	1:30	1:54
4	1:00	0:41	0:50	1:13	1:21	1:32	1:48	2:17
5	1:08	0:41	0:59	1:27	1:35	1:46	2:00	2:25
6	1:02	0:35	0:55	1:14	1:21	1:30	1:43	2:07
7	1:02	0:34	0:55	1:16	1:22	1:30	1:43	2:05
8	1:08	0:38	1:01	1:25	1:31	1:41	1:54	2:17
11	0:44	0:23	0:41	0:55	0:59	1:04	1:11	1:26

Note. SD: standard deviation

2.8 DATA FORENSICS PROGRAM

The validity of test scores depends on the integrity of the test administration. 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 the testing session, TA, and school. The flagging criteria used for these analyses are described in the following section 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 five 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 after the testing window closes.

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. The studentized residuals are computed to detect unusual residuals. 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, h_{ii} is the leverage from the regression for the ith student, and $\hat{e}_i$ is the residual for the ith student.

The variance of average residuals in the denominator is estimated in two components, conditioning on the 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),

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

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

2.8.2 Test-Taking Time

The summative assessments are not timed, and thus, individual test-taking times may vary across students. However, unusual test-taking times such as excessively shorter or longer test-taking times 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 questions, 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 question may not result in a flagged person-fit index. Of course, not all unlikely patterns indicate cheating, as in the case of a student who is able to guess a significant number of correct answers. Therefore, the evidence of person-fit index should be evaluated along with other testing irregularities to determine possible testing irregularities. The number of flagged students is summarized for every testing session, TA, and school.

The person-fit index is computed using a standardized log-likelihood statistic. Following Drasgow, Levine, and Williams (1985) and Sotaridona, Pornel, and Vallejo (2003), an 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,

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

where s = standard deviation of l_z values in an aggregate unit and n = number of students in an aggregate unit.

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, 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 the 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 section, is designed to recover from a failure of any component with little interruption. Each system is redundant, and crucial 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 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 section describes CAI's system architecture and how it recovers from device failures, Internet interruptions, and other problems.

2.9.1 High-Level System Architecture

Our 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 complex areas 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.

The key elements of the testing system, including the data integrity processes, are described in the following paragraphs.

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 of being unrecorded during testing.

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

- If connectivity is lost and restored within the designated time, 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 into 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 the 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 a redundant array of independent disks (RAID) systems to mitigate the risk of disk failure. Each response is stored on multiple independent disks.

One server operates 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 they are removed from service upon failure. Real-time student data are immediately recoverable from the satellite, backup hub, or hub (as described in the following paragraphs), 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 students to log in again within seconds or minutes of the failure without data loss. The hub manages this process. 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 equipped with RAID subsystems, providing the 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. Once the data are successfully stored, these servers notify the hub and satellites that it is safe to delete student data.

Quality Assurance System

The QA system gathers data that detect cheating, monitor real-time item function, and evaluate 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. A notification then goes out to CAI's psychometricians and project team immediately.

Database of Record

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

2.9.2 Automated Backup and Recovery

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

2.9.3 Other Disruption Prevention and Recovery Mechanisms

These testing systems are designed to be extremely fault-tolerant. The systems can withstand the 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 operate for up to 60 hours without refueling. In addition, with multiple refueling contracts 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 through their partnership 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.

- At the network level, there 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 can 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 the backup.

To summarize, the system's TDS is hosted in an industry-leading facility with redundant power, cooling systems, 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 2024–2025 OPERATIONAL TEST ADMINISTRATION

3.1 STUDENT POPULATION

All students enrolled in grades 3–8 and 11 in all public elementary and secondary schools must participate in the Smarter Balanced English language arts/literacy (ELA/L) and mathematics assessments. Before the testing window opened for the 2024–2025 test administration, the state or complex area sends CAI a student enrollment file to load to the Test Information Distribution Engine (TIDE). Using this enrollment file, the participation rates were calculated as the percentage of students who attempted the test. Tables 20 and 21 present the participation rates and the percentage of students who attempted the test by subgroups. Tables 22 and 23 present the number of Hawaiʻi students who met attemptedness requirements for scoring and reporting the results of the Smarter Balanced summative assessments.

Table 20. Participation Rates by Percentage: ELA/L

Group	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 11
All Students	94.8	95.2	95.6	94.9	94.3	94.2	93.1
Female	95.4	95.5	96.1	95.0	94.5	94.3	93.6
Male	94.3	94.9	95.2	94.9	94.2	94.1	92.5
African American	95.4	94.5	96.9	97.7	96.5	97.9	95.2
AmerIndian/Alaskan	87.5	100.0	71.4	90.5	81.0	96.0	100.0
Asian/Pacific Islander	97.3	97.8	97.4	97.5	96.9	97.4	96.0
Hispanic	94.8	95.6	95.9	94.8	93.8	93.4	92.3
Hawaiʻi Pacific Islander	90.6	89.9	91.5	89.7	89.3	89.0	89.0
White	97.0	97.0	98.1	97.8	96.7	97.0	93.1
Multi-Racial	96.4	97.0	97.0	96.8	97.0	95.5	93.8
ELL	95.4	94.6	91.8	91.0	91.2	91.1	85.2
Disadvantaged	94.3	95.0	94.9	94.3	93.4	92.4	90.3
Migrant	89.8	98.9	95.3	94.9	96.7	96.3	89.7
Disability	86.4	88.4	88.6	89.0	86.0	84.5	78.5

Note. AmerIndian/Alaskan = American Indian/Alaskan Native; ELL = English Language Learner; Disadvantaged = Economic Disadvantage Status

Table 21. Participation Rates by Percentage: Mathematics

Group	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 11
All Students	95.1	95.5	96.0	95.4	94.9	94.9	93.3
Female	95.6	95.9	96.4	95.5	95.2	95.0	93.9
Male	94.7	95.2	95.6	95.4	94.7	94.8	92.7
African American	95.9	95.9	96.9	97.7	96.5	97.9	95.2
AmerIndian/Alaskan	87.5	100.0	71.4	90.5	81.0	100.0	90.9
Asian/Pacific Islander	97.9	98.6	98.5	98.3	97.8	98.2	96.6
Hispanic	95.0	95.9	96.2	95.6	94.6	93.9	92.1
Hawaiʻi Pacific Islander	90.7	90.3	91.8	90.2	90.1	90.3	89.8
White	97.2	97.2	98.1	97.8	96.8	97.3	92.6
Multi-Racial	96.6	97.1	96.9	97.0	97.2	95.8	93.7
ELL	97.3	98.1	96.6	95.3	94.8	94.9	88.5
Disadvantaged	94.7	95.5	95.4	94.9	94.0	93.4	90.8
Migrant	89.8	98.9	95.3	95.4	96.2	96.3	89.7
Disability	87.0	88.6	88.9	89.1	86.4	85.6	78.9

Table 22. Number of Students: ELA/L

Group	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 11
All Students	12,666	12,193	12,779	12,642	11,960	11,960	11,189
Female	6,040	5,934	6,170	6,059	5,761	5,769	5,422
Male	6,626	6,259	6,609	6,582	6,199	6,190	5,767
African American	173	140	161	138	143	143	139
AmerIndian/Alaskan	14	19	5	21	17	25	12
Asian/Pacific Islander	2,608	2,759	2,939	2,969	2,931	3,158	3,426
Hispanic	2,588	2,321	2,427	2,525	2,327	2,286	1,960
Hawaiʻi Pacific Islander	3,033	2,791	3,020	2,960	2,864	2,761	2,423
White	1,513	1,408	1,444	1,360	1,173	1,283	1,144
Multi-Racial	2,737	2,755	2,783	2,669	2,504	2,304	2,085
ELL	1,494	1,361	1,098	1,088	1,238	1,166	767
Disadvantaged	5,844	5,452	5,717	5,606	5,262	4,980	4,121
Migrant	141	177	163	184	178	155	147
Disability	1,427	1,327	1,393	1,432	1,284	1,286	927

Table 23. Number of Students: Mathematics

Group	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 11
All Students	12,699	12,238	12,825	12,705	12,035	12,049	11,211
Female	6,048	5,963	6,186	6,093	5,802	5,807	5,443
Male	6,651	6,275	6,639	6,611	6,233	6,241	5,768
African American	174	142	161	137	143	143	138
AmerIndian/Alaskan	14	19	5	21	17	26	11
Asian/Pacific Islander	2,626	2,782	2,971	2,991	2,958	3,186	3,451
Hispanic	2,593	2,327	2,435	2,546	2,345	2,297	1,955
Hawaiʻi Pacific Islander	3,036	2,801	3,027	2,975	2,890	2,799	2,434
White	1,513	1,411	1,445	1,362	1,174	1,287	1,138
Multi-Racial	2,743	2,756	2,781	2,673	2,507	2,311	2,084
ELL	1,523	1,379	1,119	1,138	1,280	1,192	791
Disadvantaged	5,877	5,482	5,749	5,636	5,291	5,033	4,136
Migrant	141	175	162	185	177	155	143
Disability	1,435	1,334	1,407	1,435	1,290	1,302	930

3.2 Summary of Overall Student Performance

Tables 24–29 present a summary of the 2024–2025 summative test results for all students and by subgroup, including the average and the standard deviation of scale scores, the percentage of students in each achievement level, and the percentage of proficient students.

Figures 1 and 2 present the percentage of proficient students over the past six test administrations for all students (cohort comparisons). Figures 3 and 4 present the average scale scores in six test administrations for all students. In Figures 1–4, the 2019–2020 performance is not included because the testing was canceled due to the COVID-19 pandemic.

Appendix B, Student Performance Across Four Years for All Students and by Subgroup, provides the average and standard deviations of scale scores and the percentage of proficient students by subgroup for each test administration across four years.

Table 24. 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	12,666	2425.57	104.36	29	22	21	29	49
Female	6,040	2437.98	101.18	25	22	22	32	54
Male	6,626	2414.26	105.91	33	21	20	26	45
African American	173	2414.30	104.11	34	19	22	25	47
AmerIndian/Alaskan	14	2435.53	117.33	21	7	29	43	71
Asian/Pacific Islander	2,608	2458.71	99.95	17	19	23	40	63
Hispanic	2,588	2409.98	99.93	34	24	20	22	43
Hawaiʻi Pacific Islander	3,033	2379.77	95.70	45	25	17	13	30
White	1,513	2457.63	97.24	17	20	26	38	64
Multi-Racial	2,737	2442.44	104.24	25	19	21	35	56
ELL	1,494	2367.22	91.72	51	25	15	10	24
Disadvantaged	5,844	2395.61	98.39	39	25	19	18	36
Migrant	141	2378.22	88.86	42	29	19	10	29
Disability	1,427	2318.52	85.64	74	15	6	4	10
Grade 4								
All Students	12,193	2472.79	106.86	29	19	22	30	52
Female	5,934	2482.47	104.28	26	19	23	32	55
Male	6,259	2463.61	108.47	32	19	21	28	49
African American	140	2473.58	92.90	23	31	20	26	46
AmerIndian/Alaskan	19	2459.13	76.02	21	32	42	5	47
Asian/Pacific Islander	2,759	2503.17	105.50	20	16	23	42	64
Hispanic	2,321	2459.94	102.37	32	20	23	24	48
Hawaiʻi Pacific Islander	2,791	2421.89	99.07	49	19	17	15	32
White	1,408	2500.90	96.50	18	19	25	38	63
Multi-Racial	2,755	2490.45	104.88	22	19	23	35	58
ELL	1,361	2397.26	93.40	58	21	14	8	22
Disadvantaged	5,452	2439.79	102.71	41	21	19	19	38
Migrant	177	2422.05	100.99	50	18	17	16	33
Disability	1,327	2361.65	91.00	74	15	7	4	11
Grade 5								
All Students	12,779	2511.62	111.00	26	18	27	29	56
Female	6,170	2524.05	106.41	21	18	29	31	60
Male	6,609	2500.02	113.90	30	18	26	26	52
African American	161	2516.18	91.90	20	23	31	25	57
AmerIndian/Alaskan	5*							
Asian/Pacific Islander	2,939	2543.85	104.89	16	15	29	40	69
Hispanic	2,427	2494.41	106.21	31	20	27	22	49
Hawaiʻi Pacific Islander	3,020	2456.77	107.03	44	21	22	13	35
White	1,444	2549.93	99.05	14	15	32	39	71
Multi-Racial	2,783	2531.92	106.40	19	17	29	34	64
ELL	1,098	2404.41	93.56	65	20	12	3	15
Disadvantaged	5,717	2475.14	107.43	37	21	25	17	42
Migrant	163	2454.32	99.40	45	26	16	13	29
Disability	1,393	2385.22	95.01	73	16	8	3	11

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

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

Table 25. 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	12,642	2534.02	106.63	24	23	31	22	53
Female	6,059	2549.29	102.20	18	23	33	26	59
Male	6,582	2519.94	108.68	28	23	30	19	48
African American	138	2543.07	93.67	18	23	37	22	59
AmerIndian/Alaskan	21	2531.70	92.37	24	24	29	24	52
Asian/Pacific Islander	2,969	2566.31	101.46	14	20	35	32	67
Hispanic	2,525	2520.00	102.35	26	26	30	17	48
Hawaiʻi Pacific Islander	2,960	2480.39	98.06	42	27	24	8	32
White	1,360	2569.71	100.05	12	20	35	33	68
Multi-Racial	2,669	2552.21	103.51	18	21	34	27	60
ELL	1,088	2427.04	81.66	64	26	9	1	10
Disadvantaged	5,606	2499.09	102.42	34	27	27	12	39
Migrant	184	2474.03	98.55	42	28	26	4	30
Disability	1,432	2416.17	88.34	69	21	8	2	10
Grade 7								
All Students	11,960	2553.98	108.80	23	23	34	20	54
Female	5,761	2570.76	102.92	18	22	37	23	60
Male	6,199	2538.39	111.77	28	24	31	17	48
African American	143	2562.02	95.20	20	24	36	19	55
AmerIndian/Alaskan	17	2468.53	100.79	59	12	29	0	29
Asian/Pacific Islander	2,931	2590.20	102.09	12	19	38	30	68
Hispanic	2,327	2537.42	106.07	28	26	32	14	47
Hawaiʻi Pacific Islander	2,864	2502.00	103.21	40	27	26	7	33
White	1,173	2592.16	100.11	12	19	38	31	69
Multi-Racial	2,504	2568.68	103.01	18	22	39	22	60
ELL	1,238	2466.44	96.00	52	28	17	2	19
Disadvantaged	5,262	2522.60	107.38	33	26	30	12	41
Migrant	178	2501.35	108.20	38	29	26	7	33
Disability	1,284	2428.84	94.26	69	22	8	1	9
Grade 8								
All Students	11,960	2567.80	113.66	24	22	34	19	54
Female	5,769	2586.01	106.92	18	22	37	22	60
Male	6,190	2550.85	117.08	30	22	31	17	48
African American	143	2586.72	104.22	16	22	40	22	62
AmerIndian/Alaskan	25	2575.79	112.01	16	32	32	20	52
Asian/Pacific Islander	3,158	2606.00	107.74	14	18	39	30	68
Hispanic	2,286	2547.66	107.53	29	25	34	13	46
Hawaiʻi Pacific Islander	2,761	2510.26	106.22	42	27	24	7	31
White	1,283	2609.75	102.84	13	18	39	30	69
Multi-Racial	2,304	2579.77	109.38	20	21	38	21	59
ELL	1,166	2470.41	94.65	56	28	15	1	16
Disadvantaged	4,980	2530.27	110.00	35	26	29	10	39
Migrant	155	2514.00	106.65	37	30	24	8	32
Disability	1,286	2439.94	91.76	71	20	8	1	9

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

Table 26. Descriptive Statistics and Percentage of Students in Achievement Levels for Overall and by Subgroup: ELA/L (Grade 11)

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
				Grade 11				
All Students	11,189	2594.46	119.02	21	22	32	25	57
Female	5,422	2613.81	109.59	15	22	34	29	63
Male	5,767	2576.28	124.54	27	23	29	21	51
African American	139	2611.13	102.02	11	25	40	24	64
AmerIndian/Alaskan	12	2601.71	137.05	33	17	17	33	50
Asian/Pacific Islander	3,426	2629.90	108.60	12	18	36	34	70
Hispanic	1,960	2575.44	119.12	25	26	30	19	49
Hawaiʻi Pacific Islander	2,423	2540.27	112.35	35	29	25	11	36
White	1,144	2620.64	118.51	15	19	33	33	66
Multi-Racial	2,085	2601.59	117.88	19	20	34	27	60
ELL	767	2489.71	87.18	52	34	13	1	14
Disadvantaged	4,121	2561.13	117.31	29	26	29	16	45
Migrant	147	2551.31	109.83	30	32	26	12	38
Disability	927	2460.97	98.50	65	24	9	2	11

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

Table 27. 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	12,699	2437.52	95.93	27	20	27	26	53
Female	6,048	2436.33	91.16	27	22	28	24	52
Male	6,651	2438.61	100.07	28	19	26	28	54
African American	174	2416.35	79.57	33	26	27	14	41
AmerIndian/Alaskan	14	2427.99	81.53	29	21	29	21	50
Asian/Pacific Islander	2,626	2473.42	90.35	14	17	29	39	68
Hispanic	2,593	2419.96	90.04	33	22	27	19	45
Hawaiʻi Pacific Islander	3,036	2392.65	90.38	44	23	21	11	32
White	1,513	2465.96	88.64	16	17	31	36	67
Multi-Racial	2,743	2455.12	93.78	22	19	27	32	60
ELL	1,523	2387.76	91.40	47	23	19	11	30
Disadvantaged	5,877	2409.80	91.21	37	23	24	16	40
Migrant	141	2399.36	84.41	44	19	27	10	37
Disability	1,435	2338.03	92.88	71	15	10	5	15
Grade 4								
All Students	12,238	2485.03	96.62	21	28	25	26	50
Female	5,963	2480.89	90.69	22	31	24	23	47
Male	6,275	2488.96	101.78	21	26	25	28	53
African American	142	2475.32	80.22	18	39	25	17	42
AmerIndian/Alaskan	19	2454.63	66.43	21	37	37	5	42
Asian/Pacific Islander	2,782	2518.49	94.50	12	23	27	38	65
Hispanic	2,327	2469.64	91.89	25	31	24	20	44
Hawaiʻi Pacific Islander	2,801	2437.33	90.05	39	32	18	11	29
White	1,411	2509.69	88.35	13	25	30	32	62
Multi-Racial	2,756	2500.81	92.26	15	28	27	30	57
ELL	1,379	2422.08	91.97	45	32	14	9	23
Disadvantaged	5,482	2455.77	92.68	30	32	22	16	37
Migrant	175	2445.63	94.56	40	25	18	17	35
Disability	1,334	2388.43	92.15	61	25	10	5	14
Grade 5								
All Students	12,825	2508.56	105.07	31	24	18	26	45
Female	6,186	2506.27	98.70	31	26	19	24	43
Male	6,639	2510.69	110.65	31	23	18	29	47
African American	161	2502.40	80.38	29	34	16	21	37
AmerIndian/Alaskan	5*							
Asian/Pacific Islander	2,971	2548.67	103.18	18	21	20	40	61
Hispanic	2,435	2489.29	98.23	36	28	17	19	36
Hawaiʻi Pacific Islander	3,027	2456.01	98.74	50	26	14	11	25
White	1,445	2536.98	94.43	20	22	22	36	58
Multi-Racial	2,781	2525.35	98.67	24	24	21	31	52
ELL	1,119	2415.81	91.35	67	22	8	4	12
Disadvantaged	5,749	2474.49	100.01	43	26	16	16	31
Migrant	162	2459.68	98.26	51	28	10	12	22
Disability	1,407	2399.68	93.20	74	17	5	4	9

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

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

Table 28. 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	12,705	2521.05	118.26	32	27	18	23	41
Female	6,093	2521.76	112.20	31	28	19	22	40
Male	6,611	2520.37	123.57	33	25	18	24	42
African American	137	2521.12	104.19	29	28	26	18	43
AmerIndian/Alaskan	21	2502.77	102.46	29	43	14	14	29
Asian/Pacific Islander	2,991	2565.56	112.71	19	24	22	35	58
Hispanic	2,546	2499.98	113.03	38	29	16	17	33
Hawaiʻi Pacific Islander	2,975	2460.54	108.04	52	28	12	8	20
White	1,362	2556.68	107.89	19	28	23	31	53
Multi-Racial	2,673	2540.65	113.14	26	27	20	27	48
ELL	1,138	2416.51	102.75	72	20	5	4	8
Disadvantaged	5,636	2483.29	114.93	45	27	14	13	27
Migrant	185	2452.60	100.65	61	22	12	5	17
Disability	1,435	2394.34	106.11	78	15	4	3	7
Grade 7								
All Students	12,035	2524.92	120.51	36	26	20	18	38
Female	5,802	2523.44	117.15	37	27	20	17	37
Male	6,233	2526.29	123.55	36	25	20	19	39
African American	143	2534.36	96.20	31	32	22	15	37
AmerIndian/Alaskan	17	2460.17	98.43	59	18	24	0	24
Asian/Pacific Islander	2,958	2571.40	119.41	22	24	24	30	54
Hispanic	2,345	2502.51	112.88	43	27	18	12	30
Hawaiʻi Pacific Islander	2,890	2461.69	108.55	57	26	12	5	17
White	1,174	2567.72	110.16	23	25	22	30	52
Multi-Racial	2,507	2543.76	110.76	29	27	23	20	44
ELL	1,280	2430.44	107.06	71	18	8	3	11
Disadvantaged	5,291	2489.25	116.28	48	26	16	10	26
Migrant	177	2483.84	100.30	52	28	12	8	20
Disability	1,290	2401.13	99.10	81	13	4	2	6
Grade 8								
All Students	12,049	2535.60	131.11	41	24	17	19	35
Female	5,807	2537.54	124.69	40	25	17	18	35
Male	6,241	2533.79	136.80	42	23	16	19	36
African American	143	2533.26	114.31	37	31	17	14	31
AmerIndian/Alaskan	26	2527.38	109.98	35	27	31	8	38
Asian/Pacific Islander	3,186	2589.36	132.39	26	21	21	32	53
Hispanic	2,297	2507.11	117.58	49	25	15	11	26
Hawaiʻi Pacific Islander	2,799	2468.16	113.41	63	22	9	6	15
White	1,287	2576.54	119.35	27	25	21	26	48
Multi-Racial	2,311	2548.90	126.03	36	26	18	20	38
ELL	1,192	2438.75	111.77	74	15	6	4	10
Disadvantaged	5,033	2493.64	122.00	54	22	13	10	23
Migrant	155	2470.01	116.33	66	19	8	7	15
Disability	1,302	2398.91	99.43	86	10	3	1	4

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

Table 29. Descriptive Statistics and Percentage of Students in Achievement Levels for Overall and by Subgroup: Mathematics (Grade 11)

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
Grade 11								
All Students	11,211	2547.41	126.81	49	25	16	10	26
Female	5,443	2549.47	119.72	49	25	17	8	26
Male	5,768	2545.46	133.14	50	25	15	11	25
African American	138	2552.31	109.84	51	26	14	9	23
AmerIndian/Alaskan	11	2507.77	119.02	55	36	0	9	9
Asian/Pacific Islander	3,451	2592.16	125.27	34	28	23	15	38
Hispanic	1,955	2520.82	114.87	58	24	13	5	18
Hawaiʻi Pacific Islander	2,434	2489.47	109.69	70	19	8	3	11
White	1,138	2572.81	129.80	41	26	19	14	32
Multi-Racial	2,084	2551.93	125.61	47	27	16	10	26
ELL	791	2458.24	106.12	82	12	5	2	6
Disadvantaged	4,136	2511.77	118.58	62	22	11	5	16
Migrant	143	2485.37	101.65	76	13	8	3	11
Disability	930	2421.72	92.39	91	7	2	1	2

