Idaho Standards Achievement Tests in English Language Arts and Mathematics

2023–2024 Technical Report

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

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EXHIBITS

1. OVERVIEW

This report provides a technical summary of the 2023–2024 Idaho administration of the Smarter Balanced summative assessments in English language arts/literacy (ELA/L) and mathematics in grades 3–8, 11. All students enrolled in grades 3–8 and 11 in all public schools were required by the State Board of Education (SBOE) to participate in the Smarter Balanced summative assessments. The report includes eight chapters: Overview, Test Administration, Summary of 2023–2024 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 tests and data on students' performance are provided in Appendix A, Summary of the 2023–2024 Interim Assessments. The data included in this report are based on the Idaho data for the Smarter Balanced assessment in ELA/L and mathematics.

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

1.1 SMARTER BALANCED ASSESSMENTS IN IDAHO

In 2010, the Smarter Balanced Assessment Consortium (SBAC) began developing a next-generation assessment system. The assessments were designed to measure the new Common Core State Standards (CCSS) in English language arts/literacy (ELA/L) and mathematics for grades 3–8 and high school, and to provide valid, reliable, and fair test scores for student academic achievement. The Smarter Balanced assessments consist of the end-of-year summative assessment designed for accountability purposes and the optional interim assessments designed to support teaching and learning throughout the year. The summative assessments are used to determine student achievement based on the Idaho Content Standards and to track student progress toward college and career readiness in ELA/L and mathematics. The summative assessments consist of two parts:

- 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 respond to real-world problems. PTs can best be described as collections of items and activities that are coherently connected to a single theme or scenario. They are used to better measure capacities such as depth of understanding, research skills, and complex analysis, which cannot be adequately assessed with selected- or constructed-response items. Some PT items can be scored by the computer, but most are hand-scored.

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

Idaho administered three types of interim assessments as fixed-form tests developed by Smarter Balanced Assessment Consortium:

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

In addition, Idaho created and administered the **Shortened Interim Comprehensive Assessments (SICAs)** by dropping all short answer items and the PT component from the ICAs to assess student performance with reduced testing time.

The Idaho State Board of Education (SBOE) formally adopted the CCSS in ELA/L and mathematics on August 12, 2010 (SBOE meeting minutes, 2010). The Idaho Content Standards were updated on August 13, 2015 (SBOE meeting minutes, 2015). The Idaho Content Standards define the knowledge and skills that students need to succeed in college and careers. These standards include rigorous content and application of knowledge through higher-order skills, and they align with college and workforce expectations. Idaho was one of 19 jurisdictions (18 states and the U.S. Virgin Islands) leading the development of the ELA/L and mathematics assessments.

The statewide assessments were first administered to students in spring 2015. Starting in spring 2015, Idaho adopted the Smarter Balanced full blueprint, requiring all students in grades 3–11 in public elementary and secondary schools to be assessed, with the accountability grade in high school set at grade 10. Testing at grades 9 and 11 was optional.

In the 2019–2020 school year, the U.S. Department of Education waived testing requirements due to the COVID-19 pandemic (https://www2.ed.gov/policy/gen/guid/secletter/200320.html). For the 2020–2021 school year, the U.S. Department of Education did not grant waivers for standardized testing but did waive certain accountability requirements (e.g., mandatory high participation rates) due to the impacts of the pandemic in many states. Starting in the 2021–2022 school year, all students were again required to take ELA/L and mathematics summative assessments.

In the 2020–2021 school year, Idaho adopted the Smarter Balanced adjusted blueprint. In the following 2021–2022 school year, Idaho continued to use the adjusted blueprint and also permitted remote testing. In 2022–2023, Idaho made three changes: (1) changed the accountability grade in high school from 10 to 11, (2) removed testing for grades 9 and 10, and (3) changed the summative test blueprint from the adjusted blueprint back to the full blueprint. In the 2023–2024 school year, Idaho changed back to the Smarter Balanced adjusted blueprints for grades 3–8 and 11.

AIR delivered the assessments through the 2019–2020 school year. Starting with the 2020–2021 school year, Cambium Assessment, Inc. (CAI) delivered and scored the assessments and produced score reports. The transition from AIR to CAI for the Idaho assessment program did not entail any major changes in contractors or require new staff members, allowing for continuity in the Idaho assessment program. Measurement Incorporated (MI) scored the hand-scored items.

2. TEST ADMINISTRATION

2.1 TESTING WINDOWS

The 2023–2024 Idaho Standards Achievement Tests (ISATs) English language arts/literacy (ELA/L) and mathematics assessments testing window spanned approximately three months for the online summative assessments and approximately nine months for the interim assessments. The paper-pencil fixed-form tests for the summative assessments were administered over an eight-week period during the online summative assessment testing window. Table 1 shows the testing windows for both online and paper-pencil summative and interim assessments.

Table 1. 2023–2024 ISAT Testing Windows

Tests	Grade	Start Date	End Date	Mode
Summative Assessments	3–8, 11	3/11/2024	5/24/2024	Online Adaptive
	3–8, 11	4/1/2024	5/24/2024	Paper Fixed-Form
Interim Comprehensive Assessments, Shortened Interim Comprehensive Assessments	3–11*	9/11/2023 6/3/2024	2/23/2024 7/26/2024	Online Fixed-Form
Interim Assessment Blocks, Focused Interim Assessment Blocks	3–8, 11	9/11/2023 6/3/2024	2/23/2024 7/26/2024	Online Fixed-Form

* Grade 9 and 10 tests were available from 9/11/2023 to 7/26/2024.

2.2 TEST OPTIONS AND ADMINISTRATIVE ROLES

The ISAT ELA/L and mathematics assessments are administered primarily online. To ensure that all eligible students in the tested grades were given the opportunity to take the ISAT ELA/L and mathematics assessments, several assessment options were available for the 2023–2024 administration to accommodate students' needs. Table 2 lists the testing options that were offered in 2023–2024. A testing option was selected by content area. Once an option was selected, it applied to all tests in the content area.

Table 2. Summary of Tests and Testing Options in 2023–2024

Assessments	Test Options	Test Mode
Summative Assessments	English	Online Adaptive
	Braille	Online Adaptive
	Spanish (mathematics only)	Online Adaptive
	English	Online Fixed Form
	Braille	Online Fixed Form
	Spanish (mathematics only)	Online Fixed Form
	Braille	Paper
	Regular Print Fixed-Form	Paper
	Large Print Fixed-Form	Paper
Interim Assessments	English	Online
	Spanish (mathematics only)	Online
	Braille	Online

To ensure standardized administration conditions, teachers (TEs) and test administrators (TAs) follow procedures outlined in the *ISAT Summative Test Administration Manual (TAM)*. TEs and TAs must review the manual before testing to ensure that the testing room is prepared appropriately (e.g., removing certain classroom posters, arranging desks) and read the boxed directions verbatim to students before and during testing to maintain standardized administration conditions. Reading the *ISAT Summative Test Administration Manual* was included in the readiness checklist for each user role. Make-up procedures should be established for any students who are absent on testing day(s).

2.2.1 Administrative Roles

The key personnel involved with test administration are District Administrators (DAs), District Test Coordinators (DCs), School Test Coordinators (SCs), Teachers (TEs), and Test Administrators (TAs). The main responsibilities of these key personnel are described below. More detailed descriptions can be found in the *ISAT Summative Test Administration Manual (TAM)* provided online at https://idaho.portal.cambiumast.com/resource-item/en/isat-summative-test-administration-manual.

District Administrator (DA)

The DA's role is assigned by the Idaho State Department of Education (The Department) and is usually the district superintendent. The DA is authorized to add users to the Test Information Distribution Engine (TIDE) and to assign them any role except that of a DA. DAs and DCs share many of the same test administration responsibilities. Their primary responsibility is to coordinate the administration of the ISAT ELA/L and mathematics assessments in the district.

District Test Coordinator (DC)

The DC's primary responsibility is to coordinate the administration of the ISAT ELA/L and mathematics assessments in the district. For smaller districts and charter schools, the DC is often also the SC.

DCs are also responsible for performing the following functions:

- Reviewing all state and Smarter Balanced policies and test administration documents
- Reviewing scheduling and test requirements with SCs, TEs, and TAs
- Working with SCs and Technology Coordinators to ensure that all systems, including the CAI Secure Browser, are properly installed and functioning
- Importing users (SCs, TEs, and TAs) into TIDE
- Entering and verifying all student information, eligibility, and test settings in TIDE
- Scheduling and administering training sessions for all SCs, TEs, TAs, and Technology Coordinators
- Ensuring that all personnel are trained on how to properly administer the ISAT ELA/L and mathematics assessments
- Monitoring the secure administration of the tests
- Investigating and recording all testing improprieties, incidents, and breaches reported by TEs/TAs
- Attending to any secure materials in accordance with state and Smarter Balanced policies

School Test Coordinator (SC)

The SC's primary responsibilities are to coordinate the administration of the ISAT ELA/L and mathematics assessments and ensure that testing within his or her school is conducted in accordance with the test procedures and security policies established by the Department.

SCs are responsible for performing the following functions:

- Establishing a testing schedule with DCs, TEs, and TAs based on testing windows
- Working with technology staff to ensure timely computer setups and installations
- Working with TEs and TAs to review student information in TIDE to ensure that student information and test settings for designated supports and accommodations are correctly applied
- Entering student test settings in TIDE
- Identifying students who may require designated supports and test accommodations and ensuring that procedures for testing these students follow state and Smarter Balanced policies
- Attending all district training sessions and reviewing all state and Smarter Balanced policies and test administration documents
- Ensuring that all TEs and TAs attend school or district training sessions 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, incidents, and breaches reported by the TEs and TAs
- Attending to any secure material in accordance with state and Smarter Balanced policies

Teacher (TE)

A TE responsible for administering the ISAT ELA/L and mathematics assessments must have the same qualifications as a TA. They also have the same test administration responsibilities as those outlined below under TA. TEs can view student results when they are made available. This role may also be assigned to teachers who do not administer the test but will need access to student results.

Test Administrator (TA)

TAs are primarily responsible for administering the ISAT ELA/L and mathematics assessments. This role is designed for TAs, such as technology staff, who administer tests but should not have access to student results.

TAs are responsible for performing the following functions:

- Completing ISAT ELA/L and mathematics assessments administration training
- Reviewing all state and Smarter Balanced policies and test administration documents before administering any ISAT ELA/L and mathematics assessments
- Viewing student information before testing to ensure a student receives the proper test with the appropriate supports (TAs should report any potential data errors to SCs and DCs as appropriate)
- Administering the ISAT ELA/L and mathematics assessments
- Reporting all potential test security incidents to the SCs and DCs in a manner consistent with Smarter Balanced, state, and district policies

2.2.2 Online Administration

Within the state's testing window, schools can set testing schedules, allowing students to test in intervals (e.g., multiple sessions) rather than in one long period, minimizing the interruption of classroom instruction and efficiently using its facility. With online testing, schools do not need to address the on-site storage and security issues associated with large shipments of printed testing materials.

SCs oversee all aspects of testing at their schools and serve as the main point of contact; TEs and TAs administer the assessments only. TEs and TAs are trained in the online testing requirements and the mechanics of starting, pausing, and ending a test session. Training materials for the test administration are available online and at regional face-to-face training sessions. All school personnel who serve as test proctors are required to complete an online TA Certification Course before testing begins. Upon completion of this course, staff members receive a certificate and authorization to log in to the online testing system. School personnel that were proctoring tests remotely were required to take the TA Certification Course for Remote Testing.

The interim assessments were administered both in-person and remotely in the 2023–2024 school year. The interim assessments could be accessed in a conventional browser. The summative assessments were administered in-person only, and the Secure Browser was required for the summative assessments.

To start a test session, the TE or TA must first access the TA Interface of the online testing system using his or her own computer. A test session ID is generated when the test session is created. Students who are taking the assessment with the TE or TA need to enter their Education Unique Identification (EDUID) number, first name, and test session ID into the Student Interface using computers provided by the school. The TE or TA then verifies that the students are taking the appropriate assessments with the appropriate accessibility feature(s) (see Section 2.6 for a list of accommodations). Students can begin testing only when the TE or TA confirms the settings. The TE or TA will then read aloud the *Test Administration Script* in the *ISAT Summative Test Administration Manual (TAM)* to the student(s) and guide them through the login process.

For students that are testing remotely, teachers need to communicate links to the test session, session IDs, and EDUIDs to their students so students can take tests that were scheduled in advance. This information should not be shared over unsecured communication methods like personal email or text messages. Instead, teachers should communicate this information to students using a secure method, such as whichever classroom management system teachers and students are already using for instructional purposes.

Once an assessment begins, the student must answer all test items presented on a page before proceeding to the next page. Skipping items is not permitted. For the online computer-adaptive test (CAT), students are allowed to review and edit previously answered items, as long as these items are in the same test session and this session has not been paused for more than 20 minutes. Students may review and edit previously completed responses until they submit the assessment. During an active CAT session, even if a student reviews and changes the response to a previously answered item the responses to any following items to which the student already responded remain the same. No new items are assigned to this student because he or she changed an answer. For example, a student paused for 10 minutes after completing item 10. After the pause, the student returned to item 5 and changed the answer. If the response change in item 5 changed the item score from incorrect to correct, the student's overall score improves; however, there is no change in items 6–10.

For the performance tasks (PTs), there is no pause rule, but the same rules that apply to CATs for reviews and changes to responses also apply to PTs.

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

During a test session, TEs or TAs may pause the test for a student or group of students to provide a break. It is up to the TEs or TAs to determine an appropriate stopping point. However, for ELA/L and mathematics CATs, the assessments cannot be paused for more than 20 minutes to ensure the integrity of the test scores or testing. If an assessment is paused for more than 20 minutes, the student will resume testing at the next unanswered item where the test was paused. The student may not view or edit any previous responses.

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

2.2.3 Paper-Pencil Test Administration

The paper-pencil versions of the ISAT ELA/L and mathematics assessments are provided as an option for students who do not have access to a computer or students with blindness or visual impairments. Paper-pencil versions are also offered as a designated support for students who benefit from the paper and pencil modality. For Idaho, paper-pencil tests were offered in regular print, braille, and large print formats.

In a district with student(s) who need to take the paper-pencil version of a test, the DA or DC must submit a request to the Department for appropriate materials on behalf of the student(s). If the request is approved, the testing contractor will ship the appropriate test booklets, receipt instructions, and return instructions to the district.

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

After the student has completed the assessments, DAs, DCs, SCs, TEs, and/or TAs must transcribe the student's responses into the Data Entry Interface (DEI) and return the test booklets to the testing vendor. The testing vendor will score the hand-scored items. Once the hand-scored items are scored, scores will be combined with the machine-scored items, and the final score will appear in the Reporting System.

The total number of students who took paper-pencil tests is presented in Table 3.

Table 3. Number of Students Who Took Paper-Pencil Tests
in 2023–2024 Summative Test Administration

Subject	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 11	Total
ELA/L	1	1	1	1	18	16	2	40
Mathematics	1	1	1	1	18	17	2	41

2.2.4 Braille Test Administration

The adaptive braille test was available with the same test blueprint in English in both ELA/L and mathematics. In the 2017–2018 test administration, Smarter Balanced added the Braille Hybrid Adaptive Test (Braille HAT) for mathematics. The Braille HAT consists of a fixed-form segment, a 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 the Department. All items on the Braille HAT can be presented to the students using a Refreshable Braille Display (RBD). In the 2023–2024 school year, the Braille HAT assessment was not offered in Idaho.

The braille interface is described below:

- 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 via a braille embosser through the adaptive online summative test and a fixed-form PT. The following braille codes and formats are offered for math:
    - UEB contracted with Nemeth
    - UEB uncontracted with Nemeth
    - UEB contracted with UEB mathematics (technical)
    - UEB uncontracted with UEB mathematics (technical)
- 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, TEs or TAs must ensure that technical requirements are met. These requirements apply to the student's computer, the TE/TA's computer, and any assistive braille technologies used in conjunction with the braille interface.

2.3 Training and Information for Test Coordinators and Administrators

All DAs, DCs, and SCs oversee all aspects of testing at their LEAs and serve as the main point of contact, while TEs and TAs administer the online assessments. The online TA Certification Course, webinars, manuals, and regional training sites are used to train TEs and TAs on the online testing requirements and the mechanics of starting, pausing, and ending a test session. Training materials for test administration are available online (https://idaho.portal.cambiumast.com/resources).

2.3.1 Online Training

Multiple training opportunities are offered online.

TA Certification Course

All school personnel who serve as test proctors are required to complete an online TA Certification Course to administer assessments. This web-based course is about 20 minutes long and covers information on testing policies and steps for administering a test session in the online system. The course is interactive, requiring participants to practice starting test sessions under different scenarios. Throughout the training and at the end of the course, participants are required to answer multiple-choice items about the information provided. Completion of the TA Certification Course is tracked online in TIDE.

Modules

The following training modules were created to help users in the field understand the overall ISAT ELA/L and mathematics assessments as well as how each system works. The modules are provided as PowerPoint presentations.

Accessing Portal Resources Video Tutorial: This video tutorial provides guidance on how to access different resources on the Idaho portal.

Assessment Viewing Application (AVA) Module: This module explains how to navigate AVA, which allows authorized users to view the Interim Assessments for administrative and instructional purposes.

Authoring Training Module: This module trains users on how to use the Authoring System.

Authoring Tutorials: These tutorials explain different features in the Authoring System. The following tutorials are available:

1. Basic Use Dashboard
2. Everything Items
3. Add Images to Items
4. Everything Tests
5. Share Content
6. TDS Session
7. Reporting

Braille Training Module: This presentation provides detailed information on administering tests to students using braille.

Interim Assessment Implementation Video Tutorial: This video tutorial outlines the tasks for administering Interim Assessments. The optional Interim Assessments are given to students throughout the year to help teachers monitor student progress. This video also provides information on available materials and resources specific to Interim Assessments.

ISAT Supports and Accommodations Presentation: This presentation provides guidance to educators on the use of allowable universal tools, designated supports, and accommodations.

Reporting Training Module: This module is designed to help district-level personnel, school-level staff, and classroom teachers navigate and view student performance reports with the Reporting System.

Reporting Tutorials: These tutorials explain how to navigate the Reporting System. The following tutorials are available:

1. The Basics
2. View Results
3. ISRs and SDFs
4. Interims and Item-Level Data

Student Interface Training Module: This module demonstrates how students can navigate the practice tests, interim assessments, and summative assessments offered through CAI.

Technology Requirements for Online Testing Module: This module provides current information about technology requirements, site readiness, supported devices, and CAI Secure Browser installation, and is designed to help Technology Coordinators prepare for the administration of online tests.

Test Administrator Interface Training Module: This module provides an overview of the test session setup and student sign-in process.

TIDE Training Presentation: This presentation provides guidance on TIDE. It was designed train all TIDE users in tasks that must be completed before, during, and after testing. The presentation is divided into different sections that are customized for DAs, DCs, SCs, TEs, and Tas and can be used to train other users.

Practice and Training Test Site

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

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

A student can log in directly to the practice and training test site as a "Guest" without a TA-generated test session ID, or the student can log in using a training test session created by the TE or TA in the TA Training

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

Manuals and User Guides

The following manuals and user guides are available on the ISAT portal (https://idaho.portal.cambiumast.com/resources):

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

The *Test Administrator (TA) User Guide* is designed to help users navigate TDS, including the Student Interface and the TA Interface, and to help support TAs in managing and administering online testing for students.

The *Authoring User Guide* provides guidance on how to create items, build tests, and share content with others.

The *Interim Guide for Test Administration* describes the interim assessments and provides administration details and policy information for DCs and SCs regarding policies and procedures for the interim assessments.

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

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

The *ISAT Summative Test Administration Manual* provides information for DCs and SCs regarding policies and procedures for the 2023–2024 ISAT ELA/L and mathematics assessments. This manual also provides information for TEs and TAs administering the ISAT ELA/L and mathematics assessments.

The *Technology Guide* includes instructions for set up and configuration of devices and assistive technologies for online testing. This guide is used with operating system-specific manuals to provide information about hardware, software, and network configurations for running various testing applications provided by CAI.

The *Assistive Technology Manual for Windows & macOS* provides an overview of the embedded and non-embedded assistive technology tools that can be used to help students with accessibility needs complete online tests in the Test Delivery System (TDS).

The *Systems and User Roles Chart* offers an overview of Idaho's current assessment systems, detailing how users can access each system and the tasks they are authorized to perform.

The *Date Entry Interface (DEI) User Guide* provides guidance for users entering student information into the Data Entry Interface.

All manuals and user guides pertaining to 2023–2024 online testing are available on the Idaho Portal. DAs, DCs, and SCs can use these manuals and user guides to train TEs and TAs regarding test administration policies and procedures.

Quick Guides

The following quick guides were created to highlight the most important information for all the systems used for the ISAT ELA/L and mathematics assessments.

The *Reporting Quick Guide* provides instructions on how to use the system, specifically for accessing the interim and summative assessment results, navigating reports, setting up individual student reports, exporting and printing data, and scoring interim assessments.

The *Dual Enrollments in TIDE Quick Guide* describes a feature in TIDE for users to have the ability to enroll students in multiple districts or schools.

The *Test Administration (TA) Quick Guide* provides information to help users access and navigate the TA Interface.

The *Test Information Distribution Engine (TIDE) Quick Guide* assists users with managing accounts and settings for users and students in TIDE.

The *ISAT Practice Tests Quick Guide* provides directions for administer practice tests for the ISAT ELA/L, mathematics, and science assessments.

The *Configurations for iPads Quick Guide* contains configurations for iPads and instructions for online testing using iOS operating systems.

The *Configurations, Troubleshooting, and Advanced Secure Browser Installation Quick Guide for Chrome OS* contains configurations, network troubleshooting, and advanced secure browser installation instructions for online testing using Chrome operating systems.

The *Configurations, Troubleshooting, and Advanced Secure Browser Installation Quick Guide for Mac* contains configurations, network troubleshooting, and advanced secure browser installation instructions for online testing using Mac operating systems.

The *Configurations, Troubleshooting, and Advanced Secure Browser Installation Quick Guide for Windows* contains configurations, network troubleshooting, and advanced secure browser installation instructions for online testing using Windows operating systems.

The *Speech-to-Text (STT) Quick Guide* provides guidance for test administrators on how to support students using Speech-to-Text during testing.

The *TIDE Test Improprieties Quick Guide* provides guidance for test administrators who need to submit a test impropriety in TIDE.

The *Embedded Supports and Accommodations Quick Guide* describes the universal tools, designated supports, and accommodations entered in TIDE that students are permitted to use while testing.

2.4 TEST SECURITY

All test items, test materials, and student-level testing information for each Idaho assessment are considered secure materials. The importance of maintaining test security and the integrity of test items is stressed throughout the webinar training sessions and in the user guides, modules, and manuals. Features within the testing system also protect test security. This section describes system security, student confidentiality, and policies on testing incidents, improprieties, and breaches.

2.4.1 Student-Level Testing Confidentiality

All secured websites and software systems enforce role-based security models that protect individual privacy and confidentiality in a manner consistent with the Family Educational Rights and Privacy Act (FERPA) and other federal and state laws. Idaho Code §33-133 specifically states that student data privacy is a top priority for the state of Idaho, ensuring that confidential student information is protected. Secure transmission and password-protected access are basic features of the current system and ensure authorized data access. All aspects of the system, including item development and review, test delivery, and reporting, are secured by password-protected logins. In addition, CAI's systems use role-based security models that ensure users access only the data to which they are entitled and may edit data according to their user rights only.

