Hawai‘i Smarter Balanced Assessments

2021–2022 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

LIST OF APPENDICES

1. OVERVIEW

This report provides a technical summary of Hawai‘i's 2021–2022 administration of the Smarter Balanced summative assessments in English language arts/literacy (ELA/L) and mathematics in grades 3–8 and 11. This report includes nine chapters, including: Overview, Test Administration, Comparability of the Shortened and Full Blueprints, Summary of the 2021–2022 Operational Test Administration, Validity, Reliability, Scoring, Reporting and Interpreting Scores, and Quality Control Procedures. For the interim assessments, the number of students who took the Interim Comprehensive Assessments (ICAs) and the Interim Assessment Blocks (IABs) and their performance are provided in Appendix A, Summary of the 2021–2022 Interim Assessments. The data included in this report are based on Hawai‘i's data for the Smarter Balanced assessments in ELA/L and mathematics.

While this report includes information on all aspects of the technical quality of the Smarter Balanced test administration in Hawai‘i, the information on item and test development, item content review, field-test administration, item data review, item calibrations, content-alignment study, standard setting, and other validity information can be found in the overall Smarter Balanced technical report. The Smarter Balanced technical report includes all aspects of the technical qualities of 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).

1.1 SMARTER BALANCED ASSESSMENTS IN HAWAI‘I

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. Hawai‘i is 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 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.

The Hawai'i State Board of Education formally adopted the CCSS in ELA/L and mathematics on June 18, 2010. All students in Hawai'i, including students with significant cognitive disabilities who are eligible to take the Hawai'i State Alternate Assessment (an alternate assessment based on Alternate Academic Achievement Standards), are taught the same academic content standards. The Hawai'i 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 Hawai'i 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 Hawai'i statewide assessments in ELA/L and mathematics aligned with the CCSS were administered for the first time in spring 2015 to students in grades 3–8 and 11 in all public elementary and secondary schools.

The American Institutes for Research (AIR) delivered 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.

In the 2019–2020 school year, the U.S. Department of Education (USED) granted a waiver from testing requirements due to the COVID-19 pandemic (https://www2.ed.gov/policy/gen/guid/secletter/200320.html). In the 2020–2021 school year, ED did not grant waivers for standardized testing but did waive certain accountability requirements (e.g., mandatory high participation rates) due to the impact of the pandemic in many states, resulting in lower participation rates than in previous years.

In the 2021–2022 school year, the overall participation rates increased, ranging from 92.8%–94.7% in grades 3–8 and 87.5%–88.7% in grade 11; 1%–3% in grades 3–8; and 6% in grade 11, which are lower than the 2018–2019 participation rates.

Starting with the 2020–2021 Smarter Balanced summative test administration, Hawai'i shortened the full test blueprints for ELA/L and mathematics and allowed school districts to administer remote test administration. The rationale for implementing the short blueprints is provided in Section 1.2, Hawaiʻi's Shortened Blueprint Rationale. The technical qualities of the shortened blueprint are presented in Chapter 3, Comparability of the Shortened and Full Blueprints. The technical information of the full blueprint is shown in Chapter 3 only. The remaining chapters include information on the shortened blueprints, implemented in 2021–2022.

1.2 HAWAIʻI'S SHORTENED BLUEPRINT RATIONALE

1.2.1 Stakeholder Interest and Motivation to Shorten Test Length

A statewide survey was conducted by Ward Research in January 2016 to gather feedback on education issues in Hawaiʻi. The survey revealed that 44% of respondents felt there was an excessive emphasis on standardized testing in the state. Additionally, in July 2016, a survey of principals conducted by Ward Research found that 84% of respondents believed the Hawaiʻi Department of Education (HIDOE) should consider changes to the Smarter Balanced Assessment (SBA), and 85% agreed that testing time should be reduced.

In April 2016, Governor David Ige of Hawaiʻi convened a task force to create a blueprint for the state's public schools that aligned with the Every Student Succeeds Act (ESSA) and offered opportunities for educational transformation. The task force included the Governor Ige, State Board of Education members, State Department of Education leadership, and members of the governor's ESSA team. The collaborative planning framework developed by the task force included the expectation that educational assessments would be designed to efficiently assess student learning and minimize testing time.

1.2.2 HIDOE Research and Consideration of Models

In its efforts to explore new ways to evaluate student learning, HIDOE engaged in research and collaboration with its Technical Advisory Committee (TAC) to explore potential alternate approaches. Stakeholder meetings were held in 2019 to prepare for the development of an application for the Federal Innovative Assessment Demonstration Authority (IADA). HIDOE also contracted with the Center for Assessment to assist in creating an IADA model. It was determined that the model would include a shortened summative assessment that met all the requirements of ESSA Section 1111(b)(2)(B).

To demonstrate its commitment, HIDOE pledged that its shortened summative assessments administered for accountability purposes would

- meet the technical quality sufficient for each purpose required under and consistent with the provisions of the Every Student Succeeds Act;
- result in an overall scale score and proficiency level for each student;
- be aligned to the state-adopted content standards, provide coherent and timely information about student attainment of such standards, and measure the breadth and depth of Hawaiʻi state-adopted content standards;
- be valid and reliable, consistent with relevant, nationally recognized professional and technical testing standards; objectively measure academic achievement, knowledge, and skills; and will not evaluate or assess personal or family beliefs and attitudes, or publicly disclose personally identifiable information (PII);

- appropriately provide universal tools, designated supports, and accommodations (as verified) for students with disabilities under the Individuals with Disability Education Improvement Act (IDEA) and Section 504 of the Rehabilitation Act of 1973, including English language learners (ELLs) with disabilities, to measure their academic achievement;
- provide family reports (paper) to parents and provide access to online reports to teachers, principals, and other school leaders as soon as practicable after the CAT is administered, scored and quality checked; and
- support reporting overall scores by school and statewide for subgroups, as appropriate, as required by the ESSA.

HIDOE's 2020 IADA application was not approved by the U.S. Department of Education due to uncertainties associated with the use of classroom-based assessments administered throughout the school year.

1.2.3 HIDOE's Pursuit of Flexibilities in Assessment During the Pandemic

As a result of the COVID-19 Pandemic, USED provided flexibilities for statewide assessments in 2021. HIDOE engaged in discussions with its TAC and decided to pursue flexibility in the length of its summative assessments. The proposed shortened summative blueprints that were part of HIDOE's IADA model were adopted for the 2021 administration of the statewide assessments.

1.2.4 HIDOE's Adopted Shortened Smarter Balanced Summative Blueprints

Following the analysis of the 2021 administration of the shortened summative blueprints, HIDOE concluded that it would be feasible to report subcategory results for both ELA/literacy and mathematics at the individual level. After consulting with its TAC, HIDOE decided to proceed with defending its shortened summative blueprints for peer review by the USED. As a result, the same abbreviated blueprints were administered in SY 2021–2022.

While the full version Smarter Balanced Assessments (SBA) are an important tool to measure student progress and guide instruction, lengthy and stressful testing experiences can create unnecessary anxiety for students and may not accurately measure student learning. Therefore, a shortened blueprint for the SBA that focuses on reducing testing time and student testing anxiety is beneficial for several reasons.

First, it frees up valuable instructional time for teachers and students. Long testing periods can disrupt the regular school routine and cause students to fall behind on coursework, potentially missing out on valuable learning opportunities. By shortening the blueprint, schools can ensure that assessments are not taking away from important classroom instruction and learning experiences.

Second, a shorter testing period can help reduce student testing anxiety. Students may experience anxiety due to the length of the test, pressure to perform well, or fear of the unknown. By creating a shorter testing period, students may feel less overwhelmed and anxious, leading to a more positive testing experience and potentially more accurate results.

Finally, a shortened version of the SBA blueprint can still effectively measure student learning and offer useful insights into their progress. This can be achieved by eliminating questions or tasks that require a long time to answer but do not contribute to the testing experience, thereby significantly reducing testing time. Moreover, reducing the number of CAT questions in proportion to the full blueprint can still yield a valid and reliable measure of overall proficiency. In fact, removing time-consuming items and decreasing the number of test questions may provide a more precise measure of student learning, as it enables them to demonstrate mastery of essential concepts without being overwhelmed by a lengthy test.

In conclusion, a shortened blueprint for the SBA that focuses on reducing testing time and student testing anxiety is a valuable solution to several common problems associated with standardized testing. By freeing up instructional time, reducing student anxiety, and providing an accurate measure of student learning, a targeted and efficient SBA can benefit students, teachers, and schools alike.

1.3 Changes in the Summative Test Blueprints

Starting with the 2020–2021 summative assessment, Smarter Balanced offered member states a new adjusted blueprint for the summative assessments in ELA/L and mathematics. The adjusted blueprint was designed to meet their assessment needs while addressing the challenges created by the COVID-19 pandemic. In the adjusted blueprint, the CAT portion of the blueprint was reduced by approximately 50% of the test’s length, but the blueprints associated with the PTs were not adjusted.

Similar to Smarter Balanced, Hawai‘i also shortened the CAT blueprints for ELA/L and mathematics. Hawai‘i’s shortened blueprints were almost identical to the Smarter Balanced adjusted blueprint, except for removing the PTs in mathematics. In mathematics, Hawai‘i removed the PTs to reduce the testing time given that the targets covered in PTs were also covered in the CAT portion of the blueprint. For the Hawai‘i shortened blueprint, students received an overall scale score and an overall achievement level only in 2020–2021 but claim performance categories for Claims 1 and 2 in ELA/L and Claim 1 in mathematics were also provided in 2021–2022. The shortened blueprint allowed Hawai‘i to assess students’ progress with acceptable test reliability while significantly reducing testing time.

The impact of the Hawai‘i shortened blueprint is provided in Chapter 3, Comparability of the Shortened and Full Blueprints.

1.4 Remote Testing

Starting with the 2020–2021 testing cycle, HIDOE allowed remote test administration, which was intended as an option for parents who declined to have their child tested in person on a school campus but still wished for their students to take the assessment, and who could provide and agree to all requirements for remote test administration.

In the 2021–2022 test administration, a total of 76 students in ELA/L and 81 students in mathematics took the summative tests remotely.

2. TEST ADMINISTRATION

2.1 TESTING WINDOWS

The 2021–2022 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. 2021–2022 Testing Windows

Tests	Grade	Start Date	End Date	Mode
Summative Assessments	3–8	2/22/2022	5/27/2022	Online Adaptive
		3/14/2022 (Multi-track)	6/17/2022 (Multi-track)	
	11	2/22/2022	5/27/2022	Online Adaptive
		11/21/2021 (Block Scheduled)	5/27/2022 (Block Scheduled)	
	3–8, 11	2/22/2022	5/13/2022	Paper Fixed-Form
	3–8, 11	2/22/2022	6/17/2022	Remote Online Adaptive
	3–8, 11	2/22/2022	5/13/2022	Braille Paper Fixed-Form
Interim Comprehensive Assessments	3–8, 11	8/17/2021	7/22/2022	Online Fixed-Form
Interim Assessment Blocks	3–8, 11	8/17/2021	7/22/2022	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 2021–2022. A testing option is selected by content area. Once an option is selected, it applied to all tests in the content area.

Table 2. 2021–2022 Testing Options

Assessments	Testing 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/resources/resources-2021-2022/smarter-balanced-summative-test-administration-manual-2021-2022.

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 Hawai‘i 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/resources/resources-2021-2022/remote-summative-test-administration-2021-2022.

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 of 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. Subsequent to 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 2021–2022 is presented in Table 3.

Table 3. Number of Students Who Took Tests Remotely in the 2021–2022 Summative Test Administration

Subject	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 11	Total
ELA/L	15	16	13	15	7	9	1	76
Mathematics	16	18	15	16	6	9	1	81

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 2021-22. 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 and were all braille paper-pencil versions of the tests.

Table 4. Number of Students Who Took Paper-Pencil Tests in the 2021–2022 Summative Test Administration

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

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 2021-22. This was not an option for administration in Hawai‘i in 2021–2022, 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

TAs must complete an online TA Certification Course every year in order to administer assessments. This web-based course is about 30–45 minutes long and covers information on testing policies and the steps for administering a test session in the online testing system. The course is 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 course. A second TA Certification Course of about 20 minutes is required for TAs administering tests in a remote format. For 2021–2022, TAs administering remote tests were required to take both 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 2020, 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 Hawai'i 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 2021–2022 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 *Centralized Reporting Trainings and Webinars* webpage provides links to short tutorial videos on the following aspects of Centralized Reporting: How to Create, Manage, and Edit Rosters; How to Access Centralized Reporting for Schools; How to Access Longitudinal Reports; How to Access Centralized Reporting for Teachers; How to Access Centralized Reporting for Districts; How to Modify Scores; How to Export and Print Student Data; How to Handscore Unscored Items; and How to Set Up Your Reports So They Make Sense.

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 2021–2022. The first series of virtual statewide trainings was held September 13–14, 2021. The second series of virtual statewide trainings was held January 24–February 1, 2022. 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 virtual statewide trainings was held August 18–October 12, 2021. These training sessions focused specifically on accessibility and accommodations for all Hawai'i statewide assessments, including the Smarter Balanced summative and interim 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 high school at public or public charter schools in Hawaiʻi are required to participate in the Smarter Balanced ELA/L and mathematics assessments, except the following:

- Students with significant cognitive disabilities who meet the criteria for a state-selected or state-developed ELA/L alternate assessment based on the extensions of the Common Core standards or Hawaiʻi Content and Performance Standards (HCPS) III (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/resources/resources-2021-2022/usability,-accessibility,-and-accommodations-guidelines-2021-2022.

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 2021–2022 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/resources/resources-2021-2022/smarter-balanced-summative-test-administration-manual-2021-2022.

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. 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 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.

Mathematics Tools. These digital tools (e.g., embedded ruler, embedded protractor) are used for measurements related to mathematics items. They are available only with the specific items 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. 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). 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.

Translations (Glossaries) (for mathematics). 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 (Dual Language) (for mathematics). 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. 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.

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.

Read-Aloud (for mathematics and ELA/L items, but not for 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 items). 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 non-writing items). Students dictate their responses to a human who records verbatim what they dictate. The scribe must be trained and qualified and must follow the administration guidelines provided in the *Smarter Balanced Online Summative Test Administration Manual*.

Separate Setting. 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.

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 the fingertips. Graphic material (e.g., maps, charts, graphs, diagrams, illustrations) is presented in a raised format (paper or thermoform). Contracted and non-contracted braille is available; Nemeth 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 stims). Printed text may appear on the computer screen as audio materials are presented.

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.

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 the 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 grades 6–8 and 11 mathematics tests). 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.

Mathematics Manipulatives. This accommodation 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.

Multiplication Table (grade 4 and above mathematics tests). 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 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 writing items). Students dictate their responses to a human who records verbatim what they dictate. The scribe must be trained and qualified and must follow the administration guidelines provided in the *Smarter Balanced Online Summative Test Administration Manual*.

Speech-to-Text. Voice recognition allows students to use their voices as 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 Prediction. 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 2021–2022 administration. Tables 6–11 provide the numbers of students who utilized any of the offered accommodations and designated supports. 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. SY 2021–2022 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 Stimuli Global Notes[3] Highlighter Keyboard Navigation Line Reader Mark for Review Mathematics Tools[4] Spellcheck Strikethrough Thesaurus[2] Writing Tools[5] Zoom	Color Contrast Illustration Glossaries[6] Masking Mouse Pointer Streamline Text-to-Speech[7] Translated Test Directions[6] Translations (Glossaries)[6] Translations (Dual Language)[6] Turn Off Any Universal Tools	American Sign Language[8] Braille Braille Transcript[9] Closed Captioning[9] Text-to-Speech[10]
	Non-Embedded	
Breaks English Dictionary[2] Scratch Paper Thesaurus[2]	Amplification Bilingual Dictionary[2] Color Contrast Color Overlay Illustration Glossaries[11] Magnification Medical Supports Noise Buffers Read-Aloud[12] Read-Aloud in Spanish[6] Scribe[13] Separate Setting Simplified Test Directions Translated Student Interface Messages Translations (Glossaries)[11]	100s Number Table Abacus Alternate Response Options[14] Braille[15] Calculator[1] Mathematics Manipulatives[16] Multiplication Table Print-on-Demand Read-Aloud[17] Scribe[2] Speech-to-Text Word Prediction

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

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

[2] For ELA/L performance task full-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 ELA/L performance task (PT) stimuli, ELA/L PT and CAT items (not ELA/L CAT reading passages), and mathematics stimuli and items: must be set in TIDE before test begins

[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 items on the paper-pencil test

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

[13] For ELA/L non-writing items and mathematics items

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

[15] For paper-pencil assessments

[16] 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

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

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

Accommodations	**Grade**						
	3	**4**	**5**	**6**	**7**	**8**	**11**
Embedded Accommodations							
American Sign Language	1	5	2	4	11	5	4
Braille							1
Braille Transcript		1					
Closed Captioning	10	12	7	9	20	16	4
Text-to-Speech: Passages and Items	1	4	3	1	1	4	3
Non-Embedded Accommodations							
Alternate Response Options	1	3	1	2			1
Print-on-Demand: Stimuli & Items		1	1		1	1	
Read-Aloud Passages		4	3	6			2
Scribe (Full-Write)	1	3	4	2	2	3	1
Speech-to-Text	4	3	9	7	3	2	1
Word Prediction						1	

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

Designated Supports	**Subgroup**	**Grade**						
		3	**4**	**5**	**6**	**7**	**8**	**11**
Color Contrast	Overall	5	15	12	4	3	5	1
	ELL		1	1		1	1	
	Disability	5	8	7	3	3	1	1
Masking	Overall	142	34	45	76	26	67	1
	ELL	16	1	1	15	14	14	
	Disability	30	21	12	35	12	39	1
Mouse Pointer	Overall	1		4	11	2	5	
	ELL			1	3	1	2	
	Disability			3	9	2	1	
Streamline	Overall	84	54	44	52	14	10	14
	ELL	9	7	3	8	3	1	4
	Disability	44	30	23	39	12	9	14
Text-to-Speech: Items	Overall	3,432	2,829	2,921	1,940	897	962	57
	ELL	808	699	684	553	328	371	19
	Disability	774	801	826	625	363	366	41
Text-to-Speech: Stimuli	Overall	5	5	21	4	1	1	3
	ELL	2	1		1	1	1	2
	Disability	1	2	3				1
Text-to-Speech: Stimuli and Items	Overall	3,542	2,915	3,015	2,196	1,008	1,139	57
	ELL	833	729	713	578	336	381	20
	Disability	806	829	836	685	387	399	42

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

Designated Supports	**Subgroup**	**Grade**						
		3	**4**	**5**	**6**	**7**	**8**	**11**
Amplification	Overall		1	1	1			
	ELL							
	Disability							
Bilingual Dictionary	Overall		2	3	12	7	7	8
	ELL		1	3	11	7	7	8
	Disability					1		
Color Contrast	Overall	2	1	1				
	ELL		1					
	Disability	1	1	1				
Color Overlay	Overall		1	1				3
	ELL							
	Disability			1				3
Magnification	Overall	5	24	2		1	2	2
	ELL						1	
	Disability	3	4	2		1		2
Medical Supports	Overall		3	1	1	2		
	ELL							
	Disability							
Noise Buffers	Overall		4				1	1
	ELL							
	Disability						1	
Read-Aloud Items	Overall	94	132	102	19	13	6	12
	ELL	17	23	16	3	2		6
	Disability	44	60	70	14	11	5	10
Read-Aloud Stimuli	Overall	83	99	71	17	9	4	18
	ELL	14	16	8	2	1		12
	Disability	34	53	41	13	7	3	10
Scribe (Not Full-Write)	Overall	2	2	2	2		2	2
	ELL	1						
	Disability	1	1	2	1		2	2
Separate Setting	Overall	276	266	351	214	168	165	63
	ELL	32	47	50	30	15	9	9
	Disability	166	194	265	164	131	132	41
Simplified Test Directions	Overall	251	244	244	64	26	27	41
	ELL	65	47	54	6	10	2	10
	Disability	52	67	68	48	21	21	31
Translated Student Interface Messages	Overall			1	1			
	ELL			1				
	Disability							

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

Accommodations	**Grade**						
	3	**4**	**5**	**6**	**7**	**8**	**11**
Embedded Accommodations							
American Sign Language	1	5	2	4	11	5	4
Non-Embedded Accommodations							
100s Number Table	20	22	18	9	3		
Abacus		2	2	1			
Alternate Response Options	1	3	1	1			1
Calculator				3		1	
Math Manipulatives	13	9	7	2		1	
Multiplication Table			4	6			1
Print-on-Demand: Stimuli & Items			1		1	1	
Speech-to-Text	4	4	9	8	3	1	1
Word Prediction						1	

Table 10. Total Students with Allowed Embedded Designated Supports: Mathematics

Designated Supports	**Subgroup**	**Grade**						
		3	**4**	**5**	**6**	**7**	**8**	**11**
Color Contrast	Overall	5	15	10	4		3	1
	ELL		1	1			1	
	Disability	4	8	5	3			1
Illustration Glossaries	Overall	54	96	48	202	141	192	
	ELL	39	51	45	143	126	165	
	Disability	2	15	1	30	27	26	
Masking	Overall	143	30	45	77	25	65	1
	ELL	14	1	1	15	15	14	
	Disability	32	18	11	35	10	37	1
Mouse Pointer	Overall	1		4	11		7	
	ELL			1	3		3	
	Disability			3	9		2	
Streamline	Overall	84	55	43	51	13	10	13
	ELL	9	6	3	8	3	1	4
	Disability	43	32	23	38	11	9	13
Text-to-Speech: Items	Overall	6	5	5		1		
	ELL		1					
	Disability	1		3		1		
Text-to-Speech: Stimuli	Overall	2	5	4	2	1		
	ELL	2		2	1			
	Disability			1	1			
Text-to-Speech: Stimuli and Items	Overall	3,749	3,031	3,141	2,315	1,036	1,161	64
	ELL	896	770	726	591	351	383	22
	Disability	842	856	873	689	406	409	47
Translations (Glossaries): Spanish	Overall	2	1	5	19	5	8	
	ELL	2		5	17	5	7	
	Disability				2	2	2	
Translations (Glossaries): Other Languages	Overall	10	19	22	75	60	98	
	ELL	10	17	20	65	58	90	
	Disability	1			7	3	4	
Translations (Dual Language): Spanish	Overall			4	1	2	2	
	ELL			4	1	2	2	
	Disability							

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

Designated Supports	**Subgroup**	**Grade**						
		3	**4**	**5**	**6**	**7**	**8**	**11**
Amplification	Overall		1	1	1			
	ELL							
	Disability							
Color Contrast	Overall	2	1					
	ELL		1					
	Disability	1	1					
Color Overlay	Overall		1					4
	ELL							
	Disability							4
Illustration Glossaries	Overall	2	1	2	8	4	2	
	ELL	2	1	1	1	2	1	
	Disability				5	3	2	
Magnification	Overall	5	24	1		1	1	2
	ELL						1	
	Disability	4	4	1		1		2
Medical Supports	Overall		3	1	1	1		
	ELL							
	Disability							
Noise Buffers	Overall		5				1	1
	ELL							
	Disability		1				1	
Read-Aloud Items	Overall	93	116	92	18	11	6	10
	ELL	14	16	12	2	2		3
	Disability	42	62	63	14	9	5	10
Read-Aloud Items (Spanish)	Overall	2	3	2				
	ELL	1	2	2				
	Disability							
Read-Aloud Stimuli	Overall	90	101	70	18	9	5	18
	ELL	13	15	7	2	1		11
	Disability	38	54	42	14	7	4	10
Read-Aloud Stimuli (Spanish)	Overall	2	3	2				
	ELL	1	2	2				
	Disability							
Scribe	Overall	1	3	2	2		1	2
	ELL							
	Disability	1	2	2	1		1	2
Separate Setting	Overall	273	271	349	213	163	161	66
	ELL	33	48	51	30	15	9	9
	Disability	162	198	262	160	128	133	43
Simplified Test Directions	Overall	249	242	246	60	26	28	41
	ELL	65	44	51	6	10	2	10
	Disability	50	68	70	45	21	22	31
Translated Student Interface Messages	Overall		1	2	1			
	ELL		1	1				
	Disability							

Designated Supports	Subgroup	Grade						
		3	4	5	6	7	8	11
Translations (Glossaries): Spanish	Overall							2
	ELL							2
	Disability							
Translations (Glossaries): Other Languages	Overall		3	2			1	3
	ELL		3	1			1	3
	Disability							

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 12 and 13 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 12. Test-Taking Time: ELA/L

Grade	**Average Testing Time (hh:mm)**	Standard Deviation **of Testing Time (hh:mm)**	**Testing Time in Percentiles (hh:mm)**				
			75th	**80th**	**85th**	**90th**	**95th**
			Overall Test				
3	2:32	1:39	3:11	3:33	3:58	4:34	5:42
4	2:49	1:48	3:32	3:55	4:23	5:03	6:18
5	2:48	1:39	3:32	3:52	4:19	4:55	5:56
6	2:50	1:37	3:29	3:47	4:12	4:48	5:51
7	2:37	1:24	3:14	3:31	3:52	4:25	5:17
8	2:35	1:21	3:13	3:30	3:52	4:19	5:05
11	1:57	0:59	2:26	2:38	2:51	3:09	3:42
			CAT Component				
3	0:54	0:31	1:05	1:10	1:17	1:30	1:51
4	0:57	0:33	1:09	1:15	1:23	1:34	1:56
5	0:58	0:32	1:10	1:16	1:24	1:36	1:56
6	1:05	0:34	1:18	1:24	1:31	1:43	2:03
7	1:00	0:29	1:13	1:18	1:25	1:35	1:53
8	0:59	0:28	1:12	1:17	1:23	1:32	1:49
11	0:47	0:22	0:58	1:02	1:06	1:13	1:25
			PT Component				
3	1:39	1:19	2:08	2:24	2:45	3:17	4:11
4	1:52	1:25	2:26	2:44	3:05	3:37	4:35
5	1:50	1:18	2:25	2:41	3:01	3:27	4:19
6	1:45	1:14	2:14	2:29	2:48	3:14	4:05
7	1:37	1:05	2:04	2:17	2:35	3:00	3:44
8	1:36	1:03	2:05	2:18	2:34	2:59	3:37
11	1:10	0:45	1:30	1:40	1:51	2:06	2:29

Table 13. Test-Taking Time: Mathematics

Grade	**Average Testing Time (hh:mm)**	**SD of Testing Time (hh:mm)**	**Testing Time in Percentiles (hh:mm)**				
			75th	**80th**	**85th**	**90th**	**95th**
			Overall Test (CAT Component)				
3	0:56	0:36	1:08	1:16	1:25	1:37	2:04
4	0:59	0:37	1:13	1:20	1:30	1:43	2:08
5	1:06	0:39	1:22	1:29	1:39	1:55	2:21
6	1:05	0:36	1:18	1:24	1:32	1:44	2:07
7	1:01	0:31	1:14	1:20	1:27	1:38	2:00
8	1:07	0:33	1:23	1:29	1:36	1:48	2:07
11	0:50	0:26	1:02	1:08	1:14	1:23	1:39

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 throughout the testing window.

2.8.1 Changes in Student Performance

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

A large score gain or loss in student scores between administration years is detected by examining the residuals for outliers. The residuals are computed as the observed value minus the regression model's predicted value. 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, and $\hat{e}_i$ is the residual for the *i*th student.

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

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

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

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. COMPARABILITY OF THE SHORTENED AND FULL BLUEPRINTS

The purpose of the shortened blueprint was to assess students’ progress with acceptable test reliability while significantly reducing testing time. For the English language arts/literacy (ELA/L) shortened blueprint, the computer-adaptive test (CAT) portion of the blueprint was reduced, but the performance task (PT) component was kept as is. For the mathematics shortened blueprint, the CAT component of the blueprint was reduced, and the PT component was excluded. In mathematics, Hawai‘i removed the PTs to reduce the testing time further since all targets and Depth of Knowledge (DOK) measures associated with the PTs were covered in the CAT.

For the Hawai‘i shortened blueprint, the blueprint constraints for claims in the Smarter Balanced full-test blueprint were reduced proportionately. This was implemented to achieve a shorter CAT blueprint that contains the same claims and DOK levels with the same relative coverage as the full-test blueprint. The target requirements were adjusted to accommodate the claim and DOK requirements of the shortened test blueprint while still covering as many targets as possible. In mathematics, fewer CAT items were removed in Claim 2 and Claim 4 to compensate for the removal of PT items.

The comparability of the shortened and full blueprints is examined for blueprint constraints, target coverages, reliability of scores, and student performance. The major impact on testing time is also presented in this section.

The impact of the shortened blueprint on student performance was examined by projecting the Hawai‘i shortened blueprint to the pre-pandemic summative data with full blueprints (2018–2019 ELA/L summative data and 2016–2017 mathematics summative data) because student performance on the full and shortened blueprints cannot be directly compared between years due to the pandemic effect on student performance and student participation rates. The projected data allowed us to compare the performance of two blueprints on the same students. In mathematics, the 2016–2017 summative data were used because Hawai‘i used the Smarter Balanced full blueprint until 2016–2017 and removed the PTs in 2017–2018.

In ELA/L, for each CAT, the individual items and passages that fit the Hawai‘i shortened blueprint were selected randomly within a claim and a target and combined with the PT to form a projected estimate of the student’s performance on the Hawai‘i shortened blueprint test. In mathematics, for each CAT, the individual items that fit the Hawai‘i shortened blueprint were selected randomly within a claim.

The impact on student performance was for the overall test scores only, not for claim scores. The reliabilities of the projected scores were compared with the 2021–2022 reliabilities to verify the validity of the projected scores (refer to Appendix B).

3.1 BLUEPRINTS

3.1.1 ELA/L Blueprints

Tables 14 and 15 present the number of items for the total test and claims and the proportion of the items in each claim to the total test length for the Hawai‘i shortened blueprint, the Smarter Balanced adjusted blueprint, and the Smarter Balanced full blueprint. The PT is a common component in all of these blueprints. The Smarter Balanced adjusted blueprint is provided as a reference to compare with the Hawai‘i shortened blueprint. The Hawai‘i shortened blueprint is the same as the Smarter Balanced adjusted blueprint, except for a slightly longer test length by two items. Fewer CAT items were removed to compensate for the removal of PT items because the initial plan was to remove the PT component. The

Hawai‘i Department of Education (HIDOE) decided later to keep the PT as is without adjusting the CAT blueprints associated with PTs. Tables 16–17 exhibit how every blueprint constraint in the Smarter Balanced full blueprint was reduced in the Hawai‘i shortened blueprint.

Table 14. Number of Items by Claim: ELA/L

Grade	**Hawai’i Shortened Blueprint (CAT + PT)**					**Smarter Balanced Adjusted Blueprint (CAT + PT)**					**Smarter Balanced Full Blueprint (CAT + PT)**				
	Total Test	**C1**	**C2**	**C3**	**C4**	**Total Test**	**C1**	**C2**	**C3**	**C4**	**Total Test**	**C1**	**C2**	**C3**	**C4**
3	24	8	6	4	6	22	8	5	4	5	38–41	14–16	7	8–9	9
4	24	8	6	4	6	22	8	5	4	5	38–41	14–16	7	8–9	9
5	24	8	6	4	6	22	8	5	4	5	38–41	14–16	7	8–9	9
6	26	10	6	4	6	24	10	5	4	5	38–42	14–17	7	8–9	9
7	26	10	6	4	6	24	10	5	4	5	38–42	14–17	7	8–9	9
8	26	10	6	4	6	24	10	5	4	5	38–42	14–17	7	8–9	9
11	26	10	6	4	6	24	10	5	4	5	39–41	15–16	7	8–9	9

Note. Full-write item is counted as one item.

Table 15. Percentage of Items by Claim: ELA/L

Grade	**Hawai’i Shortened Blueprint (CAT + PT)**				**Smarter Balanced Adjusted Blueprint (CAT + PT)**				**Smarter Balanced Full Blueprint (CAT + PT)**			
	C1	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**
3	33%	25%	17%	25%	36%	23%	18%	23%	34–42%	17–18%	20–24%	22–24%
4	33%	25%	17%	25%	36%	23%	18%	23%	34–42%	17–18%	20–24%	22–24%
5	33%	25%	17%	25%	36%	23%	18%	23%	34–42%	17–18%	20–24%	22–24%
6	38%	23%	15%	23%	42%	21%	17%	21%	33–45%	17–18%	19–24%	21–24%
7	38%	23%	15%	23%	42%	21%	17%	21%	33–45%	17–18%	19–24%	21–24%
8	38%	23%	15%	23%	42%	21%	17%	21%	33–45%	17–18%	19–24%	21–24%
11	38%	23%	15%	23%	42%	21%	17%	21%	37–41%	17–18%	20–23%	22–23%

Note. Full-write item is counted as one item.

Table 16. Changes in Test Blueprints: ELA/L (Grades 3–5)

Claim	Content Category/Target	Hawai'i Shortened Blueprint		Smarter Balanced Full Blueprint	
		CAT	PT	CAT	PT
	Total Test	**22**	**2**	**36–39**	**2**
1	**Literary Text**	4		7–8	
	Target 2: Central Ideas	1–3		1–2	
	Target 4: Reasoning and Evidence			1–2	
	Targets 1, 3, 5, 6, and 7	1–3		3–6	
	Long Literary Text Passage	1		1	
	Short Literary Text Passage			1	
	Informational Text	4		7–8	
	Target 9: Central Ideas	1–3		1–2	
	Target 11: Reasoning and Evidence			1–2	
	Targets 8, 10, 12, 13, and 14	1–3		3–6	
	Long Informational Text Passage	1		1	
	Short Informational Text Passage			1	
	DOK 2	≥ 4		≥ 7	
	DOK 3 or Higher	≥ 1		≥ 2	
2	**Writing**	5	1	6	1
	Target 1, 3, or 6: Organization/Purpose	1		1	
	Target 1, 3, or 6: Evidence/Elaboration	1		1	
	Target 8: Language and Vocabulary Use	1		1	
	Target 9: Edit/Clarify	2		3	
	DOK 2	≥ 2		≥ 2	
	Targets (2, 4, or 7), 8, and 9		1		1
	DOK 4		1		1
3	**Listening**	4		8–9	
	Target 4: Listen/Interpret	4		8–9	
	DOK 2 or Higher	≥ 2		≥ 3	
	Listening Passage	2		3–4	
4	**Research**	5	1	8	1
	Target 2: Interpret & Integrate Information	1–2		2–3	
	Target 3: Analyze Information/Sources	1–2	1	2–3	1
	Target 4: Use Evidence	1–2		2–3	
	DOK 3 or 4		1		1

Table 17. Changes in Test Blueprints: ELA/L (Grades 6–8, 11)

Claim	Content Category/Target	Hawaiʻi Short Blueprint		Smarter Balanced Full Blueprint	
		CAT	**PT**	**CAT**	**PT**
	Total Test	**24**	**2**	**36–40 (37–39 [a])**	**2**
1	**Literary Text**	4		4	
	Target 2: Central Ideas	1–3		1	
	Target 4: Reasoning and Evidence			1	
	Targets 1, 3, 5, 6, and 7	1–3		2–5	
	Target 2 or 4 short text (DOK3 or 4)	0		0–1	
	Long Literary Text Passage	1		1	
	Informational Text	6		10–12 [b]	
	Target 9: Central Ideas	2–4		2–5 (2–4 [a])	
	Target 11: Reasoning and Evidence				
	Targets 8, 10, 12, 13, and 14	2–4		7–10	
	Target 9 or 11 short text (DOK3 or 4)	0		0–1	
	Long Informational Text Passage	1		1	
	Short Informational Text Passage	1		2	
	DOK 1	≤ 3 (≤ 2 [a])		≤ 5 (≤ 4 [a])	
	DOK 3 or Higher	≥ 1 (≥ 2 [a])		≥ 2 (≥ 3 [a])	
2	**Writing**	5	1	6	1
	Target 1, 3, or 6: Organization/Purpose	1[c]		1	
	Target 1, 3, or 6: Evidence/Elaboration	1[c]		1	
	Target 8: Language and Vocabulary Use	1		1	
	Target 9: Edit/Clarify	2		3	
	DOK 2	≥ 2		≥ 2	
	DOK 3 or Higher (Brief-Write items)	0		1	
	Brief Writes (DOK3, Targets 1,3, or 6)	0		1	
	Targets (2, 4, or 7), 8, and 9		1		1
	DOK 4		1		1
3	**Listening**	4		8–9	
	Target 4: Listen/Interpret	4		8–9	
	DOK 2 or Higher	≥ 2		≥ 3 (≥ 4 [a])	
	Listening Passage	2		3–4	
4	**Research**	5	1	8	1
	Target 2: Analyze and Integrate Information	1–2		2–3	
	Target 3: Evaluate Information/Sources	1–2	1	2–3	1
	Target 4: Use Evidence	1–2		2–3	
	DOK 3 or 4		1		1

[a] Required items in parentheses are for grade 11.

[b] Required items for Informational Text are 10–12 in grades 6 and 7, 12 in grade 8, and 11–12 in grade 11.

[c] In the Hawaiʻi short blueprint item pool, all items in Claim 2 targets 1, 3, and 6 are DOK 2 items.

Table 18 presents the target coverage in each test by claim for the Hawai'i shortened blueprint, and Smarter Balanced adjusted blueprint and full blueprint. The table includes the total number of targets specified in the blueprints and the mean number of unique targets administered to each test. 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 were not expected to be covered in every test, but were expected to be covered at the aggregate level. In Claim 1, the number of targets covered in each test was expected to be fewer in both the Smarter Balanced adjusted blueprint and the Hawai'i shortened blueprint than in the Smarter Balanced full blueprint, given the reduced items in Claim 1. The average number of unique targets assessed within each claim in the Hawai'i shortened blueprint are similar to the Smarter Balanced adjusted blueprint for ELA/L.

Table 18. Average Number of Unique Targets Assessed Within Each Claim: ELA/L

Grade	Total Targets Specified in Blueprint (CAT+PT)				Hawai'i Short Blueprint (CAT+PT)				Smarter Balanced Adjusted Blueprint (CAT+PT)				Smarter Balanced Full Blueprint (CAT+PT)			
	C1	**C2***	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**
3	14	8	1	3	7.6	5	1	3	7.6	5	1	3	10.6	5	1	3
4	14	8	1	3	7.8	5	1	3	7.8	5	1	3	10.5	5	1	3
5	14	8	1	3	7.3	5	1	3	7.3	5	1	3	11.4	5	1	3
6	14	8	1	3	8.9	5	1	3	9.3	5	1	3	10.2	5	1	3
7	14	8	1	3	9.1	5	1	3	9.2	5	1	3	10.7	5	1	3
8	14	8	1	3	9.0	5	1	3	8.8	5	1	3	10.7	5	1	3
11	14	7	1	3	8.4	5	1	3	8.3	5	1	3	10.0	5	1	3

*Note: In Claim 2, Targets 1, 3, 6, 8, and 9 were assessed in the CAT segment, while Targets 2, 4, and 7 (Targets 4 and 7 in grade 11) were assessed in the PT segment. Each PT form assessed one target of Targets 2, 4, or 7.

3.1.2 Mathematics Blueprints

Tables 19–20 present the number of items for the total test and claims and the proportion of the items in each claim to the total test length for the Hawai'i shortened blueprint, the Smarter Balanced adjusted blueprint, and the Smarter Balanced full blueprint. The PT was kept in the Smarter Balanced full blueprint, and the adjusted blueprint but was removed from the Hawai'i shortened blueprint. The CAT covered targets in Claims 1–4 while each PT form covered targets in Claims 2, 3, or 4. The Smarter Balanced adjusted blueprint is provided as a reference to compare with the Hawai'i shortened blueprint. Tables 21–24 present the required CAT items for each blueprint constraint in the Hawai'i shortened blueprint and the Smarter Balanced full blueprint.

Table 19. Number of Items by Claim: Mathematics

Grade	Hawai'i Shortened Blueprint (CAT only)					Smarter Balanced Adjusted Blueprint (CAT + PT)					Smarter Balanced Full Blueprint (CAT + PT)				
	Total Test	**C1**	**C2**	**C3**	**C4**	**Total Test**	**C1**	**C2**	**C3**	**C4**	**Total Test**	**C1**	**C2**	**C3**	**C4**
3	22	12	2	5	3	21–23	10	2–3	4–6	3–5	35–40	17–20	4–5	8–10	4–6
4	22	12	2	5	3	21–23	10	2–3	4–6	3–5	35–40	17–20	4–5	8–10	4–6
5	22	12	2	5	3	21–23	10	2–3	4–6	3–5	35–40	17–20	4–5	8–10	4–6
6	22	12	2	5	3	20–23	9–10	2–3	4–6	3–5	34–39	16–19	4–5	8–10	4–6
7	22	12	2	5	3	21–23	10	2–3	4–6	3–5	35–40	17–20	4–5	8–10	4–6
8	22	12	2	5	3	21–23	10	2–3	4–6	3–5	35–40	17–20	4–5	8–10	4–6
11	24	14	2	5	3	22–24	11	2–3	4–6	3–5	37–42	19–22	4–5	8–10	4–6

Table 20. Percentage of Items by Claim: Mathematics

Grade	**Hawai’i Shortened Blueprint (CAT only)**				**Smarter Balanced Adjusted Blueprint (CAT + PT)**				**Smarter Balanced Full Blueprint (CAT + PT)**			
	C1	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**
3	55%	9%	23%	14%	43–48%	10–13%	19–26%	14–22%	49–50%	11–13%	23–25%	11–15%
4	55%	9%	23%	14%	43–48%	10–13%	19–26%	14–22%	49–50%	11–13%	23–25%	11–15%
5	55%	9%	23%	14%	43–48%	10–13%	19–26%	14–22%	49–50%	11–13%	23–25%	11–15%
6	55%	9%	23%	14%	43–45%	10–13%	20–26%	15–22%	47–49%	12–13%	24–26%	12–15%
7	55%	9%	23%	14%	43–48%	10–13%	19–26%	14–22%	49–50%	11–13%	23–25%	11–15%
8	55%	9%	23%	14%	43–48%	10–13%	19–26%	14–22%	49–50%	11–13%	23–25%	11–15%
11	58%	8%	21%	13%	46–50%	9–13%	18–25%	14–21%	51–52%	11–12%	22–24%	11–14%

Table 21. Blueprint Requirements for Claim 1: Mathematics (Grades 3–4)

Grade 3			**Grade 4**		
Claim 1 Content / Target	**Hawai‘i Shortened Blueprint**	**Smarter Balanced Full Blueprint**	**Claim 1 Content / Target**	**Hawai‘i Shortened Blueprint**	**Smarter Balanced Full Blueprint**
Total Test	22	31–34	Total Test	22	31–34
Overall	12	17–20	Overall	12	17–20
DOK 2 or Higher	≥ 4	≥ 7	DOK 2 or Higher	≥ 4	≥ 7
Priority Cluster	9	13–15	*Priority Cluster*	9	13–15
Targets B, C, G, I	4	5–6	Targets A, E, F	5	8–9
Targets D, F	4	5–6	Target G	2	2–3
Target A	1	2–3	Target D	1	1–2
Supporting Cluster	3	4–5	Target H	1	1
Targets E, J, K	2	3–4	*Supporting Cluster*	3	4–5
Target H	1	1	Targets I, K	1	2–3
			Targets B, C, J	1	1
			Target L	1	1

Table 22. Blueprint Requirements for Claim 1: Mathematics (Grades 5–6)

Grade 5			**Grade 6**		
Claim 1 Content / Target	**Hawai‘i Shortened Blueprint**	**Smarter Balanced Full Blueprint**	**Claim 1 Content / Target**	**Hawai‘i Shortened Blueprint**	**Smarter Balanced Full Blueprint**
Total Test	22	31–34	Total Test	22	30–33
Overall	12	17–20	Overall	12	16–19
DOK 2 or Higher	≥ 4	≥ 7	DOK 2 or Higher	≥ 4	≥ 7
Priority Cluster	9	13–15	*Priority Cluster*	9	12–15
Targets E, I	4	5–6	Targets E, F	4	5–6
Target F	3	4–5	Target A	2	3–4
Targets C, D	2	3–4	Targets G, B	2	2
Supporting Cluster	3	4–5	Target D	1	2
Targets J, K	2	2–3	*Supporting Cluster*	3	4–5
Targets A, B, G, H	1	2	Targets C, H, I, J	3	4–5

Table 23. Blueprint Requirements for Claim 1: Mathematics (Grades 7–8, 11)

Claim 1 Content / Target	**Grade 7**		**Grade 8**		**Grade 11**	
	Hawai‘i Shortened Blueprint	**Smarter Balanced Full Blueprint**	**Hawai‘i Shortened Blueprint**	**Smarter Balanced Full Blueprint**	**Hawai‘i Shortened Blueprint**	**Smarter Balanced Full Blueprint**
Total Test	22	31–34	22	31–34	24	33–36
Overall	12	17–20	12	17–20	14	19–22
DOK 2 or Higher	≥ 4	≥ 7	≥ 4	≥ 7	≥ 4	≥ 7
Priority Cluster	9	13–15	9	13–15	10	14–16
Targets A, D	5	8–9	3	5–6		
Targets B, C	4	5–6	3	5–6		
Targets C, D			3	5–6		
Targets B, E, G			3	5–6		
Targets F, H			3	2–3		
Targets D, E					1–2	2
Target F					1	1
Targets G, H, I					3	4–5
Target J					1–2	2
Target K					1–2	2
Targets L, M, N					2	3–4
Supporting Cluster	3	4–5	3	4–5	4	5–6
Targets E, F	2	2–3				
Targets G, H, I	1	1–2				
Targets A, I, J			3	4–5		
Target O					0–2	2
Target P					0–2	1–2
Targets A, B					0–1	1
Target C					0–1	1

Table 24. Blueprint Requirements for Claims 2, 3, and 4: Mathematics (Grades 3–8, 11)

Claim	Content / Target	Hawai‘i Shortened Blueprint	Smarter Balanced Full Blueprint	
		CAT	CAT	PT
	Total Test	22–24	30–36*	4–6*
2 & 4	Overall	5	6	2–4
	DOK 3 or Higher	≥ 2	≥ 2	
	2. Target A	1	2	1–2
	2. Targets B, C, D	1	1	
	4. Targets A, D	1	1	1–3
	4. Targets B, E	1	1	
	4. Targets C, F	1	1	
	4. Target G	0	0	
3	Overall	5	8	0–2
	DOK 3 or Higher	≥ 2	≥ 2	
	Targets A, D	2 (1–3) [+]	3 (2–4)[+]	0–2
	Targets B, E	2 (1–3) [+]	3 (2–4)[+]	
	Targets C, F	1 (0–2) [+]	2 (1–3)[+]	

* Total test length is computed by adding ranges specified in Smarter Balanced blueprint document.

[+] The item distribution is in parentheses due to the no-calculator segment in Claim 3 in grades 6 and 11.

Table 25 presents the target coverage in each test by claim for the Hawai‘i shortened blueprint and Smarter Balanced adjusted blueprint and full blueprint. The table includes the total number of targets specified in the blueprints and the mean number of targets administered to each test. Similar to ELA/L, the Smarter Balanced blueprints for mathematics did not require every target to be covered in each test, therefore it was expected that the number of targets covered in each test would vary slightly across individual tests. Although the target coverage varied somewhat across individual tests, all targets were covered at an aggregate level for both the Smarter Balanced full blueprint and Hawai‘i shortened blueprint tests.

Table 25. Average Number of Unique Targets Assessed Within Each Claim: Mathematics

Grade	Total Targets Specified in Blueprint				Hawai‘i Shortened Blueprint (CAT)				Smarter Balanced Adjusted Blueprint (CAT+PT)				Smarter Balanced Full Blueprint (CAT+PT)			
	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4
3	11	4	6	6	10.0	2	4.2	3	9.0	1.5	4.1	2.6	10.7	2.2	5.5	3.4
4	12	4	6	6	9.0	2	4.2	3	9.0	1.9	4.1	2.9	10.0	2.2	5.5	3.5
5	11	4	6	6	8.0	2	4.0	3	8.0	1.7	4.1	2.8	9.0	2.1	5.5	3.6
6	10	4	7	6	9.0	2	3.5	3	8.6	1.8	3.9	2.4	10.0	2.3	5.4	3.4
7	9	4	7	6	6.9	2	3.6	3	6.6	1.6	4.0	2.8	8.0	2.2	5.3	3.6
8	10	4	7	6	10.0	2	3.8	3	9.0	1.6	4.1	3.3	10.0	2.1	5.5	3.7
11	16	4	7	6	13.4	2	3.7	3	9.7	1.7	3.8	2.7	14.8	2.3	5.4	3.5

3.2 RELIABILITY

For reliability, the marginal reliability was computed for the scale scores. Marginal reliability is a measure of the overall reliability of an assessment based on the average conditional standard error of measurement (CSEM), estimated at different points on the ability scale, for all students.

Table 26 presents the marginal reliability coefficients and the average CSEMs for the Hawaiʻi shortened blueprint and the Smarter Balanced full blueprint. In ELA/L, although the CAT length decreased, the total test level reliability coefficients for the shortened test were still high, ranging from 0.88 to 0.89, which is just below the reliability of the test with the full blueprint, 0.92. In mathematics, despite of the test length reduction in both CAT and PT, the reliability coefficients were still high, ranging from 0.86 to 0.91.