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

Figure 1. Percentage Proficient Across Years: ELA/L

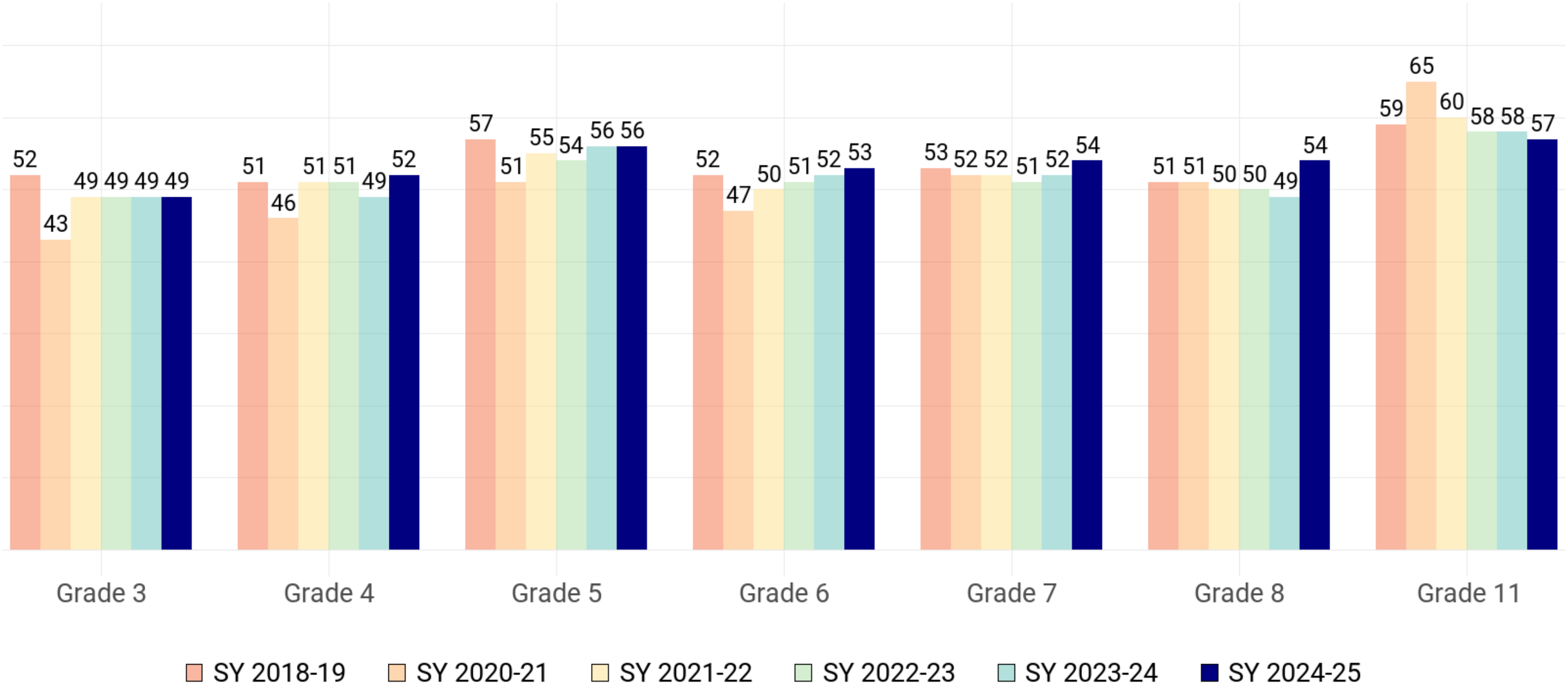


Figure 2. Percentage Proficient Across Years: Mathematics

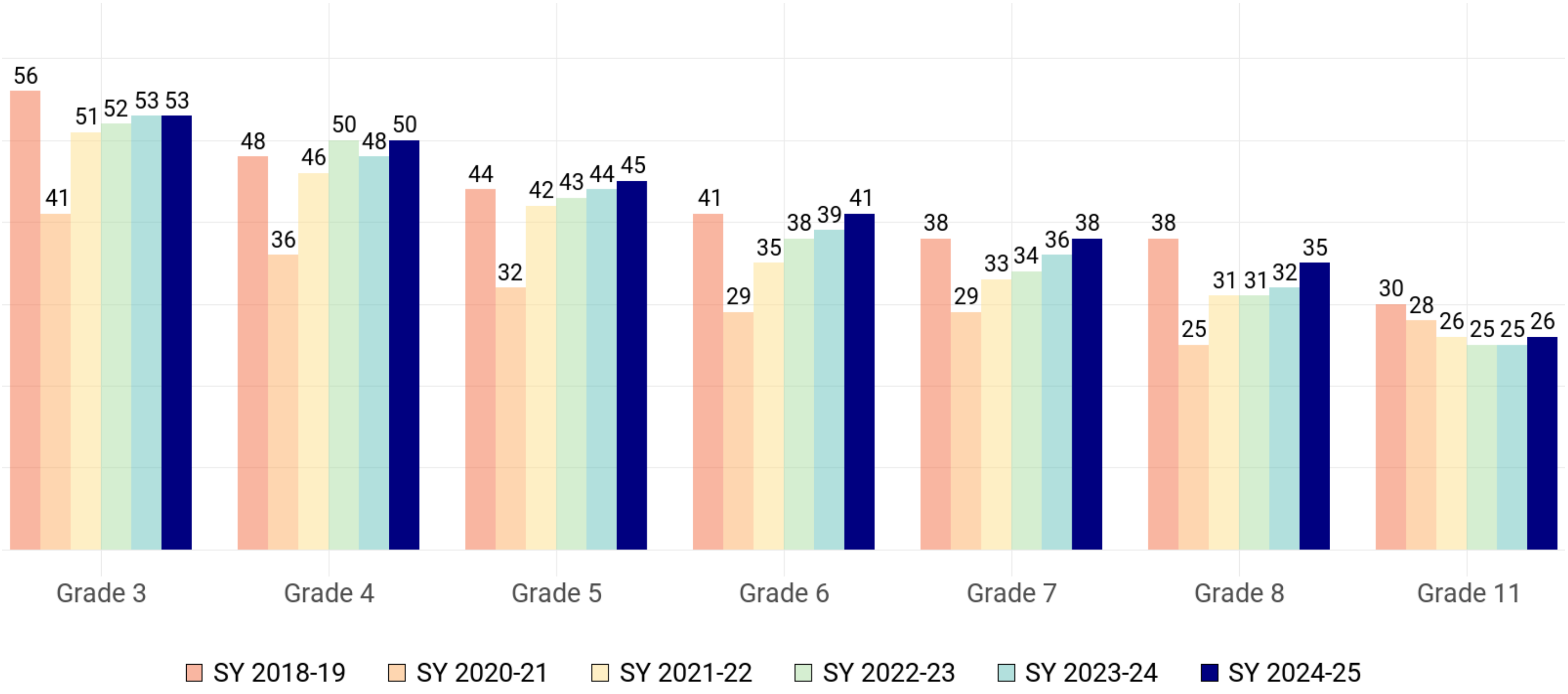


Figure 3. Average Scale Score Across Years: ELA/L

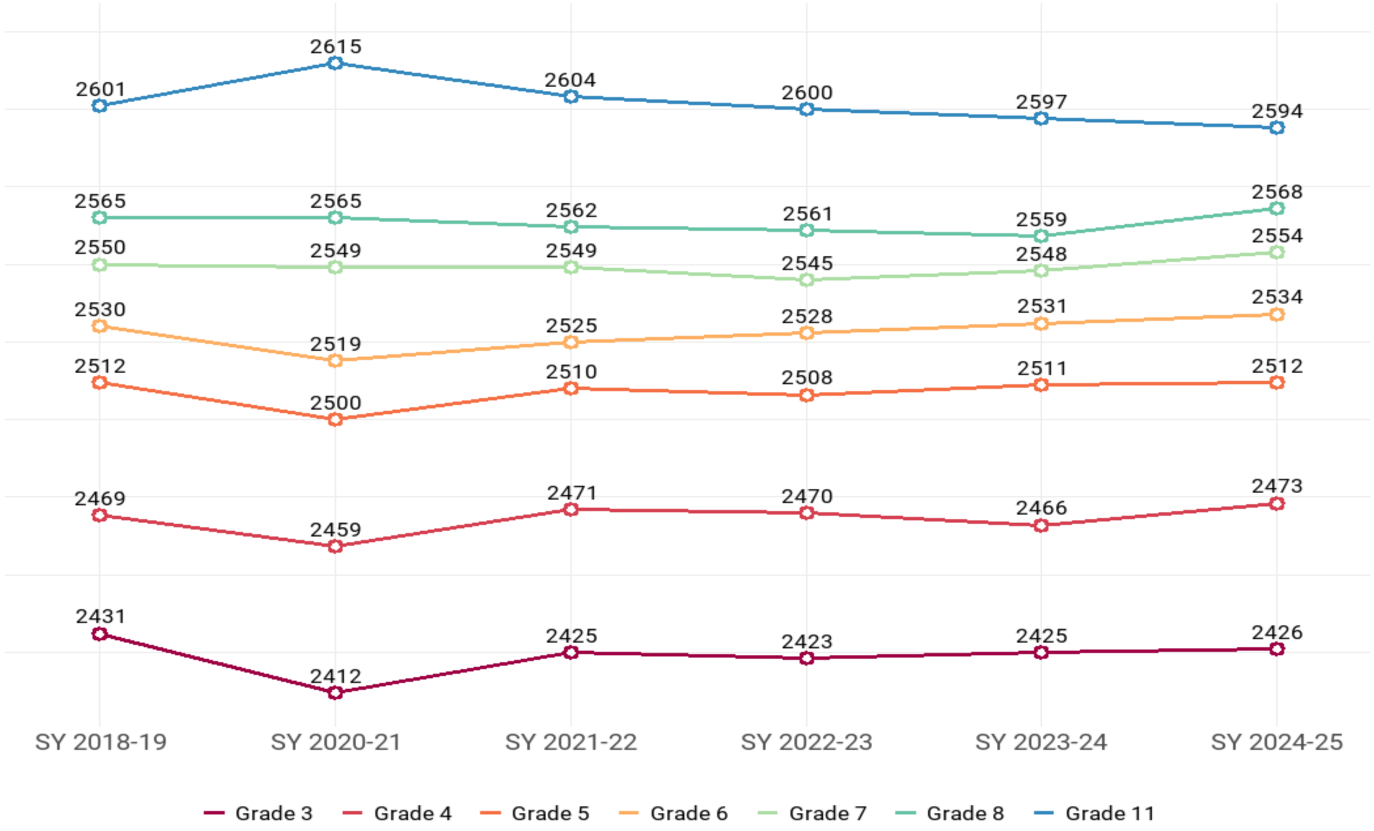
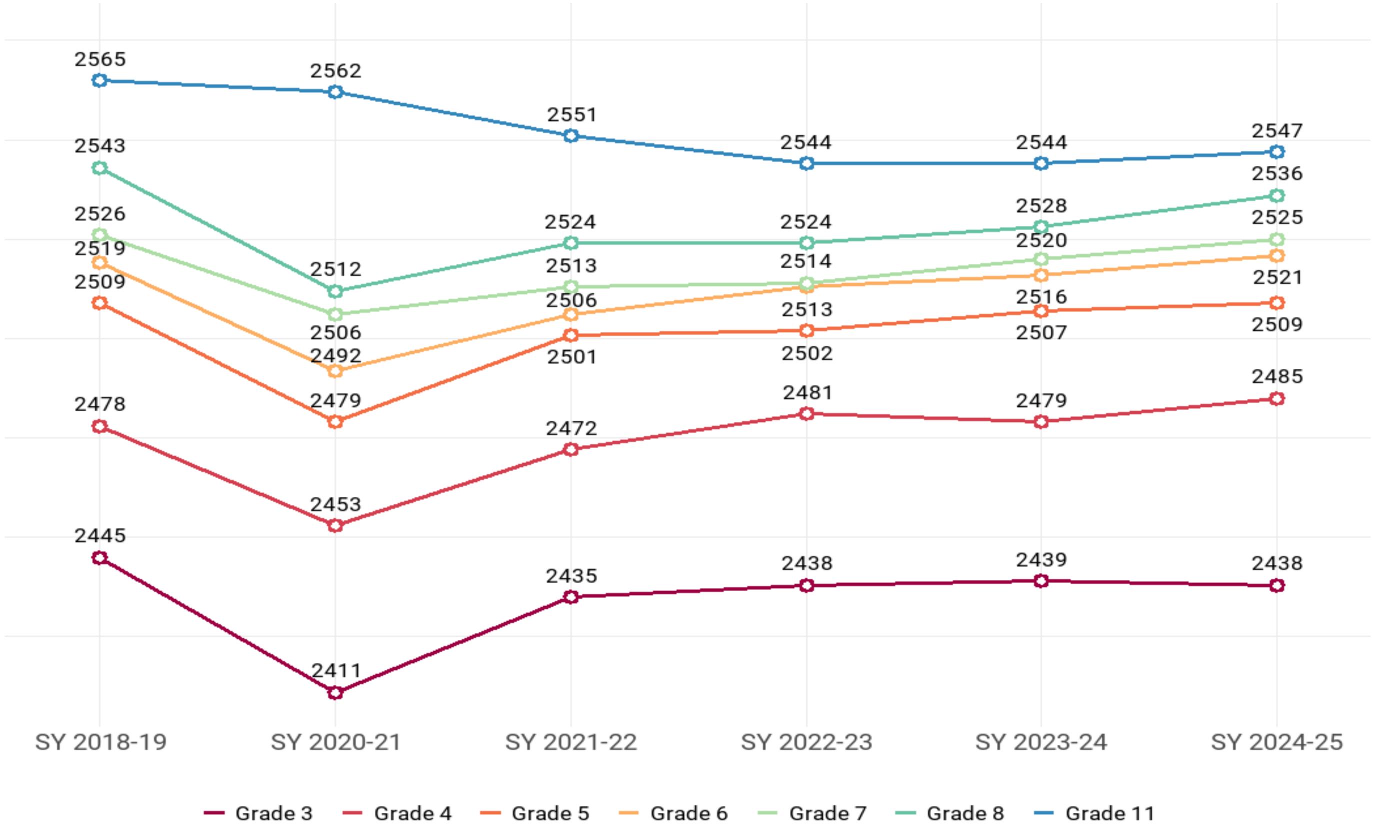


Figure 4. Average Scale Score Across Years: Mathematics



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, taking into account the standard error of measurement (SEM) of the claim score: (1) Below Standard, (2) At/Near Standard, or (3) Above Standard (see Section 6.5, Rules for Calculating Strengths and Weaknesses for Claim Scores, for the rules). Given the reduction in the number of items in Hawaiʻi's shortened blueprints, the reliabilities for claim scores are low, especially for Claim 3 and Claim 4 in ELA/L and Claims 2 and 4 combined and Claim 3 in mathematics. Therefore, starting with 2021–2022, the performance category for claim scores were reported only for Claims 1 and 2 in ELA/L and Claim 1 in mathematics at individual student level. Table 30 presents the distribution of performance categories for the reported claims.

Table 30. Percentage of Students in Performance Categories by Claim

Grade	Performance Category	ELA/L		Mathematics
		Claim 1: Reading	Claim 2: Writing	Claim 1: Concepts and Procedures
3	Below	22	28	27
	At/Near	60	51	40
	Above	18	22	33
4	Below	19	26	27
	At/Near	61	54	40
	Above	20	21	33
5	Below	20	23	32
	At/Near	59	51	40
	Above	22	26	28
6	Below	26	24	38
	At/Near	54	54	38
	Above	20	23	24
7	Below	21	22	39
	At/Near	59	51	40
	Above	19	26	21
8	Below	26	25	39
	At/Near	54	54	42
	Above	20	22	19
11	Below	20	20	54
	At/Near	58	53	34
	Above	22	27	12

3.3 DISTRIBUTION OF STUDENT ABILITY AND ITEM DIFFICULTY

Figures 5–10 display the empirical distribution of the Hawaiʻi student scale scores in the 2024–2025 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 shifted 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 level, the student ability distribution is shifted to the left for all claims except for Claim 2 grades 4–7 in ELA/L. In mathematics, the student ability distribution is shifted to the left for all claims except for Claim 1 in grades 3–5. 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, item difficulties) to better measure low-performing students.

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

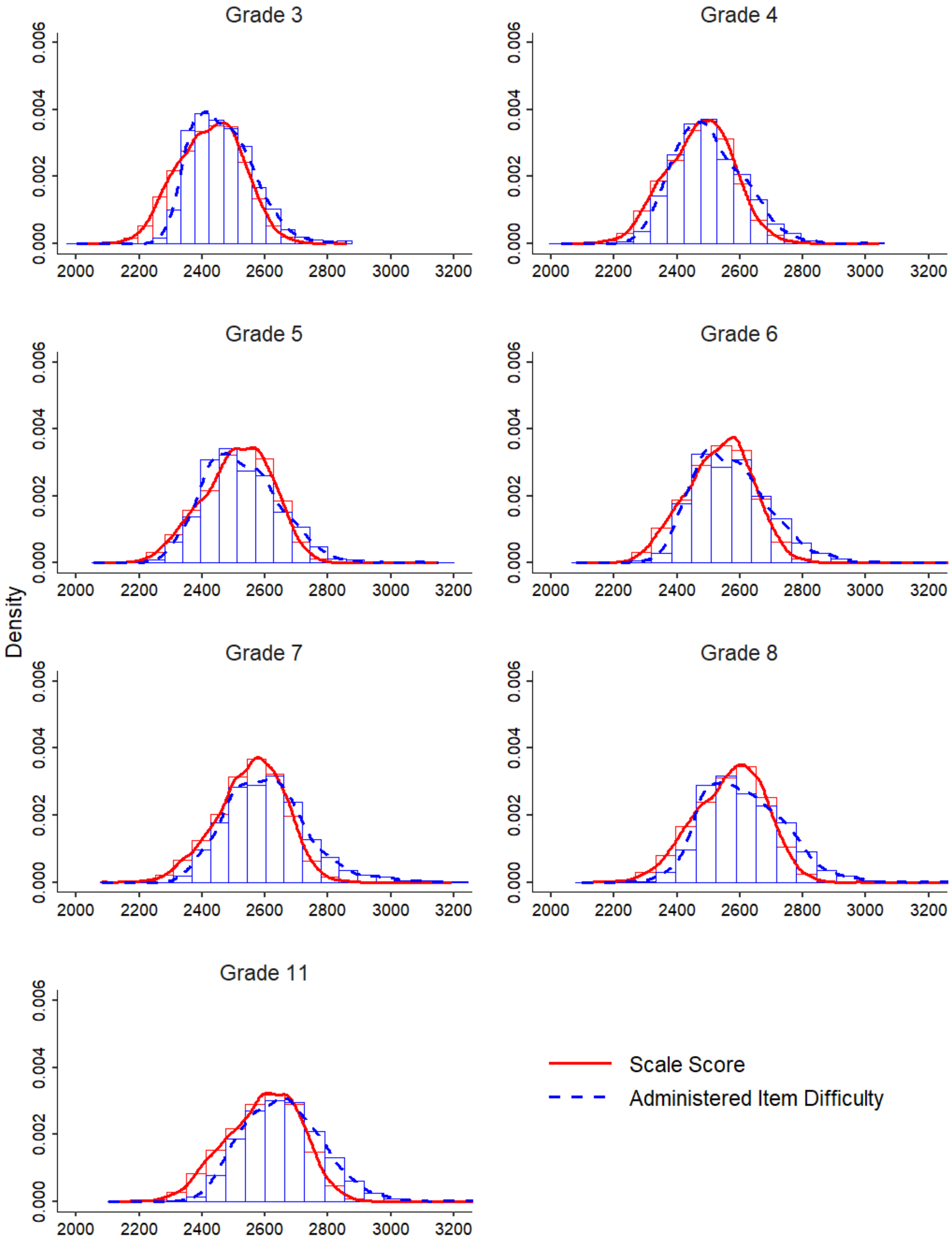


Figure 6. Student Ability—Item Difficulty Distribution by Claim: ELA/L (Grades 3–5)

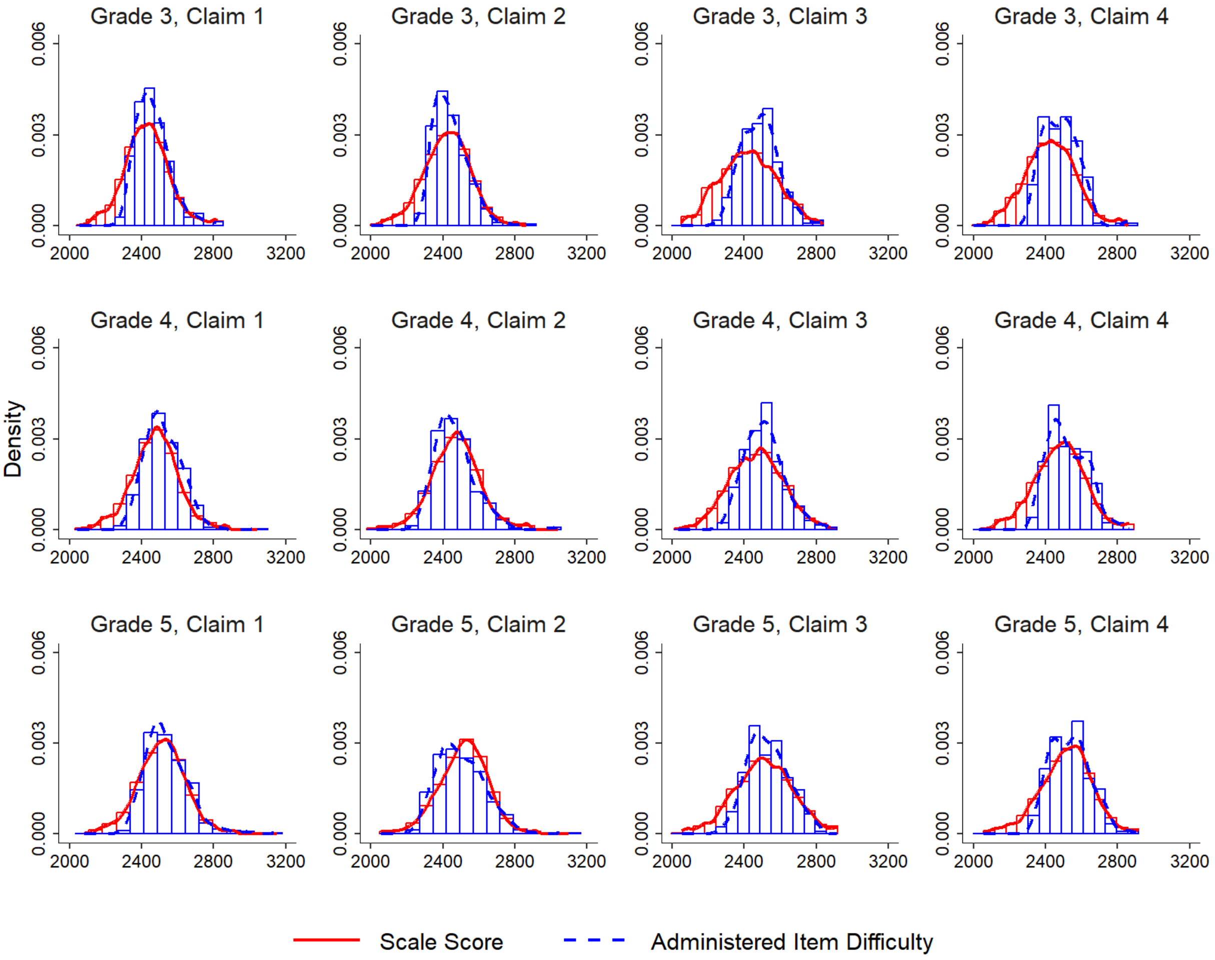


Figure 7. Student Ability—Item Difficulty Distribution by Claim: ELA/L (Grades 6–8, 11)

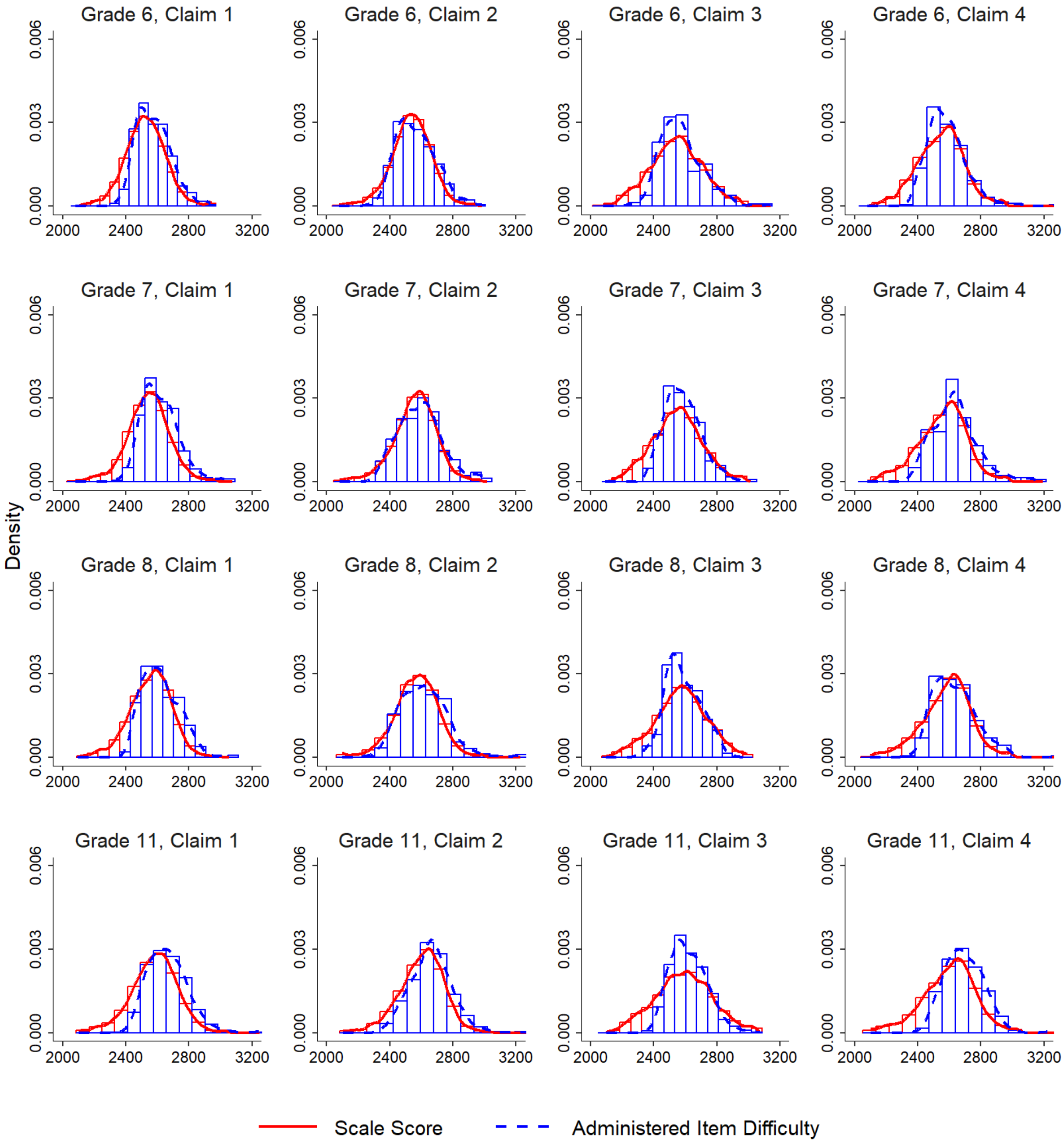


Figure 8. Student Ability—Item Difficulty Distribution: Mathematics

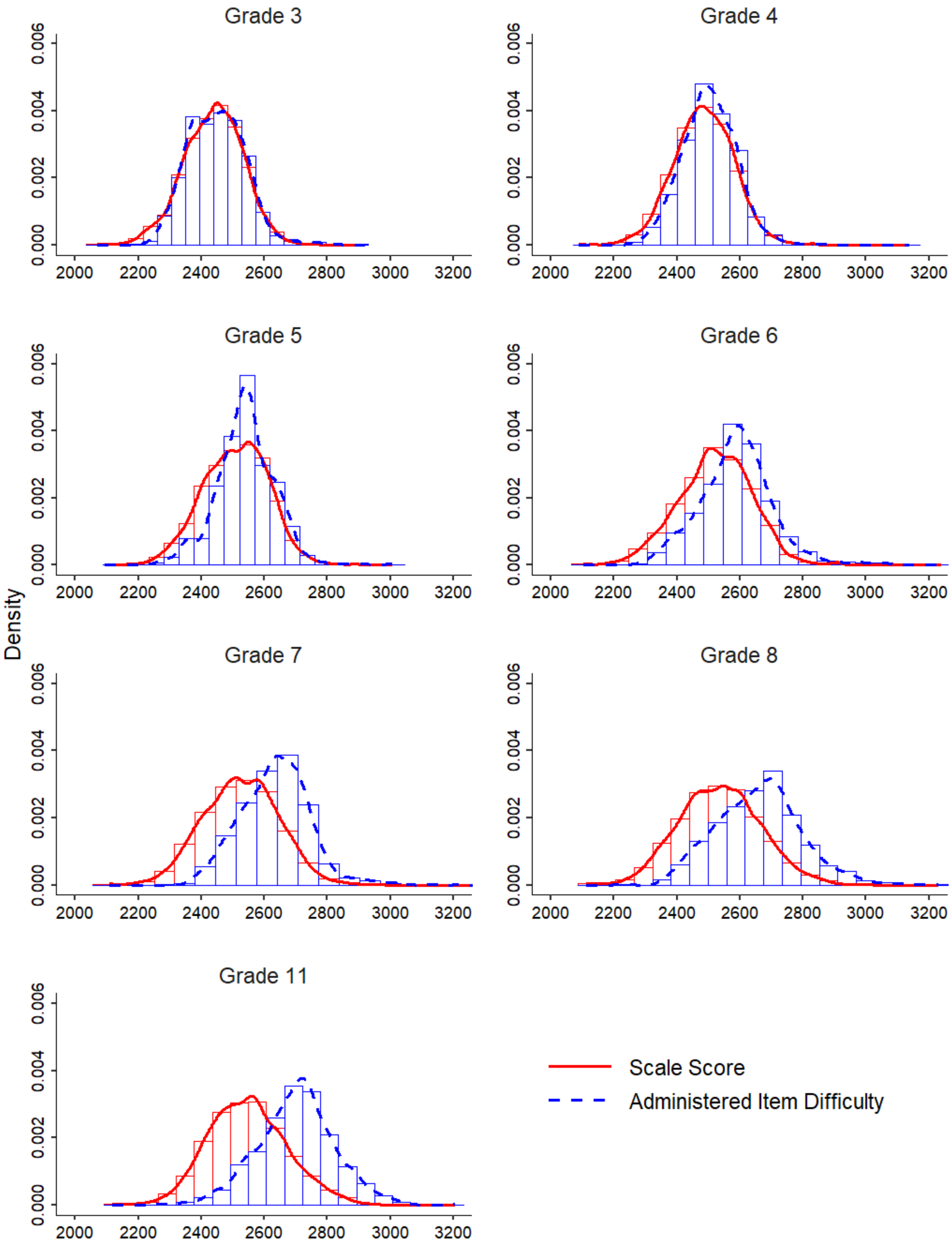


Figure 9. Student Ability—Item Difficulty Distribution by Claim: Mathematics (Grades 3–5)