There are three elements related to ensuring the correct students are accessing appropriate test content:

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

FERPA prohibits the public disclosure of student information or test results. Examples of prohibited practices include

- providing login information (username and password) to other authorized TIDE users or to unauthorized individuals;
- sending a student's name and EDUID number together in an email message; and
- having students log in and test under another student's EDUID number.

Except for authorized individuals with an appropriate need-to-know status, test materials and score reports that identify student names with test scores are not exposed.

All students, including homeschooled students, must be enrolled or registered at their testing schools in order to take the online or paper-pencil assessments. Student enrollment information, including demographic data, is uploaded by the LEA into TIDE during the school year. DA, DCs, and SCs should update and verify the accuracy of student information within TIDE throughout the school year.

Students log in to the online assessment using their legal first name, EDUID number, and a test session ID. Only students can log in to an online test session. TEs, TAs, and 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. For the paper-pencil versions of the assessments, DAs, DCs, SCs, TEs, or TAs are required to enter student responses into the DEI in order to receive student scores.

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

2.4.2 System Security

The objective of system security is to ensure that data are protected and accessed by authorized user groups. System security focuses on protecting data along with preserving the intended integrity of both data and systems. This includes ensuring that personal information is secured, data transfers (whether sent or received) remain unaltered, the data source is known, any service or activity can only be performed by a specific, designated user.

A hierarchy of control: As described in Section 2.2, DAs, DCs, SCs, TAs, and TEs have well-defined roles and access to the testing system. Districts are responsible for adding users and ensuring access to the CAI systems through TIDE. DAs must contact the Department to be assigned that role in TIDE. DAs are then responsible for selecting and entering DC and SC information into TIDE, and DCs and SCs are responsible for entering TA and TE information into TIDE. Throughout the year, the DAs, DCs, and SCs are also expected to delete user information in TIDE for staff members who have transferred to other schools, resigned, or no longer serve as TAs or TEs.

Password protection: Passwords are required of all users and roles when logging in to access any CAI system. Newly added users receive passwords through their personal email addresses or school-assigned email addresses.

Secure Browser: A key role of the Technology Coordinator is to ensure that the CAI Secure Browser is properly installed on the computers used for administration of the online assessments. Developed by CAI, the Secure Browser prevents students from accessing other computers, opening Internet applications, and copying test information. The Secure Browser suppresses access to commonly used browsers such as Internet Explorer, Chrome, and Firefox. It also prevents students from searching for answers on the Internet or communicating with other students. The summative assessments can be accessed only through the Secure Browser.

2.4.3 Security of the Testing Environment

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

Testing personnel are reminded in online trainings, face-to-face trainings, and user manuals that assessments should be administered in testing rooms that allow students enough space and avoid crowding. Good lighting, ventilation, and freedom from noise and interruption are important factors to consider when selecting testing rooms.

TEs and TAs must establish procedures to maintain a quiet environment during each test session, recognizing that some students may finish more quickly than others. The Department that all students, including those who finish early, stay in the testing room until the end of the session. They should engage in a quiet, academic activity. Allowable breaks, as outlined in specific Test Administration Manuals, are encouraged. More specific administrative rules which safeguard the integrity of Idaho assessment in the public school (08.02.03.111.), and provide well-defined policies, procedures, and guidance to support assessment safeguards can be found in the *Assessment Integrity Guide*.

If a student needs to leave the room for a brief time, TEs or TAs are required to pause the student's assessment. For the CAT, if the pause lasts longer than 20 minutes, the student can continue with the rest of the assessment in a new test session, but the system will not allow the student to return to the items presented before the pause. This measure is implemented to prevent students from looking up or verifying answers in between testing sessions.

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 help answer test items should be removed or covered. This rule applies to rubrics, vocabulary charts, student work, posters, graphs, content-area strategy charts, etc. The cell phones of both testing personnel and students must be turned off and stored out of sight in the testing room. TEs and TAs are encouraged to minimize access to the testing rooms by posting signs in halls and entrances to promote optimum testing conditions; it is also recommended that a "TESTING—DO NOT DISTURB" sign is posted on testing room doors.

Seating Arrangements

Because the online CAT is adaptive, it is unlikely a student will see the same test items as other students. For the PTs, different forms are spiraled within a classroom so that students receive different forms of the PT. However, TEs and TAs should provide adequate spacing between students' seats so students will not be tempted to look at the answers of others and to discourage students from communicating.

After the Test

At the end of a test session, TEs or TAs must walk through the classroom to pick up any scratch paper and any papers displaying students' personal information, including EDUID numbers and names. These materials should immediately be securely shredded or stored in a locked area. Regardless of assessment or content area, all printed reading passages and items provided as accommodations for individual students and settings must be shredded immediately after the test session ends.

For the paper-pencil versions, the *Paper-Pencil Test Administration Manual* provides specific instructions for how to securely package and return materials to the testing contractor's office once student responses have been entered into the DEI site.

2.4.4 Test Security Violations

Everyone involved in test administration, including proctoring the assessments, is responsible for understanding test security procedures. Prohibited practices found in the *Assessment Integrity Guide* and the *ISAT Test Administration Manual* fall into one of three categories:

1. **Incident**: An unusual circumstance 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. These circumstances can be corrected and contained at the local level.
2. **Impropriety**: An unusual circumstance that impacts an individual or group of students who are testing and may potentially affect student performance on the test, test security, or test validity. These circumstances can be corrected and contained at the local level.
3. **Breach (Test Security Violation)**: An event that poses a threat to the validity of the test. Examples may include such situations as a release of secure materials or a security/system risk. These circumstances have external implications for all states using the same items and may result in a decision to remove the test item(s) from the available secure bank, at cost to the State.

District and school personnel must document all test security incidents in the test security incident log on the SDE website (https://apps.sde.idaho.gov/testincidentlog). This log is the record for all test security incidents and should be maintained at the district level and submitted to SDE as incidents occur throughout testing.

2.5 STUDENT PARTICIPATION

All students (including retained students) currently enrolled in grades 3–8 and 11 at public schools in Idaho are required to participate in the ISAT ELA/L and mathematics assessments. Students take the assessment(s) corresponding to their grade-level enrollment. Students in grades 9 and 10 are ineligible for the summative assessments but may take interim assessments at the school/LEA's discretion. Students in grades 9 and 10 may take the grade 11 ISAT ELA/L and mathematics assessments if their teacher believes they are capable of doing so, and if they have already received instruction on all standards in the subject area.

2.5.1 Homeschooled Students

While not required, students who are homeschooled may participate in the ISAT ELA/L and mathematics assessments at the request of their parent or guardian and at the discretion of the LEA.

2.5.2 Exempt Students

The following students are exempt from participating in the ISAT ELA/L and mathematics assessments:

- Foreign exchange students who are enrolled in a U.S. school.
- English learners (ELs) who enrolled in a U.S. school within the last 12 months before the beginning of testing have a one-time exemption; these students may instead participate in the English language proficiency assessment consistent with state and federal policies. ELs are not exempt from completing the mathematics assessment.
- A student with significant cognitive disabilities who meets the criteria for a state-selected or state-developed ELA/L and mathematics alternative assessment based on alternative achievement standards (approximately 1% or fewer of the student population). The Idaho Alternate Assessments (IDAA) are available for students meeting these criteria. Students meeting these criteria are not exempt from testing; they are exempt only from completing the ISAT ELA/L and mathematics assessments.

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

2.6 Online Testing Features and Testing Accommodations

The Smarter Balanced Assessment Consortium's *Usability, Accessibility, and Accommodations Guidelines* (UAAG) focus on the accessibility features (i.e., universal tools, designated supports, and accommodations) available to students for the ISAT assessments, including ELA/L and mathematics. The UAAG 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 ISAT assessments. The UAAG are also intended for assessment staff and administrators who oversee decisions regarding instruction and assessment.

The UAAG apply to all students. They emphasize an individualized approach to the implementation of accessibility features and assessment practices to meet the diverse needs of students participating in large-scale content assessments.

The UAAG can be used by classroom teachers, English language development educators, special education teachers, and instructional assistants when selecting and administering accessibility features, including universal tools, designated supports, and accommodations, that are appropriate for individual students.

The summative assessments contain both embedded and non-embedded accessibility features. Embedded resources are part of the CAI administration system, whereas non-embedded resources are provided outside of that system.

The Department, DAs, DCs, and SCs can set embedded and non-embedded designated supports and accommodations based on their specific user roles. Designated supports and accommodations must be set in TIDE before starting a test session.

Each embedded and non-embedded universal tool is available to all students during a test session. A TE or TA can deactivate any of the preselected universal tools in the TA Interface of the testing system for a student. Deactivating a universal tool may be appropriate if a student could be distracted by access to that tool during a test session.

For additional information about the availability of accessibility features, including universal tools, designated supports and accommodations, refer to the *Usability, Accessibility, and Accommodations Guidelines (UAAG)* at https://idaho.portal.cambiumast.com/resource-item/en/usability-accessibility-and-accommodations-guidelines-sy23-24.

2.6.1 Online Universal Tools for All Students

Universal tools are a type of accessibility feature provided on the ISAT assessments. They can be an embedded or non-embedded component 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 2023–2024 test administration, the following universal tools were available for all students. For specific information on how to access and use these features, refer to the *Usability, Accessibility, and Accommodations Guidelines* at https://idaho.portal.cambiumast.com/resource-item/en/usability-accessibility-and-accommodations-guidelines-sy23-24.

Embedded Universal Tools

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

Calculator (for calculator-allowed mathematics items only in grades 6–8, 11): Students can access an embedded on-screen digital calculator for calculator-allowed items by clicking the Calculator button. This tool is only available with the specific items that the Smarter Balanced Item Specifications indicate are 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 a 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 on-screen English dictionary is available for the full-write portion of an ELA/L performance task.

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

Expandable Passages and/or Items: Each passage/stimulus and/or associated item can be expanded so that it takes up a larger portion of the screen, requiring less scrolling by the student.

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

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

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

Line Reader: This tool is used to highlight an individual line of text in a passage or test item.

Mark for Review: The student can mark an item for review in order to return to it later. However, for the CAT, if the assessment is paused for more than 20 minutes, students will not be allowed to return to marked test items.

Math Tools: These digital tools (e.g., embedded ruler or protractor) are used for measurements and are available only with the mathematics items for which the Smarter Balanced Item Specifications deem them to be appropriate.

Spellcheck: This tool is used to check 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 for which the Smarter Balanced Item Specifications indicated that it would be appropriate. Spellcheck is bundled with other embedded writing tools for all full-write portions of a performance task (planning, drafting, revising, and editing). A full-write is the second part of a PT.

Strikethrough: This tool allows students to cross out response options.

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

Writing Tools: Selected writing tools (e.g., bold, italics, bullets, and undo/redo) are available for all student-generated responses.

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

Non-Embedded Universal Tools

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

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

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

2.6.2 Designated Supports and Accommodations

Designated Supports

Designated supports for the ISAT ELA/L and mathematics assessments are accessibility features that are available for use by any student for whom the need has been indicated by an educator (or team of educators) with the parent/guardian and student. Scores achieved by students using designated supports will be included for federal accountability purposes. It is recommended that a consistent process is used to determine appropriate designated supports for individual students. To aid their decisions, all educators should be trained on the process and understand the range of designated supports available.

Embedded Designated Supports

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

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

Language/Presentation (for mathematics items): Dual language translations are a linguistic support available for some students. They provide a full translation of each English test item and stimulus.

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

Mouse Pointer: This embedded support allows the mouse pointer to be set to a larger size and for the color to be changed to help students find the mouse pointer more readily.

Streamlined Interface Mode: This accommodation provides an alternative, simplified format of the testing interface in which the items are displayed below the stimuli.

Text-to-Speech (for mathematics stimuli and items and ELA/L items; not for ELA/L reading passages): Text is read aloud to the student via embedded TTS technology in English for both ELA/L and math. Text-to-Speech is also provided in Spanish for math. The student can control the speed, pause the voice, and raise or lower the volume of the voice via a volume control.

Translated Test Directions (for mathematics items): Translation of test directions is a language support available before beginning the actual test items. As an embedded designated support, translated test directions are automatically a part of the dual language translations designated support.

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

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

Non-Embedded Designated Supports

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

Bilingual Dictionary: A bilingual/dual language word-to-word dictionary is a language support and can be provided for the full-write portion of an ELA/L PT.

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

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

Illustration Glossaries (for mathematics items on the paper-pencil tests): The illustration glossaries are a language support provided for selected construct-irrelevant terms for math. Illustrations for these terms appear in a supplement to the paper-pencil test and are identified by item number.

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

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

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

Paper-and Pencil Assessment: A paper-based version of the ISAT assessment may be made available to students as an alternative to the computerbased assessment.

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

Read Aloud (for mathematics stimuli and items and ELA/L items; not for ELA/L reading passages): Text is read aloud to the student by a trained and qualified human reader who follows the administration guidelines provided in the *ISAT Online Summative Test Administration Manual* and *Read-Aloud Guidelines*. All or portions of the content may be read aloud. LEAs and teachers can refer to the *Guidelines for Choosing the Read-Aloud Accommodation* when deciding if this accommodation is appropriate for a student.

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

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

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

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

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

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

Translations (Glossaries) (for mathematics items on the 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.

Accommodations

Accommodations are changes in procedures or materials that increase equitable access during the ISAT ELA/L and mathematics assessments. Assessment accommodations generate valid assessment results for students who need them allowingthese students to show what they know and can do. Accommodations are available for students with documented IEPs or Section 504 Plans. Consortium-approved accommodations do not compromise the learning expectations, construct, grade-level standard, or intended outcome of the assessments.

Embedded Accommodations

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

Braille: This is a raised-dot code that individuals read with their fingertips. Graphic material (e.g., maps, charts, graphs, diagrams, illustrations) is presented in a raised format (paper or thermoform). The following codes are available for the ELA/L: Unified English Braille (UEB) uncontracted and UEB contracted. The following codes are available for the mathematics paper-pencil assessment: UEB uncontracted with Nemeth, UEB contracted with Nemeth, UEB uncontracted with UEB mathematics, and UEB contracted with UEB mathematics.

Braille Transcript (for ELA/L listening items and mathematics items): This is a braille transcript of the closed captioning created for the listening passages.

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

Speech-to-Text: Voice recognition allows students to use their voices as devices to input information into 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.

Text-to-Speech (for ELA/L reading passages): Text is read aloud to the student via embedded TTS 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: This is a paper-based table listing numbers from 1–100 available from Smarter Balanced for reference.

Abacus: For students who typically use an abacus, this tool may be used in place of scratch paper.

Alternate Response Options: Alternate response options include but are not limited to adapted keyboards, large keyboards, StickyKeys, 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: Unified English Braille (UEB) uncontracted and UEB contracted. The following codes are available for the mathematics paper-pencil assessment: UEB uncontracted with Nemeth, UEB contracted with Nemeth, UEB uncontracted with UEB mathematics, and UEB contracted with UEB mathematics.

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

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

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

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

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

Speech-to-Text: Voice recognition allows students to use their voices as devices to input information into 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 may be used, as well as speech output built into the program that reads back the information the student has written. 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 4 presents a list of universal tools, designated supports, and accommodations that were offered in the 2023–2024 administration. Tables 5–10 provide the number of students who utilized any of the offered accommodations and designated supports.

Table 4. SY 2023–2024 Universal Tools, Designated Supports, and Accommodations

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

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

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

[2] For full-write portion of ELA/L performance tasks

[3] For ELA/L performance tasks

[4] Includes embedded ruler, embedded protractor

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

[6] For mathematics items

[7] For mathematics stimuli and items and ELA/L items (not for ELA/L reading passages). Must be set in TIDE by district- or school-level user and must be set before the test begins. Also available in Spanish for mathematics tests.

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

[9] For ELA/L listening items

[10] For ELA/L reading passages, all grades. Must be set in TIDE by district- or school-level user and must be set before the test begins.

[11] For mathematics paper-pencil tests

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

[13] For mathematics tests, all grades

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

[15] Includes adapted keyboards, large keyboards, Sticky Keys, Mouse Keys, Filter Keys, adapted mouse, touchscreen, head wand, and switches.

[16] For paper-pencil assessments

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

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

Accommodations	Grade						
	3	4	5	6	7	8	11
Embedded Accommodations							
American Sign Language	10	9	6	13	5	2	6
Braille	0	1	0	0	0	0	0
Braille Transcript	0	1	0	0	1	0	0
Closed Captioning	27	30	30	57	39	29	32
Speech-to-Text	416	528	657	614	572	510	176
Text-to-Speech: Reading Passages and Items	1,137	1,295	1,459	1,230	1,080	1,076	568
Non-Embedded Accommodations							
Alternate Response Options	21	23	14	10	45	27	6
Print-on-Demand	14	9	5	5	39	22	3
Read Aloud: Passages	318	316	336	256	199	192	64
Scribe (Writing)	0	0	0	0	0	68	26
Speech-to-Text	286	289	397	315	270	268	92
Word Prediction	45	67	77	58	72	73	25

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

Designated Supports	Subgroup	Grade						
		3	4	5	6	7	8	11
Color Contrast	Overall	14	11	12	13	16	24	24
	EL	4	0	1	2	3	7	1
	Special Education	8	5	6	8	8	10	15
Masking	Overall	142	159	186	241	196	158	145
	EL	34	32	23	74	92	51	39
	Special Education	96	114	149	138	114	100	99
Mouse Pointer	Overall	17	22	17	7	11	7	2
	EL	1	3	3	0	1	1	0
	Special Education	13	21	15	5	9	6	2
Streamlined Interface Mode	Overall	20	10	32	45	59	66	54
	EL	0	0	5	6	15	10	9
	Special Education	17	8	28	33	51	57	35
Text-to-Speech: CAT Items	Overall	2,058	2,050	1,887	1,543	1,495	1,426	827
	EL	712	711	549	484	473	461	277
	Special Education	625	797	762	734	767	731	443
Text-to-Speech: PT Stimuli and Items	Overall	3,193	3,310	3,303	2,690	2,541	2,483	1,286
	EL	831	842	715	637	634	620	321
	Special Education	1,708	1,999	2,106	1,811	1,727	1,682	893

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

Designated Supports	**Subgroup**	**Grade**						
		3	**4**	**5**	**6**	**7**	**8**	**11**
Amplification	Overall	10	4	12	4	10	11	9
	EL	3	0	2	3	1	3	2
	Special Education	6	2	4	1	4	7	3
Bilingual Dictionary	Overall	101	93	95	83	100	75	92
	EL	98	91	94	80	93	70	86
	Special Education	12	6	19	12	14	11	15
Color Contrast	Overall	4	6	8	7	34	24	4
	EL	0	1	1	2	10	3	1
	Special Education	3	4	5	3	34	18	3
Color Overlay	Overall	4	12	10	11	41	22	5
	EL	0	3	1	0	10	3	0
	Special Education	2	6	6	6	38	18	3
Magnification	Overall	8	10	12	12	42	25	10
	EL	0	2	4	5	13	5	1
	Special Education	5	5	8	8	37	19	6
Medical Device	Overall	6	10	28	17	47	31	14
	EL	0	1	1	2	10	4	0
	Special Education	0	2	7	2	32	17	1
Noise Buffers	Overall	119	135	138	149	105	118	64
	EL	15	13	17	17	19	10	11
	Special Education	89	111	109	111	82	68	45
Printed Test Directions in English	Overall	1	2	2	4	31	16	2
	EL	0	0	0	0	10	4	0
	Special Education	0	1	1	3	30	15	0
Read Aloud: Items	Overall	335	329	276	235	238	263	138
	EL	75	77	21	44	49	41	23
	Special Education	225	217	214	189	191	194	111
Scribe (Non-Writing)	Overall	100	101	93	80	84	58	15
	EL	11	10	10	9	18	6	0
	Special Education	87	87	78	70	76	49	12
Separate Setting	Overall	2,175	2,462	2,524	2,261	2,155	2,068	1,508
	EL	306	349	344	339	366	329	234
	Special Education	1,561	1,800	1,837	1,612	1,541	1,482	1,055
Simplified Test Directions	Overall	926	1,137	1,020	923	858	779	515
	EL	290	375	265	303	301	256	173
	Special Education	580	727	730	670	586	532	347
Translated Test Directions	Overall	52	73	42	82	77	69	57
	EL	48	62	36	79	71	59	53
	Special Education	7	9	5	7	3	2	7
Translated Test Directions in ASL	Overall	3	2	1	6	1	2	3
	EL	0	0	0	1	1	1	1
	Special Education	3	2	1	6	1	1	3

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

Accommodations	**Grade**						
	3	**4**	**5**	**6**	**7**	**8**	**11**
Embedded Accommodations							
American Sign Language	12	9	6	13	5	2	6
Braille	0	1	0	1	0	0	0
Speech-to-Text	405	488	604	566	535	472	148
Non-Embedded Accommodations							
100s Number Table	677	706	676	487	319	252	67
Abacus	16	16	22	10	36	18	5
Alternate Response Options	18	19	14	9	42	25	7
Calculator	90	134	260	445	622	702	500
Multiplication Table	573	1,079	1,311	1,194	1,207	1,207	371
Print-on-Demand	12	9	5	5	42	25	4
Speech-to-Text	250	235	314	249	215	225	76
Word Prediction	45	56	61	41	55	54	11

Table 9. Mathematics Total Students with Allowed Embedded Designated Supports

Designated Supports	**Subgroup**	**Grade**						
		3	**4**	**5**	**6**	**7**	**8**	**11**
Color Contrast	Overall	14	9	12	12	18	26	24
	EL	4	0	1	2	4	7	1
	Special Education	8	4	6	8	9	10	15
Illustration Glossaries	Overall	287	295	230	212	209	190	85
	EL	276	284	217	201	192	182	64
	Special Education	20	41	25	29	36	29	27
Language/Presentation: Spanish	Overall	182	225	194	195	189	208	90
	EL	174	210	180	186	183	199	81
	Special Education	5	6	6	4	3	3	2
Masking	Overall	145	161	187	246	217	176	144
	EL	34	35	27	82	112	68	40
	Special Education	99	115	146	136	115	102	98
Mouse Pointer	Overall	18	21	16	7	11	7	2
	EL	1	3	2	0	1	1	0
	Special Education	14	20	14	5	9	6	2
Streamlined Interface Mode	Overall	18	11	28	46	59	61	43
	EL	0	0	5	6	16	8	2
	Special Education	16	9	25	34	51	53	26
Text-to-Speech: Stimuli and Items	Overall	3,543	3,560	3,552	2,919	2,680	2,610	1,364
	EL	985	1,020	874	771	747	727	366
	Special Education	1,749	2,051	2,170	1,879	1,744	1,680	910
Translation (Glossary): Spanish	Overall	279	358	318	288	278	243	153
	EL	271	344	301	269	266	237	135
	Special Education	23	41	47	23	25	24	29
Translation (Glossary): Other Languages	Overall	25	26	23	27	31	22	6
	EL	25	24	23	27	27	22	6
	Special Education	1	2	0	1	3	0	1

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

Designated Supports	**Subgroup**	**Grade**						
		3	**4**	**5**	**6**	**7**	**8**	**11**
Amplification	Overall	9	5	11	5	10	11	9
	EL	2	0	2	3	1	3	2
	Special Education	6	3	4	2	4	6	3
Color Contrast	Overall	4	6	8	6	34	23	5
	EL	0	1	1	1	10	3	1
	Special Education	3	4	5	2	34	18	4
Color Overlay	Overall	4	11	10	12	38	22	5
	EL	0	3	1	0	10	3	0
	Special Education	2	5	6	6	35	19	3
Illustration Glossaries	Overall	13	27	2	13	36	23	34
	EL	12	25	2	12	15	10	28
	Special Education	4	4	0	5	31	16	10
Magnification	Overall	8	10	12	13	41	28	12
	EL	0	2	4	5	12	6	1
	Special Education	5	6	8	9	36	21	7
Medical Device	Overall	5	9	26	14	47	32	13
	EL	0	1	1	2	10	3	0
	Special Education	0	1	7	2	32	18	1
Noise Buffers	Overall	119	130	134	142	105	117	64
	EL	19	13	18	18	20	10	11
	Special Education	87	105	105	104	80	68	45
Printed Test Directions in English	Overall	0	1	3	2	30	16	2
	EL	0	0	0	0	10	3	0
	Special Education	0	1	2	1	30	15	0
Read Aloud: Stimuli and Items	Overall	364	362	321	293	259	282	152
	EL	76	81	33	51	55	45	31
	Special Education	248	245	248	237	207	208	115
Read Aloud in Spanish: Stimuli and Items	Overall	51	65	30	21	21	13	11
	EL	43	60	27	20	18	11	11
	Special Education	3	5	3	2	1	3	1
Scribe	Overall	123	134	117	91	93	70	20
	EL	14	9	9	9	18	7	0
	Special Education	106	116	102	78	84	61	14
Separate Setting	Overall	2,190	2,445	2,543	2,284	2,199	2,110	1,498
	EL	349	380	387	375	401	369	239
	Special Education	1,522	1,748	1,813	1,593	1,552	1,483	1,049
Simplified Test Directions	Overall	955	1197	1099	978	919	831	521
	EL	339	450	341	368	360	308	181
	Special Education	559	712	731	652	586	531	348
Translated Test Directions	Overall	110	142	112	150	136	128	53
	EL	102	129	103	145	126	116	49
	Special Education	8	10	8	6	5	4	5

Designated Supports	**Subgroup**	**Grade**						
		3	**4**	**5**	**6**	**7**	**8**	**11**
Translated Test Directions in ASL	Overall	3	2	1	5	1	3	3
	EL	0	0	0	1	1	2	1
	Special Education	3	2	1	5	1	1	3
Translation (Glossary): Spanish	Overall	88	100	92	89	72	41	63
	EL	86	93	90	82	67	39	60
	Special Education	10	7	17	10	9	5	9
Translation (Glossary): Other Languages	Overall	1	7	1	5	3	2	8
	EL	1	6	1	5	1	2	8
	Special Education	0	2	0	0	0	0	1

2.7 TESTING TIME

The online testing system captures item response time by calculating, in milliseconds, the amount of time spent on an item page. Items can appear on a page in one of two ways: for discrete items, each item appears on the screen/page one item at a time, whereas stimulus-based items appear on the screen/page together. Item page time is calculated as: the time spent on one item for discrete items and the time spent on all items associated with a stimulus for stimulus-based items. For each student, the total time taken to complete the test is computed by adding up the page time for all items and item groups (stimulus-based items).