Table 26. Marginal Reliability and Average Conditional Standard Error of Measurement, Overall Test

Grade	Hawaiʻi Shortened Blueprint			Smarter Balanced Full Blueprint		
	Items	Reliability	Average CSEM	Items	Reliability	Average CSEM
ELA/L						
3	24	0.89	33.78	38–41	0.92	25.80
4	24	0.88	36.04	38–41	0.92	27.73
5	24	0.89	35.33	38–41	0.92	27.33
6	26	0.89	34.91	38–42	0.92	28.39
7	26	0.88	36.98	38–42	0.92	29.18
8	26	0.88	36.91	38–42	0.92	29.20
11	26	0.88	40.69	39–41	0.92	32.47
Mathematics						
3	22	0.91	28.25	35–40	0.95	19.65
4	22	0.91	27.65	35–40	0.94	19.55
5	22	0.90	31.80	35–40	0.94	22.44
6	22	0.88	39.32	34–39	0.94	25.37
7	22	0.87	42.47	35–40	0.93	28.53
8	22	0.86	46.80	35–40	0.93	31.05
11	24	0.87	43.97	37–42	0.92	33.08

The CSEMs across total scale scores are displayed in Figures 1 and 2. The vertical dotted lines indicate Level 2, Level 3, and Level 4 cuts. 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. For the Level 3 (proficiency) cut, the CSEM and classification accuracy and consistency are provided in Table 27. The classification accuracy and consistency for the shortened blueprint is high. The reliability of the overall scores and the consistency and accuracy classifications at Level 3 cut are acceptable to assess student's progress, similar to the Smarter Balanced full blueprint.

Figure 1. Conditional Standard Error of Measurements Across Estimated Score Range: ELA/L

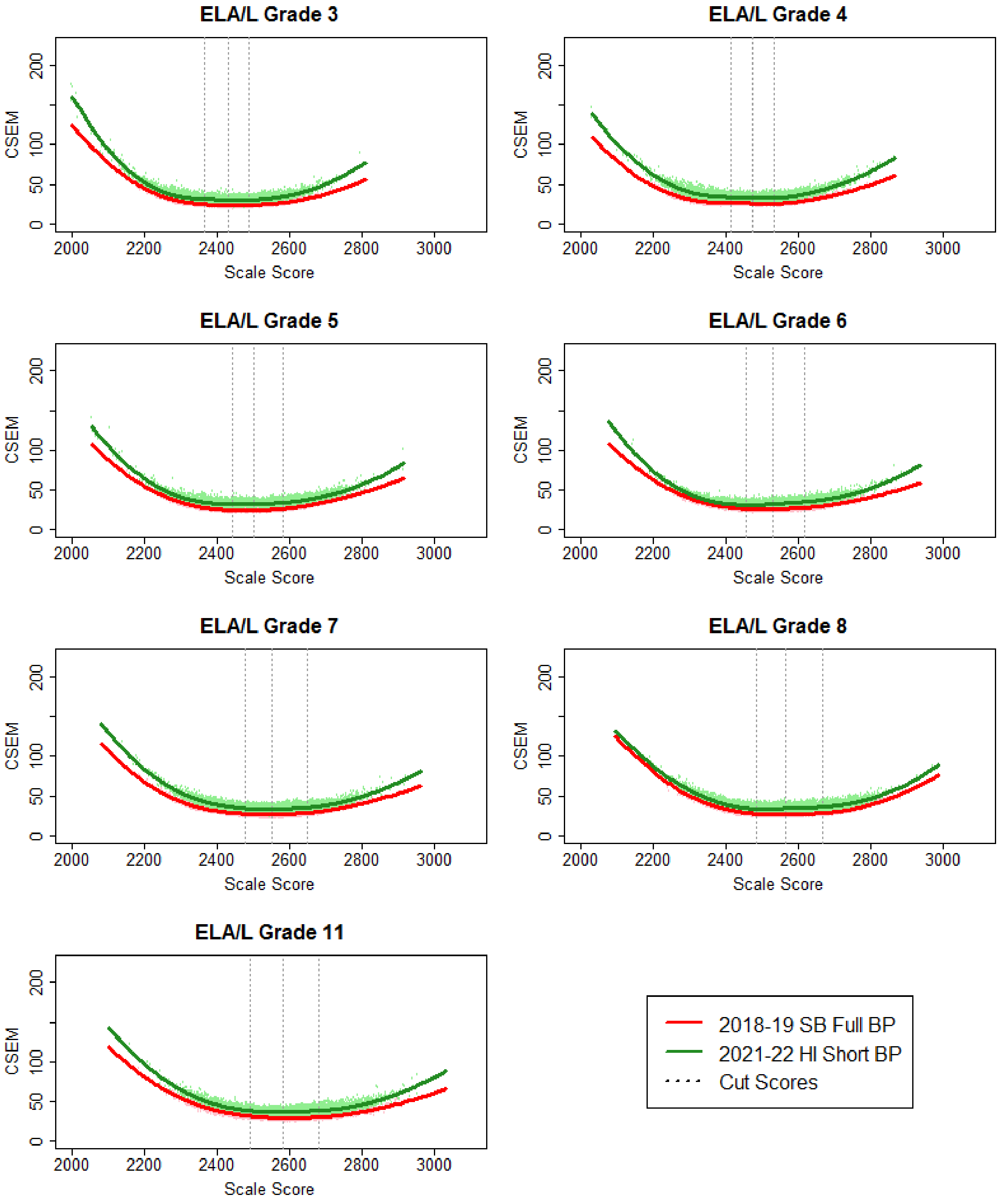


Figure 2. Conditional Standard Error of Measurements Across Estimated Score Range: Mathematics

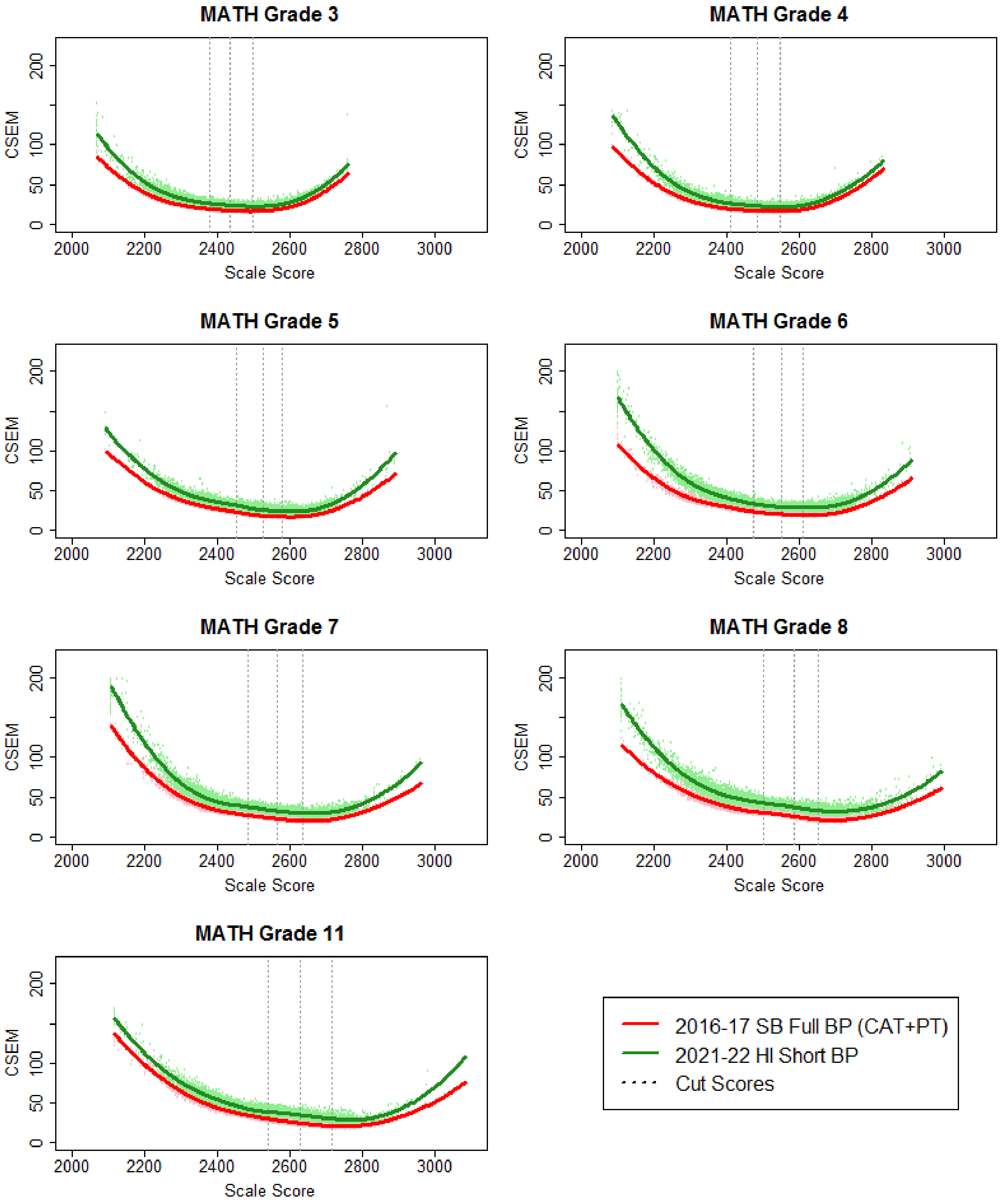


Table 27. Average CSEM and Classification Accuracy and Consistency At Level 3 Cut

Grade	**Hawai'i Shortened Blueprint**			**Smarter Balanced Full Blueprint**		
	Level 3 Cut	**% Accuracy**	**% Consistency**	**Level 3 Cut**	**% Accuracy**	**% Consistency**
ELA/L						
3	30.5	91	87	23.3	93	90
4	33.3	90	87	26.3	93	89
5	32.7	91	88	24.8	93	90
6	33.0	91	88	26.0	92	89
7	33.1	91	87	26.9	92	89
8	35.0	91	87	26.9	93	90
11	38.2	91	87	29.8	93	90
Mathematics						
3	24.1	92	89	17.7	93	91
4	23.3	92	89	17.4	93	91
5	26.7	92	89	18.7	94	91
6	29.6	92	88	20.9	93	91
7	33.1	91	87	22.3	94	91
8	37.7	92	88	25.7	94	92
11	35.8	92	89	25.1	94	92

Table 28–29 present the marginal reliability coefficients and the average CSEMs for claim scores. The claim scores with the shortened blueprint are, as expected, much less reliable than the full blueprint, especially in Claim 3 listening in ELA/L and Claims 2 and 4 in mathematics. For the shortened blueprint, the claim scores were reported for Claim 1 reading and Claim 2 writing in ELA/L and Claim 1 in mathematics for individual students. The claim-level scores were reported in three performance categories, taking into account the standard error of measurement of each student's scale score in a claim.

Table 28. Marginal Reliability and Average CSEM by Claims: ELA/L

Grade	Claim	Hawai'i Shortened Blueprint			Smarter Balanced Full Blueprint		
		Items	Reliability	Average CSEM	Items	Reliability	Average CSEM
3	Claim 1: Reading	8	0.62	76.45	14–16	0.77	49.81
	Claim 2: Writing	6	0.72	66.77	7	0.74	58.63
	Claim 3: Listening	4	0.28	122.95	8–9	0.62	79.08
	Claim 4: Research	6	0.62	82.92	9	0.74	62.93
4	Claim 1: Reading	8	0.60	81.87	14–16	0.77	52.57
	Claim 2: Writing	6	0.70	72.58	7	0.74	64.48
	Claim 3: Listening	4	0.30	123.91	8–9	0.63	84.68
	Claim 4: Research	6	0.59	92.15	9	0.74	66.05
5	Claim 1: Reading	8	0.61	83.67	14–16	0.76	57.57
	Claim 2: Writing	6	0.74	69.41	7	0.73	64.50
	Claim 3: Listening	4	0.33	127.84	8–9	0.65	83.18
	Claim 4: Research	6	0.64	81.04	9	0.78	58.29
6	Claim 1: Reading	10	0.69	70.59	14–17	0.76	57.60
	Claim 2: Writing	6	0.72	69.48	7	0.71	62.48
	Claim 3: Listening	4	0.30	133.51	8–9	0.62	89.05
	Claim 4: Research	6	0.59	90.50	9	0.73	66.29
7	Claim 1: Reading	10	0.63	82.97	14–17	0.79	55.89
	Claim 2: Writing	6	0.72	71.56	7	0.74	67.16
	Claim 3: Listening	4	0.29	125.93	8–9	0.58	88.43
	Claim 4: Research	6	0.61	93.81	9	0.73	71.10
8	Claim 1: Reading	10	0.66	75.71	14–17	0.77	58.89
	Claim 2: Writing	6	0.70	73.37	7	0.72	68.65
	Claim 3: Listening	4	0.30	131.37	8–9	0.61	84.42
	Claim 4: Research	6	0.59	94.19	9	0.74	67.02
11	Claim 1: Reading	10	0.65	85.07	15–16	0.78	62.05
	Claim 2: Writing	6	0.71	77.51	7	0.73	71.48
	Claim 3: Listening	4	0.32	145.47	8–9	0.61	98.86
	Claim 4: Research	6	0.59	102.58	9	0.72	76.78

Table 29. Marginal Reliability and Average Conditional Standard Error of Measurement for Reporting Category: Mathematics

Grade	Claim	Hawai‘i Shortened Blueprint			Smarter Balanced Full Blueprint		
		Items	Reliability	Average CSEM	Items	Reliability	Average CSEM
3	Claim 1	12	0.84	41.61	17–20	0.90	28.52
	Claims 2 & 4	5	0.60	68.69	8–10	0.73	49.64
	Claim 3	5	0.58	72.17	8–10	0.74	50.61
4	Claim 1	12	0.84	41.05	17–20	0.90	28.21
	Claims 2 & 4	5	0.55	69.88	8–10	0.76	46.61
	Claim 3	5	0.62	67.85	8–10	0.74	50.06
5	Claim 1	12	0.83	45.83	17–20	0.89	31.54
	Claims 2 & 4	5	0.46	83.93	8–10	0.67	58.35
	Claim 3	5	0.56	86.24	8–10	0.72	59.59
6	Claim 1	12	0.81	55.77	16–19	0.90	37.19
	Claims 2 & 4	5	0.44	97.47	8–10	0.73	61.24
	Claim 3	5	0.46	103.31	8–10	0.74	61.00
7	Claim 1	12	0.78	61.50	17–20	0.88	39.61
	Claims 2 & 4	5	0.39	104.94	8–10	0.62	79.70
	Claim 3	5	0.46	106.07	8–10	0.64	78.20
8	Claim 1	12	0.77	66.93	17–20	0.88	44.35
	Claims 2 & 4	5	0.44	99.26	8–10	0.65	81.53
	Claim 3	5	0.39	121.12	8–10	0.71	77.87
11	Claim 1	14	0.80	57.37	19–22	0.89	42.08
	Claims 2 & 4	5	0.53	121.09	8–10	0.60	104.14
	Claim 3	5	0.48	125.60	8–10	0.59	96.72

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

3.3 STUDENT PERFORMANCE

The impact of the shortened blueprint on student performance was examined by comparing the performance between the Smarter Balanced full blueprint and the projected Hawaiʻi shortened blueprint for all students and by subgroups. The differences between the shortened blueprint and the full blueprint were examined in average scale scores, the effect size of the difference in average scale scores, and the percentage of students who met proficiency (percentage proficient). Cohen's *d* (Cohen, 1988) was used as the effect size to measure the difference between the two means.

$$Cohen's\ d = \frac{\overline{x}_1 - \overline{x}_2}{s} \text{ and } s = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$$

where $\overline{x}_1$ and $\overline{x}_2$ are the means of the two samples; and s_1 and s_2 are the standard deviations of the two samples.

In addition, the agreements between the four achievement levels (4 x 4) and between the two proficiency levels (Proficient or Not Proficient, 2 x 2) were examined.

3.3.1 ELA/L

Tables 30–32 present the differences in the average scale scores, the associated effect sizes, and the percentage proficient (Level 3 or 4) between the shortened blueprint and the full blueprint in ELA/L. For the effect sizes, Cohen suggested that d = 0.2 be considered a "small" effect size, 0.5 represents a "medium" effect size, and 0.8 a "large" effect size. This means that if the difference between two groups' means is less than 0.2 standard deviations, the difference is negligible, even if it is statistically significant.

In ELA/L, the effect sizes in all students and subgroups are negligible, ranging from -0.05 to 0.09. In general, the effect sizes were small in all students and subgroups in all grades. Although the effect sizes are negligible, all effect sizes are positive in all grades except for a few subgroups and are slightly larger in grades 7 and 11. Students in all grades, and particularly in upper grades, tend to perform slightly better on the PT component of the assessment relative to the CAT component. Because the PT component is a larger percentage of the entire test for the shortened blueprints, the percentages of proficiency are slightly higher with the shortened blueprint, with larger differences in the upper grades.

Nonetheless, the agreement between the four achievement levels (4 x 4) and proficiency (Proficient or Not Proficient, 2 x 2) between the shortened blueprint and the full blueprint, as shown in Table 33, were high for all students and subgroups. The scale score distributions for the full blueprint and the projected short blueprint, as shown in Figure 3 and Figure 4, were very similar with high correlations from 0.97 to 0.98.

Table 30. Student Performance for Overall and by Subgroup: ELA/L (Grades 3–5)

Subgroup	N	2018–2019 Projected Hawaiʻi Shortened Blueprint			2018–2019 Smarter Balanced Full Blueprint			Diff (Shortened – Full)		Effect Size (Cohen's d)
		Scale Score Mean	Scale Score SD	% Prof	Scale Score Mean	Scale Score SD	% Prof	Scale Score Mean	% Prof	
Grade 3										
All Students	14,364	2431.8	96.6	52.5	2431.8	92.8	52.4	0.0	0.1	0.00
Female	7,003	2441.8	93.5	56.6	2441.1	90.1	56.5	0.7	0.1	0.01
Male	7,361	2422.3	98.6	48.6	2423.0	94.5	48.4	-0.7	0.2	-0.01
African American	209	2432.6	86.9	56.9	2433.8	82.4	51.2	-1.2	5.7	-0.01
Asian/Pacific	3,417	2456.5	93.4	63.1	2455.3	90.3	62.9	1.2	0.2	0.01
Hawaiʻi Pacific	3,450	2389.5	92.0	33.1	2390.0	86.6	32.8	-0.5	0.3	-0.01
Hispanic	2,755	2420.0	94.6	47.7	2420.4	90.4	47.6	-0.4	0.1	0.00
White	1,698	2461.5	89.4	66.4	2462.7	87.2	66.6	-1.2	-0.2	-0.01
Multi-Racial	2,817	2447.2	93.2	59.3	2446.6	89.6	59.8	0.6	-0.5	0.01
ELL	1,808	2371.8	82.7	25.1	2372.5	77.3	25.3	-0.7	-0.2	-0.01
Disadvantaged	6,762	2400.7	93.5	38.5	2401.2	88.7	38.3	-0.5	0.2	-0.01
Migrant	181	2365.2	95.6	26.5	2369.8	88.1	27.6	-4.6	-1.1	-0.05
Disability	1,285	2323.7	82.0	9.7	2327.7	74.3	9.3	-4.0	0.4	-0.05
Grade 4										
All Students	11,341	2471.6	102.1	52.5	2469.3	99.8	51.3	2.3	1.2	0.02
Female	5,448	2484.6	99.6	57.3	2480.9	97.5	56.0	3.7	1.3	0.04
Male	5,893	2459.5	103.0	48.0	2458.6	100.7	47.0	0.9	1.0	0.01
African American	196	2460.0	88.6	46.4	2456.3	87.8	44.4	3.7	2.0	0.04
Asian/Pacific	2,627	2496.0	100.6	63.7	2494.4	98.3	62.8	1.6	0.9	0.02
Hawaiʻi Pacific	2,729	2429.6	96.8	33.9	2425.6	93.4	32.3	4.0	1.6	0.04
Hispanic	2,146	2459.7	98.2	46.9	2457.5	94.4	46.3	2.2	0.6	0.02
White	1,558	2505.3	94.4	67.3	2505.1	93.0	66.4	0.2	0.9	0.00
Multi-Racial	2,062	2483.5	101.2	57.7	2481.5	98.5	56.2	2.0	1.5	0.02
ELL	1,273	2390.5	84.3	18.6	2387.0	79.7	15.8	3.5	2.8	0.04
Disadvantaged	5,397	2441.2	98.8	39.7	2438.7	95.5	38.2	2.5	1.5	0.03
Migrant	153	2406.8	93.7	23.5	2403.6	86.0	21.6	3.2	1.9	0.04
Disability	1,168	2360.9	83.2	9.2	2361.8	77.1	8.4	-0.9	0.8	-0.01
Grade 5										
All Students	14,741	2515.4	100.8	57.4	2512.2	99.6	56.7	3.2	0.7	0.03
Female	7,162	2529.5	97.0	63.0	2525.2	95.5	61.7	4.3	1.3	0.04
Male	7,579	2502.1	102.5	52.2	2499.9	101.8	52.0	2.2	0.2	0.02
African American	240	2519.3	88.8	59.2	2515.7	87.5	57.1	3.6	2.1	0.04
Asian/Pacific	3,701	2540.3	97.1	67.3	2537.4	95.8	67.1	2.9	0.2	0.03
Hawaiʻi Pacific	3,604	2468.0	95.3	36.3	2464.6	94.5	35.4	3.4	0.9	0.04
Hispanic	2,598	2503.5	97.6	52.8	2499.8	96.3	52.2	3.7	0.6	0.04
White	1,782	2552.1	94.7	74.2	2550.4	92.0	73.8	1.7	0.4	0.02
Multi-Racial	2,795	2531.0	95.2	65.0	2527.2	94.3	63.8	3.8	1.2	0.04
ELL	1,312	2417.7	76.6	13.3	2412.8	74.2	12.0	4.9	1.3	0.06
Disadvantaged	6,880	2484.1	98.9	44.4	2480.3	97.8	43.3	3.8	1.1	0.04
Migrant	195	2456.3	94.1	32.3	2450.3	92.8	28.2	6.0	4.1	0.06
Disability	1,405	2393.0	79.2	8.8	2389.6	77.2	8.7	3.4	0.1	0.04

Table 31. Student Performance for Overall and by Subgroup: ELA/L (Grades 6–8)

Subgroup	**N**	**2018–2019 Projected Hawai‘i Shortened Blueprint**			**2018–2019 Smarter Balanced Full Test Blueprint**			**Diff (Shortened – Full)**		**Effect Size (Cohen’s d)**
		Scale Score Mean	**Scale Score SD**	**% Prof**	**Scale Score Mean**	**Scale Score SD**	**% Prof**	**Scale Score Mean**	**% Prof**	
Grade 6										
All Students	14,064	2532.2	99.6	53.5	2530.3	97.6	52.4	1.9	1.1	0.02
Female	6,814	2547.4	95.7	60.4	2545.4	94.2	59.2	2.0	1.2	0.02
Male	7,250	2517.9	101.1	47.1	2516.1	98.7	46.1	1.8	1.0	0.02
African American	210	2536.1	87.6	57.6	2533.0	87.2	56.2	3.1	1.4	0.04
Asian/Pacific	3,697	2555.6	94.2	64.0	2553.8	92.8	63.1	1.8	0.9	0.02
Hawai‘i Pacific	3,520	2484.9	95.4	33.5	2483.1	92.8	31.9	1.8	1.6	0.02
Hispanic	2,537	2522.2	96.9	49.4	2520.8	94.5	48.5	1.4	0.9	0.01
White	1,566	2573.0	93.0	71.1	2571.2	91.7	70.8	1.8	0.3	0.02
Multi-Racial	2,514	2548.3	94.0	59.0	2545.5	91.6	57.8	2.8	1.2	0.03
ELL	946	2421.2	81.2	9.2	2419.6	75.6	8.2	1.6	1.0	0.02
Disadvantaged	6,612	2500.4	97.0	40.1	2498.5	94.5	38.9	1.9	1.2	0.02
Migrant	213	2474.5	94.9	29.6	2472.8	90.8	27.7	1.7	1.9	0.02
Disability	1,412	2419.0	85.5	9.4	2418.0	82.0	8.4	1.0	1.0	0.01
Grade 7										
All Students	13,450	2556.5	106.0	55.5	2551.2	102.8	53.0	5.3	2.5	0.05
Female	6,401	2576.6	100.1	63.5	2569.7	98.1	60.8	6.9	2.7	0.07
Male	7,049	2538.2	107.8	48.2	2534.4	104.1	45.9	3.8	2.3	0.04
African American	212	2567.8	97.9	61.3	2562.6	96.3	59.4	5.2	1.9	0.05
Asian/Pacific	3,887	2583.1	98.9	66.6	2576.4	96.7	63.6	6.7	3.0	0.07
Hawai‘i Pacific	3,586	2506.7	101.0	35.1	2501.2	95.8	32.3	5.5	2.8	0.06
Hispanic	2,155	2545.9	101.1	51.1	2540.7	98.5	49.0	5.2	2.1	0.05
White	1,473	2599.5	100.2	72.8	2597.2	97.5	72.0	2.3	0.8	0.02
Multi-Racial	2,112	2572.6	102.0	61.6	2567.9	98.0	59.1	4.7	2.5	0.05
ELL	818	2443.9	90.1	12.2	2439.4	79.5	10.0	4.5	2.2	0.05
Disadvantaged	6,234	2524.1	104.2	42.1	2518.3	99.7	39.6	5.8	2.5	0.06
Migrant	174	2501.3	94.4	34.5	2494.5	93.3	31.0	6.8	3.5	0.07
Disability	1,264	2434.1	91.7	9.3	2433.6	82.2	8.1	0.5	1.2	0.01
Grade 8										
All Students	12,816	2569.2	108.0	53.0	2565.8	103.7	51.6	3.4	1.4	0.03
Female	6,169	2590.8	102.3	60.4	2585.8	98.4	59.0	5.0	1.4	0.05
Male	6,647	2549.3	109.2	46.1	2547.3	105.0	44.7	2.0	1.4	0.02
African American	232	2572.3	101.5	56.0	2569.1	100.3	53.0	3.2	3.0	0.03
Asian/Pacific	4,450	2594.8	103.5	62.5	2590.4	99.6	61.7	4.4	0.8	0.04
Hawai‘i Pacific	4,204	2524.8	101.5	35.4	2522.1	96.3	33.5	2.7	1.9	0.03
Hispanic	1,246	2563.2	104.1	51.2	2560.7	100.1	49.4	2.5	1.8	0.02
White	1,522	2605.3	104.3	68.1	2602.5	99.9	67.2	2.8	0.9	0.03
Multi-Racial	1,126	2591.4	102.9	61.6	2587.5	99.0	60.1	3.9	1.5	0.04
ELL	777	2457.8	81.9	8.4	2457.9	74.8	7.2	-0.1	1.2	0.00
Disadvantaged	5,658	2534.0	104.2	38.7	2531.1	99.5	36.7	2.9	2.0	0.03
Migrant	200	2488.0	104.2	25.5	2489.8	97.4	24.5	-1.8	1.0	-0.02
Disability	1,239	2444.2	85.4	7.5	2445.3	77.9	6.4	-1.1	1.1	-0.01

Table 32. Student Performance for Overall and by Subgroup: ELA/L (Grade 11)

Subgroup	N	2018–2019 Projected Hawai‘i Shortened Blueprint			2018–2019 Smarter Balanced Full Blueprint			Diff (Shortened – Full)		Effect Size (Cohen's d)
		Scale Score Mean	Scale Score SD	% Prof	Scale Score Mean	Scale Score SD	% Prof	Scale Score Mean	% Prof	
Grade 11										
All Students	10,647	2608.8	113.5	61.8	2602.2	112.0	59.3	6.6	2.5	0.06
Female	5,222	2628.4	103.3	69.2	2619.9	103.4	66.2	8.5	3.0	0.08
Male	5,425	2589.9	119.5	54.8	2585.1	117.2	52.6	4.8	2.2	0.04
African American	228	2602.3	107.6	59.2	2594.7	105.4	57.5	7.6	1.7	0.07
Asian/Pacific	4,177	2630.1	104.9	69.6	2623.4	104.2	67.0	6.7	2.6	0.06
Hawai‘i Pacific	3,119	2559.7	112.1	44.0	2552.4	108.0	40.9	7.3	3.1	0.07
Hispanic	863	2600.9	108.4	60.4	2593.7	107.5	57.0	7.2	3.4	0.07
White	1,278	2649.8	107.3	75.6	2645.2	107.2	74.0	4.6	1.6	0.04
Multi-Racial	951	2630.0	113.1	69.2	2623.5	111.6	67.6	6.5	1.6	0.06
ELL	594	2488.7	81.5	10.6	2481.7	75.7	8.6	7.0	2.0	0.09
Disadvantaged	3,914	2577.6	112.4	50.1	2570.6	109.8	46.9	7.0	3.2	0.06
Migrant	126	2557.3	110.6	44.4	2550.2	104.5	40.5	7.1	3.9	0.07
Disability	852	2473.5	99.5	14.1	2470.2	92.3	13.0	3.3	1.1	0.03

Table 33. Percentage of Agreements in Four Achievement Levels (4 x 4)
and Meets Proficiency (2 x 2): ELA/L

Subgroup	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 11	
	% Ach Level	% Prof	% Ach Level	% Prof	% Ach Level	% Prof	% Ach Level	% Prof	% Ach Level	% Prof	% Ach Level	% Prof	% Ach Level	% Prof
All Students	84.6	94.4	82.7	93.7	83.9	94.2	85.7	94.2	85.2	94.1	87.1	95.4	85.1	94.4
Female	84.0	94.1	82.6	93.8	83.7	94.4	85.5	94.3	84.3	93.8	86.2	95.5	84.7	94.7
Male	85.2	94.7	82.7	93.5	84.2	94	86	94.2	86.1	94.3	88.0	95.4	85.5	94.0
African American	84.2	93.3	81.1	89.8	82.5	94.6	83.8	91.9	86.8	95.3	88.8	96.1	83.3	93.0
Asian/Pacific	84.2	94.2	84.4	94.6	84.6	94.4	86.4	95.1	84.2	93.8	86.8	95.6	85.3	94.5
Hawai‘i Pacific	83.8	94.4	81.5	93.7	84.8	93.9	85.0	93.1	84.6	93.3	86.8	95.4	83.9	93.6
Hispanic	85.3	94.9	82.9	93.6	83.8	94.1	86.0	94.1	86.2	94.1	88.0	95.2	84.9	93.9
White	85.5	95.1	81.3	93.4	84.5	95.1	86.5	94.7	86.4	95.7	86.7	95.3	86.5	95.3
Multi-Racial	84.7	93.8	82.8	93.2	82.3	93.7	84.7	94.2	85.2	94.1	86.2	95.8	84.6	94.8
ELL	84.9	94.8	85.7	95.0	86.5	95.6	90.7	97.6	88.4	95.8	91.1	97.8	86.2	96.3
Disadvantaged	84.7	94.3	82.3	93.3	84.2	94.2	85.6	94.0	84.9	93.6	87.9	95.6	84.1	94.1
Migrant	89.0	95.6	82.4	91.5	82.6	93.8	88.7	94.4	82.8	93.1	91.0	97.0	85.7	96.0
Disability	91.3	96.8	90.2	96.8	91.0	97.1	91.3	97.5	90.3	97.4	91.4	98.1	89.7	95.7

Note: "% Ach Level" is the percentage of students with the same achievement level on both tests and "% Prof" is the percentage of students with the same proficiency status on both tests.

Figure 3. Scale Score Comparison Between Smarter Balanced Full Blueprint and Projected Hawai'i Shortened Blueprint: ELA/L

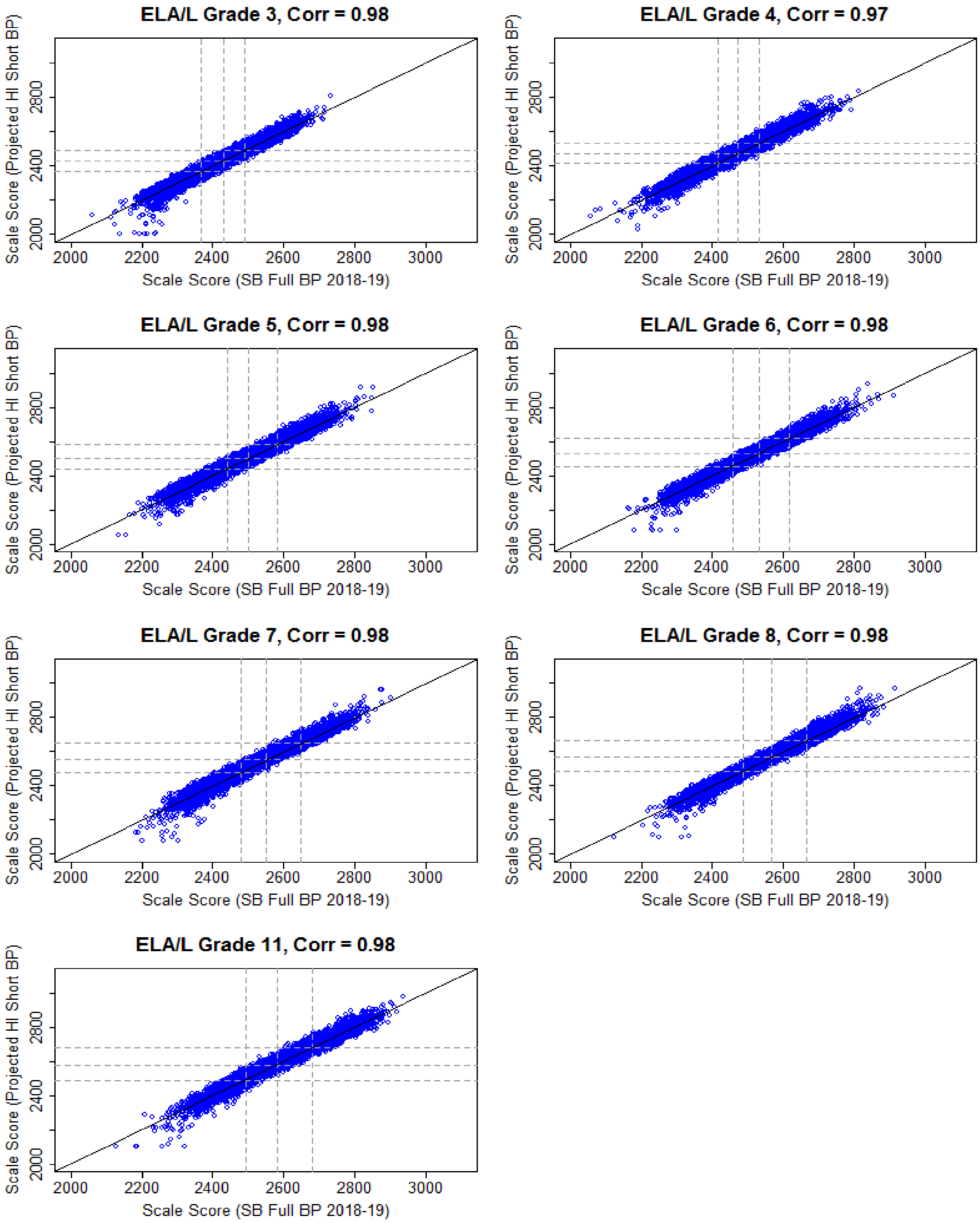
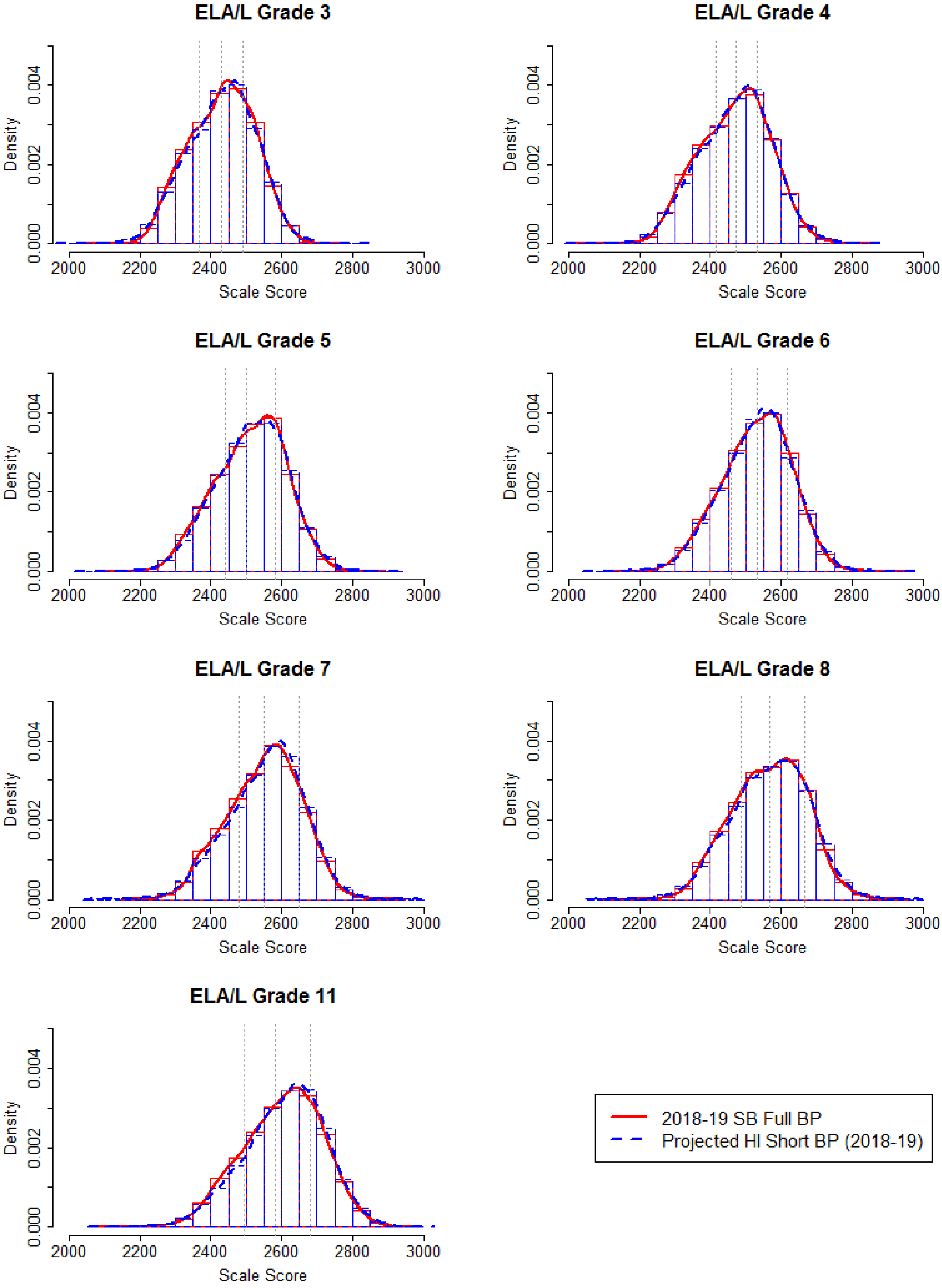


Figure 4. Scale Score Distribution for Smarter Balanced Full Blueprint and Projected Hawai'i Shortened Blueprint: ELA/L



3.3.2 Mathematics

Tables 34–36 present the differences in the average scale scores, the associated effect sizes, and the percentage proficient (Level 3 or 4) between the shortened blueprint and the full blueprint in mathematics. In mathematics, the effect sizes in all students and subgroups are negligible, smaller than ELA/L, ranging from -0.07 to 0.04, in all students and subgroups in all grades. The percentages of proficient were small in all grades, but the percentage proficient for English language learners was 1.4%–1.8% higher for the shortened blueprint in grades 4, 5, 8, and 11.

The agreement between the four achievement levels (4 x 4) and proficiency (Proficient or Not Proficient, 2 x 2) between the shortened blueprint and the full blueprint, as shown in Table 37, were high for all students and subgroups. The scale score distributions for the full blueprint and the projected short blueprint, as shown in Figure 5 and Figure 6, were very similar with high correlations from 0.97 to 0.98.

Table 34. Student Performance for Overall and by Subgroup: Mathematics (Grades 3–4)

Subgroup	N	2016–2017 Projected Hawai‘i Shortened Blueprint			2016–2017 Smarter Balanced Full Blueprint			Diff (Shortened – Full)		Effect Size (Cohen's d)
		Scale Score Mean	Scale Score SD	% Prof	Scale Score Mean	Scale Score SD	% Prof	Scale Score Mean	% Prof	
Grade 3										
All Students	14,824	2437.9	87.3	51.6	2438.1	84.1	52.6	-0.2	-1.0	0.00
Female	7,211	2437.8	83.8	51.7	2438.3	80.4	52.5	-0.5	-0.8	-0.01
Male	7,613	2438.0	90.4	51.5	2438.0	87.5	52.7	0.0	-1.2	0.00
African American	275	2427.2	77.6	44.0	2426.3	75.5	46.5	0.9	-2.5	0.01
Asian/Pacific	3,547	2464.4	83.7	64.1	2464.4	79.8	65.0	0.0	-0.9	0.00
Hawai‘i Pacific	3,627	2401.5	82.3	34.4	2401.6	79.9	34.8	-0.1	-0.4	0.00
Hispanic	2,672	2424.0	84.0	44.4	2424.8	80.4	44.7	-0.8	-0.3	-0.01
White	1,890	2462.1	85.3	64.2	2462.2	81.6	66.3	-0.1	-2.1	0.00
Multi-Racial	2,791	2449.5	84.4	57.1	2449.7	81.2	58.7	-0.2	-1.6	0.00
ELL	1,599	2386.5	77.4	24.7	2385.6	75.4	24.6	0.9	0.1	0.01
Disadvantaged	7,472	2412.9	83.9	39.5	2413.5	81.1	40.0	-0.6	-0.5	-0.01
Migrant	160	2388.5	79.2	26.9	2385.0	78.9	24.4	3.5	2.5	0.04
Disability	1,206	2338.1	86.7	11.4	2339.4	84.2	10.9	-1.3	0.5	-0.02
Grade 4										
All Students	14,690	2476.3	85.2	47.5	2476.8	83.1	48.0	-0.5	-0.5	0.00
Female	7,063	2477.7	80.1	47.5	2478.4	78.1	48.1	-0.7	-0.6	-0.01
Male	7,627	2475.1	89.6	47.5	2475.2	87.4	47.8	-0.1	-0.3	0.00
African American	249	2462.9	83.9	44.6	2464.9	81.2	44.6	-2.0	0.0	-0.02
Asian/Pacific	3,752	2501.7	81.3	59.9	2502.1	79.4	60.4	-0.4	-0.5	0.00
Hawai‘i Pacific	3,630	2440.2	81.6	29.6	2440.0	80.1	29.7	0.2	-0.1	0.00
Hispanic	2,612	2464.6	80.2	40.8	2465.7	78.3	41.3	-1.1	-0.5	-0.01
White	1,771	2501.3	82.3	60.4	2501.8	78.2	61.7	-0.5	-1.3	-0.01
Multi-Racial	2,653	2485.9	82.6	52.6	2486.4	80.0	53.0	-0.5	-0.4	-0.01
ELL	824	2395.5	79.6	11.9	2394.1	76.2	10.3	1.4	1.6	0.02
Disadvantaged	7,336	2451.7	82.3	35.4	2451.9	80.5	35.7	-0.2	-0.3	0.00
Migrant	160	2435.6	85.3	28.1	2435.8	83.9	26.3	-0.2	1.8	0.00
Disability	1,290	2381.9	83.5	10.2	2381.6	79.3	9.5	0.3	0.7	0.00

Table 35. Student Performance for Overall and by Subgroup: Mathematics (Grades 5–7)

Subgroup	**N**	**2016–2017 Projected Hawai‘i Shortened Blueprint**			**2016–2017 Smarter Balanced Full Blueprint**			**Diff (Shortened – Full)**		**Effect Size (Cohen's d)**
		Scale Score Mean	**Scale Score SD**	**% Prof**	**Scale Score Mean**	**Scale Score SD**	**% Prof**	**Scale Score Mean**	**% Prof**	
Grade 5										
All Students	14,495	2504.8	94.1	42.4	2505.2	89.9	42.2	-0.4	0.2	0.00
Female	6,995	2506.8	89.9	42.4	2507.5	85.6	42.2	-0.7	0.2	-0.01
Male	7,500	2503.0	97.9	42.4	2503.0	93.7	42.1	0.0	0.3	0.00
African American	248	2490.9	86.9	33.5	2490.1	79.2	30.6	0.8	2.9	0.01
Asian/Pacific	3,984	2533.7	91.1	54.9	2533.1	86.6	54.9	0.6	0.0	0.01
Hawai‘i Pacific	3,894	2464.2	88.9	24.9	2465.3	84.1	24.6	-1.1	0.3	-0.01
Hispanic	2,304	2491.2	89.4	35.7	2492.3	85.7	35.4	-1.1	0.3	-0.01
White	1,723	2529.7	89.5	54.0	2530.5	85.6	54.4	-0.8	-0.4	-0.01
Multi-Racial	2,309	2520.6	89.3	49.5	2520.5	86.1	48.9	0.1	0.6	0.00
ELL	657	2412.2	82.3	9.3	2412.0	77.5	7.5	0.2	1.8	0.00
Disadvantaged	7,158	2478.1	91.0	30.5	2478.7	86.3	29.9	-0.6	0.6	-0.01
Migrant	152	2454.0	86.2	22.4	2456.5	78.7	18.4	-2.5	4.0	-0.03
Disability	1,276	2400.3	79.5	6.3	2401.7	73.2	5.3	-1.4	1.0	-0.02
Grade 6										
All Students	13,795	2519.0	110.0	40.8	2520.0	106.2	40.9	-1.0	-0.1	-0.01
Female	6,610	2528.3	104.1	43.8	2529.1	100.9	43.9	-0.8	-0.1	-0.01
Male	7,185	2510.5	114.6	38.1	2511.6	110.1	38.0	-1.1	0.1	-0.01
African American	275	2512.2	110.1	38.9	2510.3	109.6	37.8	1.9	1.1	0.02
Asian/Pacific	4,514	2548.0	102.7	51.3	2548.5	99.2	51.7	-0.5	-0.4	-0.01
Hawai‘i Pacific	4,766	2475.9	107.5	24.8	2477.4	102.8	24.6	-1.5	0.2	-0.01
Hispanic	1,303	2506.7	108.1	35.9	2508.3	104.1	35.2	-1.6	0.7	-0.02
White	1,728	2557.2	98.6	55.5	2557.7	95.5	56.2	-0.5	-0.7	-0.01
Multi-Racial	1,158	2541.9	103.5	50.3	2543.1	99.2	50.1	-1.2	0.2	-0.01
ELL	699	2395.1	105.1	6.4	2396.9	99.2	6.3	-1.8	0.1	-0.02
Disadvantaged	6,816	2486.6	108.6	28.7	2487.8	104.0	28.0	-1.2	0.7	-0.01
Migrant	155	2451.5	98.9	13.5	2455.6	95.4	14.8	-4.1	-1.3	-0.04
Disability	1,334	2390.4	105.8	5.6	2392.4	97.6	4.9	-2.0	0.7	-0.02
Grade 7										
All Students	13,190	2522.5	114.8	36.4	2525.0	109.3	36.4	-2.5	0.0	-0.02
Female	6,307	2530.7	110.3	38.3	2533.1	105.6	38.6	-2.4	-0.3	-0.02
Male	6,883	2514.9	118.2	34.7	2517.5	112.1	34.4	-2.6	0.3	-0.02
African American	231	2515.3	99.7	29.9	2520.0	89.0	28.1	-4.7	1.8	-0.05
Asian/Pacific	4,447	2554.0	111.8	47.5	2556.8	106.4	48.0	-2.8	-0.5	-0.03
Hawai‘i Pacific	4,511	2476.0	107.4	20.4	2478.5	101.3	19.9	-2.5	0.5	-0.02
Hispanic	1,169	2506.0	111.6	29.9	2509.0	105.5	30.0	-3.0	-0.1	-0.03
White	1,676	2563.0	101.0	50.3	2563.6	95.8	50.2	-0.6	0.1	-0.01
Multi-Racial	1,120	2543.2	112.6	44.4	2546.1	107.1	44.7	-2.9	-0.3	-0.03
ELL	704	2409.0	103.8	7.2	2413.0	97.0	6.4	-4.0	0.8	-0.04
Disadvantaged	6,319	2487.8	111.6	24.4	2490.7	105.3	24.3	-2.9	0.1	-0.03
Migrant	154	2473.1	110.7	18.8	2475.8	101.2	18.8	-2.7	0.0	-0.03
Disability	1,295	2393.2	100.0	4.1	2399.4	90.1	3.9	-6.2	0.2	-0.06

Table 36. Student Performance for Overall and by Subgroup: Mathematics (Grades 8, 11)