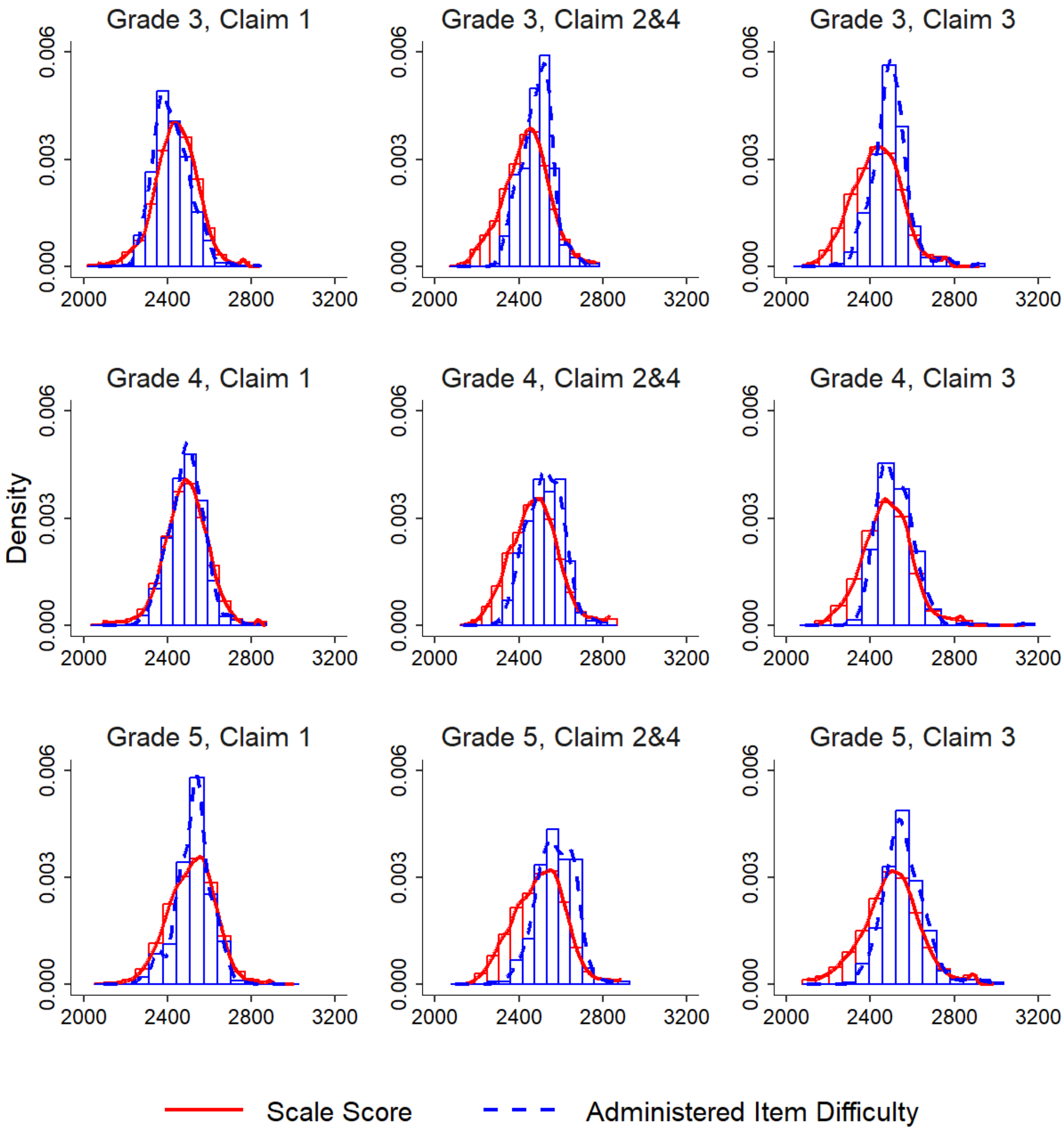
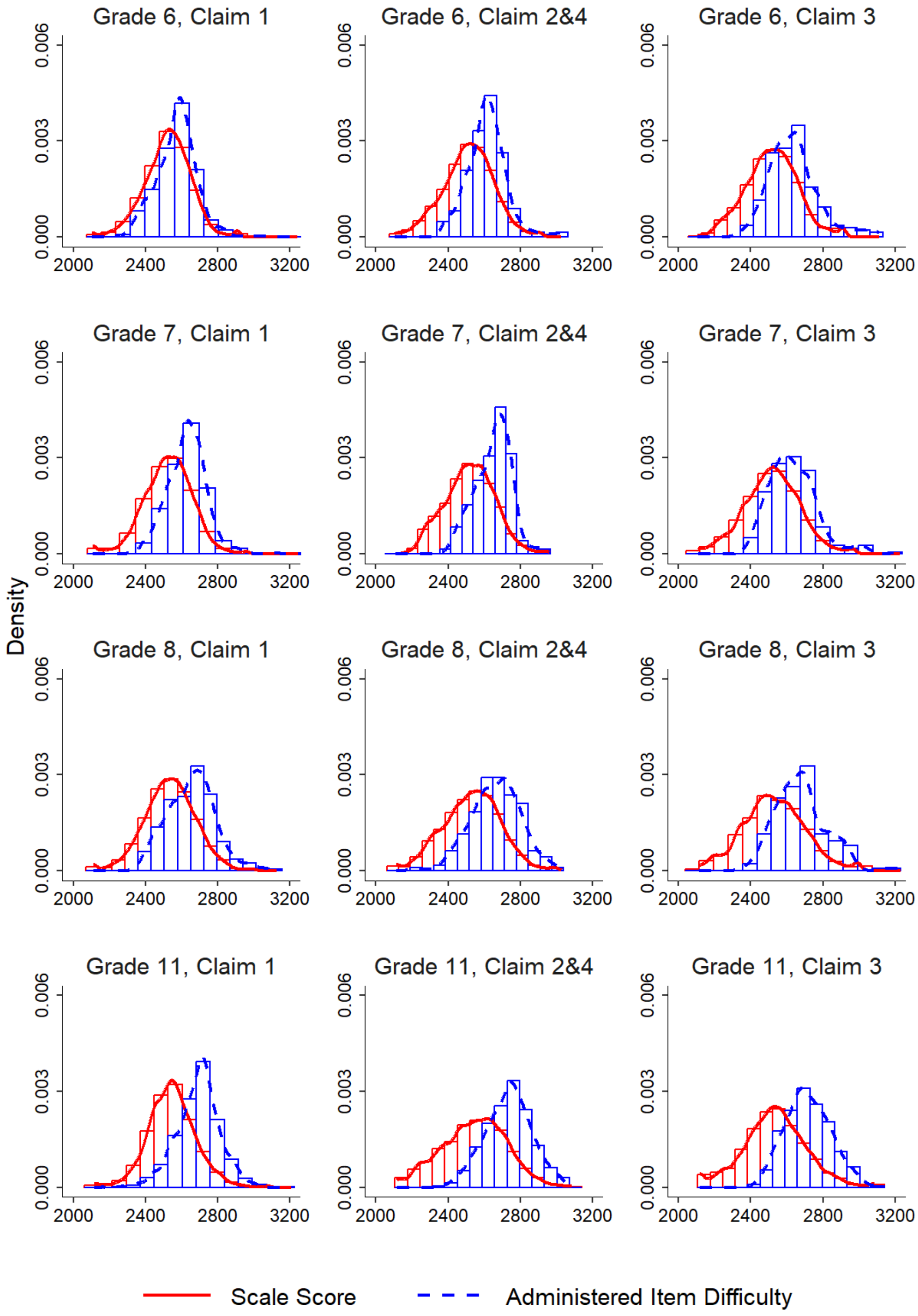


Figure 10. Student Ability—Item Difficulty Distribution by Claim: Mathematics (Grades 6–8, 11)



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 intercorrelations 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 a fixed-form test. The content coverage in all PT forms is the same. The test blueprint constraints for CAT and PT can be found at: https://smarterbalanced.alohahsap.org/resource-list/en/hawaii-shortened-summative-assessment-final-blueprints.

In the adaptive item-selection algorithm, item selection takes place in two discrete stages: blueprint satisfaction and match-to-ability. The blueprints specify a range of items to be administered in each claim, content domain/standard, and target. 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.

For the Smarter Balanced item pool, all items are developed in English. A portion of the English item pool was transcribed in braille or translated into Spanish to accommodate students who use braille and students who require tests administered in Spanish. The ELA/L pool is available in English and braille. The mathematics pool is available in English, braille, and Spanish. For each of these pools, a portion of items in each pool was further divided to accommodate American sign language (ASL), translations glossaries, and illustration glossaries. The translations glossaries and illustration glossaries were for mathematics items while the ASL was for mathematics items and listening items in ELA/L. Since the accommodated pools are small, some tests that use one or more accommodations to filter the pool have violations in some blueprint constraints.

Tables 31– 33 present the percentage of tests aligned with the ELA/L CAT test blueprint constraints for claims, targets, DOK, and number of passages. All tests met the blueprint requirements except for Claim 1 target or DOK requirements in one test each in grades 5 and 6 due to the uneven distribution of items across targets and DOKs within and across passages.

Tables 34–36 provide the percentage of tests aligned with the test blueprint constraints for the mathematics CAT for claim, DOK, and target constraints. All tests met all blueprint constraints, except for a few tests in grades 3, 4, 5, 7, and 8. The violations appeared on tests due to the application of pool filters limiting the item pool. Pool filters, such as using 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 31. Percentage of CAT Delivered Tests Meeting Blueprint Requirements: ELA/L (Grades 3–5)

Claim	Content Category/Target	Required Items/Passages	%BP Match		
			Grade 3	Grade 4	Grade 5
1	**Literary Text**	4	100.00	100.00	100.00
	Target 2: Central Ideas Target 4: Reasoning and Evaluation	1–3	100.00	100.00	100.00
	Targets 1, 3, 5, 6, and 7	1–3	100.00	100.00	100.00
	Long Literary Text Passage Short Literary Text Passage	1	100.00	100.00	100.00
	Informational Text	4	100.00	100.00	100.00
	Target 9: Central Ideas Target 11: Reasoning and Evaluation	1–3	100.00	100.00	100.00
	Targets 8, 10, 12, 13, and 14	1–3	100.00	100.00	100.00
	Long Informational Text Passage Short Informational Text Passage	1	100.00	100.00	100.00
	DOK 2	≥ 4	100.00	100.00	99.99
	DOK 3 or 4	≥ 1	100.00	100.00	100.00
2	**Writing**	5	100.00	100.00	100.00
	Target 1, 3, or 6: Organization/Purpose	1	100.00	100.00	100.00
	Target 1, 3, or 6: Evidence/Elaboration	1	100.00	100.00	100.00
	Target 8: Language and Vocabulary Use	1	100.00	100.00	100.00
	Target 9: Edit/Clarify	2	100.00	100.00	100.00
	DOK 2	≥ 2	100.00	100.00	100.00
3	**Listening**	4	100.00	100.00	100.00
	Target 4: Listen/Interpret	4	100.00	100.00	100.00
	DOK 2 or Higher	≥ 2	100.00	100.00	100.00
	Listening Passage	2	100.00	100.00	100.00
4	**Research**	5	100.00	100.00	100.00
	Target 2: Interpret and Integrate Information	1–2	100.00	100.00	100.00
	Target 3: Analyze Information/Sources	1–2	100.00	100.00	100.00
	Target 4: Use Evidence	1–2	100.00	100.00	100.00

Table 32. Percentage of CAT Delivered Tests Meeting Blueprint Requirements: ELA/L (Grades 6–8)

Claim	**Content Category/Target**	**Required Items/Passages**	**%BP Match**		
			Grade 6	**Grade 7**	**Grade 8**
1	**Literary Text**	4	100.00	100.00	100.00
	Target 2: Central Ideas Target 4: Reasoning and Evaluation	1–3	100.00	100.00	100.00
	Targets 1, 3, 5, 6, and 7	1–3	100.00	100.00	100.00
	Long Literary Text Passage	1	100.00	100.00	100.00
	Informational Text	6	100.00	100.00	100.00
	Target 9: Central Ideas Target 11: Reasoning and Evaluation	2–4	99.99	100.00	100.00
	Targets 8, 10, 12, 13, and 14	2–4	99.99	100.00	100.00
	Long Informational Text Passage	1	100.00	100.00	100.00
	Short Informational Text Passage	1	100.00	100.00	100.00
	DOK 1	≤ 3	100.00	100.00	100.00
	DOK 3 or Higher	≥ 1	100.00	100.00	100.00
2	**Writing**	5	100.00	100.00	100.00
	Target 1, 3, or 6: Organization/Purpose	1	100.00	100.00	100.00
	Target 1, 3, or 6: Evidence/Elaboration	1	100.00	100.00	100.00
	Target 8: Language and Vocabulary Use	1	100.00	100.00	100.00
	Target 9: Edit/Clarify	2	100.00	100.00	100.00
	DOK 2	≥ 2	100.00	100.00	100.00
3	**Listening**	4	100.00	100.00	100.00
	Target 4: Listen/Interpret	4	100.00	100.00	100.00
	DOK 2 or Higher	≥ 2	100.00	100.00	100.00
	Listening Passage	2	100.00	100.00	100.00
4	**Research**	5	100.00	100.00	100.00
	Target 2: Analyze/Integrate Information	1–2	100.00	100.00	100.00
	Target 3: Evaluate Information/Sources	1–2	100.00	100.00	100.00
	Target 4: Use Evidence	1–2	100.00	100.00	100.00

Table 33. Percentage of CAT Delivered Tests Meeting Blueprint Requirements: ELA/L (Grade 11)

Claim	**Content Category/Target**	**Required Items/Passages**	**%BP Match**
			Grade 11
1	**Literary Text**	4	100.00
	Target 2: Central Ideas Target 4: Reasoning and Evaluation	1–3	100.00
	Targets 1, 3, 5, 6, and 7	1–3	100.00
	Long Literary Text Passage	1	100.00
	Informational Text	6	100.00
	Target 9: Central Ideas Target 11: Reasoning and Evaluation	2–4	100.00
	Targets 8, 10, 12, 13, and 14	2–4	100.00
	Long Informational Text Passage	1	100.00
	Short Informational Text Passage	1	100.00
	DOK 1	≤ 2	100.00
	DOK 3 or Higher	≥ 2	100.00
2	**Writing**	5	100.00
	Target 1, 3, or 6: Organization/Purpose	1	100.00
	Target 1, 3, or 6: Evidence/Elaboration	1	100.00
	Target 8: Language and Vocabulary Use	1	100.00
	Target 9: Edit/Clarify	2	100.00
	DOK 2	≥ 2	100.00
3	**Listening**	4	100.00
	Target 4: Listen/Interpret	4	100.00
	DOK 2 or Higher	≥ 2	100.00
	Listening Passage	2	100.00
4	**Research**	5	100.00
	Target 2: Analyze/Integrate Information	1–2	100.00
	Target 3: Evaluate Information/Sources	1–2	100.00
	Target 4: Use Evidence	1–2	100.00

Table 34. Percentage of CAT Delivered Tests Meeting Blueprint Requirements for Claims and Targets: Mathematics (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	12	100.00	12	100.00	12	100.00
	DOK 2 or Higher	≥ 4	100.00	≥ 4	100.00	≥ 4	100.00
	Priority Cluster	9	100.00				
	Targets B, C, G, I	4	100.00				
	Targets D, F	4	100.00				
	Target A	1	100.00				
	Supporting Cluster	3	100.00				
	Targets E, J, K	2	100.00				
	Target H	1	100.00				
	Priority Cluster			9	100.00		
	Targets A, E, F			5	100.00		
	Target G			2	100.00		
	Target D			1	100.00		
	Target H			1	100.00		
	Supporting Cluster			3	100.00		
	Targets I, K			1	100.00		
	Targets B, C, J			1	100.00		
	Target L			1	100.00		
	Priority Cluster					9	100.00
	Targets E, I					4	100.00
	Target F					3	100.00
	Targets C, D					2	100.00
	Supporting Cluster					3	100.00
	Targets J, K					2	100.00
	Targets A, B, G, H					1	100.00
2 & 4	Overall	5	100.00	5	99.99	5	100.00
	DOK 3 or Higher	≥ 2	99.99	≥ 2	99.93	≥ 2	99.88
	2. Target A	1	100.00	1	99.97	1	100.00
	2. Targets B, C, D	1	100.00	1	99.98	1	100.00
	4. Targets A, D	1	99.99	1	100.00	1	99.99
	4. Targets B, E	1	100.00	1	100.00	1	99.99
	4. Targets C, F	1	99.99	1	100.00	1	100.00
3	Overall	5	100.00	5	99.99	5	100.00
	DOK 3 or Higher	≥ 2	100.00	≥ 2	100.00	≥ 2	100.00
	Targets A, D	2	100.00	2	99.99	2	100.00
	Targets B, E	2	100.00	2	100.00	2	100.00
	Targets C, F	1	100.00	1	100.00	1	100.00

Table 35. Percentage of CAT Delivered Tests Meeting Blueprint Requirements for Claims and Targets: Mathematics (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	12	100.00	12	100.00	12	100.00
	DOK 2 or Higher	≥ 4	100.00	≥ 4	100.00	≥ 4	100.00
	Priority Cluster	9	100.00				
	Targets E, F	4	100.00				
	Target A	2	100.00				
	Targets G, B	2	100.00				
	Target D	1	100.00				
	Supporting Cluster	3	100.00				
	Targets C, H, I, J	3	100.00				
	Priority Cluster			9	99.64		
	Targets A, D			5	100.00		
	Targets B, C			4	99.64		
	Supporting Cluster			3	99.64		
	Targets E, F			2	99.64		
	Targets G, H, I			1	100.00		
	Priority Cluster					9	99.98
	Targets C, D					3	99.97
	Targets B, E, G					3	99.98
	Targets F, H					3	100.00
	Supporting Cluster					3	99.98
	Targets A, I, J					3	99.98
2 & 4	Overall	5	100.00	5	100.00	5	100.00
	DOK 3 or Higher	≥ 2	100.00	≥ 2	99.97	≥ 2	100.00
	2. Target A	1	100.00	1	100.00	1	100.00
	2. Targets B, C, D	1	100.00	1	100.00	1	100.00
	4. Targets A, D	1	100.00	1	100.00	1	100.00
	4. Targets B, E	1	100.00	1	100.00	1	100.00
	4. Targets C, F	1	100.00	1	100.00	1	100.00
3	Overall	5	100.00	5	100.00	5	100.00
	DOK 3 or Higher	≥ 2	100.00	≥ 2	99.99	≥ 2	100.00
	Targets A, D	2	100.00	2	100.00	2	100.00
	Targets B, E	2	100.00	2	100.00	2	100.00
	Targets C, F, G	1	100.00	1	100.00	1	100.00

Table 36. Percentage of CAT Delivered Tests Meeting Blueprint Requirements for Claims and Targets: Mathematics (Grade 11)

Claim	**Content Domain**	**Grade 11**	
		Required Items	**% BP Match**
1	Overall	14	100.00
	DOK 2 or Higher	≥ 4	100.00
	Priority Cluster	10	100.00
	Targets D, E	1–2	100.00
	Target F	1	100.00
	Targets G, H, I	3	100.00
	Target J	1–2	100.00
	Target K	1–2	100.00
	Targets L, M, N	2	100.00
	Supporting Cluster	4	100.00
	Target O	0–2	100.00
	Target P	0–2	100.00
	Targets A, B	0–1	100.00
	Target C	0–1	100.00
2 & 4	Overall	5	100.00
	DOK 3 or Higher	≥ 2	100.00
	2. Target A	1	100.00
	2. Targets B, C, D	1	100.00
	4. Targets A, D	1	100.00
	4. Targets B, E	1	100.00
	4. Targets C, F	1	100.00
3	Overall	5	100.00
	DOK 3 or Higher	≥ 2	100.00
	Targets A, D	2	100.00
	Targets B, E	2	100.00
	Targets C, F, G	1	100.00

Table 37 summarizes target coverage by claim and includes the average and range of the number of unique targets administered in each delivered CAT component. The Smarter Balanced blueprints for ELA/L did not require every target to be covered in a claim; therefore, all targets listed in the blueprint are not expected to be covered in every test. Although the target coverage varies somewhat across individual tests, all targets are covered at an aggregate level across all tests combined.

Table 37. Average and Range of the Number of Unique Targets Assessed within Each Claim Across All Delivered CAT Tests

Grade	Total Targets in BP				Average				Range (Minimum–Maximum)			
	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4
ELA/L												
3	14	5	1	3	7.4	4.0	1.0	3.0	4–8	4–4	1–1	3–3
4	14	5	1	3	7.8	4.0	1.0	3.0	6–8	4–4	1–1	3–3
5	14	5	1	3	7.4	4.0	1.0	3.0	5–8	4–4	1–1	3–3
6	14	5	1	3	9.2	4.0	1.0	3.0	7–10	4–4	1–1	3–3
7	14	5	1	3	9.3	4.0	1.0	3.0	7–10	4–4	1–1	3–3
8	14	5	1	3	9.1	4.0	1.0	3.0	8–10	4–4	1–1	3–3
11	14	5	1	3	8.3	4.0	1.0	3.0	6–10	4–4	1–1	3–3
Mathematics												
3	11	4	6	6	10.0	2.0	4.1	3.0	9–10	2–2	3–5	3–3
4	12	4	6	6	9.0	2.0	4.1	3.0	9–9	1–2	3–5	3–3
5	11	4	6	6	8.0	2.0	4.1	3.0	7–8	2–2	3–5	2–3
6	10	4	7	6	9.0	2.0	3.8	3.0	8–9	2–2	3–5	3–3
7	9	4	7	6	6.9	2.0	4.1	3.0	6–7	2–2	3–5	3–3
8	10	4	7	6	10.0	2.0	4.2	3.0	8–10	2–2	3–5	3–3
11	16	4	7	6	12.7	2.0	3.9	3.0	10–14	2–2	3–5	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 distinct claims under each subject. The item selection algorithm prioritizes blueprint matching to ensure 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 assessments measure a single underlying ability, and that 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 38 and 39. The correction for attenuation indicates what the correlation would be if claim scores could be measured with perfect reliability and corrected (adjusted) for measurement error estimates.

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

When corrected for attenuation (above diagonal), the correlations among claim scores are higher than observed correlations. The disattenuated correlations are quite high in both subjects, showing evidence of unidimensional tests. The correction for attenuation is large in both ELA/L and mathematics because the marginal reliabilities of claim scores are low due to the reduction in the test length.

Table 38. Correlations Among Claims: ELA/L

Grade	**Claim**	**Observed & Disattenuated Correlation**			
		Claim 1	**Claim 2**	**Claim 3**	**Claim 4**
3	Claim 1: Reading		0.92	1	0.95
	Claim 2: Writing	0.62		1	0.93
	Claim 3: Listening	0.48	0.49		1
	Claim 4: Research	0.57	0.61	0.47	
4	Claim 1: Reading		0.92	1	0.94
	Claim 2: Writing	0.61		1	0.92
	Claim 3: Listening	0.53	0.52		1
	Claim 4: Research	0.56	0.60	0.5	
5	Claim 1: Reading		0.91	1	0.94
	Claim 2: Writing	0.61		1	0.94
	Claim 3: Listening	0.53	0.53		1
	Claim 4: Research	0.58	0.63	0.52	
6	Claim 1: Reading		0.89	1	0.92
	Claim 2: Writing	0.63		1	0.91
	Claim 3: Listening	0.53	0.52		1
	Claim 4: Research	0.59	0.60	0.48	
7	Claim 1: Reading		0.85	1	0.93
	Claim 2: Writing	0.58		1	0.91
	Claim 3: Listening	0.52	0.50		1
	Claim 4: Research	0.57	0.60	0.48	
8	Claim 1: Reading		0.90	1	0.92
	Claim 2: Writing	0.63		1	0.94
	Claim 3: Listening	0.55	0.52		1
	Claim 4: Research	0.59	0.62	0.49	
11	Claim 1: Reading		0.88	1	0.93
	Claim 2: Writing	0.60		0.98	0.92
	Claim 3: Listening	0.49	0.47		1
	Claim 4: Research	0.58	0.60	0.47	

Table 39. Correlations among Claims: Mathematics

Grade	**Claim**	**Observed & Disattenuated Correlation**		
		Claim 1	**Claims 2 & 4**	**Claim 3**
3	Claim 1		1	1
	Claims 2 & 4	0.74		1
	Claim 3	0.71	0.65	
4	Claim 1		1	0.98
	Claims 2 & 4	0.69		1
	Claim 3	0.73	0.63	
5	Claim 1		1	1
	Claims 2 & 4	0.69		1
	Claim 3	0.69	0.6	
6	Claim 1		1	1
	Claims 2 & 4	0.70		1
	Claim 3	0.70	0.6	
7	Claim 1		1	1
	Claims 2 & 4	0.70		1
	Claim 3	0.66	0.57	
8	Claim 1		1	0.97
	Claims 2 & 4	0.72		1
	Claim 3	0.60	0.54	
11	Claim 1		0.98	0.93
	Claims 2 & 4	0.64		0.97
	Claim 3	0.60	0.5	

Legend:
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 testing, items administered vary among students, so the amount of measurement error differs from one test to another, which yields conditional standard errors of measurement (CSEM).

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

5.1 MARGINAL RELIABILITY

For reliability, the marginal reliability was computed for the scale scores, taking into account the varying measurement errors across the ability range. Marginal reliability is a measure of the overall reliability of an assessment based on the average CSEM, estimated at different points on the ability scale, for all students.

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

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

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 the IRT, SEM is estimated as a function of test information provided by a given set of items that make up the test. In computer-adaptive testing (CAT), items administered vary among all students, so the SEM also can vary among students, which yields CSEM. The average CSEM can be computed as

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

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

Table 40 presents the marginal reliability coefficients and the average CSEM for the total scale scores.

Table 40. Marginal Reliability: 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	12,666	24	0.89	2425.57	104.36	35.06
4	12,193	24	0.88	2472.79	106.86	36.42
5	12,779	24	0.89	2511.62	111.00	36.51
6	12,642	26	0.89	2534.02	106.63	35.44
7	11,960	26	0.88	2553.98	108.80	36.94
8	11,960	26	0.89	2567.80	113.66	37.86
11	11,189	26	0.88	2594.46	119.02	41.58
			Mathematics			
3	12,699	22	0.91	2437.52	95.93	28.03
4	12,238	22	0.92	2485.03	96.62	27.50
5	12,825	22	0.91	2508.56	105.07	31.83
6	12,705	22	0.91	2521.05	118.26	35.55
7	12,035	22	0.89	2524.92	120.51	39.13
8	12,049	22	0.89	2535.60	131.11	43.88
11	11,211	24	0.88	2547.41	126.81	44.26

5.2 STANDARD ERROR CURVES

Figures 11 and 12 present plots of the CSEM of scale scores across the range of ability. 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 score, 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 shown 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 CSEM curves that are flatter. The Smarter Balanced assessments focus on increasing precision where it is most needed, i.e., the 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 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 Error of Measurement: ELA/L

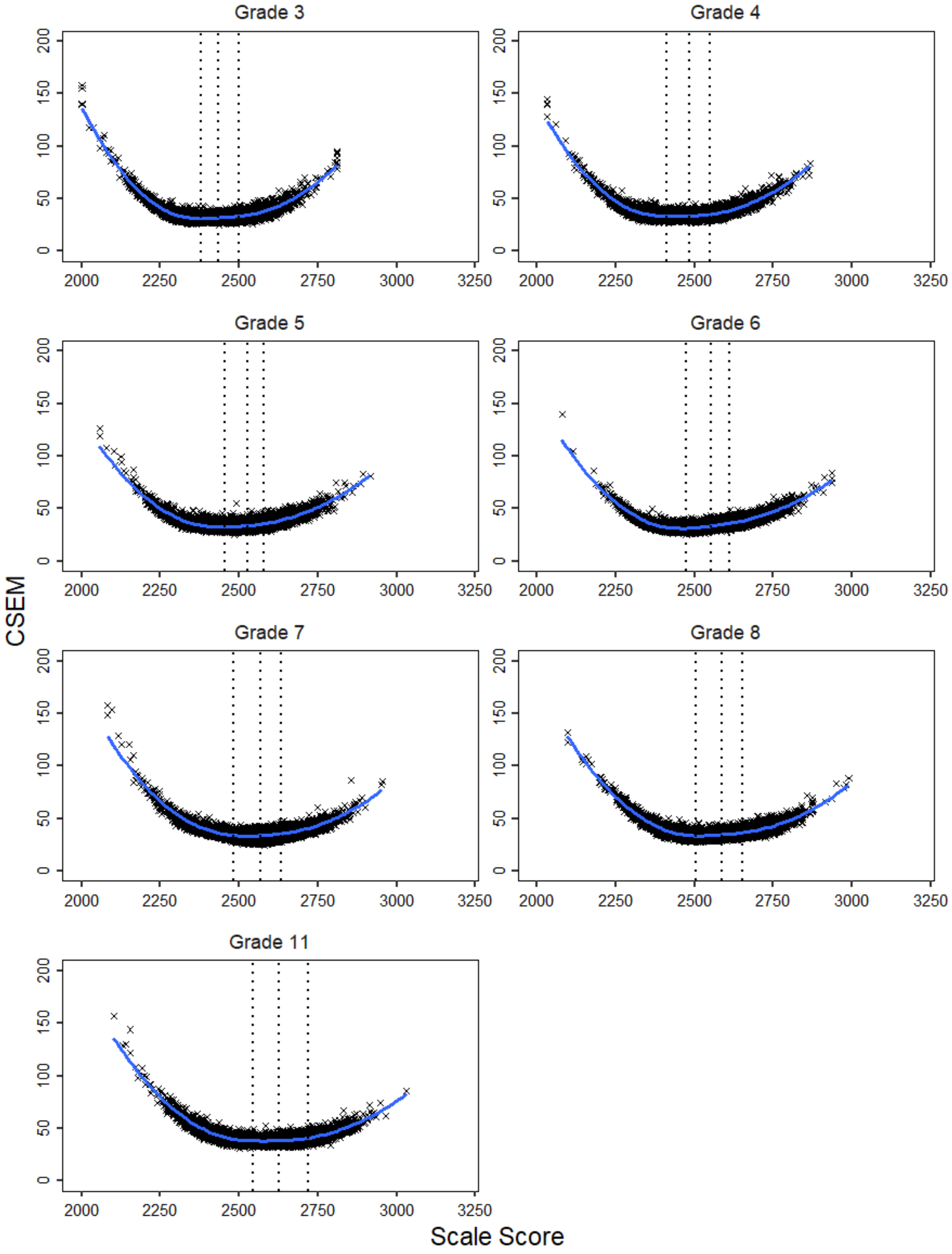
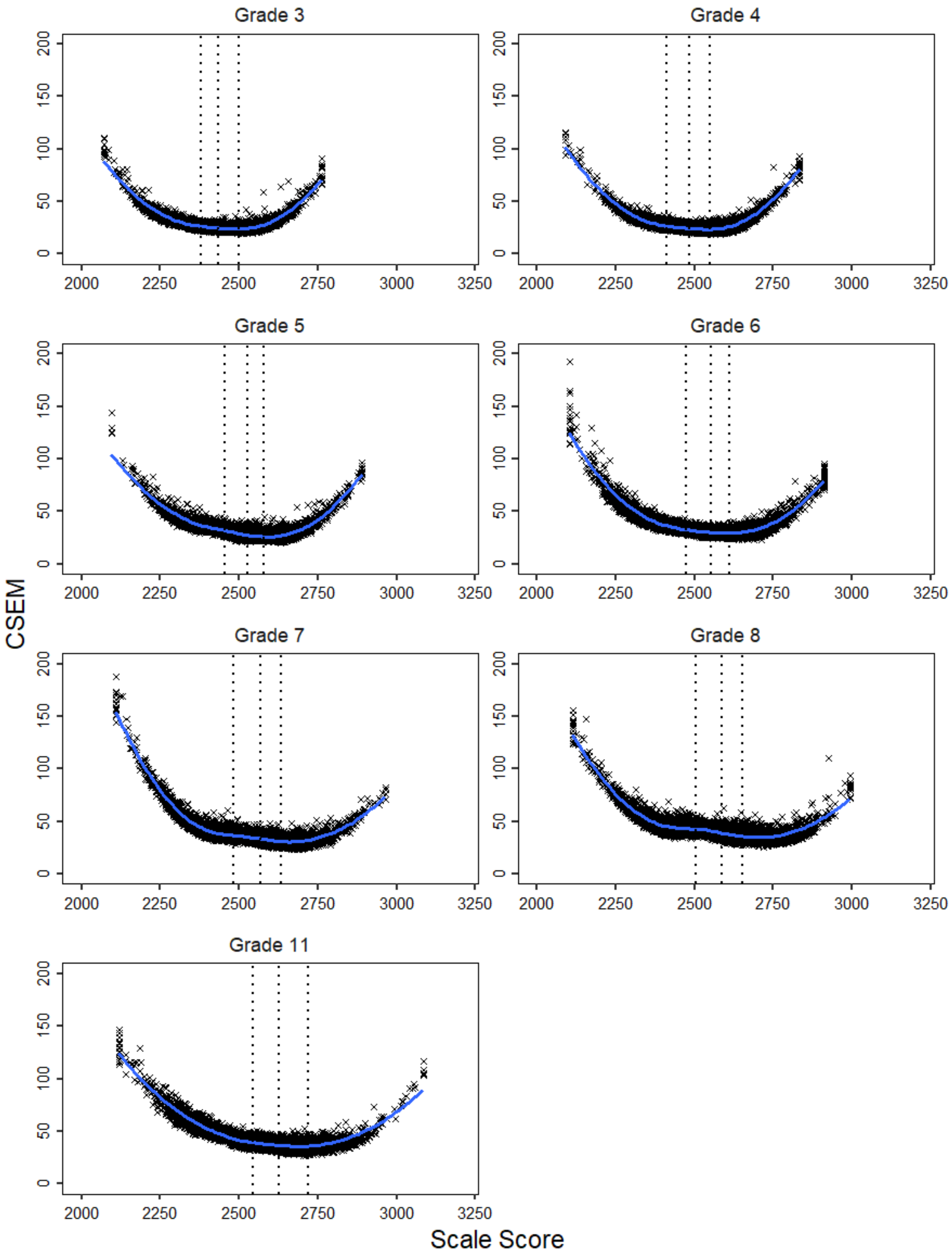


Figure 12. Conditional Standard Error of Measurement: Mathematics



The CSEMs presented in Figures 11 and 12 are summarized in Tables 41 and 42. Table 41 provides the average CSEM for all scale scores and by achievement level. Table 42 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 for most grades in ELA/L and all grades in mathematics. Average CSEMs at all cut scores are larger at Level 4 cut scores in ELA/L but larger at Level 2 cut scores in mathematics.