The ISAT assessments are not timed, and an individual student may need more or less time overall. The length of a test session is determined by an LEA's or school's testing schedule. Testing schedules are typically developed by SCs, TEs, and/or TAs who are knowledgeable about the school's instructional schedule and the timing needed for each ISAT assessment. Students should be allowed extra time as needed, and TEs or TAs should use their best professional judgment when allowing students extra time.

Tables 11 and 12 present the average testing time and the testing time by percentile for the overall test, the CAT component, and the PT component.

Table 11. ELA/L Testing Times

Grade	Average Testing Time (hh:mm)	SD of Testing Time (hh:mm)	Median Testing Time (hh:mm)	Testing Time in Percentiles (hh:mm)				
				75th	80th	85th	90th	95th
Overall Test								
3	2:41	1:45	2:17	3:22	3:43	4:10	4:50	6:02
4	2:53	1:45	2:31	3:37	3:57	4:23	5:02	6:10
5	2:53	1:47	2:31	3:35	3:55	4:22	5:00	6:11
6	2:45	1:31	2:27	3:23	3:40	4:03	4:36	5:34
7	2:29	1:23	2:14	3:01	3:15	3:34	4:02	4:55
8	2:20	1:18	2:06	2:53	3:08	3:28	3:57	4:44
11	1:43	0:59	1:36	2:10	2:21	2:33	2:50	3:22
CAT Component								
3	0:57	0:35	0:49	1:08	1:14	1:23	1:34	1:57
4	0:59	0:33	0:52	1:11	1:17	1:25	1:38	1:59
5	1:00	0:32	0:54	1:14	1:19	1:27	1:38	1:58
6	1:11	0:34	1:05	1:26	1:32	1:40	1:52	2:13
7	1:02	0:30	0:58	1:16	1:22	1:28	1:38	1:55
8	0:58	0:28	0:54	1:11	1:16	1:22	1:31	1:46
11	0:48	0:24	0:46	1:00	1:04	1:09	1:16	1:28
PT Component								
3	1:44	1:23	1:24	2:15	2:32	2:54	3:27	4:25
4	1:54	1:24	1:35	2:28	2:43	3:04	3:36	4:30
5	1:53	1:26	1:33	2:24	2:40	3:02	3:33	4:31
6	1:34	1:07	1:19	2:00	2:13	2:30	2:54	3:41
7	1:26	1:03	1:13	1:47	1:58	2:13	2:33	3:14
8	1:22	0:58	1:10	1:44	1:55	2:10	2:32	3:10
11	0:55	0:41	0:48	1:13	1:19	1:28	1:41	2:05

Table 12. Mathematics Testing Times

Grade	Average Testing Time (hh:mm)	SD of Testing Time (hh:mm)	Median Testing Time (hh:mm)	Testing Time in Percentiles (hh:mm)				
				75th	80th	85th	90th	95th
Overall Test								
3	1:24	0:55	1:10	1:44	1:55	2:09	2:30	3:09
4	1:29	0:55	1:15	1:49	2:00	2:15	2:37	3:15
5	1:40	1:05	1:25	2:05	2:17	2:32	2:55	3:38
6	1:31	0:51	1:20	1:51	2:00	2:13	2:31	3:02
7	1:09	0:38	1:01	1:23	1:30	1:40	1:52	2:16
8	1:11	0:39	1:04	1:27	1:34	1:44	1:58	2:21
11	0:53	0:29	0:49	1:07	1:12	1:19	1:28	1:44
CAT Component								
3	0:46	0:31	0:37	0:56	1:02	1:11	1:23	1:43
4	0:49	0:32	0:41	1:00	1:06	1:14	1:26	1:47
5	0:50	0:33	0:43	1:02	1:08	1:17	1:28	1:49
6	0:46	0:26	0:40	0:56	1:01	1:08	1:17	1:32
7	0:42	0:23	0:38	0:52	0:56	1:01	1:09	1:24
8	0:43	0:23	0:38	0:52	0:57	1:03	1:11	1:25
11	0:30	0:16	0:27	0:37	0:40	0:44	0:49	0:57
PT Component								
3	0:38	0:30	0:30	0:49	0:55	1:02	1:14	1:36
4	0:40	0:30	0:32	0:50	0:56	1:04	1:17	1:39
5	0:50	0:41	0:40	1:03	1:11	1:21	1:35	2:02
6	0:45	0:32	0:37	0:56	1:01	1:09	1:21	1:41
7	0:26	0:20	0:22	0:33	0:37	0:41	0:48	1:01
8	0:28	0:21	0:24	0:36	0:40	0:44	0:52	1:05
11	0:23	0:17	0:20	0:31	0:34	0:38	0:44	0:54

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 any inferences based on those test scores. Multiple facets ensure proper test administration, including clear policies, effective TA training, and tools to identify possible testing incidents, including improprieties or breaches.

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

Analyses are performed at the student level and summarized for each aggregate unit, including testing session, TA, and school. Flagging criteria used for these analyses are described below and are configurable by an authorized user. When the aggregate unit size is small, the aggregate unit is flagged if the percentage of flagged students is greater than 50% in the analysis. The default small aggregate unit size is 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 review and ensure the test integrity after the testing window closes.

2.8.1 Changes in Student Performance

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

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

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

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

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

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

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

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

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

The 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

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

The person-fit index is computed using a standardized log-likelihood statistic. Following Drasgow, Levine, and Williams (1985) and Sotaridona, Pornel, and Vallejo (2003), an aberrant response pattern is defined as a deviation from the expected item score model. Snijders (2001) showed that the distribution of l_z is asymptotically normal (i.e., with an increasing number of administered items). Even at shorter test lengths of 8 or 15 items, the "asymptotic error probabilities are quite reasonable for nominal Type I error probabilities of 0.10 and 0.05" (Snijders, 2001).

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

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

where *s* is the standard deviation of $l_z$values in an aggregate unit and *n* is the number of students in an aggregate unit. The QA report includes a list of the flagged aggregate units.

2.8.4 Item Response Change

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

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

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

2.9 Prevention and Recovery of Disruptions in the Test Delivery System

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

CAI has developed a unique monitoring system that is extremely sensitive to changes in server performance. Most monitoring systems provide warnings when something is going wrong. The CAI system does, too, but it also provides warnings when any given server performs differently from its performance over the few hours prior or differently than the other servers performing the same jobs. Subtle changes in performance often precede actual failure by hours or days, allowing CAI to detect 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 system architecture and how it recovers from device failures, Internet interruptions, and other problems.

2.9.1 High-Level System Architecture

CAI's architecture provides the redundancy, robustness, and reliability required for 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.

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

The following sections describe the key elements of CAI's testing system, including the data integrity processes, fault tolerance, and automated recovery built into each component of the system.

Student Machine

Student responses are conveyed to CAI's servers in real time. 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 for confirmation of successfully stored data from the server. If confirmation is not received within the designated time (usually set to 30–90 seconds), the system will prevent the student from doing any more work until connectivity is restored. The student is offered the choice of asking the system to try again or pausing the test and returning at a later time. For example:

- If connectivity is lost and restored within the designated time period, the student may be unaware of the momentary interruption.
- If connectivity cannot be silently restored, the student is prevented from testing and given the option of logging out or retrying the save.
- If the system fails completely, upon logging back into the system, the student returns to the item at which the failure occurred.

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

Test Delivery Satellites

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

One server for every four satellites serves as a backup hub. This server continually monitors and stores all changed student response data from the satellites, creating an additional copy of the real-time data. In the unlikely event of failure, data are completely protected. Satellites are automatically monitored, and upon failure, they are removed from service. Real-time student data are immediately recoverable from the satellite, backup hub, or hub (described in the following section), 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. This process is managed by the hub. Data will remain on the satellites until the satellite receives notice from the demographic and history servers that the data are safely stored on those disks.

Hub

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

Demographic and History Servers

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

Quality Assurance System

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

Database of Record

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

2.9.2 Automated Backup and Recovery

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

2.9.3 Other Disruption Prevention and System Recovery Measures

CAI's testing systems are designed to be extremely fault tolerant and can withstand failure of any component with little or no service interruption. This robustness is achieved through redundancy. Key redundant systems are as follows:

- The system's hosting provider has redundant power generators that can continue to operate for up to 60 hours without refueling. With the multiple refueling contracts that are in place, these generators can operate indefinitely.
- The hosting provider has multiple redundancies in the flow of information to and from the system's data centers by partnering with nine different network providers. Each fiber carrier must enter the data center at separate physical points, protecting the data center from a complete service failure caused by an unlikely network cable cut.
- There are redundant firewalls and load balancers throughout the environment on the network level.
- The system uses redundant power and switching within all server cabinets.

- Data are protected by both a full weekly backup and incremental nightly backups. Should a catastrophic event occur, CAI is able to reconstruct real-time data using the data retained on the TDS satellites and hubs.
- The server backup agents send alerts to notify system administration staff in the event of a backup error, at which time they will inspect the error to determine whether the backup was successful or if they need to rerun it.

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

3. SUMMARY OF 2023–2024 OPERATIONAL TEST ADMINISTRATION

3.1 STUDENT POPULATION

All students enrolled in grades 3–8 and 11 in all public elementary and secondary schools must participate in the Idaho Standards Achievement Test (ISAT) English language arts/literacy (ELA/L) and mathematics assessments.

Before the testing window opens, the Idaho Department of Education (the Department) or LEAs send Cambium Assessment, Inc. (CAI) a student enrollment file to load to the Test Information Distribution Engine (TIDE). Using this enrollment file, the participation rates are calculated as the percentage of students who attempted the test. Tables 13 and 14 present the participation rates for the ELA/L and mathematics ISATs by subgroup. Tables 15 and 16 present the demographic composition of Idaho students who meet attemptedness requirements for scoring and reporting the results of the summative assessments.

Table 13. Participation Rates by Percentage for the ISAT ELA/L Summative Assessment

Group	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 11
All Students	97.5	97.6	97.4	97.3	96.5	96.2	91.2
Female	97.8	97.7	97.7	97.6	96.5	96.1	90.5
Male	97.2	97.4	97.1	97.1	96.4	96.2	91.8
African American	92.5	92.3	92.1	93.3	92.8	92.0	88.6
AI/AN	93.2	96.3	94.9	95.5	95.7	94.7	89.9
Asian	95.4	95.0	94.3	93.9	96.7	95.7	91.1
Hispanic	95.8	96.2	96.0	96.1	95.8	95.3	92.8
Pacific Islander	97.0	99.1	96.0	98.2	97.2	98.0	90.3
White	98.3	98.3	98.1	98.0	96.9	96.6	90.9
EL	90.7	90.6	91.2	92.0	91.5	92.0	91.9
Special Education	91.9	93.4	92.4	92.4	90.6	90.8	85.8
Section 504	98.8	98.4	99.2	97.9	97.4	97.0	91.8

Legend. African American = Black or African American; AI/AN = American Indian or Alaska Native; Pacific Islander = Native Hawaiian or Other Pacific Islander; EL = English learners

Table 14. Participation Rates by Percentage for the ISAT Mathematics Summative Assessment

Group	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 11
All Students	98.1	98.3	97.9	97.9	96.9	96.6	91.5
Female	98.5	98.4	98.1	98.1	96.9	96.5	91.1
Male	97.8	98.2	97.7	97.7	96.9	96.6	91.9
African American	97.5	98.7	96.9	96.3	96.4	96.7	89.8
AI/AN	93.7	96.3	95.7	96.0	95.7	93.1	92.3
Asian	97.3	96.9	97.5	96.1	98.5	97.5	92.4
Hispanic	98.1	98.5	98.1	98.1	97.3	97.1	93.7
Pacific Islander	97.0	98.6	96.0	98.2	97.7	98.5	90.8
White	98.3	98.3	98.0	97.9	96.9	96.4	91.0
EL	97.9	98.4	98.2	98.3	97.0	97.7	94.5
Special Education	92.0	93.3	92.2	92.4	90.9	90.8	85.7
Section 504	99.6	98.9	99.1	97.8	97.4	96.6	93.4

Table 15. Number of Students for the ISAT ELA/L Summative Assessment

Group	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 11
All Students	23,374	23,631	23,742	23,513	23,766	23,923	22,710
Female	11,507	11,477	11,701	11,436	11,710	11,623	11,052
Male	11,867	12,154	12,041	12,077	12,056	12,300	11,658
African American	261	276	269	251	281	279	300
AI/AN	207	238	241	191	243	232	204
Asian	251	249	298	263	261	266	283
Hispanic	4,464	4,564	4,426	4,435	4,649	4,570	4,376
Pacific Islander	261	215	193	213	208	194	174
White	17,666	17,979	18,229	18,077	18,040	18,305	17,320
EL	2,006	2,092	2,125	2,140	2,199	2,231	1,832
Special Education	2,912	3,077	2,998	2,697	2,623	2,542	1,946
Section 504	719	903	1,080	1,259	1,375	1,423	1,457

Table 16. Number of Students for the ISAT Mathematics Summative Assessment

Group	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 11
All Students	23,524	23,806	23,864	23,631	23,859	24,013	23,022
Female	11,591	11,555	11,748	11,490	11,748	11,665	11,231
Male	11,933	12,251	12,116	12,141	12,111	12,348	11,791
African American	274	295	285	259	293	293	301
AI/AN	208	238	242	192	243	228	204
Asian	255	253	308	269	265	271	287
Hispanic	4,566	4,676	4,520	4,526	4,714	4,652	4,409
Pacific Islander	261	214	193	213	209	195	174
White	17,671	17,982	18,200	18,059	18,022	18,270	17,587
EL	2,159	2,271	2,288	2,288	2,330	2,371	1,884
Special Education	2,913	3,084	2,999	2,689	2,622	2,534	1,942
Section 504	731	908	1,084	1,264	1,378	1,419	1,483

3.2 SUMMARY OF OVERALL STUDENT PERFORMANCE

Tables 17–22 present a summary of the 2023–2024 summative test results for all students and by subgroup, including the average and the standard deviation of scale scores, the percentage of students in each achievement level, and the percentage of proficient students.

Figures 1 and 2 present the percentage of proficient students over the past five years for all students (cohort comparisons) in grades 3–8 and 11. Figures 3 and 4 present the average scale scores over the past five years for all students in grades 3–8 and 11. For grade 11, student performance prior to the 2022-2023 school year was not included because the 2022–2023 school year was the first year of administering grade 11 tests as the accountability grade in high school. In Figures 1–4, the 2019–2020 performance is not included because testing was canceled due to the COVID-19 pandemic. Appendix B 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 17. Descriptive Statistics and Percentage of Students in Achievement Levels for Overall and by Subgroup: ELA/L (Grades 3–5)

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
Grade 3								
All Students	23,374	2421.51	97.51	29	23	22	25	48
Female	11,507	2427.68	96.46	26	23	23	27	50
Male	11,867	2415.52	98.16	31	24	22	24	45
African American	261	2376.99	98.49	48	19	18	15	33
AI/AN	207	2369.38	90.11	52	24	12	12	24
Asian	251	2454.00	98.52	18	19	27	36	63
Hispanic	4,464	2385.66	93.75	43	25	17	14	32
Pacific Islander	261	2415.39	93.43	26	30	25	19	44
White	17,666	2431.80	95.94	25	23	24	28	52
EL	2,006	2361.21	90.84	54	23	14	9	22
Special Education	2,912	2342.96	91.93	63	20	10	7	17
Section 504	719	2410.62	95.43	32	25	22	21	43
Grade 4								
All Students	23,631	2465.51	102.71	31	20	23	27	49
Female	11,477	2472.81	101.11	28	20	23	29	52
Male	12,154	2458.61	103.73	33	20	22	25	47
African American	276	2414.89	101.41	53	19	13	15	28
AI/AN	238	2419.31	93.75	52	21	14	13	27
Asian	249	2506.65	97.74	15	19	23	43	66
Hispanic	4,564	2426.59	98.11	45	21	19	14	33
Pacific Islander	215	2460.61	104.29	34	17	27	23	49
White	17,979	2476.54	101.03	26	20	24	30	54
EL	2,092	2400.57	97.35	56	20	16	8	24
Special Education	3,077	2373.12	98.21	69	15	10	6	16
Section 504	903	2461.96	94.25	31	24	23	22	45
Grade 5								
All Students	23,742	2505.19	106.65	27	20	29	25	53
Female	11,701	2513.82	105.24	24	19	29	27	56
Male	12,041	2496.80	107.34	30	20	28	22	50
African American	269	2441.43	104.23	47	23	21	9	30
AI/AN	241	2452.14	103.29	47	22	19	12	31
Asian	298	2537.03	121.09	21	10	32	37	69
Hispanic	4,426	2460.18	100.14	43	22	24	11	35
Pacific Islander	193	2492.08	112.30	32	18	28	22	50
White	18,229	2517.51	104.40	23	19	30	28	58
EL	2,125	2436.51	102.41	53	21	18	8	26
Special Education	2,998	2399.45	95.87	69	17	10	4	14
Section 504	1,080	2496.85	98.38	28	24	29	19	49

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

Table 18. Descriptive Statistics and Percentage of Students in Achievement Levels for Overall and by Subgroup: ELA/L (Grades 6–8)

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
				Grade 6				
All Students	23,513	2529.11	100.41	24	24	32	20	52
Female	11,436	2540.91	98.23	20	23	34	23	57
Male	12,077	2517.93	101.17	28	25	30	17	47
African American	251	2477.04	99.55	46	21	25	8	33
AI/AN	191	2474.26	98.20	43	24	26	6	32
Asian	263	2572.28	104.21	13	19	31	37	68
Hispanic	4,435	2486.42	97.45	38	29	24	9	33
Pacific Islander	213	2538.05	95.37	20	26	32	21	54
White	18,077	2540.36	97.75	20	23	34	22	57
EL	2,140	2467.74	100.18	48	26	19	7	26
Special Education	2,697	2416.60	86.49	71	19	8	2	10
Section 504	1,259	2516.13	92.53	26	30	30	14	44
				Grade 7				
All Students	23,766	2556.99	107.76	23	22	35	20	56
Female	11,710	2571.22	104.04	18	21	37	23	61
Male	12,056	2543.18	109.50	27	22	34	17	51
African American	281	2499.60	117.09	41	22	30	7	36
AI/AN	243	2508.80	104.80	39	25	27	9	36
Asian	261	2596.89	114.14	15	15	38	33	70
Hispanic	4,649	2511.91	108.25	37	26	27	10	37
Pacific Islander	208	2550.67	108.38	25	25	32	18	50
White	18,040	2569.77	103.69	18	21	38	23	61
EL	2,199	2485.02	111.03	48	24	21	7	28
Special Education	2,623	2437.40	98.32	67	21	11	2	12
Section 504	1,375	2544.13	97.13	25	27	34	13	47
				Grade 8				
All Students	23,923	2564.33	109.78	23	24	35	17	53
Female	11,623	2580.34	105.63	18	23	38	20	59
Male	12,300	2549.19	111.46	28	26	33	14	47
African American	279	2505.32	123.65	43	26	22	10	32
AI/AN	232	2520.87	109.12	37	31	23	9	32
Asian	266	2602.40	120.28	16	15	35	34	69
Hispanic	4,570	2521.48	107.19	36	28	28	8	36
Pacific Islander	194	2571.78	99.52	16	28	41	14	55
White	18,305	2576.04	106.88	20	23	38	19	57
EL	2,231	2498.54	114.29	45	26	22	7	29
Special Education	2,542	2435.78	98.02	72	20	7	1	8
Section 504	1,423	2553.75	97.75	25	29	34	12	46

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

Table 19. Descriptive Statistics and Percentage of Students in Achievement Levels for Overall and by Subgroup: 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	22,710	2598.36	122.68	20	21	32	27	59
Female	11,052	2616.68	115.23	15	20	34	31	65
Male	11,658	2580.99	126.93	25	22	30	23	53
African American	300	2514.31	130.92	45	20	25	10	35
AI/AN	204	2547.35	115.76	31	26	31	11	43
Asian	283	2636.13	135.81	16	13	31	40	71
Hispanic	4,376	2553.48	117.62	30	27	28	14	43
Pacific Islander	174	2592.37	118.23	21	26	28	25	52
White	17,320	2611.32	120.19	17	20	33	30	63
EL	1,832	2519.34	123.39	42	27	22	10	32
Special Education	1,946	2460.15	105.27	64	23	11	2	13
Section 504	1,457	2588.14	116.80	20	25	33	22	55

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

Table 20. Descriptive Statistics and Percentage of Students in Achievement Levels for Overall and by Subgroup: Mathematics (Grades 3–5)