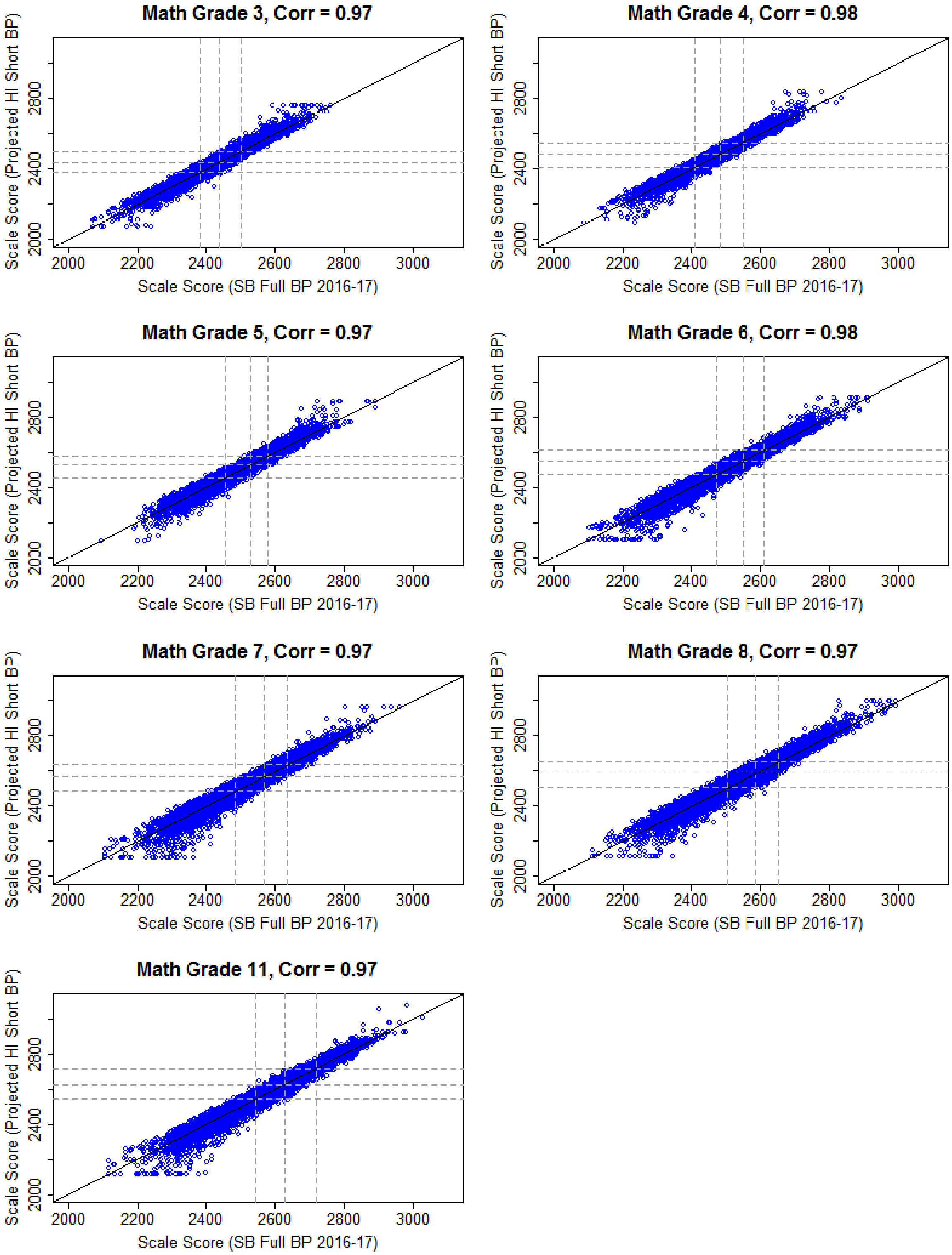
Subgroup	N	2016–2017 Projected Hawai‘i Shortened Blueprint			2016–2017 Smarter Balanced Full Blueprint			Diff (Shortened – Full)		Effect Size (Cohen’s d)
		Scale Score Mean	Scale Score SD	% Prof	Scale Score Mean	Scale Score SD	% Prof	Scale Score Mean	% Prof	
Grade 8										
All Students	12,510	2542.9	125.6	37.4	2543.7	120.6	37.4	-0.8	0.0	-0.01
Female	6,035	2556.3	121.7	41.6	2556.9	116.1	41.7	-0.6	-0.1	0.00
Male	6,475	2530.5	128.0	33.5	2531.4	123.5	33.4	-0.9	0.1	-0.01
African American	229	2539.9	106.9	34.5	2537.8	101.3	30.6	2.1	3.9	0.02
Asian/Pacific	4,498	2579.8	125.7	49.6	2580.3	120.7	49.7	-0.5	-0.1	0.00
Hawai‘i Pacific	4,067	2490.0	112.2	19.7	2491.5	106.1	19.4	-1.5	0.3	-0.01
Hispanic	1,077	2530.1	119.7	33.4	2530.0	114.9	33.4	0.1	0.0	0.00
White	1,502	2572.9	122.5	48.5	2573.4	118.1	49.3	-0.5	-0.8	0.00
Multi-Racial	1,101	2561.3	115.3	42.7	2561.8	111.5	42.9	-0.5	-0.2	0.00
ELL	692	2434.4	112.7	9.7	2435.4	104.3	7.9	-1.0	1.8	-0.01
Disadvantaged	5,729	2508.9	119.9	26.0	2509.4	114.4	25.7	-0.5	0.3	0.00
Migrant	137	2475.1	102.5	12.4	2475.0	101.7	16.8	0.1	-4.4	0.00
Disability	1,251	2401.8	100.0	3.9	2404.8	91.4	3.4	-3.0	0.5	-0.03
Grade 11										
All Students	10,550	2564.2	126.6	31.5	2565.9	120.5	31.1	-1.7	0.4	-0.01
Female	5,261	2575.9	117.6	33.9	2576.6	112.4	33.5	-0.7	0.4	-0.01
Male	5,289	2552.6	134.0	29.1	2555.2	127.2	28.7	-2.6	0.4	-0.02
African American	212	2551.8	115.8	25.9	2552.8	110.0	25.0	-1.0	0.9	-0.01
Asian/Pacific	4,381	2596.5	124.3	41.9	2597.2	118.8	41.2	-0.7	0.7	-0.01
Hawai‘i Pacific	3,024	2511.8	118.0	15.7	2514.6	109.2	15.3	-2.8	0.4	-0.03
Hispanic	783	2538.7	120.1	23.4	2541.1	114.9	22.6	-2.4	0.8	-0.02
White	1,247	2588.4	118.0	38.2	2590.7	114.0	38.5	-2.3	-0.3	-0.02
Multi-Racial	846	2577.5	123.9	33.9	2578.0	119.2	34.4	-0.5	-0.5	0.00
ELL	349	2428.4	106.6	4.0	2434.0	96.6	2.6	-5.6	1.4	-0.06
Disadvantaged	4,141	2536.1	125.4	23.7	2538.6	118.9	23.0	-2.5	0.7	-0.02
Migrant	91	2488.6	121.0	11.0	2492.1	109.2	13.2	-3.5	-2.2	-0.03
Disability	901	2410.4	107.6	2.1	2417.3	93.7	1.7	-6.9	0.4	-0.07

Table 37. Percentage of Agreements in Four Achievement Levels (4 x 4) and Meets Proficiency (2 x 2): Mathematics

Subgroup	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 11	
	% Ach Level	% Prof	% Ach Level	% Prof	% Ach Level	% Prof	% Ach Level	% Prof	% Ach Level	% Prof	% Ach Level	% Prof	% Ach Level	% Prof
All Students	82.7	93.4	84.9	94.0	83.6	94.3	84.7	94.0	84.5	94.2	83.3	94.2	85.7	94.2
Female	82.4	93.5	84.1	93.7	83.2	94.1	83.8	93.5	83.5	93.7	82.9	94.4	84.9	93.7
Male	83.0	93.3	85.6	94.4	84.1	94.5	85.6	94.4	85.4	94.6	83.6	94.1	86.5	94.7
African American	74.5	90.2	84.3	94.4	83.5	90.7	87.3	93.1	84.4	92.2	76.4	88.2	88.2	94.3
Asian/Pacific	82.8	93.5	84.4	93.7	83.5	94.3	84.2	93.8	83.5	93.9	82.4	94	84.9	93.4
Hawai‘i Pacific	82.6	93.1	86.0	95.0	84.7	95.4	86.0	94.8	86.0	95.3	85.6	95.7	87.9	95.5
Hispanic	82.3	93.0	83.7	93.2	83.7	94.3	83.7	93.5	84.3	93.5	85.4	95.5	86.5	95.4
White	84.6	94.7	85.0	94.1	81.4	92.7	82.7	93.1	82.2	92.4	81.9	92.8	82.4	92.9
Multi-Racial	82.7	93.3	85.1	94.1	83.8	93.9	84.9	93.7	85.5	94.1	79.3	91.6	84.9	93.4
ELL	83.6	93.3	86.7	96.2	89.0	97.0	94.7	98.4	92.8	98.0	91.6	98.0	94.3	98.0
Disadvantaged	82.2	93.2	84.6	94.1	84.3	94.9	85.4	94.6	85.7	95.1	84.3	95.0	87.0	94.9
Migrant	86.3	97.5	81.9	93.1	83.6	93.4	84.5	97.4	87.0	94.8	81.0	92.7	89.0	95.6
Disability	88.5	97.0	90.2	97.6	91.1	97.7	92.4	98.1	93.8	98.5	93.5	98.7	96.8	99.1

Note: "% Ach Level" is the percentage of students with the same achievement level on both tests, and "% Prof" is the percentage of students with the same proficiency status on both tests.

Figure 5. Scale Score Comparison Between Smarter Balanced Full Blueprint and Hawai‘i Shortened Blueprint: Mathematics



Math Grade 3, Corr = 0.97

Math Grade 4, Corr = 0.98

Math Grade 5, Corr = 0.97

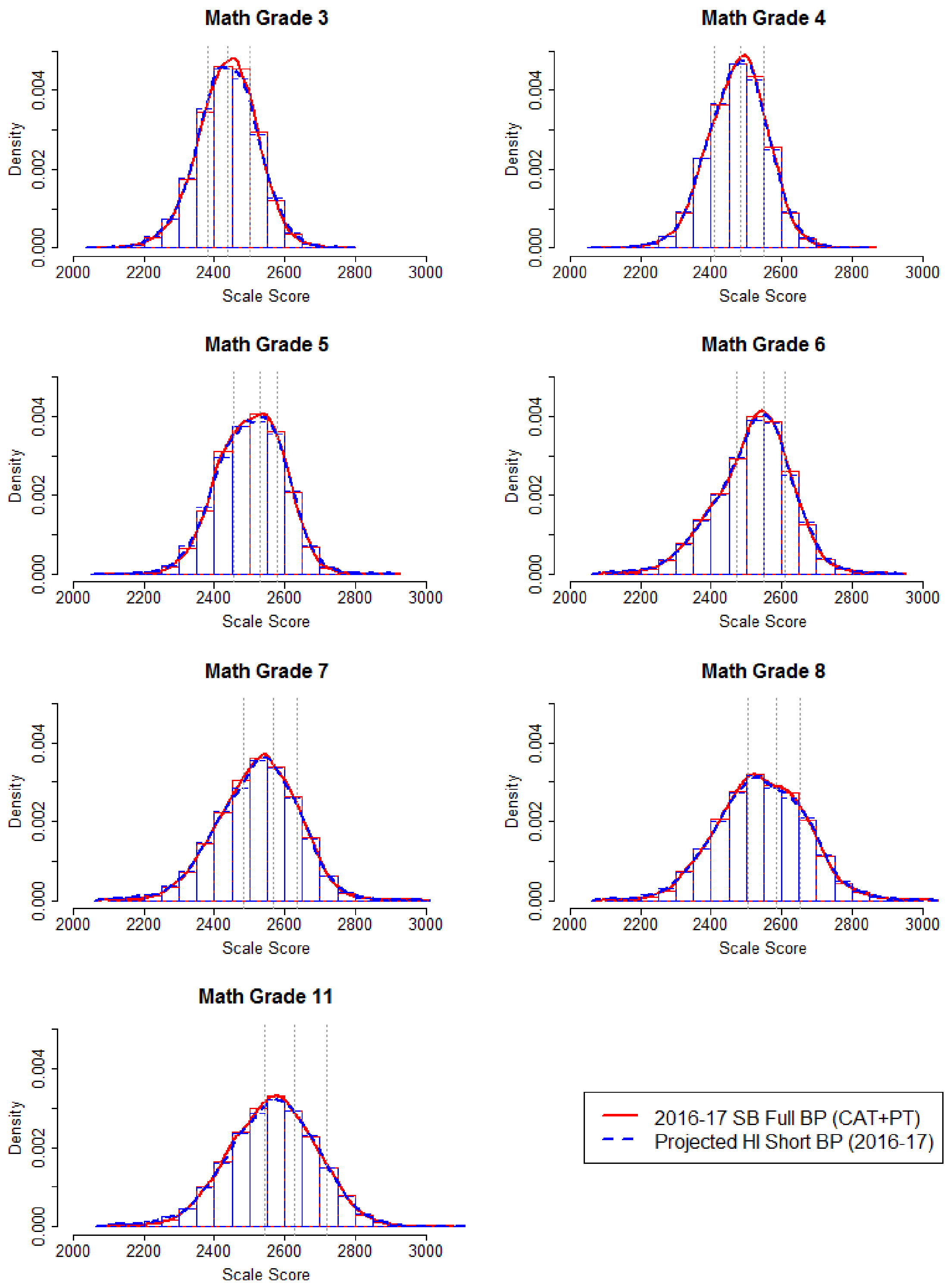
Math Grade 6, Corr = 0.98

Math Grade 7, Corr = 0.97

Math Grade 8, Corr = 0.97

Math Grade 11, Corr = 0.97

Figure 6. Scale Score Distribution for Smarter Balanced Full Blueprint and Hawai'i Shortened Blueprint: Mathematics



3.4 TESTING TIME

The major benefit of the shortened blueprint was the significant decrease in testing time. Overall testing time was greatly reduced for all grades. Tables 38–39 show the testing time for the shortened blueprint decreased by 47–98 minutes in ELA/L and 63–134 minutes in mathematics. The larger decrease in mathematics is because the shortened blueprint (BP) does not include the PT component, in addition to the shortened CAT component. In 2021–2022, students used fewer pauses during testing, i.e., finishing a test in one seating without taking pauses. Although the time in pauses is not included in computing the overall testing time, when students take more pauses, their testing time tends to be longer.

Table 38. Changes in Average Testing Time: ELA/L

Grade	2018–2019 Full BP			2021–2022 Hawai'i Shortened BP			Full BP – 2021–2022 Shortened BP
	Overall	CAT	PT	Overall	CAT	PT	Overall
3	3:49	1:44	2:05	2:32	0:54	1:39	1:17
4	4:02	1:48	2:14	2:49	0:57	1:52	1:13
5	4:26	1:58	2:27	2:48	0:58	1:50	1:38
6	4:09	1:50	2:19	2:50	1:05	1:45	1:19
7	3:43	1:37	2:06	2:37	1:00	1:37	1:06
8	3:41	1:38	2:03	2:35	0:59	1:36	1:06
11	2:44	1:18	1:26	1:57	0:47	1:10	0:47

Table 39. Changes in Average Testing Time: Mathematics

Grade	2016–2017 SB Full BP			2021–2022 Hawai'i Shortened BP	Full BP – 2021–2022 Shortened BP
	Overall (CAT + PT)	CAT	PT	Overall (CAT)	Overall
3	2:36	1:33	1:02	0:56	1:40
4	2:37	1:38	0:58	0:59	1:38
5	3:20	1:48	1:32	1:06	2:14
6	3:02	1:51	1:11	1:05	1:57
7	2:11	1:32	0:38	1:01	1:10
8	2:32	1:44	0:48	1:07	1:25
11	1:53	1:18	0:34	0:50	1:03

3.5 SUMMARY

The shortened blueprint specifies the same constraints in the full blueprint with items representing the breadth and depth of the test blueprints. Due to the reduction in the CAT length, the total number of targets covered at individual tests is 1–2 targets fewer than the full blueprint tests; however, all targets are covered at an aggregate level for both the Smarter Balanced full blueprint and the Hawaiʻi shortened blueprint tests.

The reliability coefficients for the shortened blueprints were high, ranging from 0.88–0.89 in ELA/L and 0.86–0.91 in mathematics. The classification of the proficiency cut (Level 3 or higher) was also high, 87%–89% for the consistency classification and 90%-92% for the accuracy classification, which is 1–3% lower than the classifications for the full blueprint tests.

The impact of the shortened blueprint on student performance was negligible in both ELA/L and mathematics. The agreement between the four achievement levels (4 x 4) and proficiency (Proficient or Not Proficient, 2 x 2) between the shortened blueprint and the full blueprint were high for all students and subgroups. The scale score distributions for the full blueprint and the projected short blueprint were highly correlated, from 0.97 to 0.98.

The major benefit of the shortened blueprint was the significant decrease in testing time. Overall testing time was greatly reduced for all grades, with a decrease in testing time by 47–98 minutes in ELA/L and 63–134 minutes in mathematics.

Overall, the results of the comparability of the shortened and full blueprints demonstrated that the shortened blueprint assessed student's progress with high test reliability with no significant impact on student performance while significantly reducing testing time.

4. SUMMARY OF 2021–2022 OPERATIONAL TEST ADMINISTRATION

4.1 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 2021–2022 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 40 and 41 present the participation rates and the percentage of students who attempted the test by subgroups. Tables 42 and 43 present the number of Hawai'i students who met attemptedness requirements for scoring and reporting the results of the Smarter Balanced summative assessments.

Table 40. Participation Rates by Percentage: ELA/L

Group	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 11
All Students	94.3	94.1	94.5	94.2	92.8	92.9	87.5
Female	94.3	93.9	94.8	94.3	93.1	93.2	87.4
Male	94.2	94.2	94.1	94.0	92.5	92.6	87.5
African American	95.0	95.7	98.8	95.5	94.6	95.0	87.8
AmerIndian/Alaskan	93.8	88.2	94.1	93.8	84.6	94.4	83.3
Asian/Pacific Islander	97.5	97.3	97.4	97.8	97.5	97.7	93.5
Hispanic	94.1	93.5	93.7	93.3	92.5	91.8	84.3
Hawai'i Pacific Islander	89.2	89.1	89.9	89.8	87.6	87.3	79.5
White	95.7	95.4	95.8	95.4	94.3	94.0	88.0
Multi-Racial	96.0	96.4	96.3	95.5	93.3	94.1	90.9
ELL	95.0	93.7	94.1	94.3	92.5	92.0	80.7
Disadvantaged	93.8	93.1	93.6	93.3	90.6	90.7	82.2
Migrant	94.2	95.4	92.6	94.1	91.2	93.0	82.8
Disability	86.7	86.6	85.6	87.1	83.0	80.7	66.2

Note: AmerIndian/Alaskan = American Indian/Alaskan Native

Table 41. Participation Rates by Percentage: Mathematics

Group	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 11
All Students	94.7	94.5	94.7	94.5	93.2	93.3	88.7
Female	94.7	94.2	95.1	94.6	93.5	93.5	88.9
Male	94.6	94.7	94.4	94.4	92.9	93.1	88.5
African American	95.0	95.7	98.8	96.0	93.2	95.0	89.4
AmerIndian/Alaskan	100.0	82.4	94.1	93.8	92.3	94.4	86.7
Asian/Pacific Islander	98.2	97.8	97.9	97.9	97.5	97.8	94.6
Hispanic	94.4	93.7	93.8	93.5	93.0	92.4	85.3
Hawai'i Pacific Islander	89.8	89.6	90.2	90.4	88.6	88.4	81.4
White	96.0	95.9	96.2	96.1	94.3	94.2	88.8
Multi-Racial	96.1	96.6	96.5	95.7	93.6	93.9	91.1
ELL	96.9	96.0	96.0	95.3	94.2	94.4	84.0
Disadvantaged	94.1	93.5	93.8	93.7	91.2	91.3	83.8
Migrant	94.8	95.4	91.3	94.6	92.9	92.5	83.5
Disability	87.1	87.2	85.8	87.2	83.2	81.5	68.3

Table 42. Number of Students: ELA/L

Group	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 11
All Students	12,991	12,819	13,058	12,841	9,922	12,456	10,033
Female	6,208	6,173	6,316	6,234	4,745	6,076	4,924
Male	6,783	6,646	6,742	6,607	5,177	6,380	5,109
African American	157	158	166	173	146	182	165
AmerIndian/Alaskan	15	15	16	16	13	17	26
Asian/Pacific Islander	2,969	2,964	3,221	3,296	2,498	3,475	4,024
Hispanic	2,576	2,493	2,495	2,395	1,909	2,202	986
Hawaiʻi Pacific Islander	2,983	2,987	3,081	3,143	2,458	2,955	2,716
White	1,428	1,441	1,507	1,407	1,183	1,383	1,157
Multi-Racial	2,863	2,761	2,572	2,411	1,715	2,242	959
ELL	1,790	1,655	1,460	1,411	1,107	1,198	549
Disadvantaged	5,776	5,646	5,681	5,748	4,454	5,439	3,499
Migrant	145	165	139	191	155	200	126
Disability	1,205	1,306	1,338	1,336	1,125	1,217	771

Table 43. Number of Students: Mathematics

Group	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 11
All Students	13,041	12,872	13,096	12,888	9,959	12,511	10,171
Female	6,231	6,190	6,336	6,255	4,761	6,101	4,999
Male	6,810	6,682	6,760	6,633	5,198	6,410	5,172
African American	157	159	165	174	143	182	168
AmerIndian/Alaskan	16	14	16	16	14	17	27
Asian/Pacific Islander	2,990	2,979	3,235	3,302	2,498	3,479	4,072
Hispanic	2,581	2,498	2,497	2,401	1,921	2,216	995
Hawaiʻi Pacific Islander	2,998	3,008	3,090	3,163	2,484	2,993	2,783
White	1,432	1,448	1,515	1,417	1,181	1,389	1,163
Multi-Racial	2,867	2,766	2,578	2,415	1,718	2,235	963
ELL	1,812	1,681	1,464	1,423	1,126	1,211	572
Disadvantaged	5,797	5,676	5,698	5,781	4,482	5,471	3,566
Migrant	146	165	137	192	158	199	124
Disability	1,212	1,321	1,336	1,340	1,128	1,231	790

4.2 SUMMARY OF OVERALL STUDENT PERFORMANCE

Tables 44–49 present a summary of the 2021–2022 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 7 and 8 present the percentage of proficient students over the past seven years for all students (cohort comparisons). Figures 9 and 10 present the average scale scores in seven years for all students. In Figures 7–10, the 2019–2020 performance is not included because the testing was canceled due to the COVID-19 pandemic.

Appendix C, 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 44. 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	12,991	2425.19	101.40	28	22	22	27	49
Female	6,208	2436.00	100.52	25	21	23	31	54
Male	6,783	2415.28	101.20	32	24	21	24	45
African American	157	2434.34	87.95	19	25	29	26	55
AmerIndian/Alaskan	15	2413.59	71.26	27	27	40	7	47
Asian/Pacific Islander	2,969	2457.40	95.96	17	21	24	38	62
Hispanic	2,576	2409.87	98.87	33	24	21	22	43
Hawai‘i Pacific Islander	2,983	2374.54	91.99	48	24	17	11	28
White	1,428	2455.10	94.85	18	19	25	38	63
Multi-Racial	2,863	2442.96	99.21	22	22	23	33	56
ELL	1,790	2373.77	92.77	48	24	17	11	28
Disadvantaged	5,776	2389.96	95.88	41	25	18	16	34
Migrant	145	2363.73	93.31	50	26	13	10	23
Disability	1,205	2318.99	82.74	73	17	6	3	9
Grade 4								
All Students	12,819	2470.92	103.20	30	19	23	29	51
Female	6,173	2482.27	100.63	26	19	24	32	56
Male	6,646	2460.38	104.44	34	19	22	26	48
African American	158	2452.24	94.21	34	27	18	21	39
AmerIndian/Alaskan	15	2459.94	93.53	27	27	20	27	47
Asian/Pacific Islander	2,964	2499.91	98.55	20	16	24	40	64
Hispanic	2,493	2455.49	100.04	35	20	22	23	45
Hawai‘i Pacific Islander	2,987	2424.32	95.43	47	21	19	13	32
White	1,441	2503.27	97.82	18	17	26	39	65
Multi-Racial	2,761	2488.40	101.31	23	18	24	34	58
ELL	1,655	2413.56	92.92	52	21	17	10	27
Disadvantaged	5,646	2437.25	97.73	42	21	21	17	38
Migrant	165	2416.76	96.14	52	18	16	13	30
Disability	1,306	2357.07	85.96	76	15	6	3	9
Grade 5								
All Students	13,058	2509.88	107.82	27	18	28	27	55
Female	6,316	2524.35	104.32	22	18	29	31	60
Male	6,742	2496.33	109.28	32	18	27	23	50
African American	166	2498.96	95.99	28	25	28	19	47
AmerIndian/Alaskan	16	2535.91	73.84	6	38	25	31	56
Asian/Pacific Islander	3,221	2542.73	103.18	17	17	28	39	67
Hispanic	2,495	2497.46	105.20	30	19	29	22	50
Hawai‘i Pacific Islander	3,081	2457.08	101.86	46	21	22	12	34
White	1,507	2545.01	94.36	14	15	34	37	71
Multi-Racial	2,572	2524.01	104.52	22	17	30	31	61
ELL	1,460	2428.87	91.62	55	23	18	4	23
Disadvantaged	5,681	2473.41	103.57	38	21	25	15	40
Migrant	139	2440.37	97.71	52	22	17	9	27
Disability	1,338	2392.25	90.72	72	17	9	3	12

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

Table 45. 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	12,841	2525.04	104.80	26	24	30	20	50
Female	6,234	2538.36	101.64	21	24	32	23	55
Male	6,607	2512.46	106.19	31	24	28	17	45
African American	173	2530.16	99.86	24	25	31	21	51
AmerIndian/Alaskan	16	2501.58	111.22	31	31	19	19	38
Asian/Pacific Islander	3,296	2553.32	102.31	18	21	33	28	61
Hispanic	2,395	2510.21	100.54	30	27	28	14	43
Hawai‘i Pacific Islander	3,143	2474.89	98.83	44	26	21	8	29
White	1,407	2565.96	92.43	12	20	36	31	67
Multi-Racial	2,411	2542.37	98.92	20	22	34	24	58
ELL	1,411	2435.64	81.62	61	26	12	1	13
Disadvantaged	5,748	2490.63	99.13	38	27	25	11	35
Migrant	191	2458.03	87.27	50	27	18	5	23
Disability	1,336	2408.44	83.10	73	19	7	1	8
Grade 7								
All Students	9,922	2548.91	108.29	25	23	34	18	52
Female	4,745	2563.13	104.65	21	22	36	21	57
Male	5,177	2535.87	109.92	30	23	32	15	47
African American	146	2558.72	95.24	16	27	42	14	56
AmerIndian/Alaskan	13	2604.01	99.35	8	15	46	31	77
Asian/Pacific Islander	2,498	2580.46	103.77	16	19	39	26	65
Hispanic	1,909	2534.42	106.54	29	25	31	14	45
Hawai‘i Pacific Islander	2,458	2497.28	100.96	41	28	25	6	31
White	1,183	2593.36	98.41	12	18	41	29	70
Multi-Racial	1,715	2561.15	101.85	21	23	36	20	57
ELL	1,107	2459.89	91.89	55	29	14	2	16
Disadvantaged	4,454	2514.97	105.21	36	26	29	10	38
Migrant	155	2485.57	96.57	43	32	22	3	25
Disability	1,125	2433.59	89.47	70	20	9	1	10
Grade 8								
All Students	12,456	2561.71	107.21	24	25	34	16	50
Female	6,076	2577.17	100.67	18	25	37	19	56
Male	6,380	2546.99	111.11	30	25	31	14	45
African American	182	2571.46	91.27	18	27	41	15	55
AmerIndian/Alaskan	17	2565.67	94.03	24	24	35	18	53
Asian/Pacific Islander	3,475	2595.18	101.74	15	21	40	24	65
Hispanic	2,202	2545.53	105.00	29	28	31	12	43
Hawai‘i Pacific Islander	2,955	2509.28	99.87	41	30	24	5	29
White	1,383	2593.60	99.50	14	23	40	24	64
Multi-Racial	2,242	2574.33	102.95	19	26	36	19	54
ELL	1,198	2476.26	87.93	53	31	15	1	16
Disadvantaged	5,439	2528.26	104.90	35	28	28	9	37
Migrant	200	2482.63	96.07	54	26	16	5	21
Disability	1,217	2439.03	90.23	72	20	7	1	8

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

Table 46. 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	10,033	2604.42	115.29	17	23	33	27	60
Female	4,924	2622.00	109.08	13	21	35	31	66
Male	5,109	2587.47	118.53	22	24	31	23	54
African American	165	2588.48	118.70	20	27	28	25	53
AmerIndian/Alaskan	26	2620.77	88.21	12	19	46	23	69
Asian/Pacific Islander	4,024	2630.72	106.82	11	20	36	34	69
Hispanic	986	2585.76	112.97	20	26	33	21	54
Hawaiʻi Pacific Islander	2,716	2556.16	111.56	29	29	28	14	42
White	1,157	2632.01	114.01	13	16	34	37	71
Multi-Racial	959	2618.93	116.27	16	19	33	32	65
ELL	549	2488.30	88.55	50	33	16	1	17
Disadvantaged	3,499	2571.24	114.68	25	28	30	18	47
Migrant	126	2547.18	114.86	34	28	25	13	38
Disability	771	2465.30	94.44	62	26	10	2	11

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

Table 47. 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	13,041	2435.11	94.91	27	22	26	25	51
Female	6,231	2433.19	91.57	28	22	26	24	50
Male	6,810	2436.88	97.85	27	21	26	26	52
African American	157	2435.19	80.45	22	29	29	20	49
AmerIndian/Alaskan	16	2413.14	76.90	19	38	38	6	44
Asian/Pacific Islander	2,990	2471.57	88.62	14	19	29	37	66
Hispanic	2,581	2419.47	91.66	33	24	25	19	44
Hawai'i Pacific Islander	2,998	2385.10	88.24	48	23	20	9	29
White	1,432	2461.49	85.60	16	21	28	35	63
Multi-Racial	2,867	2450.41	90.78	21	20	29	30	59
ELL	1,812	2393.63	94.97	44	23	20	13	33
Disadvantaged	5,797	2402.09	91.27	40	24	22	14	36
Migrant	146	2364.62	85.52	58	19	16	6	23
Disability	1,212	2338.40	89.39	70	15	11	4	15
Grade 4								
All Students	12,872	2472.36	92.57	25	29	25	20	46
Female	6,190	2469.16	88.31	26	30	25	18	44
Male	6,682	2475.32	96.26	25	27	25	23	48
African American	159	2454.37	74.54	27	40	23	10	33
AmerIndian/Alaskan	14	2445.66	102.75	29	43	14	14	29
Asian/Pacific Islander	2,979	2504.42	89.10	15	25	29	31	60
Hispanic	2,498	2455.37	87.80	31	31	23	15	38
Hawai'i Pacific Islander	3,008	2426.68	85.90	43	32	17	8	25
White	1,448	2502.18	87.09	14	25	31	29	60
Multi-Racial	2,766	2488.41	87.35	19	27	29	25	54
ELL	1,681	2424.60	87.14	45	31	15	8	24
Disadvantaged	5,676	2441.08	87.73	37	32	20	11	31
Migrant	165	2421.16	82.74	49	30	13	8	21
Disability	1,321	2375.15	84.11	70	20	7	3	10
Grade 5								
All Students	13,096	2501.03	100.61	33	26	18	23	42
Female	6,336	2500.13	96.61	32	27	18	22	40
Male	6,760	2501.87	104.21	33	24	18	25	43
African American	165	2482.25	84.25	39	36	12	13	25
AmerIndian/Alaskan	16	2505.84	59.56	25	25	50	0	50
Asian/Pacific Islander	3,235	2541.80	96.49	19	22	21	38	59
Hispanic	2,497	2482.28	93.74	39	29	16	16	32
Hawai'i Pacific Islander	3,090	2450.65	93.07	53	26	12	9	21
White	1,515	2529.28	90.42	19	26	23	31	54
Multi-Racial	2,578	2512.97	97.34	28	26	20	27	47
ELL	1,464	2434.20	90.95	60	24	11	5	16
Disadvantaged	5,698	2465.68	95.79	46	27	14	13	27
Migrant	137	2431.35	97.92	65	19	9	7	16
Disability	1,336	2400.52	87.92	76	16	4	4	7

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

Table 48. 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	12,888	2505.77	114.36	37	28	17	18	35
Female	6,255	2505.31	110.78	37	29	17	17	34
Male	6,633	2506.20	117.65	38	26	17	19	36
African American	174	2503.53	102.38	36	33	16	16	31
AmerIndian/Alaskan	16	2457.90	170.87	50	19	6	25	31
Asian/Pacific Islander	3,302	2543.30	108.85	25	27	21	27	48
Hispanic	2,401	2484.33	111.10	44	29	15	12	27
Hawaiʻi Pacific Islander	3,163	2450.56	107.55	57	26	11	6	17
White	1,417	2550.18	102.47	21	29	20	29	50
Multi-Racial	2,415	2522.51	106.19	31	29	20	21	40
ELL	1,423	2419.27	100.31	71	21	6	3	8
Disadvantaged	5,781	2468.48	109.44	51	26	13	10	22
Migrant	192	2427.37	106.06	65	24	8	3	11
Disability	1,340	2386.34	102.52	81	14	3	1	5
Grade 7								
All Students	9,959	2513.27	117.48	39	28	18	15	33
Female	4,761	2511.05	115.13	39	29	18	14	32
Male	5,198	2515.30	119.58	39	26	18	16	34
African American	143	2503.71	101.92	41	34	18	8	26
AmerIndian/Alaskan	14	2555.80	125.59	36	7	14	43	57
Asian/Pacific Islander	2,498	2556.89	116.86	26	26	22	26	49
Hispanic	1,921	2492.25	107.76	46	29	16	9	25
Hawaiʻi Pacific Islander	2,484	2456.55	106.48	59	26	10	5	15
White	1,181	2555.99	106.48	23	29	24	23	47
Multi-Racial	1,718	2526.41	112.56	33	29	21	16	38
ELL	1,126	2424.36	108.98	73	18	6	3	10
Disadvantaged	4,482	2477.36	112.15	52	27	13	8	21
Migrant	158	2449.79	97.50	61	28	9	3	11
Disability	1,128	2396.74	97.90	82	15	2	1	4
Grade 8								
All Students	12,511	2524.30	123.71	43	26	16	15	31
Female	6,101	2526.71	119.05	42	27	16	14	31
Male	6,410	2522.01	127.96	44	25	16	15	31
African American	182	2536.41	117.35	41	23	21	15	36
AmerIndian/Alaskan	17	2520.79	106.45	47	35	6	12	18
Asian/Pacific Islander	3,479	2570.96	122.36	29	25	20	25	45
Hispanic	2,216	2498.61	113.68	51	28	12	9	21
Hawaiʻi Pacific Islander	2,993	2463.29	109.87	65	23	8	4	13
White	1,389	2558.88	116.41	29	28	24	19	43
Multi-Racial	2,235	2536.41	117.74	38	28	18	16	34
ELL	1,211	2433.40	110.45	75	17	5	3	8
Disadvantaged	5,471	2486.52	117.97	56	25	11	8	19
Migrant	199	2458.21	108.39	69	20	7	4	11
Disability	1,231	2400.48	102.04	86	11	2	1	3

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

Table 49. 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	10,171	2550.90	120.01	46	28	17	8	26
Female	4,999	2555.46	113.36	44	29	19	7	26
Male	5,172	2546.49	125.97	48	26	16	9	25
African American	168	2529.33	119.11	49	35	9	7	16
AmerIndian/Alaskan	27	2543.56	100.69	59	26	4	11	15
Asian/Pacific Islander	4,072	2583.42	115.86	35	30	23	12	35
Hispanic	995	2526.20	113.11	57	26	13	5	18
Hawai'i Pacific Islander	2,783	2496.59	109.47	65	24	9	2	11
White	1,163	2575.74	115.90	37	30	23	11	34
Multi-Racial	963	2569.86	117.46	40	28	21	11	31
ELL	572	2463.52	103.71	79	16	5	1	6
Disadvantaged	3,566	2518.03	116.02	58	24	12	5	17
Migrant	124	2480.23	98.51	69	26	4	2	6
Disability	790	2412.50	93.65	92	6	2	0	2

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

Figure 7. Percentage Proficient Across Years: ELA/L

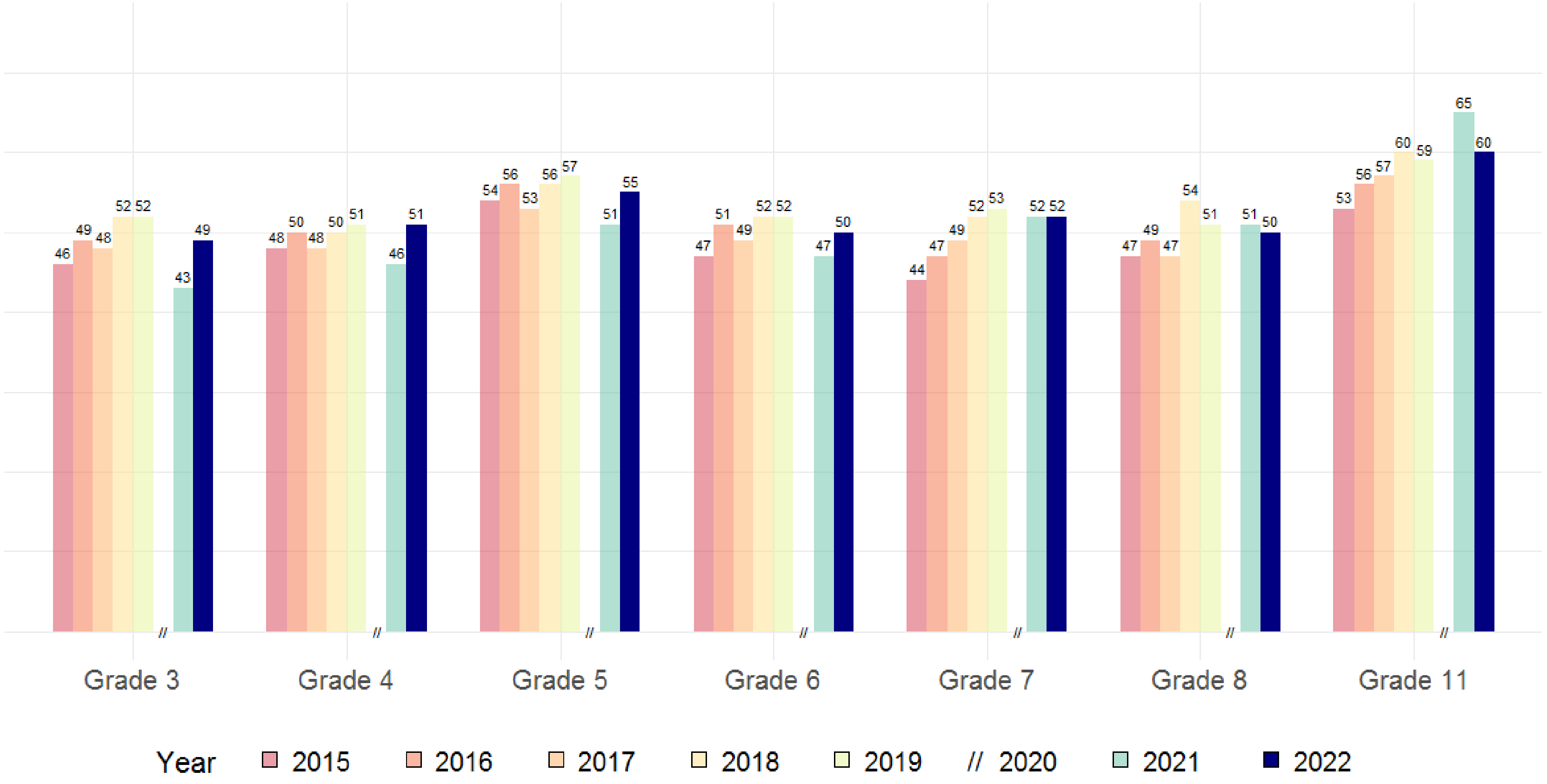


Figure 8. Percentage Proficient Across Years: Mathematics

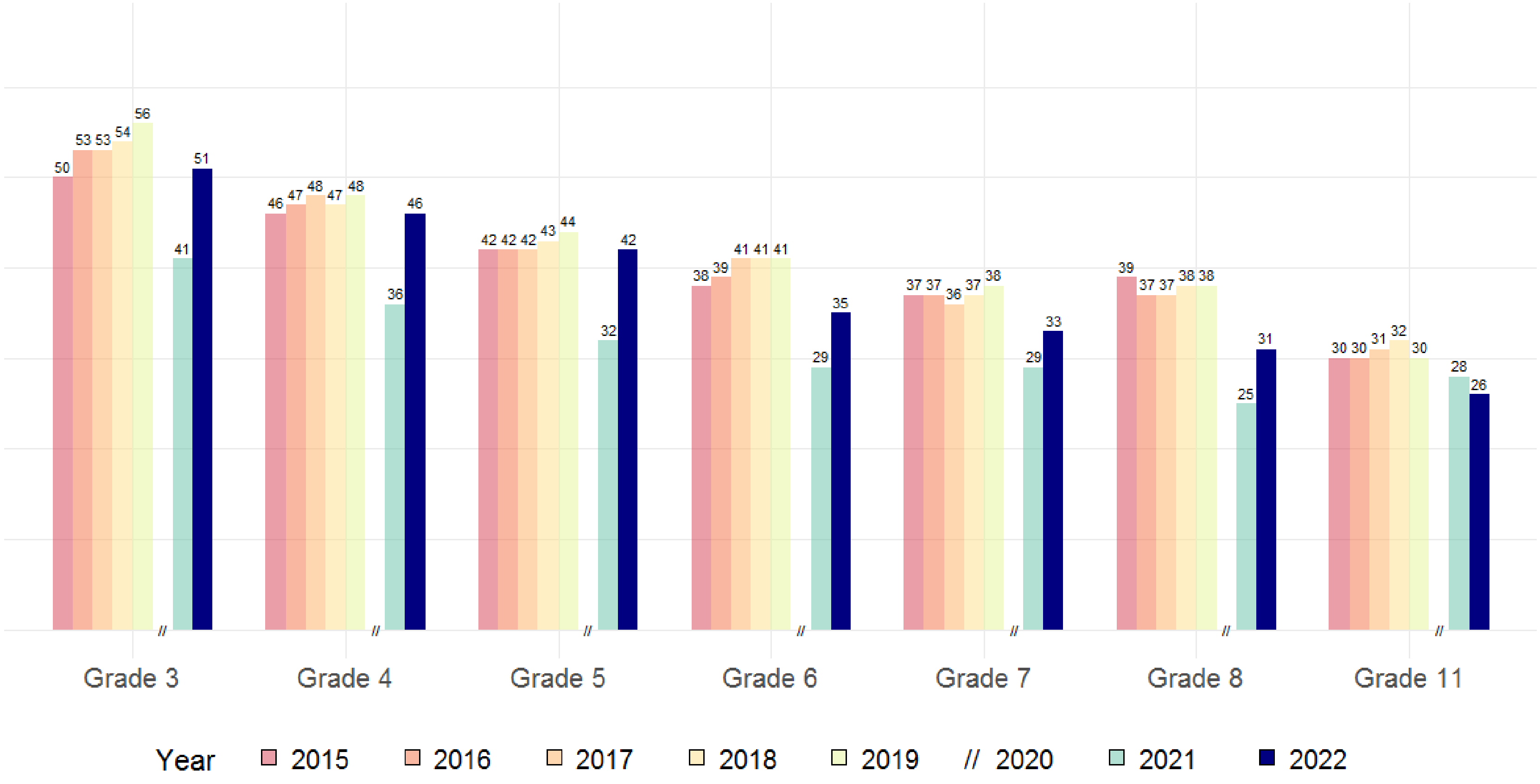


Figure 9. Average Scale Score Across Years: ELA/L

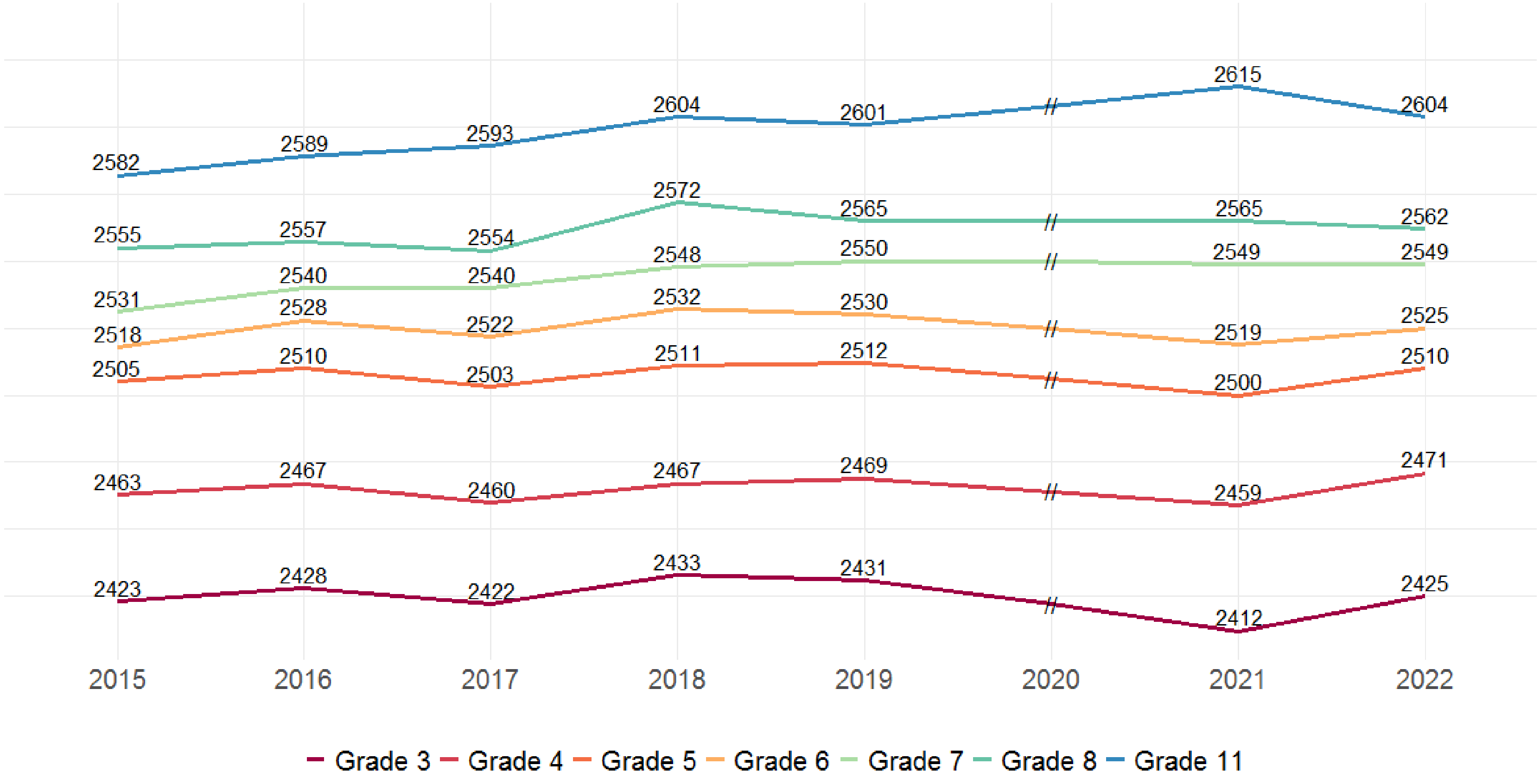
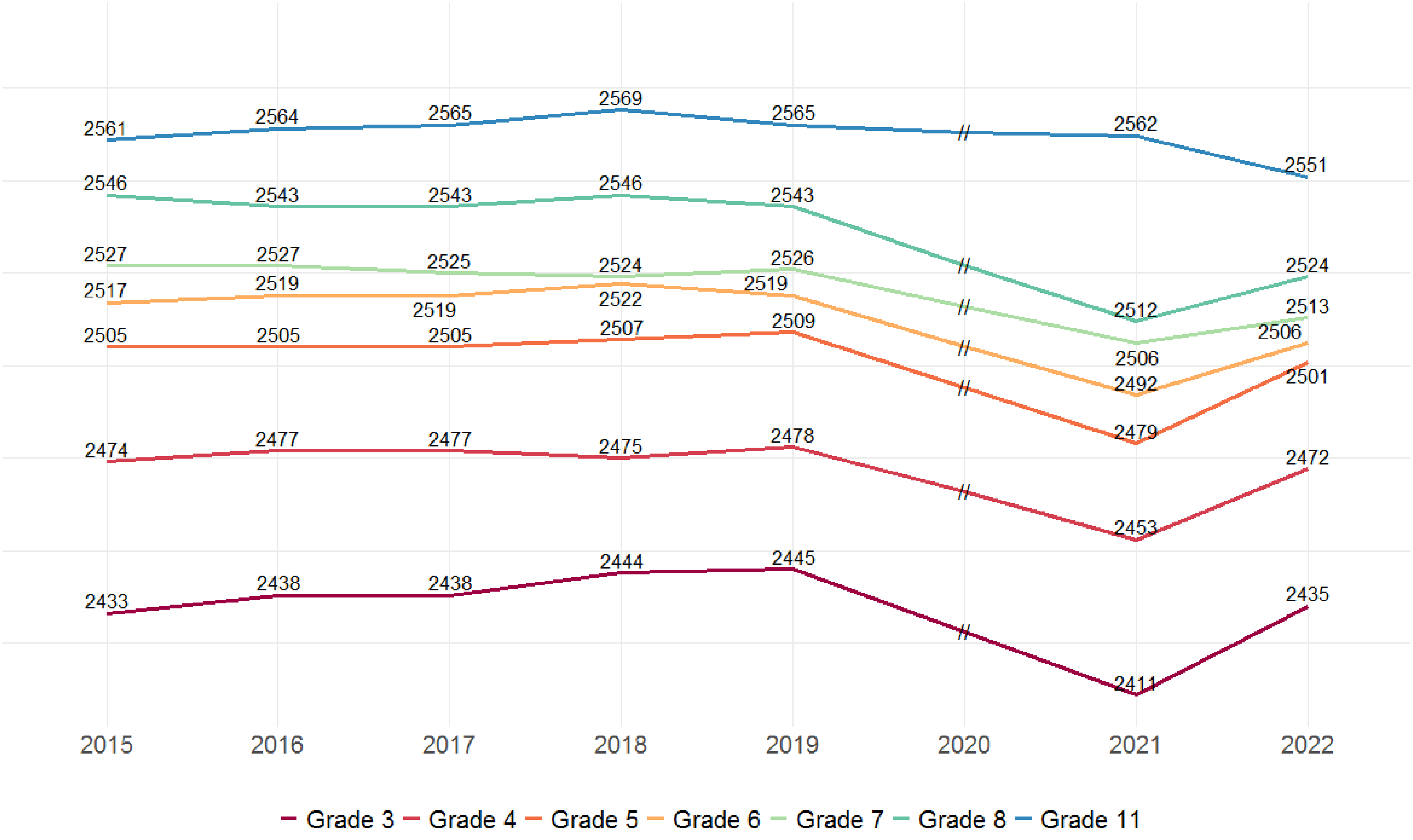


Figure 10. Average Scale Score Across Years: Mathematics



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 7.5, 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, in 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 50 presents the distribution of performance categories for the reported claims.