Table 41. Average Conditional Standard Error of Measurement by Achievement Level

Grade	Level 1	Level 2	Level 3	Level 4	Average CSEM
			ELA/L		
3	37.75	31.64	32.26	36.58	35.06
4	37.91	33.37	33.72	38.60	36.42
5	37.16	32.94	34.21	40.04	36.51
6	35.60	31.57	34.41	40.15	35.44
7	41.16	33.42	34.59	39.41	36.94
8	42.30	33.62	35.38	40.63	37.86
11	47.92	38.40	38.48	42.45	41.58
			Mathematics		
3	33.53	25.26	24.20	27.37	28.03
4	32.92	25.41	23.93	27.94	27.50
5	38.17	29.43	26.65	29.09	31.83
6	42.79	31.19	29.72	33.37	35.55
7	47.40	35.32	32.14	32.23	39.13
8	50.35	41.58	36.95	36.45	43.88
11	50.09	38.24	35.89	39.23	44.26

Table 42. Average Conditional Standard Error of Measurement at Each Achievement-Level Cut and Difference of the SEMs Between Two Cuts

Grade	L2 Cut	L3 Cut	L4 Cut	|L2-L3|	|L3-L4|	|L2-L4|
			ELA/L			
3	31.77	32.19	32.87	0.42	0.68	1.09
4	32.77	33.51	34.34	0.74	0.83	1.57
5	32.50	33.04	35.82	0.54	2.78	3.32
6	31.46	32.63	36.72	1.17	4.09	5.26
7	34.44	33.68	35.19	0.77	1.51	0.75
8	33.97	34.04	36.49	0.07	2.45	2.52
11	39.51	38.90	38.36	0.61	0.54	1.15
			Mathematics			
3	26.13	24.45	23.82	1.68	0.63	2.31
4	26.66	24.07	23.22	2.59	0.86	3.45
5	32.43	27.11	26.81	5.32	0.30	5.62
6	32.42	30.24	29.21	2.18	1.03	3.21
7	36.03	33.12	30.67	2.91	2.46	5.37
8	42.25	39.68	35.21	2.57	4.47	7.04
11	38.94	36.57	34.96	2.37	1.61	3.97

5.3 RELIABILITY OF ACHIEVEMENT CLASSIFICATION

When student performance is reported in terms of achievement levels, the 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 indexes consider the accuracy and consistency of classifications.

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

The classification index can be examined in terms of the classification accuracy and the classification consistency. The term classification accuracy refers to the agreement between classifications that were made based on the form actually taken and 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 based on an alternative 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, item parameters, and assumed underlying latent ability distribution as described in this section. The true score is an expected value of the test score with a measurement error.

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

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

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 that probability subtracted from 1 is the estimate of the chance that the student is correctly classified as being below the cut score. Using this logic, the various classification probabilities can be defined.

The probability of the ith student being classified at achievement level $l (l = 1,2,\cdots,L)$ based on the cut scores cut_{l-1} and cut_l, given the student's item scores $\mathbf{z}_i = (z_{i1},\cdots,z_{iJ})$ and item parameters $\mathbf{b} = (\mathbf{b}_1,\cdots,\mathbf{b}_J)$ and using the J administered items, can be estimated as

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

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

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

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

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

where d stands for dichotomous and p stands for polytomous items; $\mathbf{b}_j = (a_j, b_j, c_j)$ if the jth item is a dichotomous item, and $\mathbf{b}_j = (a_j, b_{j1}, \ldots, b_{jK_i})$ if the jth item is a polytomous item; a_j is the item's discrimination parameter (for Rasch model, a_j=1), c_j is the guessing parameter (for Rasch and 2PL models, c_j=0), 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 number of students at achievement level lm, pl_i is the ith student's achievement level, and p_{im} is the probability of the ith 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 ith student being classified at achievement level l and m, respectively, based on observed scores and hypothetical scores from an 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 the overall scale scores. Table 43 provides the percentages of classification accuracy and consistency for overall, by achievement level, and at proficiency cut score.

The overall classification index ranged from 74% to 80% for accuracy and from 66% to 72% for 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 [-∞, L2 cut; L4 cut, ∞] 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 93% for accuracy and from 87% to 90% for consistency.

The 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 indexes by subgroup are provided in Appendix C, Classification Accuracy and Consistency Index by Subgroup.

Table 43. Classification Accuracy and Consistency

Grade	Achievement Level	ELA/L		Mathematics	
		% Accuracy	**% Consistency**	**% Accuracy**	**% Consistency**
3	Overall	75	67	77	69
	L1	89	83	85	79
	L2	61	50	63	50
	L3	57	46	70	60
	L4	86	79	88	82
	Proficiency Cut	91	88	92	89
4	Overall	74	67	79	71
	L1	89	82	87	79
	L2	55	44	72	63
	L3	57	46	70	59
	L4	85	79	88	82
	Proficiency Cut	91	87	92	89
5	Overall	76	67	78	70
	L1	89	82	88	82
	L2	58	46	68	56
	L3	66	55	59	48
	L4	85	78	88	81
	Proficiency Cut	91	87	92	89
6	Overall	76	67	78	70
	L1	89	82	90	84
	L2	66	55	68	59
	L3	69	60	60	48
	L4	84	74	87	80
	Proficiency Cut	91	87	92	88
7	Overall	76	67	78	70
	L1	88	80	89	84
	L2	64	52	66	55
	L3	72	63	62	52
	L4	82	72	87	79
	Proficiency Cut	91	87	91	87
8	Overall	77	68	76	68
	L1	88	82	88	83
	L2	66	54	60	49
	L3	72	64	57	45
	L4	82	72	87	79
	Proficiency Cut	91	88	91	88
11	Overall	75	66	80	72
	L1	86	79	91	86
	L2	66	54	64	53
	L3	69	59	68	56
	L4	83	75	85	76
	Proficiency Cut	91	87	93	90

5.4 Reliability for Subgroups

The reliability of test scores is also computed by subgroup. Tables 44–51 present the marginal reliability coefficients by gender, ethnicity groups, ELLs, disadvantaged (free or reduced lunch), migrant, and students with disabilities. The reliability coefficients are similar across subgroups but somewhat lower for the ELL and students with disabilities subgroups. A large percentage of students in these subgroups received Level 1 with large CSEMs.

Table 44. Marginal Reliability Coefficients for Overall and by Subgroup: ELA/L (Grades 3–4)

Subgroup	**Grade 3**					**Grade 4**				
	N	**MR**	**SS**	**SD**	**CSEM**	**N**	**MR**	**SS**	**SD**	**CSEM**
All Students	12,666	0.89	2425.57	104.36	35.06	12,193	0.88	2472.79	106.86	36.42
Female	6,040	0.88	2437.98	101.18	34.82	5,934	0.88	2482.47	104.28	36.19
Male	6,626	0.89	2414.26	105.91	35.27	6,259	0.89	2463.61	108.47	36.63
African American	173	0.88	2414.30	104.11	36.63	140	0.85	2473.58	92.90	35.81
AmerIndian/Alaskan	14	0.90	2435.53	117.33	36.37	19	0.80	2459.13	76.02	34.24
Asian/Pacific Islander	2,608	0.88	2458.71	99.95	35.13	2,759	0.88	2503.17	105.50	36.89
Hispanic	2,588	0.88	2409.98	99.93	34.96	2,321	0.88	2459.94	102.37	36.18
Hawaiʻi Pacific Islander	3,033	0.86	2379.77	95.70	35.40	2,791	0.87	2421.89	99.07	36.17
White	1,513	0.87	2457.63	97.24	34.56	1,408	0.86	2500.90	96.50	36.13
Multi-Racial	2,737	0.89	2442.44	104.24	34.85	2,755	0.88	2490.45	104.88	36.59
ELL	1,494	0.85	2367.22	91.72	35.75	1,361	0.84	2397.26	93.40	37.37
Disadvantaged	5,844	0.87	2395.61	98.39	35.12	5,452	0.88	2439.79	102.71	36.25
Migrant	141	0.85	2378.22	88.86	34.49	177	0.87	2422.05	100.99	35.97
Disability	1,427	0.79	2318.52	85.64	39.16	1,327	0.82	2361.65	91.00	38.97

Legend. MR: Marginal Reliability; SS: Scale Score Mean; SD: Standard Deviation of Scale Score; CSEM: Mean of Conditional Standard Error of Measurement

Table 45. Marginal Reliability Coefficients for Overall and by Subgroup: ELA/L (Grades 5–6)

Subgroup	**Grade 5**					**Grade 6**				
	N	**MR**	**SS**	**SD**	**CSEM**	**N**	**MR**	**SS**	**SD**	**CSEM**
All Students	12,779	0.89	2511.62	111.00	36.51	12,642	0.89	2534.02	106.63	35.44
Female	6,170	0.88	2524.05	106.41	36.52	6,059	0.88	2549.29	102.20	35.44
Male	6,609	0.90	2500.02	113.90	36.49	6,582	0.89	2519.94	108.68	35.44
African American	161	0.85	2516.18	91.90	35.12	138	0.86	2543.07	93.67	35.58
AmerIndian/Alaskan	5*					21	0.85	2531.70	92.37	35.65
Asian/Pacific Islander	2,939	0.88	2543.85	104.89	36.98	2,969	0.87	2566.31	101.46	36.25
Hispanic	2,427	0.89	2494.41	106.21	35.85	2,525	0.88	2520.00	102.35	34.90
Hawaiʻi Pacific Islander	3,020	0.89	2456.77	107.03	36.16	2,960	0.88	2480.39	98.06	34.40
White	1,444	0.86	2549.93	99.05	37.30	1,360	0.87	2569.71	100.05	36.35
Multi-Racial	2,783	0.88	2531.92	106.40	36.60	2,669	0.88	2552.21	103.51	35.69
ELL	1,098	0.84	2404.41	93.56	37.37	1,088	0.82	2427.04	81.66	34.85
Disadvantaged	5,717	0.89	2475.14	107.43	35.94	5,606	0.88	2499.09	102.42	34.74
Migrant	163	0.87	2454.32	99.40	35.39	184	0.88	2474.03	98.55	34.62
Disability	1,393	0.84	2385.22	95.01	38.36	1,432	0.83	2416.17	88.34	36.52

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

Table 46. Marginal Reliability Coefficients for Overall and by Subgroup: ELA/L (Grades 7–8)

Subgroup	Grade 7					Grade 8				
	N	MR	SS	SD	CSEM	N	MR	SS	SD	CSEM
All Students	11,960	0.88	2553.98	108.80	36.94	11,960	0.89	2567.80	113.66	37.86
Female	5,761	0.87	2570.76	102.92	36.67	5,769	0.88	2586.01	106.92	37.21
Male	6,199	0.89	2538.39	111.77	37.20	6,190	0.89	2550.85	117.08	38.46
African American	143	0.86	2562.02	95.20	36.09	143	0.88	2586.72	104.22	36.66
AmerIndian/Alaskan	17	0.87	2468.53	100.79	37.00	25	0.89	2575.79	112.01	36.71
Asian/Pacific Islander	2,931	0.87	2590.20	102.09	36.99	3,158	0.88	2606.00	107.74	37.85
Hispanic	2,327	0.88	2537.42	106.07	36.91	2,286	0.88	2547.66	107.53	37.21
Hawaiʻi Pacific Islander	2,864	0.87	2502.00	103.21	37.36	2,761	0.87	2510.26	106.22	38.90
White	1,173	0.86	2592.16	100.11	36.84	1,283	0.87	2609.75	102.84	37.59
Multi-Racial	2,504	0.87	2568.68	103.01	36.53	2,304	0.88	2579.77	109.38	37.48
ELL	1,238	0.84	2466.44	96.00	38.70	1,166	0.81	2470.41	94.65	41.32
Disadvantaged	5,262	0.88	2522.60	107.38	37.23	4,980	0.88	2530.27	110.00	38.36
Migrant	178	0.88	2501.35	108.20	37.87	155	0.87	2514.00	106.65	37.90
Disability	1,284	0.81	2428.84	94.26	41.35	1,286	0.79	2439.94	91.76	41.94

Table 47. Marginal Reliability Coefficients for Overall and by Subgroup: ELA/L (Grade 11)

Subgroup	Grade 11				
	N	MR	SS	SD	CSEM
All Students	11,189	0.88	2594.46	119.02	41.58
Female	5,422	0.86	2613.81	109.59	40.79
Male	5,767	0.88	2576.28	124.54	42.31
African American	139	0.84	2611.13	102.02	40.39
AmerIndian/Alaskan	12	0.92	2601.71	137.05	39.03
Asian/Pacific Islander	3,426	0.86	2629.90	108.60	40.80
Hispanic	1,960	0.88	2575.44	119.12	41.61
Hawaiʻi Pacific Islander	2,423	0.86	2540.27	112.35	42.68
White	1,144	0.88	2620.64	118.51	41.43
Multi-Racial	2,085	0.87	2601.59	117.88	41.70
ELL	767	0.76	2489.71	87.18	42.62
Disadvantaged	4,121	0.87	2561.13	117.31	42.19
Migrant	147	0.86	2551.31	109.83	41.24
Disability	927	0.76	2460.97	98.50	48.20

Table 48. Marginal Reliability Coefficients for Overall and by Subgroup: Mathematics (Grades 3–4)

Subgroup	Grade 3					Grade 4				
	N	MR	SS	SD	CSEM	N	MR	SS	SD	CSEM
All Students	12,699	0.91	2437.52	95.93	28.03	12,238	0.92	2485.03	96.62	27.50
Female	6,048	0.91	2436.33	91.16	27.64	5,963	0.91	2480.89	90.69	27.01
Male	6,651	0.92	2438.61	100.07	28.38	6,275	0.92	2488.96	101.78	27.96
African American	174	0.89	2416.35	79.57	26.58	142	0.89	2475.32	80.22	26.68
AmerIndian/Alaskan	14	0.89	2427.99	81.53	26.92	19	0.85	2454.63	66.43	26.06
Asian/Pacific Islander	2,626	0.91	2473.42	90.35	27.55	2,782	0.92	2518.49	94.50	27.29
Hispanic	2,593	0.90	2419.96	90.04	28.35	2,327	0.91	2469.64	91.89	27.64
Hawaiʻi Pacific Islander	3,036	0.89	2392.65	90.38	29.67	2,801	0.90	2437.33	90.05	28.52
White	1,513	0.91	2465.96	88.64	26.77	1,411	0.91	2509.69	88.35	26.46
Multi-Racial	2,743	0.92	2455.12	93.78	27.08	2,756	0.91	2500.81	92.26	27.10
ELL	1,523	0.89	2387.76	91.40	29.86	1,379	0.89	2422.08	91.97	29.90
Disadvantaged	5,877	0.90	2409.80	91.21	28.67	5,482	0.91	2455.77	92.68	27.99
Migrant	141	0.89	2399.36	84.41	27.43	175	0.91	2445.63	94.56	27.72
Disability	1,435	0.85	2338.03	92.88	35.72	1,334	0.88	2388.43	92.15	32.51

Table 49. Marginal Reliability Coefficients for Overall and by Subgroup: Mathematics (Grades 5–6)

Subgroup	Grade 5					Grade 6				
	N	MR	SS	SD	CSEM	N	MR	SS	SD	CSEM
All Students	12,825	0.91	2508.56	105.07	31.83	12,705	0.91	2521.05	118.26	35.55
Female	6,186	0.90	2506.27	98.70	31.36	6,093	0.90	2521.76	112.20	34.73
Male	6,639	0.91	2510.69	110.65	32.27	6,611	0.91	2520.37	123.57	36.30
African American	161	0.86	2502.40	80.38	30.15	137	0.89	2521.12	104.19	34.19
AmerIndian/Alaskan	5*					21	0.89	2502.77	102.46	33.49
Asian/Pacific Islander	2,971	0.91	2548.67	103.18	30.86	2,991	0.91	2565.56	112.71	34.22
Hispanic	2,435	0.89	2489.29	98.23	32.21	2,546	0.90	2499.98	113.03	36.22
Hawaiʻi Pacific Islander	3,027	0.88	2456.01	98.74	34.15	2,975	0.87	2460.54	108.04	38.21
White	1,445	0.90	2536.98	94.43	30.38	1,362	0.90	2556.68	107.89	33.91
Multi-Racial	2,781	0.90	2525.35	98.67	30.71	2,673	0.91	2540.65	113.14	34.20
ELL	1,119	0.84	2415.81	91.35	36.79	1,138	0.83	2416.51	102.75	42.76
Disadvantaged	5,749	0.89	2474.49	100.01	33.14	5,636	0.89	2483.29	114.93	37.31
Migrant	162	0.88	2459.68	98.26	33.73	185	0.86	2452.60	100.65	37.06
Disability	1,407	0.82	2399.68	93.20	39.20	1,435	0.82	2394.34	106.11	45.20

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

Table 50. Marginal Reliability Coefficients for Overall and by Subgroup: Mathematics (Grades 7–8)

Subgroup	**Grade 7**					**Grade 8**				
	N	**MR**	**SS**	**SD**	**CSEM**	**N**	**MR**	**SS**	**SD**	**CSEM**
All Students	12,035	0.89	2524.92	120.51	39.13	12,049	0.89	2535.60	131.11	43.88
Female	5,802	0.89	2523.44	117.15	38.73	5,807	0.88	2537.54	124.69	42.87
Male	6,233	0.90	2526.29	123.55	39.50	6,241	0.89	2533.79	136.80	44.79
African American	143	0.87	2534.36	96.20	35.06	143	0.86	2533.26	114.31	42.83
AmerIndian/Alaskan	17	0.83	2460.17	98.43	40.03	26	0.84	2527.38	109.98	43.87
Asian/Pacific Islander	2,958	0.90	2571.40	119.41	36.84	3,186	0.90	2589.36	132.39	41.36
Hispanic	2,345	0.87	2502.51	112.88	40.12	2,297	0.86	2507.11	117.58	44.62
Hawaiʻi Pacific Islander	2,890	0.84	2461.69	108.55	43.92	2,799	0.82	2468.16	113.41	47.88
White	1,174	0.90	2567.72	110.16	35.48	1,287	0.88	2576.54	119.35	40.96
Multi-Racial	2,507	0.89	2543.76	110.76	36.73	2,311	0.88	2548.90	126.03	43.07
ELL	1,280	0.80	2430.44	107.06	47.79	1,192	0.80	2438.75	111.77	50.49
Disadvantaged	5,291	0.87	2489.25	116.28	41.88	5,033	0.86	2493.64	122.00	46.18
Migrant	177	0.84	2483.84	100.30	39.55	155	0.84	2470.01	116.33	47.21
Disability	1,290	0.73	2401.13	99.10	51.31	1,302	0.71	2398.91	99.43	53.88

Table 51. Marginal Reliability Coefficients for Overall and by Subgroup: Mathematics (Grade 11)

Subgroup	**Grade 11**				
	N	**MR**	**SS**	**SD**	**CSEM**
All Students	11,211	0.88	2547.41	126.81	44.26
Female	5,443	0.87	2549.47	119.72	43.38
Male	5,768	0.89	2545.46	133.14	45.08
African American	138	0.86	2552.31	109.84	41.69
AmerIndian/Alaskan	11	0.85	2507.77	119.02	45.39
Asian/Pacific Islander	3,451	0.89	2592.16	125.27	41.47
Hispanic	1,955	0.85	2520.82	114.87	45.20
Hawaiʻi Pacific Islander	2,434	0.81	2489.47	109.69	48.13
White	1,138	0.89	2572.81	129.80	43.31
Multi-Racial	2,084	0.88	2551.93	125.61	43.82
ELL	791	0.78	2458.24	106.12	50.15
Disadvantaged	4,136	0.85	2511.77	118.58	46.51
Migrant	143	0.78	2485.37	101.65	47.39
Disability	930	0.62	2421.72	92.39	56.63

5.5 RELIABILITY FOR CLAIM SCORES

The marginal reliability, average and standard deviation of scale scores, and average of CSEM are also computed for claim scores by test and grade. In mathematics, Claims 2 and 4 are combined to have enough items to generate a score. Given the reduction in the small number of items in the Hawaiʻi shortened blueprint, the reliabilities for claim scores are low, especially for Claim 3 and Claim 4 in ELA/L and Claims 2 and 4 combined and Claim 3 in mathematics. In 2024–2025, the performance category for claim scores was reported at the individual student level for only Claims 1 and 2 in ELA/L and Claim 1 in mathematics.

Tables 52 and 53 present the marginal reliability coefficients and descriptive statistics by claim in ELA/L and mathematics, respectively.

Table 52. 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	8	0.61	2428.05	124.37	77.26
	Claim 2: Writing	6	0.73	2421.59	129.65	67.06
	Claim 3: Listening	4	0.26	2423.14	153.61	132.06
	Claim 4: Research	6	0.59	2425.72	138.75	88.61
4	Claim 1: Reading	8	0.61	2475.11	129.97	81.25
	Claim 2: Writing	6	0.72	2469.00	136.36	72.19
	Claim 3: Listening	4	0.34	2467.50	149.42	121.68
	Claim 4: Research	6	0.59	2477.99	141.88	91.23
5	Claim 1: Reading	8	0.62	2512.17	133.35	81.77
	Claim 2: Writing	6	0.73	2512.31	136.64	70.47
	Claim 3: Listening	4	0.36	2506.89	162.40	129.97
	Claim 4: Research	6	0.62	2513.59	145.26	89.34
6	Claim 1: Reading	10	0.70	2525.39	128.10	70.48
	Claim 2: Writing	6	0.72	2536.02	128.61	67.74
	Claim 3: Listening	4	0.34	2542.83	165.30	133.80
	Claim 4: Research	6	0.59	2545.35	147.08	94.42
7	Claim 1: Reading	10	0.64	2543.08	133.03	79.88
	Claim 2: Writing	6	0.73	2559.42	137.53	71.29
	Claim 3: Listening	4	0.32	2548.01	156.40	129.03
	Claim 4: Research	6	0.59	2557.48	154.64	98.42
8	Claim 1: Reading	10	0.68	2556.93	133.47	75.29
	Claim 2: Writing	6	0.72	2567.75	141.06	75.11
	Claim 3: Listening	4	0.35	2571.53	164.92	133.32
	Claim 4: Research	6	0.60	2583.88	154.24	97.01
11	Claim 1: Reading	10	0.65	2582.01	147.48	86.84
	Claim 2: Writing	6	0.71	2600.89	143.94	77.50
	Claim 3: Listening	4	0.33	2589.14	180.46	148.05
	Claim 4: Research	6	0.60	2598.14	167.90	106.27

Table 53. 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	12	0.85	2441.32	104.63	40.52
	Claims 2 & 4	5	0.60	2435.20	110.44	70.19
	Claim 3	5	0.59	2430.30	117.05	75.05
4	Claim 1	12	0.86	2488.88	106.64	39.88
	Claims 2 & 4	5	0.55	2474.66	115.96	78.02
	Claim 3	5	0.64	2480.81	118.12	71.19
5	Claim 1	12	0.84	2514.85	114.63	45.63
	Claims 2 & 4	5	0.49	2497.58	124.58	88.98
	Claim 3	5	0.57	2499.97	139.54	91.54
6	Claim 1	12	0.85	2523.92	129.34	49.52
	Claims 2 & 4	5	0.52	2512.11	142.20	98.17
	Claim 3	5	0.54	2516.73	145.39	99.10
7	Claim 1	12	0.82	2524.93	135.36	57.45
	Claims 2 & 4	5	0.51	2519.93	138.08	96.72
	Claim 3	5	0.52	2518.49	156.30	107.89
8	Claim 1	12	0.82	2535.78	142.10	61.05
	Claims 2 & 4	5	0.56	2529.69	159.46	105.68
	Claim 3	5	0.47	2528.81	172.28	124.94
11	Claim 1	14	0.82	2546.45	134.94	57.72
	Claims 2 & 4	5	0.52	2543.89	180.79	125.53
	Claim 3	5	0.52	2533.32	177.72	123.61

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

6 SCORING

The Smarter Balanced Assessment Consortium (SBAC) 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 Claims 1 and 2 in English language arts/literacy (ELA/L) and Claim 1 in mathematics. This section describes the rules used to generate the scores and the handscoring procedure. The rules and procedures for generating scores are the same in all operational administration years.

6.1 ESTIMATING STUDENT ABILITY USING MAXIMUM LIKELIHOOD ESTIMATION

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

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

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

where $\mathbf{b}_i = \left(b_{i,1}, \dots, b_{i,m_i}\right)$ for the ith 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 person j, and k indexes the step of item i.