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
Grade 3								
All Students	23,524	2430.14	90.28	28	23	28	21	50
Female	11,591	2423.83	87.28	30	24	28	18	46
Male	11,933	2436.27	92.68	26	21	29	24	53
African American	274	2367.08	102.51	53	17	24	7	30
AI/AN	208	2381.92	85.95	52	21	17	10	27
Asian	255	2460.90	100.46	20	16	27	38	65
Hispanic	4,566	2393.26	87.14	44	24	21	11	32
Pacific Islander	261	2423.79	85.85	27	30	25	19	43
White	17,671	2441.29	87.55	23	22	31	24	55
EL	2,159	2372.62	87.84	55	22	15	8	23
Special Education	2,913	2354.85	96.39	62	18	13	7	20
Section 504	731	2424.79	85.50	31	23	26	20	46
Grade 4								
All Students	23,806	2475.85	91.38	24	28	26	22	48
Female	11,555	2469.50	86.91	25	31	26	19	44
Male	12,251	2481.84	95.03	22	26	26	25	51
African American	295	2420.90	94.55	47	28	15	9	24
AI/AN	238	2430.11	81.37	43	32	15	9	24
Asian	253	2519.06	103.58	16	17	26	41	68
Hispanic	4,676	2435.13	87.43	40	32	19	9	29
Pacific Islander	214	2466.83	92.53	28	32	19	21	40
White	17,982	2487.94	88.38	19	28	28	25	53
EL	2,271	2416.39	86.97	50	30	14	6	20
Special Education	3,084	2393.93	92.87	60	23	11	6	16
Section 504	908	2474.92	82.68	22	33	25	19	44
Grade 5								
All Students	23,864	2499.64	101.28	32	26	18	23	41
Female	11,748	2493.62	97.61	34	28	18	20	38
Male	12,116	2505.47	104.38	30	25	19	26	44
African American	285	2423.87	110.10	65	20	7	8	15
AI/AN	242	2445.35	93.80	56	27	9	8	17
Asian	308	2534.83	120.86	26	16	19	39	58
Hispanic	4,520	2454.31	93.88	51	27	12	10	22
Pacific Islander	193	2490.25	98.02	36	26	18	20	38
White	18,200	2512.64	98.46	27	27	20	26	47
EL	2,288	2434.26	96.90	60	23	9	7	17
Special Education	2,999	2401.40	96.18	74	16	6	5	10
Section 504	1,084	2493.28	92.06	35	29	18	18	36

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

Table 21. 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	23,631	2516.23	111.95	33	27	20	20	40
Female	11,490	2512.78	108.95	34	28	19	19	38
Male	12,141	2519.49	114.63	32	27	20	22	42
African American	259	2438.81	124.14	59	24	8	9	17
AI/AN	192	2450.68	102.84	58	28	7	7	15
Asian	269	2578.22	123.42	20	21	16	43	59
Hispanic	4,526	2461.48	109.16	53	26	13	8	21
Pacific Islander	213	2522.94	103.57	30	29	21	20	41
White	18,059	2531.29	106.84	27	28	22	24	45
EL	2,288	2441.44	115.34	61	23	9	7	16
Special Education	2,689	2393.37	108.42	79	14	4	3	8
Section 504	1,264	2507.75	100.14	34	31	19	16	34
Grade 7								
All Students	23,859	2537.46	115.31	31	27	22	20	42
Female	11,748	2531.96	113.34	32	28	22	18	40
Male	12,111	2542.80	116.96	29	26	23	22	45
African American	293	2459.84	121.79	58	23	12	8	19
AI/AN	243	2478.42	112.99	52	26	14	9	22
Asian	265	2588.63	135.04	22	15	22	40	63
Hispanic	4,714	2482.67	112.80	50	26	15	9	23
Pacific Islander	209	2532.41	117.31	33	28	21	19	40
White	18,022	2553.68	109.92	25	27	25	23	48
EL	2,330	2458.11	115.94	60	23	11	6	17
Special Education	2,622	2412.88	106.29	77	15	5	3	8
Section 504	1,378	2527.80	102.16	32	34	21	14	34
Grade 8								
All Students	24,013	2549.76	128.08	36	25	18	21	39
Female	11,665	2547.85	123.30	35	26	19	20	38
Male	12,348	2551.56	132.43	36	24	18	23	40
African American	293	2468.40	127.58	61	19	13	6	20
AI/AN	228	2493.07	125.04	57	22	12	10	21
Asian	271	2618.84	155.47	24	15	18	44	62
Hispanic	4,652	2490.31	117.46	55	24	12	8	21
Pacific Islander	195	2546.34	118.47	36	30	16	18	34
White	18,270	2566.44	124.72	30	25	20	25	45
EL	2,371	2468.95	125.10	65	19	9	8	17
Special Education	2,534	2407.37	109.00	83	12	3	2	5
Section 504	1,419	2535.00	110.71	40	30	16	15	31

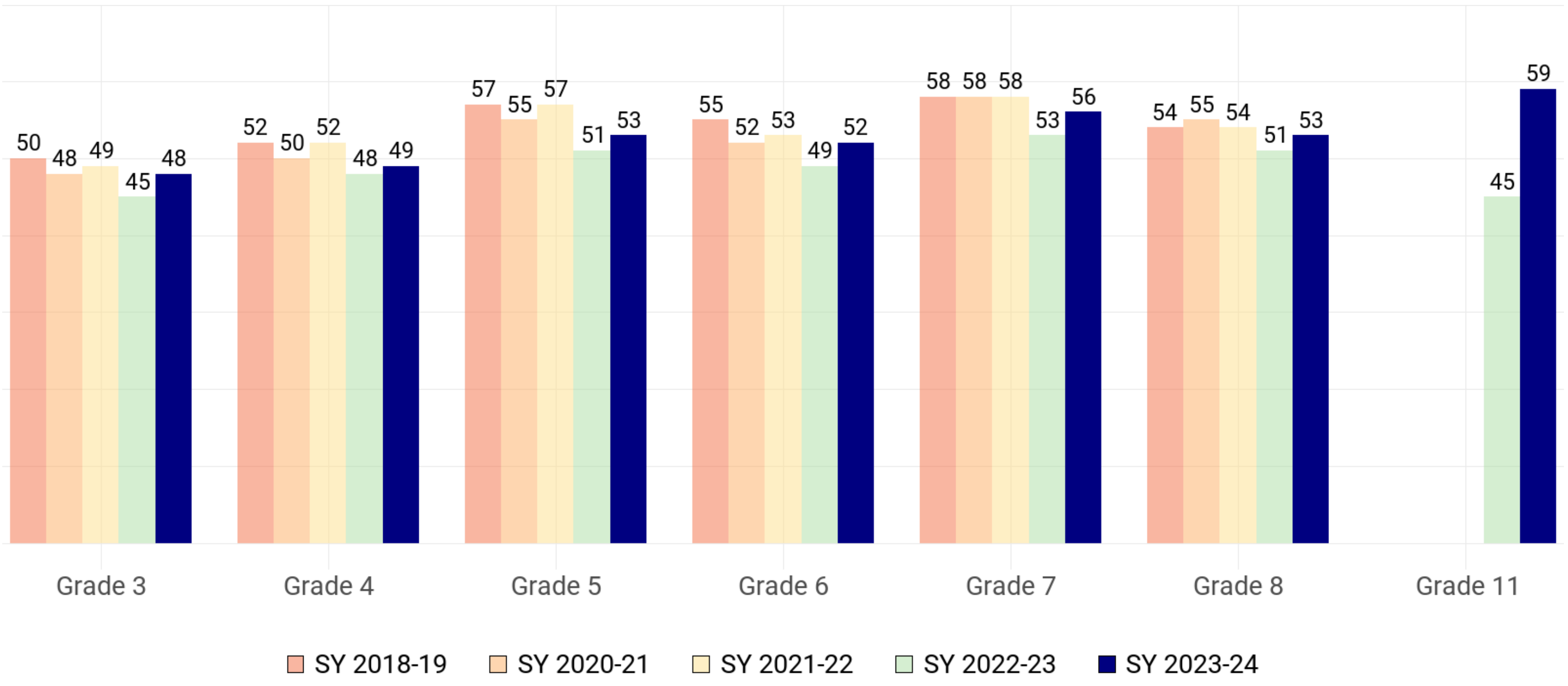
Note. The percentage of each achievement level may not add up to 100% due to rounding.

Table 22. 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	23,022	2562.86	131.45	44	24	19	12	31
Female	11,231	2560.38	121.71	44	27	19	10	29
Male	11,791	2565.24	140.07	44	23	20	14	33
African American	301	2471.49	128.02	72	17	8	2	11
AI/AN	204	2484.91	121.73	69	19	9	3	12
Asian	287	2631.31	148.59	27	22	22	30	51
Hispanic	4,409	2505.65	115.96	65	21	10	4	14
Pacific Islander	174	2544.02	141.14	48	22	18	12	30
White	17,587	2578.93	129.64	38	26	22	14	36
EL	1,884	2485.17	120.00	72	16	8	3	11
Special Education	1,942	2419.23	105.19	90	7	2	0	2
Section 504	1,483	2545.47	123.14	51	26	15	9	23

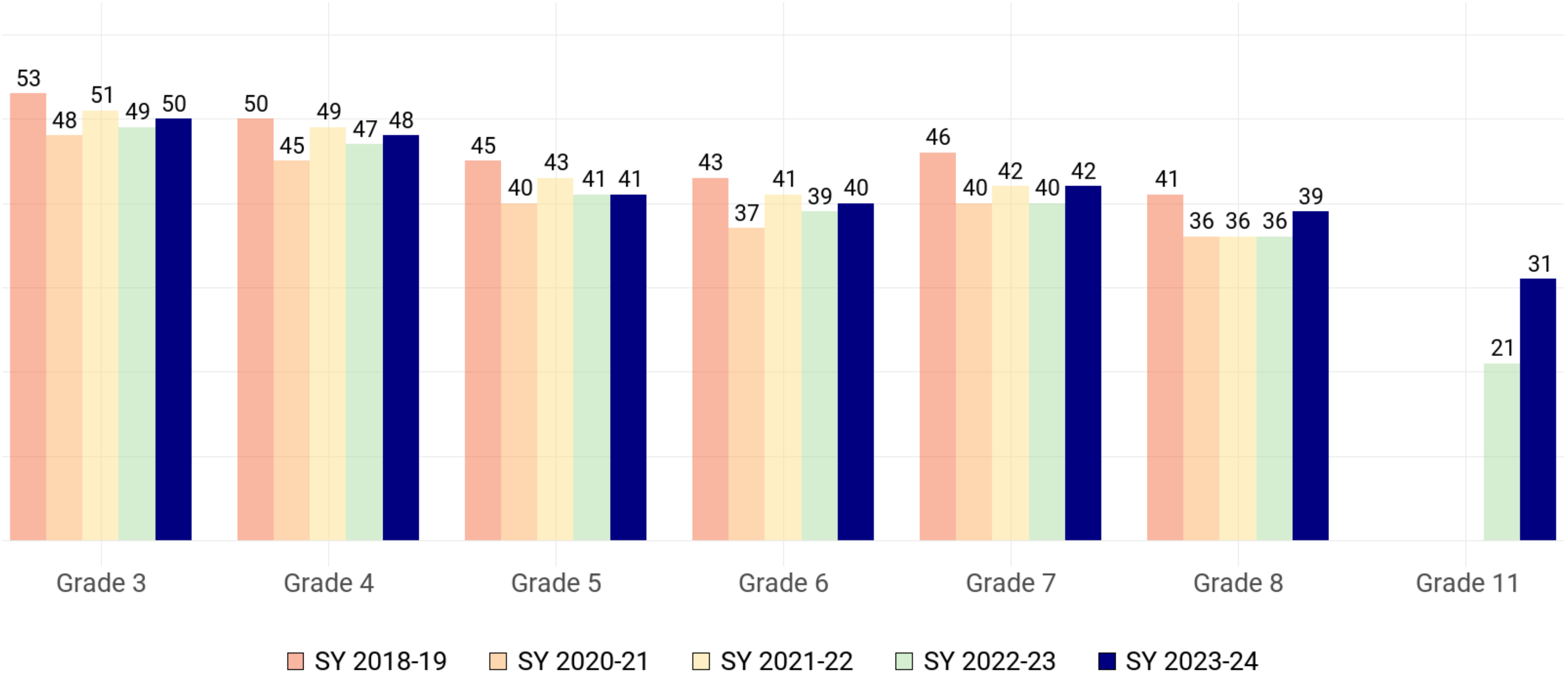
Note. The percentage of each achievement level may not add up to 100% due to rounding.

Figure 1. ELA/L Percent Proficient Across Years



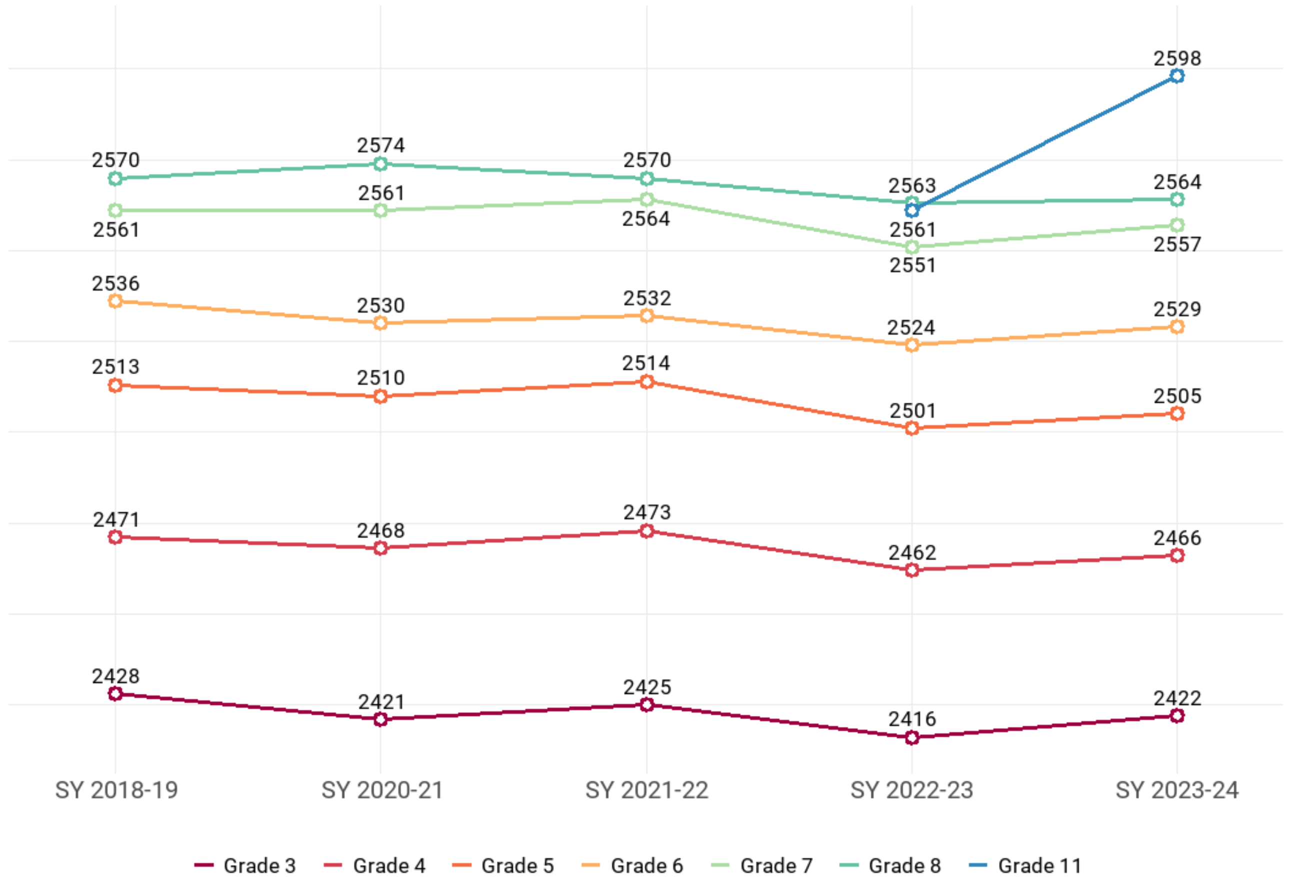
Note. For grade 11, student performance prior to SY 2022-2023 was not included because the 2022–2023 school year was the first year of administering grade 11 tests as the accountability grade in high school.

Figure 2. Mathematics Percent Proficient Across Years



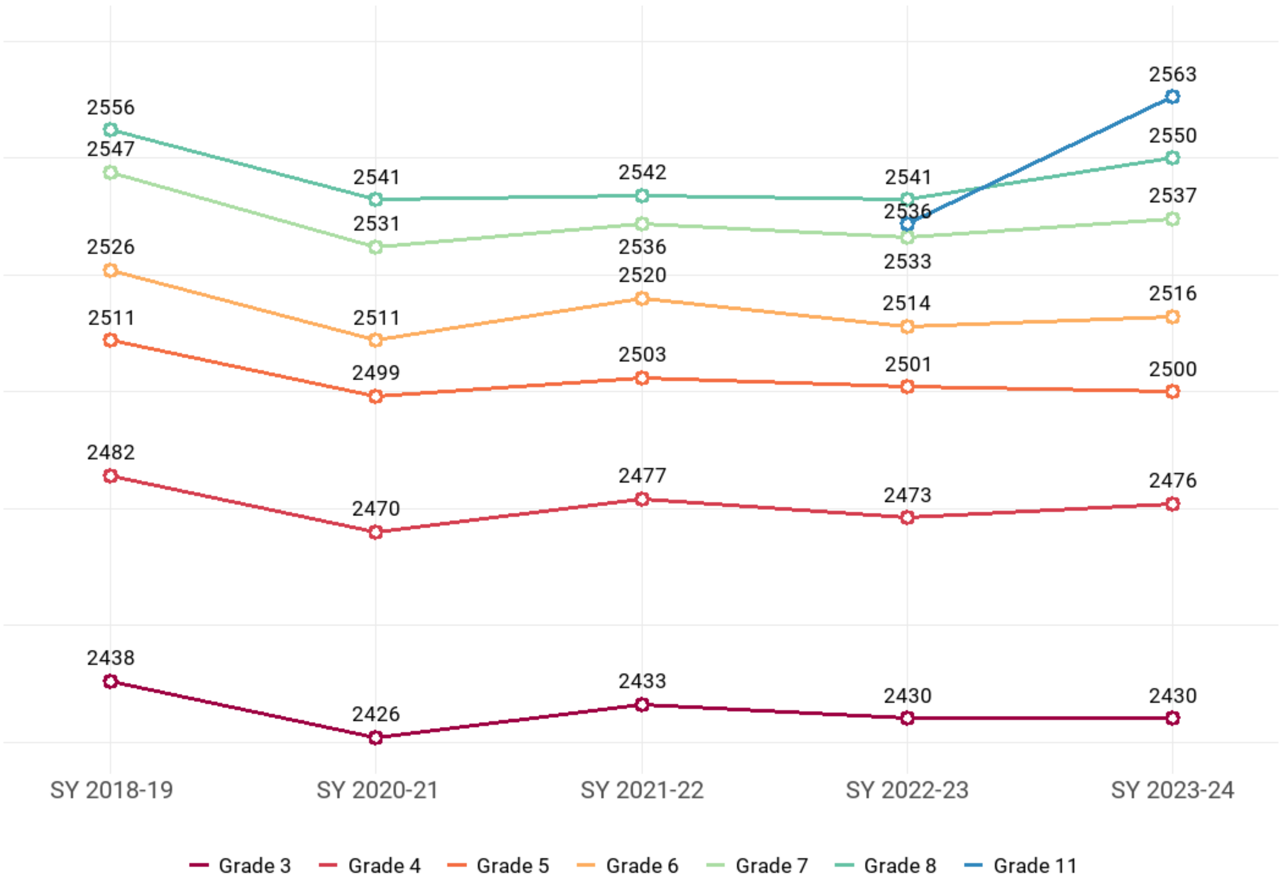
Note. For grade 11, student performance prior to SY 2022-2023 was not included because the 2022–2023 school year was the first year of administering grade 11 tests as the accountability grade in high school.

Figure 3. ELA/L Average Scale Score Across Years



Note. For grade 11, student performance prior to SY 2022-2023 was not included because the 2022–2023 school year was the first year of administering grade 11 tests as the accountability grade in high school.

Figure 4. Mathematics Average Scale Score Across Years



Note. For grade 11, student performance prior to SY 2022-2023 was not included because the 2022–2023 school year was the first year of administering grade 11 tests as the accountability grade in high school.

3.3 Distribution of Student Ability and Item Difficulty

Figures 5–10 show the empirical distribution of the Idaho student scale scores in the 2023–2024 test administration and the distribution of the administered item difficulty parameters for overall and by claim. Overall, the student ability distribution is generally 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, the student ability distribution is generally shifted to the left in claim 4 for all grades in ELA/L. In mathematics, the student ability distribution is shifted to the left for all claims except for claim 1 in all grades. The Smarter Balanced Assessment Consortium plans to add additional easy items to the pool and to augment the pool in proportion to the test blueprint constraints (e.g., content, Depth of Knowledge [DOK], item type, item difficulties) to better measure low-performing students.