Table 50. 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	20	29	27
	At/Near	60	50	40
	Above	20	21	32
4	Below	18	25	32
	At/Near	62	53	40
	Above	20	21	29
5	Below	19	24	35
	At/Near	59	50	40
	Above	22	26	25
6	Below	28	29	42
	At/Near	54	52	39
	Above	18	19	19
7	Below	22	24	43
	At/Near	60	52	39
	Above	18	24	18
8	Below	25	26	42
	At/Near	56	56	43
	Above	19	18	15
11	Below	17	17	52
	At/Near	59	52	36
	Above	24	31	12

4.3 DISTRIBUTION OF STUDENT ABILITY AND ITEM DIFFICULTY

Figures 11–16 display the empirical distribution of the Hawaiʻi student scale scores in the 2021–2022 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 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 shifted to the left in Claims 1 (Reading) and 4 (Research) in upper grades for ELA/L. In mathematics, the student ability distribution 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 11. Student Ability—Item Difficulty Distribution: ELA/L

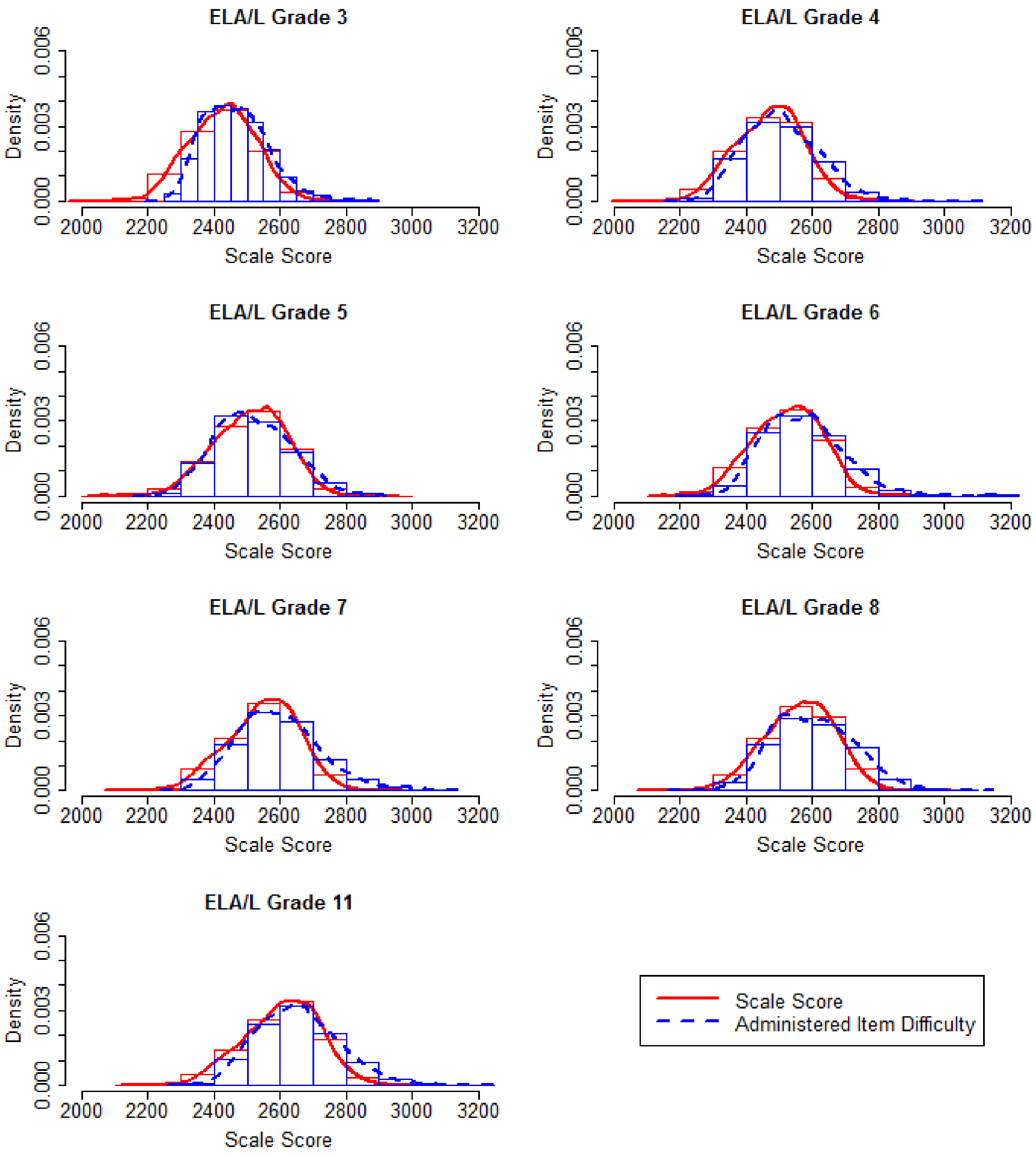


Figure 12. Student Ability—Item Difficulty Distribution by Claim: ELA/L (Grades 3–5)

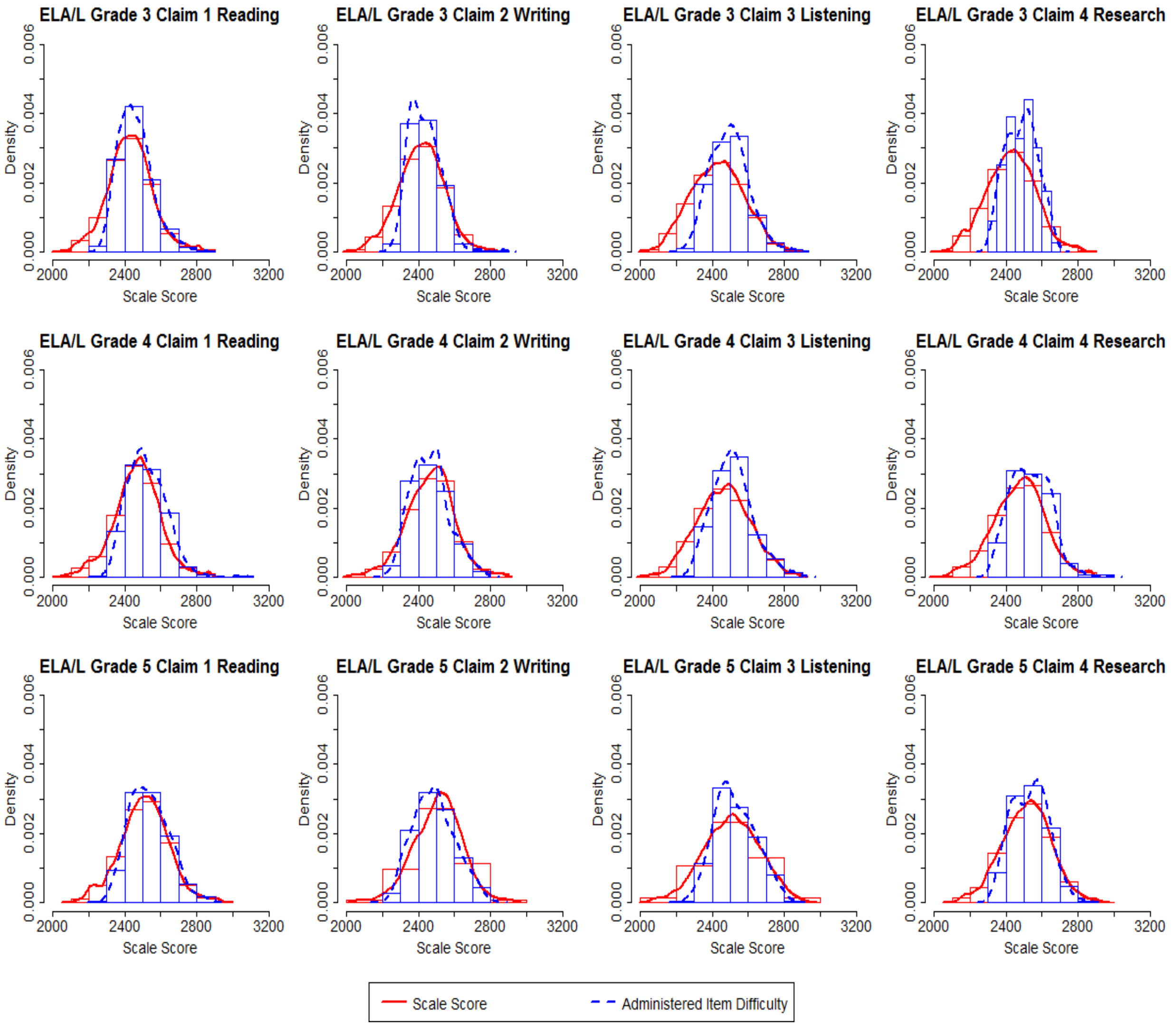
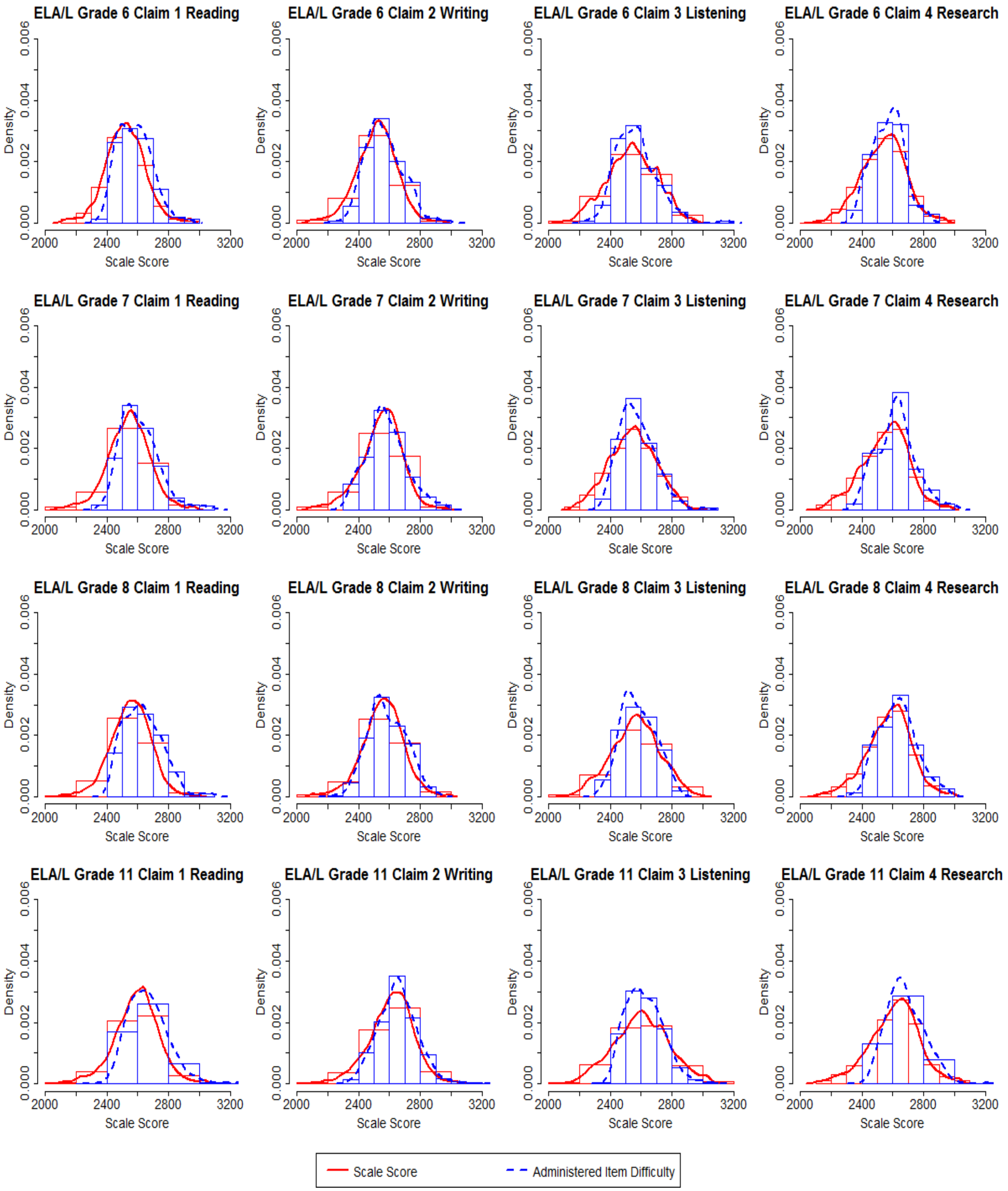


Figure 13. Student Ability—Item Difficulty Distribution by Claim: ELA/L (Grades 6–8, and 11)



ELA/L Grade 6 Claim 1 Reading

ELA/L Grade 6 Claim 2 Writing

ELA/L Grade 6 Claim 3 Listening

ELA/L Grade 6 Claim 4 Research

ELA/L Grade 7 Claim 1 Reading

ELA/L Grade 7 Claim 2 Writing

ELA/L Grade 7 Claim 3 Listening

ELA/L Grade 7 Claim 4 Research

ELA/L Grade 8 Claim 1 Reading

ELA/L Grade 8 Claim 2 Writing

ELA/L Grade 8 Claim 3 Listening

ELA/L Grade 8 Claim 4 Research

ELA/L Grade 11 Claim 1 Reading

ELA/L Grade 11 Claim 2 Writing

ELA/L Grade 11 Claim 3 Listening

ELA/L Grade 11 Claim 4 Research

Figure 14. Student Ability—Item Difficulty Distribution: Mathematics

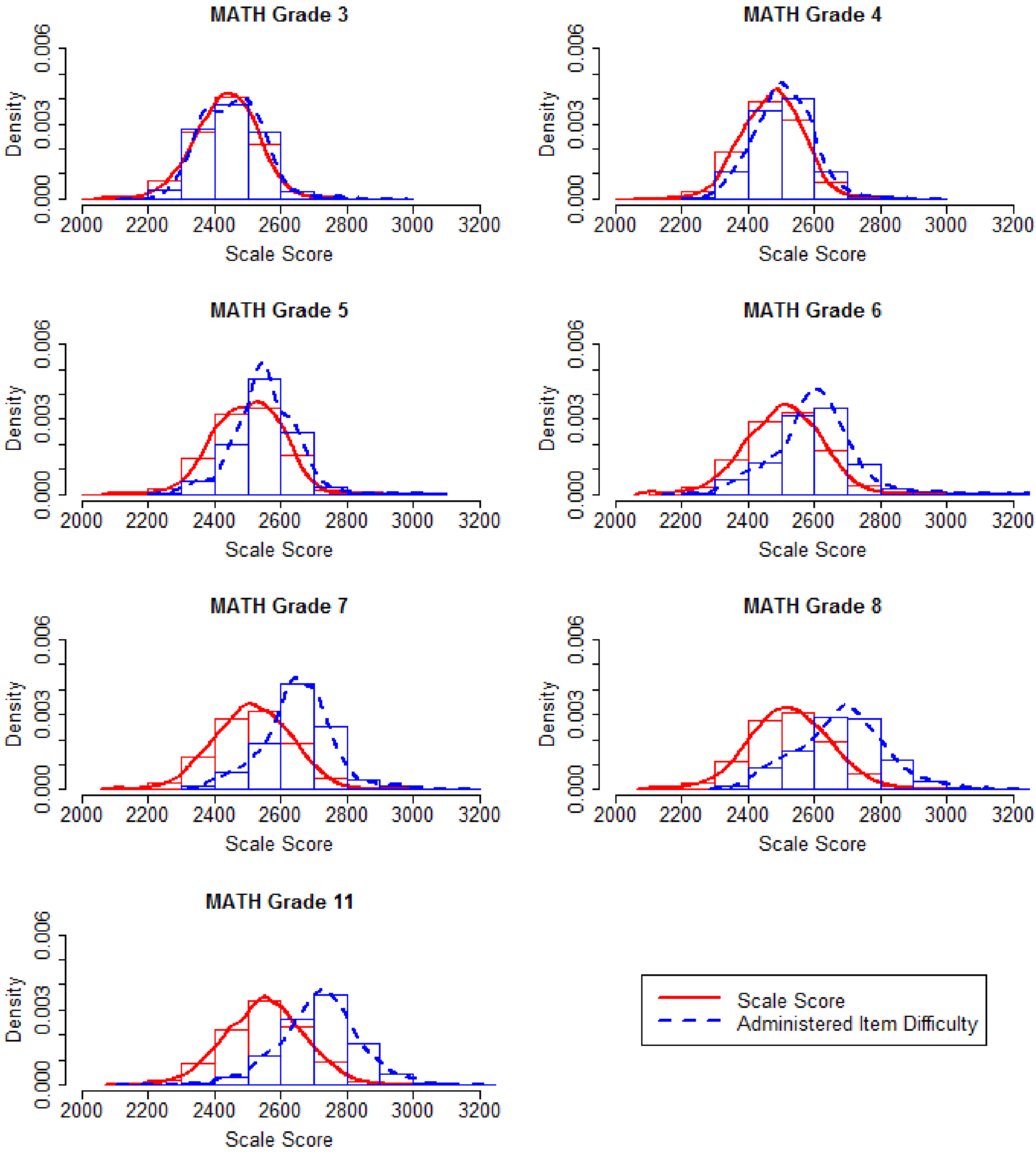


Figure 15. Student Ability—Item Difficulty Distribution by Claim: Mathematics (Grades 3–5)

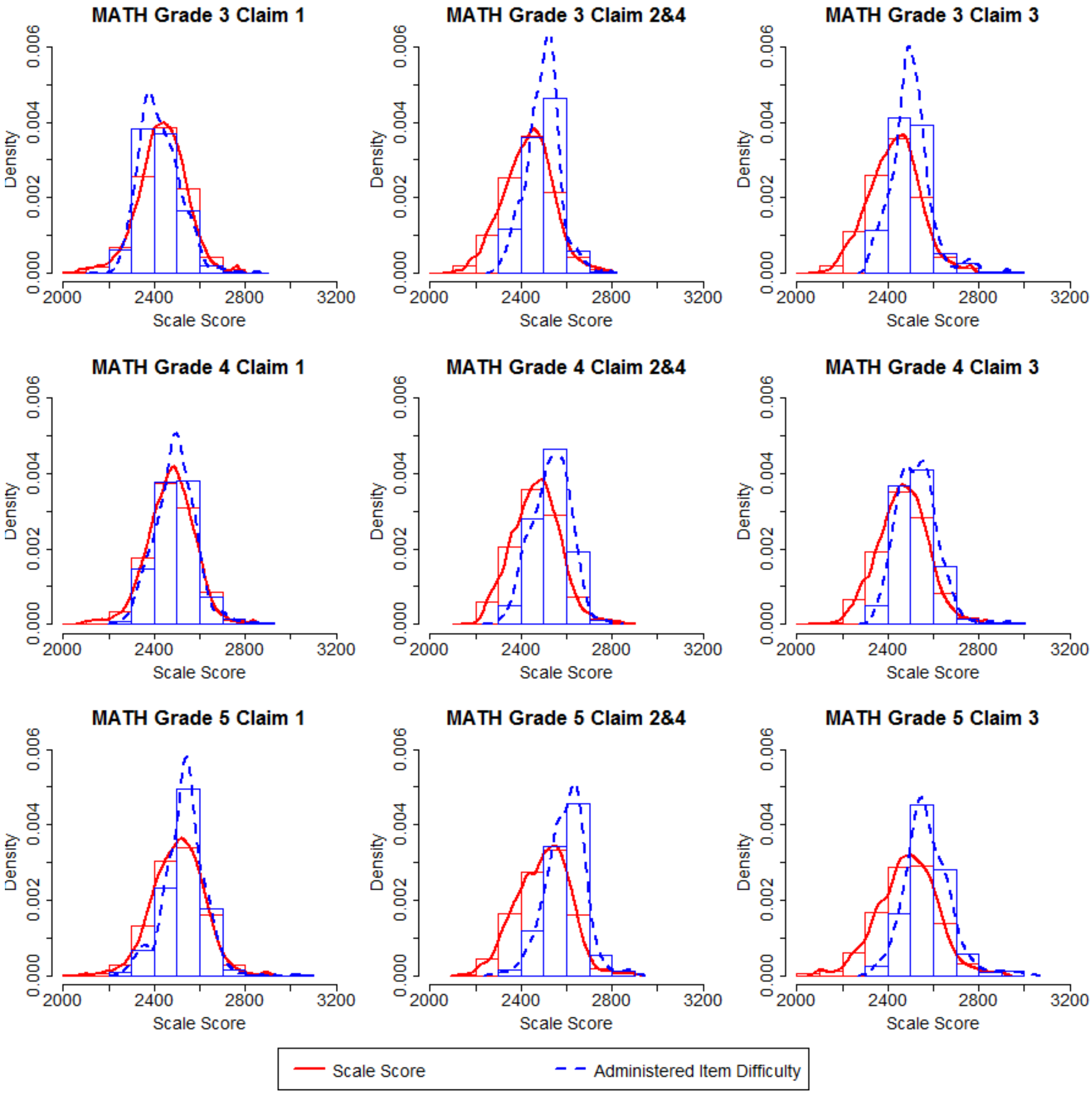
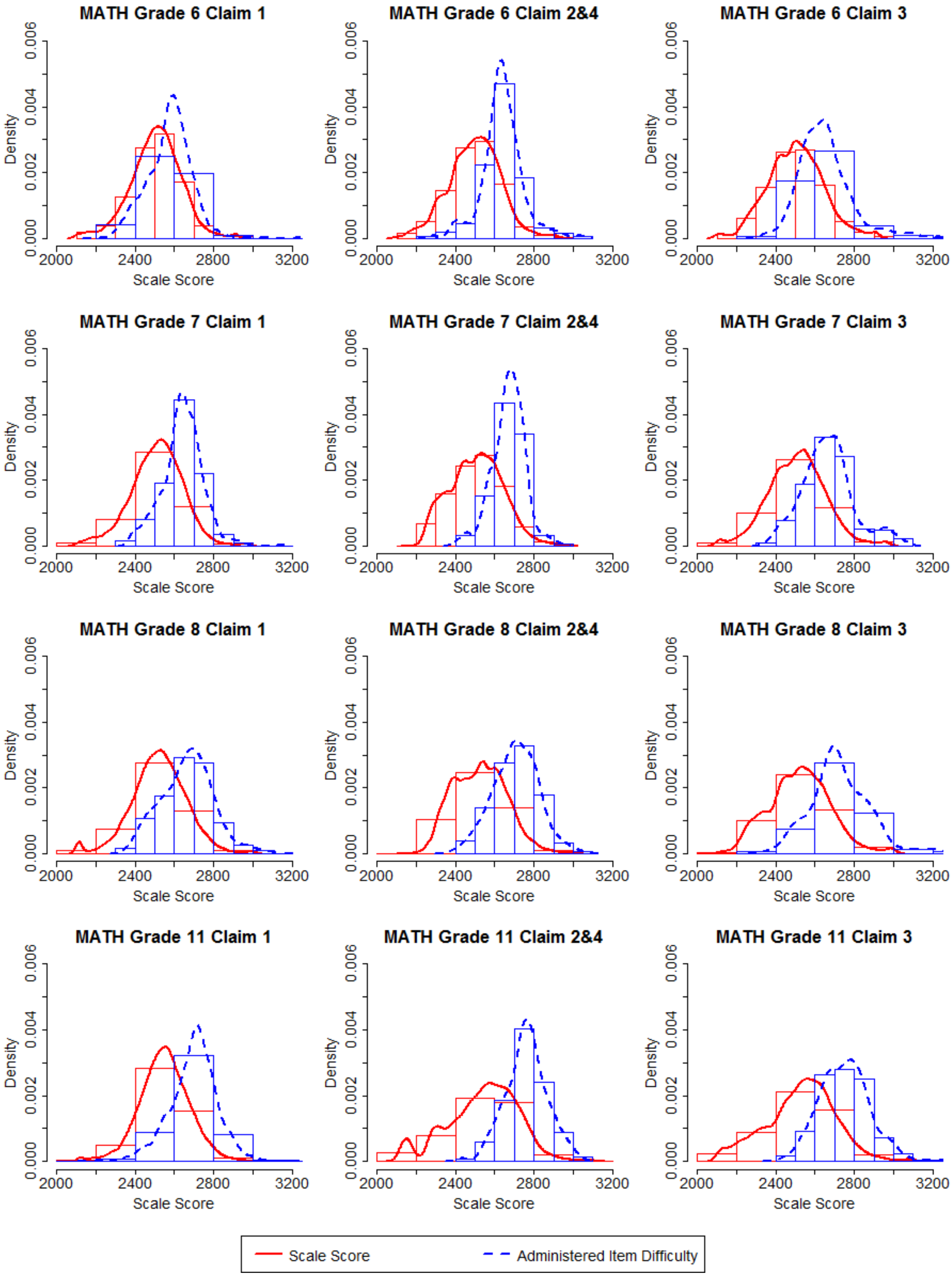


Figure 16. Student Ability—Item Difficulty Distribution by Claim: Mathematics (Grades 6–8, 11)



5. 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
- Relations to Other Variables (External Structure)

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

5.1 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.

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, few tests have violations in a few blueprint constraints.

Tables 51–55 present the percentages of tests aligned with the CAT blueprint constraints. All tests, except for a few tests, met all constraints. Few tests with blueprint violations are indicated in percentages smaller than 100. The blueprint violations were from the small pools with accommodations. The violations involved administering one item more or one or two items fewer than the blueprint requirements.

Tables 51 and 52 present the percentages of tests aligned with the ELA/L CAT test blueprint constraints for items in claims, targets, DOK, and number of passage requirement. Tables 53–55 provide the percentages of tests aligned with the test blueprint constraints for the mathematics CAT for claims, DOK, and target constraints.

Table 51. Percentage of 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	100	100
	Target 2: Central Ideas	1–3	100	100	100
	Target 4: Reasoning and Evaluation				
	Targets 1, 3, 5, 6, and 7	1–3	100	100	100
	Long Literary Text Passage	1	100	100	100
	Short Literary Text Passage				
	Informational Text	4	100	100	100
	Target 9: Central Ideas	1–3	100	100	100
	Target 11: Reasoning and Evaluation				
	Targets 8, 10, 12, 13, and 14	1–3	100	100	100
	Long Informational Text Passage	1	100	100	100
	Short Informational Text Passage				
	DOK 2	≥ 4	100	99.98	99.99
	DOK 3 or 4	≥ 1	100	100	100
2	**Writing**	5	100	100	100
	Target 1, 3, or 6: Organization/Purpose	1	100	100	100
	Target 1, 3, or 6: Evidence/Elaboration	1	100	100	100
	Target 8: Language and Vocabulary Use	1	100	100	100
	Target 9: Edit/Clarify	2	100	100	100
	DOK 2 or Higher	≥ 2	100	100	100
3	**Listening**	4	100	100	100
	Target 4: Listen/Interpret	4	100	100	100
	DOK 2 or Higher	≥ 2	100	100	100
	Listening Passage	2	100	100	100
4	**Research**	5	100	100	100
	Target 2: Interpret and Integrate Information	1–2	100	100	100
	Target 3: Analyze Information/Sources	1–2	100	100	100
	Target 4: Use Evidence	1–2	100	100	100

Table 52. Percentage of Delivered Tests Meeting Blueprint Requirements: ELA/L (Grades 6–8, 11)

Claim	Content Category/Targets	Required Items/Passages in Grades 6–8	Required Items/Passages in Grade 11	%BP Match			
				Grade 6	Grade 7	Grade 8	Grade 11
1	**Literary Text**	4	4	100	100	100	100
	Target 2: Central Ideas Target 4: Reasoning and Evaluation	1–3	1–3	100	100	100	100
	Targets 1, 3, 5, 6, and 7	1–3	1–3	100	100	100	100
	Long Literary Text Passage	1	1	100	100	100	100
	Informational Text	6	6	100	100	100	100
	Target 9: Central Ideas Target 11: Reasoning and Evaluation	2–4	2–4	100	99.99	100	100
	Targets 8, 10, 12, 13, and 14	2–4	2–4	100	99.99	100	100
	Long Informational Text Passage	1	1	100	100	100	100
	Short Informational Text Passage	1	1	100	100	100	100
	DOK 1	≤ 3	≤ 2	100	100	100	100
	DOK 3 or Higher	≥ 1	≥ 2	100	100	100	100
2	**Writing**	5	5	100	100	100	100
	Target 1, 3, or 6: Organization/Purpose	1	1	100	100	100	100
	Target 1, 3, or 6: Evidence/Elaboration	1	1	100	100	100	100
	Target 8: Language and Vocabulary Use	1	1	100	100	100	100
	Target 9: Edit/Clarify	2	2	100	100	100	100
	DOK 2	≥ 2	≥ 2	100	100	100	100
3	**Listening**	4	4	100	100	100	100
	Target 4: Listen/Interpret	4	4	100	100	100	100
	DOK 2 or Higher	≥ 2	≥ 2	100	100	100	100
	Listening Passage	2	2	100	100	100	100
4	**Research**	5	5	100	100	100	100
	Target 2: Analyze/Integrate Information	1–2	1–2	100	100	100	100
	Target 3: Evaluate Information/Sources	1–2	1–2	100	100	100	100
	Target 4: Use Evidence	1–2	1–2	100	100	100	100

Table 53. Percentage of Delivered Tests Meeting Blueprint Requirements for Claims and Targets: Mathematics (Grades 3–5)

Claim	Content / Target	Grade 3		Grade 4		Grade 5	
		Required Items	%Blueprint Match	Required Items	%Blueprint Match	Required Items	%Blueprint Match
1	Overall	12	100	12	100	12	100
	DOK 2 or Higher	≥ 4	100	≥ 4	100	≥ 4	100
	Priority Cluster	9	100				
	Targets B, C, G, I	4	100				
	Targets D, F	4	100				
	Target A	1	100				
	Supporting Cluster	3	100				
	Targets E, J, K	2	100				
	Target H	1	100				
	Priority Cluster			9	100		
	Targets A, E, F			5	100		
	Target G			2	100		
	Target D			1	100		
	Target H			1	100		
	Supporting Cluster			3	100		
	Targets I, K			1	100		
	Targets B, C, J			1	100		
	Target L			1	100		
	Priority Cluster					9	100
	Targets E, I					4	100
	Target F					3	100
	Targets C, D					2	100
	Supporting Cluster					3	100
	Targets J, K					2	100
	Targets A, B, G, H					1	100
2&4	Overall	5	100	5	100	5	99.96
	DOK 3 or Higher	≥ 2	100	≥ 2	99.91	≥ 2	99.99
	2. Target A	1	100	1	100	1	99.99
	2. Targets B, C, D	1	100	1	100	1	99.98
	4. Targets A, D	1	100	1	100	1	99.97
	4. Targets B, E	1	99.99	1	100	1	99.98
	4. Targets C, F	1	99.99	1	100	1	99.97
3	Overall	5	100	5	100	5	99.96
	DOK 3 or Higher	≥ 2	100	≥ 2	100	≥ 2	100
	Targets A, D	2	100	2	100	2	99.97
	Targets B, E	2	100	2	100	2	99.99
	Targets C, F	1	100	1	100	1	100

Table 54. Percentage of Delivered Tests Meeting Blueprint Requirements for Claims and Targets: Mathematics (Grades 6–8)

Claim	Content / Target	Grade 6		Grade 7		Grade 8	
		Required Items	%Blueprint Match	Required Items	%Blueprint Match	Required Items	%Blueprint Match
1	Overall	12	100	12	100	12	100
	DOK 2 or Higher	≥ 4	100	≥ 4	100	≥ 4	100
	Priority Cluster	9	100				
	Targets E, F	4	100				
	Target A	2	100				
	Targets G, B	2	100				
	Target D	1	100				
	Supporting Cluster	3	100				
	Targets C, H, I, J	3	100				
	Priority Cluster			9	99.46		
	Targets A, D			5	100		
	Targets B, C			4	99.46		
	Supporting Cluster			3	99.46		
	Targets E, F			2	99.40		
	Targets G, H, I			1	99.92		
	Priority Cluster					9	99.98
	Targets C, D					3	99.68
	Targets B, E, G					3	99.70
	Targets F, H					3	100
	Supporting Cluster					3	99.98
	Targets A, I, J					3	99.98
2&4	Overall	5	100	5	100	5	99.96
	DOK 3 or Higher	≥ 2	100	≥ 2	99.76	≥ 2	99.89
	2. Target A	1	100	1	100	1	100
	2. Targets B, C, D	1	100	1	99.99	1	100
	4. Targets A, D	1	100	1	100	1	99.77
	4. Targets B, E	1	100	1	99.98	1	99.91
	4. Targets C, F	1	100	1	99.99	1	99.90
3-Calc	Overall	4	100	5	100	5	99.96
	DOK 3 or Higher	≥ 2	100	≥ 2	100	≥ 2	100
	Targets A, D	1–2	100	2	100	2	99.96
	Targets B, E	1–2	100	2	100	2	100
	Targets C, F, G	0–1	100	1	100	1	100
3-No Calc	Overall	1	100				

Table 55. Percentage of Delivered Tests Meeting Blueprint Requirements for Claims and Targets: Mathematics (Grade 11)

Claim	**Content / Target**	**Grade 11**	
		Required Items	**%Blueprint Match**
1	Overall	14	100
	DOK 2 or Higher	≥ 4	100
	Priority Cluster	10	100
	Targets D, E	1–2	100
	Target F	1	100
	Targets G, H, I	3	100
	Target J	1–2	100
	Target K	1–2	100
	Targets L, M, N	2	100
	Supporting Cluster	4	100
	Target O	0–2	100
	Target P	0–2	100
	Targets A, B	0–1	99.98
	Target C	0–1	100
2&4	Overall	5	100
	DOK 3 or Higher	≥ 2	100
	2. Target A	1	100
	2. Targets B, C, D	1	100
	4. Targets A, D	1	100
	4. Targets B, E	1	100
	4. Targets C, F	1	100
3-Calc	Overall	4	100
	DOK 3 or Higher	≥ 2	100
	Targets A, D	1–2	100
	Targets B, E	1–2	100
	Targets C, F, G	0–1	100
3-No Calc	Overall	1	100

Table 56 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 56. Average and Range of the Number of Unique Targets Assessed Within Each Claim Across All Delivered CAT Components

Grade	**Total Targets in Blueprint**				**Mean**				**Range (Minimum–Maximum)**			
	C1	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**
ELA/L												
3	14	5	1	3	7.5	4.0	1.0	3.0	6–8	4–4	1–1	3–3
4	14	5	1	3	7.9	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	8.9	4.0	1.0	3.0	6-–0	4–4	1–1	3–3
7	14	5	1	3	9.2	4.0	1.0	3.0	8–10	4–4	1–1	3–3
8	14	5	1	3	9.0	4.0	1.0	3.0	7–10	4–4	1–1	3–3
11	14	5	1	3	8.3	4.0	1.0	3.0	5–10	4–4	1–1	3–3
Mathematics												
3	11	4	6	6	10.0	2.0	4.2	3.0	9–10	2–2	3–5	3–3
4	12	4	6	6	9.0	2.0	4.1	3.0	8–9	2–2	3–5	3–3
5	11	4	6	6	8.0	2.0	4.0	3.0	8–8	2–3	3–5	2–4
6	10	4	7	6	9.0	2.0	3.6	3.0	9–9	2–2	2–5	3–3
7	9	4	7	6	6.9	2.0	3.7	3.0	6–7	1–2	3–5	3–3
8	10	4	7	6	10.0	2.0	3.9	3.0	7–10	2–2	3–5	2–4
11	16	4	7	6	13.6	2.0	3.7	3.0	11–14	2–2	2–5	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.

5.2 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 57 and 58. 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 57. Correlations Among Claims: ELA/L

Grade	Claim	Observed and Disattenuated Correlation			
		Claim 1	Claim 2	Claim 3	Claim 4
3	Claim 1: Reading		0.91	1	0.95
	Claim 2: Writing	0.61		1	0.92
	Claim 3: Listening	0.50	0.50		1
	Claim 4: Research	0.58	0.61	0.49	
4	Claim 1: Reading		0.90	1	0.95
	Claim 2: Writing	0.58		1	0.89
	Claim 3: Listening	0.50	0.47		1
	Claim 4: Research	0.57	0.57	0.48	
5	Claim 1: Reading		0.90	1	0.96
	Claim 2: Writing	0.60		1	0.92
	Claim 3: Listening	0.52	0.51		1
	Claim 4: Research	0.60	0.63	0.52	
6	Claim 1: Reading		0.88	1	0.93
	Claim 2: Writing	0.62		1	0.91
	Claim 3: Listening	0.54	0.50		1
	Claim 4: Research	0.60	0.59	0.48	
7	Claim 1: Reading		0.87	1	0.94
	Claim 2: Writing	0.59		1	0.91
	Claim 3: Listening	0.52	0.49		1
	Claim 4: Research	0.59	0.60	0.49	
8	Claim 1: Reading		0.88	1	0.92
	Claim 2: Writing	0.60		1	0.92
	Claim 3: Listening	0.52	0.49		1
	Claim 4: Research	0.58	0.59	0.47	
11	Claim 1: Reading		0.86	1	0.92
	Claim 2: Writing	0.59		0.99	0.92
	Claim 3: Listening	0.51	0.47		1
	Claim 4: Research	0.57	0.59	0.46	

Table 58. Correlations Among Claims: Mathematics

Grade	Claim	Observed and Disattenuated Correlation		
		Claim 1	**Claims 2 & 4**	**Claim 3**
3	Claim 1		1	1
	Claims 2 & 4	0.75		1
	Claim 3	0.71	0.65	
4	Claim 1		1	1
	Claims 2 & 4	0.72		1
	Claim 3	0.74	0.66	
5	Claim 1		1	1
	Claims 2 & 4	0.70		1
	Claim 3	0.69	0.62	
6	Claim 1		1	1
	Claims 2 & 4	0.68		1
	Claim 3	0.67	0.58	
7	Claim 1		1	1
	Claims 2 & 4	0.67		1
	Claim 3	0.64	0.56	
8	Claim 1		1	1
	Claims 2 & 4	0.68		1
	Claim 3	0.60	0.56	
11	Claim 1		0.97	0.94
	Claims 2 & 4	0.63		0.96
	Claim 3	0.58	0.48	

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

5.3 EVIDENCE ON RELATIONS TO OTHER VARIABLES

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

Evidence for convergent and discriminant validity is determined by examining the patterns of correlations between Smarter Balanced assessments and performance on other tests. Observed correlations should be limited only by the unreliability of the measures.

When both assessments measure student achievement in common subject areas, as with, for example, test scores based on mathematics in the Smarter Balanced summative test and the Algebra I and Algebra II End-of-Course (EOC) tests, we expect test scores between the common subject-area assessments to be substantially correlated. In addition, we expect that the magnitude of observed correlations between test scores in different subject areas will be lower than correlations between test scores in a common subject area.

The relationship between the Smarter Balanced scores and the Algebra I and II scores was examined to evaluate the convergent and discriminant aspects of validity using grade 8 and grade 11 assessment data—Smarter Balanced mathematics and Hawai'i Algebra I and II EOC test scores for two different traits (contents) and the Smarter Balanced ELA/L. In examining the convergent and discriminant aspects of validity, Algebra I (grade 8) and II (grade 11) EOC test scores were considered.

It was expected that the correlation between the Smarter Balanced mathematics scores and the Algebra I and II scores for the same subject (convergent validity) would be moderate and higher than the correlation between Smarter Balanced ELA/L and Smarter Balanced mathematics (discriminant validity). That is, the correlation between two tests measuring the same content would be higher than the correlation between tests measuring different contents. For Algebra I and II EOC test, the scores would show a higher correlation with the Smarter Balanced mathematics scores than with the Smarter Balanced ELA/L scores (discriminant validity).

The results are provided in Table 59. In most scenarios, the results are as would be expected given the criteria set forth by Campbell and Fiske (1959), providing the validity evidence.

First, the reliability coefficients (numbers in boldface) were higher than the convergent and discriminant coefficients for all tests.

Second, the scores between similar traits measured by the different methods correlated more highly with each other than they did with different traits measured by the same method. This is the evidence needed for convergent validity (numbers underlined). For example, the correlation between the Smarter Balanced mathematics and Algebra I in grade 8 scores is 0.84. This is higher than the correlation between the Smarter Balanced ELA/L and Smarter Balanced mathematics scores (r = 0.61) and between the Smarter Balanced ELA/L and Hawai'i Algebra I EOC test scores (r = 0.62). The same pattern is shown in grade 11 Algebra II EOC scores. The correlation between the Smarter Balanced mathematics and Algebra II score is 0.69 which is higher than the correlation between the Smarter Balanced ELA/L and Smarter Balanced mathematics scores (r = 0.54) and between the Smarter Balanced ELA/L and Hawai'i Algebra II EOC test scores (r = 0.50).

Last, the correlations of scores between different traits are lower than the correlations between similar traits. This is the evidence needed for discriminant validity (numbers in a rectangle). The correlations between the Smarter Balanced ELA/L scores and the Smarter Balanced mathematics and Algebra I and II EOC test scores in a rectangle are lower than the underlined correlations.

Overall, the observed pattern of correlations in each multitrait-multimethod matrix conforms to the criteria expected for convergent and discriminant validity.

Table 59. Relationship Among the Smarter Balanced, Algebra I, and Algebra II Test Scores

Test/Subject	SB ELA/L	SB Mathematics	EOC Algebra
Grade 8 (N = 1,497)			
SB ELA/L	**0.78**		
SB Mathematics	0.61	**0.86**	
Algebra I	0.62	0.84	**0.91**
Grade 11 (N = 607)			
SB ELA/L	**0.82**		
SB Mathematics	0.54	**0.79**	
Algebra II	0.50	0.69	**0.84**

6. 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.

6.1 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 ($\bar{\rho}$) is defined as

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

where N is the number of students, $CSEM_i$ is the CSEM of the scale score for student i, and $\sigma^2$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

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

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

Table 60 presents the marginal reliability coefficients and the average CSEM for the total scale scores.

Table 60. 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	12,991	24	0.89	2425.19	101.40	33.78
4	12,819	24	0.88	2470.92	103.20	36.04
5	13,058	24	0.89	2509.88	107.82	35.33
6	12,841	26	0.89	2525.04	104.80	34.91
7	9,922	26	0.88	2548.91	108.29	36.98
8	12,456	26	0.88	2561.71	107.21	36.91
11	10,033	26	0.88	2604.42	115.29	40.69
Mathematics						
3	13,041	22	0.91	2435.11	94.91	28.25
4	12,872	22	0.91	2472.36	92.57	27.65
5	13,096	22	0.90	2501.03	100.61	31.80
6	12,888	22	0.88	2505.77	114.36	39.32
7	9,959	22	0.87	2513.27	117.48	42.47
8	12,511	22	0.86	2524.30	123.71	46.80
11	10,171	24	0.87	2550.90	120.01	43.97

6.2 STANDARD ERROR CURVES

Figures 17 and 18 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 17 and 18. 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 17. Conditional Standard Error of Measurement: ELA/L

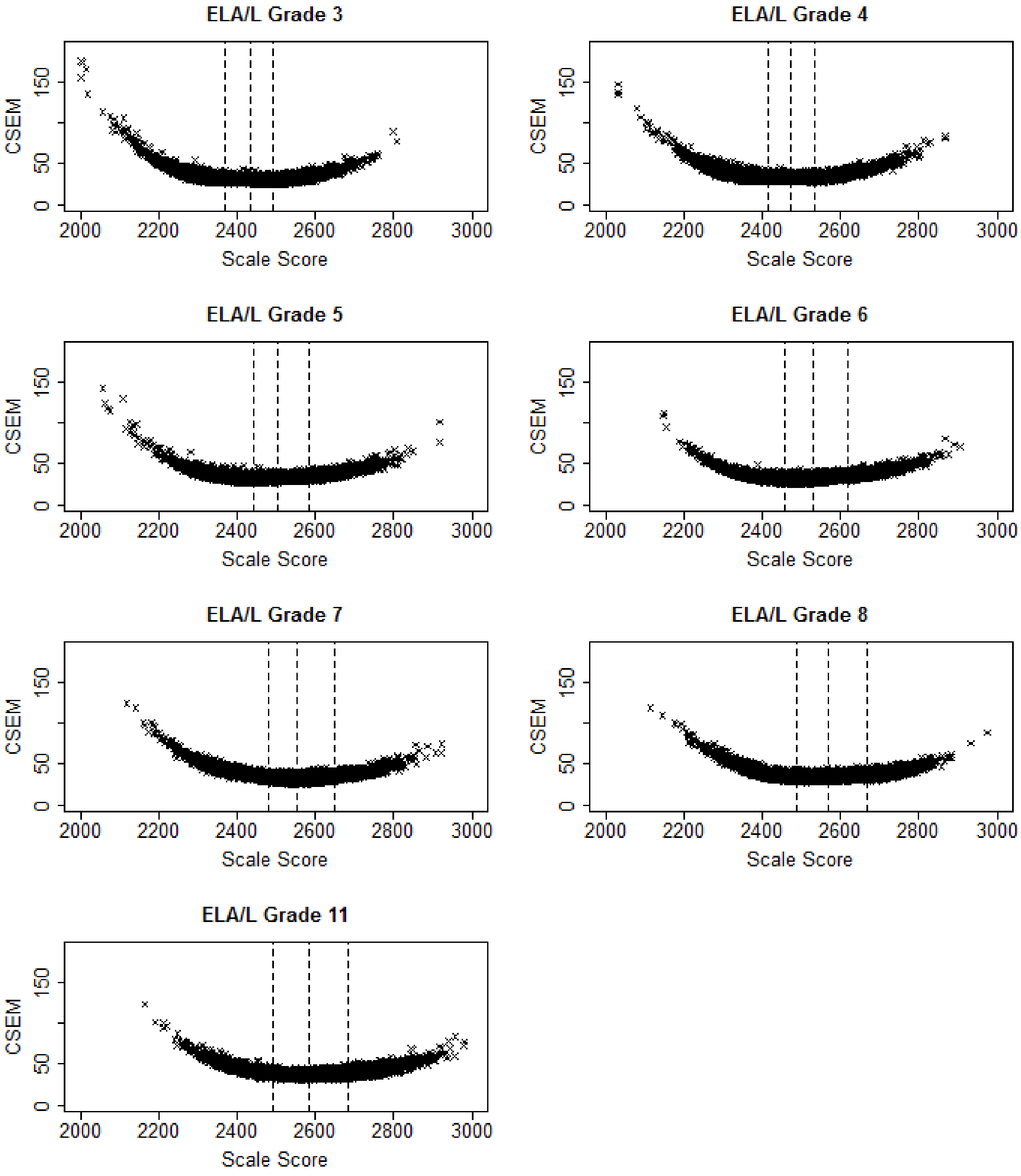
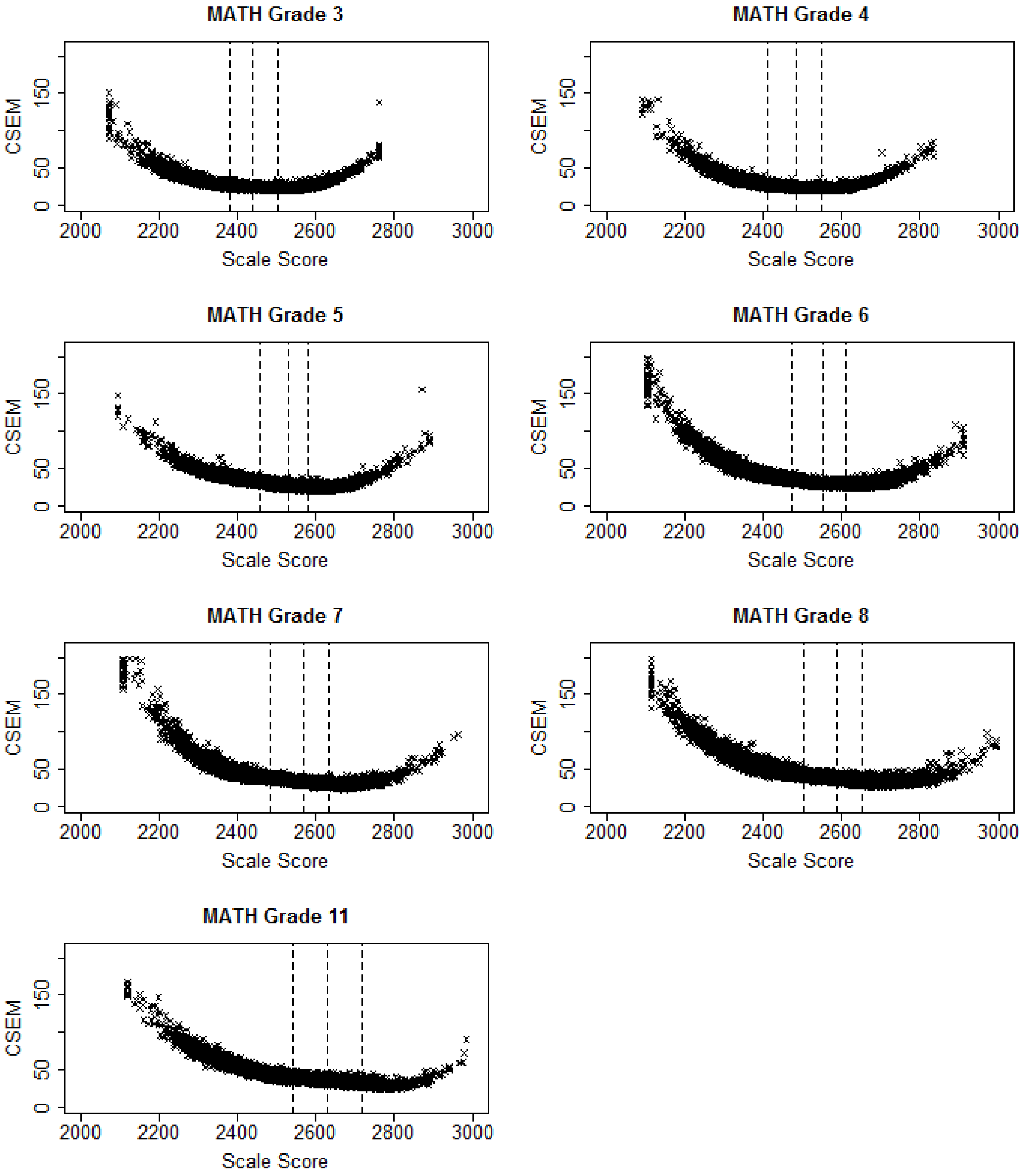


Figure 18. Conditional Standard Error of Measurement: Mathematics



The CSEMs presented in Figures 17 and 18 are summarized in Tables 61 and 62. Table 61 provides the average CSEM for all scale scores and by achievement level. Table 62 presents the average CSEMs at each cut score and the difference in average CSEMs between two cut scores. As shown in Figures 17 and 18, the greatest average CSEM is in Level 1 in both ELA/L and mathematics. Average CSEMs at all cut scores are similar in ELA/L, but larger in Level 2 cut scores in mathematics.