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

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

in the case of items with two or more points,

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

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

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

where m_i is the maximum possible score point (starting from 0) for the ith item, and D is the scale factor, 1.7. The SE is calculated based on the answered item(s) only for both complete and incomplete tests. The upper bound of the SE is set to 2.5 on the θ metric. Any value larger than 2.5 is truncated at 2.5 on the θ 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 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. Although the update of the ability estimates is performed at each iteration, the overall 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 SBAC. Table 54 presents the scaling constants for each subject for the theta-to-scale score linear transformation. Scale scores are rounded to an integer.

Table 54. Vertical Scaling Constants on the Reporting Metric

Subject	Grade	Slope (a)	Intercept (b)
ELA/L	3–8, 11	85.8	2508.2
Mathematics	3–8, 11	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, SE_{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 55 provides three achievement standards for each grade and content area.

Table 55. Cut Scores in Scale Scores

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

6.3 LOWEST/HIGHEST OBTAINABLE SCORES

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 enough easy or difficult items to measure low- and high-performing students, the standard error could be large in the low and high ends of the ability range. SBAC decided to truncate extreme, unreliable student ability estimates. Table 56 presents the lowest obtainable theta (LOT) and scale score (LOSS) and the highest obtainable theta (HOT) and scale score (HOSS) in both theta and scale score metrics. 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 total scores. The standard error for the LOT and HOT is computed using the LOT and HOT ability estimates given the administered items.

Table 56. Lowest and Highest Obtainable Scores

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

6.4 SCORING ALL CORRECT AND ALL INCORRECT CASES

In the item response theory (IRT) maximum likelihood 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) and the lowest obtainable scores (LOT and LOSS) were assigned in the 2014–2015 administration. Since the 2015–2016 administration, all incorrect and correct cases were scored by either adding 0.5 to or subtracting 0.5 from an item score with the smallest item discrimination parameter among the administered operational items (computer-adaptive testing [CAT] and performance tasks [PTs]) for a student.

6.5 Rules for Calculating Strengths and Weaknesses for Claim Scores

In ELA/L, claim scores are computed and reported for Claims 1 and 2 at the individual student level; in mathematics, claim scores are computed and reported for Claim 1 only. For the claim, three performance categories, indicating relative strength and weakness, 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.

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 (i.e., benchmark) 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. An adaptive test, however, offers a tremendous opportunity for target-level data at the class, school, and complex-area level. 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 claim (four claims) for ELA/L and in Claim 1 only for mathematics. Target scores can be computed for any aggregate group of students, and Chapter 7: Reporting and Interpreting Scores provides details on which aggregate groups of students have target scores computed and who has access to the reports.

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 $p_{ij} = p(z_{ij} = 1)$, indicating the probability that student j responds correctly to item i, z_{ij} represents the jth student's score on the ith 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 generalized partial credit model (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

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

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

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

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

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

For an aggregate unit, a target score is computed by averaging the individual student target scores for the 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 difference from zero in these aggregates may indicate that a roster, teacher, school, complex, or complex area 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 a target within an aggregate group, a minimum amount of precision is required to report target performance for the group. There are no requirements for a minimum number of items or students.

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

- If $\bar{\delta}_{Tg} \geq +1 * se(\bar{\delta}_{Tg})$, then performance is *better* than on the overall test.
- If $\bar{\delta}_{Tg} \leq -1 * se(\bar{\delta}_{Tg})$, then performance is *worse* than on the overall test.
- Otherwise, performance is *similar to* performance on the test as a whole.
- If $se(\bar{\delta}_{Tg}) > 0.2$, data are insufficient.

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

By defining $p_{ij} = p(z_{ij} = 1)$, indicating the probability that student j responds correctly to item i, z_{ij} represents the jth student's score on the ith 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:

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

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

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

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

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

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

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

For an aggregate unit, a target score is computed by averaging the individual student target scores for the 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 difference from zero in these aggregates may indicate that a class, teacher, school, complex, or complex area 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/weaknesses, the following are reported:

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

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

6.7 HANDSCORING

Constructed response short-answer (SA) items and essay (i.e., full write) items in English language arts/literacy (ELA/L) and SA items in 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 and automated scoring using the Project Essay Grade (PEG) engine. Some Smarter Balanced member states have elected to use handscoring exclusively, while others have elected to use a hybrid automated scoring/handscoring approach. Hawai‘i has elected to use a hybrid automated scoring/handscoring approach. The methods and results for handscoring and hybrid automated scoring are described in the following sections.

For 2024–2025 summative tests, there were a total of 514 ELA/L SA items, 193 ELA/L essay items, and 347 mathematics SA items administered from the 2025 Smarter Balanced summative item pool. Table 57 shows the number of handscored items administered from the Smarter Balanced summative operational item pool, by grade and subject.

Table 57. Administered Handscored Items in Smarter Balanced Summative Item Pool, by Grade and Subject

Grade	ELA/L		Mathematics
	Short-Answer	Essay	
3	44	27	50
4	50	28	53
5	48	29	86
6	79	21	51
7	74	29	28
8	92	30	35
11	127	29	44
Total	514	193	347

All guidelines for handscoring responses were specified by Smarter Balanced. Outlined below is the handscoring process MI followed in spring 2025 in accordance with the Smarter Balanced guidelines. This process applied to the scoring of all student’s constructed responses for ELA/L SA and essay items and mathematics SA items. This section describes rater selection, rater training, qualification and scoring, rater monitoring, evaluation, feedback, and rater agreement for handscoring.

6.7.1 Rater Selection

Rater pool and supplement

MI has developed a pool of approximately five thousand raters experienced in scoring the Smarter Balanced assessments. MI first recruited qualified raters who had experience scoring these assessments. Rater accuracy data, collected during prior administration scoring, was used to prioritize recruitment of the most accurate, experienced 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, 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 handscoring requirements.

Rater and team leader requirements

At minimum, all raters were required to possess 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 to monitor the raters, MI scoring leadership reviewed records of all returning staff. They looked for people who were experienced team leaders with a record of good performance on previous projects, and they also considered raters who had been recommended for promotion to the team leader position or otherwise displayed exemplary performance.

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, Qualification, and Scoring

Rater groups

Once hired, raters were assigned to a scoring group corresponding to the subject/grade that they were deemed best suited to score. Raters were trained to score a specific item group of either SA (research, brief write, reading, and mathematics) or essay (i.e., full-write) 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 were assigned a unique ID used to track their scoring work throughout the scoring effort. The number of items an individual rater scored was minimized to allow the rater to more quickly develop experience scoring responses to a small number of items.

Training modules and materials

When beginning 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 hosts 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

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. Many of these sets were created during the original field-test scoring in 2014 and were approved by Smarter Balanced. Additional sets were created as new items were field-tested. The same anchor sets are used each year.

Additionally, MI conducts an annual review of the rater agreement and scoring materials 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, as indicated by suboptimal rater accuracy (based on validity responses) and/or rater agreement. Supplemental materials may address item- or response-specific concerns. 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 uncommon 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—such as full write conventions—or to help them more accurately identify responses that should be flagged as non-scorable.

The VSC score resource library

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 associated supplemental materials.

Training and practice

All raters, regardless of experience, were required to train on all anchor and training sets. Following training and practice, all raters were required to pass a qualification to prove that they understood and could apply the criteria accurately. The scoring director and team leaders had access to all practice and qualification results, which were reviewed to identify frequently mis-scored responses and inform initial monitoring and feedback needs.

Until a rater had trained and qualified successfully, the rater was not permitted to score operational student responses. Training was structured 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.

Training time

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, while training for essay items took up to five days to complete. Raters generally worked 3–7 hours per day. The hours worked per day were flexible, based on the raters' shift preference and item(s) being scored. At a minimum, most raters scored 20 hours per week (day shift) or 15 hours per week (evening shift), with many scoring over 30 hours per week (day shift) or 20 hours per week (evening shift).

Qualification

Training and qualification design varied slightly depending on Smarter Balanced item type:

- ELA/L full write: Raters trained and qualified on a baseline training lesson for a grade and writing purpose (e.g., grade 3 narrative, grade 6 argumentative, etc.). After qualifying on the baseline, raters then completed qualifying sets for each item associated with that grade and purpose. Raters could only score those items for which they passed the qualifying set.
- 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 permitted 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 permitted the rater to score that item and all items associated with it; for items with no associated items, training was for the specific item.

An additional validation stage was implemented to supplement the training and qualification process for full write, brief write, and research raters. After completing these initial steps, all prospective raters were required to score a set of validity responses. As in the qualification stage, raters were required to meet established accuracy standards during this validation in order to be approved to score operational responses for a given item.

Raters who failed to meet accuracy standards on the validity responses received continued retraining and were given additional opportunities to improve. Those who were unable to meet the required standards—despite having passed the qualification stage—were disqualified from scoring that item.

Scoring

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 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.

In addition to item-specific scoring expectations, a variety of substantive procedural and policy information was provided to each trainee during training. These included instructions for how to identify and flag certain types of responses as well as how to communicate with leadership during hand scoring.

Flagging nonscorable responses

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, such as essays, 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.

The training process, including this additional information, ensured that raters were fully prepared to handscore responses and understood all responsibilities and scoring requirements before they began operational scoring.

Minimizing rater bias

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 provided any other identifying information as part of their response, the raters had no knowledge of student characteristics. Second, all raters were trained using materials provided by Smarter Balanced, 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 continue to exhibit bias after receiving a reasonable amount of feedback, they were dismissed.

Score accuracy

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 Monitoring, Feedback, and Evaluation

During scoring, rater monitoring using validity responses and second read is performed, and rater performance metrics are generated and evaluated. Additionally, automated feedback based on recent rater performance is provided.

Rater monitoring

During operational scoring, five percent of the responses scored comprised pre-approved validity responses. Validity responses serve as benchmark responses as the most appropriate score for each validity response is predetermined by key stakeholders. 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 validity pool includes anchor validity responses originating from the field test administration.[1] The pool of validity responses is selected to be generally representative of operational responses, while ensuring sufficient examples of each score point. Validity results compare the score assigned by a rater to a validity response with the benchmark score of the same response. Validity responses provide a more direct measurement of rating quality than measures of inter-rater reliability (Raczynski et al., 2015).

Scoring accuracy

Scoring accuracy during handscoring was maintained by continuously assessing rater performance using validity responses. MI specifically evaluated how closely raters' scores aligned with the benchmark scores of these validity responses. Key performance measures included the agreement between rater and benchmark scores, quantified using Quadratic Weighted Kappa (QWK)[2], and the comparison of standardized mean differences (SMD) between the distributions of benchmark and rater-assigned scores.

Rater accuracy calibration and second read procedures

MI calibrates validity responses 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 raters are then fitted to a Generalized Partial Credit Model (GPCM). Utilizing the resulting IRT parameters, MI calculates accuracy values for each rater based on a given set of validity responses. This calculation is conducted several times each day during scoring, providing real-time measures of rater accuracy.

In addition to validity responses, 15% of handscored responses received blind second reads, the results of which were used to calculate inter-rater reliability. To support interpretability, second reads were conducted exclusively by expert (i.e., highly accurate) raters, described below.

The VSC system automatically and randomly routed the requisite number of responses to raters for second reads and validity in an inconspicuous manner. In this way 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.

Rater performance evaluation

The system automatically generated performance metrics several times a day based on the most recent data, providing raters and scoring managers with daily, automated summaries of rater performance. This ensured that all handscoring staff were kept informed of their current performance and any issues that needed attention. In addition to these daily summaries, detailed manager-level reports were produced to identify raters who required retraining or, if necessary, removal due to accuracy or productivity concerns. These

[1] Responses and results of the 2014–15 Smarter Balanced field test administration were used to derive the base scale to which subsequent item parameters are aligned.

[2] QWK is a measure used to assess the agreement between two raters, accounting for the possibility of agreement occurring by chance and giving more weight to larger discrepancies between ratings.

reports enabled scoring management to direct scoring leaders to specific VSC reports, allowing them to pinpoint the areas where individual raters needed improvement.

The monitoring system afforded the objective, dynamic identification of the most accurate raters, referred to as “expert raters.” Specifically, expert raters are those who demonstrate highly accurate and consistent scoring of validity responses. 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.

Automated feedback

During scoring, raters received automated feedback based on recent performance. The automated feedback system identifies raters who require additional feedback—based on accuracy metrics—and automatically notifies them to review a set of responses that reflect their observed scoring challenge(s). 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.

VSC provided real-time reports throughout the scoring effort. These reports were available for access by handscoring management and clients. 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 help reinforce the lines for the various score-points and re-anchor the raters to the proper position, arresting groupwide drift.

Reports using item-level accuracy expectations identified any items not meeting the expected levels of agreement. Specifically, these reports indicated the difference between expected accuracy and current accuracy for each item. In this way, reports informed improvements to the scoring accuracy of all items.

Automated removal of raters and score resets were performed when item and rater performance failed to meet accuracy expectations. In these cases, all responses scored by a rater during a period of poor performance were reset and redistributed to other qualified raters for rescoring. By limiting raters to scoring relatively fewer items, this approach also maximized accuracy across items.

In addition to automated feedback, scoring leadership provided individualized feedback to raters based on their performance. Specifically, leadership reviewed the rater’s performance on validity responses to look for a trend that suggested the rater had drifted from the anchored responses. If such a trend was present, leadership tailored feedback specific to that rater, typically by presenting them with live responses they had mis-scored in a way that was reflective of their overall drift from the anchor set criteria and by providing targeted, thoughtful rationales for the “correct” scores.

Finally, as a supplement to automated assessments, 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 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, or foreign-language responses) 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 combined data across all states and territories that participated in the 2024–2025 summative assessment.

In ELA/L essay (i.e., full writes) item responses were scored in three dimensions: conventions (0–2 rubric), evidence/elaboration (1–4 rubric), and organization/purpose (1–4 rubric). 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.

Table 58 through Table 60 provide a summary of the human-human IRR based on items with a sample size greater than or equal to 50. For Mathematics and ELA/L essay items, the tables show the majority of the items administered. For ELA/L SA items, relatively fewer items reached a sample size greater than or equal to 50, and thus a subset of the items administered are represented in the tables. The IRR is presented with the mean percent exact agreement, minimum and maximum percent exact agreements, combined percent exact and adjacent agreement, and the mean, minimum and maximum QWK. Additionally, the Tables present the average number of responses, as well as minimum and maximum number of responses to a given item

Table 58. 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	30	310.5	56	800	77.7	57.3	93.1	100	0.68	0.43	0.88
4	36	257.4	52	675	75.0	47.3	86.3	100	0.69	0.43	0.86
5	37	307.7	56	925	72.0	46.6	86.4	100	0.69	0.36	0.91
6	65	278.2	50	1247	73.3	47.8	90.9	100	0.66	0.35	0.92
7	70	306.4	51	1565	72.8	56.9	85.3	100	0.68	0.46	0.83
8	78	296.3	58	1089	72.4	58.6	84.3	100	0.70	0.49	0.83
11	67	209.6	50	1100	74.2	56.8	93.8	100	0.71	0.36	0.90

Table 59. Inter-Rater Agreement for ELA/L Essay Items

Grade	Trait	Number of Items	Number of Responses			% Exact			% (Exact + Adjacent)	QWK		
			Mean	Min	Max	Mean	Min	Max		Mean	Min	Max
3	Conventions	27	508.5	144	831	67.5	59.8	74.5	100	0.67	0.60	0.73
	Evid/Elab	27	508.5	144	831	71.0	59.6	84.7	100	0.72	0.61	0.84
	Org/Purp	27	508.5	144	831	70.9	58.5	84.5	100	0.72	0.61	0.84
4	Conventions	28	504.9	86	938	66.0	55.4	82.1	100	0.70	0.61	0.87
	Evid/Elab	28	504.9	86	938	68.8	58.8	79.0	100	0.75	0.65	0.89
	Org/Purp	28	504.9	86	938	68.8	59.1	77.1	100	0.75	0.66	0.88
5	Conventions	29	472.8	103	758	68.3	62.1	75.3	100	0.68	0.59	0.77
	Evid/Elab	29	472.8	103	758	68.3	57.0	76.3	100	0.77	0.65	0.89
	Org/Purp	29	472.8	103	758	68.5	57.2	76.1	100	0.77	0.65	0.88
6	Conventions	21	564.3	97	997	68.8	63.4	73.2	100	0.70	0.62	0.76
	Evid/Elab	21	564.3	97	997	70.5	63.8	78.2	100	0.77	0.68	0.89
	Org/Purp	21	564.3	97	997	70.8	63.8	78.6	100	0.77	0.68	0.89
7	Conventions	29	433.8	95	797	69.9	63.8	78.0	100	0.68	0.54	0.79
	Evid/Elab	29	433.8	95	797	71.0	60.0	80.7	100	0.76	0.65	0.88
	Org/Purp	29	433.8	95	797	71.2	61.3	80.1	100	0.77	0.67	0.88
8	Conventions	30	409.4	84	794	72.0	62.9	82.0	100	0.69	0.59	0.78
	Evid/Elab	30	409.4	84	794	70.9	63.2	80.8	100	0.78	0.73	0.86
	Org/Purp	30	409.4	84	794	71.1	63.0	81.1	100	0.78	0.73	0.87
11	Conventions	29	312.1	181	658	70.2	63.7	76.1	100	0.72	0.67	0.79
	Evid/Elab	29	312.1	181	658	75.5	68.7	82.7	100	0.82	0.75	0.86
	Org/Purp	29	312.1	181	658	75.6	68.7	82.3	100	0.82	0.75	0.86

Note. Evid/Elab: Evidence/Elaboration, Org/Purp: Organization/Purpose

Table 60. 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	12	698.3	418	1210	93.4	86.9	98.3	100	NA	NA	NA
4	0–1	10	767.7	589	1201	89.7	82.5	97.1	100	NA	NA	NA
5	0–1	12	580.6	417	1053	93.5	83.6	97.8	100	NA	NA	NA
6	0–1	10	1066.0	329	2111	96.9	85.0	99.7	100	NA	NA	NA
7	0–1	12	1374.7	620	2099	95.1	86.9	99.2	100	NA	NA	NA
8	0–1	10	1747.1	934	2114	86.8	79.1	98.3	100	NA	NA	NA
11	0–1	15	557.9	51	1605	95.5	91.6	100.0	100	NA	NA	NA
3	0–2	34	915.5	129	1762	91.3	79.6	99.7	100	0.92	0.81	0.98
4	0–2	39	875.9	187	1625	92.1	82.9	99.6	100	0.91	0.73	1.00
5	0–2	65	731.2	402	1293	88.7	78.5	96.7	100	0.86	0.59	0.97
6	0–2	41	1397.4	636	1928	89.2	76.6	99.3	100	0.86	0.70	0.99
7	0–2	15	1591.7	749	2075	89.6	84.2	94.2	100	0.84	0.61	0.94
8	0–2	21	1420.4	785	2376	88.6	76.3	98.8	100	0.86	0.70	0.98
11	0–2	21	825.6	288	1717	92.0	78.8	99.4	100	0.85	0.57	0.97
3	0-3	4	951.3	323	1739	89.9	86.4	94.7	100	0.95	0.92	0.98
4	0-3	4	554.8	501	687	89.6	87.0	93.0	100	0.95	0.94	0.97
5	0-3	9	789.6	273	1244	87.3	81.2	96.0	100	0.90	0.86	0.95
7	0-3	1	1955.0	1955	1955	93.0	93.0	93.0	100	0.93	0.93	0.93
8	0-3	4	1913.8	1789	1992	82.4	78.3	89.1	100	0.94	0.93	0.97
11	0-3	7	1586.0	1410	1753	87.1	78.6	93.1	100	0.87	0.80	0.91

Note. QWK is not presented for 0–1 items due to the binary score scale.

6.8 AUTOMATED SCORING

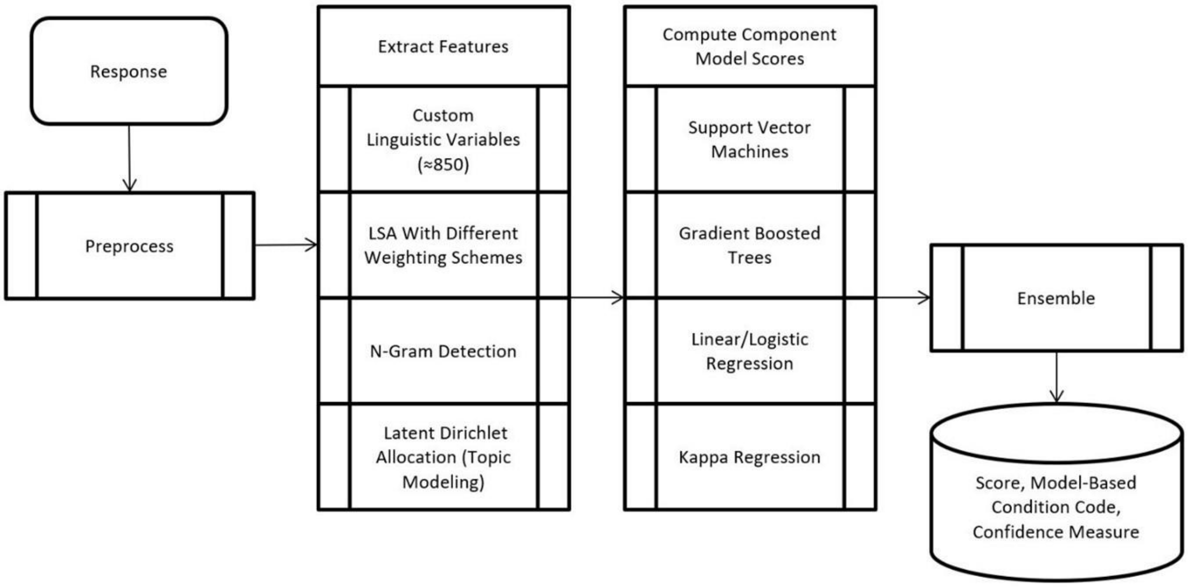
MI's PEG automated scoring technology was used to score eligible SA and essay items in ELA/L and SA items in mathematics. This section describes PEG, the model training and validation sample and process, the automated scoring process, and the human-machine (HM) agreement statistics.

6.8.1 Project Essay Grade

Figure 13 presents the architecture of MI's PEG engine. During engine training, this architecture allows PEG to generate hundreds of custom linguistic (rule-based) features, which are determined by codified English linguistic rules such as syntax and semantics and extracted from representative student responses. In addition to rule-based features, PEG also includes features extracted by Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) procedures.

PEG's item and trait specific scoring models use computed features from the training responses along with the scores assigned to them by expert human raters. Using hundreds of parameterizations across several machine-learning algorithms, via cross-validation and optimization, PEG determines which algorithms best predict the expert-assigned scores. These algorithms draw on many of the latest advances in the field of machine learning to generate linear and non-linear classification and regression models. These approaches typically result in 100 candidate models for a single item or trait. PEG then uses an ensembling procedure to combine the best models into a robust final model. The ensembling procedure utilizes linear regression, where the objective is to maximize a continuous relaxation of QWK, thus maximizing PEG's agreement with the expert human raters.

Figure 13. PEG Architecture


Response
Preprocess
Extract Features
Custom Linguistic Variables (≈850)
LSA With Different Weighting Schemes
N-Gram Detection
Latent Dirichlet Allocation (Topic Modeling)
Compute Component Model Scores
Support Vector Machines
Gradient Boosted Trees
Linear/Logistic Regression
Kappa Regression
Ensemble
Score, Model-Based Condition Code, Confidence Measure

The sections that follow describe the process used to train and validate the engine, followed by a description and results of the hybrid human-automated scoring process.

6.8.2 Model Training and Validation

Automated scoring models were not created for items that had an insufficient quantity of training responses. This was the case for items with low exposure to students, as dictated by the adaptive testing algorithm. Table 61 shows that pretrained models existed for 650 items, thus, no additional training was conducted in preparation for the spring 2025 administration. The remainder of this section describes the process used to train and validate the existing models associated with the 650 items.

Table 61. Number of Items Eligible for Model Training, by Grade and Subject Area

Grade	Item with existing model			Items without models		
	ELA/L		Mathematics	ELA/L		Mathematics
	Short-Answer	Essay		Short-Answer	Essay	
3	13	19	40	0	0	0
4	15	22	42	0	0	0
5	14	21	69	0	0	0
6	34	16	45	0	0	0
7	43	19	21	0	0	0
8	50	15	31	0	0	0
11	55	21	45	0	0	0
Total	224	133	293	0	0	0

Training Data

Student responses used for training and validation were sourced from the 2018–2019 through 2023–2024 Smarter Balanced operational test administrations. Responses were randomly sampled from available on-grade responses in the operational population. For each item, the sample included 1,500–2,000 responses, stratified by score point. The score of record used to train the engine was the score assigned to each response by an expert rater.

For each item, the sample was divided as follows:

- Approximately 85% of the responses were assigned to a training set used to build the model.
- Approximately 15% of the responses were assigned to a validation set used to evaluate the accuracy of the model.

Model Training

Essay scoring model

Component model training requires inputs of response "features." For items that assess writing quality (e.g., essays), PEG processes the responses and calculates approximately 850 linguistic variables that describe the responses in mathematical terms. These variables range in complexity from simple to highly complex. Examples of simple variables are measures such as word count or sentence length, word choice and spelling errors, and the number and severity of grammatical errors. The most complex variables measure patterns that represent style, fluidity, smoothness of transitions, clarity of communication, and other sophisticated concepts.

To build an essay scoring model, PEG examines the variables and text features of responses, correlates them with the human scores previously assigned, and identifies those variables that have high predictive value.

Content scoring model

For content-based items (e.g., SA mathematics items), the number of variables is unknown until the models are built. Because the content varies significantly from item to item, and therefore from model to model, PEG examines training responses and identifies the variables that most accurately capture the content in question. To do this, MI uses techniques like LSA, N-Gram Detection, and LDA. To further refine the variable generation process, MI built a computer language to perform a simultaneous search over semantic, lexicographic and syntactic features of responses.

To build a content scoring model, PEG analyzes training responses and calculates features that pertain to the content in question. PEG then sends the features to hundreds of different algorithms that compete to see which algorithms best associate the features with the human-assigned scores. These algorithms draw on many of the latest advances in the field of machine learning to generate both linear and non-linear models. Examples of approaches used include Support Vector Machines, Gradient Boosted Trees, and various regression approaches.

Component models

Note that building component models for each item—and for multi-dimensional items, each trait or dimension—prevents variables from being generalized across items or traits, allowing PEG to faithfully reproduce humans' application of the scoring rubrics. This means that the resultant models are reasonably robust to gaming attempts, as each represents a unique valuation of the item- (or trait-) specific text features similarly valued by expert professional raters.

The approaches just described typically result in 100 models for a single item or essay trait. Ensembling is the process of selecting the "best of the best" models, to result in a small set of strong, yet dissimilar component models. A linear-kappa regression is used to determine the model ensembling weights. The more accurate a given model is, the more weight it carries in the final score decision.

Scoring a response involves first preprocessing the response. The purpose of preprocessing is twofold: (1) create raw and canonical representations of the response from which features can be extracted, and (2) filter out responses for which the scoring model does not apply (e.g., blank or insufficient responses). The response is then scored with the associated component models. A final score is produced performing a weighted sum using the ensembling weights.

Model Validation

Model validation involved a two-phase approach: an initial validation using held-out training data and a secondary validation using operational data from the current administration.

Initial Validation

Initial validation was conducted by applying each model to score a respective validation set of responses. The validation set is independent of the training set, in that none of the responses it contains have been used to build the model. It should be noted that two or more professional raters will not always agree on what score to give a student's response; therefore, modeling is considered successful when the engine produces scores that agree with professional raters to the same or greater extent than the raters agree with each other. The initial evaluation was made using the criteria shown in Table 62, based on criteria proposed by Williamson, Xi, and Breyer (2012). While Williamson et al. (2012) recommend a QWK of 0.70 between human and machine scores for normally distributed data, a QWK threshold of 0.65 was adopted due to the prevalence of skewed distributions in response data. For human-human score agreement, the degradation (QWK) criterion of 0.07 is slightly more stringent than proposed by Williamson et al. (2012). The evaluation process was used for both the item-specific scoring models and the condition code models.

Table 62. Initial Model Evaluation Criteria

Criterion	Threshold
Agreement of automated scores with human scores	$QWK_{H:M} \geq 0.65$
Degradation from the human-human score agreement	$QWK_{H:H} - QWK_{H:M} < 0.07$
Standardized mean score difference between human and automated scores	$\lvert SMD_{H:M}\rvert < 0.15$

Note. QWK = Quadratic weighted kappa. SMD = Standardized mean difference. H:H = human:human. H:M = human:machine.