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

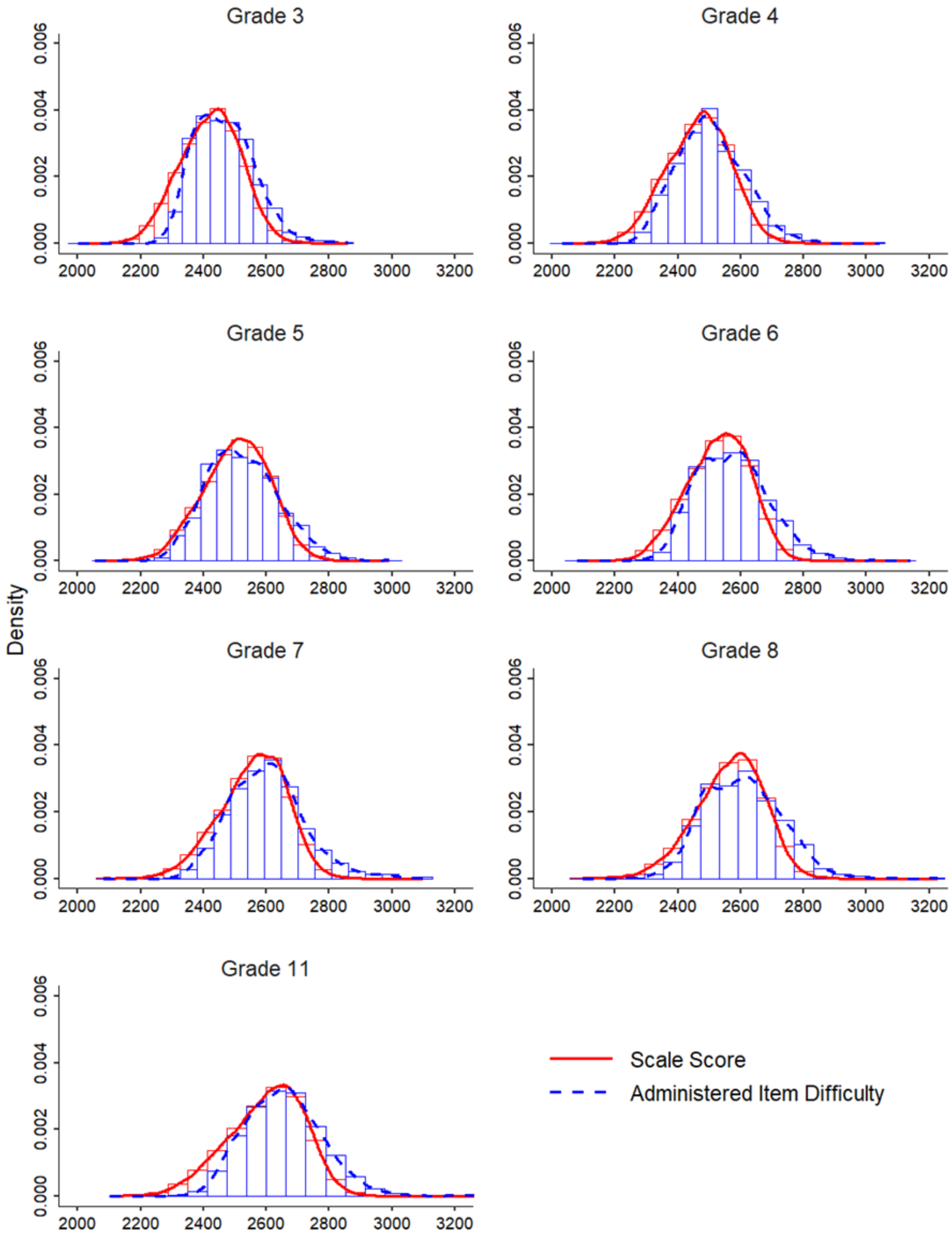


Figure 6. Student Ability–Item Difficulty Distribution by Claim: ELA/L (Grades 3–5)

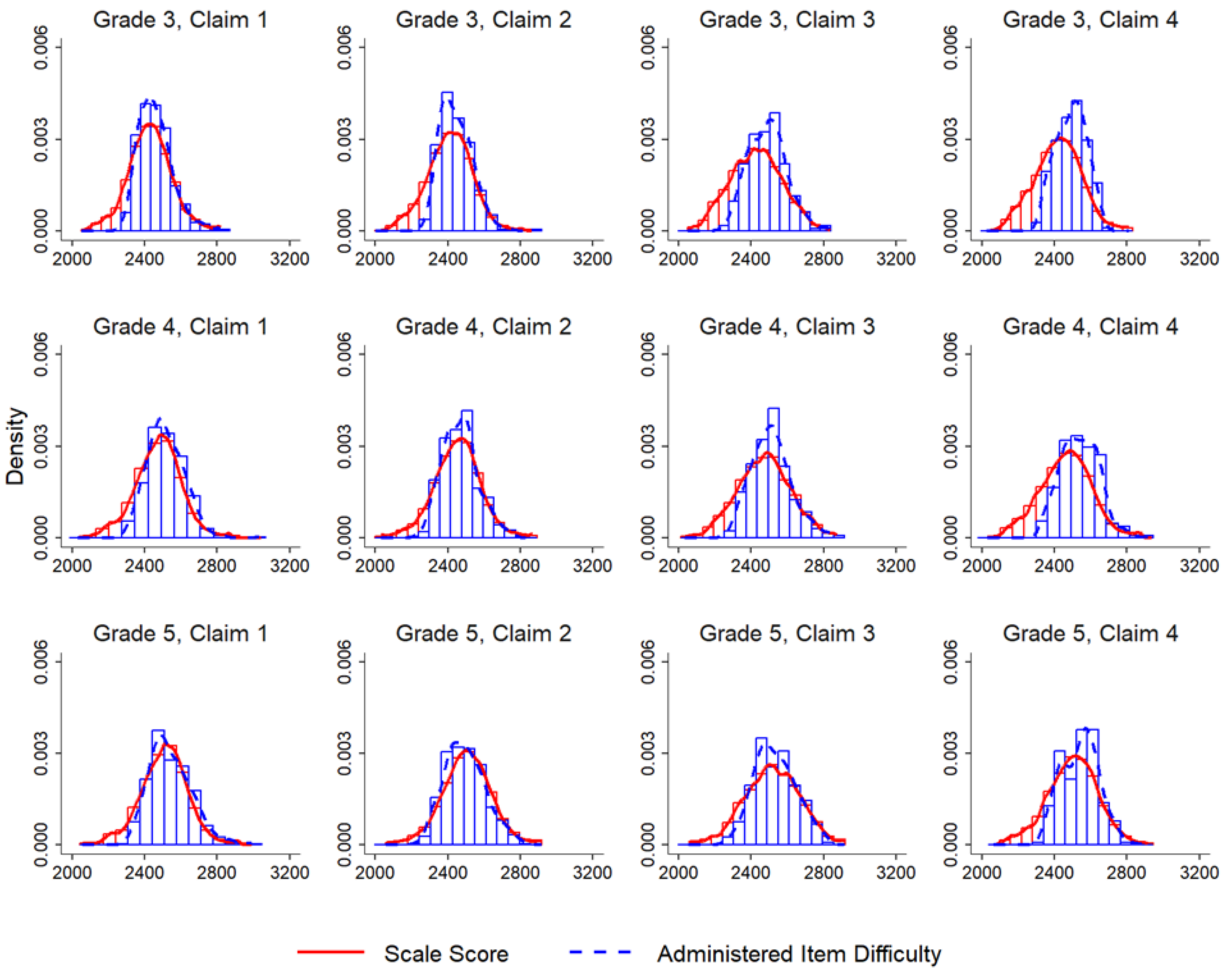


Figure 7. Student Ability–Item Difficulty Distribution by Claim: ELA/L (Grades 6–8, 11)

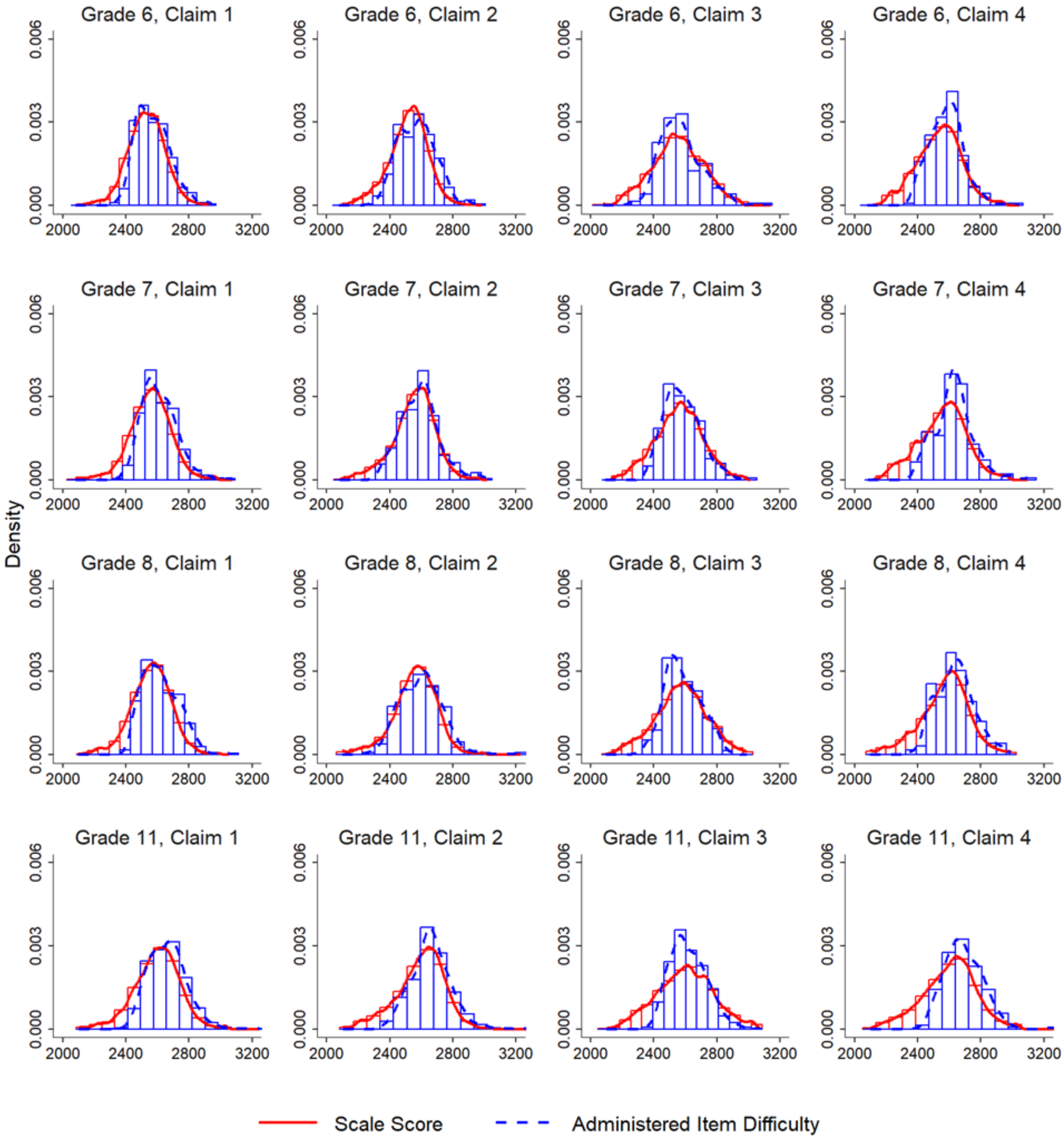


Figure 8. Student Ability–Item Difficulty Distribution for Mathematics

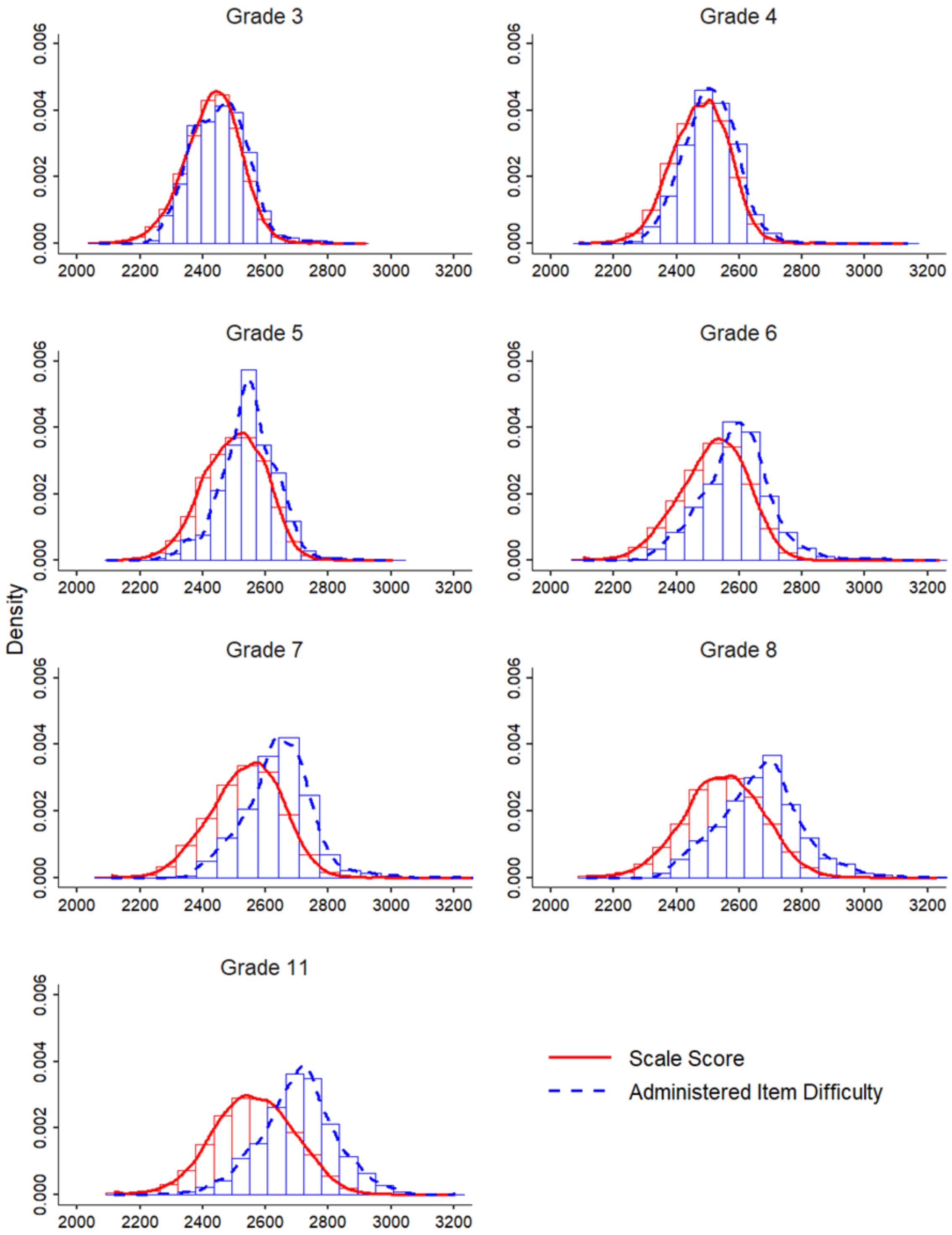


Figure 9. Student Ability–Item Difficulty Distribution by Claim: Mathematics (Grades 3–5)

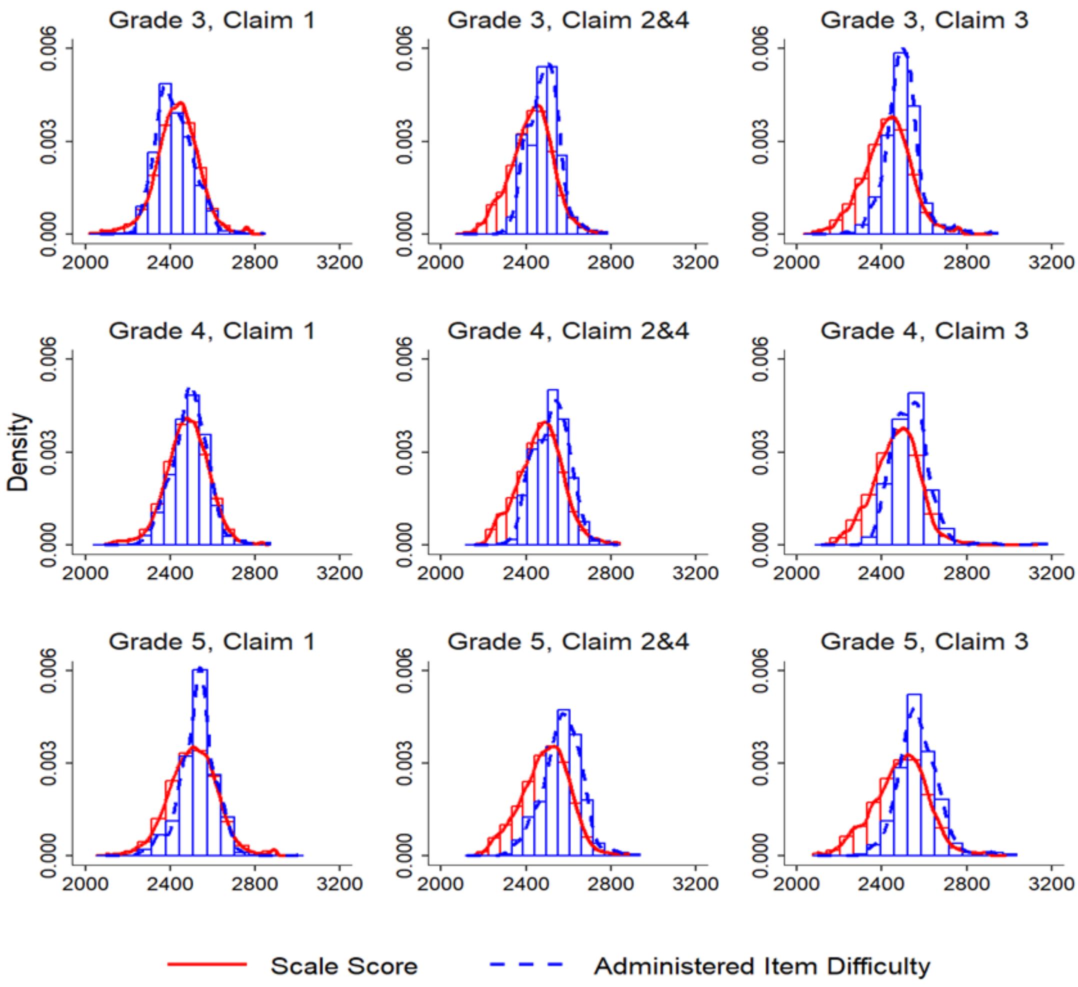
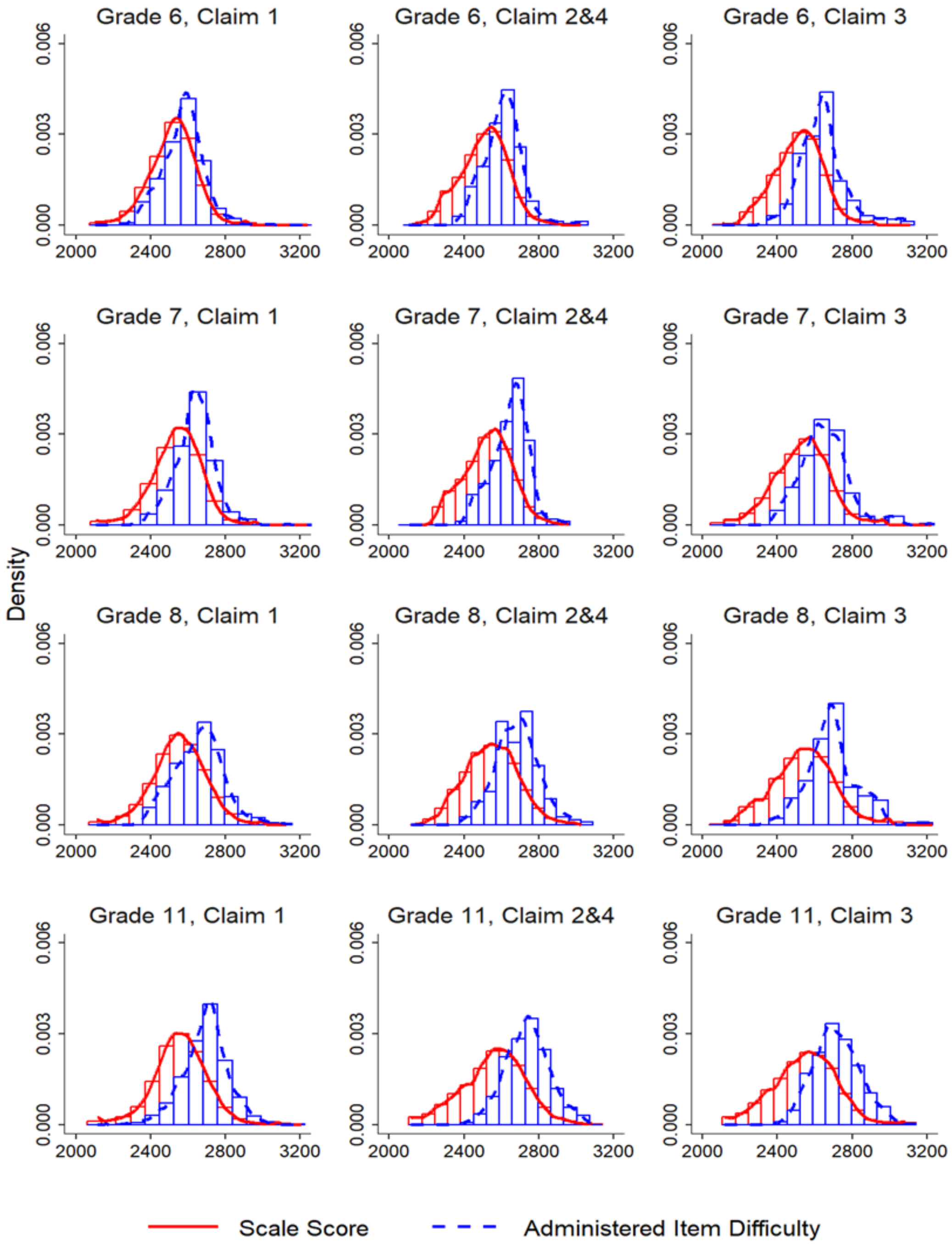


Figure 10. Student Ability–Item Difficulty Distribution by Claim: Mathematics (Grades 6–8, 11)



4. VALIDITY

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

Validity evidence provided in this chapter is as follows:

- Test Content
- Internal Structure

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

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

4.1 EVIDENCE ON TEST CONTENT

The ISAT ELA/L and mathematics 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 while meeting the blueprint specifications. The Smarter Balanced blueprints specify a range of items to be administered in each claim, content domain/standards, and targets. Moreover, blueprints constrain the Depth of Knowledge (DOK) along with item and passage types. For the PT, each student is administered a fixed-form test. The content coverage in all PT forms is the same. The test blueprint constraints for CAT and PT can be found at: https://www.sde.idaho.gov/assessment/isat-cas/.

Tables 23 and 24 present the percentages of tests aligned with the English language arts/literacy (ELA/L) CAT blueprint constraints for items in claims, targets, DOK, and the number of passages requirement. All tests met the blueprint requirements, except for a few tests in grades 5 and 8. Although rare, a few tests administered one item fewer or more than required. These few violations could happen while selecting items that align with the blueprint constraints and adapt to a student's ability. This is primarily due to the uneven distribution of items across targets and DOKs, within and across the passages, and a shortage of easy items. Tables 25–27 provide the percentages of tests aligned with the blueprint constraints for the mathematics CAT for claims, DOK, and target. In mathematics, all tests adhered to the blueprint requirements, except for a few tests in grades 7 and 8 where blueprint violations occurred due to the application of pool filters limiting the item pool. Pool filters—such as using an alternative language like Braille or Spanish, or only items with illustration or language glossaries—can significantly reduce the accommodated CAT item pool. This reduction may prevent the test from meeting all blueprint requirements, especially if multiple pool filters are employed on the same test.