Table 61. Average Conditional Standard Error of Measurement by Achievement Level

Grade	Level 1	Level 2	Level 3	Level 4	Average CSEM
			ELA/L		
3	38.24	31.11	30.46	33.46	33.78
4	38.81	33.67	33.03	36.78	36.04
5	36.60	32.33	33.24	37.97	35.33
6	35.77	31.87	33.93	38.49	34.91
7	41.95	33.38	34.18	38.90	36.98
8	40.45	34.00	35.12	39.27	36.91
11	46.50	37.69	37.94	42.30	40.69
			Mathematics		
3	35.63	25.22	23.45	26.12	28.25
4	35.56	24.83	22.91	25.49	27.65
5	39.87	28.59	25.23	26.55	31.80
6	50.18	31.58	29.27	32.13	39.32
7	53.66	35.16	31.28	32.13	42.47
8	56.75	40.18	35.29	34.49	46.80
11	52.26	37.62	33.25	30.78	43.97

Table 62. 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	31.38	30.50	30.64	0.88	0.14	0.74
4	33.98	33.34	33.63	0.64	0.29	0.35
5	32.72	32.71	34.31	0.01	1.60	1.59
6	32.05	33.00	35.04	0.95	2.04	2.99
7	35.22	33.06	35.21	2.16	2.15	0.01
8	33.83	34.95	36.22	1.12	1.27	2.39
11	39.26	38.21	38.67	1.05	0.46	0.59
			Mathematics			
3	26.38	24.08	23.14	2.30	0.94	3.24
4	26.75	23.29	22.74	3.46	0.55	4.01
5	31.61	26.65	25.23	4.96	1.42	6.38
6	33.55	29.59	29.15	3.96	0.44	4.40
7	37.44	33.05	31.31	4.39	1.74	6.13
8	42.94	37.65	33.78	5.29	3.87	9.16
11	39.83	35.82	31.57	4.01	4.25	8.26

6.3 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 & Wingersky, 1979; Subkoviak, 1976; Livingston & Lewis, 1995). For the CAT, because the adaptive testing algorithm constructs a test form unique to each student, the classification indexes are computed based on all sets of items administered across students using an IRT-based method (Guo, 2006).

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

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

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

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

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

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

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

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

where d stands for dichotomous and p stands for polytomous items; $\mathbf{b}_j = \left(a_j, b_j, c_j\right)$ if the jth item is a dichotomous item, and $\mathbf{b}_j = (a_j, b_{j1}, \dots, b_{jK_i})$ 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 *i*th student's achievement level, and p_{im} is the probability of the *i*th student being classified at achievement level *m*. In the above table, the row represents the observed level, and the column represents the expected level.

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

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

and the overall classification accuracy is estimated by

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

where N is the total number of students. Because classifying students as proficient or not proficient is such a high-stakes decision, classification accuracy is also considered at the proficiency level by repeating the process for overall classification accuracy of achievement levels but with the four achievement levels collapsed into two proficiency categories: proficient (achievement levels 3 and 4) and not proficient (achievement levels 1 and 2).

Classification Consistency

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

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

where $n_{clm} = \sum_{i=1}^{N} p_{il} p_{im}$. p_{il} and p_{im} are the probabilities of the 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 63 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 79% for accuracy and from 66% to 71% for the consistency across all grades and subjects. For achievement levels, the classification index is higher in L1 and L4 than in L2 and L3. The higher accuracy at L1 and L4 is due to the fact that the intervals used to compute the classification probabilities for students in L1 and L4 [$-\infty,$ 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 90% to 92% for accuracy and from 87% to 89% 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 D, Classification Accuracy and Consistency Index by Subgroup.

Table 63. Classification Accuracy and Consistency

Grade	Achievement Level	ELA/L		Mathematics	
		% Accuracy	% Consistency	% Accuracy	% Consistency
3	Overall	76	68	77	69
	L1	89	82	86	79
	L2	62	51	64	51
	L3	59	48	71	61
	L4	87	80	88	82
	Proficiency Cut	91	87	92	89
4	Overall	74	66	79	71
	L1	88	82	87	81
	L2	55	43	73	63
	L3	57	46	71	61
	L4	85	78	87	80
	Proficiency Cut	90	87	92	89
5	Overall	76	68	78	70
	L1	88	82	88	82
	L2	58	46	68	57
	L3	67	56	61	49
	L4	85	78	88	81
	Proficiency Cut	91	88	92	89
6	Overall	76	67	78	70
	L1	88	82	89	84
	L2	65	54	68	58
	L3	69	60	60	48
	L4	83	74	86	78
	Proficiency Cut	91	88	92	88
7	Overall	76	67	78	70
	L1	89	82	89	83
	L2	64	52	66	56
	L3	72	63	63	51
	L4	82	72	86	77
	Proficiency Cut	91	87	91	87
8	Overall	76	67	76	68
	L1	88	80	87	82
	L2	66	55	61	50
	L3	72	64	59	47
	L4	82	71	86	77
	Proficiency Cut	91	87	92	88
11	Overall	75	67	79	71
	L1	86	77	89	85
	L2	66	55	64	54
	L3	69	60	70	58
	L4	84	76	84	74
	Proficiency Cut	91	87	92	89

6.4 Reliability for Subgroups

The reliability of test scores is also computed by subgroup. Tables 64–71 present the marginal reliability coefficients by the subgroup: 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 64. Marginal Reliability Coefficients for Overall and by Subgroup: ELA/L (Grades 3–4)

Subgroup	**Grade 3**				**Grade 4**			
	MR	**SS**	**SD**	**CSEM**	**MR**	**SS**	**SD**	**CSEM**
All Students	0.89	2425.19	101.40	33.78	0.88	2470.92	103.20	36.04
Female	0.89	2436.00	100.52	33.60	0.87	2482.27	100.63	35.77
Male	0.89	2415.28	101.20	33.95	0.88	2460.38	104.44	36.29
African American	0.86	2434.34	87.95	32.62	0.86	2452.24	94.21	35.47
AmerIndian/Alaskan	0.81	2413.59	71.26	31.04	0.86	2459.94	93.53	34.62
Asian/Pacific	0.88	2457.40	95.96	32.81	0.87	2499.91	98.55	35.75
Hispanic	0.88	2409.87	98.87	34.15	0.87	2455.49	100.04	36.14
Hawai'i Pacific	0.86	2374.54	91.99	34.84	0.85	2424.32	95.43	36.56
White	0.88	2455.10	94.85	33.49	0.87	2503.27	97.82	35.72
Multi-Racial	0.89	2442.96	99.21	33.55	0.87	2488.40	101.31	35.90
ELL	0.86	2373.77	92.77	35.15	0.84	2413.56	92.92	36.92
Disadvantaged	0.87	2389.96	95.88	34.41	0.86	2437.25	97.73	36.27
Migrant	0.84	2363.73	93.31	37.53	0.85	2416.76	96.14	36.82
Disability	0.77	2318.99	82.74	39.62	0.78	2357.07	85.96	40.27

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

Table 65. Marginal Reliability Coefficients for Overall and by Subgroup: ELA/L (Grades 5–6)

Subgroup	**Grade 5**				**Grade 6**			
	MR	**SS**	**SD**	**CSEM**	**MR**	**SS**	**SD**	**CSEM**
All Students	0.89	2509.88	107.82	35.33	0.89	2525.04	104.80	34.91
Female	0.89	2524.35	104.32	35.31	0.88	2538.36	101.64	34.92
Male	0.90	2496.33	109.28	35.35	0.89	2512.46	106.19	34.91
African American	0.87	2498.96	95.99	34.59	0.88	2530.16	99.86	34.47
AmerIndian/Alaskan	0.79	2535.91	73.84	34.02	0.90	2501.58	111.22	35.23
Asian/Pacific	0.88	2542.73	103.18	35.54	0.88	2553.32	102.31	35.27
Hispanic	0.89	2497.46	105.20	35.13	0.88	2510.21	100.54	34.70
Hawai'i Pacific	0.88	2457.08	101.86	35.34	0.88	2474.89	98.83	34.75
White	0.86	2545.01	94.36	35.35	0.86	2565.96	92.43	35.08
Multi-Racial	0.89	2524.01	104.52	35.31	0.88	2542.37	98.92	34.78
ELL	0.85	2428.87	91.62	35.56	0.82	2435.64	81.62	34.91
Disadvantaged	0.88	2473.41	103.57	35.18	0.88	2490.63	99.13	34.56
Migrant	0.87	2440.37	97.71	35.37	0.85	2458.03	87.27	34.16
Disability	0.83	2392.25	90.72	37.67	0.80	2408.44	83.10	36.82

Table 66. Marginal Reliability Coefficients for Overall and by Subgroup: ELA/L (Grades 7–8)

Subgroup	**Grade 7**				**Grade 8**			
	MR	**SS**	**SD**	**CSEM**	**MR**	**SS**	**SD**	**CSEM**
All Students	0.88	2548.91	108.29	36.98	0.88	2561.71	107.21	36.91
Female	0.88	2563.13	104.65	36.70	0.87	2577.17	100.67	36.49
Male	0.89	2535.87	109.92	37.23	0.89	2546.99	111.11	37.31
African American	0.86	2558.72	95.24	35.98	0.85	2571.46	91.27	35.64
AmerIndian/Alaskan	0.86	2604.01	99.35	36.86	0.86	2565.67	94.03	35.22
Asian/Pacific	0.88	2580.46	103.77	36.55	0.87	2595.18	101.74	36.77
Hispanic	0.88	2534.42	106.54	37.57	0.88	2545.53	105.00	36.79
Hawaiʻi Pacific	0.86	2497.28	100.96	37.86	0.86	2509.28	99.87	37.72
White	0.86	2593.36	98.41	36.27	0.87	2593.60	99.50	36.49
Multi-Racial	0.87	2561.15	101.85	36.23	0.87	2574.33	102.95	36.53
ELL	0.81	2459.89	91.89	39.88	0.81	2476.26	87.93	37.93
Disadvantaged	0.87	2514.97	105.21	37.73	0.87	2528.26	104.90	37.31
Migrant	0.85	2485.57	96.57	37.94	0.84	2482.63	96.07	37.86
Disability	0.80	2433.59	89.47	40.49	0.79	2439.03	90.23	40.92

Table 67. Marginal Reliability Coefficients for Overall and by Subgroup: ELA/L (Grade 11)

Subgroup	**Grade 11**			
	MR	**SS**	**SD**	**CSEM**
All Students	0.88	2604.42	115.29	40.69
Female	0.86	2622.00	109.08	40.39
Male	0.88	2587.47	118.53	40.99
African American	0.88	2588.48	118.70	40.54
AmerIndian/Alaskan	0.81	2620.77	88.21	38.69
Asian/Pacific Islander	0.86	2630.72	106.82	40.30
Hispanic	0.87	2585.76	112.97	40.59
Hawaiʻi Pacific Islander	0.86	2556.16	111.56	41.10
White	0.87	2632.01	114.01	40.98
Multi-Racial	0.88	2618.93	116.27	41.03
ELL	0.76	2488.30	88.55	43.12
Disadvantaged	0.87	2571.24	114.68	41.10
Migrant	0.87	2547.18	114.86	41.14
Disability	0.78	2465.30	94.44	44.73

Table 68. Marginal Reliability Coefficients for Overall and by Subgroup: Mathematics (Grades 3–4)

Subgroup	**Grade 3**				**Grade 4**			
	MR	**SS**	**SD**	**CSEM**	**MR**	**SS**	**SD**	**CSEM**
All Students	0.91	2435.11	94.91	28.25	0.91	2472.36	92.57	27.65
Female	0.91	2433.19	91.57	28.06	0.90	2469.16	88.31	27.38
Male	0.92	2436.88	97.85	28.43	0.92	2475.32	96.26	27.90
African American	0.85	2435.19	80.45	30.89	0.87	2454.37	74.54	27.05
AmerIndian/Alaskan	0.87	2413.14	76.90	27.22	0.92	2445.66	102.75	29.76
Asian/Pacific	0.91	2471.57	88.62	26.87	0.91	2504.42	89.10	26.25
Hispanic	0.90	2419.47	91.66	28.29	0.89	2455.37	87.80	28.46
Hawaiʻi Pacific	0.88	2385.10	88.24	30.83	0.88	2426.68	85.90	30.17
White	0.90	2461.49	85.60	26.50	0.91	2502.18	87.09	26.20
Multi-Racial	0.91	2450.41	90.78	27.51	0.91	2488.41	87.35	26.25
ELL	0.90	2393.63	94.97	30.72	0.88	2424.60	87.14	30.55
Disadvantaged	0.89	2402.09	91.27	29.86	0.89	2441.08	87.73	29.05
Migrant	0.83	2364.62	85.52	35.06	0.87	2421.16	82.74	29.81
Disability	0.84	2338.40	89.39	35.96	0.81	2375.15	84.11	36.70

Table 69. Marginal Reliability Coefficients for Overall and by Subgroup: Mathematics (Grades 5–6)

Subgroup	**Grade 5**				**Grade 6**			
	MR	**SS**	**SD**	**CSEM**	**MR**	**SS**	**SD**	**CSEM**
All Students	0.90	2501.03	100.61	31.80	0.88	2505.77	114.36	39.32
Female	0.89	2500.13	96.61	31.48	0.88	2505.31	110.78	38.91
Male	0.91	2501.87	104.21	32.10	0.89	2506.20	117.65	39.69
African American	0.86	2482.25	84.25	31.36	0.87	2503.53	102.38	37.07
AmerIndian/Alaskan	0.77	2505.84	59.56	28.49	0.89	2457.90	170.87	57.70
Asian/Pacific	0.90	2541.80	96.49	29.80	0.89	2543.30	108.85	35.57
Hispanic	0.88	2482.28	93.74	32.19	0.86	2484.33	111.10	41.42
Hawaiʻi Pacific	0.86	2450.65	93.07	35.36	0.82	2450.56	107.55	45.06
White	0.89	2529.28	90.42	29.46	0.89	2550.18	102.47	34.68
Multi-Racial	0.90	2512.97	97.34	30.70	0.88	2522.51	106.19	36.43
ELL	0.84	2434.20	90.95	36.84	0.76	2419.27	100.31	49.00
Disadvantaged	0.87	2465.68	95.79	34.06	0.85	2468.48	109.44	42.71
Migrant	0.86	2431.35	97.92	37.28	0.80	2427.37	106.06	47.56
Disability	0.78	2400.52	87.92	41.25	0.71	2386.34	102.52	55.50

Table 70. Marginal Reliability Coefficients for Overall and by Subgroup: Mathematics (Grades 7–8)

Subgroup	**Grade 7**				**Grade 8**			
	MR	**SS**	**SD**	**CSEM**	**MR**	**SS**	**SD**	**CSEM**
All Students	0.87	2513.27	117.48	42.47	0.86	2524.30	123.71	46.80
Female	0.87	2511.05	115.13	42.17	0.85	2526.71	119.05	46.12
Male	0.87	2515.30	119.58	42.74	0.86	2522.01	127.96	47.44
African American	0.85	2503.71	101.92	39.75	0.85	2536.41	117.35	45.76
AmerIndian/Alaskan	0.92	2555.80	125.59	35.84	0.84	2520.79	106.45	42.62
Asian/Pacific	0.89	2556.89	116.86	38.47	0.88	2570.96	122.36	42.08
Hispanic	0.84	2492.25	107.76	43.50	0.82	2498.61	113.68	48.63
Hawaiʻi Pacific	0.79	2456.55	106.48	48.67	0.77	2463.29	109.87	52.92
White	0.88	2555.99	106.48	37.17	0.86	2558.88	116.41	42.91
Multi-Racial	0.87	2526.41	112.56	40.88	0.85	2536.41	117.74	45.62
ELL	0.76	2424.36	108.98	53.78	0.73	2433.40	110.45	57.02
Disadvantaged	0.83	2477.36	112.15	46.10	0.82	2486.52	117.97	50.55
Migrant	0.76	2449.79	97.50	48.09	0.76	2458.21	108.39	53.03
Disability	0.64	2396.74	97.90	58.39	0.63	2400.48	102.04	62.37

Table 71. Marginal Reliability Coefficients for Overall and by Subgroup: Mathematics (Grade 11)

Subgroup	**Grade 11**			
	MR	**SS**	**SD**	**CSEM**
All Students	0.87	2550.90	120.01	43.97
Female	0.86	2555.46	113.36	42.98
Male	0.87	2546.49	125.97	44.91
African American	0.86	2529.33	119.11	44.97
AmerIndian/Alaskan	0.82	2543.56	100.69	42.36
Asian/Pacific Islander	0.88	2583.42	115.86	40.66
Hispanic	0.84	2526.20	113.11	45.65
Hawaiʻi Pacific Islander	0.80	2496.59	109.47	49.28
White	0.87	2575.74	115.90	41.74
Multi-Racial	0.87	2569.86	117.46	41.80
ELL	0.74	2463.52	103.71	52.55
Disadvantaged	0.84	2518.03	116.02	46.84
Migrant	0.76	2480.23	98.51	48.65
Disability	0.60	2412.50	93.65	59.53

6.5 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 Hawai'i 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 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. Tables 72 and 73 present the marginal reliability coefficients and descriptive statistics by claim in ELA/L and mathematics, respectively.

Table 72. Marginal Reliability Coefficients for Claim Scores: 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	0.62	2430.42	123.30	76.45
	Claim 2: Writing	6	0.72	2414.35	125.93	66.77
	Claim 3: Listening	4	0.28	2430.39	144.61	122.95
	Claim 4: Research	6	0.62	2427.26	133.73	82.92
4	Claim 1: Reading	8	0.60	2469.83	129.61	81.87
	Claim 2: Writing	6	0.70	2465.59	132.94	72.58
	Claim 3: Listening	4	0.30	2464.67	148.11	123.91
	Claim 4: Research	6	0.59	2475.70	143.18	92.15
5	Claim 1: Reading	8	0.61	2509.23	134.69	83.67
	Claim 2: Writing	6	0.74	2508.27	134.90	69.41
	Claim 3: Listening	4	0.33	2508.72	156.22	127.84
	Claim 4: Research	6	0.64	2513.55	135.99	81.04
6	Claim 1: Reading	10	0.69	2520.33	127.13	70.59
	Claim 2: Writing	6	0.72	2516.28	131.06	69.48
	Claim 3: Listening	4	0.30	2540.81	159.67	133.51
	Claim 4: Research	6	0.59	2540.02	142.04	90.50
7	Claim 1: Reading	10	0.63	2543.19	137.15	82.97
	Claim 2: Writing	6	0.72	2546.98	135.91	71.56
	Claim 3: Listening	4	0.29	2543.58	149.44	125.93
	Claim 4: Research	6	0.61	2552.60	150.28	93.81
8	Claim 1: Reading	10	0.66	2555.08	130.70	75.71
	Claim 2: Writing	6	0.70	2556.92	134.24	73.37
	Claim 3: Listening	4	0.30	2564.07	157.14	131.37
	Claim 4: Research	6	0.59	2576.34	147.26	94.19
11	Claim 1: Reading	10	0.65	2593.77	143.50	85.07
	Claim 2: Writing	6	0.71	2611.74	144.39	77.51
	Claim 3: Listening	4	0.32	2603.02	177.05	145.47
	Claim 4: Research	6	0.59	2609.99	159.64	102.58

Table 73. Marginal Reliability Coefficients for Claim Scores: Mathematics

Grade	Claim	Number of Items Specified in Test Blueprint	Marginal Reliability	Scale Score Mean	Scale Score SD	Average CSEM
3	Claim 1	12	0.84	2437.12	105.37	41.61
	Claims 2 & 4	5	0.60	2431.87	108.06	68.69
	Claim 3	5	0.58	2429.60	111.42	72.17
4	Claim 1	12	0.84	2474.30	103.65	41.05
	Claims 2 & 4	5	0.55	2466.73	104.09	69.88
	Claim 3	5	0.62	2467.97	109.73	67.85
5	Claim 1	12	0.83	2505.70	110.93	45.83
	Claims 2 & 4	5	0.46	2497.09	114.38	83.93
	Claim 3	5	0.56	2487.68	129.40	86.24
6	Claim 1	12	0.81	2508.09	127.72	55.77
	Claims 2 & 4	5	0.44	2497.82	130.50	97.47
	Claim 3	5	0.46	2503.06	140.99	103.31
7	Claim 1	12	0.78	2512.49	132.05	61.50
	Claims 2 & 4	5	0.39	2507.85	134.14	104.94
	Claim 3	5	0.46	2509.24	144.72	106.07
8	Claim 1	12	0.77	2522.50	138.72	66.93
	Claims 2 & 4	5	0.44	2524.52	133.19	99.26
	Claim 3	5	0.39	2522.24	155.30	121.12
11	Claim 1	14	0.80	2550.03	127.12	57.37
	Claims 2 & 4	5	0.53	2541.60	176.21	121.09
	Claim 3	5	0.48	2529.04	174.85	125.60

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

7. 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.

7.1 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 *j*th person's score pattern for *I* items is

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

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

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

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

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

in the case of items with two or more points,

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

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

Standard Error of Measurement

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

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

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

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

where m_i is the maximum possible score point (starting from 0) for the 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.

7.2 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 74 presents the scaling constants for each subject for the theta-to-scale score linear transformation. Scale scores are rounded to an integer.

Table 74. 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, SS_{θ} is the standard error of the ability estimate on the θ scale, and a is the slope of the scaling constant that transforms θ into the reporting scale.

The scale scores are mapped into four achievement levels using three achievement standards (i.e., cut scores). Table 75 provides three achievement standards for each grade and content area.

Table 75. 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

7.3 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 76 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 76. Extended Lowest and Highest Obtainable Scores

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

7.4 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.

7.5 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.

7.6 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 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).

7.6.1 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 *j*th student's score on the *i*th item. For items with one score point, the 2PL IRT model is used to calculate the expected score on item *i* for student *j* with estimated ability $\hat{\theta}_j$ as:

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

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

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

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

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

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

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

For an aggregate unit, a target score is computed by averaging the individual student target scores for 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 statistically significant 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 target-level strengths/weaknesses, the following are reported:

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

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

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

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

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

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

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.

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

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

A statistically significant difference from zero in these aggregates may indicate that a class, teacher, school, 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.

7.7 HANDSCORING

Constructed-response short-answer (SA) items and essay (i.e., full-write) items in ELA/L and short-answer (SA) items in mathematics for the summative assessments administered by 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. The methods and results used for handscoring and autoscoring are described in the following sections.

For handscoring items, CAI generated the total number of items and the summary of rater agreements across all states and territories that participated in the 2021–2022 summative assessments in grades 3–8 and 11. Grade 11 data are based on the students in grades 9, 10, and 11.

For the 2021–2022 summative operational item pool, there were a total of 616 SA items and 198 essay items in ELA/L and 345 SA items in mathematics. Table 77 shows the number of items by grade and subject.

Table 77. Number of Handscored Items in 2021–2022 Smarter Balanced Summative Item Pool, by Grade and Subject

Grade	**ELA/L**		**Mathematics**
	Short Answer	**Essay**	
3	67	25	46
4	75	29	52
5	83	30	74
6	69	22	52
7	70	30	35
8	76	33	41
11	176	29	45
Total	616	198	345

All guidelines for handscoring responses were specified by Smarter Balanced. Outlined below is the handscoring process MI followed in spring 2022 in accordance with the Smarter Balanced guidelines. This process applied to the scoring of all student constructed responses for ELA/L SA and essay items and mathematics SA items.

7.7.1 Rater Selection

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

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

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

In selecting team leaders who will 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.

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.

7.7.2 Rater Training and Scoring

All raters hired to score the Smarter Balanced assessments were trained using the rubric(s), anchor sets, and training/qualifying sets provided by Smarter Balanced. These sets were created during the original field-test scoring in 2014 and approved by Smarter Balanced. The same anchor sets are used each year. Additionally, MI conducts an annual review of the rater agreement and scoring materials in order to inform the development of item-specific, supplemental training materials. Supplemental materials are developed each summer and implemented in the subsequent operational administration.

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

All raters, regardless of experience, were required to train on all anchor and training sets. Following training, all raters were required to pass the qualification sets in order to prove that they understood and could apply the criteria accurately. Until a rater had trained and qualified successfully, the rater was not permitted to score any student responses. MI carefully orchestrated training so that raters understood that all scoring decisions must be grounded in the training materials. In addition, raters learned how to navigate the anchor set, developed the knowledge and flexibility needed to evaluate or escalate a variety of responses, and retained the necessary consistency to score all responses accurately.

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

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

Training design varied slightly depending on Smarter Balanced item type:

- ELA/L essay: 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 in that grade and purpose. Raters could only score those items for which they have 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 qualified the rater to score all items in that grade band and target.
- Mathematics SA: Raters trained and qualified on baseline lessons within a specific grade band. Qualification on a baseline lesson qualified the rater to score that item and all items associated with it; for items with no associated items, training was for the specific item.

Rater training time varied by grade and content area. Training for ELA/L brief write, ELA/L reading, research SA, and mathematics SA items could typically be accomplished in one day, while training for essay items took up to five days to complete. Raters generally worked 6.5 hours per day, excluding breaks. Evening shift raters worked 3.75 hours, excluding breaks.

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

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

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

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

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

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

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

Finally, a series of automated score verifications were implemented to further ensure the accuracy of scores. For example, a blank check was conducted, which reset scores when a condition code of "blank" was assigned to a response that had one or more characters in the response string (e.g., a response comprised of spaces or tabs). In this case, only after three independent raters had assigned a condition code of "blank" to a response that appeared blank, but which included characters in the response string, was the score recorded. A similar check was run when a score or condition code other than "blank" was assigned to a response that included no characters in the response string. Automatic resetting of double-scored responses when two raters assigned 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.

7.7.3 Rater Statistics and Monitoring

At a minimum, 10–15% (depending on state contractual requirements) of the handscored responses received blind double reads. Additionally, 5% of the responses scored comprised pre-approved validity responses. MI's VSC system automatically and randomly routed the requisite number of responses to raters for second reads and validity in an inconspicuous manner. Raters had no means of discerning whether they were scoring a first read, a second read, or a validity response. This system also prohibited raters from being eligible to score second reads for responses they had already scored.

MI's VSC scoring system randomly seeds validity responses among operational responses during scoring. A small set of validity responses is provided by Smarter Balanced for all vendors to use, and these are supplemented with responses selected and approved by MI scoring management. The “true” scores for these responses are entered into a validity database. Validity responses are indistinguishable from operational responses.

VSC reports provided real-time reports throughout the scoring effort. These reports were available for access by handscoring management. Inter-rater reliability reports provided the percentage of exact, adjacent, and non-adjacent agreement for scorable responses. Validity performance reports provided the percentage of exact, adjacent, and non-adjacent agreement for validity responses and were used to monitor drift. Score point frequency distribution reports provided the percentage per score point and included the mean and standard deviation for each item.

Years of Smarter Balanced handscoring has allowed MI to amass a longitudinal dataset of rater performance data. MI’s rater monitoring system uses validity responses calibrated to fit a unidimensional item response theory (IRT) model for each content area/item type. Extensive metrics (inter-rater reliability, calibrated validity, and sub-pools for monitoring drift) calculated by the monitoring system were used to ensure accuracy and productivity throughout the handscoring of a project. The system generated automated measures of rater performance drawing on validity, inter-rater reliability (IRR), and other performance data. Raters and scoring managers received daily, automated messages summarizing raters’ performance, ensuring all handscoring staff were aware of current performance and any issues that required attention. Additional outputs were also provided in manager-level reports and used to identify raters who required retraining and/or removal due to issues with accuracy and/or production. These data allowed scoring management to direct scoring leaders in review of specific VSC reports in order to determine the specific areas of attention required for any raters.

The monitoring system afforded the objective, dynamic identification of the most accurate and productive raters, referred to as “advanced raters.” Advanced rater status changed daily based on current rater performance to ensure that any rater drift did not negatively impact scoring accuracy. Advanced rater status was a precondition for conducting second readings.

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.

7.7.4 Rater Retraining and Dismissal

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

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

7.7.5 Rater Agreement

Rater IRR was computed based only on scorable responses (numeric scores) scored by two independent raters. Non-scorable responses (e.g., off-topic, off-purpose, foreign-language responses) that were scored by scoring leadership—and not by two independent raters—were excluded from IRR computations. For the handscored items, the human-human agreement was computed based on the combined data across all states and territories that participated in the 2021–2022 summative assessments.

In ELA/L, essay (i.e., full-write) 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. Condition codes were scored as zero.

Tables 78 through 80 provide a summary of the human-human IRR based on items with a sample size greater than 50. The IRR is presented with mean of percentage exact agreement, minimum and maximum percentage exact agreements, combined percentage exact and adjacent agreement, and the mean, minimum and maximum quadratic weighted kappa (QWK). The average number of responses, as well as minimum and maximum number of responses to a given item, are presented, as well.

Table 78. 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	578.9	73	1086	72.4	65.4	84.1	100.0	0.68	0.44	0.78
4	42	487.2	54	1127	70.1	58.3	86.6	100.0	0.68	0.42	0.81
5	37	532.5	77	1023	68.7	54.3	82.3	100.0	0.70	0.39	0.86
6	43	896.7	70	3247	70.5	61.5	84.8	100.0	0.67	0.48	0.86
7	48	828.4	90	3769	69.5	58.4	84.7	100.0	0.67	0.47	0.82
8	55	774.2	67	3261	69.3	56.1	83.6	100.0	0.68	0.47	0.85
11	99	415.7	51	966	68.6	53.9	86.9	100.0	0.70	0.42	0.89

Table 79. Inter-Rater Agreement for ELA/L Essay Items

Grade	**Dimension**	**Number of Items**	**Number of Responses**			**%Exact**			**%(Exact+ Adjacent)**	**QWK**		
			Mean	**Min**	**Max**	**Mean**	**Min**	**Max**		**Mean**	**Min**	**Max**
3	Conventions	25	685.2	385	980	60.6	52.2	65.2	97.4	0.55	0.46	0.68
	Evid/Elab	25	685.2	385	980	63.9	52.5	71.7	96.7	0.61	0.45	0.77
	Org/Purp	25	685.2	385	980	63.9	52.9	71.6	96.7	0.61	0.46	0.76
4	Conventions	29	675.1	359	915	56.1	48.4	65.4	95.1	0.52	0.40	0.65
	Evid/Elab	29	675.1	359	915	60.9	53.3	66.5	96.0	0.62	0.53	0.78
	Org/Purp	29	675.1	359	915	60.9	52.3	66.5	96.1	0.63	0.52	0.77
5	Conventions	29	760.7	422	992	61.9	53.0	69.7	97.7	0.51	0.33	0.60
	Evid/Elab	29	760.7	422	992	60.8	53.0	65.3	97.1	0.67	0.54	0.76
	Org/Purp	29	760.7	422	992	61.3	53.2	66.9	97.2	0.67	0.54	0.75
6	Conventions	22	934.4	607	1152	60.6	53.0	67.2	96.9	0.55	0.49	0.61
	Evid/Elab	22	934.4	607	1152	66.1	54.6	72.3	97.9	0.68	0.56	0.74
	Org/Purp	22	934.4	607	1152	66.0	54.9	72.9	98.0	0.68	0.52	0.74
7	Conventions	30	711.4	423	857	64.3	56.8	73.6	98.1	0.53	0.39	0.69
	Evid/Elab	30	711.4	423	857	64.6	53.0	73.1	98.1	0.68	0.59	0.77
	Org/Purp	30	711.4	423	857	65.2	53.8	74.8	98.2	0.69	0.60	0.77
8	Conventions	33	669.1	449	830	68.2	55.4	76.8	98.4	0.54	0.42	0.64
	Evid/Elab	33	669.1	449	830	63.3	55.0	73.8	97.9	0.67	0.54	0.77
	Org/Purp	33	669.1	449	830	63.2	53.2	72.9	98.1	0.68	0.60	0.77
11	Conventions	29	701.0	616	782	70.7	64.0	76.6	98.6	0.60	0.47	0.67
	Evid/Elab	29	701.0	616	782	62.4	54.3	70.5	98.6	0.72	0.61	0.79
	Org/Purp	29	701.0	616	782	62.4	54.2	70.2	98.6	0.72	0.60	0.79

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

Table 80. 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	8	1382.9	1036	1676	92.5	91.0	96.1	100.0	0.83	0.77	0.91
4	0–1	10	1269.2	1192	1402	88.3	82.2	94.2	100.0	0.69	0.58	0.88
5	0–1	9	1238.2	1091	1351	92.2	84.2	97.6	100.0	0.71	0.41	0.95
6	0–1	12	1299.3	756	1990	97.1	94.7	100.0	100.0	0.73	0.45	1.00
7	0–1	10	1643.4	1137	2025	95.1	87.9	98.2	100.0	0.76	0.33	0.92
8	0–1	15	1980.6	1844	2118	92.2	85.0	98.4	100.0	0.76	0.55	0.95
11	0–1	16	1328.8	83	1694	92.9	87.0	100.0	100.0	0.74	0.60	1.00
3	0–2	32	1477.9	343	1962	90.4	81.2	99.2	100.0	0.92	0.80	0.97
4	0–2	38	1255.8	321	1607	89.2	78.4	99.7	100.0	0.88	0.47	1.00
5	0–2	57	1237.1	482	1507	88.8	78.7	97.2	100.0	0.87	0.56	0.97
6	0–2	40	1798.2	1503	2039	88.1	72.5	98.4	100.0	0.85	0.72	0.98
7	0–2	24	1646.8	1338	1952	91.4	81.6	96.6	100.0	0.86	0.58	0.97
8	0–2	26	1817.5	1586	2124	90.3	84.4	99.1	100.0	0.86	0.75	0.99
11	0–2	22	1477.1	948	2021	91.0	76.2	99.2	100.0	0.87	0.53	0.98
3	0–3	6	1098.2	680	1764	91.6	89.2	94.7	100.0	0.96	0.94	0.98
4	0–3	4	1223.5	1140	1381	84.8	83.5	86.7	100.0	0.93	0.92	0.94
5	0–3	8	1205.3	774	1453	88.2	85.2	98.4	100.0	0.89	0.74	0.97
7	0–3	1	1762.0	1762	1762	87.1	87.1	87.1	100.0	0.88	0.88	0.88
11	0–3	7	1610.3	1516	1895	87.5	80.7	91.3	100.0	0.90	0.88	0.92

7.8 AUTOMATED SCORING

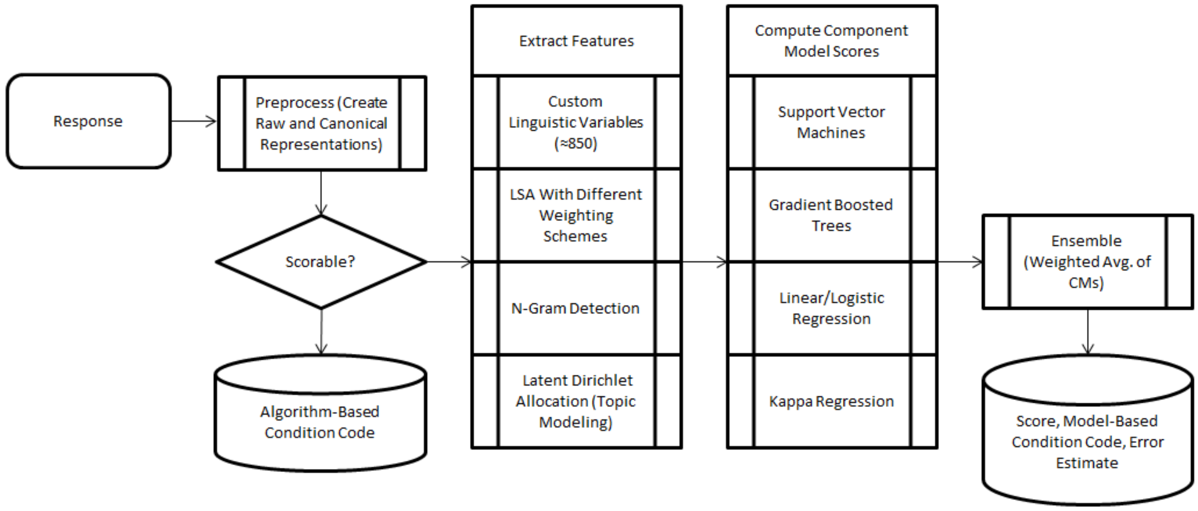
MI's PEG automated scoring technology was used to score eligible SA and essay items in ELA/L and SA items in mathematics. This section describes PEG, the training and validation sample and process, and the automated scoring process. This section concludes with the human-machine (HM) agreement statistics.

7.8.1 Project Essay Grade

MI's Project Essay Grade (PEG) automated scoring engine uses a supervised learning method involving Natural Language Processing, syntactic analysis, and Latent Semantic Analysis to model the relations among text features (i.e., elements of text) and human scores. For a detailed description of PEG modeling, see Bunch, Vaughn, and Miel, 2016. PEG measures thousands of response features, both surface and complex, and employs a host of algorithms to determine the mapping from features to scores so as to minimize error with expert raters. After extracting features on responses for which gold-standard human scores are available, PEG proceeds with a supervised learning approach to train a number of statistical models. These models draw on many of the latest advances in the field of machine learning to generate both linear and non-linear models. The strongest models are then automatically blended to create a final model that retains the best elements from the various algorithms. The reliability and criterion validity of PEG scoring have been confirmed in multiple empirical studies (e.g., Keith, 2003; Shermis, Koch, Page, Keith, & Harrington, 2002).

Figure 19 presents an overview of the PEG engine. Building an automated scoring solution is a multi-step process that includes component model training, ensembling, and scoring. The sections that follow describe this process and how it was used to extract features from responses and assign scores (or condition codes), as appropriate.

Figure 19. PEG Engine Overview


Response
Preprocess (Create Raw and Canonical Representations)
Scorable?
Algorithm-Based Condition Code
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 (Weighted Avg. of CMs)
Score, Model-Based Condition Code, Error Estimate

PEG is used in both formative and summative assessment contexts. In total, PEG has been used successfully in schools and districts in 27 states and several countries. In spring 2022, PEG provided nearly 10 million summative assessment scores for students across the United States.

7.8.2 Model Training and Validation

Sample

Automated scoring models were not—and could not—be created for items that had an insufficient quantity of training responses. This was this case for items that had low exposure to students, as dictated by the adaptive testing algorithm. Additionally, mathematics performance task items that had multiple parts with scoring dependencies were not considered for automated scoring. A total of 767 items (out of the 1,044 total items) were initially identified as eligible for automated scoring for spring 2022, as shown in Table 81.

Table 81. Number of Items Eligible for Automated Scoring, by Grade and Subject Area

Grade	ELA/L		Mathematics
	Short-Answer	Essay	
3	29	18	44
4	38	22	51
5	34	18	64
6	43	12	52
7	49	20	30
8	58	19	41
11	61	22	42
Total	312	131	324

Training Data

Student responses used for training and validation were sourced from the 2014 Smarter Balanced field test as well as the 2016–2017, 2017–2018, 2018–2019, 2020–2021, and 2021–2022 Smarter Balanced operational test administrations. Field test responses were randomly sampled from the available on-grade responses in either the standard setting sample or the census sample. Operational test responses were also randomly sampled from available on-grade responses in the operational population. For all items, the sample included 1,500–2,000 responses, stratified by score point. The score of record used to train the engine was either the matched or resolution score for responses scored by two or more raters, or the score assigned by an expert rater. Expert raters are raters identified as highly accurate using calibrated validity responses (i.e., raters for whom MI has empirical evidence of high accuracy).

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

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.

For content-based items (e.g., SA mathematics items), the number of variables is unknown until the models are built. Because 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 Latent Semantic Analysis, N-Gram Detection, and Latent Dirichlet Allocation (a type of topic modeling). To further refine the variable generation process, MI built a computer language to perform a simultaneous search over semantic, lexographical, and syntactic features of responses.

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

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.

Note that building component models for each item—and for multi-dimensional items, each trait or dimension—prevents variables from being generalized between 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 professional raters.

The approaches just described typically results 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. Two or more professional raters will not always agree on what score to give a student's response; therefore, when the engine produces scores that agree with professional raters to the same or greater extent than the raters agree with each other, modeling is considered successful. The initial evaluation was made using the criteria shown in Table 82. This evaluation process was used for both the item-specific scoring models and the condition code models. Note that the absolute QWK criterion (.65) is slightly lower than that recommended by Williamson, Xi, and Breyer (2012) and the relative QWK criterion (.07) is slightly more stringent. The SMD criterion matches that of Williamson et al. (2012).

Table 82. 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 in what was a pilot integration of bias analyses into the initial validation process using available data from the previous summative administration. Items passing initial validation were considered; only items with spring 2021 student data from California were analyzed in this pilot study. While this was a subsample (*n*=107) of the items subject to initial validation, the pilot study represented MI's best effort to ensure that items showing evidence of bias were excluded from the items eligible for automated scoring during the spring 2022 administration.

As noted, spring 2021 student data from California was analyzed. MI received separate datafiles containing (1) hand-score data and (2) student demographic data associated with responses. Table 83 shows the demographic variables and categories. A crosswalk was used to link the handscored and demographic data. Matched data existed for 107 items.

Table 83. 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

Handscore data consisted of scores assigned by a pool of ETS raters. However, automated scoring models were trained exclusively using scores assigned by MI's expert raters. MI confirmed differences in agreement (i.e., QWK) existed among the two rater populations and the engine for a subset of the 107 items. Items that exhibited large agreement differences between the two groups of raters were excluded from the matched data. Item exclusion was determined using the criterion |engine holdout set HM QWK – subgroup data HM QWK | > 0.1. Of the 107 total items, 54 were eligible for analysis. While this data cleaning step was necessary during this pilot to support valid interpretations of bias analysis results, it will not be required in subsequent administrations; beginning in 2022, all data required for these analyses was produced by MI's expert raters.

For each item, analysis was performed on a subgroup if the number of observations (i.e., HM scores) was at least 10. A subgroup was flagged for bias if |SMD| ≥ 0.125 and if the SMD was significant at an overall significance level of 95%. A Bonferroni correction was used to adjust the significance level for each subgroup comparison. An item was flagged for bias if any subgroup comparison associated with the item was flagged. Of the 54 items eligible for analysis, 15 (27.8%) were flagged for bias as part of the initial validation and excluded from automated scoring.

Table 84 presents overall results of the initial validation. Models associated with 549 of the 767 items (71.6%) passed all initial validation criteria and the pilot bias evaluation criteria.

Table 84. Summary of Initial 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	29	18	44	10	18	41
4	38	22	51	14	20	48
5	34	18	64	12	13	58
6	43	12	52	34	10	18
7	49	20	30	43	16	14
8	58	19	41	48	16	18
11	61	22	42	58	20	20
Total	312	131	324	219	113	217

Secondary Validation

All models associated with items that passed initial validation were subject to a secondary validation at the start of the spring 2022 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 interval, 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 85 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 85. Secondary Validation Criteria

Criterion	**Threshold**
Agreement of automated scores with human scores	$QWK_{H:M} \geq 0.65$
Standardized mean score difference between human and automated scores	$\lvert SMD_{H:M} \rvert \leq 0.15$

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

Table 86 presents the secondary validation results. Of the 549 items with models subject to secondary validation, models associated with 407 of the items (74.1%) passed all secondary evaluation criteria.

Table 86. 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	10	18	41	9	9	30
4	14	20	48	11	14	38
5	12	13	58	7	7	32
6	34	10	18	24	9	17
7	43	16	14	32	11	14
8	48	16	18	28	14	15
11	58	20	20	47	15	19
Total	219	113	217	158	79	165

Live Training and Validation

Additionally, in April of 2022 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 2022 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 2022 operational responses offered potential to improve model performance. All models associated with these items were thus trained using either exclusively 2022 responses (when a minimum of 1,400 2022 responses/item existed) or 2022 responses supplemented with 2021 responses. In either case, the validation sets consisted exclusively of 2022 responses. Because live validation involved operational data, it was unnecessary to conduct a secondary validation.

Table 87 summarizes the results of the live training and validation. Of the 225 items associated with models that underwent live training and validation, models associated with 74 of the items (32.9%) passed all evaluation criteria. While this pass rate is considerably lower than the pass rates observed during the initial (71.6%) and secondary (74.1%) validation efforts, it is most likely explained by the nature of the items modeled. Specifically, since all item models in this sample had failed a prior validation, by design the sample consisted of difficult-to-model items.

Table 87. 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	1	2	14	0	0	8
4	2	2	14	0	1	5
5	4	3	31	0	0	5
6	4	0	35	4	0	15
7	11	1	18	8	1	3
8	16	1	26	15	1	6
11	15	2	23	1	0	6
Total	53	11	161	28	3	48

Following initial validation, secondary validation, and live training and validation, a total of 481 items, comprised of 186 ELA/L SA, 82 essay, and 213 mathematics SA, were scored using a hybrid process, described next.

7.8.3 Automated Scoring Processes

Hybrid Scoring Process

As 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 were handscored by the larger pool of raters. These raters were monitored using validity responses and backreads conducted by expert raters, and they and their supervisors (team leaders, scoring directors) received automated, daily reports of their performance (i.e., accuracy and productivity).

In the hybrid model, responses were first pre-processed for automated scoring and to filter alert responses and certain non-scorable cases (e.g., insufficient text to score or high proportion of copied prompt text). This is achieved through the use of a series of three-digit flags used to indicate condition codes as defined in the handscoring criteria (see Table 88 and Table 89). 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. Responses were then sent to the automated scoring engine, where text features are extracted, the scoring model(s) applied, and responses assigned a score and measure of score confidence (i.e., an error estimate based on response features). 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 on a day-to-day basis 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 MI's handscoring system.

Table 88. 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, 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, 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 89. Model Setting

MI RECOMMENDED VALUES	FLAG IMPACTED	DESCRIPTION	VALUES
MIN_WORDS = 0–15	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 = PREDC,LIST027,5,LIST028, 3,X_ALERT0,1,X_ALERT1, 2,X_ALERT2,3,X_ALERT3, 1	500 501–599	Current setting (PREDC...1) is for the standard alert scan.	Standard settings in place
PLAG = prompt.txt, 0.5	620	Prompt text is provided by the client and included in model configuration.	50% characters triggers 620

Scoring Infrastructure

During the automated scoring process, response data are transferred from CAI to MI's IT project team. They 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 were then transferred from PEG to the IT project team, who ultimately sent the data/scores back to CAI.

Quality Assurance

MI's hybrid scoring approach included numerous quality assurance steps. First, 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. Second, MI conducted a secondary validation using the first 500 student responses received during the administration window to confirm that each model performed as expected on 2021–2022 operational responses. Third, 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 formal process for informing clients when student responses reflect a possibly dangerous situation for the test taker. Specifically, MI employed a set of alert procedures to notify the client of responses indicating endangerment, abuse, or psychological and/or emotional difficulties. PEG employed a rule-based detection system to flag responses that are indicative of potentially dangerous situations. Responses flagged by PEG as possible alerts were reviewed by scoring leadership, who decided whether each response should be forwarded to the client. Once vetted, all alerts were provided to CAI, who associated the pertinent student information with the response(s) and contacts the state. In addition, CAI separately evaluates all responses and student-generated text for possible alerts.

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.

7.8.4 Human-Machine Agreement

This section summarizes the human-machine agreement for all items scored using a hybrid process in spring 2022, including (1) items passing initial model validation, (2) items passing secondary validation, and (3) items passing live validation.

Tables 90 through 92 present the human-machine agreement on the initial and secondary validation samples for ELA/L SA items, ELA/L essay items, and mathematics SA items, respectively.