Bias Considerations

Subgroup differences in responses to constructed response items can introduce construct-irrelevant variance in scores, in turn threatening valid score interpretations. MI investigated potential sources of bias annually, for newly modeled items, as part of the initial validation process using available data from previous summative administration. Table 63 shows the demographic variables and categories considered. MI received separate data files containing handscore data and student demographic data associated with responses.

Table 63. Demographic Variables and Categories

Demographic Variable	Categories
Gender	Male Female
Race/Ethnicity	American Indian or Alaska Native Asian Native Hawaiian or Pacific Islander Filipino Hispanic or Latino Black or African American White Two or More Races
LEP Status	LEP Non LEP

For each new item being modeled, we analyzed a subgroup if there were at least 10 observations (human–machine score pairs). A subgroup was flagged for potential bias if the absolute SMD between human and machine scores exceeded 0.125 and the difference was statistically significant, controlling the family-wise error rate at $\alpha = 0.05$ via a Bonferroni correction (i.e., using a Bonferroni-adjusted two-sided α for each subgroup comparison).

Secondary Validation

All models associated with items that passed initial validation were subject to a secondary validation at the start of the spring 2025 administration using an early sample of operational responses from that administration. This sample was comprised of the first available 500 responses/item across states, at a minimum. Responses from this sample were scored by both the automated scoring engine and an expert rater. During this stage the human score was reported as the score of record. If the PEG scores were found to be consistent with the scores assigned by the expert raters, subsequent student responses for a given item were scored by PEG using a hybrid human-automated scoring approach. If not, the item was handscored. Table 64 presents the secondary validation criteria. Note that since expert raters are the only humans that score the secondary validation sample, a second human score is not collected and thus QWK degradation is not part of the criteria.

Table 64. Secondary Validation Criteria

Criterion	Threshold
Agreement of automated scores with human scores	$\text{QWK}_{\text{H:M}} \geq 0.65$
Standardized mean score difference between human and automated scores	$\lvert\text{SMD}_{\text{H:M}}\rvert \leq 0.15$

Note. QWK = Quadratic weighted kappa. SMD = Standardized mean difference. H:M = human:machine.

Table 65 presents the secondary validation results. Of the 650 items with existing models subject to secondary validation, models associated with 540 of the items (83.1%) passed all secondary evaluation criteria.

Table 65. Summary of Secondary Validation Results, by Grade and Subject Area

Grade	Items with all Models Passing Initial Validation Criteria			Items with all Models Passing Secondary Validation Criteria		
	ELA/L		Mathematics	ELA/L		Mathematics
	Short-Answer	Essay		Short-Answer	Essay	
3	13	19	40	9	12	38
4	15	22	42	14	19	37
5	14	21	69	12	11	60
6	34	16	45	31	10	41
7	43	19	21	31	14	18
8	50	15	31	43	13	24
11	55	21	45	50	14	39
Total	224	133	293	190	93	257

Live Training and Validation

Additionally, in April–May 2025 when operational scoring was underway, a live training and validation effort was undertaken for those handscored items lacking validated models from prior efforts but having sufficient 2025 operational responses to train and validate new models. In general, these items were associated with models that had previously failed an initial and/or secondary validation. In such cases, training with 2025 operational responses offered potential to improve model performance. All models associated with these items were thus trained using either exclusively 2025 responses (when a minimum of 1,400 2025 responses/item existed) or 2025 responses supplemented with 2024 responses. In either case, the validation sets consisted exclusively of 2025 responses. Because this live validation involved operational data, it was unnecessary to conduct a secondary validation.

Table 66 summarizes the results of the live training and validation. Of the 261 items associated with models that underwent live training and validation, models associated with 214 of the items (82%) passed all evaluation criteria. Following initial validation, secondary validation, and live training and validation, a total of 754 items, comprised of 271 ELA/L SA, 173 essay, and 310 mathematics SA, were scored using a hybrid process, described next.

Table 66. Summary of Live Training and Validation Results, by Grade and Subject Area

Grade	Items Trained			Items with all Models Passing Initial Validation Criteria		
	ELA/L		Mathematics	ELA/L		Mathematics
	Short-Answer	Essay		Short-Answer	Essay	
3	8	10	12	8	9	7
4	3	4	13	3	3	8
5	4	15	26	4	15	10
6	14	10	10	11	10	6
7	21	13	10	19	13	8
8	28	15	11	25	14	10
11	13	16	5	11	16	4
Total	91	83	87	81	80	53

6.8.3 Automated Scoring Processes

Hybrid Scoring Process

As all models associated with a given item passed secondary validation (or live validation), subsequent student responses were scored using a hybrid human-automated scoring approach. If all models associated with a given item did not pass secondary validation, responses associated with the item continued to be handscored by the larger pool of raters. These raters were monitored and evaluated as described in the handscoring section above.

Figure 14 shows the response routing rules under the hybrid scoring process. In the hybrid model, responses with associated scoring models were first pre-processed for automated scoring; “alert” responses with an alert and certain non-scorable cases (e.g., insufficient text to score or high proportion of copied prompt text) were filtered and flagged. Table 67 and 68 present the flags and model settings used to indicate condition codes as defined in the handscoring criteria. For example, PEG flags responses that lack proper development, lack enough content to be scored, are written in an unsupported language, or contain vulgar language or other alert words or phrases that indicate that the response should be reviewed by the client. Standard scoring responses were then sent to the automated scoring engine, where text features were extracted, the scoring model(s) applied, and responses assigned a score and a measure of score confidence. Low-confidence responses straddle the lines between score point values on a rubric and are difficult to score accurately because they exhibit characteristics of multiple score points. Higher-confidence responses received the engine score as the score of record, while lower-confidence responses were routed directly to expert raters, who assigned the score of record. Note that the expert rater pool was dynamic, and raters were added or removed several times each day based on their current performance. Overall, approximately 15% of responses to engine-scored items were flagged as low confidence and scored by expert raters. Upon receipt and validation of each response, MI routed responses for those items eligible for automated scoring to PEG and the remainder of the responses to the VSC handscoring system.

Figure 14. Response Routing Rules

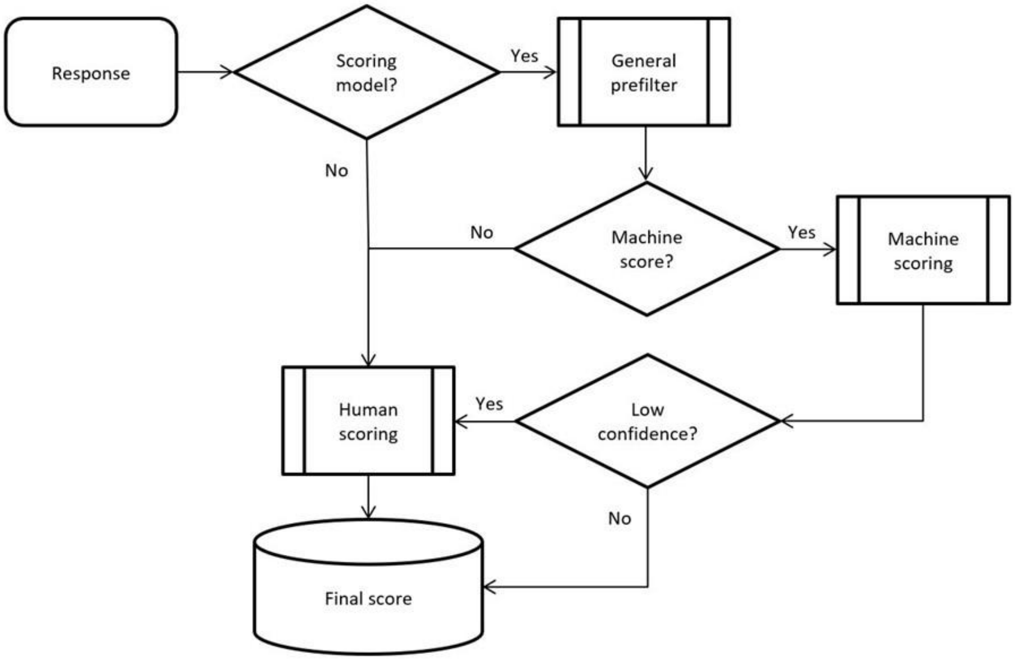


Table 67. Flags Currently Established

FLAG	USAGE DESCRIPTION	*SCORABLE
0	Standard scoring	YES
200	Too few words (i.e., blank, or extremely short response)	NO
240	Too long (i.e., too many characters submitted; 30,000 characters is the current limit)	NO
250	Expected essay fields are null or empty; set when nulls are discovered within the processing pipeline. Not client configurable.	NO
400	Unexpected item_id (i.e., the item_id is not one of the items PEG AI has modeled)	NO
500	Scorable alert (i.e., an essay which seems perfectly scorable, but happens to contain alert language); client may configure alert scanning to “on” or “off”, but other changes are not recommended.	YES
501-599	Non-scorable alert (i.e., alert language was detected, and the essay could not be scored). If alert scanning is “on”, then any code in the 500-599 range is possible. Not client configurable.	NO
620	Applies when the ratio of copied characters exceeds specified threshold (e.g., 0.5 means 50%). Can be used for all Smarter items for which prompt content was provided.	YES
650	Insufficient Condition Code (I): Response holds strong general resemblance to those marked ‘Insufficient’ by human readers, but is nonetheless PEG scorable (and, so scores are provided). *PEG Configuration:* Item agnostic; but for 2021 onwards, applicable to ELA/L items only.	YES
660	Language Non-English Condition Code (L): Response holds strong general resemblance to those marked ‘Non-English’ by human readers, but is nonetheless PEG scorable (and, so scores are provided). *PEG Configuration:* Item agnostic; but for 2021 onwards, applicable to ELA/L items only.	YES
670	Off-Topic: Applicable to ELA/L essays only and is item specific in the PEG environment.	YES
680	Off-Mode: Applicable to ELA/L essays only and is item specific in the PEG environment.	YES
900	Timeout (i.e., unable to complete essay score prediction within time limits). Not client configurable.	NO
950	System error processing essay (i.e., internal PEG error). Not client configurable.	NO

Note. Scorable flags indicate instances where PEG will return both the applicable flag and a score.

Table 68. Model Settings

TYPE	ASSOCIATED FLAG(S)	DESCRIPTION	VALUES
Minimum Words	200	Triggers if there are fewer than the associated value of word-tokens in a response. The flag may also appear regardless of setting if the response is blank.	0-15
Alert	501-599	Current setting (PREDC...1) is for the standard alert scan.	Standard settings in place
Plagiarism	620	Prompt and source material text is included in model configuration.	50% of prompt and source material characters triggers flag

Scoring Infrastructure

During the automated scoring process, response data are transferred from CAI to MI's IT project team. Data are then passed to PEG from the IT project team via an internal server, at which point they are processed through the PEG Streaming Scoring Service—a cloud-deployed, horizontally scalable, distributed parallel computing application. Scored batches were typically completed within one day. All data are then transferred from PEG to the IT project team, who ultimately sends the data/scores back to CAI.

Score Delivery

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

Quality Assurance

MI's hybrid scoring approach included numerous quality assurance steps. First, models were trained exclusively using scores assigned by expert raters and the associated responses. Second, each automated scoring model was subjected to an evaluation process, as described in the model validation section. This involved evaluating the quality of the human-scored training data, as well as comparing the performance of the engine to the performance of expert raters. Third, for models trained using responses from prior administrations, the generalizability of each model to the 2024–2025 operational responses were confirmed via a secondary validation. Finally, quality was further assured during scoring by routing a minimum of 15% of the responses that were most different from the training responses to expert raters and assigning the human score.

"Alert" Procedures

MI implemented a process for routing any student responses flagged by the automated scoring engine as possible alerts to expert raters for review. Any responses identified as alerts by expert raters and/or scoring leadership are sent to CAI, who associates the pertinent student information with the response(s) and contacts the state for follow-up.

6.8.4 Human-Machine Agreement

This section summarizes the human-machine (HM) agreement for all items scored using a hybrid process in spring 2025, including (1) items passing initial model validation, (2) items passing secondary validation, and (3) items passing live validation.

Table 69 through Table 71 present the HM agreement rates on the initial and secondary validation samples for ELA/L SA items, ELA/L essay items, and mathematics SA items, respectively. The HM agreement was computed based on the combined data across all states with hybrid scoring in the 2024–2025 summative assessment.

Table 69. Human-Machine Agreement for ELA/L Short-Answer Items on Initial and Secondary Validation Samples, by Grade

Grade	Initial Validation				Secondary Validation			
	Number of Items	% Exact	% (Exact+ Adjacent)	QWK	Number of Items	% Exact	% (Exact+ Adjacent)	QWK
3	9	79.6	99.6	0.81	9	83.3	99.8	0.80
4	14	80.1	99.8	0.84	14	82.0	99.8	0.82
5	12	75.4	99.6	0.81	12	79.1	99.9	0.82
6	31	78.7	99.5	0.81	31	79.0	99.6	0.76
7	31	76.3	99.4	0.79	31	78.1	99.6	0.76
8	43	76.2	99.5	0.78	43	75.4	99.6	0.76
11	50	77.2	99.5	0.79	50	76.4	99.6	0.77

Table 70. Human-Machine Agreement for ELA/L Essay Items on Initial and Secondary Validation Samples, by Grade

Grade	Trait	Initial Validation				Secondary Validation			
		Number of Items	% Exact	% (Exact+ Adjacent)	QWK	Number of Items	% Exact	% (Exact+ Adjacent)	QWK
3	Conventions	12	72.1	99.8	0.73	12	72.0	99.5	0.71
	Evid/Elab	12	77.7	99.3	0.83	12	76.2	99.4	0.78
	Org/Purp	12	75.3	99.7	0.82	12	76.9	99.6	0.79
4	Conventions	18	69.5	98.9	0.74	18	68.9	99.1	0.72
	Evid/Elab	18	72.7	99.6	0.84	18	73.7	99.5	0.79
	Org/Purp	18	72.2	99.2	0.83	18	74.5	99.5	0.79
5	Conventions	10	71.7	99.6	0.70	10	71.1	99.7	0.71
	Evid/Elab	10	72.0	99.2	0.81	10	73.3	99.5	0.81
	Org/Purp	10	72.2	99.6	0.83	10	73.6	99.4	0.81
6	Conventions	10	75.5	99.0	0.72	10	75.5	99.4	0.74
	Evid/Elab	10	71.4	98.7	0.78	10	74.7	99.4	0.79
	Org/Purp	10	74.5	98.7	0.81	10	75.6	99.6	0.80
7	Conventions	14	76.1	99.7	0.70	14	73.0	99.5	0.71
	Evid/Elab	14	75.6	99.7	0.83	14	77.7	99.8	0.82
	Org/Purp	14	75.6	99.6	0.84	14	77.6	99.7	0.81
8	Conventions	13	77.0	99.1	0.71	13	74.4	99.6	0.72
	Evid/Elab	13	73.7	99.1	0.82	13	76.0	99.7	0.82
	Org/Purp	13	75.1	99.7	0.84	13	76.0	99.8	0.82
11	Conventions	14	79.1	99.7	0.75	14	78.4	99.7	0.73
	Evid/Elab	14	76.5	99.7	0.86	14	76.0	99.7	0.83
	Org/Purp	14	76.4	99.7	0.86	14	76.1	99.8	0.83

Note. Evid/Elab: Evidence/Elaboration, Org/Purp: Organization/Purpose

Table 71. Human-Machine Agreement for Mathematics Items on Initial and Secondary Validation Samples, by Grade

Grade	Score Point Range	Initial Validation				Secondary Validation			
		Number of Items	% Exact	% (Exact+ Adjacent)	QWK	Number of Items	% Exact	% (Exact+ Adjacent)	QWK
3	0-1	11	94.2	100.0	NA	11	95.5	100.0	NA
4	0-1	5	91.0	100.0	NA	5	94.2	100.0	NA
5	0-1	9	92.6	100.0	NA	9	94.3	100.0	NA
6	0-1	8	96.6	100.0	NA	8	96.0	100.0	NA
7	0-1	8	96.9	100.0	NA	8	95.8	100.0	NA
8	0-1	4	90.2	100.0	NA	4	92.7	100.0	NA
11	0-1	15	95.6	100.0	NA	15	95.5	100.0	NA
3	0-2	23	90.8	99.3	0.91	23	91.8	99.4	0.90
4	0-2	28	91.0	99.7	0.91	28	92.8	99.8	0.90
5	0-2	48	88.3	99.6	0.88	48	87.4	99.6	0.85
6	0-2	33	88.9	99.6	0.86	33	89.1	99.5	0.85
7	0-2	10	87.0	99.4	0.80	10	87.0	99.5	0.79
8	0-2	18	89.1	99.8	0.89	18	90.7	99.7	0.88
11	0-2	17	89.1	99.4	0.88	17	90.8	99.5	0.87
3	0-3	4	91.1	99.8	0.96	4	91.7	99.8	0.95
4	0-3	4	87.9	99.8	0.94	4	87.5	99.7	0.92
5	0-3	3	90.9	98.4	0.94	3	90.3	98.9	0.91
8	0-3	2	78.2	98.0	0.88	2	75.0	97.9	0.88
11	0-3	5	85.5	99.0	0.89	5	89.8	99.3	0.90

Note. QWK is not presented for 0-1 items due to the binary score scale.

Table 72 through Table 74 present the HM agreement rates on the live validation samples for ELA/L SA items, ELA/L essay items, and mathematics SA items, respectively. Recall live training did not involve secondary validation since 2024–2025 operational data were used to build the models.

Table 72. Human-Machine Agreement for ELA/L Short-Answer Items on Live Validation Sample, by Grade

Grade	Live Validation			
	Number of Items	% Exact	% (Exact+ Adjacent)	QWK
3	8	80.2	99.6	0.78
4	3	78.9	99.4	0.78
5	4	78.0	99.9	0.80
6	11	80.0	99.7	0.75
7	19	78.1	99.6	0.74
8	25	78.0	99.7	0.75
11	11	77.9	99.8	0.74

Table 73. Human-Machine Agreement for ELA/L Essay Items on Live Validation Sample, by Grade

Grade	Trait	Live Validation			
		Number of Items	% Exact	% (Exact+ Adjacent)	QWK
3	Conventions	8	75.0	99.6	0.72
	Evid/Elab	8	76.6	99.2	0.77
	Org/Purp	8	75.9	99.1	0.77
4	Conventions	3	73.0	99.4	0.74
	Evid/Elab	3	68.5	97.6	0.72
	Org/Purp	3	70.6	98.2	0.76
5	Conventions	12	74.5	99.7	0.71
	Evid/Elab	12	74.3	99.5	0.80
	Org/Purp	12	75.1	99.6	0.81
6	Conventions	10	76.1	99.6	0.74
	Evid/Elab	10	73.0	99.6	0.81
	Org/Purp	10	73.0	99.7	0.80
7	Conventions	12	75.5	99.8	0.73
	Evid/Elab	12	76.4	99.8	0.84
	Org/Purp	12	76.7	99.4	0.84
8	Conventions	14	76.8	99.7	0.73
	Evid/Elab	14	75.4	99.7	0.83
	Org/Purp	14	74.7	99.8	0.83
11	Conventions	16	77.3	99.4	0.73
	Evid/Elab	16	77.1	99.9	0.85
	Org/Purp	16	77.1	99.8	0.85

Notes.
1). Evid/Elab: Evidence/Elaboration, Org/Purp: Organization/Purpose.
*2).*The number of items is slightly fewer than in Table 10 for grades 3, 5, and 7 due to missing data.

Table 74. Human-Machine Agreement for Mathematics Items on Live Validation Samples, by Grade

Grade	Score Point Range	Live Validation			
		Number of Items	% Exact	% (Exact+ Adjacent)	QWK
3	0-1	1	92.9	100.0	NA
4	0-1	5	89.6	100.0	NA
5	0-1	2	94.4	100.0	NA
6	0-1	1	86.4	100.0	NA
7	0-1	3	94.9	100.0	NA
8	0-1	5	85.3	100.0	NA
3	0-2	1	92.0	100.0	NA
5	0-2	6	88.8	99.7	0.89
6	0-2	3	86.0	99.1	0.83
7	0-2	7	85.1	99.4	0.77
8	0-2	5	89.8	99.1	0.84
11	0-2	4	87.2	98.6	0.86
5	0-3	3	81.2	99.4	0.82
7	0-3	2	91.1	100.0	0.91
8	0-3	1	79.5	97.7	0.86
11	0-3	1	88.2	98.1	0.86

Note. QWK is not presented for 0-1 items due to the binary score scale.

6.8.5 Recommendations

The 2023–2024 summative administration identified two key areas for improvement: (a) strengthening automated oversight and refining accuracy monitoring, which led to expanding the additional rater-validation stage to all ELA/L item types and adding mean-score distribution checks alongside QWK; and (b) addressing production risk by improving rater availability and hours worked. The 2023–2024 technical report also noted variability in some ELA short-answer items, where relatively low minimum QWK values indicated a need for targeted calibration.

In 2025, MI piloted a core pool of seasonal, full-time contractors with guaranteed hours and introduced higher minimum work-hour requirements. While uptake into the core pool was lower than anticipated, those who accepted the role performed well, and the minimum-hours requirement was successfully implemented. These changes supported consistency and reliability; however, there remain opportunities to further improve score accuracy and manage workflows to ensure that 100% of responses are scored on time.

Maintaining and refining the core seasonal contractor model could strengthen performance and reliability. A guaranteed-hours arrangement for raters and team leaders—paired with higher compensation tied to training participation, qualification results, production, and accuracy—would help retain skilled staff. Recruitment efforts could begin earlier in the year, with contingent offers extended ahead of peak scoring months to secure commitments. Completion incentives could reinforce season-long engagement, while flexibility in scheduling (including evening and weekend options) should be balanced with strict enforcement of the 37.5-hour weekly minimum. Clear accountability measures and early identification of underperforming contractors would allow for timely reassignment or contract termination, keeping the pool productive and high-quality throughout the scoring season.

In addition, certain ELA short-answer items with a history of lower QWK values should continue to be a focus for targeted calibration. Building on the enhanced validity pool and monitoring tools introduced in 2025, MI could deploy supplemental materials—such as annotated exemplars, guided scoring exercises, and focused discussions of common scoring pitfalls—for these specific items. Intensified calibration at the start of the scoring season, followed by ongoing monitoring and rapid feedback when accuracy falls below thresholds, would help sustain reliability for these more challenging items and reinforce consistency across the scoring population.

Finally, reducing production pressure could be supported by improving rater preparation prior to operational scoring. Training could incorporate learning-science principles while continuing to comply with Smarter Balanced training requirements, emphasizing both accuracy and efficiency from the outset. Strategies might include breaking content into manageable portions, providing scaffolded examples, and offering regular deliberate practice with varied responses and timely feedback. Encouraging raters to reflect on and explain their scoring decisions could further strengthen consistency. Advancement through training could remain contingent on meeting clear performance thresholds, with targeted remediation for those who do not qualify.

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 handscored items are scored. Because the score reports on students' performance are updated every time students complete tests and handscored items are scored, authorized users (e.g., school principals, teachers) can readily access information on students' test performance and use it to improve student learning. In addition to individual student's score reports, the CRS also produces aggregate score reports by class, school, complex, complex area, and state. 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 detailed description of the types of scores reported in the CRS and how to interpret and use these scores.

7.1 CENTRALIZED REPORTING SYSTEM

The CRS is designed to help educators and students answer questions about how well students have performed on the English language arts/literacy (ELA/L) and mathematics assessments. The CRS is an online tool that provides all stakeholders with timely, relevant score reports. The CRS for the Smarter Balanced assessments was 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 avoid comparing dissimilar elements.

Generally, the CRS provides two categories of online score reports: (1) aggregate score reports, and (2) student score reports. Table 75 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 in the CRS.

Table 75. Types of Online Score Reports by Level of Aggregation

Level of Aggregation	Types of Online Score Reports
State Complex Area Complex School Teacher Roster	• Number of students tested and percentage of proficient students (for overall students and by subgroup) • Average scale score and standard error of average scale score on the overall test and claim (for overall students and by subgroup) • Percentage of students at each achievement level on the overall test (for overall students and by subgroup) • Performance category in each target (for overall students) • On-demand student roster report
Student	• Total scale score and standard error of measurement • Achievement level for the overall score and claim scores with achievement-level descriptors • Average scale scores and standard errors of average scale scores for individual complex, complex areas, and states • Writing performance descriptors and scores by dimensions

Aggregate score reports at a selected aggregate level are provided for overall students and by subgroup. Users can see student assessment results by any of the subgroups. Table 76 presents the types of subgroups and subgroup categories provided in the CRS.

Table 76. Types of Subgroups with Subgroup Categories

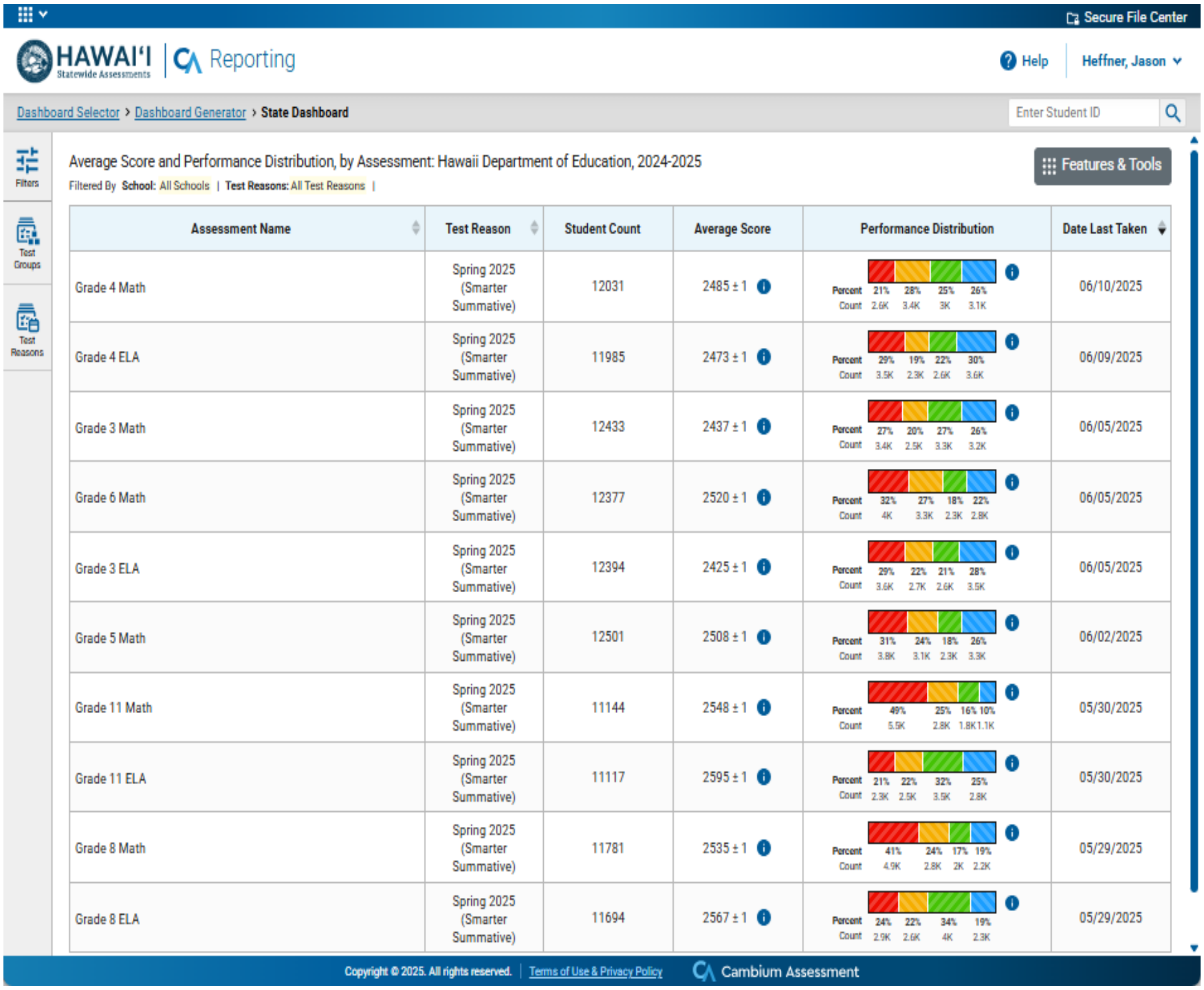
Subgroup	Subgroup Category
Gender	Male Female
ELL	Yes No
Disability	01 - Autism 02 - Deaf-Blindness 03 - Deafness 04 - Developmental Delay (Age 3-5) 05 - Developmental Delay (Age 6-8) 06 - Emotional Disturbance 07 - Hearing Impaired 08 - Mental Retardation 09 - Multiple Disability 10 - Orthopedic Impairment 11 - Other Health Impairment 12 - Specific Learning Disability 13 - Speech/Language Impairment 14 - Traumatic Brain Injury 15 - Visual Impairment including Blindness 16 - Autism Spectrum Disorder 17 - Other Health Disability 18 - Speech or Language Disability 19 - Intellectual Disability 20 - Visual Disability Including Blindness 21 - Hard of Hearing 22 - Orthopedic Disability
Migrant Status	Yes No
Disadvantaged	C, D, E, F, R, 1, 2, 3
Ethnicity	American Indian/Alaskan Native Asian/Pacific Islander African American Hispanic Hawaiʻi Pacific Islander White Multi-Racial

7.1.1 Dashboard

The CRS provides a state dashboard for authorized state-level users to track student performance for a test across the entire state. The dashboard summarizes students' performance for both ELA/L and mathematics in each grade, including (1) student count, (2) average score and standard error of the average score, (3) percentage and counts of students at each achievement level, and (4) test date last taken.