Table 23. Percentage of ELA/L CAT Delivered Tests Meeting Blueprint Requirements for Each Claim and the Number of Passages Administered (Grades 3–5)

Claim	**Content Category/Target**	**Required Items/Passages**	**% BP Match**		
			Grade 3	**Grade 4**	**Grade 5**
1	**Literary Text**	4	100.00	100.00	100.00
	Target 2: Central Ideas Target 4: Reasoning and Evidence	1–3	100.00	100.00	100.00
	Targets 1, 3, 5, 6, and 7	1–3	100.00	100.00	100.00
	Long Literary Text Passage Short Literary Text Passage	1	100.00	100.00	100.00
	Informational Text	4	100.00	100.00	100.00
	Target 9: Central Ideas Target 11: Reasoning and Evidence	1–3	100.00	100.00	100.00
	Targets 8, 10, 12, 13, and 14	1–3	100.00	100.00	100.00
	Long Informational Text Passage Short Informational Text Passage	1	100.00	100.00	100.00
	DOK 2	≥ 4	100.00	100.00	99.98
	DOK 3 or Higher	≥ 1	100.00	100.00	100.00
2	**Writing**	4	100.00	100.00	100.00
	Target 1, 3, or 6: Organization/Purpose	1	100.00	100.00	100.00
	Target 1, 3, or 6: Evidence/Elaboration	1	100.00	100.00	100.00
	Target 8: Language and Vocabulary Use	1	100.00	100.00	100.00
	Target 9: Edit/Clarify	1	100.00	100.00	100.00
	DOK 2	≥ 2	100.00	100.00	100.00
3	**Listening**	4	100.00	100.00	100.00
	Target 4: Listen/Interpret	4	100.00	100.00	100.00
	DOK 2 or Higher	≥ 2	100.00	100.00	100.00
	Listening Passage	2	100.00	100.00	100.00
4	**Research**	4	100.00	100.00	100.00
	Target 2: Interpret and Integrate Information	1–2	100.00	100.00	100.00
	Target 3: Analyze Information/Sources	1–2	100.00	100.00	100.00
	Target 4: Use Evidence	1–2	100.00	100.00	100.00

Table 24. Percentage of ELA/L CAT Delivered Tests Meeting Blueprint Requirements for Each Claim and the Number of Passages Administered (Grades 6–8, 11)

Claim	Content Category/Target	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	100.00	100.00	100.00	100.00
	Target 2: Central Ideas Target 4: Reasoning and Evidence	1–3	1–3	100.00	100.00	100.00	100.00
	Targets 1, 3, 5, 6, and 7	1–3	1–3	100.00	100.00	100.00	100.00
	Target 2 or 4 Short Text	0–1	0–1	100.00	100.00	100.00	100.00
	Long Literary Text Passage	1	1	100.00	100.00	100.00	100.00
	Informational Text	6	6	100.00	100.00	100.00	100.00
	Target 9: Central Ideas Target 11: Reasoning and Evidence	2–4	2–4	100.00	100.00	100.00	100.00
	Targets 8, 10, 12, 13, and 14	2–4	2–4	100.00	100.00	100.00	100.00
	Target 9 or 11 Short Text	0–1	0–1	100.00	100.00	100.00	100.00
	Long Informational Text Passage	1	1	100.00	100.00	100.00	100.00
	Short Informational Text Passage	1	1	100.00	100.00	100.00	100.00
	DOK 1	≤ 3	≤ 2	100.00	100.00	100.00	100.00
	DOK 3 or Higher	≥ 1	≥ 2	100.00	100.00	100.00	100.00
2	**Writing**	4	4	100.00	100.00	100.00	100.00
	Target 1, 3, or 6: Organization/Purpose	1	1	100.00	100.00	100.00	100.00
	Target 1, 3, or 6: Evidence/Elaboration	1	1	100.00	100.00	100.00	100.00
	Target 8: Language and Vocabulary Use	1	1	100.00	100.00	100.00	100.00
	Target 9: Edit/Clarify	1	1	100.00	100.00	100.00	100.00
	DOK 2	≥ 1	≥ 1	100.00	100.00	100.00	100.00
	DOK 3	1	1	100.00	100.00	99.99	100.00
	Brief Write	1	1	100.00	100.00	99.99	100.00
3	**Listening**	4	4	100.00	100.00	100.00	100.00
	Target 4: Listen/Interpret	4	4	100.00	100.00	100.00	100.00
	DOK 2 or Higher	≥ 2	≥ 2	100.00	100.00	100.00	100.00
	Listening Passage	2	2	100.00	100.00	100.00	100.00
4	**Research**	4	4	100.00	100.00	100.00	100.00
	Target 2: Analyze and Integrate Information	1–2	1–2	100.00	100.00	100.00	100.00
	Target 3: Evaluate Information/Sources	1–2	1–2	100.00	100.00	100.00	100.00
	Target 4: Use Evidence	1–2	1–2	100.00	100.00	100.00	100.00

Table 25. Percentage of Mathematics CAT Delivered Tests Meeting Blueprint Requirements for Claims and Targets (Grades 3–5)

Claim	**Content Domain**	**Grade 3**		**Grade 4**		**Grade 5**	
		Required Items	**% BP Match**	**Required Items**	**% BP Match**	**Required Items**	**% BP Match**
1	Overall	10	100.00	10	100.00	10	100.00
	DOK 2 or Higher	≥ 4	100.00	≥ 4	100.00	≥ 4	100.00
	Priority Cluster	7	100.00				
	Targets B, C, G, I	3	100.00				
	Targets D, F	3	100.00				
	Target A	1	100.00				
	Supporting Cluster	3	100.00				
	Targets E, J, K	2	100.00				
	Target H	1	100.00				
	Priority Cluster			7	100.00		
	Targets A, E, F			3	100.00		
	Target G			2	100.00		
	Target D			1	100.00		
	Target H			1	100.00		
	Supporting Cluster			3	100.00		
	Targets I, K			1	100.00		
	Targets B, C, J			1	100.00		
	Target L			1	100.00		
	Priority Cluster					7	100.00
	Targets E, I					3	100.00
	Target F					2	100.00
	Targets C, D					2	100.00
	Supporting Cluster					3	100.00
	Targets J, K					2	100.00
	Targets A, B, G, H					1	100.00
2 and 4	Overall	3	100.00	3	100.00	3	100.00
	DOK 3 or Higher	≥ 1	100.00	≥ 1	100.00	≥ 1	100.00
	2. Target A	0–1	100.00	0–1	100.00	0–1	100.00
	2. Targets B, C, D	0–1	100.00	0–1	100.00	0–1	100.00
	4. Targets A, D	0–1	100.00	0–1	100.00	0–1	100.00
	4. Targets B, E	0–1	100.00	0–1	100.00	0–1	100.00
	4. Targets C, F	0–1	100.00	0–1	100.00	0–1	100.00
3	Overall	4	100.00	4	100.00	4	100.00
	DOK 3 or Higher	≥ 1	100.00	≥ 1	100.00	≥ 1	100.00
	Targets A, D	1–2	100.00	1–2	100.00	1–2	100.00
	Targets B, E	1–2	100.00	1–2	100.00	1–2	100.00
	Targets C, F	1	100.00	1	100.00	1	100.00

Table 26. Percentage of Mathematics CAT Delivered Tests Meeting Blueprint Requirements for Claims and Targets (Grades 6–8)

Claim	Content Domain	Grade 6		Grade 7		Grade 8	
		Required Items	% BP Match	Required Items	% BP Match	Required Items	% BP Match
1	Overall	9–10	100.00	9–10	100.00	9–10	100.00
	DOK 2 or Higher	≥ 4	100.00	≥ 4	100.00	≥ 4	100.00
	Priority Cluster	6–7	100.00				
	Targets E, F	3	100.00				
	Target A	1–2	100.00				
	Targets B, G	1–2	100.00				
	Target D	1	100.00				
	Supporting Cluster	3	100.00				
	Targets C, H, I, J	3	100.00				
	Priority Cluster			7	99.98		
	Targets A, D			4	100.00		
	Targets B, C			3	99.98		
	Supporting Cluster			3	99.98		
	Targets E, F			2	99.97		
	Targets G, H, I			1	99.99		
	Priority Cluster					7	100.00
	Targets C, D					3	99.99
	Targets B, E, G					3	99.99
	Targets F, H					1	100.00
	Supporting Cluster					3	100.00
	Targets A, I, J					3	100.00
2 and 4	Overall	3	100.00	3	100.00	3	100.00
	DOK 3 or Higher	≥ 1	100.00	≥ 1	100.00	≥ 1	100.00
	2. Target A	0–1	100.00	0–1	100.00	0–1	100.00
	2. Targets B, C, D	0–1	100.00	0–1	100.00	0–1	100.00
	4. Targets A, D	0–1	100.00	0–1	100.00	0–1	100.00
	4. Targets B, E	0–1	100.00	0–1	100.00	0–1	100.00
	4. Targets C, F	0–1	100.00	0–1	100.00	0–1	100.00
3	Overall	4	100.00	4	100.00	4	100.00
	DOK 3 or Higher	≥ 1	100.00	≥ 1	100.00	≥ 1	100.00
	Targets A, D	1–2	100.00	1–2	100.00	1–2	100.00
	Targets B, E	1–2	100.00	1–2	100.00	1–2	100.00
	Targets C, F, G	1	100.00	1	99.99	1	100.00

Table 27. Percentage of Mathematics CAT Delivered Tests Meeting Blueprint Requirements for Claims and Targets (Grade 11)

Claim	Content Domain	Grade 11	
		Required Items	**% BP Match**
1	Overall	11	100.00
	DOK 2 or higher	≥ 4	100.00
	Priority Cluster	8	100.00
	Targets D, E	1–2	100.00
	Target F	0–1	100.00
	Targets G, H, I	2	100.00
	Target J	0–2	100.00
	Target K	0–2	100.00
	Targets L, M, N	2	100.00
	Supporting Cluster	3	100.00
	Target O	0–2	100.00
	Target P	0–2	100.00
	Targets A, B	0–1	100.00
	Target C	0–1	100.00
2 and 4	Overall	3	100.00
	DOK 3 or higher	≥ 1	100.00
	2. Target A	0–1	100.00
	2. Targets B, C, D	0–1	100.00
	4. Targets A, D	0–1	100.00
	4. Targets B, E	0–1	100.00
	4. Targets C, F	0–1	100.00
3	Overall	4	100.00
	DOK 3 or higher	≥ 1	100.00
	Targets A, D	1–2	100.00
	Targets B, E	1–2	100.00
	Targets C, F, G	0–1	100.00

Table 28 summarizes the target coverage by claim and includes the average and range of the number of unique targets administered in each delivered CAT component. Since the test blueprint is not required to cover all targets in each test, it is expected that the number of targets covered varies across tests. Although the target coverage varies somewhat across individual tests, all targets are covered at an aggregate level across all tests combined.

Table 28. 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	7.6	4.0	1.0	3.0	5–8	4–4	1–1	3–3
4	14	5	1	3	7.8	4.0	1.0	3.0	6–8	4–4	1–1	3–3
5	14	5	1	3	7.6	4.0	1.0	3.0	6–8	4–4	1–1	3–3
6	14	5	1	3	9.2	4.0	1.0	3.0	7–10	4–4	1–1	3–3
7	14	5	1	3	9.4	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	6–10	4–4	1–1	3–3
Mathematics												
3	11	4	6	6	9.0	1.0	3.6	2.0	9–9	1–1	3–4	2–2
4	12	4	6	6	9.0	1.0	3.6	2.0	8–9	1–1	3–4	2–2
5	11	4	6	6	8.0	1.0	3.4	2.0	8–8	1–1	3–4	2–2
6	10	4	7	6	8.6	1.0	3.4	2.0	8–9	1–1	3–4	2–2
7	9	4	7	6	6.3	1.0	3.4	2.0	5–7	1–1	2–4	2–2
8	10	4	7	6	9.0	1.0	3.5	2.0	7–9	1–1	3–4	2–2
11	11	4	7	6	10.1	1.0	3.4	2.0	7–11	1–1	3–4	2–2

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

4.2 Evidence on Internal Structure

The measurement model used in the ISAT ELA/L and mathematics 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). A high correlation among claim scores is evidence that the ISAT ELA/L and mathematics assessment measures 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, disattenuated correlation), are presented in Tables 29 and 30. The correction for attenuation indicates what the correlation would be if claim scores could be measured with perfect reliability, corrected (adjusted) for measurement error estimates.

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

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

Table 29. Correlations Among Claim Scores for ELA/L

Grade	Claim	Observed & Disattenuated Correlation			
		Claim 1	Claim 2	Claim 3	Claim 4
3	Claim 1: Reading		0.91	1	1
	Claim 2: Writing	0.58		1	1
	Claim 3: Listening	0.47	0.46		1
	Claim 4: Research	0.53	0.56	0.44	
4	Claim 1: Reading		0.93	1	0.99
	Claim 2: Writing	0.59		1	0.96
	Claim 3: Listening	0.48	0.47		1
	Claim 4: Research	0.53	0.54	0.43	
5	Claim 1: Reading		0.89	1	1
	Claim 2: Writing	0.58		1	0.96
	Claim 3: Listening	0.52	0.50		1
	Claim 4: Research	0.57	0.59	0.49	
6	Claim 1: Reading		0.89	1	0.96
	Claim 2: Writing	0.62		1	0.96
	Claim 3: Listening	0.52	0.48		1
	Claim 4: Research	0.57	0.57	0.45	
7	Claim 1: Reading		0.89	1	0.96
	Claim 2: Writing	0.61		1	0.96
	Claim 3: Listening	0.49	0.48		1
	Claim 4: Research	0.56	0.59	0.44	
8	Claim 1: Reading		0.90	1	0.95
	Claim 2: Writing	0.63		1	0.96
	Claim 3: Listening	0.54	0.51		1
	Claim 4: Research	0.57	0.59	0.46	
11	Claim 1: Reading		0.90	1	0.96
	Claim 2: Writing	0.64		1	0.98
	Claim 3: Listening	0.51	0.50		1
	Claim 4: Research	0.58	0.62	0.45	

Table 30. Correlations Among Claim Scores for Mathematics

Grade	Claim	Observed & Disattenuated Correlation		
		Claim 1	**Claims 2 & 4**	**Claim 3**
3	Claim 1		0.99	1
	Claims 2 & 4	0.73		1
	Claim 3	0.68	0.68	
4	Claim 1		0.99	1
	Claims 2 & 4	0.73		1
	Claim 3	0.71	0.69	
5	Claim 1		1	1
	Claims 2 & 4	0.70		1
	Claim 3	0.67	0.65	
6	Claim 1		1	1
	Claims 2 & 4	0.73		1
	Claim 3	0.68	0.66	
7	Claim 1		1	1
	Claims 2 & 4	0.71		1
	Claim 3	0.64	0.62	
8	Claim 1		1	1
	Claims 2 & 4	0.70		1
	Claim 3	0.67	0.65	
11	Claim 1		1	0.98
	Claims 2 & 4	0.67		1
	Claim 3	0.61	0.59	

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

5. RELIABILITY

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

The reliability evidence of the Idaho Standards Achievement Test (ISAT) summative assessments is provided with marginal reliability, CSEM, and classification accuracy and consistency in each achievement level.

5.1 MARGINAL RELIABILITY

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

The marginal reliability ($\bar{\rho}$) 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 CSEM. In IRT, CSEM is estimated as a function of test information provided by a given set of items that makes up the test. In the 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 31 presents the marginal reliability coefficients and the average CSEM for the total scale scores.

Table 31. Marginal Reliability for ELA/L and Mathematics

Grade	*N*	Number of Items Specified in Test Blueprint	Marginal Reliability	Scale Score Mean	Scale Score SD	Average CSEM
			ELA/L			
3	23,374	22	0.87	2421.51	97.51	35.73
4	23,631	22	0.86	2465.51	102.71	37.98
5	23,742	22	0.88	2505.19	106.65	37.65
6	23,513	24	0.88	2529.11	100.41	35.08
7	23,766	24	0.88	2556.99	107.76	37.38
8	23,923	24	0.88	2564.33	109.78	37.61
11	22,710	24	0.88	2598.36	122.68	41.75
			Mathematics			
3	23,524	21–23	0.91	2430.14	90.28	27.48
4	23,806	21–23	0.91	2475.85	91.38	28.06
5	23,864	21–23	0.89	2499.64	101.28	32.93
6	23,631	20–23	0.90	2516.23	111.95	35.65
7	23,859	20–23	0.89	2537.46	115.31	38.97
8	24,013	20–23	0.88	2549.76	128.08	43.78
11	23,022	22–24	0.87	2562.86	131.45	48.27

5.2 STANDARD ERROR CURVES

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

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

A consequence of the selection algorithm's prioritization of meeting blueprint requirements is that student ability near the low and high extremes of the scale is measured with relatively less precision. This produces the expected u-curve shape for the CSEM plots in Figures 11 and 12. An adaptive test with an infinitely large item pool and a selection algorithm that focused on maximizing information over blueprint requirements would produce CSEM curves that are more flat. The ISATs focus on increasing precision where it is most needed, ability scores near and in between the cut scores. It is worth noting that larger standard errors are observed at the lower ends of the score distribution, relative to the higher ends. This occurs because the item pools currently have a shortage of very easy items that are better targeted toward these lower-achieving students. Content experts use this information to consider how to further target and populate item pools.

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

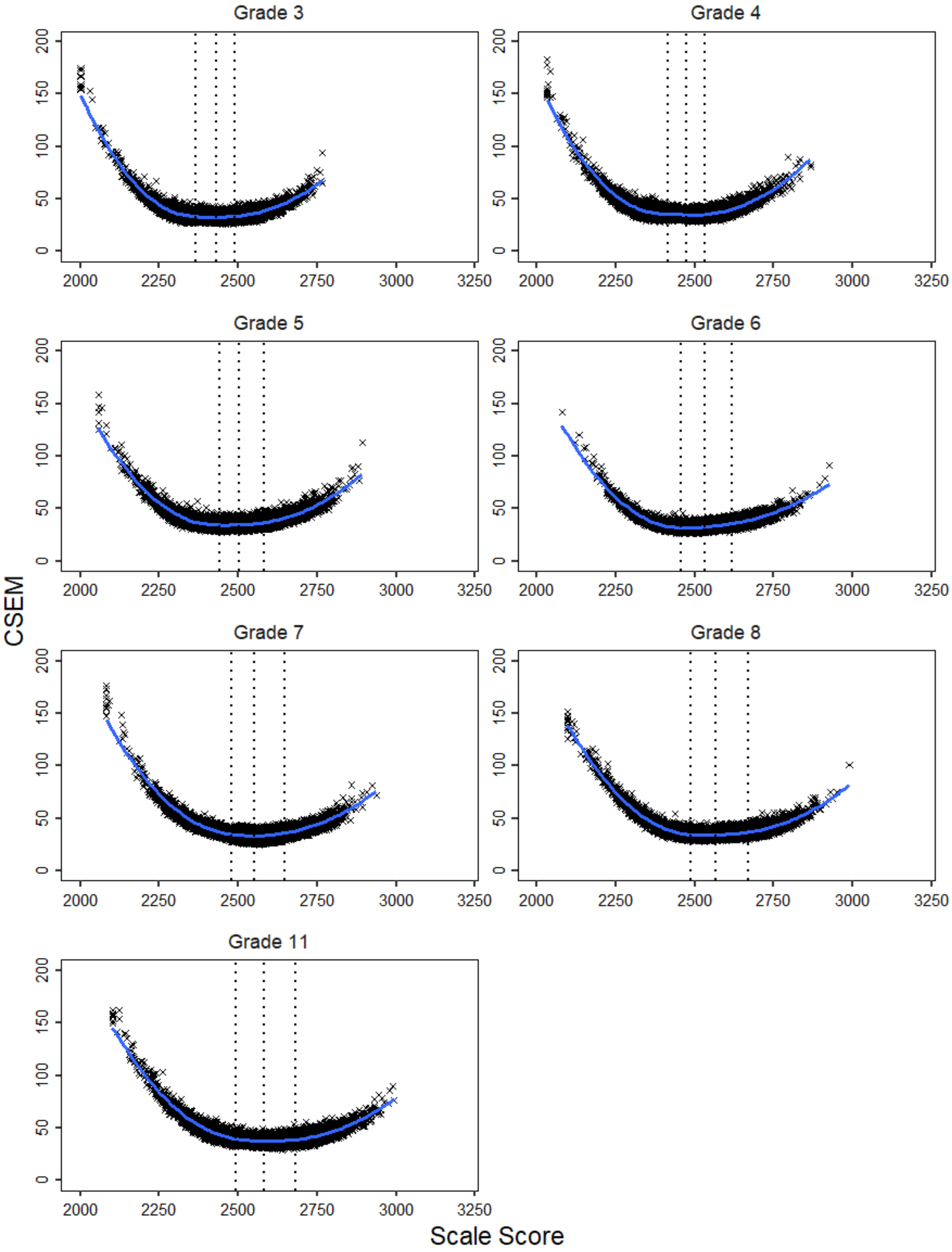
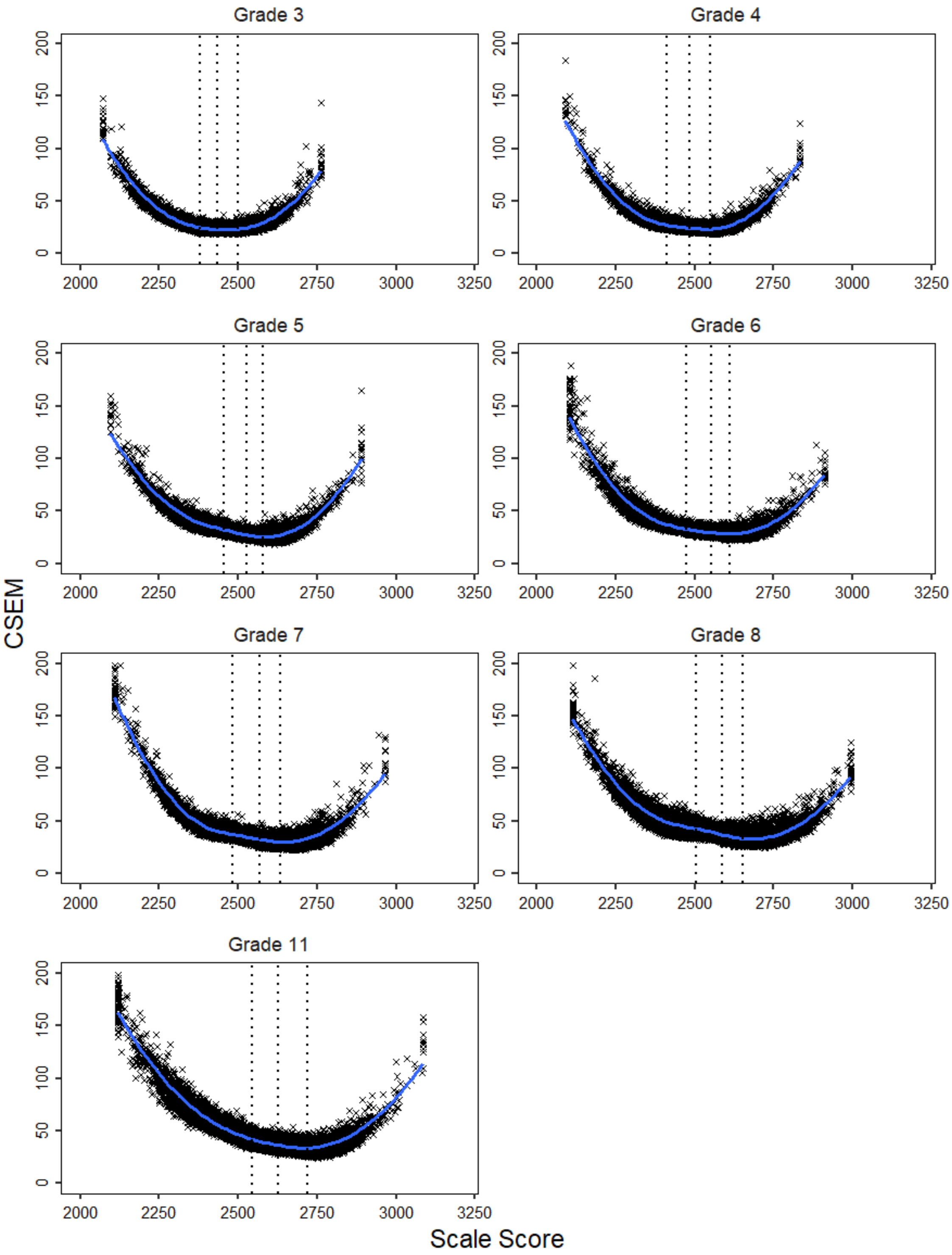


Figure 12. Conditional Standard Error of Measurement for Mathematics



The CSEMs presented in Figures 11 and 12 are summarized in Tables 32 and 33. Table 32 provides the average CSEM for all scale scores and by achievement level. Table 33 presents the average CSEMs at each cut score and the difference in average CSEMs between two cut scores. As shown in Figures 11 and 12, the greatest average CSEM is in Level 1 for most grades in ELA/L and all grades in mathematics. Average CSEMs at all cut scores are similar in ELA/L, but larger in Level 2 cut scores in mathematics. All CSEMs are reported in the scale score metric.