Table 90. 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 & Adj.	QWK	Number of Items	% Exact	% Exact & Adj.	QWK
3	9	81.0	99.7	0.83	9	80.9	99.2	0.73
4	11	80.5	99.8	0.85	11	77.7	99.3	0.77
5	7	74.9	99.9	0.84	7	75.8	99.6	0.79
6	24	77.9	99.7	0.79	24	78.8	99.6	0.74
7	32	77.2	99.6	0.78	32	78.3	99.4	0.73
8	28	76.5	99.6	0.79	28	77.2	99.5	0.74
11	47	76.6	99.6	0.79	47	75.4	99.4	0.75

Table 91. 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 & Adj.**	**QWK**	**Number of Items**	**% Exact**	**% Exact & Adj.**	**QWK**
3	Conventions	9	73.8	99.4	0.75	9	69.5	99.6	0.71
3	Evid/Elab	9	75.2	98.9	0.80	9	77.5	99.2	0.75
3	Org/Purp	9	75.3	98.9	0.80	9	77.1	99.2	0.74
4	Conventions	14	70.8	99.5	0.76	14	68.5	99.3	0.71
4	Evid/Elab	14	73.9	99.3	0.83	14	73.1	99.5	0.77
4	Org/Purp	14	72.9	99.4	0.82	14	73.2	99.5	0.78
5	Conventions	7	73.1	99.7	0.69	7	69.4	99.6	0.68
5	Evid/Elab	7	73.8	99.1	0.82	7	76.0	99.4	0.79
5	Org/Purp	7	73.2	99.5	0.83	7	75.0	99.5	0.78
6	Conventions	9	75.4	99.3	0.73	9	70.5	98.7	0.71
6	Evid/Elab	9	71.7	98.5	0.78	9	76.6	99.6	0.79
6	Org/Purp	9	70.8	99.3	0.79	9	76.8	99.6	0.79
7	Conventions	11	76.8	99.6	0.71	11	73.8	99.6	0.71
7	Evid/Elab	11	75.1	99.5	0.83	11	77.1	99.7	0.79
7	Org/Purp	11	74.9	99.7	0.84	11	77.5	99.6	0.79
8	Conventions	14	77.2	99.2	0.71	14	74.4	99.6	0.71
8	Evid/Elab	14	73.9	99.3	0.83	14	78.7	99.8	0.81
8	Org/Purp	14	74.7	99.5	0.84	14	79.2	99.9	0.82
11	Conventions	15	79.6	99.7	0.75	15	78.1	99.5	0.71
11	Evid/Elab	15	76.6	99.6	0.86	15	72.8	99.8	0.81
11	Org/Purp	15	76.4	99.7	0.86	15	72.4	99.7	0.81

Table 92. 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 & Adj.	QWK	Number of Items	% Exact	% Exact & Adj.	QWK[a]
3	0-1	6	94.1	100	0.89	6	93.5	100	NA
4	0-1	10	90.9	100	0.8	10	92.7	100	NA
5	0-1	5	94.2	100	0.83	5	95.3	100	NA
6	0-1	4	99.6	100	0.97	4	99.8	100	NA
7	0-1	3	97.6	100	0.85	3	99.2	100	NA
8	0-1	3	87.6	100	0.73	3	93.6	100	NA
11	0-1	12	95.3	100	0.85	12	93.9	100	NA
3	0-2	20	91.6	99.5	0.93	20	91.4	99.7	0.91
4	0-2	24	91.1	99.8	0.92	24	92.5	99.7	0.89
5	0-2	25	87.5	99.6	0.89	25	87.6	99.6	0.83
6	0-2	13	89.7	99.8	0.90	13	89.9	99.9	0.88
7	0-2	11	88.7	99.7	0.86	11	90.7	99.9	0.82
8	0-2	12	89.4	99.7	0.90	12	91.3	99.6	0.85
11	0-2	7	84.4	99.4	0.84	7	82.8	99.4	0.79
3	0-3	4	92.6	100	0.98	4	94.2	99.4	0.98
4	0-3	4	87.9	99.8	0.94	4	84.3	99.2	0.90
5	0-3	2	90.9	98.4	0.94	2	87.9	98.4	0.90

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

Tables 93 through 95 present the HM agreement 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 a secondary validation since it involved operational data.

Table 93. Human-Machine Agreement for ELA/L Short-Answer Items on Live Validation Sample, by Grade

Grade	Live Validation			
	Number of Items	% Exact	% Exact & Adj.	QWK
6	4	74.2	99.6	0.78
7	8	72.9	99.1	0.74
8	15	74.4	99.2	0.77
11	1	76.5	100	0.78

Table 94. Human-Machine Agreement for ELA/L Essay Items on Live Validation Sample, by Grade

Grade	Trait	Live Validation			
		Number of Items	**% Exact**	**% Exact & Adj.**	**QWK**
4	Conventions	1	65.2	98.4	0.71
4	Evid/Elab	1	69.5	99.6	0.84
4	Org/Purp	1	69.3	98.4	0.81
7	Conventions	1	68.2	99.6	0.68
7	Evid/Elab	1	75.2	99.6	0.84
7	Org/Purp	1	77.3	99.6	0.86
8	Conventions	1	80.6	100	0.72
8	Evid/Elab	1	75.2	99.6	0.88
8	Org/Purp	1	73.2	100	0.88

Table 95. Human-Machine Agreement for Mathematics Items on Live Validation Samples, by Grade

Grade	Score Point Range	Live Validation			
		Number of Items	**% Exact**	**% Exact & Adj.**	**QWK[a]**
3	0-1	2	93.5	100	NA
6	0-1	3	95.2	100	NA
7	0-1	3	98.1	100	NA
8	0-1	2	87.9	100	NA
11	0-1	1	93.0	100	NA
3	0-2	4	86.9	98.5	0.87
4	0-2	5	91.8	99.7	0.91
5	0-2	5	90.3	99.8	0.90
6	0-2	12	89.6	99.5	0.86
8	0-2	4	86.9	99.0	0.77
11	0-2	4	95.6	99.3	0.91
3	0-3	2	88.0	99.5	0.93
11	0-3	1	74.0	96.7	0.82

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

7.8.5 Recommendations

The primary recommendation following the spring 2020 administration was to increase the amount of automated scoring to provide greater value to those states using hybrid scoring. MI made substantial strides in increasing the number of automated scoring models by automating its training procedures and by creating models for all independent items with sufficient training responses. The present results indicate success in this area, as 46.1% (481/1,044) of handscored items were scored using a hybrid process in 2022 vs. 7.7% (85/1,100) in 2021.

There are several new recommendations for future administrations. In spring 2022, the average HH agreement remained lower than observed during pre-pandemic administrations. Extending scoring and reporting timelines would allow for more practice per rater and likely support greater accuracy. If this is not possible, hiring additional raters will be required to better position MI to score the majority of responses in a short period of time. Next year, MI should revisit pay rates and incentives in light of 2023 market conditions to optimally attract and retain this population. In addition, MI should consider additional assessments of rater quality that can be administered to raters immediately after qualification.

8. REPORTING AND INTREPRETING 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.

8.1 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 the 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 96 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 96. 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 97 presents the types of subgroups and subgroup categories provided in the CRS.

Table 97. Types of Subgroups

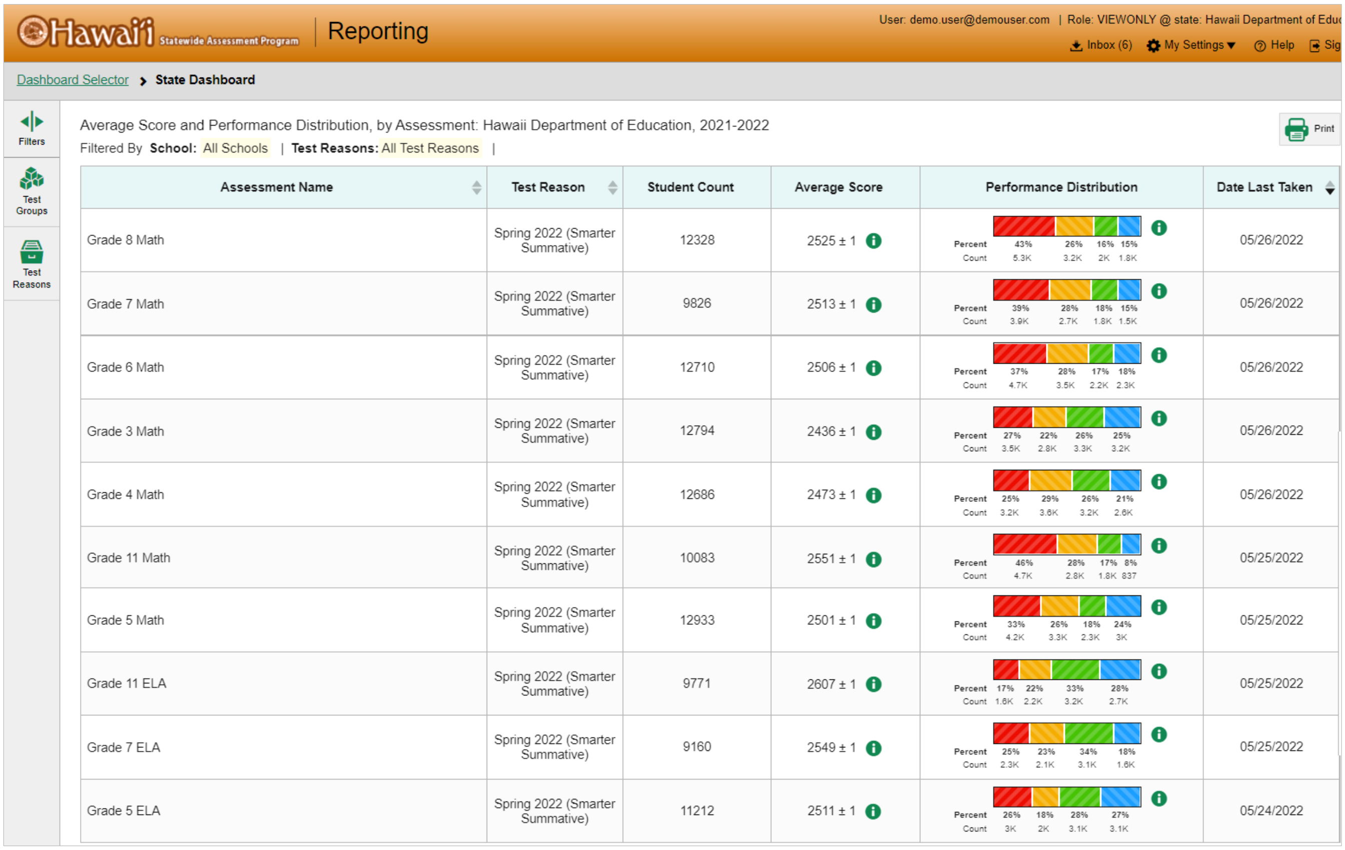
Subgroup	**Subgroup Category**
Gender	Male Female
ELL	ELL Not ELL
Disability	With Disability No Disability
Migrant Status	Migrant Not Migrant
Disadvantaged	Disadvantaged Not Disadvantaged
Ethnicity	American Indian/Alaskan Native Asian/Pacific Islander African American Hispanic Hawai'i Pacific Islander White Multi-Racial

8.1.1 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


Hawai'i Statewide Assessment Program
Reporting
User: demo.user@demouser.com | Role: VIEWONLY @ state: Hawaii Department of Educ
Inbox (6)
My Settings ▾
Help
Sig
Dashboard Selector
State Dashboard
Filters
Test Groups
Test Reasons
Print

Average Score and Performance Distribution, by Assessment: Hawaii Department of Education, 2021-2022

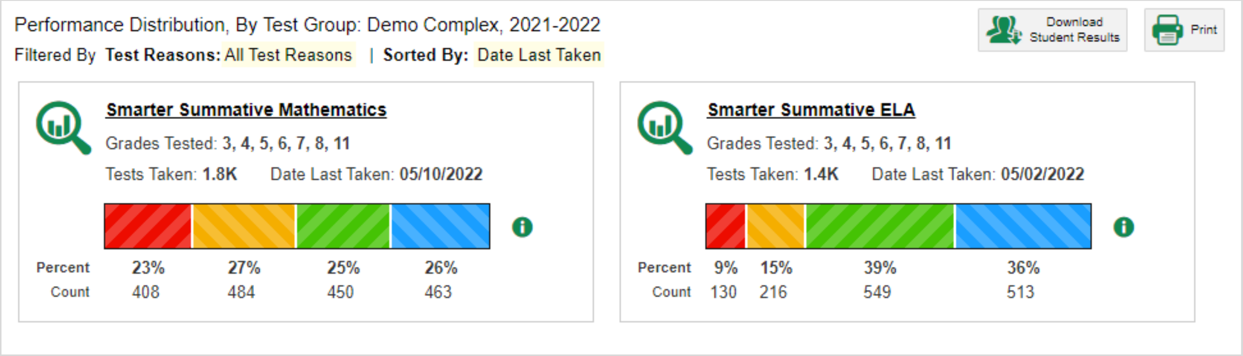
Filtered By **School:** All Schools | **Test Reasons:** All Test Reasons |

Assessment Name	Test Reason	Student Count	Average Score	Performance Distribution	Date Last Taken
Grade 8 Math	Spring 2022 (Smarter Summative)	12328	2525 ± 1	Percent 43% 26% 16% 15% Count 5.3K 3.2K 2K 1.8K	05/26/2022
Grade 7 Math	Spring 2022 (Smarter Summative)	9826	2513 ± 1	Percent 39% 28% 18% 15% Count 3.9K 2.7K 1.8K 1.5K	05/26/2022
Grade 6 Math	Spring 2022 (Smarter Summative)	12710	2506 ± 1	Percent 37% 28% 17% 18% Count 4.7K 3.5K 2.2K 2.3K	05/26/2022
Grade 3 Math	Spring 2022 (Smarter Summative)	12794	2436 ± 1	Percent 27% 22% 26% 25% Count 3.5K 2.8K 3.3K 3.2K	05/26/2022
Grade 4 Math	Spring 2022 (Smarter Summative)	12686	2473 ± 1	Percent 25% 29% 26% 21% Count 3.2K 3.6K 3.2K 2.6K	05/26/2022
Grade 11 Math	Spring 2022 (Smarter Summative)	10083	2551 ± 1	Percent 46% 28% 17% 8% Count 4.7K 2.8K 1.8K 837	05/25/2022
Grade 5 Math	Spring 2022 (Smarter Summative)	12933	2501 ± 1	Percent 33% 26% 18% 24% Count 4.2K 3.3K 2.3K 3K	05/25/2022
Grade 11 ELA	Spring 2022 (Smarter Summative)	9771	2607 ± 1	Percent 17% 22% 33% 28% Count 1.6K 2.2K 3.2K 2.7K	05/25/2022
Grade 7 ELA	Spring 2022 (Smarter Summative)	9160	2549 ± 1	Percent 25% 23% 34% 18% Count 2.3K 2.1K 3.1K 1.6K	05/25/2022
Grade 5 ELA	Spring 2022 (Smarter Summative)	11212	2511 ± 1	Percent 26% 18% 28% 27% Count 3K 2K 3.1K 3.1K	05/24/2022

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 Complex, 2021-2022
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: **1.8K** Date Last Taken: **05/10/2022**

Percent	23%	27%	25%	26%
Count	408	484	450	463

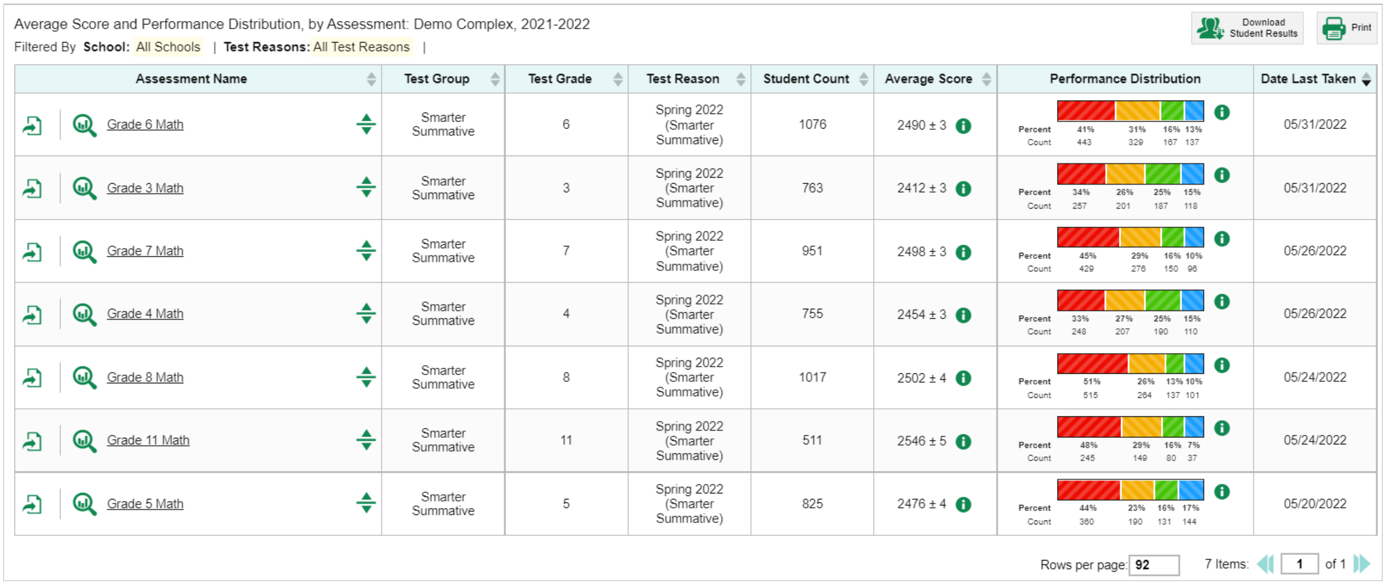
Smarter Summative ELA
Grades Tested: **3, 4, 5, 6, 7, 8, 11**
Tests Taken: **1.4K** Date Last Taken: **05/02/2022**

Percent	9%	15%	39%	36%
Count	130	216	549	513

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



Average Score and Performance Distribution, by Assessment: Demo Complex, 2021-2022
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 6 Math	Smarter Summative	6	Spring 2022 (Smarter Summative)	1076	2490 ± 3	Percent 41% 31% 16% 13% Count 443 329 167 137	05/31/2022
Grade 3 Math	Smarter Summative	3	Spring 2022 (Smarter Summative)	763	2412 ± 3	Percent 34% 26% 25% 15% Count 257 201 187 118	05/31/2022
Grade 7 Math	Smarter Summative	7	Spring 2022 (Smarter Summative)	951	2498 ± 3	Percent 45% 29% 16% 10% Count 429 276 150 96	05/26/2022
Grade 4 Math	Smarter Summative	4	Spring 2022 (Smarter Summative)	755	2454 ± 3	Percent 33% 27% 25% 15% Count 248 207 190 110	05/26/2022
Grade 8 Math	Smarter Summative	8	Spring 2022 (Smarter Summative)	1017	2502 ± 4	Percent 51% 26% 13% 10% Count 515 264 137 101	05/24/2022
Grade 11 Math	Smarter Summative	11	Spring 2022 (Smarter Summative)	511	2546 ± 5	Percent 48% 29% 16% 7% Count 245 149 80 37	05/24/2022
Grade 5 Math	Smarter Summative	5	Spring 2022 (Smarter Summative)	825	2476 ± 4	Percent 44% 23% 16% 17% Count 360 190 131 144	05/20/2022

Rows per page: 92 7 Items: 1 of 1

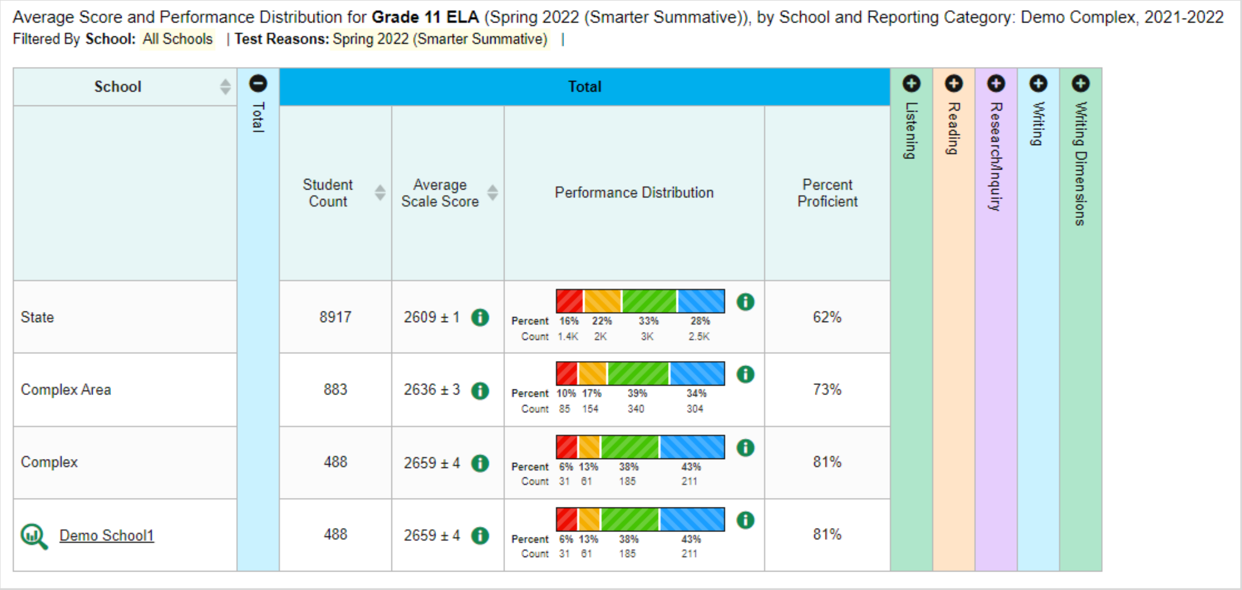
8.1.2 Aggregate Score Reports: Overall Performance

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 ELA/L at the complex level, and Exhibit 5 presents an example summary for grade 11 by gender.

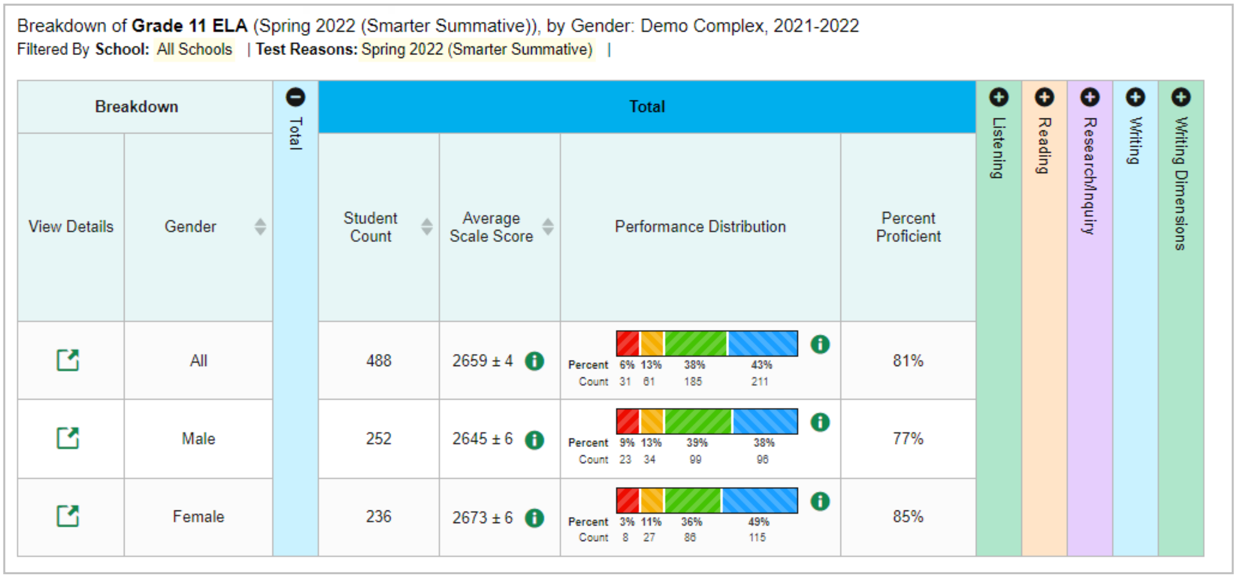
Exhibit 4. Overall Performance Summary Results for Grade 11 ELA/L: Complex Level



Average Score and Performance Distribution for **Grade 11 ELA** (Spring 2022 (Smarter Summative)), by School and Reporting Category: Demo Complex, 2021-2022
Filtered By **School:** All Schools | **Test Reasons:** Spring 2022 (Smarter Summative) |

School	Total	Total				Listening	Reading	ResearchInquiry	Writing	Writing Dimensions
		Student Count	Average Scale Score	Performance Distribution	Percent Proficient					
State		8917	2609 ± 1	Percent 16% 22% 33% 28% Count 1.4K 2K 3K 2.5K	62%					
Complex Area		883	2636 ± 3	Percent 10% 17% 39% 34% Count 85 154 340 304	73%					
Complex		488	2659 ± 4	Percent 6% 13% 38% 43% Count 31 61 185 211	81%					
Demo School1		488	2659 ± 4	Percent 6% 13% 38% 43% Count 31 61 185 211	81%					

Exhibit 5. Overall Performance Summary Results for Grade 11 ELA/L by Gender: Complex Level



Breakdown of **Grade 11 ELA** (Spring 2022 (Smarter Summative)), by Gender: Demo Complex, 2021-2022
Filtered By **School:** All Schools | **Test Reasons:** Spring 2022 (Smarter Summative) |

Breakdown		Total	Total				Listening	Reading	ResearchInquiry	Writing	Writing Dimensions
View Details	Gender		Student Count	Average Scale Score	Performance Distribution	Percent Proficient					
	All		488	2659 ± 4	Percent 6% 13% 38% 43% Count 31 61 185 211	81%					
	Male		252	2645 ± 6	Percent 9% 13% 39% 38% Count 23 34 99 96	77%					
	Female		236	2673 ± 6	Percent 3% 11% 36% 49% Count 8 27 86 115	85%					

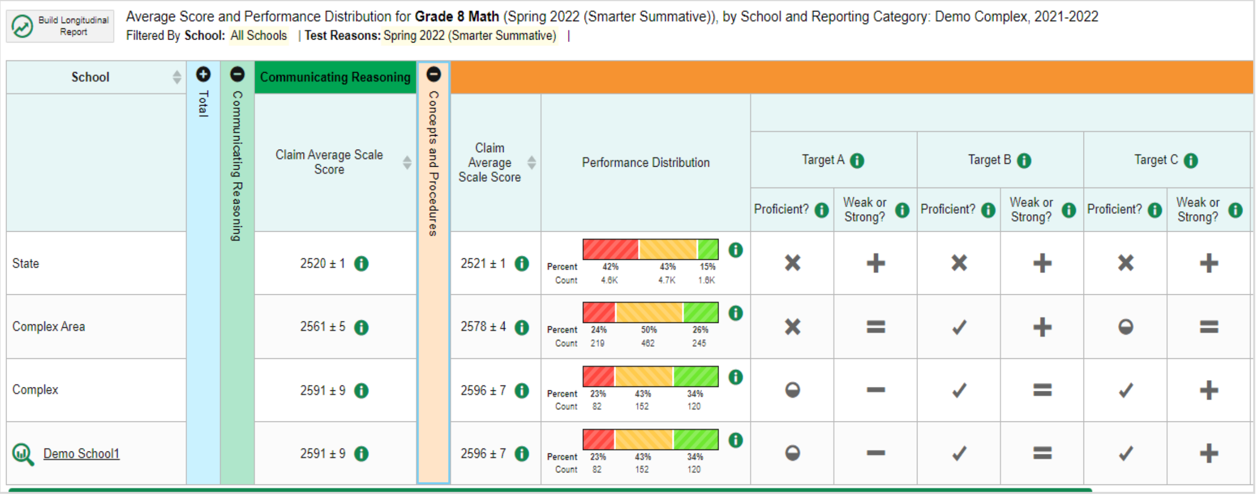
8.1.3 Aggregate Score Reports: Claim and Target Performance

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 8 mathematics at the complex level.

Exhibit 6. Claim and Target Level Results for Grade 8 Mathematics: Complex Level



Average Score and Performance Distribution for **Grade 8 Math** (Spring 2022 (Smarter Summative)), by School and Reporting Category: Demo Complex, 2021-2022
Filtered By **School:** All Schools | **Test Reasons:** Spring 2022 (Smarter Summative) |

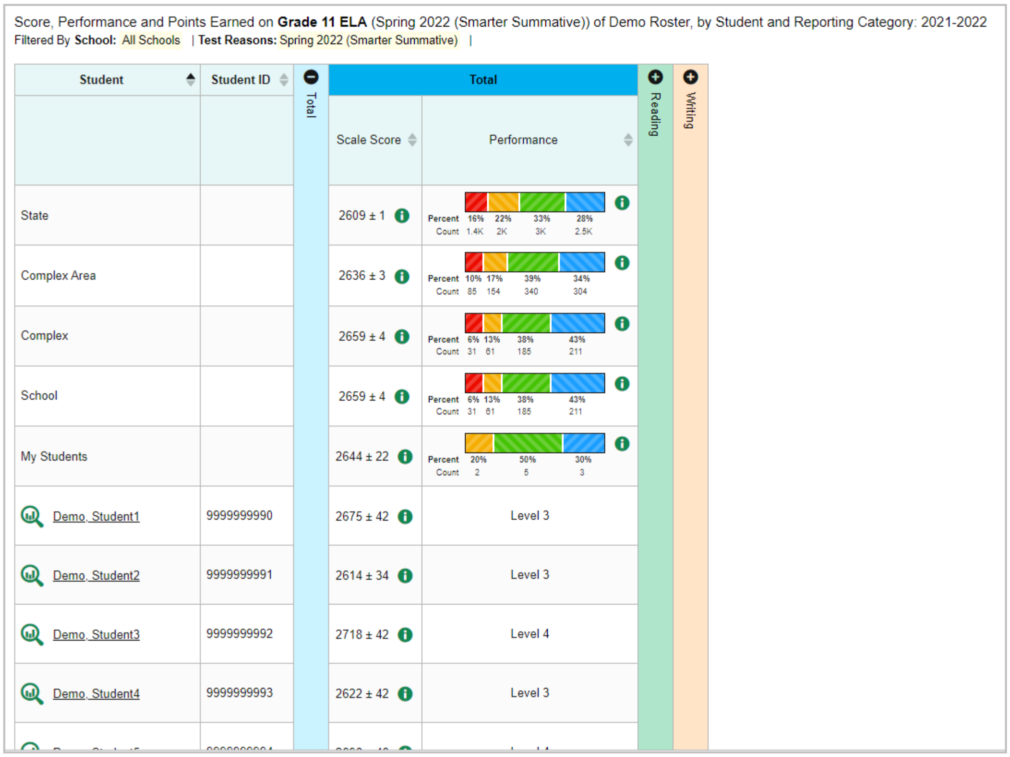
School	Total	Communicating Reasoning	Communicating Reasoning	Concepts and Procedures								
			Claim Average Scale Score		Claim Average Scale Score	Performance Distribution	Target A		Target B		Target C	
							Proficient?	Weak or Strong?	Proficient?	Weak or Strong?	Proficient?	Weak or Strong?
State			2520 ± 1		2521 ± 1	Percent 42% 43% 15% Count 4.6K 4.7K 1.8K	✗	+	✗	+	✗	+
Complex Area			2561 ± 5		2578 ± 4	Percent 24% 50% 26% Count 219 462 245	✗	=	✓	+	◒	=
Complex			2591 ± 9		2596 ± 7	Percent 23% 43% 34% Count 82 152 120	◒	−	✓	=	✓	+
Demo School1			2591 ± 9		2596 ± 7	Percent 23% 43% 34% Count 82 152 120	◒	−	✓	=	✓	+

8.1.4 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 ELA/L summative assessment.

Exhibit 7. Roster Performance Report for Grade 11 ELA/L



Score, Performance and Points Earned on **Grade 11 ELA** (Spring 2022 (Smarter Summative)) of Demo Roster, by Student and Reporting Category: 2021-2022
Filtered By **School:** All Schools | **Test Reasons:** Spring 2022 (Smarter Summative) |

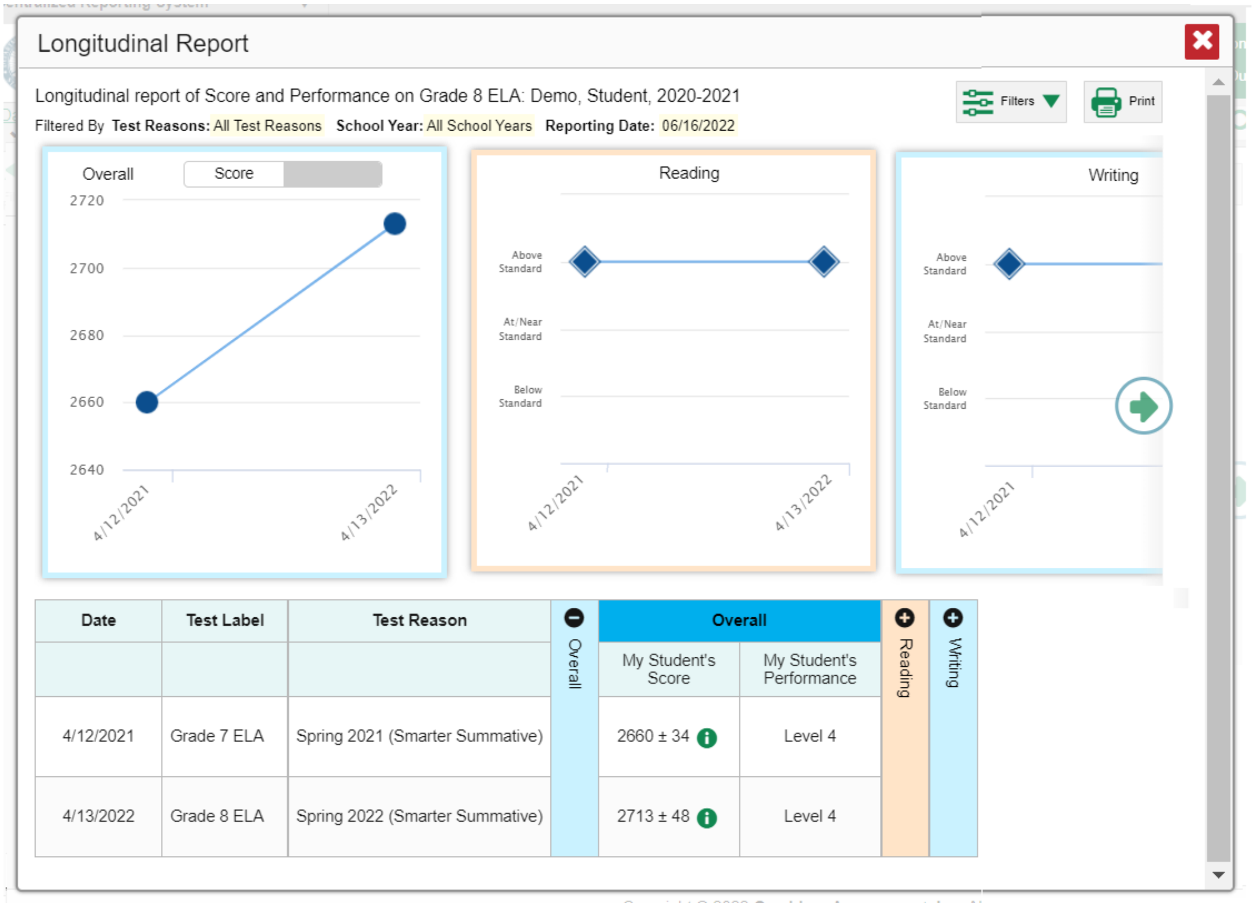
Student	Student ID	Total	Total		Reading	Writing
			Scale Score	Performance		
State			2609 ± 1	Percent 16% 22% 33% 28% Count 1.4K 2K 3K 2.5K		
Complex Area			2636 ± 3	Percent 10% 17% 39% 34% Count 85 154 340 304		
Complex			2659 ± 4	Percent 6% 13% 38% 43% Count 31 61 185 211		
School			2659 ± 4	Percent 6% 13% 38% 43% Count 31 61 185 211		
My Students			2644 ± 22	Percent 20% 50% 30% Count 2 5 3		
Demo, Student1	9999999990		2675 ± 42	Level 3		
Demo, Student2	9999999991		2614 ± 34	Level 3		
Demo, Student3	9999999992		2718 ± 42	Level 4		
Demo, Student4	9999999993		2622 ± 42	Level 3		

8.1.5 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 for overall score and by claim score.

Exhibit 8 presents an example trend report page for ELA/L at the individual student level.

Exhibit 8. Trend Report for ELA/L: Student Level



Longitudinal report of Score and Performance on Grade 8 ELA: Demo, Student, 2020-2021

Filtered By **Test Reasons:** All Test Reasons **School Year:** All School Years **Reporting Date:** 06/16/2022

Date	Test Label	Test Reason	Overall	Overall		Reading	Writing
				My Student's Score	My Student's Performance		
4/12/2021	Grade 7 ELA	Spring 2021 (Smarter Summative)		2660 ± 34	Level 4		
4/13/2022	Grade 8 ELA	Spring 2022 (Smarter Summative)		2713 ± 48	Level 4		

8.1.6 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).

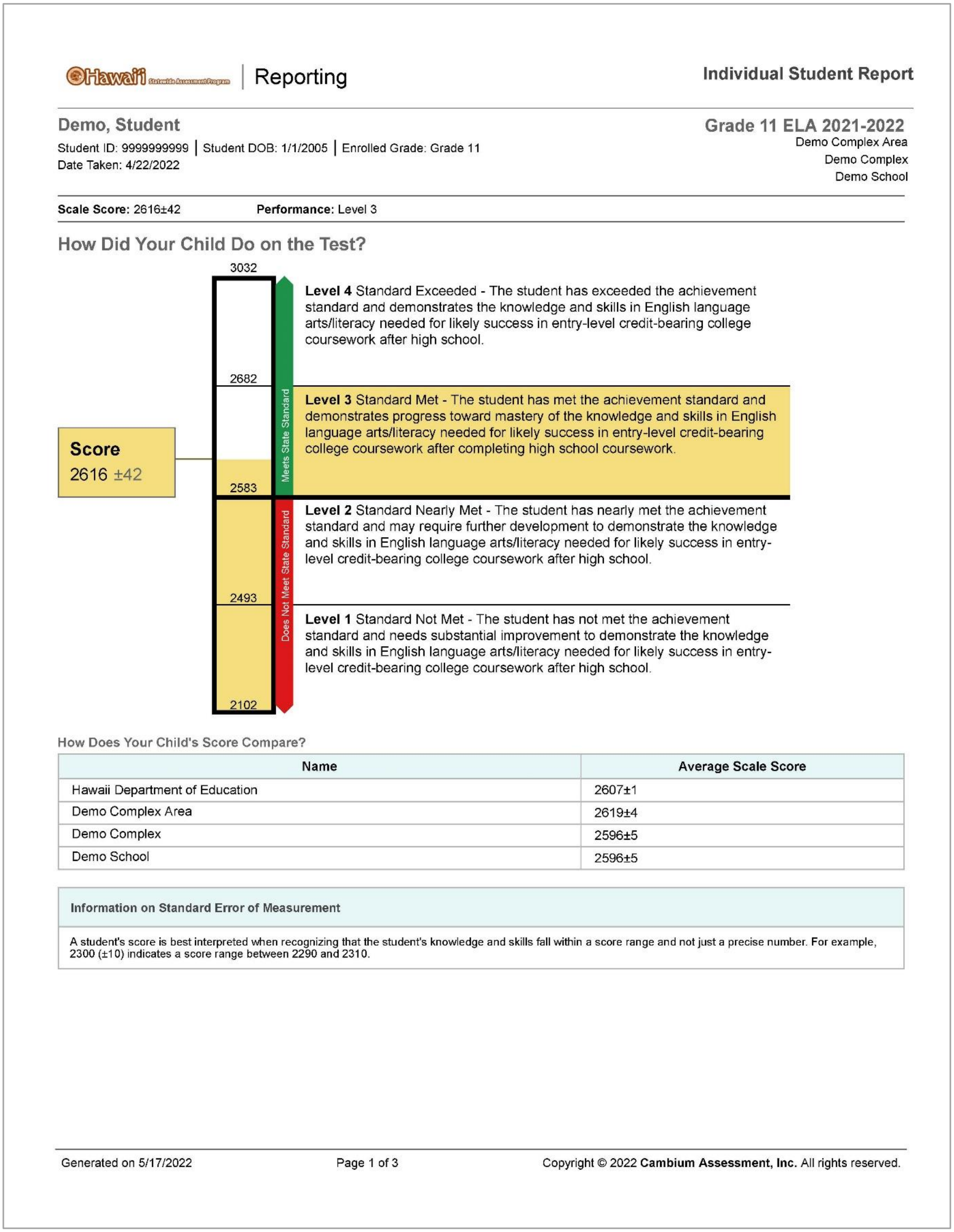
On the first page of the ISR, the student's name, scale score with the SEM, achievement level, and reported Lexile® measure for ELA/L are shown at the top of the page. In the middle section, the student's performance is described in detail using a barrel chart. In the barrel chart, the student's scale 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 next 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. The third page shows the student's performance on claims (i.e., Claims 1 and 2 for ELA/L 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.

Exhibit 9 presents a sample ISR for grade 11 ELA/L.

Exhibit 9. Individual Student Report for Grade 11 ELA/L



Hawaii Statewide Assessment Program | Reporting

Individual Student Report

Demo, Student

Student ID: 9999999999 | Student DOB: 1/1/2005 | Enrolled Grade: Grade 11
Date Taken: 4/22/2022

Grade 11 ELA 2021-2022
Demo Complex Area
Demo Complex
Demo School

Scale Score: 2616±42 **Performance:** Level 3

How Did Your Child Do on the Test?

Score
2616 ±42

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

Level 3 Standard Met - 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 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 English language arts/literacy 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 English language arts/literacy needed for likely success in entry-level credit-bearing college coursework after high school.

How Does Your Child's Score Compare?

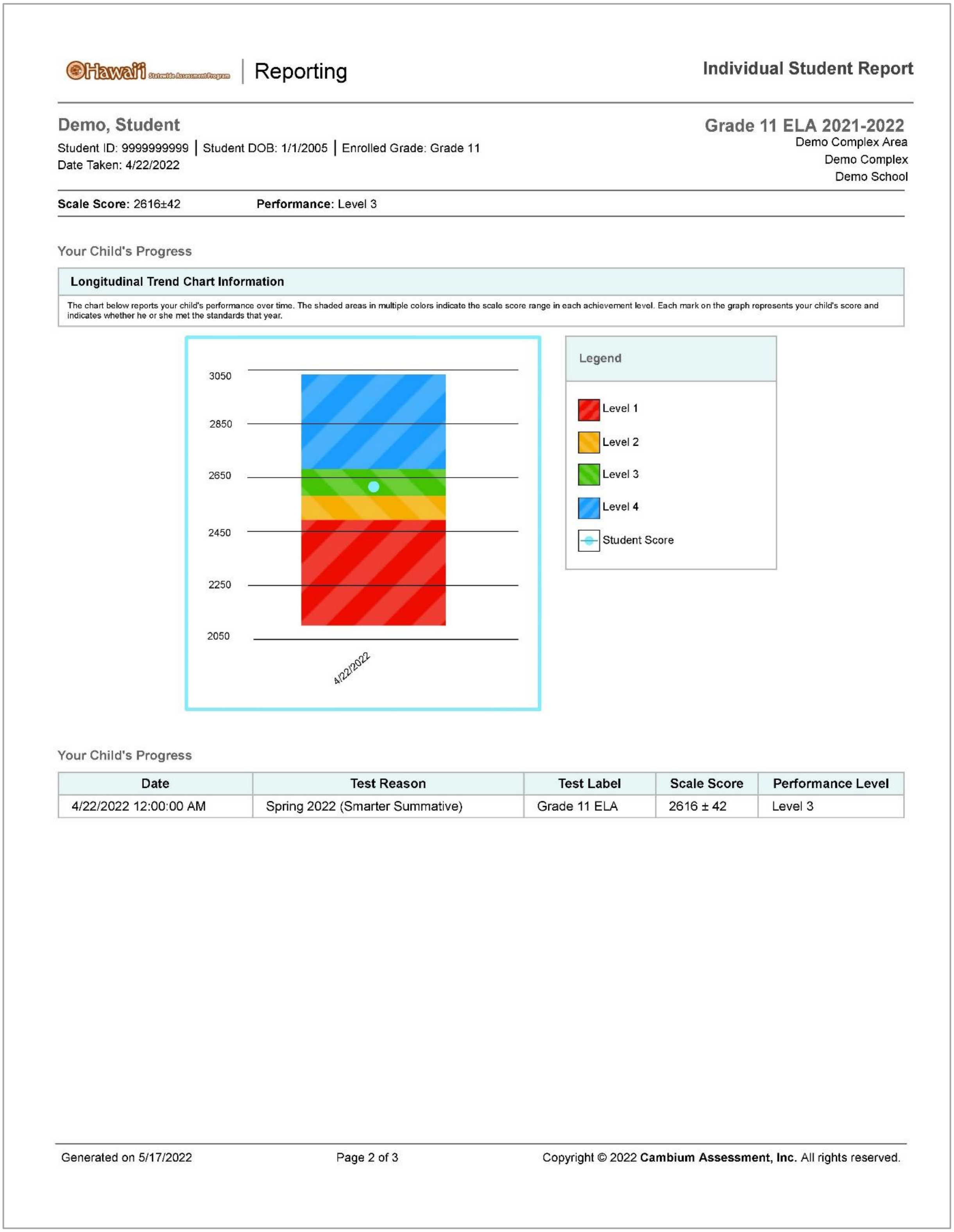
Name	Average Scale Score
Hawaii Department of Education	2607±1
Demo Complex Area	2619±4
Demo Complex	2596±5
Demo School	2596±5

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 5/17/2022 Page 1 of 3 Copyright © 2022 **Cambium Assessment, Inc.** All rights reserved.

Exhibit 9. Individual Student Report for Grade 11 ELA/L (Continued)



Hawaii Statewide Assessment Program | Reporting

Individual Student Report

Demo, Student

Student ID: 9999999999 | Student DOB: 1/1/2005 | Enrolled Grade: Grade 11
Date Taken: 4/22/2022

Grade 11 ELA 2021-2022
Demo Complex Area
Demo Complex
Demo School

Scale Score: 2616±42 **Performance:** Level 3

Your Child's Progress

Longitudinal Trend Chart Information

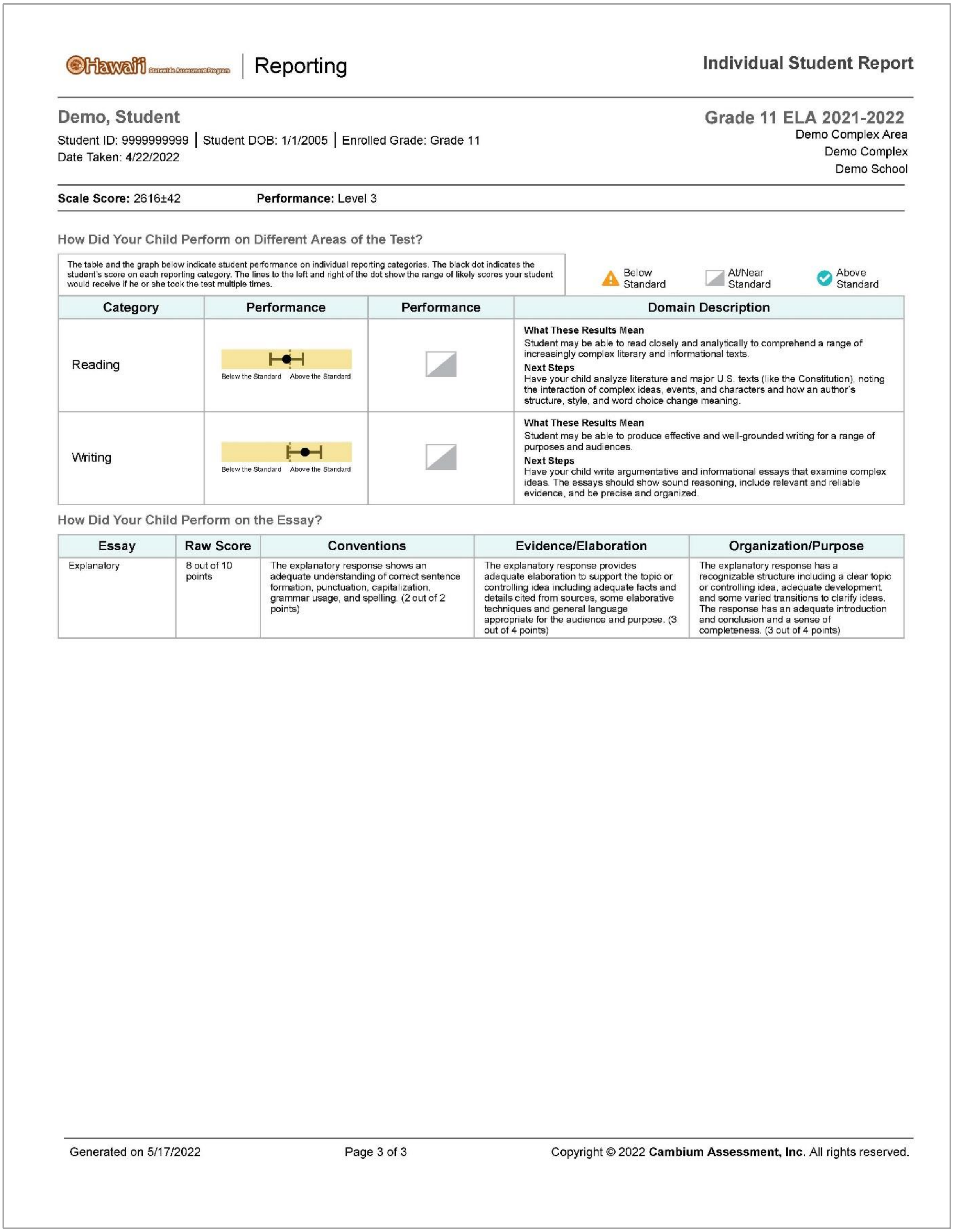
The chart below reports your child's performance over time. The shaded areas in multiple colors indicate the scale score range in each achievement level. Each mark on the graph represents your child's score and indicates whether he or she met the standards that year.

Your Child's Progress

Date	Test Reason	Test Label	Scale Score	Performance Level
4/22/2022 12:00:00 AM	Spring 2022 (Smarter Summative)	Grade 11 ELA	2616 ± 42	Level 3

Generated on 5/17/2022 Page 2 of 3 Copyright © 2022 **Cambium Assessment, Inc.** All rights reserved.

Exhibit 9. Individual Student Report for Grade 11 ELA/L (Continued)



Reporting

Individual Student Report

Demo, Student

Student ID: 9999999999 | Student DOB: 1/1/2005 | Enrolled Grade: Grade 11
Date Taken: 4/22/2022

Grade 11 ELA 2021-2022
Demo Complex Area
Demo Complex
Demo School

Scale Score: 2616±42 **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	Domain Description
Reading	Below the Standard Above the Standard	◪	**What These Results Mean** Student may be able to read closely and analytically to comprehend a range of increasingly complex literary and informational texts. **Next Steps** Have your child analyze literature and major U.S. texts (like the Constitution), noting the interaction of complex ideas, events, and characters and how an author's structure, style, and word choice change meaning.
Writing	Below the Standard Above the Standard	◪	**What These Results Mean** Student may be able to produce effective and well-grounded writing for a range of purposes and audiences. **Next Steps** Have your child write argumentative and informational essays that examine complex ideas. The essays should show sound reasoning, include relevant and reliable evidence, and be precise and organized.