Exhibit 1 presents a sample state dashboard page.

Exhibit 1. Dashboard: State Level



Dashboard Selector > Dashboard Generator > State Dashboard

Average Score and Performance Distribution, by Assessment: Hawaii Department of Education, 2024-2025

Filtered By School: All Schools | Test Reasons: All Test Reasons |

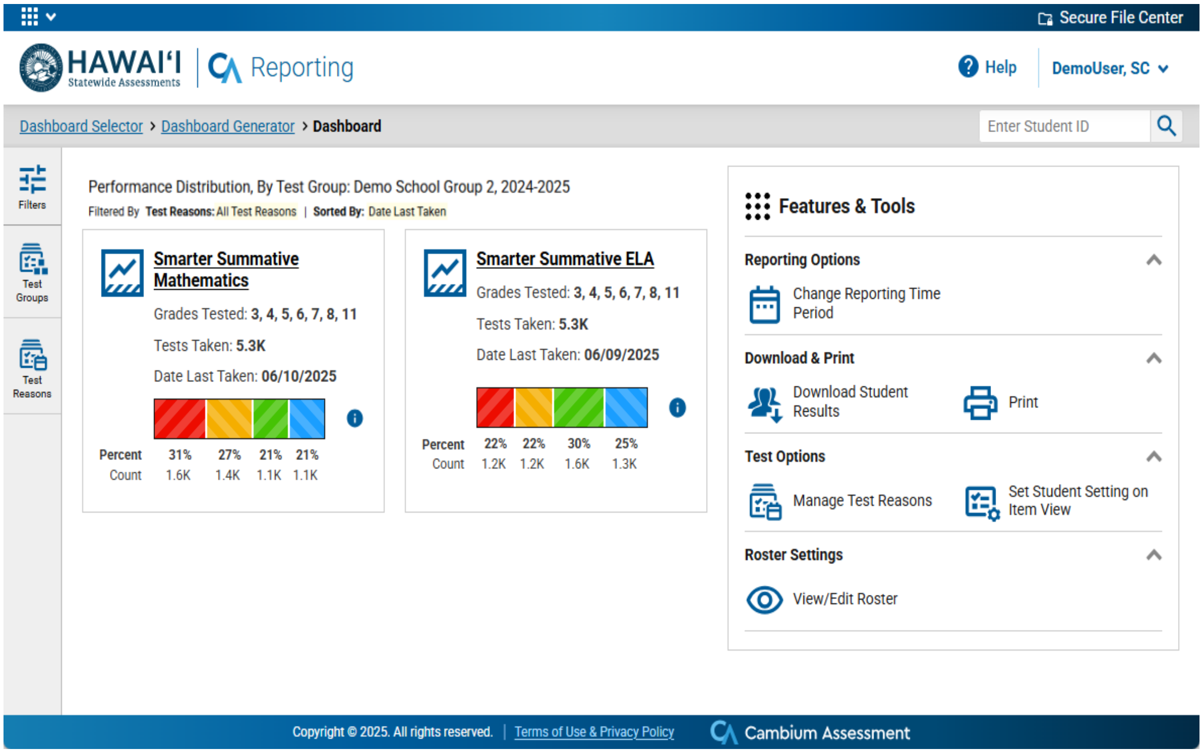
Assessment Name	Test Reason	Student Count	Average Score	Performance Distribution	Date Last Taken
Grade 4 Math	Spring 2025 (Smarter Summative)	12031	2485 ± 1	Percent 21% 28% 25% 26% Count 2.6K 3.4K 3K 3.1K	06/10/2025
Grade 4 ELA	Spring 2025 (Smarter Summative)	11985	2473 ± 1	Percent 29% 19% 22% 30% Count 3.5K 2.3K 2.6K 3.6K	06/09/2025
Grade 3 Math	Spring 2025 (Smarter Summative)	12433	2437 ± 1	Percent 27% 20% 27% 26% Count 3.4K 2.5K 3.3K 3.2K	06/05/2025
Grade 6 Math	Spring 2025 (Smarter Summative)	12377	2520 ± 1	Percent 32% 27% 18% 22% Count 4K 3.3K 2.3K 2.8K	06/05/2025
Grade 3 ELA	Spring 2025 (Smarter Summative)	12394	2425 ± 1	Percent 29% 22% 21% 28% Count 3.6K 2.7K 2.6K 3.5K	06/05/2025
Grade 5 Math	Spring 2025 (Smarter Summative)	12501	2508 ± 1	Percent 31% 24% 18% 26% Count 3.8K 3.1K 2.3K 3.3K	06/02/2025
Grade 11 Math	Spring 2025 (Smarter Summative)	11144	2548 ± 1	Percent 49% 25% 16% 10% Count 5.5K 2.8K 1.8K 1.1K	05/30/2025
Grade 11 ELA	Spring 2025 (Smarter Summative)	11117	2595 ± 1	Percent 21% 22% 32% 25% Count 2.3K 2.5K 3.5K 2.8K	05/30/2025
Grade 8 Math	Spring 2025 (Smarter Summative)	11781	2535 ± 1	Percent 41% 24% 17% 19% Count 4.9K 2.8K 2K 2.2K	05/29/2025
Grade 8 ELA	Spring 2025 (Smarter Summative)	11694	2567 ± 1	Percent 24% 22% 34% 19% Count 2.9K 2.6K 4K 2.3K	05/29/2025

Copyright © 2025. All rights reserved. | Terms of Use & Privacy Policy   Cambium Assessment

When authorized users at the complex area, complex, school, and teacher level log in to the CRS, the dashboard page shows the overall test results for all tests that the students have taken grouped by test family (i.e., 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. State personnel and complex area personnel would select a specific complex to view the aggregate results.

Exhibit 2 presents a sample dashboard page at the complex level.

Exhibit 2. Dashboard: Complex Level



Performance Distribution, By Test Group: Demo School Group 2, 2024-2025

Filtered By **Test Reasons:** All Test Reasons | **Sorted By:** Date Last Taken

Smarter Summative Mathematics

Grades Tested: **3, 4, 5, 6, 7, 8, 11**

Tests Taken: **5.3K**

Date Last Taken: **06/10/2025**

Percent	31%	27%	21%	21%
Count	1.6K	1.4K	1.1K	1.1K

Smarter Summative ELA

Grades Tested: **3, 4, 5, 6, 7, 8, 11**

Tests Taken: **5.3K**

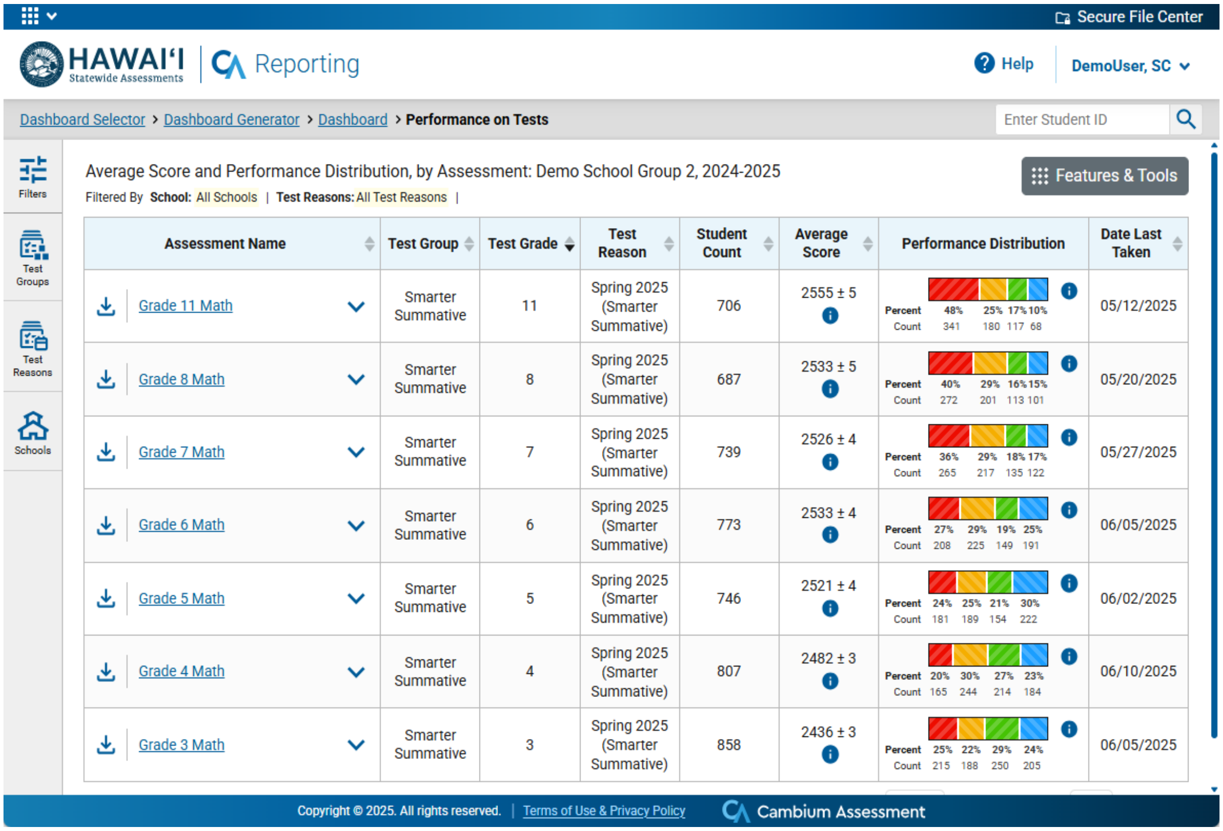
Date Last Taken: **06/09/2025**

Percent	22%	22%	30%	25%
Count	1.2K	1.2K	1.6K	1.3K

When a user clicks on a test family for further exploration, he or she will be taken to a detailed dashboard, where the results will be displayed 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) the number of students tested, (2) average score and standard error of the means, and (3) percentage and counts of students at each performance level.

Exhibit 3 presents a sample detailed dashboard page for Smarter Balanced summative mathematics at the complex level.

Exhibit 3. Detailed Dashboard: Complex Level



Dashboard Selector > Dashboard Generator > Dashboard > Performance on Tests

Average Score and Performance Distribution, by Assessment: Demo School Group 2, 2024-2025

Filtered By **School:** All Schools | **Test Reasons:** All Test Reasons |

Assessment Name	Test Group	Test Grade	Test Reason	Student Count	Average Score	Performance Distribution	Date Last Taken
Grade 11 Math	Smarter Summative	11	Spring 2025 (Smarter Summative)	706	2555 ± 5	Percent 48% 25% 17% 10% Count 341 180 117 68	05/12/2025
Grade 8 Math	Smarter Summative	8	Spring 2025 (Smarter Summative)	687	2533 ± 5	Percent 40% 29% 16% 15% Count 272 201 113 101	05/20/2025
Grade 7 Math	Smarter Summative	7	Spring 2025 (Smarter Summative)	739	2526 ± 4	Percent 36% 29% 18% 17% Count 265 217 135 122	05/27/2025
Grade 6 Math	Smarter Summative	6	Spring 2025 (Smarter Summative)	773	2533 ± 4	Percent 27% 29% 19% 25% Count 208 225 149 191	06/05/2025
Grade 5 Math	Smarter Summative	5	Spring 2025 (Smarter Summative)	746	2521 ± 4	Percent 24% 25% 21% 30% Count 181 189 154 222	06/02/2025
Grade 4 Math	Smarter Summative	4	Spring 2025 (Smarter Summative)	807	2482 ± 3	Percent 20% 30% 27% 23% Count 165 244 214 184	06/10/2025
Grade 3 Math	Smarter Summative	3	Spring 2025 (Smarter Summative)	858	2436 ± 3	Percent 25% 22% 29% 24% Count 215 188 250 205	06/05/2025

Copyright © 2025. All rights reserved. | Terms of Use & Privacy Policy Cambium Assessment

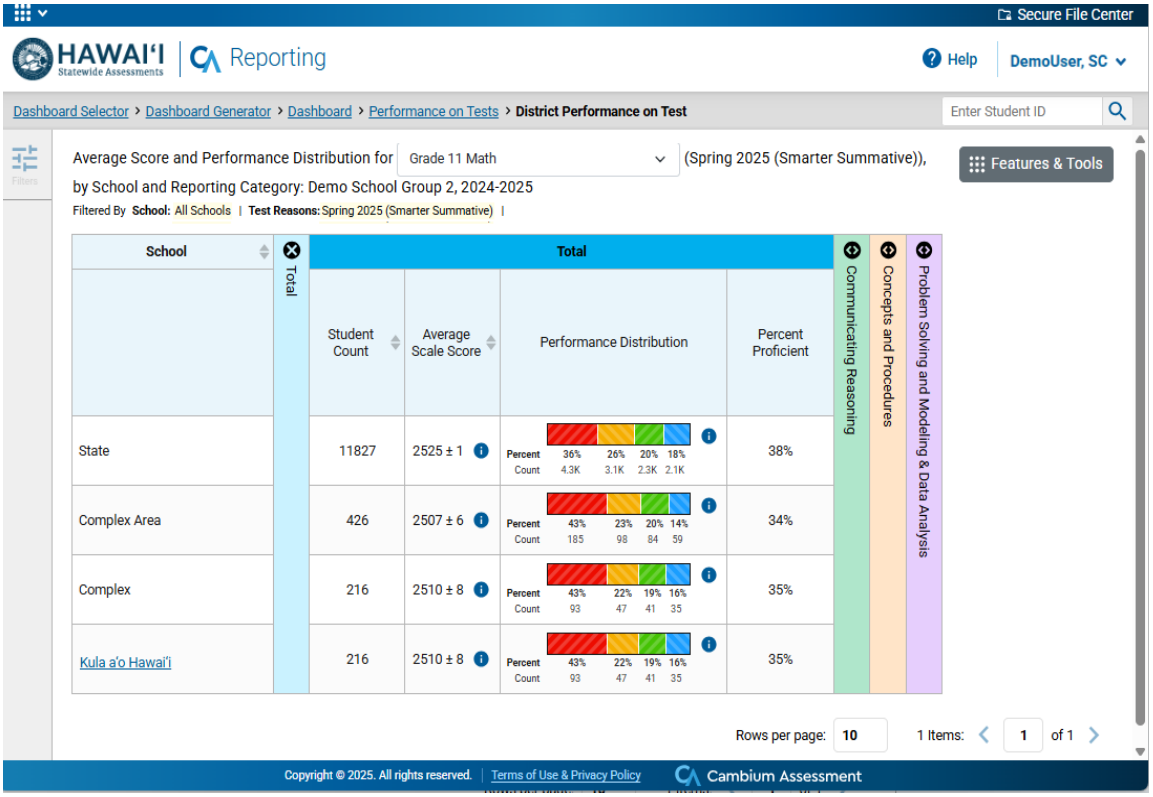
7.1.2 Aggregate Score Reports: Overall Performance

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 state and the aggregate unit both above and below the selected aggregate. For example, if a complex is selected, the summary results of the state and individual schools within the complex are provided as well as the complex summary results so that complex 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) the 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 4 presents a sample overall performance summary results page for grade 11 mathematics at the complex level, and Exhibit 5 presents an example summary for grade 11 mathematics by gender.

Exhibit 4. Overall Performance Summary Results for Grade 11 Mathematics: Complex Level



Dashboard Selector › Dashboard Generator › Dashboard › Performance on Tests › **District Performance on Test**

Average Score and Performance Distribution for Grade 11 Math (Spring 2025 (Smarter Summative)), by School and Reporting Category: Demo School Group 2, 2024-2025

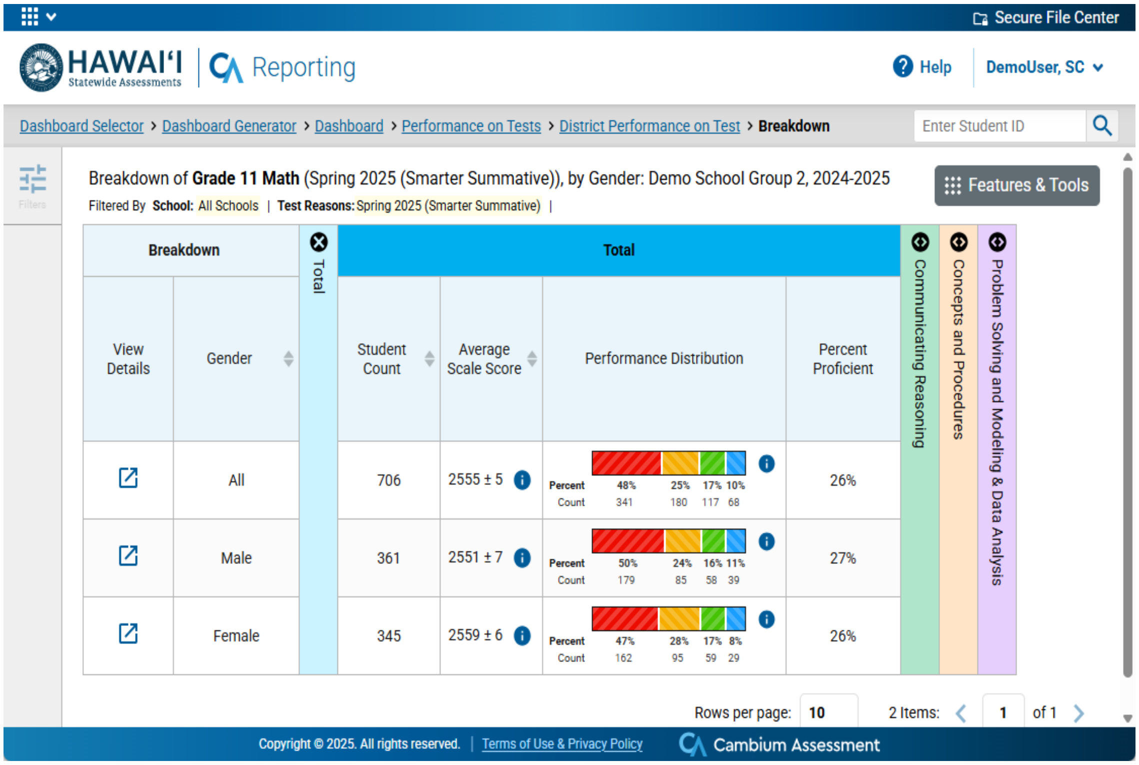
Filtered By **School:** All Schools | **Test Reasons:** Spring 2025 (Smarter Summative) |

School	Total	Student Count	Average Scale Score	Performance Distribution	Percent Proficient	Communicating Reasoning	Concepts and Procedures	Problem Solving and Modeling & Data Analysis
State		11827	2525 ± 1	Percent 36% 26% 20% 18% Count 4.3K 3.1K 2.3K 2.1K	38%			
Complex Area		426	2507 ± 6	Percent 43% 23% 20% 14% Count 185 98 84 59	34%			
Complex		216	2510 ± 8	Percent 43% 22% 19% 16% Count 93 47 41 35	35%			
Kula a'o Hawai'i		216	2510 ± 8	Percent 43% 22% 19% 16% Count 93 47 41 35	35%			

Rows per page: 10 1 Items: ‹ 1 of 1 ›

Copyright © 2025. All rights reserved. | Terms of Use & Privacy Policy Cambium Assessment

Exhibit 5. Overall Performance Summary Results for Grade 11 Mathematics by Gender: Complex Level



Dashboard Selector › Dashboard Generator › Dashboard › Performance on Tests › District Performance on Test › **Breakdown**

Breakdown of **Grade 11 Math** (Spring 2025 (Smarter Summative)), by Gender: Demo School Group 2, 2024-2025

Filtered By **School:** All Schools | **Test Reasons:** Spring 2025 (Smarter Summative) |

Breakdown		Total	Total				Communicating Reasoning	Concepts and Procedures	Problem Solving and Modeling & Data Analysis
View Details	Gender		Student Count	Average Scale Score	Performance Distribution	Percent Proficient			
	All		706	2555 ± 5	Percent 48% 25% 17% 10% Count 341 180 117 68	26%			
	Male		361	2551 ± 7	Percent 50% 24% 16% 11% Count 179 85 58 39	27%			
	Female		345	2559 ± 6	Percent 47% 28% 17% 8% Count 162 95 59 29	26%			

Rows per page: 10 2 Items: ‹ 1 of 1 ›

Copyright © 2025. All rights reserved. | Terms of Use & Privacy Policy Cambium Assessment

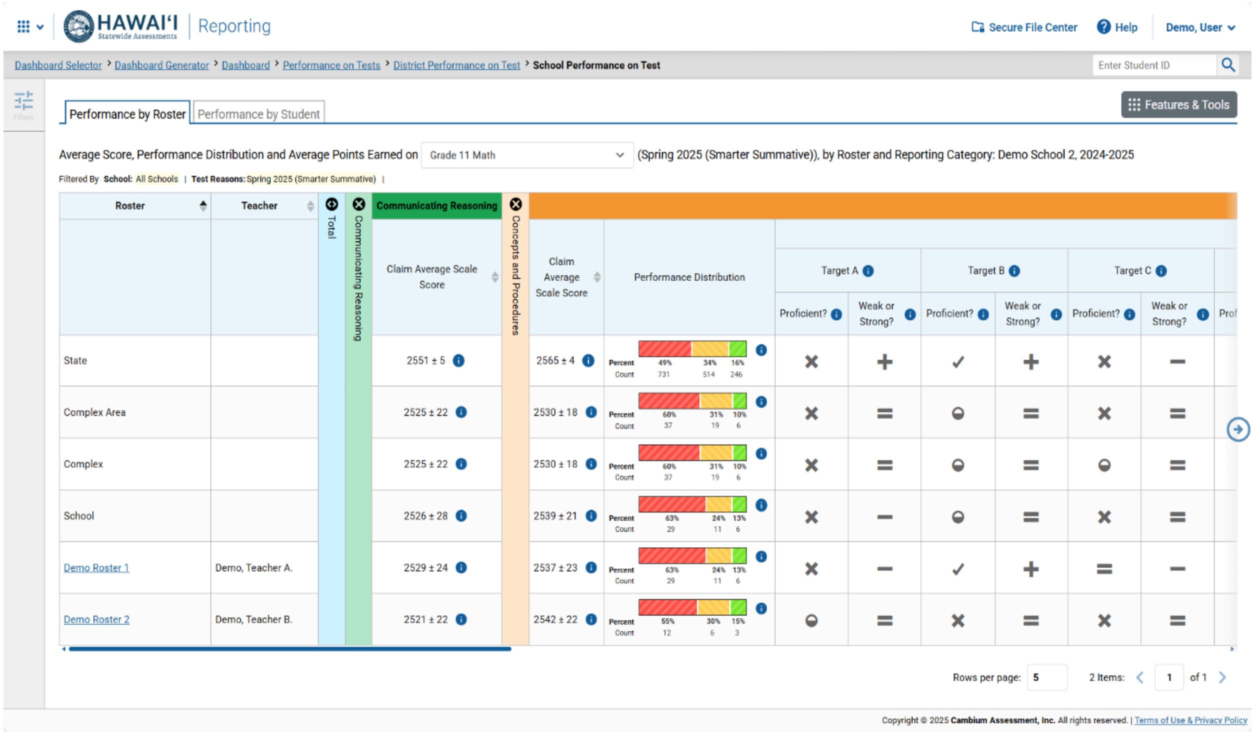
7.1.3 Aggregate Score Reports: Claim and Target Performance

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, both the average scale score and standard error of the average scale score 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 (checkmark), 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, for the state, and for 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 6 presents a sample claim and target-level results page for grade 11 mathematics at the complex level.

Exhibit 6. Claim and Target Level Results for Grade 11 Mathematics: Complex Level



Dashboard Selector › Dashboard Generator › Dashboard › Performance on Tests › District Performance on Test › School Performance on Test

Average Score, Performance Distribution and Average Points Earned on Grade 11 Math (Spring 2025 (Smarter Summative)), by Roster and Reporting Category: Demo School 2, 2024-2025

Filtered By School: All Schools | Test Reasons: Spring 2025 (Smarter Summative) |

Roster	Teacher	Total	Communicating Reasoning	Communicating Reasoning	Concepts and Procedures			Target A		Target B		Target C		
				Claim Average Scale Score		Claim Average Scale Score	Performance Distribution	Proficient?	Weak or Strong?	Proficient?	Weak or Strong?	Proficient?	Weak or Strong?	Prof
State				2551 ± 5		2565 ± 4	Percent 49% 34% 16% Count 731 514 246	✕	+	✓	+	✕	−	
Complex Area				2525 ± 22		2530 ± 18	Percent 60% 31% 10% Count 37 19 6	✕	=	◒	=	✕	=	
Complex				2525 ± 22		2530 ± 18	Percent 60% 31% 10% Count 37 19 6	✕	=	◒	=	◒	=	
School				2526 ± 28		2539 ± 21	Percent 63% 24% 13% Count 29 11 6	✕	−	◒	=	✕	=	
Demo Roster 1	Demo, Teacher A.			2529 ± 24		2537 ± 23	Percent 63% 24% 13% Count 29 11 6	✕	−	✓	+	=	−	
Demo Roster 2	Demo, Teacher B.			2521 ± 22		2542 ± 22	Percent 55% 30% 15% Count 12 6 3	◒	=	✕	=	✕	=	

Rows per page: 5   2 Items: ‹ 1 of 1 ›

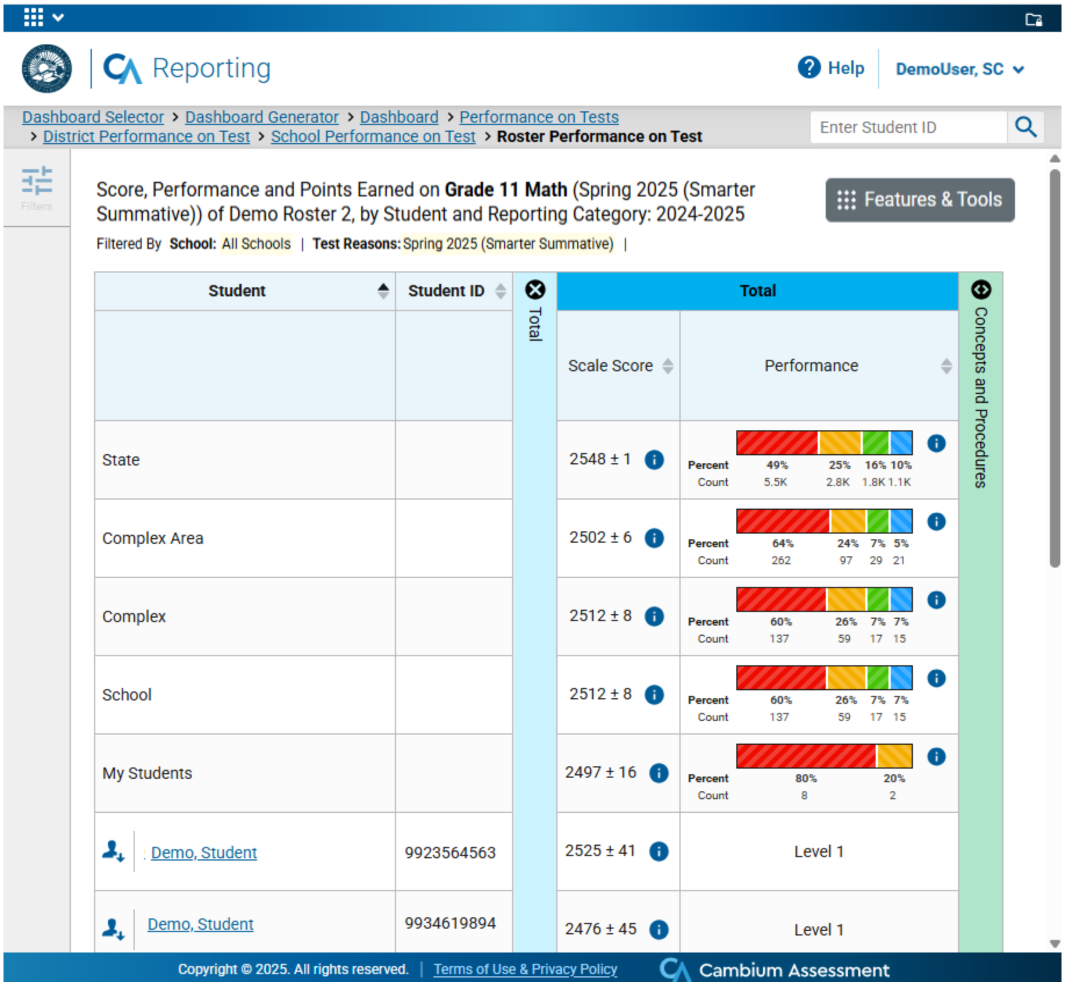
Copyright © 2025 Cambium Assessment, Inc. All rights reserved. | Terms of Use & Privacy Policy

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, and (2) the performance level.