Table 32. Average Conditional Standard Error of Measurement by Achievement Level

Grade	Level 1	Level 2	Level 3	Level 4	Average CSEM
			ELA/L		
3	39.48	32.79	32.94	36.25	35.73
4	41.13	35.24	34.77	38.82	37.98
5	39.18	34.32	35.37	40.88	37.65
6	36.52	31.87	34.00	38.63	35.08
7	43.16	33.70	34.23	39.46	37.38
8	43.27	33.87	34.95	39.64	37.61
11	49.86	38.09	37.86	42.31	41.75
			Mathematics		
3	33.54	23.77	23.11	27.75	27.48
4	35.32	25.50	23.70	27.05	28.06
5	40.72	29.66	26.24	28.74	32.93
6	44.54	30.95	28.88	31.12	35.65
7	49.72	34.95	30.97	32.75	38.97
8	53.32	40.13	34.81	36.31	43.78
11	59.30	39.02	34.82	37.68	48.27

Table 33. Average Conditional Standard Error of Measurement at Each Achievement Level Cut Score and Difference Between the SEMs for Two Cuts

Grade	L2 Cut	L3 Cut	L4 Cut	|L2–L3|	|L3–L4|	|L2–L4|
			ELA/L			
3	33.70	32.58	33.44	1.12	0.86	0.26
4	35.72	35.02	34.50	0.70	0.52	1.22
5	34.32	34.65	36.11	0.33	1.46	1.79
6	31.54	32.42	35.41	0.88	2.99	3.88
7	34.72	33.70	36.06	1.02	2.36	1.34
8	34.65	33.96	36.32	0.69	2.37	1.68
11	40.03	37.62	38.96	2.41	1.34	1.07
			Mathematics			
3	24.78	23.12	23.16	1.66	0.04	1.62
4	27.29	24.68	23.35	2.61	1.33	3.94
5	32.73	27.33	25.22	5.41	2.10	7.51
6	32.80	29.48	28.08	3.32	1.40	4.72
7	37.36	32.28	30.35	5.08	1.93	7.01
8	43.21	36.84	33.54	6.37	3.30	9.68
11	42.25	36.69	33.17	5.56	3.51	9.07

5.3 RELIABILITY OF ACHIEVEMENT CLASSIFICATION

When student performance is reported in terms of achievement levels, a reliability of achievement classification is computed in terms of the probabilities of accurate and consistent classification of students as specified in Standard 2.16 in the *Standards for Educational and Psychological Testing* (AERA, APA, and NCME, 2014). The indices consider the accuracy and consistency of classifications.

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

The classification index can be examined in terms of the classification accuracy and the classification consistency. Classification accuracy refers to the agreement between the classifications based on the form actually taken and the classifications that would be made on the basis of the test takers' true scores if their true scores could somehow be known. Classification consistency refers to the agreement between the classifications based on the form (adaptively administered items) actually taken and the classifications that would be made on the basis of an alternate form (another set of adaptively administered items given the same ability), that is, the percentages of students who would be 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 on the basis of students' item scores and the item parameters, along with the assumed underlying latent ability distribution as described in the following paragraph. The true score is an expected value of the test score with a measurement error.

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

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

Instead of assuming a normal distribution of $\hat{\theta}_i \sim N\left(\theta_i, se^2(\hat{\theta}_i)\right)$, the above probabilities can be estimated directly using the likelihood function.

The likelihood function of theta given a student's item scores represents the likelihood of the student's ability at that theta value. Integrating the likelihood values over the range of theta at and above the cut point (with proper normalization) represents the probability of the student's latent ability or the true score being at or above that cut point. If a student with estimated theta is below the cut point, a probability of being at or above the cut point is an estimate of the chance that this student is misclassified as below the cut, and one minus that probability is the estimate of the chance that the student is correctly classified as below the cut score. Using this logic, the various classification probabilities can be defined.

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

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

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

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

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

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

where d stands for dichotomous and p stands for polytomous items; $\mathbf{b}_j = \left(a_j,b_j,c_j\right)$ if the *j*th item is a dichotomous item, and $\mathbf{b}_j = (a_j,b_{j1},\ldots,b_{jK_i})$ if the *j*th item is a polytomous item; a_j is the item's discrimination parameter (for Rasch model, $a_j = 1$), c_j is the guessing parameter (for Rasch and two-parameter logistic [2PL] models, $c_j = 0$), and D is 1.7 for non-Rasch models and 1 for Rasch model.

Classification Accuracy

Using p_{il}, a $L \times L$ table can be constructed as

$$\begin{pmatrix} n_{a11} & \cdots & n_{a1L} \\ \vdots & \vdots & \vdots \\ n_{aL1} & \cdots & n_{aLL} \end{pmatrix},$$

where $n_{alm} = \sum_{pl_i=l} p_{im}$. n_{alm} is the expected number of students at achievement level *lm*, pl_i is the *i*th student's achievement level, and p_{im} are the probabilities of the *i*th student being classified at achievement level *m*. In the given 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 $n_{clm} = \sum_{i=1}^{N} p_{il} p_{im} \cdot p_{il}$ and p_{im} are the probabilities of the ith student being classified at achievement levels 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 overall scale scores. Table 34 provides the percentages of classification accuracy and consistency for overall, by achievement level, and at proficiency cut score.

The overall classification index ranged from 73% to 79% for accuracy and from 65% to 71% for consistency across all grades and subjects. For achievement levels, the classification index is higher in L1 and L4 than in L2 and L3. The higher accuracy at L1 and L4 is due to the fact that the intervals used to compute the classification probabilities for students in L1 and L4 $[-\infty,$ L2 cut; L4 cut, $\infty]$ are wider than the intervals used to compute the classification probabilities for students in L2 and L3 [L2 cut, L3 cut; L3 cut, L4 cut]. The misclassification probability tends to be higher for narrower intervals. Classification accuracy and classification consistency at the proficiency cut scores were high, ranging from 90% to 92% for accuracy and from 86% to 89% for consistency.

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

Table 34. Classification Accuracy and Consistency

Grade	Achievement Level	ELA/L		Mathematics	
		% Accuracy	% Consistency	% Accuracy	% Consistency
3	Overall	73	65	77	69
	L1	88	80	86	80
	L2	60	49	66	53
	L3	56	45	71	62
	L4	84	76	86	79
	Proficiency Cut	90	86	92	88
4	Overall	73	65	78	70
	L1	88	81	87	80
	L2	53	42	72	62
	L3	55	45	71	60
	L4	84	76	86	80
	Proficiency Cut	90	86	92	88
5	Overall	74	65	77	69
	L1	88	80	88	82
	L2	56	44	67	57
	L3	65	54	60	48
	L4	84	76	87	80
	Proficiency Cut	90	86	92	89
6	Overall	75	66	77	69
	L1	88	80	90	84
	L2	66	54	69	59
	L3	69	60	61	49
	L4	81	71	86	78
	Proficiency Cut	91	87	91	88
7	Overall	76	67	77	68
	L1	88	80	88	82
	L2	64	52	66	56
	L3	72	63	64	53
	L4	82	72	86	78
	Proficiency Cut	91	87	91	87
8	Overall	76	67	76	67
	L1	88	80	87	81
	L2	66	55	62	50
	L3	73	64	59	48
	L4	81	70	88	80
	Proficiency Cut	91	87	91	88
11	Overall	76	67	79	71
	L1	87	80	89	85
	L2	66	54	63	52
	L3	69	60	69	58
	L4	84	76	86	76
	Proficiency Cut	91	88	92	89

5.4 Reliability for Subgroups

The reliability of test scores is also computed by subgroup. Tables 35–42 present the marginal reliability coefficients and average CSEMs by subgroup. The reliability coefficients are similar across subgroups except for some subgroups with low performance (e.g., English learner [EL], special education) in some grades, a large percentage of students in Level 1 with large CSEMs.

Table 35. Marginal Reliability Coefficients for Overall and by Subgroup: ELA/L (Grades 3–4)

Subgroup	Grade 3					Grade 4				
	N	MR	SS	SD	CSEM	N	MR	SS	SD	CSEM
All Students	23,374	0.87	2421.51	97.51	35.73	23,631	0.86	2465.51	102.71	37.98
Female	11,507	0.86	2427.68	96.46	35.56	11,477	0.86	2472.81	101.11	37.78
Male	11,867	0.87	2415.52	98.16	35.91	12,154	0.86	2458.61	103.73	38.17
African American	261	0.86	2376.99	98.49	36.88	276	0.85	2414.89	101.41	38.99
AI/AN	207	0.84	2369.38	90.11	36.41	238	0.83	2419.31	93.75	38.38
Asian	251	0.87	2454.00	98.52	35.76	249	0.85	2506.65	97.74	37.66
Hispanic	4,464	0.85	2385.66	93.75	36.36	4,564	0.85	2426.59	98.11	38.39
Pacific Islander	261	0.84	2415.39	93.43	37.38	215	0.87	2460.61	104.29	37.64
White	17,666	0.86	2431.80	95.94	35.52	17,979	0.86	2476.54	101.03	37.83
EL	2,006	0.83	2361.21	90.84	37.50	2,092	0.83	2400.57	97.35	40.08
Special Education	2,912	0.82	2342.96	91.93	39.03	3,077	0.82	2373.12	98.21	41.91
Section 504	719	0.85	2410.62	95.43	36.61	903	0.84	2461.96	94.25	37.26

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

Table 36. Marginal Reliability Coefficients for Overall and by Subgroup: ELA/L (Grades 5–6)

Subgroup	Grade 5					Grade 6				
	N	MR	SS	SD	CSEM	N	MR	SS	SD	CSEM
All Students	23,742	0.88	2505.19	106.65	37.65	23,513	0.88	2529.11	100.41	35.08
Female	11,701	0.87	2513.82	105.24	37.67	11,436	0.87	2540.91	98.23	35.05
Male	12,041	0.88	2496.80	107.34	37.63	12,077	0.88	2517.93	101.17	35.10
African American	269	0.86	2441.43	104.23	38.52	251	0.88	2477.04	99.55	34.83
AI/AN	241	0.87	2452.14	103.29	37.65	191	0.87	2474.26	98.20	34.96
Asian	298	0.89	2537.03	121.09	39.83	263	0.88	2572.28	104.21	36.34
Hispanic	4,426	0.86	2460.18	100.14	37.44	4,435	0.87	2486.42	97.45	34.93
Pacific Islander	193	0.89	2492.08	112.30	38.02	213	0.87	2538.05	95.37	34.91
White	18,229	0.87	2517.51	104.40	37.64	18,077	0.87	2540.36	97.75	35.10
EL	2,125	0.86	2436.51	102.41	38.43	2,140	0.87	2467.74	100.18	35.59
Special Education	2,998	0.83	2399.45	95.87	39.81	2,697	0.81	2416.60	86.49	37.23
Section 504	1,080	0.86	2496.85	98.38	36.99	1,259	0.86	2516.13	92.53	34.81

Table 37. Marginal Reliability Coefficients for Overall and by Subgroup: ELA/L (Grades 7–8)

Subgroup	**Grade 7**					**Grade 8**				
	N	**MR**	**SS**	**SD**	**CSEM**	**N**	**MR**	**SS**	**SD**	**CSEM**
All Students	23,766	0.88	2556.99	107.76	37.38	23,923	0.88	2564.33	109.78	37.61
Female	11,710	0.87	2571.22	104.04	37.10	11,623	0.88	2580.34	105.63	37.20
Male	12,056	0.88	2543.18	109.50	37.66	12,300	0.88	2549.19	111.46	37.99
African American	281	0.87	2499.60	117.09	41.44	279	0.88	2505.32	123.65	42.78
AI/AN	243	0.87	2508.80	104.80	37.98	232	0.88	2520.87	109.12	38.49
Asian	261	0.89	2596.89	114.14	38.33	266	0.89	2602.40	120.28	39.19
Hispanic	4,649	0.87	2511.91	108.25	38.48	4,570	0.87	2521.48	107.19	38.25
Pacific Islander	208	0.88	2550.67	108.38	36.99	194	0.86	2571.78	99.52	37.38
White	18,040	0.87	2569.77	103.69	37.01	18,305	0.88	2576.04	106.88	37.32
EL	2,199	0.87	2485.02	111.03	40.30	2,231	0.87	2498.54	114.29	40.44
Special Education	2,623	0.81	2437.40	98.32	42.82	2,542	0.80	2435.78	98.02	43.97
Section 504	1,375	0.86	2544.13	97.13	36.42	1,423	0.86	2553.75	97.75	36.64

Table 38. Marginal Reliability Coefficients for Overall and by Subgroup: ELA/L (Grade 11)

Subgroup	**Grade 11**				
	N	**MR**	**SS**	**SD**	**CSEM**
All Students	22,710	0.88	2598.36	122.68	41.75
Female	11,052	0.87	2616.68	115.23	41.04
Male	11,658	0.89	2580.99	126.93	42.41
African American	300	0.88	2514.31	130.92	46.19
AI/AN	204	0.86	2547.35	115.76	42.97
Asian	283	0.90	2636.13	135.81	43.16
Hispanic	4,376	0.87	2553.48	117.62	42.29
Pacific Islander	174	0.88	2592.37	118.23	41.07
White	17,320	0.88	2611.32	120.19	41.49
EL	1,832	0.87	2519.34	123.39	44.82
Special Education	1,946	0.79	2460.15	105.27	47.78
Section 504	1,457	0.87	2588.14	116.80	41.48

Table 39. Marginal Reliability Coefficients for Overall and by Subgroup: Mathematics (Grades 3–4)

Subgroup	**Grade 3**					**Grade 4**				
	N	**MR**	**SS**	**SD**	**CSEM**	**N**	**MR**	**SS**	**SD**	**CSEM**
All Students	23,524	0.91	2430.14	90.28	27.48	23,806	0.91	2475.85	91.38	28.06
Female	11,591	0.90	2423.83	87.28	27.27	11,555	0.90	2469.50	86.91	27.68
Male	11,933	0.91	2436.27	92.68	27.68	12,251	0.91	2481.84	95.03	28.41
African American	274	0.89	2367.08	102.51	34.55	295	0.87	2420.90	94.55	34.31
AI/AN	208	0.88	2381.92	85.95	30.02	238	0.87	2430.11	81.37	28.94
Asian	255	0.92	2460.90	100.46	28.86	253	0.92	2519.06	103.58	29.91
Hispanic	4,566	0.89	2393.26	87.14	28.64	4,676	0.88	2435.13	87.43	30.18
Pacific Islander	261	0.90	2423.79	85.85	26.83	214	0.91	2466.83	92.53	28.01
White	17,671	0.90	2441.29	87.55	26.99	17,982	0.90	2487.94	88.38	27.29
EL	2,159	0.88	2372.62	87.84	30.49	2,271	0.86	2416.39	86.97	32.30
Special Education	2,913	0.88	2354.85	96.39	33.67	3,084	0.86	2393.93	92.87	34.51
Section 504	731	0.90	2424.79	85.50	26.72	908	0.90	2474.92	82.68	26.76

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

Table 40. Marginal Reliability Coefficients for Overall and by Subgroup: Mathematics (Grades 5–6)

Subgroup	**Grade 5**					**Grade 6**				
	N	**MR**	**SS**	**SD**	**CSEM**	**N**	**MR**	**SS**	**SD**	**CSEM**
All Students	23,864	0.89	2499.64	101.28	32.93	23,631	0.90	2516.23	111.95	35.65
Female	11,748	0.89	2493.62	97.61	32.84	11,490	0.89	2512.78	108.95	35.38
Male	12,116	0.90	2505.47	104.38	33.01	12,141	0.90	2519.49	114.63	35.90
African American	285	0.86	2423.87	110.10	41.14	259	0.87	2438.81	124.14	44.71
AI/AN	242	0.85	2445.35	93.80	35.87	192	0.85	2450.68	102.84	39.38
Asian	308	0.92	2534.83	120.86	34.16	269	0.92	2578.22	123.42	34.86
Hispanic	4,520	0.86	2454.31	93.88	35.51	4,526	0.87	2461.48	109.16	40.05
Pacific Islander	193	0.89	2490.25	98.02	32.85	213	0.89	2522.94	103.57	33.69
White	18,200	0.89	2512.64	98.46	32.00	18,059	0.90	2531.29	106.84	34.16
EL	2,288	0.85	2434.26	96.90	38.08	2,288	0.86	2441.44	115.34	43.77
Special Education	2,999	0.82	2401.40	96.18	41.19	2,689	0.80	2393.37	108.42	47.99
Section 504	1,084	0.88	2493.28	92.06	32.46	1,264	0.88	2507.75	100.14	34.14

Table 41. Marginal Reliability Coefficients for Overall and by Subgroup: Mathematics (Grades 7–8)

Subgroup	**Grade 7**					**Grade 8**				
	N	**MR**	**SS**	**SD**	**CSEM**	**N**	**MR**	**SS**	**SD**	**CSEM**
All Students	23,859	0.89	2537.46	115.31	38.97	24,013	0.88	2549.76	128.08	43.78
Female	11,748	0.88	2531.96	113.34	39.16	11,665	0.88	2547.85	123.30	43.31
Male	12,111	0.89	2542.80	116.96	38.77	12,348	0.89	2551.56	132.43	44.22
African American	293	0.84	2459.84	121.79	48.45	293	0.84	2468.40	127.58	51.75
AI/AN	243	0.84	2478.42	112.99	44.86	228	0.85	2493.07	125.04	49.10
Asian	265	0.91	2588.63	135.04	40.03	271	0.92	2618.84	155.47	44.45
Hispanic	4,714	0.85	2482.67	112.80	43.88	4,652	0.83	2490.31	117.46	47.84
Pacific Islander	209	0.89	2532.41	117.31	38.49	195	0.87	2546.34	118.47	42.89
White	18,022	0.89	2553.68	109.92	37.20	18,270	0.88	2566.44	124.72	42.38
EL	2,330	0.83	2458.11	115.94	47.86	2,371	0.83	2468.95	125.10	51.59
Special Education	2,622	0.75	2412.88	106.29	52.68	2,534	0.72	2407.37	109.00	57.29
Section 504	1,378	0.86	2527.80	102.16	37.81	1,419	0.85	2535.00	110.71	43.24

Table 42. Marginal Reliability Coefficients for Overall and by Subgroup: Mathematics (Grade 11)

Subgroup	**Grade 11**				
	N	**MR**	**SS**	**SD**	**CSEM**
All Students	23,022	0.87	2562.86	131.45	48.27
Female	11,231	0.85	2560.38	121.71	47.24
Male	11,791	0.88	2565.24	140.07	49.22
African American	301	0.76	2471.49	128.02	62.14
AI/AN	204	0.78	2484.91	121.73	57.35
Asian	287	0.91	2631.31	148.59	45.24
Hispanic	4,409	0.79	2505.65	115.96	53.24
Pacific Islander	174	0.87	2544.02	141.14	50.98
White	17,587	0.87	2578.93	129.64	46.53
EL	1,884	0.77	2485.17	120.00	57.13
Special Education	1,942	0.59	2419.23	105.19	67.16
Section 504	1,483	0.84	2545.47	123.14	49.19

5.5 RELIABILITY FOR CLAIM SCORES

The marginal reliability, average and standard deviation of scale scores, and average of CSEM are also computed for claim scores by test and grade. In mathematics, claims 2 and 4 are combined to have enough items to generate a score. Given the small number of items, the reliabilities for claim scores are low, thus they were not reported at student level. Tables 43 and 44 present the marginal reliability coefficients and descriptive statistics by claim in ELA/L and mathematics, respectively.

Table 43. Marginal Reliability Coefficients for Claim Scores in ELA/L

Grade	Claim	Number of Items Specified in Test Blueprint	Marginal Reliability	Scale Score Mean	Scale Score SD	Average CSEM
3	Claim 1: Reading	8	0.60	2428.68	121.17	76.42
	Claim 2: Writing	5	0.67	2409.42	124.54	71.11
	Claim 3: Listening	4	0.25	2427.73	142.59	123.46
	Claim 4: Research	5	0.46	2420.12	131.44	96.59
4	Claim 1: Reading	8	0.60	2475.24	129.45	81.98
	Claim 2: Writing	5	0.67	2451.84	132.02	76.37
	Claim 3: Listening	4	0.30	2471.25	147.79	123.74
	Claim 4: Research	5	0.48	2462.39	144.46	104.06
5	Claim 1: Reading	8	0.60	2507.41	132.29	83.32
	Claim 2: Writing	5	0.70	2502.53	137.57	75.64
	Claim 3: Listening	4	0.33	2515.24	153.53	125.29
	Claim 4: Research	5	0.53	2500.88	140.00	95.61
6	Claim 1: Reading	10	0.70	2530.41	119.85	65.62
	Claim 2: Writing	5	0.70	2517.44	123.68	67.91
	Claim 3: Listening	4	0.28	2551.01	161.37	136.57
	Claim 4: Research	5	0.49	2535.72	142.98	101.78
7	Claim 1: Reading	10	0.65	2555.64	131.62	77.96
	Claim 2: Writing	5	0.73	2552.30	136.05	71.31
	Claim 3: Listening	4	0.29	2559.08	151.17	127.77
	Claim 4: Research	5	0.52	2557.71	154.33	106.65
8	Claim 1: Reading	10	0.68	2560.33	127.53	71.68
	Claim 2: Writing	5	0.71	2560.04	137.51	73.78
	Claim 3: Listening	4	0.35	2572.48	162.83	131.59
	Claim 4: Research	5	0.52	2570.43	155.99	107.82
11	Claim 1: Reading	10	0.68	2598.07	144.95	82.07
	Claim 2: Writing	5	0.73	2594.66	154.19	79.83
	Claim 3: Listening	4	0.33	2598.02	180.18	147.38
	Claim 4: Research	5	0.54	2599.90	170.17	115.05

Table 44. Marginal Reliability Coefficients for Claim Scores in Mathematics

Grade	Claim	Number of Items Specified in Test Blueprint	Marginal Reliability	Scale Score Mean	Scale Score SD	Average CSEM
3	Claim 1	10	0.80	2433.86	102.55	45.49
	Claims 2 & 4	6–8	0.67	2426.42	102.22	58.37
	Claim 3	5–6	0.58	2425.91	112.85	72.97
4	Claim 1	10	0.80	2480.80	102.55	45.39
	Claims 2 & 4	5–7	0.68	2470.40	106.48	60.08
	Claim 3	5–6	0.59	2469.18	108.88	69.32
5	Claim 1	10	0.79	2507.22	117.16	54.03
	Claims 2 & 4	5–7	0.62	2491.70	114.34	70.18
	Claim 3	5–6	0.54	2488.13	129.82	88.43
6	Claim 1	10	0.80	2520.67	125.08	55.37
	Claims 2 & 4	6–7	0.64	2508.02	127.75	76.89
	Claim 3	5–7	0.52	2514.00	130.52	90.20
7	Claim 1	10	0.78	2541.39	132.06	61.99
	Claims 2 & 4	6–7	0.58	2528.40	128.70	83.62
	Claim 3	4–6	0.51	2531.67	148.63	104.09
8	Claim 1	10	0.78	2552.37	142.97	67.49
	Claims 2 & 4	5–7	0.54	2546.43	142.55	96.99
	Claim 3	5–6	0.53	2539.01	162.48	111.05
11	Claim 1	11	0.76	2557.71	144.38	70.16
	Claims 2 & 4	5–7	0.56	2559.20	169.88	112.51
	Claim 3	5–6	0.50	2547.92	170.22	119.78

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

6. SCORING

The Smarter Balanced Assessment Consortium provided the vertically scaled item parameters by linking across all grades using common items in adjacent grades. All scores are estimated based on these item parameters. Each student received an overall scale score, an overall achievement level, and a performance category for each claim. This section describes the rules used in generating scores, as well as the hand-scoring procedure.

6.1 ESTIMATING STUDENT ABILITY USING MAXIMUM LIKELIHOOD ESTIMATION

The ISAT ELA/L and mathematics 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

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

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

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

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

in the case of items with two or more points,

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

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

Standard Error of Measurement

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

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

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

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

where m_i is the maximum possible score point (starting from 0) for the ith item, and D is the scale factor, 1.7. The SE is calculated based only on the answered item(s) for both complete and incomplete tests. The upper bound of the SE is set to 2.5 on 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 and claim ability estimates after each item is answered. When a previously answered item is changed, the proficiency estimate is adjusted to account for the changed responses when the next new item is selected. While the update of the ability estimates is performed at each iteration, the overall and claim scores are recalculated using all data at the end of the assessment for the final score.