How Did Your Child Perform on the Essay?

Essay	Raw Score	Conventions	Evidence/Elaboration	Organization/Purpose
Explanatory	8 out of 10 points	The explanatory response shows an adequate understanding of correct sentence formation, punctuation, capitalization, grammar usage, and spelling. (2 out of 2 points)	The explanatory response provides adequate elaboration to support the topic or controlling idea including adequate facts and details cited from sources, some elaborative techniques and general language appropriate for the audience and purpose. (3 out of 4 points)	The explanatory response has a recognizable structure including a clear topic or controlling idea, adequate development, and some varied transitions to clarify ideas. The response has an adequate introduction and conclusion and a sense of completeness. (3 out of 4 points)

Generated on 5/17/2022 Page 3 of 3 Copyright © 2022 **Cambium Assessment, Inc.** All rights reserved.

8.2 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.

8.2.1 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.

8.2.2 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.

8.2.3 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 considered to be on track to demonstrate progress toward mastery of the knowledge and skills necessary for college and career readiness.

8.2.4 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.

8.2.5 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.

8.2.6 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.

8.3 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' placement and retention, or teachers' instructional planning and implementation, the assessment results should not be used as the only source of information. Given that assessment results measured by a test provide limited information, other sources on student achievement such as classroom assessment and teacher evaluation should be considered when making decisions on student learning. Finally, when student performance is compared across groups, users need to consider the group size. The smaller the group size, the larger the measurement error related to these aggregate data, thus requiring interpretation with more caution.

9. 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.

9.1 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.

9.1.1 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.

9.1.2 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.

9.2 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.

9.3 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.

9.4 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 98 presents an overview of the QA reports.

Table 98. 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

9.4.1 Score Report Quality Check

Two types of score reports were produced in the Smarter Balanced summative assessments: (1) online reports, and (2) printed reports (family reports only).

9.4.1.1 Online Report Quality Assurance

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.

9.4.1.2 Paper Report Quality Assurance

Statistical Programming

The family reports contain custom programming and require rigorous QA processes to ensure their accuracy. All custom programming is guided by the detailed and precise specifications outlined in CAI's reporting specifications document. Analytic rules are programmed upon approval of the specifications, and each program is extensively tested on test decks and real data from other programs. The final programs are reviewed by two senior statisticians and one senior programmer to ensure that they implemented agreed-on procedures. Custom programming is implemented independently by two statistical programming teams working from the specifications. The scripts are released for production only when the output from both teams matches precisely.

Much of the statistical processing is repeated, and CAI has implemented a structured software development process to ensure that the repeated tasks are implemented correctly and identically each time. Small programs (called *macros*) are written to take specified data as input and produce data sets containing derived variables as output. Approximately 30 such macros reside in CAI's library for score reports. Each macro is extensively tested and stored in a central development server. Once a macro is tested and stored, changes to the macro must be approved by the director of score reporting, the director of psychometrics, and the project directors for affected projects.

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

Display Programming

The paper report development process uses graphical programming, which takes place in a Xerox-developed programming language called Variable Data Intelligent PostScript Printware (VIPP) and allows virtually infinite control of the visual appearance of the reports. After our designers create backgrounds, CAI's VIPP programmers write code that indicates where to place all variable information (data, graphics, and text) on the reports. The VIPP code is tested using both artificial and real data. CAI's data generation utilities can read the output layout specifications and generate artificial data for direct input into the VIPP programs. This allows testing of these programs to begin before the statistical programming is complete. In later stages, artificial data are generated according to the input layout and are run through the psychometric process and the score reporting statistical programs, and the output is formatted as VIPP input. This process enables CAI to test the entire system.

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

All reports containing actual data are stored in a locked storage area. Before the reports are printed, CAI provides a live data file and individual student reports with sample complex areas for HIDOE staff review. CAI will work closely with the Hawai'i to resolve questions and correct any problems. The reports will not be delivered unless the Department approves the sample reports and data file.

REFERENCES

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

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

Bunch, M. B., Vaughn, D., & Miel, S. (2016). Automated scoring in assessment systems. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.). *Technology tools for real-world skill development* (pp. 611–626). Hershey, PA: IGI Global.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

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.

Keith, T. Z. (2003). Validity and automated essay scoring systems. In M. D. Shermis & J. C. Burstein (Eds.). *Automated essay scoring: A cross-disciplinary perspective* (pp. 147–167). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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.

Shermis, M. D., Koch, C. M., Page, E. B., Keith, T. Z., & Harrington, S. (2002). Trait ratings for automated essay grading. *Educational and Psychological Measurement*, *62*(1), 5–18.

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

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

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

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

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

APPENDICES

Appendix A: Summary of the 2021–2022 Interim Assessments

The Interim Comprehensive Assessments (ICAs) were fixed-form tests for each grade and subject. Most students took ICAs once, but some students took them multiple times. Table A-1 presents the number of students who took ICAs by the number of attempts. Total number of tests indicate the total ICA tests taken by the total number of students, counting multiple attempts as multiple tests. For example, if a student took ICAs twice, the number of tests for this student is counted as two. Table A-2 summarizes student performance on ICAs for all tests taken, including the average and the standard deviation of scale scores, the percentage of tests in each achievement level, and the percentage of proficient tests.

Table A-1. Number of Students Who Took ICAs

Grade	Number of Students by Number of Attempts						Total Number of Tests Taken
	Once	Twice	Three Times	Four Times	Five Times	Total Number of Students	
ELA/L							
3	2,633	146	2	12	0	2,793	2,979
4	2,671	9	5	8	0	2,693	2,736
5	2,706	26	1	0	0	2,733	2,761
6	947	24	3	0	0	974	1,004
7	360	0	0	0	0	360	360
8	612	0	0	0	0	612	612
9	1	0	0	0	0	1	1
10	29	0	0	0	0	29	29
11	515	37	0	0	0	552	589
Mathematics							
3	2,956	182	36	10	8	3,192	3,508
4	3,178	96	11	1	15	3,301	3,482
5	2,645	141	12	9	0	2,807	2,999
6	1,642	95	1	0	0	1,738	1,835
7	536	53	1	0	0	590	645
8	799	16	0	0	0	815	831
9	78	0	0	0	0	78	78
10	163	0	0	0	0	163	163
11	757	6	0	0	0	763	769

Table A-2. ICA ELA/L and Mathematics Percentage of Tests in Achievement Levels

Subject	Grade	Total Number of Tests Taken	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
ELA/L	3	2,979	2418.98	89.25	30	25	22	23	45
	4	2,736	2448.76	93.80	38	21	21	20	41
	5	2,761	2498.78	104.24	29	18	31	22	53
	6	1,004	2502.94	107.82	33	25	25	16	42
	7	360	2569.50	101.90	19	22	38	20	59
	8	612	2576.68	99.31	18	26	40	17	56
	9	1*							
	10	29	2554.77	99.59	24	28	45	3	48
	11	589	2595.78	123.16	20	24	30	26	55
Mathematics	3	3,508	2429.95	86.31	26	27	30	16	47
	4	3,482	2458.58	90.69	28	33	26	14	40
	5	2,999	2500.83	101.64	32	30	19	19	38
	6	1,835	2513.82	114.05	33	29	20	18	38
	7	645	2567.41	128.48	24	22	26	28	54
	8	831	2521.30	118.68	43	27	18	11	30
	9	78	2481.45	151.69	65	14	10	10	21
	10	163	2563.47	110.27	33	34	24	9	33
	11	769	2551.64	127.17	44	32	17	7	24

Note: The percentage of each achievement level may not add up to 100% or percentage proficient due to rounding.
***** Suppressed data due to the small sample size, n < 10.

For the Interim Assessment Blocks (IABs), there were 14 to 15 IABs for English language arts/literacy (ELA/L) and 10 to 15 IABs for mathematics. Students were allowed to take as many IABs as they wanted, and to take the same IAB multiple times. Table A-3 shows the total number of students who took at least one IAB and the number of students by the number of distinct IABs taken. For example, in grade 3 ELA/L, a total of 3,805 students took at least one IAB. Among 3,805 students, 1,192 students took one IAB, 849 students took two distinct IABs, and so on. Tables A-4 to A-11 disaggregate the number of students in Table A-3 by each individual block. For example, 1,192 students in grade 3 ELA/L took one IAB only. Among 1,192 students, 95 students took the Brief Writes IAB, 22 students took the Editing IAB, and so on.

Tables A-12 to A-17 summarize student performance on each IAB for all tests taken, including the percentage of tests in each performance category. The total number of tests indicates the total number of IAB tests taken by all students, counting multiple attempts as multiple tests. For example, if a student took the same IAB twice, the number of tests for this student is counted as two.

Table A-3. Number of Students Who Took Distinct IABs (Grades 3–8, 11)

Grade	**Total Students with At Least One IAB**	**Number of Distinct IABs Taken**													
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
ELA/L															
3	3,805	1,192	849	508	312	287	196	141	144	73	16	9	42	6	30
4	3,715	1,105	1,133	539	362	247	165	78	31	20	11	6	10	8	
5	3,096	962	802	430	248	249	147	68	46	29	37	17	32	9	20
6	2,136	1,004	556	143	148	77	85	26	28	16	4	1	4	7	37
7	1,244	761	296	83	94	10									
8	1,488	1,168	293	26	1										
11	2,246	1,832	215	150	6	12	31								
Mathematics															
3	3,760	906	711	658	378	360	264	193	126	74	54	36			
4	3,472	1,017	1,019	531	335	222	67	71	74	51	59	2	8	16	
5	3,355	982	727	442	401	306	111	54	77	43	67	145			
6	2,441	1,200	442	249	233	145	72	36	13	20	6	25			
7	1,442	651	281	266	117	79	20	12	11	5					
8	1,054	614	352	74	10	1	3								
11	2,666	1,865	583	65	34	21	17	26	18	9	28				

Table A-4: ELA/L Number of Students Who Took Distinct IABs by Block Labels (Grades 3–4)

Grade	**Block**	**Number of Distinct IABs Taken**													
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
3	Brief Writes	95	36	54	16	47	49	31	63	45	7	6	40	6	30
	Editing	22	67	44	94	82	144	59	88	60	15	8	42	6	30
	Language and Vocabulary Use	217	221	203	177	211	145	110	117	54	15	9	42	5	30
	Listen/Interpret	124	188	215	157	186	132	94	124	41	13	8	42	6	29
	Read Informational Texts	303	312	262	163	185	105	81	105	51	9	5	28	6	30
	Read Literary Texts	149	307	252	180	179	103	100	130	63	15	9	42	6	30
	Research	1	37	6	22	43	51	57	34	25	13	8	41	6	30
	Research: Analyze Information	89	69	93	103	80	107	97	93	50	14	9	42	6	30
	Research: Interpret and Integrate	7	73	47	73	81	96	78	55	48	9	9	41	6	30
	Research: Use Evidence	6	125	42	46	46	85	61	57	47	14	9	40	5	30
	Revision	1	25	36	13	36	28	26	37	19	8	3	35	4	30
	Write & Revise Informational Texts	5	20	17	24	42	16	42	78	22	5	3	15	3	12
	Write & Revise Narratives	21	22	72	23	49	34	85	49	48	3	4	18	5	30
	Write & Revise Opinion Texts	5	48	60	31	69	21	18	41	32	7	4	10	5	30
	Performance Task	147	148	121	126	99	60	48	81	52	13	5	26	3	19
4	Brief Writes	16	50	27	56	77	71	31	13	14	3	6	10	8	
	Editing	50	64	47	74	73	107	31	13	19	10	6	10	8	
	Language and Vocabulary Use	162	241	243	202	160	130	77	28	18	11	6	10	8	
	Listen/Interpret	222	374	228	263	161	118	46	22	7	10	4	10	8	
	Read Informational Texts	406	610	329	217	202	146	74	7	6	9	3	4	8	
	Read Literary Texts	104	594	234	159	191	111	50	22	8	11	6	10	8	
	Research	1	2	7	19	73	59	39	3	2	3	5	10	8	
	Research: Analyze Information	38	9	65	84	43	18	18	25	14	2	5	8	8	
	Research: Interpret and Integrate	10	26	95	33	30	27	23	22	14	2	6	10	8	
	Research: Use Evidence	4	22	57	87	33	21	34	24	19	9	4	6		
	Revision			4	28	43	35	21	8	9	9	5	10	8	
	Write & Revise Informational Texts	9	21	37	41	8	8	13	24	18	10	4	10	8	
	Write & Revise Narratives	13	25	45	29	11	10	5	10	16	10	2	4	8	
	Write & Revise Opinion Texts	22	39	45	2	3	9	15	21	12	3	4	8	8	
	Performance Task	48	189	154	154	127	120	69	6	4	8				

Table A-5: ELA/L Number of Students Who Took Distinct IABs by Block Labels (Grades 5–6)

Grade	**Block**	**Number of Distinct IABs Taken**													
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
5	Brief Writes	6	17	10	14	13	27	32	12	21	14	13	32	9	20
	Editing	28	62	70	100	85	54	19	25	26	33	14	30	9	20
	Language and Vocabulary Use	160	200	187	151	172	120	60	43	26	36	17	32	8	20
	Listen/Interpret	321	245	88	84	164	120	47	39	23	36	17	31	8	20
	Read Informational Texts	79	317	292	176	165	111	43	19	21	32	14	5	8	20
	Read Literary Texts	66	251	288	113	176	79	29	34	24	24	13	30	6	20
	Research	14	141	27	53	32	46	16	28	8	16	7	32	9	20
	Research: Analyze Information	52	82	86	87	108	93	63	33	23	28	13	32	9	20
	Research: Interpret and Integrate	4	26	56	50	69	62	53	31	15	34	17	32	9	20
	Research: Use Evidence	2	90	19	25	46	19	21	22	26	24	13	29	9	20
	Revision	6	16	23	21	24	14	13	20	10	27	12	32	9	20
	Write & Revise Informational Texts	2	15	2	4	17	2		2	1	6	4	2	4	12
	Write & Revise Narratives		24	40	28	28	69	42	22	9	16	11	31	9	20
	Write & Revise Opinion Texts	1	6	1	18	21	23	29	17	18	14	8	32	9	20
	Performance Task	221	112	101	68	125	43	9	21	10	30	14	2	2	8
6	Brief Writes		1	8	15		8	7	25	15	4	1	4	7	37
	Editing	9	104	17	10	14	62	4	18		1		4	7	37
	Language and Vocabulary Use	326	176	102	113	74	84	26	28	16	4	1	4	7	37
	Listen/Interpret	141	41	35	97	67	31	3	10	16	3	1	3	7	37
	Read Informational Texts	209	348	60	36	11	23	25	27	14	4	1	3	6	37
	Read Literary Texts	72	302	81	53	16	61	23	15	15	3	1	3	6	37
	Research	153	6	4	7	50	5	21	12	8	2	1	3	6	37
	Research: Analyze & Integrate Info	4	8	55	25	15	65	4	16	7	1		3	6	37
	Research: Evaluate Info & Sources	3	3	9	29		5	5	22	13	3		2	7	37
	Research: Use Evidence	32	16	17	38	16	79	20	14	14	3	1	4	6	37
	Revision				2		3	3	17	8	4	1	3	7	37
	Write & Revise Explanatory Texts	6	14	8	68	49				1	2	1	4	7	37
	Write & Revise Narratives	6	18	6	14	13	64	19	4	8	2	1	4	7	37
	Performance Task	43	75	27	85	60	20	22	16	9	4	1	4	5	37

Table A-6: ELA/L Number of Students Who Took Distinct IABs by Block Labels (Grades 7–8)

Grade	**Block**	**Number of Distinct IABs Taken**													
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
7	Brief Writes	32	1	1											
	Editing		2	16	88										
	Language and Vocabulary Use	204	43	16	88										
	Listen/Interpret		9	52											
	Read Informational Texts	154	242	65	4	10									
	Read Literary Texts	76	218	60	92	10									
	Research														
	Research: Analyze & Integrate Info	18													
	Research: Evaluate Info & Sources	13	50	14	6	10									
	Research: Use Evidence	27	26	19	4										
	Revision	2													
	Write & Revise Argumentative Texts	229	1	1	5	10									
	Write & Revise Explanatory Texts	6				1									
	Write & Revise Narratives			2	5	9									
	Performance Task			3	84										
8	Brief Writes														
	Edit/Revise	260	75	6											
	Editing														
	Language and Vocabulary Use	294	10	6											
	Listen/Interpret	112	106												
	Read Informational Texts	172	106	15	1										
	Read Literary Texts	21	106	20	1										
	Research	1	6	15	1										
	Research: Analyze & Integrate Info		20												
	Research: Evaluate Info & Sources	7	76	5											
	Research: Use Evidence	3	44	6	1										
	Write & Revise Explanatory Texts	271	17												
	Write & Revise Narratives	9	18	5											
	Performance Task	18	2												

Table A-7: ELA/L Number of Students Who Took Distinct IABs by Block Labels (Grade 11)

Grade	Block	Number of Distinct IABs Taken													
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
11	Brief Writes	2	4	24											
	Editing	18	32	53	5	11	31								
	Language and Vocabulary Use	11	52	73	4	8	31								
	Listen/Interpret	1	2	1	4	11	31								
	Read Informational Texts	1,766	175	28	4	10	31								
	Read Literary Texts	5	27	20	4	10	31								
	Research	9	113	7	3	10	31								
	Research: Analyze & Integrate Info		1	70											
	Research: Evaluate Info & Sources	20	5	70											
	Research: Use Evidence		1	70											
	Revision		18	34											
	Write & Revise Argumentative Texts														
	Write & Revise Narratives														
	Performance Task														

Table A-8: Mathematics Number of Students Who Took Distinct IABs by Block Labels (Grades 3–4)

Grade	**Block**	**Number of Distinct IABs Taken**												
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
3	Four Operations	148	146	176	139	165	202	134	105	73	54	36		
	Geometry	42	86	80	138	148	105	117	113	65	40	36		
	Linear and Area Measurement	41	47	135	170	181	144	121	65	64	54	36		
	Measurement and Data	9	91	67	91	180	89	107	68	68	54	36		
	Multiplication & Division	84	195	186	104	132	84	151	110	37	54	36		
	Multiply & Divide within 100	24	164	209	110	180	131	163	107	70	54	36		
	Number and Operations–Fractions	289	228	455	227	284	220	114	113	74	54	36		
	Number and Operations in Base 10	147	174	291	187	186	173	143	107	61	53	36		
	Operational and Algebraic Thinking	78	146	257	189	170	120	110	90	46	53	36		
	Properties of Multiplication & Division	19	54	61	49	73	164	94	79	66	47	36		
	Time, Volume, and Mass	24	91	56	105	98	149	96	49	42	23	36		
	Performance Task	1		1	3	3	3	1	2					
4	Build Fractions from Unit Fractions	20	186	94	71	65	42	58	72	44	46	2	8	16
	Factors and Multiples	121	51	115	129	101	35	42	49	48	59	2	7	16
	Four Operations	166	164	61	65	47	33	60	34	42	52	1	7	16
	Fraction Equivalence and Ordering	197	165	63	143	129	56	62	71	46	46	2	8	16
	Fractions and Decimal Notation	31	146	113	89	44	22	59	46	23	59	2	6	16
	Generate and Analyze Patterns	37	9	31	4	5	6	26	9	9	20	1	6	16
	Geometry	108	156	173	146	135	42	39	66	24	59	2	8	16
	Measurement and Data	41	75	166	144	90	33	33	34	18	38	1	7	16
	Multidigit Arithmetic	43	44	44	63	22	19	6	44	36	15	2	8	16
	Number and Operations–Fractions	58	417	226	73	138	37	48	39	41	44	2	7	16
	Number and Operations in Base 10	46	211	302	212	164	31	23	46	44	59	2	8	16
	Operational and Algebraic Thinking	84	369	143	163	132	31	30	52	49	59	1	8	16
	Place Value & Multidigit Whole Numbers	65	45	62	38	38	15	11	30	35	34	2	8	16
	Performance Task													

Table A-9: Mathematics Number of Students Who Took Distinct IABs by Block Labels (Grades 5–6)

Grade	**Block**	**Number of Distinct IABs Taken**												
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
5	Add & Subtract with Equivalent Fractions	128	245	197	196	168	101	51	57	40	61	145		
	Convert Measurements	29	70	81	120	117	88	44	70	36	52	143		
	Geometry	51	65	60	117	138	69	24	70	39	65	145		
	Measurement and Data	28	10	67	81	146	33	15	46	21	42	144		
	Number and Operations–Fractions	263	345	203	137	238	67	27	61	38	67	145		
	Number and Operations in Base 10	152	222	121	206	181	34	33	38	31	65	145		
	Numerical Expressions	30	74	114	227	93	91	41	64	31	64	145		
	Operations and Algebraic Thinking	40	64	61	124	146	14	7	43	40	63	144		
	Operations with Whole Numbers & Decimals	99	249	211	137	117	86	49	73	41	66	145		
	Place Value System	153	75	77	174	108	42	44	42	35	63	145		
	Volume Concepts	9	35	134	85	78	37	43	52	35	61	145		
	Performance Task						4				1	4		
6	Algebraic Expressions	226	100	58	77	91	69	30	12	20	6	25		
	Dependent & Independent Variables	6	32	23	22	72	16	17	7	17	6	25		
	Divide Fractions by Fractions	35	65	67	85	121	64	21	13	19	5	25		
	Expressions and Equations	248	58	79	129	41	10	20	9	6	6	25		
	Geometry	73	98	99	128	32	15	12	12	20	4	25		
	Multidigit Numbers, Factors, & Multiples	144	125	53	34	38	57	33	13	20	6	25		
	One-Variable Expressions and Equations	23	55	96	85	109	23	26	3	2	4	25		
	Rational Number System II	28	1	22	18	34	55	19	5	15	3	25		
	Ratios and Proportional Relationships	336	220	122	177	66	59	30	3	19	5	25		
	Statistics and Probability	13	60	20	110	61	12	22	11	20	6	25		
	The Number System	65	69	97	66	60	51	20	11	19	6	25		
	Performance Task	3	1	11	1		1	2	5	3	3			

Table A-10: Mathematics Number of Students Who Took Distinct IABs by Block Labels (Grades 7-8)

Grade	**Block**	**Number of Distinct IABs Taken**												
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
7	Algebraic Expressions and Equations	107	42	172	95	53	7	7	8	5				
	Angles, Areas, & Volume	51	36	18	51	73	20	11	11	5				
	Equivalent Expressions	13	40	66	86	62	20	10	7	5				
	Expressions and Equations	74	157	89	3	12	19	9	11	5				
	Geometric Figures	1	2	14	22	29	17	7	10	5				
	Geometry		8	13	5	6	7	10	10	5				
	Ratios and Proportional Relationships	374	134	232	97	69	14	12	10	5				
	Statistics and Probability	4	27	2	5	20	2	6	10	5				
	The Number System	27	116	192	103	71	14	12	11	5				
	Performance Task				1									
8	Analyze and Solve Linear Equations	12	144	10	10	1	3							
	Congruence and Similarity	5	151	10	9	1	3							
	Expressions and Equations I	6	4		1		3							
	Expressions and Equations II	8	11	64	1									
	Functions	6	125	64										
	Geometry	80	8											
	Proportional Relationships, Lines, & Linear Equations	495	247	72	10	1	3							
	The Number System					1	3							
	Volumes of Cylinders, Cones, & Spheres	2	14	2	9	1	3							
	Performance Task													

Table A-11: Mathematics Number of Students Who Took Distinct IABs by Block Labels (Grade 11)

Grade	**Block**	**Number of Distinct IABs Taken**												
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
11	Algebraic Functions I	795	476	40	10	6	5	2	3					
	Algebraic Functions II	358	495	17	3			2						
	Create Equations: Linear & Exponential	48	10	11	20	14	13	23	17	8	28			
	Create Equations: Quadratic	6	28	10	13	10	10	22	18	9	28			
	Equations and Reasoning	16	5	4	9	7	5	22	12	7	28			
	Geometry & Right Angle Trigonometry	3	9	10	15	14	11	22	10	8	28			
	Geometry Congruence	577	52	16	3	2								
	Geometry Measurement & Modeling													
	Interpreting Functions	1	26	30	9	8	5	9	14	8	28			
	Number and Quantity		3	4	10	15	11	12	14	9	28			
	Seeing Structure in Expressions/Polynomial Expressions		20	36	19	12	13	22	16	9	28			
	Solve Equations & Inequalities: Linear & Exponential	21	10	5	6	5	8	22	15	8	28			
	Solve Equations & Inequalities: Quadratic		9	10	15	8	15	19	16	8	28			
	Statistics and Probability	40	23	2	4	4	6	5	9	7	28			
	Performance Task													

Table A-12: ELA/L Percentage of Tests in Performance Categories by IAB Block Labels (Grades 3–5)

Grade	Block	Total Number of Tests Taken	% Below	% At/Near	% Above
3	Brief Writes	531		84	16
	Editing	825	25	44	31
	Language and Vocabulary Use	1,666	23	50	27
	Listen/Interpret	1,421	17	53	29
	Read Informational Texts	1,782	19	55	26
	Read Literary Texts	1,714	23	46	31
	Research	384	12	44	44
	Research: Analyze Information	924	15	51	35
	Research: Interpret and Integrate	708	15	44	40
	Research: Use Evidence	613	5	60	35
	Revision	302	19	63	18
	Write & Revise Informational Texts	305	14	60	26
	Write & Revise Narratives	487	24	53	23
	Write & Revise Opinion Texts	385	22	53	25
	Performance Task	1,039		77	23
4	Brief Writes	422	21	61	18
	Editing	555	28	48	24
	Language and Vocabulary Use	1,354	23	51	25
	Listen/Interpret	1,528	24	52	24
	Read Informational Texts	2,175	18	57	25
	Read Literary Texts	1,628	28	54	17
	Research	244	15	55	30
	Research: Analyze Information	343	31	52	17
	Research: Interpret and Integrate	306	18	54	28
	Research: Use Evidence	348	20	54	26
	Revision	181	21	59	20
	Write & Revise Informational Texts	213	29	50	21
	Write & Revise Narratives	188	39	53	9
	Write & Revise Opinion Texts	191	32	59	9
	Performance Task	921		81	19
5	Brief Writes	241	15	63	22
	Editing	615	16	51	33
	Language and Vocabulary Use	1,427	24	50	26
	Listen/Interpret	1,377	18	56	26
	Read Informational Texts	1,462	12	63	25
	Read Literary Texts	1,272	17	53	30
	Research	470	23	48	29
	Research: Analyze Information	777	25	55	20
	Research: Interpret and Integrate	517	19	45	35
	Research: Use Evidence	458	25	51	25
	Revision	250	28	56	16
	Write & Revise Informational Texts	73	21	63	16
	Write & Revise Narratives	377	26	52	21
	Write & Revise Opinion Texts	256	34	50	16
	Performance Task	954	18	62	19

Note: The percentage of each performance category may not add up to 100% due to rounding.

Table A-13: ELA/L Percentage of Tests in Performance Categories by IAB Block Labels (Grades 6–8)

Grade	Block	Total Number of Tests Taken	% Below	% At/Near	% Above
6	Brief Writes	147	6	73	21
	Editing	317	7	43	51
	Language and Vocabulary Use	1,152	18	49	33
	Listen/Interpret	516	22	47	31
	Read Informational Texts	871	17	62	21
	Read Literary Texts	748	18	54	28
	Research	315	27	43	30
	Research: Analyze & Integrate Info	246	2	42	56
	Research: Evaluate Info & Sources	139	12	47	41
	Research: Use Evidence	317	13	42	45
	Revision	125	25	57	18
	Write & Revise Explanatory Texts	224	32	46	21
	Write & Revise Narratives	223	14	54	31
	Performance Task	429	27	63	10
7	Brief Writes	34	6	56	38
	Editing	106	11	65	24
	Language and Vocabulary Use	546	17	50	33
	Listen/Interpret	61	13	52	34
	Read Informational Texts	476	30	51	19
	Read Literary Texts	488	34	47	20
	Research				
	Research: Analyze & Integrate Info	18	28	33	39
	Research: Evaluate Info & Sources	93	10	73	17
	Research: Use Evidence	80	10	38	53
	Revision	2*			
	Write & Revise Argumentative Texts	246	17	79	4
	Write & Revise Explanatory Texts	7*			
	Write & Revise Narratives	16	38	31	31
	Performance Task	87	40	60	
8	Brief Writes				
	Edit/Revise	341	23	60	17
	Editing				
	Language and Vocabulary Use	323	16	53	31
	Listen/Interpret	218	17	67	17
	Read Informational Texts	303	21	57	22
	Read Literary Texts	148	25	51	24
	Research	23	17	65	17
	Research: Analyze & Integrate Info	20	25	55	20
	Research: Evaluate Info & Sources	88	22	45	33
	Research: Use Evidence	54	9	69	22
	Write & Revise Explanatory Texts	288	14	75	10
	Write & Revise Narratives	32	56	38	6
	Performance Task	20	35	65	

Note: The percentage of each performance category may not add up to 100% due to rounding.
* Suppressed data due to the small sample size, n < 10.

Table A-14: ELA/L Percentage of Tests in Performance Categories by IAB Block Labels (Grade 11)

Grade	Block	Total Number of Tests Taken	% Below	% At/Near	% Above
11	Brief Writes	30	63	37	
	Editing	150	38	56	6
	Language and Vocabulary Use	179	50	43	7
	Listen/Interpret	50	14	84	2
	Read Informational Texts	2,675	15	44	40
	Read Literary Texts	97	11	62	27
	Research	175	22	53	25
	Research: Analyze & Integrate Info	71	44	38	18
	Research: Evaluate Info & Sources	106	28	51	21
	Research: Use Evidence	71	56	30	14
	Revision	52	44	52	4
	Write & Revise Argumentative Texts				
	Write & Revise Narratives				
	Performance Task				

Note: The percentage of each performance category may not add up to 100% due to rounding.

Table A-15: Mathematics Percentage of Tests in Performance Categories by IAB Block Labels (Grades 3–5)

Grade	Block	Total Number of Tests Taken	% Below	% At/Near	% Above
3	Four Operations	1,431	29	40	31
	Geometry	1,020	17	53	30
	Linear and Area Measurement	1,122	15	39	46
	Measurement and Data	960	18	38	45
	Multiplication & Division	1,212	19	45	35
	Multiply & Divide within 100	1,417	21	29	49
	Number and Operations–Fractions	2,411	14	42	44
	Number and Operations in Base 10	1,746	25	37	38
	Operational and Algebraic Thinking	1,473	22	44	35
	Properties of Multiplication & Division	803	16	44	40
	Time, Volume, and Mass	795	16	41	42
	Performance Task	14		100	
4	Build Fractions from Unit Fractions	807	15	36	49
	Factors and Multiples	835	24	52	24
	Four Operations	797	37	35	29
	Fraction Equivalence and Ordering	1,115	29	32	39
	Fractions and Decimal Notation	737	17	43	41
	Generate and Analyze Patterns	179	16	60	24
	Geometry	1,072	11	65	24
	Measurement and Data	727	13	58	29
	Multidigit Arithmetic	365	19	51	29
	Number and Operations–Fractions	1,381	36	40	24
	Number and Operations in Base 10	1,309	26	46	28
	Operational and Algebraic Thinking	1,360	31	50	19
	Place Value & Multidigit Whole Numbers	461	21	42	36
	Performance Task				
5	Add & Subtract with Equivalent Fractions	1,788	29	32	40
	Convert Measurements	895	25	39	37
	Geometry	1,048	25	50	24
	Measurement and Data	740	29	44	28
	Number and Operations–Fractions	1,843	32	45	23
	Number and Operations in Base 10	1,509	31	44	25
	Numerical Expressions	1,095	17	36	47
	Operations and Algebraic Thinking	843	18	40	42
	Operations with Whole Numbers & Decimals	1,450	30	42	28
	Place Value System	1,016	22	35	43
	Volume Concepts	757	15	36	49
	Performance Task	9*			

Note: The percentage of each performance category may not add up to 100% due to rounding.
* Suppressed data due to the small sample size, n < 10.

Table A-16: Mathematics Percentage of Tests in Performance Categories by IAB Block Labels (Grades 6–8)

Grade	Block	Total Number of Tests Taken	% Below	% At/Near	% Above
6	Algebraic Expressions	739	18	48	35
	Dependent & Independent Variables	261	21	41	38
	Divide Fractions by Fractions	550	21	38	41
	Expressions and Equations	810	31	39	31
	Geometry	816	26	35	39
	Multidigit Numbers, Factors, & Multiples	550	28	42	29
	One-Variable Expressions and Equations	485	22	39	40
	Rational Number System II	227	18	49	33
	Ratios and Proportional Relationships	1,152	35	34	31
	Statistics and Probability	361	9	50	41
	The Number System	517	23	43	33
	Performance Task	30		100	
7	Algebraic Expressions and Equations	600	34	47	19
	Angles, Areas, & Volume	329	10	45	45
	Equivalent Expressions	335	12	44	44
	Expressions and Equations	382	38	37	25
	Geometric Figures	110	9	37	54
	Geometry	64	6	61	33
	Ratios and Proportional Relationships	1,066	20	51	29
	Statistics and Probability	103	10	41	50
	The Number System	554	25	48	27
	Performance Task	1*			
8	Analyze and Solve Linear Equations	180	22	45	33
	Congruence and Similarity	189	14	38	48
	Expressions and Equations I	14	43	36	21
	Expressions and Equations II	86	65	35	
	Functions	275	35	41	24
	Geometry	88	24	44	32
	Proportional Relationships, Lines, & Linear Equations	828	17	45	38
	The Number System	4*			
	Volumes of Cylinders, Cones, & Spheres	31		55	45
	Performance Task				

Note: The percentage of each performance category may not add up to 100% due to rounding.
* Suppressed data due to the small sample size, n < 10.

Table A-17: Mathematics Percentage of Tests in Performance Categories by IAB Block Labels (Grade 11)

Grade	Block	Total Number of Tests Taken	% Below	% At/Near	% Above
11	Algebraic Functions I	2,275	73	24	3
	Algebraic Functions II	1,361	27	62	11
	Create Equations: Linear & Exponential	196	20	48	32
	Create Equations: Quadratic	158	4	70	25
	Equations and Reasoning	116	15	25	60
	Geometry & Right Angle Trigonometry	130	12	38	51
	Geometry Congruence	1,189	12	77	11
	Geometry Measurement & Modeling				
	Interpreting Functions	139	29	55	15
	Number and Quantity	108	11	39	50
	Seeing Structure in Expressions/Polynomial Expressions	175	35	33	33
	Solve Equations & Inequalities: Linear & Exponential	129	13	43	43
	Solve Equations & Inequalities: Quadratic	128	4	45	51
	Statistics and Probability	128	12	64	24
	Performance Task				

Note: The percentage of each performance category may not add up to 100% due to rounding.

Appendix B: Reliabilities and Standard Error of Measurement Curves for the Projected and the 2021–2022 Shortened Blueprints

Table B-1. Marginal Reliability and Average Conditional Standard Error of Measurement for Overall Test and by Reporting Category: ELA/L

Grade	Claim	2018–2019 Projected Hawai‘i Shortened Blueprint			2021–2022 Hawai‘i Shortened Blueprint	
		Items	Reliability	Average CSEM	Reliability	Average CSEM
3	Total Test	24	0.89	32.35	0.89	33.78
	Claim 1	8	0.62	72.45	0.62	76.45
	Claim 2	6	0.72	63.12	0.72	66.77
	Claim 3	4	0.23	118.67	0.28	122.95
	Claim 4	6	0.62	81.21	0.62	82.92
4	Total Test	24	0.88	34.96	0.88	36.04
	Claim 1	8	0.60	79.80	0.60	81.87
	Claim 2	6	0.71	69.78	0.70	72.58
	Claim 3	4	0.26	129.52	0.30	123.91
	Claim 4	6	0.62	86.80	0.59	92.15
5	Total Test	24	0.89	33.92	0.89	35.33
	Claim 1	8	0.61	79.27	0.61	83.67
	Claim 2	6	0.70	69.48	0.74	69.41
	Claim 3	4	0.33	126.58	0.33	127.84
	Claim 4	6	0.65	76.0	0.64	81.04
6	Total Test	26	0.88	34.23	0.89	34.91
	Claim 1	10	0.66	74.56	0.69	70.59
	Claim 2	6	0.69	67.29	0.72	69.48
	Claim 3	4	0.30	131.65	0.30	133.51
	Claim 4	6	0.61	85.42	0.59	90.50
7	Total Test	26	0.89	35.82	0.88	36.98
	Claim 1	10	0.67	76.09	0.63	82.97
	Claim 2	6	0.71	71.59	0.72	71.56
	Claim 3	4	0.18	136.30	0.29	125.93
	Claim 4	6	0.62	91.16	0.61	93.81
8	Total Test	26	0.89	35.47	0.88	36.91
	Claim 1	10	0.70	73.39	0.66	75.71
	Claim 2	6	0.70	72.88	0.70	73.37
	Claim 3	4	0.24	133.66	0.30	131.37
	Claim 4	6	0.63	87.54	0.59	94.19
11	Total Test	26	0.88	39.68	0.88	40.69
	Claim 1	10	0.67	81.60	0.65	85.07
	Claim 2	6	0.70	75.86	0.71	77.51
	Claim 3	4	0.18	151.52	0.32	145.47
	Claim 4	6	0.61	97.40	0.59	102.58

Table B-2. Marginal Reliability and Average Conditional Standard Error of Measurement for Overall Test and by Reporting Category: Mathematics

Grade	Claim	2016–2017 Projected Hawai‘i Shortened Blueprint			2021–2022 Hawai‘i Shortened Blueprint	
		Items	Reliability	Average CSEM	Reliability	Average CSEM
3	Total Test	22	0.90	27.05	0.91	28.25
	Claim 1	12	0.84	39.09	0.84	41.61
	Claims 2 & 4	5	0.50	71.05	0.60	68.69
	Claim 3	5	0.52	74.80	0.58	72.17
4	Total Test	22	0.90	26.35	0.91	27.65
	Claim 1	12	0.84	38.49	0.84	41.05
	Claims 2 & 4	5	0.55	65.65	0.55	69.88
	Claim 3	5	0.55	71.82	0.62	67.85
5	Total Test	22	0.90	30.44	0.90	31.80
	Claim 1	12	0.82	43.43	0.83	45.83
	Claims 2 & 4	5	0.48	71.73	0.46	83.93
	Claim 3	5	0.51	87.68	0.56	86.24
6	Total Test	22	0.90	34.58	0.88	39.32
	Claim 1	12	0.83	49.96	0.81	55.77
	Claims 2 & 4	5	0.53	81.52	0.44	97.47
	Claim 3	5	0.53	88.15	0.46	103.31
7	Total Test	22	0.88	39.89	0.87	42.47
	Claim 1	12	0.81	54.47	0.78	61.50
	Claims 2 & 4	5	0.37	104.48	0.39	104.94
	Claim 3	5	0.39	106.75	0.46	106.07
8	Total Test	22	0.89	41.95	0.86	46.80
	Claim 1	12	0.81	60.75	0.77	66.93
	Claims 2 & 4	5	0.51	102.55	0.44	99.26
	Claim 3	5	0.50	115.15	0.39	121.12
11	Total Test	24	0.88	44.46	0.87	43.97
	Claim 1	14	0.82	56.32	0.80	57.37
	Claims 2 & 4	5	0.48	126.07	0.53	121.09
	Claim 3	5	0.40	132.76	0.48	125.60

Figure B-1. Conditional Standard Error of Measurements Across Estimated Score Range for the Hawaiʻi Projected Shortened and the 2021–2022 Shortened Blueprints: ELA/L

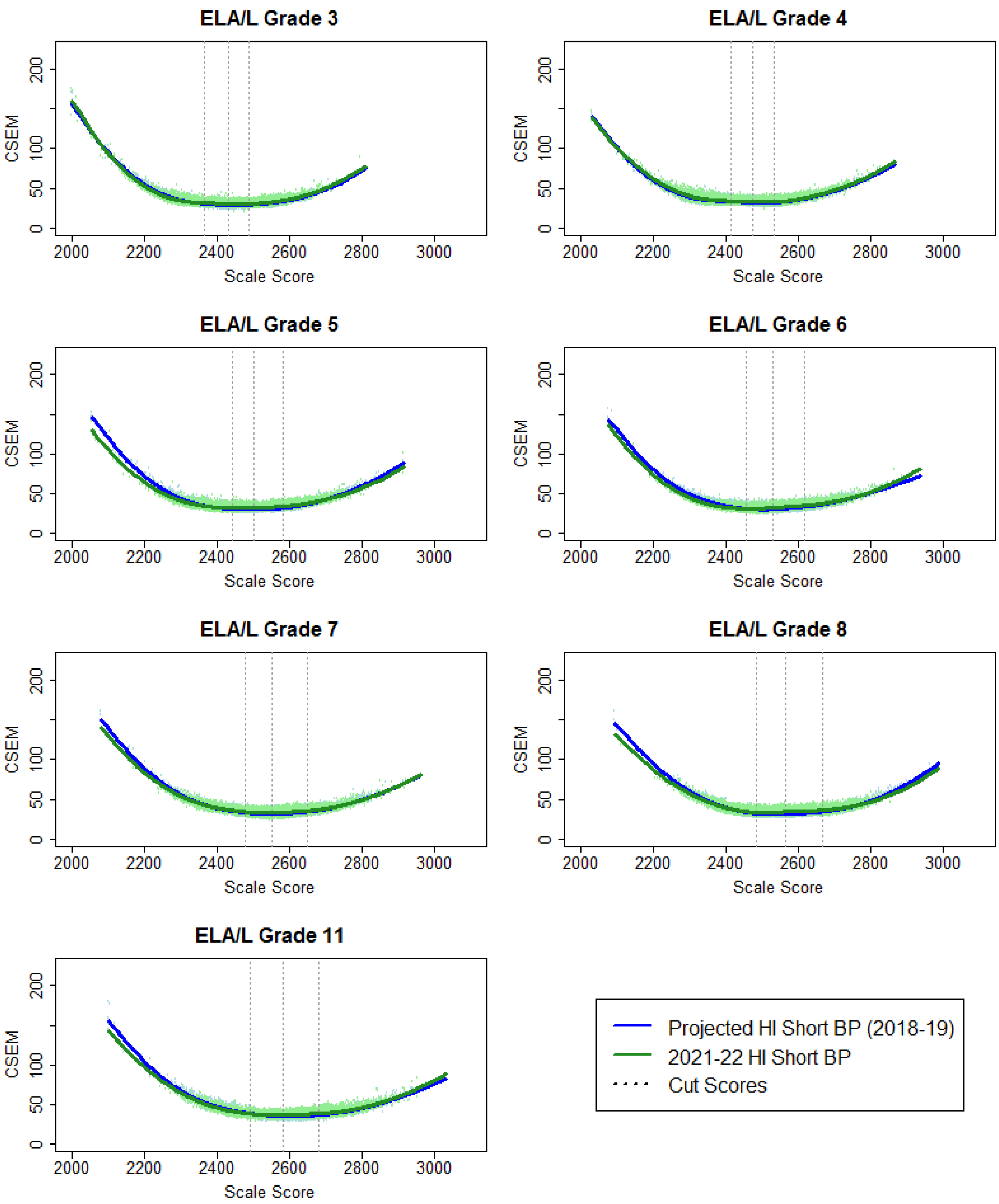
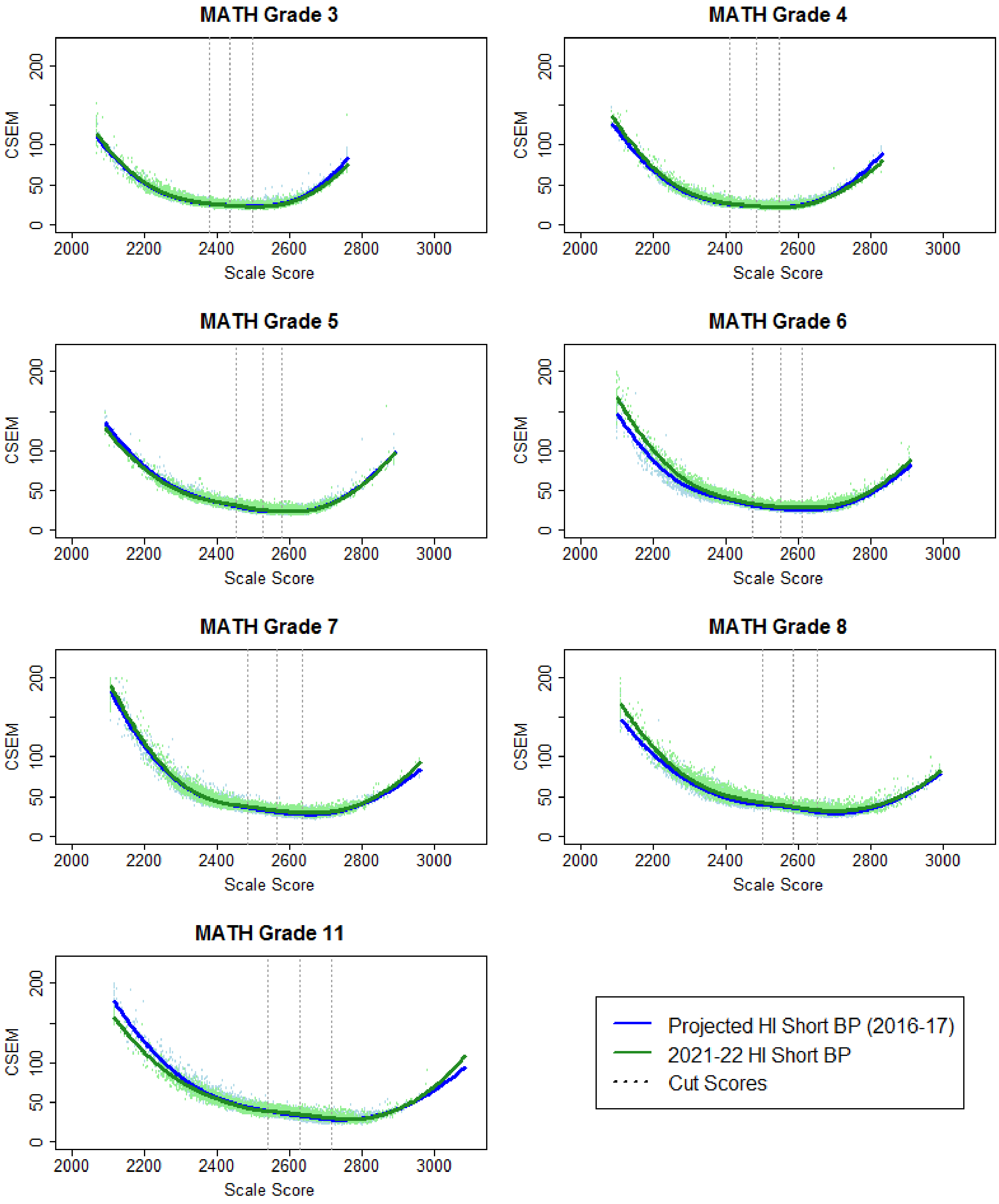


Figure B-2. Conditional Standard Error of Measurements Across Estimated Score Range for the Hawai'i Projected Shortened and the 2021–2022 Shortened Blueprints: Mathematics



Appendix C: Student Performance Across Four Years for All Students and by Subgroup

Table C-1. Student Performance Across Four Years: ELA/L (Grades 3 and 4)

Group	2017–2018				2018–2019				2020–2021				2021–2022			
	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD
Grade 3																
All Students	11,522	52	2432.9	92.3	14,398	52	2431.4	93.2	12,328	43	2412.0	98.7	12,991	49	2425.2	101.4
Female	5,552	57	2443.0	90.2	7,013	56	2440.9	90.2	5,970	47	2422.5	97.0	6,208	54	2436.0	100.5
Male	5,970	48	2423.5	93.2	7,385	48	2422.4	95.1	6,358	39	2402.1	99.3	6,783	45	2415.3	101.2
African American	222	47	2423.0	83.0	209	51	2433.8	82.4	155	35	2398.2	92.2	157	55	2434.3	88.0
AmerIndian/Alaskan	20	50	2421.7	92.3	18	61	2443.9	72.1	14	29	2392.1	104.7	15	47	2413.6	71.3
Asian/Pacific Islander	2,591	63	2455.5	89.8	3,421	63	2455.2	90.5	2,872	53	2436.0	97.1	2,969	62	2457.4	96.0
Hispanic	2,183	47	2422.2	88.1	2,764	47	2419.9	91.0	2,375	37	2398.2	94.5	2,576	43	2409.9	98.9
Hawai'i Pacific Islander	2,784	34	2392.5	85.4	3,461	33	2389.5	87.1	2,847	24	2366.6	90.0	2,983	28	2374.5	92.0
White	1,618	67	2464.8	89.0	1,703	66	2462.4	87.3	1,356	59	2447.5	95.1	1,428	63	2455.1	94.9
Multi-Racial	2,104	58	2446.2	91.1	2,822	60	2446.4	90.1	2,709	51	2429.5	95.3	2,863	56	2443.0	99.2
ELL	1,434	24	2372.5	80.5	1,817	25	2371.8	78.2	1,716	23	2362.1	91.4	1,790	28	2373.8	92.8
Disadvantaged	5,682	40	2404.6	87.1	6,785	38	2400.8	89.2	5,848	29	2380.1	92.6	5,776	34	2390.0	95.9
Migrant	126	18	2367.7	72.9	185	28	2369.0	87.7	149	19	2360.4	86.8	145	23	2363.7	93.3
Disability	1,046	9	2332.7	72.4	1,300	9	2326.3	76.0	1,156	8	2315.7	83.0	1,205	9	2319.0	82.7
Grade 4																
All Students	14,827	50	2467.5	98.6	11,358	51	2469.2	99.9	12,476	46	2458.7	100.5	12,819	51	2470.9	103.2
Female	7,200	55	2478.3	96.0	5,459	56	2480.7	97.7	6,057	51	2470.2	98.2	6,173	56	2482.3	100.6
Male	7,627	47	2457.2	99.9	5,899	47	2458.6	100.8	6,419	43	2447.8	101.4	6,646	48	2460.4	104.4
African American	252	53	2468.7	92.1	196	44	2456.3	87.8	153	46	2455.0	86.1	158	39	2452.2	94.2
AmerIndian/Alaskan	27	59	2476.3	77.7	23	65	2506.3	83.9	17	53	2481.7	73.6	15	47	2459.9	93.5
Asian/Pacific Islander	3,626	61	2492.7	93.9	2,630	63	2494.1	98.6	3,125	58	2486.3	98.5	2,964	64	2499.9	98.6
Hispanic	2,656	43	2453.2	95.4	2,149	46	2457.3	94.5	2,358	40	2444.7	97.9	2,493	45	2455.5	100.0
Hawai'i Pacific Islander	3,640	31	2421.2	91.3	2,735	32	2425.4	93.4	2,891	26	2411.6	91.7	2,987	32	2424.3	95.4
White	1,838	68	2505.5	93.6	1,561	66	2504.9	93.0	1,448	62	2493.8	89.5	1,441	65	2503.3	97.8
Multi-Racial	2,788	58	2483.5	95.3	2,064	56	2481.6	98.5	2,484	52	2471.7	99.3	2,761	58	2488.4	101.3
ELL	1,423	12	2383.5	72.5	1,277	16	2386.8	79.9	1,632	21	2400.2	88.1	1,655	27	2413.6	92.9
Disadvantaged	7,266	38	2438.1	94.8	5,410	38	2438.5	95.6	5,899	32	2425.6	94.7	5,646	38	2437.3	97.7
Migrant	164	23	2404.9	91.3	153	22	2403.6	86.0	126	15	2386.1	91.2	165	30	2416.8	96.1
Disability	1,347	8	2353.1	77.3	1,172	8	2361.6	77.1	1,256	9	2354.1	86.3	1,306	9	2357.1	86.0

2019–2020 is not included because the summative assessments were canceled due to the COVID-19 pandemic.