Exhibit 7 shows a sample roster performance report page for the grade 11 mathematics summative assessment.

Exhibit 7. Roster Performance Report for Grade 11 Mathematics


Reporting
Help
DemoUser, SC
Dashboard Selector > Dashboard Generator > Dashboard > Performance on Tests > District Performance on Test > School Performance on Test > Roster Performance on Test
Enter Student ID
Filters
Features & Tools
Concepts and Procedures

Score, Performance and Points Earned on **Grade 11 Math** (Spring 2025 (Smarter Summative)) of Demo Roster 2, by Student and Reporting Category: 2024-2025

Filtered By **School:** All Schools | **Test Reasons:** Spring 2025 (Smarter Summative) |

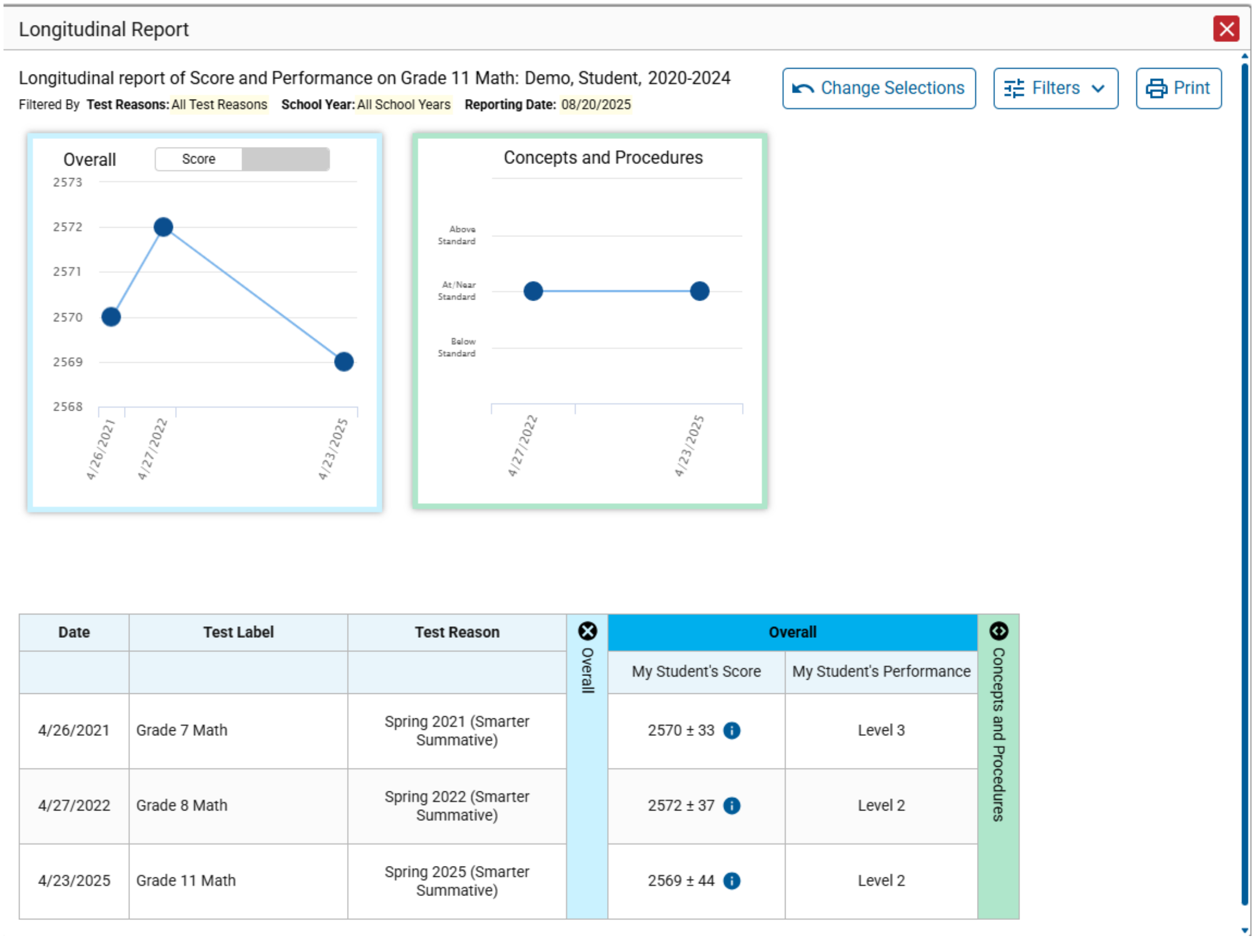
Student	Student ID	Total Scale Score	Total Performance
State		2548 ± 1	Percent 49% 25% 16% 10% Count 5.5K 2.8K 1.8K 1.1K
Complex Area		2502 ± 6	Percent 64% 24% 7% 5% Count 262 97 29 21
Complex		2512 ± 8	Percent 60% 26% 7% 7% Count 137 59 17 15
School		2512 ± 8	Percent 60% 26% 7% 7% Count 137 59 17 15
My Students		2497 ± 16	Percent 80% 20% Count 8 2
Demo, Student	9923564563	2525 ± 41	Level 1
Demo, Student	9934619894	2476 ± 45	Level 1

Copyright © 2025. All rights reserved. | Terms of Use & Privacy Policy | Cambium Assessment

7.1.5 Trend Report

The trend (i.e., longitudinal) page provides the trend of student performance for individual level and aggregate level 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 8 presents an example trend report page for mathematics at the individual student level.

Exhibit 8. Trend Report for Mathematics: Student Level



Longitudinal Report

Longitudinal report of Score and Performance on Grade 11 Math: Demo, Student, 2020-2024

Filtered By **Test Reasons:** All Test Reasons **School Year:** All School Years **Reporting Date:** 08/20/2025

Date	Test Label	Test Reason	Overall	Overall		Concepts and Procedures
				My Student's Score	My Student's Performance	
4/26/2021	Grade 7 Math	Spring 2021 (Smarter Summative)		2570 ± 33	Level 3	
4/27/2022	Grade 8 Math	Spring 2022 (Smarter Summative)		2572 ± 37	Level 2	
4/23/2025	Grade 11 Math	Spring 2025 (Smarter Summative)		2569 ± 44	Level 2	

7.1.6 Individual Student Report

An individual student report (ISR) can be generated and exported as a PDF. The ISR shows the student's overall performance on the test with detailed information on multiple pages. In each subject area, the ISR provides (1) the scale score and SEM; (2) achievement level for the overall test; (3) average scale scores for student's state, complex area, complex, and school; and (4) writing performance descriptors in each dimension (ELA/L only).

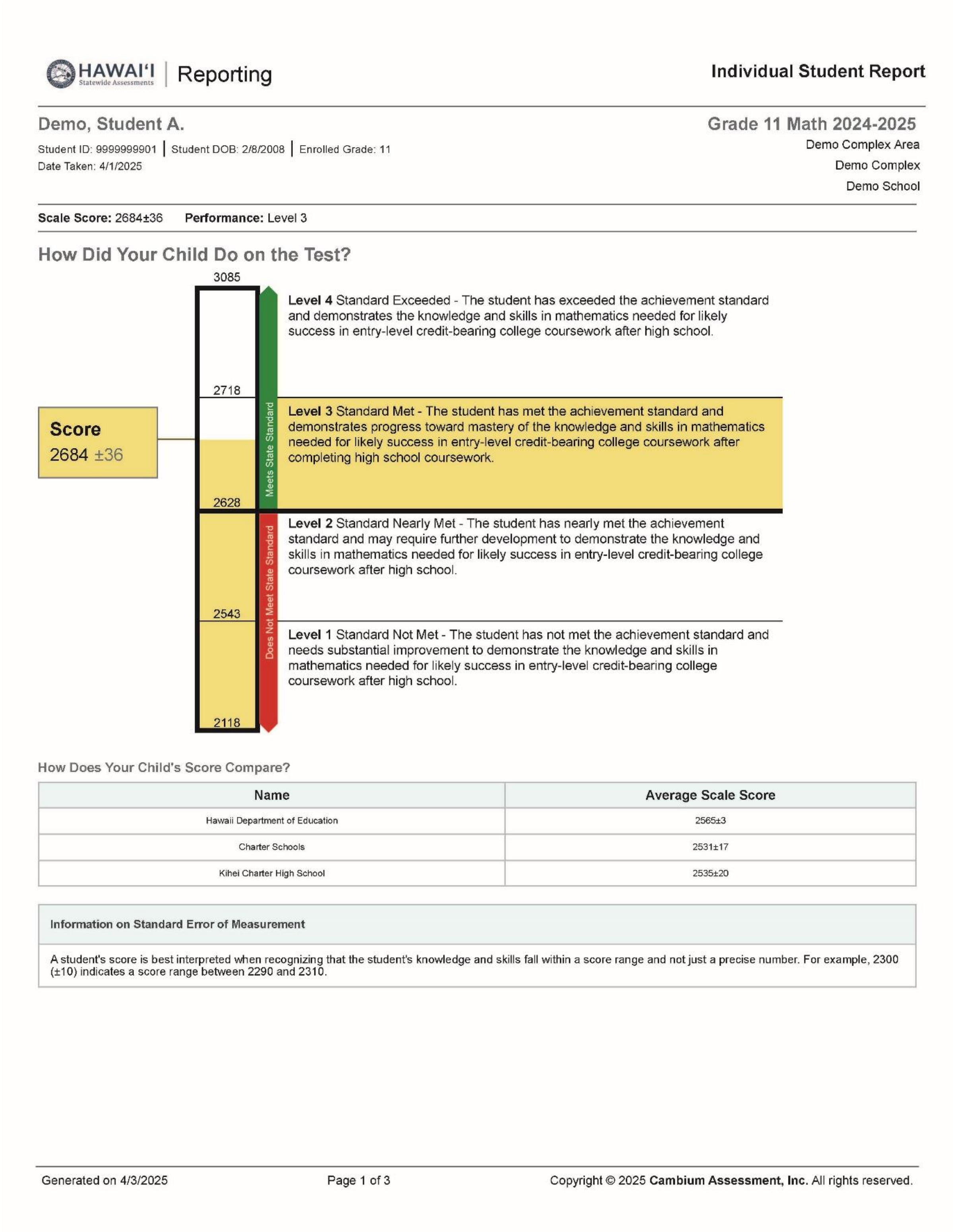
On the first page of the ISR, the student's name, scale score with the SEM, and achievement level for ELA are shown at the top of the page. In the middle section, the student's performance is described in detail using a barrel chart. In the barrel chart, the student's scale score is presented with the SEM using a "±" sign. The 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. These define the content-area knowledge, skills, and processes that test takers at the achievement level are expected to possess.

Average scale scores and standard errors of the average scale scores for the student's state, complex area, complex, and school are displayed at the bottom of the page 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 standard error of measurement of the scale score, whereas the “±” next to the average scale scores for aggregate levels represents the standard error of the average scale scores.

The second page shows the student’s performance on claims (i.e., Claims 1 and 2 for ELA and Claim 1 only for mathematics) which is displayed alongside a description of his or her performance on the claim. At the bottom of the page, the student’s performance on the different writing dimensions is displayed alongside a detailed description. The last page provides the trend of the student’s performance over time. Student scale scores and achievement levels over time are graphed, showing how the student’s scale scores changed over time and whether the student met the standards each year.

Exhibit 9 presents a sample ISR for grade 11 mathematics.

Exhibit 9. Individual Student Report for Grade 11 Mathematics



HAWAI'I Statewide Assessments | Reporting

Individual Student Report

Demo, Student A.

Student ID: 9999999901 | Student DOB: 2/8/2008 | Enrolled Grade: 11
Date Taken: 4/1/2025

Grade 11 Math 2024-2025
Demo Complex Area
Demo Complex
Demo School

Scale Score: 2684±36 **Performance:** Level 3

How Did Your Child Do on the Test?

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

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

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

Level 1 Standard Not Met - The student has not met the achievement standard and needs substantial improvement to demonstrate the knowledge and skills in mathematics needed for likely success in entry-level credit-bearing college coursework after high school.

How Does Your Child's Score Compare?

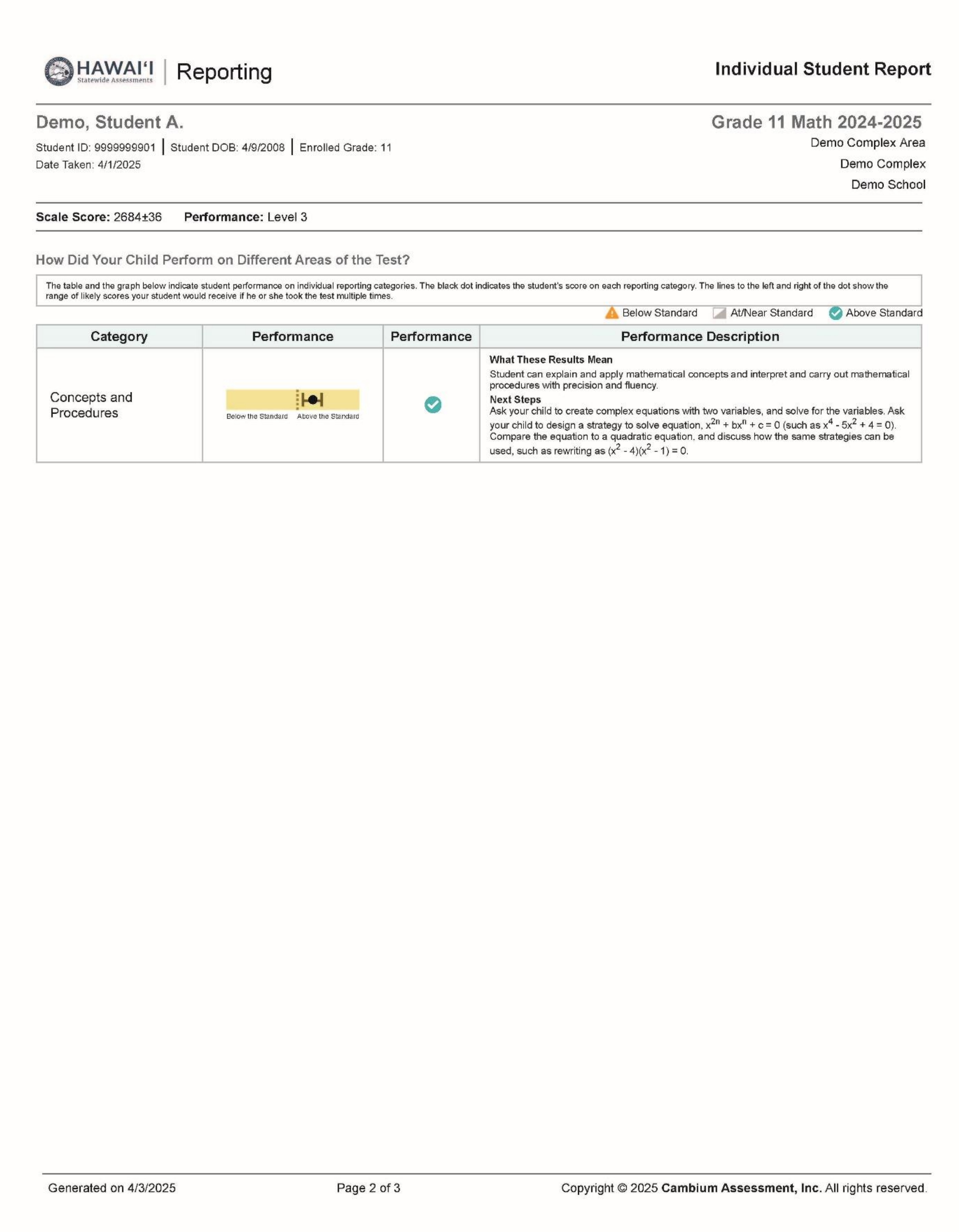
Name	Average Scale Score
Hawaii Department of Education	2565±3
Charter Schools	2531±17
Kihei Charter High School	2535±20

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.

Generated on 4/3/2025 Page 1 of 3 Copyright © 2025 **Cambium Assessment, Inc.** All rights reserved.

Exhibit 9. Individual Student Report for Grade 11 Mathematics (Continued)



HAWAIʻI Statewide Assessments | Reporting

Individual Student Report

Demo, Student A.

Student ID: 9999999901 | Student DOB: 4/9/2008 | Enrolled Grade: 11
Date Taken: 4/1/2025

Grade 11 Math 2024-2025
Demo Complex Area
Demo Complex
Demo School

Scale Score: 2684±36 **Performance:** Level 3

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

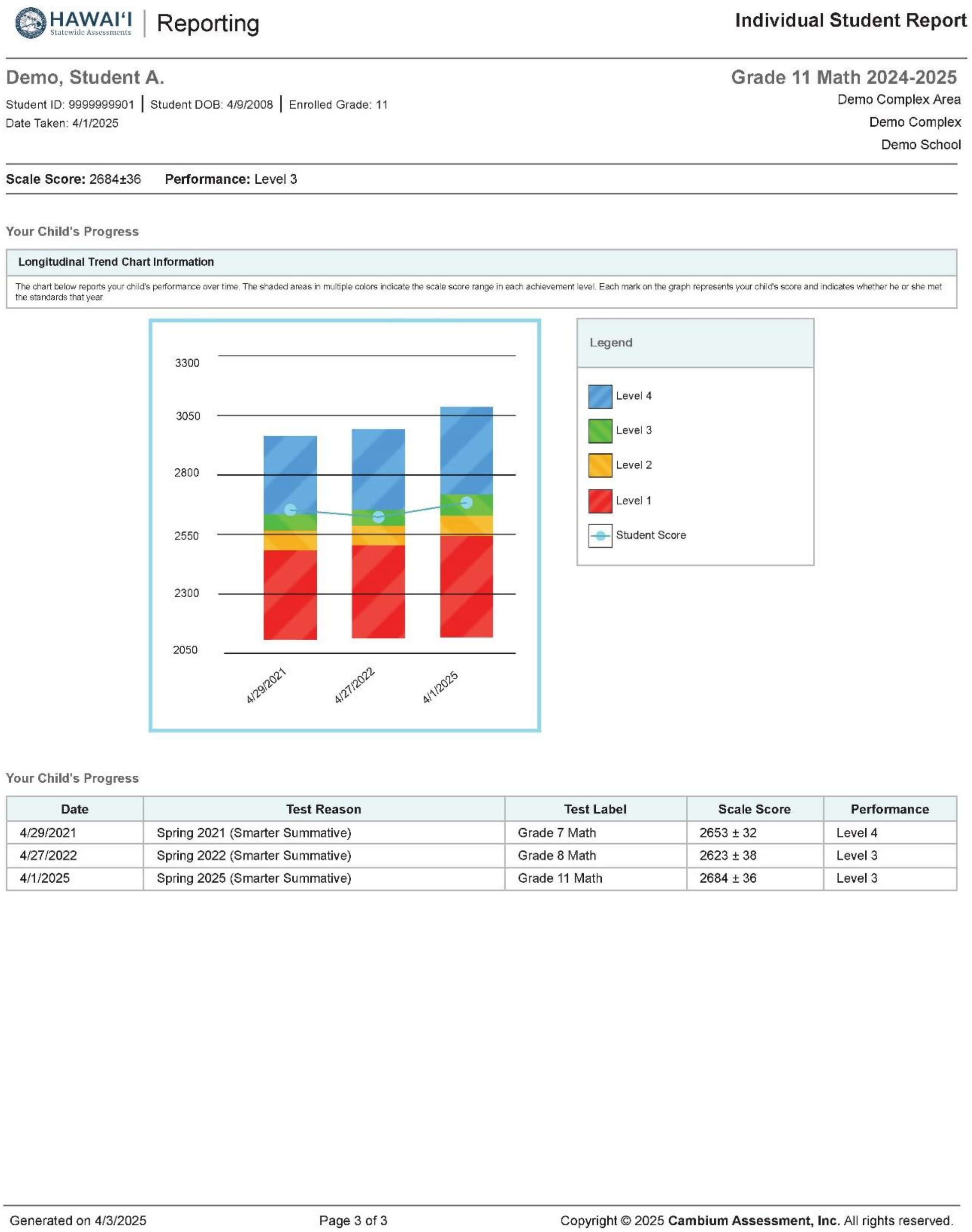
The table and the graph below indicate student performance on individual reporting categories. The black dot indicates the student's score on each reporting category. The lines to the left and right of the dot show the range of likely scores your student would receive if he or she took the test multiple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

⚠ Below Standard ▨ At/Near Standard ✓ Above Standard

Category	Performance	Performance	Performance Description
Concepts and Procedures	Below the Standard Above the Standard	✓	**What These Results Mean** Student can explain and apply mathematical concepts and interpret and carry out mathematical procedures with precision and fluency. **Next Steps** Ask your child to create complex equations with two variables, and solve for the variables. Ask your child to design a strategy to solve equation, $x^{2n} + bx^n + c = 0$ (such as $x^4 - 5x^2 + 4 = 0$). Compare the equation to a quadratic equation, and discuss how the same strategies can be used, such as rewriting as $(x^2 - 4)(x^2 - 1) = 0$.

Generated on 4/3/2025 Page 2 of 3 Copyright © 2025 **Cambium Assessment, Inc.** All rights reserved.

Exhibit 9. Individual Student Report for Grade 11 Mathematics (Continued)



HAWAIʻI Statewide Assessments | Reporting

Individual Student Report

Demo, Student A.

Student ID: 9999999901 | Student DOB: 4/9/2008 | Enrolled Grade: 11
Date Taken: 4/1/2025

Grade 11 Math 2024-2025
Demo Complex Area
Demo Complex
Demo School

Scale Score: 2684±36 **Performance:** Level 3

Your Child's Progress

Longitudinal Trend Chart Information

The chart below reports your child's performance over time. The shaded areas in multiple colors indicate the scale score range in each achievement level. Each mark on the graph represents your child's score and indicates whether he or she met the standards that year.

Legend

- 🟦 Level 4
- 🟩 Level 3
- 🟨 Level 2
- 🟥 Level 1
- ● Student Score

Your Child's Progress

Date	Test Reason	Test Label	Scale Score	Performance
4/29/2021	Spring 2021 (Smarter Summative)	Grade 7 Math	2653 ± 32	Level 4
4/27/2022	Spring 2022 (Smarter Summative)	Grade 8 Math	2623 ± 38	Level 3
4/1/2025	Spring 2025 (Smarter Summative)	Grade 11 Math	2684 ± 36	Level 3

Generated on 4/3/2025 Page 3 of 3 Copyright © 2025 **Cambium Assessment, Inc.** All rights reserved.

7.2 INTERPRETATION OF REPORTED SCORES

A student's performance on a test is reported as a scale score and an achievement level for the overall test. Students' scores and achievement levels are also summarized at the aggregate levels. The next section provides a description of 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. The interpretation of scale scores is more meaningful when the scale scores are used along with achievement levels and achievement-level descriptors.

7.2.2 Standard Error of Measurement

A scale score (observed score on any test) is an estimate of the true score. If a student takes a similar test multiple times, the resulting scale score will vary across administrations, sometimes being a little higher, a little lower, or the same. The standard error of measurement (SEM) represents the precision of the scale score, or the range in which the student would likely score if a similar test was administered 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 was tested again, it is likely that the student would receive a score between 2670 and 2690. The SEM can be different for the same scale score, depending on how closely the administered items match the student's ability.

7.2.3 Achievement Level

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

7.2.4 Performance Category for Claims

Students' performance on each claim is reported in three categories: (1) Below Standard, (2) At/Near Standard, and (3) Above Standard. Unlike the achievement level for 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 purposes. 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 complex and provide information about how a group of students in a class, school, or complex performed on each target, either relative to the proficiency standard (i.e., "Proficient?" target measure) or relative to their overall performance on the test (i.e., "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 score (i.e., the Achievement Level 3 cut). At the aggregate level, when the observed performance within a target is greater than the proficiency cut, the reporting unit shows 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 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 the observed performance within a target is greater than the expected performance, the reporting unit (e.g., roster, teacher, school, complex) shows relative strength in that target. Conversely, when the observed performance within a target is below the level expected based on overall achievement, the reporting unit shows 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, complex, complex area, and state levels to represent how a group of students performs 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 possesses. 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 for overall are reported at the aggregate level to represent how well a group of students performs.

7.3 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. Further, by narrowing 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 limited English proficiency (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, complex, and complex area 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' academic progress, 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 PROCEDURES

Quality assurance (QA) procedures are enforced throughout all stages of the Smarter Balanced assessment development, administration, scoring, and reporting of results. CAI uses a series of quality control (QC) 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 testing (CAT) component, a test configuration file is the key resource that contains all specifications for the item-selection algorithm and the scoring algorithm, such as the test blueprint, cut scores, 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.

CAI uses simulated test administrations along with the test configuration file to configure the adaptive algorithm in order 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 the 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 rules specified in scoring specifications were 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, view the same item to ensure 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 HIDOE with an opportunity to interact with the exact test that the students will use.

8.2 Quality Assurance in Document Processing

The Smarter Balanced assessments are administered primarily online; however, a few students take paper-pencil assessments. When test documents are scanned, a QC sample of documents consisting of 10 test cases per document type (normally between 500 and 600 documents) is created so that all possible responses and all demographic grids are verified, including various typical errors that required editing via Measurement Incorporated's (MI) Data Inspection, Correction, and Entry (DICE) application. This structured testing method provides exact test parameters and a methodical way of determining that the output received from the scanner(s) is correct. MI staff carefully compare 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 are correct.

8.3 Quality Assurance in Data Preparation

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

Data pass directly from the Quality Monitor System (QM) to the Database of Record (DOR), which serves as the repository for all test information from which all test information for reporting is pulled. The Data Extract Generator is the tool that is used to pull data from the DOR for delivery to HIDOE. CAI staff ensure that data in the extract files match the DOR before it is delivered.

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 testing 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 arise. The applications log not only errors and exceptions, but also latency (timing) information for crucial 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, latency data, such as data about how long it takes to load, view, or respond to an item, are captured for each assessed student. All this information is logged, enabling CAI to automatically identify schools or complex areas experiencing unusual slowdowns, often before they even notice.

A series of quality assurance reports, such as blueprint match rate, item exposure rate, and item statistics, can also be generated at any time during the online assessment window for the 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 Forensics Program.

For example, an item statistics analysis report allows psychometricians to ensure that items are performing as intended and serves as an empirical key check throughout 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 the incorrect designation of a keyed response or other scoring errors and 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 component, other reports, such as blueprint match and item exposure reports, allow psychometricians to verify that test administrations conform to the simulation results. The QA reports can be generated on any desired schedule. Item analysis and blueprint match reports are evaluated frequently at the opening of the testing window to ensure that test administrations conform to the blueprint and that items are performing as anticipated.

Table 77 presents an overview of the QA reports.

Table 77. 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 (highly 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

In the Smarter Balanced summative assessments, online score reports are generated. The system automatically assigns scores for the online assessments 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 central 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 these processes take milliseconds to complete, with CAI receiving handscores and passing them through QA validation checks in less than one second and making the composite score available in the CRS immediately.

REFERENCES

American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME]. (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.

Raczynski, K. R., Cohen, A. S., Engelhard, G., Jr., & Lu, Z. (2015). Comparing the effectiveness of self-paced and collaborative frame-of-reference training on rater accuracy in a large-scale writing assessment. *Journal of Educational Measurement*, *52*(3), 301–318.

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, DC. Retrieved from https://www2.ed.gov/policy/elsec/guid/assessguid15.pdf

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for the evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, *31*(1), 2–13.