Table C-2. Student Performance Across Four Years: ELA/L (Grades 5 and 6)

Group	2017–2018				2018–2019				2020–2021				2021–2022			
	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD
Grade 5																
All Students	14,803	56	2511.3	98.6	14,754	57	2512.1	99.7	12,712	51	2500.0	103.0	13,058	55	2509.9	107.8
Female	7,147	62	2525.9	95.0	7,169	62	2525.1	95.6	6,133	55	2512.3	100.8	6,316	60	2524.4	104.3
Male	7,656	50	2497.7	99.8	7,585	52	2499.8	101.9	6,579	47	2488.5	103.7	6,742	50	2496.3	109.3
African American	245	57	2509.3	94.3	240	57	2515.8	87.5	181	45	2491.4	91.4	166	47	2499.0	96.0
AmerIndian/Alaskan	24	58	2514.1	85.3	21	43	2508.8	112.0	19	47	2497.6	74.1	16	56	2535.9	73.8
Asian/Pacific Islander	3,849	66	2535.2	93.3	3,703	67	2537.4	95.8	3,279	63	2529.1	100.1	3,221	67	2542.7	103.2
Hispanic	2,628	51	2499.6	93.9	2,601	52	2499.6	96.4	2,331	44	2483.6	97.8	2,495	50	2497.5	105.2
Hawaiʻi Pacific Islander	3,655	35	2464.8	93.4	3,611	35	2464.5	94.5	3,033	31	2451.7	99.1	3,081	34	2457.1	101.9
White	1,726	74	2552.7	92.5	1,783	74	2550.3	92.0	1,413	67	2535.4	94.0	1,507	71	2545.0	94.4
Multi-Racial	2,676	61	2525.6	95.2	2,795	64	2527.2	94.3	2,456	57	2516.4	96.8	2,572	61	2524.0	104.5
ELL	851	6	2397.1	68.2	1,315	12	2413.0	74.1	1,441	17	2420.3	82.8	1,460	23	2428.9	91.6
Disadvantaged	7,094	43	2480.6	95.2	6,891	43	2480.2	97.8	5,932	36	2465.6	99.1	5,681	40	2473.4	103.6
Migrant	158	26	2438.9	87.7	195	28	2450.4	92.8	175	25	2434.6	100.4	139	27	2440.4	97.7
Disability	1,388	10	2396.5	80.0	1,408	9	2389.6	77.2	1,282	10	2386.5	85.4	1,338	12	2392.2	90.7
Grade 6																
All Students	13,896	52	2531.8	98.1	14,121	52	2529.8	98.0	9,506	47	2519.5	99.7	12,841	50	2525.0	104.8
Female	6,620	59	2548.5	93.3	6,832	59	2545.1	94.6	4,527	53	2533.0	97.7	6,234	55	2538.4	101.6
Male	7,276	47	2516.6	100.0	7,289	46	2515.5	99.1	4,979	41	2507.2	100.1	6,607	45	2512.5	106.2
African American	234	56	2539.5	88.7	211	56	2533.0	87.0	125	50	2530.9	87.8	173	51	2530.2	99.9
AmerIndian/Alaskan	38	39	2503.6	69.0	20	65	2551.3	81.0	12	67	2559.5	82.6	16	38	2501.6	111.2
Asian/Pacific Islander	3,878	62	2554.8	93.5	3,704	63	2553.4	93.2	2,397	57	2541.3	98.0	3,296	61	2553.3	102.3
Hispanic	2,249	48	2522.1	95.2	2,552	48	2520.2	94.7	1,787	41	2508.1	95.3	2,395	43	2510.2	100.5
Hawaiʻi Pacific Islander	3,705	32	2485.5	92.9	3,541	32	2482.4	93.5	2,294	27	2472.4	93.9	3,143	29	2474.9	98.8
White	1,599	71	2572.7	90.0	1,569	71	2571.2	91.7	1,178	64	2562.0	91.1	1,407	67	2566.0	92.4
Multi-Racial	2,193	60	2549.0	93.3	2,524	58	2545.1	91.7	1,713	52	2533.4	94.7	2,411	58	2542.4	98.9
ELL	737	8	2414.4	78.4	957	8	2417.6	78.3	1,031	9	2426.0	79.6	1,411	13	2435.6	81.6
Disadvantaged	6,614	39	2500.3	95.6	6,657	39	2497.7	95.0	4,485	33	2487.6	94.9	5,748	35	2490.6	99.1
Migrant	146	21	2469.5	78.7	215	27	2472.2	90.8	142	16	2456.3	82.6	191	23	2458.0	87.3
Disability	1,302	9	2416.6	81.3	1,424	8	2417.5	82.1	1,034	7	2412.4	76.5	1,336	8	2408.4	83.1

2019–2020 is not included because the summative assessments were canceled due to the COVID-19 pandemic.

Table C-3. Student Performance Across Four Years: ELA/L (Grades 7 and 8)

Group	**2017–2018**				**2018–2019**				**2020–2021**				**2021–2022**			
	N	**% Prof**	**Scale Score**	**SD**	**N**	**% Prof**	**Scale Score**	**SD**	**N**	**% Prof**	**Scale Score**	**SD**	**N**	**% Prof**	**Scale Score**	**SD**
Grade 7																
All Students	13,396	52	2548.4	102.9	13,536	53	2550.4	103.3	11,107	52	2548.8	102.6	9,922	52	2548.9	108.3
Female	6,388	60	2568.1	98.4	6,436	61	2569.0	98.6	5,417	58	2563.9	96.8	4,745	57	2563.1	104.6
Male	7,008	45	2530.4	103.6	7,100	46	2533.6	104.6	5,690	46	2534.4	106.0	5,177	47	2535.9	109.9
African American	256	49	2546.1	99.1	212	59	2562.6	96.3	169	59	2560.8	92.4	146	56	2558.7	95.2
AmerIndian/Alaskan	38	42	2538.2	95.5	26	31	2489.1	107.6	14	50	2558.8	64.8	13	77	2604.0	99.3
Asian/Pacific Islander	4,518	62	2572.3	97.3	3,902	63	2576.0	96.9	3,171	63	2576.6	95.8	2,498	65	2580.5	103.8
Hispanic	1,291	50	2542.6	101.6	2,168	49	2540.1	99.0	1,899	44	2534.4	96.9	1,909	45	2534.4	106.5
Hawaiʻi Pacific Islander	4,559	34	2505.3	95.6	3,630	32	2499.9	96.6	2,516	31	2496.8	98.3	2,458	31	2497.3	101.0
White	1,579	70	2591.8	98.6	1,478	72	2596.9	97.5	1,264	67	2582.2	95.3	1,183	70	2593.4	98.4
Multi-Racial	1,155	63	2572.6	97.8	2,120	59	2567.4	98.3	2,074	58	2561.1	101.7	1,715	57	2561.2	101.9
ELL	809	8	2433.2	78.4	831	10	2437.5	81.7	1,106	15	2455.4	89.2	1,107	16	2459.9	91.9
Disadvantaged	6,428	38	2516.1	98.3	6,291	39	2517.4	100.1	5,171	39	2516.0	101.3	4,454	38	2515.0	105.2
Migrant	159	28	2489.7	95.9	175	31	2493.2	94.6	187	25	2480.6	99.8	155	25	2485.6	96.6
Disability	1,323	8	2430.3	83.2	1,280	8	2433.0	82.4	1,083	8	2426.2	87.6	1,125	10	2433.6	89.5
Grade 8																
All Students	12,748	54	2571.9	101.2	12,872	51	2565.3	104.1	10,677	51	2564.8	102.7	12,456	50	2561.7	107.2
Female	6,145	62	2591.5	96.0	6,192	59	2585.4	98.6	5,067	58	2581.6	97.6	6,076	56	2577.2	100.7
Male	6,603	47	2553.5	102.6	6,680	45	2546.6	105.6	5,610	46	2549.6	104.9	6,380	45	2547.0	111.1
African American	207	60	2584.8	93.5	233	53	2569.4	100.2	154	56	2574.0	100.9	182	55	2571.5	91.3
AmerIndian/Alaskan	39	56	2567.7	96.0	36	47	2561.9	93.3	13	54	2578.1	98.4	17	53	2565.7	94.0
Asian/Pacific Islander	4,449	64	2594.5	97.1	4,461	62	2590.2	99.6	3,201	62	2589.0	99.0	3,475	65	2595.2	101.7
Hispanic	1,155	51	2563.8	99.4	1,254	49	2560.0	100.7	1,848	47	2556.4	97.4	2,202	43	2545.5	105.0
Hawaiʻi Pacific Islander	4,151	36	2527.5	94.4	4,235	33	2521.1	97.1	2,435	29	2512.6	97.3	2,955	29	2509.3	99.9
White	1,622	72	2611.5	90.5	1,525	67	2602.3	99.9	1,098	67	2601.4	93.3	1,383	64	2593.6	99.5
Multi-Racial	1,125	64	2594.8	99.7	1,128	60	2587.3	99.1	1,928	56	2576.9	99.5	2,242	54	2574.3	102.9
ELL	708	8	2455.3	74.2	783	7	2456.9	76.0	822	10	2458.7	85.8	1,198	16	2476.3	87.9
Disadvantaged	5,742	41	2539.9	98.7	5,699	37	2530.3	100.2	4,752	37	2531.2	98.3	5,439	37	2528.3	104.9
Migrant	136	36	2525.2	90.3	201	24	2489.0	97.9	160	21	2497.2	96.1	200	21	2482.6	96.1
Disability	1,233	9	2451.4	80.0	1,254	6	2444.0	78.9	1,048	9	2449.9	84.1	1,217	8	2439.0	90.2

2019–2020 is not included because the summative assessments were canceled due to the COVID-19 pandemic.

Table C-4. Student Performance Across Four Years: ELA/L (Grade 11)

Group	**2017–2018**				**2018–2019**				**2020–2021**				**2021–2022**			
	N	**% Prof**	**Scale Score**	**SD**	**N**	**% Prof**	**Scale Score**	**SD**	**N**	**% Prof**	**Scale Score**	**SD**	**N**	**% Prof**	**Scale Score**	**SD**
Grade 11																
All Students	10,272	60	2604.0	110.5	10,730	59	2601.2	112.6	7,804	65	2615.4	106.1	10,033	60	2604.4	115.3
Female	5,270	67	2621.9	101.3	5,261	66	2618.8	104.2	3,820	71	2632.1	99.2	4,924	66	2622.0	109.1
Male	5,002	53	2585.1	116.5	5,469	52	2584.2	117.6	3,984	59	2599.4	110.0	5,109	54	2587.5	118.5
African American	199	55	2597.6	107.5	229	57	2594.0	105.8	107	68	2612.8	93.4	165	53	2588.5	118.7
AmerIndian/Alaskan	27	59	2602.3	126.8	32	69	2601.7	100.3	16	81	2650.7	76.7	26	69	2620.8	88.2
Asian/Pacific Islander	4,164	68	2625.9	105.3	4,198	67	2622.5	105.1	3,384	71	2633.2	101.7	4,024	69	2630.7	106.8
Hispanic	815	59	2597.7	107.6	867	57	2593.5	107.4	659	60	2599.3	104.1	986	54	2585.8	113.0
Hawai‘i Pacific Islander	3,025	44	2559.3	105.8	3,164	40	2551.2	108.6	1,948	47	2568.7	102.9	2,716	42	2556.2	111.6
White	1,181	70	2631.7	109.8	1,288	74	2644.3	107.5	932	77	2646.5	100.3	1,157	71	2632.0	114.0
Multi-Racial	861	68	2624.5	106.8	952	68	2623.5	111.6	758	71	2631.2	103.6	959	65	2618.9	116.3
ELL	368	5	2457.4	71.8	604	8	2480.2	77.5	349	13	2488.4	84.1	549	17	2488.3	88.5
Disadvantaged	3,822	50	2576.5	108.8	3,946	47	2569.8	110.3	2,769	54	2588.9	106.2	3,499	47	2571.2	114.7
Migrant	97	37	2546.7	108.0	126	40	2550.2	104.5	103	43	2565.6	110.5	126	38	2547.2	114.9
Disability	790	11	2470.1	87.4	865	13	2468.8	93.6	548	17	2488.3	96.1	771	11	2465.3	94.4

2019–2020 is not included because the summative assessments were canceled due to the COVID-19 pandemic.

Table C-5. Student Performance Across Four Years: Mathematics (Grades 3 and 4)

Group	2017–2018				2018–2019				2020–2021				2021–2022			
	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD
Grade 3																
All Students	11,586	54	2443.6	84.9	14,454	56	2445.2	85.8	12,407	41	2411.5	92.6	13,041	51	2435.1	94.9
Female	5,580	54	2443.4	80.7	7,043	55	2443.0	81.3	5,995	39	2410.2	90.0	6,231	50	2433.2	91.6
Male	6,006	54	2443.7	88.6	7,411	57	2447.3	89.8	6,412	42	2412.6	95.0	6,810	52	2436.9	97.8
African American	223	40	2417.7	74.5	210	47	2429.3	73.6	155	28	2387.2	75.7	157	49	2435.2	80.5
AmerIndian/Alaskan	20	65	2432.9	83.2	18	56	2443.6	86.6	14	29	2402.0	124.2	16	44	2413.1	76.9
Asian/Pacific Islander	2,631	68	2472.9	83.6	3,447	69	2473.5	82.8	2,884	53	2439.8	89.9	2,990	66	2471.6	88.6
Hispanic	2,191	48	2430.8	78.7	2,770	50	2432.3	82.5	2,397	33	2395.5	87.0	2,581	44	2419.5	91.7
Hawai‘i Pacific Islander	2,794	36	2407.7	80.6	3,479	37	2406.2	79.8	2,871	20	2363.7	88.2	2,998	29	2385.1	88.2
White	1,620	65	2463.3	79.1	1,704	69	2470.8	79.7	1,356	57	2448.0	81.7	1,432	63	2461.5	85.6
Multi-Racial	2,107	60	2455.6	83.1	2,826	61	2457.3	83.3	2,730	48	2429.0	86.4	2,867	59	2450.4	90.8
ELL	1,485	32	2397.9	79.2	1,871	32	2397.3	80.6	1,736	23	2367.5	93.1	1,812	33	2393.6	95.0
Disadvantaged	5,719	42	2418.6	80.4	6,822	42	2418.3	82.4	5,905	27	2380.5	89.1	5,797	36	2402.1	91.3
Migrant	126	21	2380.2	77.3	187	29	2384.3	86.1	149	16	2358.4	82.8	146	23	2364.6	85.5
Disability	1,052	13	2350.4	79.2	1,297	14	2348.6	81.3	1,168	10	2322.4	90.2	1,212	15	2338.4	89.4
Grade 4																
All Students	14,881	47	2475.5	84.5	11,423	48	2478.3	85.5	12,521	36	2452.7	89.8	12,872	46	2472.4	92.6
Female	7,211	46	2475.2	79.0	5,489	47	2477.1	80.0	6,076	34	2451.3	85.0	6,190	44	2469.2	88.3
Male	7,670	48	2475.7	89.4	5,934	49	2479.5	90.2	6,445	38	2454.1	94.1	6,682	48	2475.3	96.3
African American	252	43	2465.1	72.5	197	40	2459.5	73.5	155	35	2449.9	71.4	159	33	2454.4	74.5
AmerIndian/Alaskan	27	44	2477.1	77.4	23	57	2512.3	80.5	17	18	2449.3	39.6	14	29	2445.7	102.7
Asian/Pacific Islander	3,668	61	2503.4	80.6	2,666	62	2506.8	85.3	3,136	50	2483.1	88.5	2,979	60	2504.4	89.1
Hispanic	2,656	41	2463.2	80.6	2,151	42	2465.4	79.2	2,375	27	2435.3	83.4	2,498	38	2455.4	87.8
Hawai‘i Pacific Islander	3,653	28	2436.6	79.5	2,752	30	2440.8	79.2	2,902	16	2408.8	81.1	3,008	25	2426.7	85.9
White	1,837	60	2500.0	80.6	1,566	61	2500.3	79.4	1,455	52	2486.4	80.9	1,448	60	2502.2	87.1
Multi-Racial	2,788	53	2486.2	81.3	2,068	52	2489.8	84.4	2,481	41	2462.9	88.5	2,766	54	2488.4	87.3
ELL	1,474	15	2410.9	73.4	1,334	19	2420.1	74.6	1,650	16	2404.1	85.2	1,681	24	2424.6	87.1
Disadvantaged	7,298	35	2450.8	81.4	5,449	35	2453.1	81.3	5,929	22	2421.7	84.0	5,676	31	2441.1	87.7
Migrant	162	17	2415.1	77.9	155	19	2420.6	73.5	127	9	2393.0	73.5	165	21	2421.2	82.7
Disability	1,346	8	2376.9	78.9	1,175	10	2386.1	76.9	1,264	8	2366.8	83.1	1,321	10	2375.1	84.1

2019–2020 is not included because the summative assessments were canceled due to the COVID-19 pandemic.

Table C-6. Student Performance Across Four Years: Mathematics (Grades 5 and 6)

Group	2017–2018				2018–2019				2020–2021				2021–2022			
	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD
Grade 5																
All Students	14,848	43	2507.2	92.7	14,814	44	2509.0	94.9	12,770	32	2478.8	97.1	13,096	42	2501.0	100.6
Female	7,170	43	2509.1	88.4	7,190	43	2508.8	90.7	6,163	29	2476.0	92.2	6,336	40	2500.1	96.6
Male	7,678	43	2505.4	96.4	7,624	45	2509.1	98.8	6,607	34	2481.4	101.3	6,760	43	2501.9	104.2
African American	246	35	2491.7	85.9	240	39	2499.2	81.6	182	18	2458.6	79.0	165	25	2482.3	84.3
AmerIndian/Alaskan	24	38	2505.4	72.8	21	43	2514.4	80.5	19	32	2486.3	91.5	16	50	2505.8	59.6
Asian/Pacific Islander	3,879	56	2538.1	89.5	3,744	59	2542.2	90.8	3,293	46	2512.1	95.4	3,235	59	2541.8	96.5
Hispanic	2,630	35	2492.5	87.3	2,604	35	2492.4	88.5	2,345	24	2459.8	92.2	2,497	32	2482.3	93.7
Hawai'i Pacific Islander	3,663	24	2465.3	87.6	3,626	25	2463.9	89.3	3,054	14	2431.5	87.7	3,090	21	2450.7	93.1
White	1,734	55	2534.1	85.6	1,784	56	2535.2	89.2	1,409	44	2512.4	89.6	1,515	54	2529.3	90.4
Multi-Racial	2,672	48	2518.2	89.4	2,795	50	2522.6	91.7	2,468	37	2493.0	92.0	2,578	47	2513.0	97.3
ELL	896	9	2420.0	75.5	1,374	11	2429.9	77.9	1,455	9	2414.9	80.2	1,464	16	2434.2	90.9
Disadvantaged	7,110	30	2479.0	89.2	6,928	31	2479.9	91.9	5,979	19	2446.1	90.3	5,698	27	2465.7	95.8
Migrant	158	20	2450.8	82.9	194	21	2457.4	92.4	181	12	2412.9	97.0	137	16	2431.3	97.9
Disability	1,389	8	2407.3	79.5	1,414	7	2399.4	77.8	1,290	5	2386.1	84.7	1,336	7	2400.5	87.9
Grade 6																
All Students	13,950	41	2521.7	105.4	14,176	41	2519.0	106.6	9,572	29	2491.5	107.4	12,888	35	2505.8	114.4
Female	6,650	44	2530.5	98.7	6,854	43	2525.8	100.9	4,547	29	2491.5	105.0	6,255	34	2505.3	110.8
Male	7,300	38	2513.6	110.7	7,322	38	2512.6	111.3	5,025	29	2491.6	109.5	6,633	36	2506.2	117.6
African American	235	34	2521.3	84.4	213	35	2508.2	104.0	125	23	2487.4	86.6	174	31	2503.5	102.4
AmerIndian/Alaskan	38	29	2498.1	81.2	20	55	2533.9	81.1	13	31	2532.3	97.4	16	31	2457.9	170.9
Asian/Pacific Islander	3,909	54	2553.1	99.2	3,726	54	2553.1	100.9	2,405	40	2523.2	106.1	3,302	48	2543.3	108.8
Hispanic	2,253	34	2507.4	99.6	2,559	35	2507.1	101.9	1,812	21	2471.6	101.3	2,401	27	2484.3	111.1
Hawai'i Pacific Islander	3,721	23	2473.1	104.7	3,558	21	2467.6	101.9	2,310	13	2442.9	98.9	3,163	17	2450.6	107.6
White	1,602	56	2556.5	94.2	1,573	53	2553.3	97.4	1,184	43	2530.6	99.3	1,417	50	2550.2	102.5
Multi-Racial	2,192	47	2538.0	99.7	2,527	46	2532.6	101.0	1,723	33	2506.7	104.2	2,415	40	2522.5	106.2
ELL	785	8	2411.0	102.7	988	8	2419.2	97.2	1,045	6	2408.0	94.6	1,423	8	2419.3	100.3
Disadvantaged	6,633	29	2489.3	105.1	6,701	27	2484.7	103.4	4,528	17	2457.5	101.7	5,781	22	2468.5	109.4
Migrant	151	17	2455.1	98.6	217	17	2462.5	96.7	143	7	2427.8	85.4	192	11	2427.4	106.1
Disability	1,307	5	2396.2	97.4	1,427	5	2397.5	96.8	1,054	3	2381.5	93.4	1,340	5	2386.3	102.5

2019–2020 is not included because the summative assessments were canceled due to the COVID-19 pandemic.

Table C-7. Student Performance Across Four Years: Mathematics (Grades 7 and 8)

Group	2017–2018				2018–2019				2020–2021				2021–2022			
	N	**% Prof**	**Scale Score**	**SD**	**N**	**% Prof**	**Scale Score**	**SD**	**N**	**% Prof**	**Scale Score**	**SD**	**N**	**% Prof**	**Scale Score**	**SD**
Grade 7																
All Students	13,441	37	2524.1	111.7	13,606	38	2526.3	113.5	11,183	29	2505.9	110.9	9,959	33	2513.3	117.5
Female	6,412	39	2531.8	108.4	6,463	40	2534.0	109.0	5,459	29	2506.3	106.7	4,761	32	2511.0	115.1
Male	7,029	35	2517.0	114.3	7,143	36	2519.4	116.9	5,724	30	2505.5	114.8	5,198	34	2515.3	119.6
African American	256	32	2516.0	108.3	210	33	2513.0	103.6	169	27	2505.5	96.8	143	26	2503.7	101.9
AmerIndian/Alaskan	39	33	2525.3	107.1	26	23	2477.9	92.9	15	40	2538.6	73.4	14	57	2555.8	125.6
Asian/Pacific Islander	4,547	49	2557.0	108.5	3,930	52	2563.2	109.4	3,174	43	2545.1	107.9	2,498	49	2556.9	116.9
Hispanic	1,280	33	2513.2	108.1	2,180	31	2509.7	105.9	1,917	22	2487.2	102.4	1,921	25	2492.2	107.8
Hawaiʻi Pacific Islander	4,586	20	2476.6	102.0	3,656	19	2470.9	105.0	2,546	12	2448.0	102.2	2,484	15	2456.5	106.5
White	1,578	51	2562.8	104.5	1,480	53	2565.5	104.7	1,268	40	2536.6	101.9	1,181	47	2556.0	106.5
Multi-Racial	1,155	45	2543.8	104.6	2,124	44	2545.2	107.0	2,094	32	2515.3	106.2	1,718	38	2526.4	112.6
ELL	846	8	2414.9	95.2	872	9	2419.1	100.6	1,110	7	2422.6	100.7	1,126	10	2424.4	109.0
Disadvantaged	6,451	24	2489.4	106.5	6,346	25	2490.5	109.5	5,234	19	2471.7	107.3	4,482	21	2477.4	112.1
Migrant	160	16	2453.4	104.0	175	13	2450.4	96.8	187	10	2436.3	97.7	158	11	2449.8	97.5
Disability	1,321	4	2401.4	90.2	1,301	4	2401.5	93.4	1,095	2	2387.8	97.9	1,128	4	2396.7	97.9
Grade 8																
All Students	12,794	38	2546.4	121.3	12,940	38	2543.0	123.4	10,742	25	2511.7	115.7	12,511	31	2524.3	123.7
Female	6,176	42	2557.6	116.9	6,234	41	2554.7	118.5	5,087	26	2516.0	111.6	6,101	31	2526.7	119.1
Male	6,618	35	2536.0	124.4	6,706	35	2532.1	126.9	5,655	24	2507.8	119.2	6,410	31	2522.0	128.0
African American	203	38	2553.2	112.8	237	30	2524.2	109.6	154	20	2505.3	108.3	182	36	2536.4	117.4
AmerIndian/Alaskan	38	39	2531.0	128.4	36	39	2544.1	122.6	13	38	2540.5	105.3	17	18	2520.8	106.5
Asian/Pacific Islander	4,476	49	2582.4	119.5	4,489	51	2581.7	120.6	3,217	37	2547.7	113.9	3,479	45	2571.0	122.4
Hispanic	1,156	33	2528.4	113.9	1,256	33	2528.5	116.7	1,856	18	2495.8	106.0	2,216	21	2498.6	113.7
Hawaiʻi Pacific Islander	4,176	21	2493.0	108.5	4,256	20	2489.5	112.1	2,467	8	2452.1	105.0	2,993	13	2463.3	109.9
White	1,621	52	2581.1	111.4	1,536	50	2579.7	115.8	1,102	35	2548.3	110.9	1,389	43	2558.9	116.4
Multi-Racial	1,124	47	2569.6	120.8	1,130	44	2560.9	117.8	1,933	28	2522.5	110.0	2,235	34	2536.4	117.7
ELL	745	10	2439.3	108.2	822	12	2445.3	106.8	832	6	2419.8	105.2	1,211	8	2433.4	110.4
Disadvantaged	5,773	26	2509.8	116.9	5,736	25	2504.7	118.2	4,810	14	2474.1	108.9	5,471	19	2486.5	118.0
Migrant	135	21	2477.7	107.3	202	15	2463.3	114.2	161	5	2443.6	98.1	199	11	2458.2	108.4
Disability	1,242	4	2409.6	91.2	1,266	4	2404.4	92.7	1,061	3	2404.2	97.7	1,231	3	2400.5	102.0

2019–2020 is not included because the summative assessments were canceled due to the COVID-19 pandemic.

Table C-8. Student Performance Across Four Years: Mathematics (Grade 11)

Group	**2017–2018**				**2018–2019**				**2020–2021**				**2021–2022**			
	N	**% Prof**	**Scale Score**	**SD**	**N**	**% Prof**	**Scale Score**	**SD**	**N**	**% Prof**	**Scale Score**	**SD**	**N**	**% Prof**	**Scale Score**	**SD**
Grade 11																
All Students	10,290	32	2569.1	118.9	10,775	30	2564.8	119.3	7,764	28	2562.1	116.8	10,171	26	2550.9	120.0
Female	5,274	34	2579.0	110.6	5,285	32	2574.1	110.2	3,804	29	2566.1	111.9	4,999	26	2555.5	113.4
Male	5,016	30	2558.7	126.1	5,490	29	2555.9	126.7	3,960	27	2558.3	121.3	5,172	25	2546.5	126.0
African American	197	22	2547.4	113.4	231	23	2537.3	115.1	109	17	2536.6	99.3	168	16	2529.3	119.1
AmerIndian/Alaskan	27	30	2560.3	117.8	33	18	2547.2	90.3	16	25	2564.3	108.0	27	15	2543.6	100.7
Asian/Pacific Islander	4,181	43	2601.5	116.4	4,221	40	2598.5	117.3	3,350	37	2589.8	115.1	4,072	35	2583.4	115.9
Hispanic	810	22	2547.1	112.6	867	21	2541.6	109.2	658	19	2535.7	105.6	995	18	2526.2	113.1
Hawai'i Pacific Islander	3,046	17	2519.9	109.6	3,196	14	2513.4	106.3	1,949	13	2508.2	105.9	2,783	11	2496.6	109.5
White	1,170	37	2589.3	115.1	1,273	40	2592.1	113.7	927	35	2589.5	116.4	1,163	34	2575.7	115.9
Multi-Racial	859	35	2584.5	111.0	954	34	2580.0	121.3	755	29	2571.3	109.8	963	31	2569.9	117.5
ELL	390	6	2451.9	102.3	626	6	2467.6	93.0	331	8	2465.7	113.2	572	6	2463.5	103.7
Disadvantaged	3,829	23	2540.2	115.8	3,958	20	2532.5	112.9	2,737	19	2534.3	112.0	3,566	17	2518.0	116.0
Migrant	97	15	2513.2	109.4	125	13	2517.1	109.3	100	12	2508.3	110.0	124	6	2480.2	98.5
Disability	797	2	2420.8	92.4	864	1	2427.9	86.0	541	2	2431.6	95.5	790	2	2412.5	93.6

2019–2020 is not included because the summative assessments were canceled due to the COVID-19 pandemic.

Appendix D: Classification Accuracy and Consistency Index by Subgroup

Table D-1. Classification Accuracy and Consistency by Subgroup: ELA/L (Grades 3–4)

Group	N	%Accuracy						%Consistency					
		All	**L1**	**L2**	**L3**	**L4**	**Proficiency Cut**	**All**	**L1**	**L2**	**L3**	**L4**	**Proficiency Cut**
Grade 3													
All Students	12,991	76	89	62	59	87	91	68	82	51	48	80	87
Female	6,208	75	88	62	59	87	91	67	81	50	48	81	87
Male	6,783	76	89	62	59	86	91	68	83	51	48	79	87
African American	157	72	89	62	58	85	88	63	77	51	49	76	83
AmerIndian/Alaskan	15	70	88*	60*	60*	97*	85	61	81*	48*	55*	59*	79
Asian/Pacific Islander	2,969	76	87	62	59	88	91	68	77	51	49	83	87
Hispanic	2,576	75	89	62	59	85	91	67	83	51	48	77	87
Hawai‘i Pacific Islander	2,983	77	90	62	59	82	92	69	86	50	48	69	88
White	1,428	75	87	62	59	86	91	67	78	50	48	81	87
Multi-Racial	2,863	75	87	62	59	88	91	67	79	51	48	82	87
ELL	1,790	77	90	62	60	82	92	70	86	50	49	70	89
Disadvantaged	5,776	76	90	62	59	84	91	69	85	51	48	75	88
Migrant	145	78	89	62	58	83	93	70	85	53	44	73	90
Disability	1,205	86	94	62	58	80	96	80	91	50	44	66	94
Grade 4													
All Students	12,819	74	88	55	57	85	90	66	82	43	46	78	87
Female	6,173	74	88	55	57	86	90	66	80	43	47	79	86
Male	6,646	74	89	55	57	84	91	67	83	43	46	77	87
African American	158	73	89	55	58	83	89	65	81	46	44	75	85
AmerIndian/Alaskan	15	72	88*	58*	50*	85*	90	64	77*	52*	36*	79*	86
Asian/Pacific Islander	2,964	74	86	55	57	86	91	66	78	43	46	81	87
Hispanic	2,493	74	88	55	57	85	90	66	82	43	47	76	87
Hawai‘i Pacific Islander	2,987	75	90	55	57	81	91	67	85	43	46	69	87
White	1,441	74	88	55	58	87	90	66	78	42	47	81	86
Multi-Racial	2,761	74	88	55	57	85	90	65	79	43	46	79	86
ELL	1,655	77	91	55	58	80	91	69	87	44	46	67	87
Disadvantaged	5,646	74	89	55	57	82	90	66	84	43	46	72	86
Migrant	165	78	92	55	55	80	92	71	88	43	44	70	89
Disability	1,306	86	94	55	57	86	95	81	92	44	41	68	93

*The classification index is based on $n < 10$.

Table D-2. Classification Accuracy and Consistency by Subgroup: ELA/L (Grades 5–6)

Group	N	%Accuracy						%Consistency					
		All	L1	L2	L3	L4	Proficiency Cut	All	L1	L2	L3	L4	Proficiency Cut
Grade 5													
All Students	13,058	76	88	58	67	85	91	68	82	46	56	78	88
Female	6,316	76	87	58	66	86	91	67	79	46	56	79	88
Male	6,742	76	89	58	67	85	91	68	84	46	57	77	88
African American	166	74	85	59	66	87	89	64	77	48	56	77	84
AmerIndian/Alaskan	16	70	63*	58*	68*	87*	90	59	41*	52*	52*	79*	85
Asian/Pacific Islander	3,221	76	87	58	67	86	92	68	78	47	56	81	89
Hispanic	2,495	75	88	58	66	84	91	67	82	46	57	75	87
Hawaiʻi Pacific Islander	3,081	77	90	58	66	82	91	69	85	46	56	72	88
White	1,507	74	85	58	66	85	91	66	75	45	56	78	87
Multi-Racial	2,572	76	88	58	67	86	91	68	81	45	57	79	88
ELL	1,460	78	91	58	66	77	91	71	87	47	56	57	88
Disadvantaged	5,681	76	89	58	66	82	91	68	84	47	56	72	87
Migrant	139	77	89	56	67	79	92	70	85	46	56	70	89
Disability	1,338	84	93	57	65	81	94	78	91	45	51	65	91
Grade 6													
All Students	12,841	76	88	65	69	83	91	67	82	54	60	74	88
Female	6,234	75	88	65	69	84	91	66	79	54	60	74	87
Male	6,607	77	89	65	69	82	92	68	83	54	60	72	88
African American	173	76	89	65	69	82	91	67	81	54	60	73	87
AmerIndian/Alaskan	16	70	82*	60*	59*	75*	89	63	77*	53*	48*	72*	86
Asian/Pacific Islander	3,296	76	88	65	69	85	91	67	79	54	60	77	87
Hispanic	2,395	76	88	66	69	82	91	67	82	55	60	70	87
Hawaiʻi Pacific Islander	3,143	78	90	65	69	80	92	70	85	55	59	66	89
White	1,407	74	85	65	68	82	91	65	74	53	59	75	87
Multi-Racial	2,411	75	87	66	69	83	91	66	79	54	60	73	87
ELL	1,411	81	91	65	69	80	94	74	87	54	56	48	91
Disadvantaged	5,748	77	89	66	69	80	91	68	84	55	59	68	88
Migrant	191	80	93	66	70	66*	93	73	87	55	59	52*	90
Disability	1,336	86	93	64	69	77	96	81	91	53	54	54	94

*The classification index is based on *n* < 10.

Table D-3. Classification Accuracy and Consistency by Subgroup: ELA/L (Grades 7–8)

Group	**N**	**%Accuracy**						**%Consistency**					
		All	**L1**	**L2**	**L3**	**L4**	**Proficiency Cut**	**All**	**L1**	**L2**	**L3**	**L4**	**Proficiency Cut**
Grade 7													
All Students	9,922	76	89	64	72	82	91	67	82	52	63	72	87
Female	4,745	76	88	64	72	82	91	67	80	52	63	72	87
Male	5,177	77	89	64	72	82	91	68	83	52	63	71	87
African American	146	74	86	63	72	86	86	65	75	52	65	73	81
AmerIndian/Alaskan	13	77	99*	61*	73*	84*	91	68	80*	46*	67*	75*	87
Asian/Pacific Islander	2,498	76	88	64	71	83	91	67	79	52	63	74	87
Hispanic	1,909	76	88	63	72	81	90	67	82	52	63	69	86
Hawai‘i Pacific Islander	2,458	78	90	64	72	79	90	69	84	53	62	63	86
White	1,183	76	88	65	72	84	92	67	77	52	64	75	88
Multi-Racial	1,715	75	88	63	72	81	91	66	79	52	64	71	87
ELL	1,107	80	92	64	71	68	92	73	87	55	57	47	89
Disadvantaged	4,454	77	89	64	72	80	90	68	84	53	63	65	86
Migrant	155	78	90	65	72	83*	90	69	84	55	61	57*	86
Disability	1,125	84	92	63	72	77	95	79	90	51	58	55	92
Grade 8													
All Students	12,456	76	88	66	72	82	91	67	80	55	64	71	87
Female	6,076	75	86	66	72	82	90	66	77	55	64	71	86
Male	6,380	77	89	66	72	82	91	68	82	55	64	71	88
African American	182	73	86	64	71	81	87	64	76	53	64	70	82
AmerIndian/Alaskan	17	75	87*	66*	69*	80*	85	65	84*	53*	59*	71*	81
Asian/Pacific Islander	3,475	76	86	66	72	83	91	66	77	54	65	73	87
Hispanic	2,202	77	89	66	73	82	90	68	81	56	64	70	87
Hawai‘i Pacific Islander	2,955	78	89	66	73	76	91	69	83	56	63	58	88
White	1,383	75	87	66	72	83	90	66	75	55	64	73	86
Multi-Racial	2,242	75	87	66	72	82	91	66	77	56	63	71	87
ELL	1,198	80	90	66	72	66	92	72	86	57	59	38	89
Disadvantaged	5,439	77	88	66	72	80	91	68	82	56	63	66	87
Migrant	200	80	89	67	75	77*	94	73	85	55	64	65*	92
Disability	1,217	86	93	65	71	79	96	80	91	53	57	55	94

*The classification index is based on *n* < 10.

Table D-4. Classification Accuracy and Consistency by Subgroup: ELA/L (Grade 11)

Group	**N**	**%Accuracy**						**%Consistency**					
		All	**L1**	**L2**	**L3**	**L4**	**Proficiency Cut**	**All**	**L1**	**L2**	**L3**	**L4**	**Proficiency Cut**
Grade 11													
All Students	10,033	75	86	66	69	84	91	67	77	55	60	76	87
Female	4,924	75	85	66	70	84	91	66	74	54	60	77	87
Male	5,109	76	87	66	69	83	91	67	80	55	59	75	87
African American	165	75	91	62	71	83	91	67	79	54	60	77	88
AmerIndian/Alaskan	26	70	73*	65*	65	85*	84	62	67*	51*	59	74*	80
Asian/Pacific Islander	4,024	75	85	66	69	85	91	66	73	54	60	77	88
Hispanic	986	75	89	66	70	82	90	66	79	56	60	73	86
Hawai'i Pacific Islander	2,716	75	87	67	70	81	91	66	80	56	60	69	87
White	1,157	76	86	66	68	84	92	68	79	53	59	78	89
Multi-Racial	959	76	84	66	70	85	91	67	76	53	60	78	88
ELL	549	78	89	66	70	61*	92	70	83	56	58	36*	88
Disadvantaged	3,499	75	87	66	69	82	90	66	79	56	59	72	87
Migrant	126	76	89	65	71	79	91	68	84	55	59	67	87
Disability	771	83	91	67	70	82	94	76	88	56	55	58	92

*The classification index is based on *n* < 10.

Table D-5. Classification Accuracy and Consistency by Subgroup: Mathematics (Grades 3–4)

Group	**N**	**%Accuracy**						**%Consistency**					
		All	**L1**	**L2**	**L3**	**L4**	**Proficiency Cut**	**All**	**L1**	**L2**	**L3**	**L4**	**Proficiency Cut**
Grade 3													
All Students	13,041	77	86	64	71	88	92	69	79	51	61	82	89
Female	6,231	77	86	64	71	87	92	69	79	51	61	81	88
Male	6,810	78	85	63	71	88	92	69	79	51	61	82	89
African American	157	75	82	65	72	85	91	65	72	55	61	75	86
AmerIndian/Alaskan	16	76	95*	64*	76*	98*	92	67	72*	58*	71*	64*	89
Asian/Pacific Islander	2,990	78	83	63	71	89	92	70	73	52	61	85	88
Hispanic	2,581	77	86	64	71	86	92	68	80	52	60	79	88
Hawai'i Pacific Islander	2,998	78	87	63	71	84	93	69	83	50	60	73	90
White	1,432	77	82	65	71	87	91	69	73	53	60	83	88
Multi-Racial	2,867	77	85	63	71	88	92	69	77	51	61	82	88
ELL	1,812	78	86	63	71	86	93	70	82	50	60	78	90
Disadvantaged	5,797	77	86	64	71	84	92	69	81	51	60	76	89
Migrant	146	81	89	66	69	80*	92	72	86	48	59	72*	90
Disability	1,212	81	87	64	71	78	96	74	86	45	60	68	94
Grade 4													
All Students	12,872	79	87	73	71	87	92	71	81	63	61	80	89
Female	6,190	78	87	72	70	87	91	70	80	63	61	79	88
Male	6,682	80	88	73	71	88	92	72	81	63	61	81	89
African American	159	76	85	72	69	84	89	67	77	63	58	75	85
AmerIndian/Alaskan	14	83	96*	70*	83*	91*	87	74	93*	68*	56*	83*	83
Asian/Pacific Islander	2,979	79	85	73	71	89	92	71	76	63	62	83	89
Hispanic	2,498	78	87	73	69	84	92	70	81	63	59	76	89
Hawai'i Pacific Islander	3,008	80	89	72	71	83	93	72	84	63	60	73	90
White	1,448	79	85	73	71	89	91	71	77	62	63	83	88
Multi-Racial	2,766	78	86	73	71	87	91	70	78	62	62	80	88
ELL	1,681	81	89	72	72	85	94	73	85	63	59	75	91
Disadvantaged	5,676	79	88	72	70	86	92	71	83	63	60	77	89
Migrant	165	80	87	72	72	82	94	72	83	60	57	73	91
Disability	1,321	86	91	72	68	85	97	80	89	58	56	75	95

*The classification index is based on *n* < 10.

Table D-6. Classification Accuracy and Consistency by Subgroup: Mathematics (Grades 5–6)

Group	N	%Accuracy						%Consistency					
		All	L1	L2	L3	L4	Proficiency Cut	All	L1	L2	L3	L4	Proficiency Cut
Grade 5													
All Students	13,096	78	88	68	61	88	92	70	82	57	49	81	89
Female	6,336	77	87	68	61	87	92	69	81	57	49	81	89
Male	6,760	79	89	68	61	88	93	71	83	57	50	82	90
African American	165	78	86	69	64	89	93	69	79	61	47	79	90
AmerIndian/Alaskan	16	65	81*	69*	55*	–	83	57	77*	54*	52*	–	78
Asian/Pacific Islander	3,235	78	85	68	61	89	92	70	78	57	50	85	89
Hispanic	2,497	77	88	68	61	86	93	69	82	57	49	78	89
Hawai‘i Pacific Islander	3,090	80	90	68	61	84	94	72	86	57	48	73	91
White	1,515	76	86	68	61	87	91	67	77	58	50	81	87
Multi-Racial	2,578	77	87	68	61	88	92	69	81	57	49	82	89
ELL	1,464	82	91	68	61	84	94	74	87	55	48	71	92
Disadvantaged	5,698	79	89	68	61	84	93	71	85	57	48	76	90
Migrant	137	82	89	68	61	85*	96	75	87	52	51	74*	94
Disability	1,336	86	92	66	62	79	97	80	90	51	44	71	96
Grade 6													
All Students	12,888	78	89	68	60	86	92	70	84	58	48	78	88
Female	6,255	77	89	68	60	86	92	69	83	58	48	78	88
Male	6,633	78	90	68	60	86	92	71	85	57	48	79	88
African American	174	77	90	68	61	85	91	69	83	59	47	77	87
AmerIndian/Alaskan	16	88	96*	69*	63*	93*	93	82	93*	59*	40*	91*	89
Asian/Pacific Islander	3,302	76	87	68	60	87	91	68	80	57	49	81	87
Hispanic	2,401	78	90	68	60	83	92	71	85	58	48	73	89
Hawai‘i Pacific Islander	3,163	81	91	68	60	84	93	74	88	57	47	71	90
White	1,417	76	85	68	61	87	91	67	75	59	48	81	87
Multi-Racial	2,415	76	88	68	60	86	91	68	81	58	48	78	88
ELL	1,423	85	92	68	59	85	95	79	90	56	43	69	92
Disadvantaged	5,781	80	90	68	60	84	93	73	87	57	48	74	90
Migrant	192	83	91	69	61	81*	94	77	89	58	46	65*	91
Disability	1,340	89	95	67	60	83	96	85	93	54	44	66	94

*The classification index is based on *n* < 10.; Cells with “–” indicate no data available.

Table D-7. Classification Accuracy and Consistency by Subgroup: Mathematics (Grades 7–8)

Group	N	%Accuracy						%Consistency					
		All	L1	L2	L3	L4	Proficiency Cut	All	L1	L2	L3	L4	Proficiency Cut
Grade 7													
All Students	9,959	78	89	66	63	86	91	70	83	56	51	77	87
Female	4,761	78	90	66	63	85	91	70	83	57	51	76	87
Male	5,198	77	89	66	63	86	91	70	84	55	52	78	88
African American	143	76	88	66	64	82	90	67	82	57	52	68	85
AmerIndian/Alaskan	14	79	84*	69*	70*	78*	98	71	85*	47*	53*	77*	96
Asian/Pacific Islander	2,498	76	87	66	63	87	91	68	80	56	52	81	87
Hispanic	1,921	78	89	66	63	82	91	70	84	56	51	71	87
Hawai'i Pacific Islander	2,484	81	91	66	63	81	92	74	87	55	49	68	89
White	1,181	75	87	66	63	87	90	66	77	57	52	79	86
Multi-Racial	1,718	76	88	65	63	85	91	67	81	56	52	75	87
ELL	1,126	86	93	67	63	83	94	81	91	54	48	70	90
Disadvantaged	4,482	79	90	65	63	85	92	72	86	55	50	73	88
Migrant	158	82	91	66	65	85*	94	75	86	57	50	64*	91
Disability	1,128	89	94	65	62	89	95	85	93	52	41	77	91
Grade 8													
All Students	12,511	76	87	61	59	86	92	68	82	50	47	77	88
Female	6,101	75	87	61	59	86	91	67	81	50	46	76	88
Male	6,410	76	88	61	59	86	92	68	83	50	47	78	89
African American	182	74	84	61	59	85	90	65	81	48	48	74	87
AmerIndian/Alaskan	17	76	90*	58*	65*	74*	95	68	81*	50*	45*	73*	92
Asian/Pacific Islander	3,479	75	86	62	59	88	91	66	78	51	48	81	87
Hispanic	2,216	76	88	61	59	84	92	68	83	50	45	72	89
Hawai'i Pacific Islander	2,993	80	89	61	59	80	94	72	86	48	43	64	92
White	1,389	73	86	61	59	86	90	64	77	50	49	76	85
Multi-Racial	2,235	74	86	61	59	85	91	65	79	50	46	75	87
ELL	1,211	84	91	60	58	85	96	77	89	45	42	69	95
Disadvantaged	5,471	78	89	61	58	83	93	70	85	49	44	72	90
Migrant	199	82	90	61	59	86*	95	75	88	47	43	70*	93
Disability	1,231	88	92	59	57	84	98	82	91	39	38	67	97

*The classification index is based on *n* < 10.

Table D-8. Classification Accuracy and Consistency by Subgroup: Mathematics (Grade 11)

Group	**N**	**%Accuracy**						**%Consistency**					
		All	**L1**	**L2**	**L3**	**L4**	**Proficiency Cut**	**All**	**L1**	**L2**	**L3**	**L4**	**Proficiency Cut**
Grade 11													
All Students	10,171	79	89	64	70	84	92	71	85	54	58	74	89
Female	4,999	78	89	64	71	82	92	70	84	55	58	70	88
Male	5,172	80	90	64	70	86	93	72	86	54	58	76	90
African American	168	80	92	63	72	81	92	73	86	57	50	77	89
AmerIndian/Alaskan	27	75	82	62*	63*	68*	92	67	82	50*	40*	72*	89
Asian/Pacific Islander	4,072	77	88	64	71	85	91	68	81	55	59	75	87
Hispanic	995	81	90	64	69	89	94	73	86	53	57	74	91
Hawaiʻi Pacific Islander	2,783	83	91	64	71	78	95	76	88	54	54	62	92
White	1,163	76	88	64	70	84	91	68	81	55	59	72	88
Multi-Racial	963	77	88	65	70	82	92	69	83	55	59	74	88
ELL	572	87	93	64	65	71*	96	82	92	51	48	58*	95
Disadvantaged	3,566	82	91	64	70	83	94	75	88	53	57	70	91
Migrant	124	85	94	65	65*	71*	95	80	91	58	42*	64*	93
Disability	790	94	96	63	61	97*	99	91	96	46	45	78*	98

*The classification index is based on *n* < 10.