6.2 Rules for Transforming Theta to Vertical Scale Scores

The student's performance in each subject is summarized in an overall test score referred to as a *scale score.* The scale scores represent a linear transformation of the ability estimates (theta scores) using the formula, $SS = a * \theta + b$. The scaling constants a and b are provided by the Smarter Balanced assessment consortium. Table 45 presents the scaling constants for each subject for the theta-to-scale score linear transformation. Scale scores are rounded to an integer.

Table 45. Vertical Scaling Constants on the Reporting Metric

Subject	Grade	Slope (a)	Intercept (b)
ELA/L	3–8, 11	85.8	2508.2
Mathematics	3–8, 11	79.3	2514.9

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

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

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

The scale scores are mapped into four achievement levels using three achievement standards (i.e., cut scores). Table 46 provides three achievement standards for each grade and content area.

Table 46. Cut Scores in Scale Scores

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

6.3 LOWEST/HIGHEST OBTAINABLE SCORES (LOSS/HOSS)

Although the observed score is measured more precisely in an adaptive test than in a fixed-form test, especially for high- and low-performing students, if the item pool does not include enough easy or difficult items to measure low- and high-performing students, the standard error can be large in low and high ends of the ability range. The Smarter Balanced Assessment Consortium decided to truncate extreme unreliable student ability estimates. Table 47 presents the lowest obtainable score (lowest obtainable theta score [LOT] or lowest obtainable scale score [LOSS]) and the highest obtainable score (highest obtainable theta score [HOT] or highest obtainable scale score [HOSS]). Estimated thetas lower than LOT or higher than HOT are truncated to the LOT and HOT values and are assigned LOSS and HOSS associated with the LOT and HOT. LOT and HOT were applied to all tests and total scores. The standard error for LOT and HOT is computed using the LOT and HOT ability estimates given the administered items.

Table 47. Extended Lowest and Highest Obtainable Scores

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

6.4 Scoring All Correct and All Incorrect Cases

In item response theory (IRT) maximum likelihood (ML) ability estimation methods, zero and perfect scores are assigned the ability of minus and plus infinity. For all correct and all incorrect cases, the highest obtainable scores (HOT and HOSS) or the lowest obtainable scores (LOT and LOSS) were assigned in the 2014–2015 test administration. Since the 2015–2016 test 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 test [CAT] and performance task [PT]) for a student.

6.5 Target Scores

The target-level reports cannot be produced for a fixed-form test because the number of items included per target (i.e., benchmark) is too low 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 district 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 only in claim 1 for mathematics.

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

6.5.1 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:

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

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

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

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

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

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*.

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

For an aggregate unit, a target score is computed by averaging 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, or district is more effective (if $\bar{\delta}_{Tg}$is positive) or less effective (negative $\bar{\delta}_{Tg}$) in teaching a given target.

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

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

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

6.5.2 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*. The value 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:

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

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

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

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

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

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

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

For an aggregate unit, a target score is computed by averaging 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 class, teacher, school, or district is more effective (if $\bar{\delta}_{Tg}$is positive) or less effective (if $\bar{\delta}_{Tg}$is negative) in teaching a given target.

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

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

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

6.6 HAND-SCORING

Constructed-response short-answer (SA) items and essay (i.e., full write) items in English language arts/literacy (ELA/L) and SA items in mathematics for the ISAT summative assessments administered by Cambium Assessment Inc. (CAI) are routed to Measurement Incorporated (MI) for scoring. MI provides hand-scoring using human raters and automated scoring using the Project Essay Grade (PEG) engine. Idaho have elected to use a hybrid automated scoring/hand-scoring approach. The methods and results for hand-scoring and hybrid automated scoring are described in the following sections.

For hand-scoring items in the 2023–2024 ISAT summative operational item pool, there were a total of 470 ELA/L SA items, 186 ELA/L essay items, and 334 mathematics items. Table 48 shows the number of hand-scored items by grade and subject.

Table 48. Number of Hand-scored Items in 2023–2024 ISAT Summative Item Pool, by Grade and Subject

Grade	ELA/L		Mathematics
	Short Answer	Essay	
3	13	25	54
4	16	27	49
5	14	27	86
6	85	20	51
7	91	29	22
8	83	29	30
11	168	29	42
Total	470	186	334

All guidelines for hand-scoring responses were specified by Smarter Balanced. Outlined below is the hand-scoring process MI followed in spring 2024 in accordance with the Smarter Balanced guidelines. This process applied to the scoring of all students constructed responses for ELA/L SA and essay items and mathematics items.

6.6.1 Rater Selection

MI has developed a pool of approximately five thousand raters experienced in scoring the Smarter Balanced assessments. MI first recruited qualified raters who had experience scoring these assessments. Rater accuracy data, collected during prior administration scoring, was used to prioritize recruitment of the most accurate, experienced raters. Once recruited, experienced raters were assigned to the content area and grade band(s) with which they were most experienced.

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

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

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

MI requires all hand-scoring project staff (scoring directors, team leaders, raters, and clerical staff) to sign a confidentiality/nondisclosure agreement before receiving any training or viewing any secure project materials. The employment agreement indicates that no participant in training and/or scoring may reveal information about the test, the scoring criteria, or the scoring methods to any person.

6.6.2 Rater Training, Qualification, 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. Many of these sets were created during the original field-test scoring in 2014 and approved by Smarter Balanced. Additional sets were created as new items were field-tested. The same anchor sets are used each year. Additionally, MI conducts an annual review of the rater agreement and scoring materials to inform the development of item-specific, supplemental training materials. Supplemental materials are developed each summer and implemented in the subsequent operational administration. These additional materials are developed with a focus on challenging areas identified during the previous operational administration, as indicated by suboptimal rater accuracy (based on validity responses) and/or rater agreement. Supplemental materials may address item- or response-specific concerns. Supplemental materials are also created for newly operational items for which MI identifies a need for additional examples. For instance, MI may find an approach to a mathematics item that was not encountered during field testing but appears frequently during operational scoring, or an uncommon but valid way to address a Research prompt that is not reflected in the existing rubric. In these cases, MI provides examples of these specific approaches along with guidance on how to score them correctly. MI also supplement materials to provide raters with additional guidance for content-wide challenging spots—such as full write conventions—or to help them more accurately identify responses that should be flagged as non-scorable.

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

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

Until a rater had trained and qualified successfully, the rater was not permitted to score operational student responses. Training was structured so that raters understood that all scoring decisions must be grounded in the training materials. In addition, raters learned how to navigate the anchor set, developed the knowledge and flexibility needed to evaluate or escalate a variety of responses, and retained the necessary consistency to score all responses accurately.

When beginning working, all scoring personnel logged in to MI's secure Scoring Resource Center (SRC). SRC includes all online training modules, serves as the portal to MI's Virtual Scoring Center (VSC) interface, and host 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 and qualification design varied slightly depending on Smarter Balanced item type:

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

An additional validation stage supplemented full write, brief write, reading, and research rater qualification. Following the training and qualification steps described above, all prospective full write, brief write, reading, and research raters were required to score, for most items, a 20-response set of pre-scored student responses sourced from the prior test administration. Like the qualification step, raters were required to meet accuracy standards during this validation to score operational responses for a given item. Any raters who failed to meet validation accuracy standards were automatically disqualified from scoring the item despite having passed qualification. This additional validation matches the full write qualification methods that have been in place since the start of Smarter Balanced scoring in 2015 and adds an additional level of quality assurance.

Rater training time varied by grade and content area. Training for SA brief write, reading, research, and mathematics items could typically be accomplished in one day, while training for essay items took up to five days to complete. Raters generally worked 3-7 hours per day. The hours worked per day were flexible, based on the raters' shift preference and item(s) being scored. At a minimum, most raters scored 15 hours per week (day shift) or 10 hours per week (evening shift), with many scoring over 30 hours per week (day shift) or 20 hours per week (evening shift).

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

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

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

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

Following training, all training materials remained available to raters throughout scoring via the VSC Score Resource Library. This library included the item and rubric, the annotated anchor and practice sets, and any associated supplemental materials.

When scoring, raters had access only to those items for which they had successfully trained and qualified. The hand-scoring 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.

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 continue to exhibit bias after receiving a reasonable amount of feedback, they were dismissed.

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

6.6.3 Rater Monitoring, Feedback, and Evaluation

During operational scoring, five percent of the responses scored comprised pre-approved validity responses. Validity responses serve as benchmark responses as the most appropriate score for each validity response is predetermined by key stakeholders. A small set of validity responses is provided by Smarter Balanced for all vendors to use, and these are supplemented with responses selected and approved by MI scoring management. The validity pool includes anchor validity responses originating from the field test administration.[1] The pool of validity responses is selected to be generally representative of operational

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

responses, while ensuring sufficient examples of each score point. Validity results compare the score assigned by a rater to a validity response with the benchmark score of the same response. Validity responses provide a more direct measurement of rating quality than measures of inter-rater reliability (Raczynski et al., 2015).

MI calibrates validity responses to fit a unidimensional Item Response Theory (IRT) model for each content area/item type. This approach involves transforming raters' validity response scores into accuracy scores. Specifically, if the rater's score matches the "true" score of the validity response, an accuracy score of 2 is assigned. If the rater's score is adjacent to the score of the validity response, an accuracy score of 1 is assigned. Otherwise, for scores that are non-adjacent, an accuracy score of 0 is assigned. All accuracy score data for validity responses and raters are then fitted to a Generalized Partial Credit Model (GPCM) IRT model. Utilizing the resulting IRT parameters, MI calculates accuracy values for each rater based on a given set of validity responses. This calculation is conducted several times each day during scoring, providing real-time measures of rater accuracy.

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

The VSC system automatically and randomly routed the requisite number of responses to raters for second reads and validity in an inconspicuous manner. In this way raters had no means of discerning whether they were scoring a first read, a second read, or a validity response. This system also prohibited raters from being eligible to score second reads for responses they had already scored.

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

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

The monitoring system afforded the objective, dynamic identification of the most accurate raters, referred to as "expert raters." Specifically, expert raters are those who demonstrate highly accurate and consistent scoring of validity responses. Rater status changed daily based on current rater performance to ensure that any rater drift did not negatively impact scoring accuracy. Expert rater status was a precondition for conducting second readings.

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

During scoring, raters received automated feedback system based on recent performance. The automated feedback system identifies raters who require additional feedback—based on accuracy metrics—and automatically generates a custom set of responses for the rater to review. The system functions at the item level, thus providing feedback even to those raters with relatively high accuracy when the data identifies there are one or more items on which they can improve.

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

Reports using item-level accuracy expectations identified any items not meeting the expected levels of agreement. Specifically, these reports indicated the difference between expected accuracy and current accuracy for each item. Expected accuracy was defined based on historical data; in some cases (e.g., most Mathematics items) expected accuracy exceeded Smarter Balanced's minimum accuracy thresholds. In this way, reports informed improvements to the scoring accuracy of all items.

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

In addition to the automated feedback, scoring leadership provided individualized feedback to raters based on their performance. Specifically, leadership reviewed the rater's mis-scored validity responses and associated data and looked for a trend that suggests the rater has drifted from the anchored responses. If such a trend is present, leadership can tailor feedback specific to that rater, typically by presenting them with live responses they have mis-scored in a way that is reflective of their overall drift from the anchor set criteria and providing targeted, thoughtful rationales for the "correct" scores.

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

6.6.4 Rater Agreement

Rater inter-rater reliability (IRR) was computed based only on scorable responses (numeric scores) scored by two independent raters. Non-scorable responses (e.g., off-topic, off-purpose, or foreign-language responses) were scored by scoring leadership per the hand-scoring rules—and not by one expert and one random rater—and were thus excluded from IRR computations. For the hand-scored items, the human-human agreement was computed based on the 2023–2024 ISAT summative assessment.

In ELA/L essay (i.e., full writes) item responses were scored in three dimensions: conventions (0–2 rubric), evidence/elaboration (1–4 rubric), and organization/purpose (1–4 rubric). All ELA/L SA items were scored using a 0–2 rubric. Mathematics SA items were scored using 0–1, 0–2, or 0–3 rubrics.

Tables 49 through 51 provide a summary of the human-human IRR based on items with a sample size greater than or equal to 50. For Mathematics and ELA/L essay items, the tables show the majority of the items administered. For ELA/L SA items, relatively fewer items reached a sample size greater than or equal to 50, and thus a subset of the items administered are represented in the tables. The IRR is presented with mean of percent exact agreement, minimum and maximum percent exact agreements, combined percent exact and adjacent agreement, and the mean, minimum and maximum QWK. The average number of responses, as well as minimum and maximum number of responses to a given item are presented as well.

Table 49. 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	2	97.0	65	129	73.7	72.1	76.9	100.0	0.76	0.72	0.85
4	5	112.6	75	131	76.7	65.7	82.4	100.0	0.77	0.67	0.82
5	3	102.3	85	132	70.0	63.5	73.3	100.0	0.73	0.67	0.75
6	12	258.8	136	566	69.9	59.6	89.7	100.0	0.65	0.57	0.84
7	20	120.2	62	402	75.1	65.1	85.5	100.0	0.66	0.38	0.86
8	28	112.0	53	425	69.7	60.4	77.9	100.0	0.65	0.46	0.83
11	23	107.3	50	194	71.8	59.1	80.2	100.0	0.70	0.53	0.82

Table 50. Inter-Rater Agreement for ELA/L Essay Items

Grade	Trait	Number of Items	Number of Responses			%Exact			%(Exact+ Adjacent)	QWK		
			Mean	Min	Max	Mean	Min	Max		Mean	Min	Max
3	Conventions	22	94.0	57	115	70.8	62.5	79.0	100.0	0.69	0.60	0.80
	Evid/Elab	22	94.0	57	115	68.7	53.0	77.2	100.0	0.70	0.59	0.87
	Org/Purp	22	94.0	57	115	68.6	53.0	77.4	100.0	0.70	0.57	0.85
4	Conventions	26	92.2	53	117	67.1	56.8	80.5	100.0	0.71	0.65	0.83
	Evid/Elab	26	92.2	53	117	68.4	56.6	80.8	100.0	0.73	0.57	0.88
	Org/Purp	26	92.2	53	117	68.1	56.6	82.2	100.0	0.73	0.58	0.88
5	Conventions	24	107.5	54	122	69.0	55.9	77.9	100.0	0.67	0.57	0.78
	Evid/Elab	24	107.5	54	122	65.5	55.5	74.2	100.0	0.75	0.66	0.82
	Org/Purp	24	107.5	54	122	65.7	57.7	75.9	100.0	0.75	0.66	0.83
6	Conventions	16	145.2	99	168	73.4	65.4	85.4	100.0	0.70	0.57	0.78
	Evid/Elab	16	145.2	99	168	67.7	58.9	76.8	100.0	0.75	0.71	0.80
	Org/Purp	16	145.2	99	168	66.9	58.9	77.4	100.0	0.75	0.69	0.79
7	Conventions	24	96.8	57	119	71.4	59.6	83.8	100.0	0.70	0.52	0.84
	Evid/Elab	24	96.8	57	119	72.3	65.9	85.7	100.0	0.77	0.69	0.86
	Org/Purp	24	96.8	57	119	71.8	63.1	85.7	100.0	0.77	0.64	0.85
8	Conventions	25	108.8	80	125	76.1	61.6	84.1	100.0	0.68	0.56	0.81
	Evid/Elab	25	108.8	80	125	69.9	58.9	88.6	100.0	0.75	0.68	0.87
	Org/Purp	25	108.8	80	125	70.0	59.8	87.5	100.0	0.75	0.68	0.85
11	Conventions	25	96.4	80	109	74.0	62.8	85.1	100.0	0.69	0.57	0.84
	Evid/Elab	25	96.4	80	109	73.5	67.0	80.0	100.0	0.79	0.72	0.85
	Org/Purp	25	96.4	80	109	73.7	67.0	80.0	100.0	0.79	0.72	0.85

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

Table 51. Inter-Rater Agreement for Mathematics Items

Grade	**Score Point Range**	**Number of Items**	**Number of Responses**			**%Exact**			**%(Exact+ Adjacent)**	**QWK**[a]		
			Mean	**Min**	**Max**	**Mean**	**Min**	**Max**		**Mean**	**Min**	**Max**
3	0–1	5	149.6	128	164	93.2	87.6	97.7	100.0	NA	NA	NA
4	0–1	4	173.3	156	181	86.6	84.0	88.5	100.0	NA	NA	NA
5	0–1	6	114.7	60	127	92.0	76.7	97.6	100.0	NA	NA	NA
6	0–1	6	173.5	106	226	97.6	96.9	99.1	100.0	NA	NA	NA
7	0–1	5	236.2	194	275	97.8	95.0	100.0	100.0	NA	NA	NA
8	0–1	8	284.6	247	313	88.1	81.1	97.6	100.0	NA	NA	NA
11	0–1	5	180.0	120	209	94.3	85.6	98.9	100.0	NA	NA	NA
3	0–2	14	148.8	50	177	90.4	54.0	96.8	100.0	0.86	0.10	0.97
4	0–2	12	176.4	164	186	91.5	79.7	98.9	100.0	0.85	0.68	0.98
5	0–2	37	122.3	65	137	89.1	74.0	98.9	100.0	0.86	0.57	0.98
6	0–2	29	203.4	183	222	88.1	75.4	98.1	100.0	0.76	0.53	0.97
7	0–2	10	263.9	134	289	91.0	85.7	97.8	100.0	0.83	0.57	0.97
8	0–2	9	297.2	250	317	91.5	81.9	96.8	100.0	0.83	0.60	0.97
11	0–2	12	222.1	186	264	94.9	82.4	99.5	100.0	0.89	0.62	0.98
3	0-3	2	163.0	159	167	90.2	88.6	91.8	100.0	0.93	0.92	0.94
5	0-3	7	125.9	120	130	84.1	74.6	93.7	100.0	0.87	0.77	0.96
7	0-3	1	293.0	293	293	91.1	91.1	91.1	100.0	0.93	0.93	0.93
8	0-3	2	287.5	259	316	81.9	80.3	83.2	100.0	0.92	0.89	0.95
11	0-3	6	236.7	222	254	88.1	84.0	91.9	100.0	0.86	0.82	0.92

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

6.7 AUTOMATED SCORING

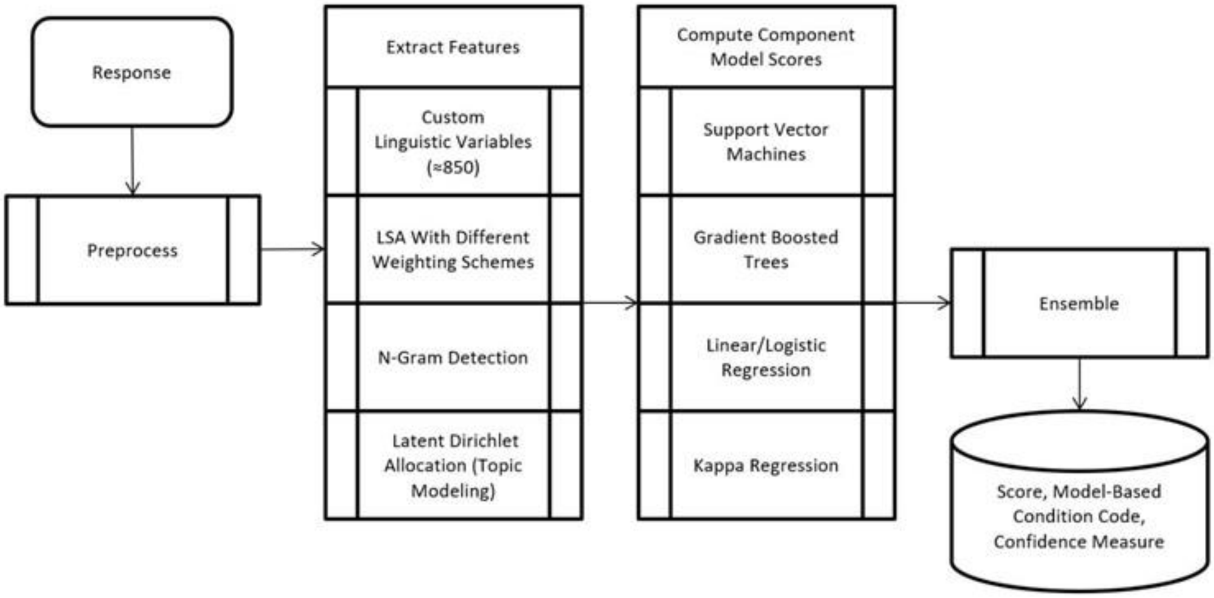
MI's Project Essay Grade (PEG) automated scoring technology was used to score eligible short-answer (SA) and essay items in ELA/L and SA items in mathematics. This section describes PEG, the training and validation sample and process, and the automated scoring process, concluding with the human-machine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

6.7.1 Project Essay Grade

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

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

Figure 13. PEG Architecture


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

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

6.7.2 Model Training and Validation

Sample

Automated scoring models were not created for items that had an insufficient quantity of training responses. This was this case for items with 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. Table 52 shows that pretrained models existed for 595 items, thus, no additional training was conducted in preparation for the spring 2024 administration. The remainder of this section describes the process used to train and validate the 595 existing models.

Table 52. Number of Items Eligible for Automated Scoring, by Grade and Subject Area

Grade	Items With Existing Models			Items Without Models		
	ELA/L		Mathematics	ELA/L		Mathematics
	Short-Answer	Essay		Short-Answer	Essay	
3	12	13	44	0	0	0
4	13	16	42	0	0	0
5	13	10	50	0	0	0
6	32	10	41	0	0	0
7	45	17	15	0	0	0
8	49	14	24	0	0	0
11	80	17	38	0	0	0
Total	244	97	254	0	0	0

Training Data

Student responses used for training and validation were sourced from the 2018–2019, 2020–2021, 2021–2022, and 2022–2023 Smarter Balanced operational test administrations. Responses were 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 the score assigned to each response by an expert rater.

For each item, the sample was divided as follows:

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

Model Training

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 the content varies significantly from item to item, and therefore from model to model, PEG examines training responses and identifies the variables that most accurately capture the content in question. To do this, MI uses techniques like LSA, N-Gram Detection, and LDA. To further refine the variable generation process, MI built a computer language to perform a simultaneous search over semantic, lexicographic and syntactic features of responses.

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

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 across items or traits, allowing PEG to faithfully reproduce humans' application of the scoring rubrics. This means that the resultant models are reasonably robust to gaming attempts, as each represents a unique valuation of the item- (or trait-) specific text features similarly valued by expert professional raters.

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

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

Model Validation

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

Initial Validation

Initial validation was conducted by applying each model to score a respective validation set of responses. The validation set is independent of the training set, in that none of the responses it contains have been used to build the model. Two or more professional raters will not always agree on what score to give a student's response; therefore, modeling is considered successful when the engine produces scores that agree with professional raters to the same or greater extent than the raters agree with each other. The initial evaluation was made using the criteria shown in Table 53, based on criteria proposed by Williamson, Xi, and Breyer (2012). While Williamson et al. (2012) recommend an agreement between human and machine scores of 0.70 quadratic weighted kappa (QWK) for normally distributed data, a QWK threshold of 0.65 was adopted due to the prevalence of skewed distributions in response data. The degradation (QWK) criterion of .07 is slightly more stringent than proposed by Williamson et al. (2012). The evaluation process was used for both the item-specific scoring models and the condition code models.

Table 53. Initial Model Evaluation Criteria

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

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

Bias Considerations. Subgroup differences in responses to constructed response items can introduce construct-irrelevant variance in scores, in turn threatening valid score interpretations. MI investigated potential sources of bias annually, for newly modeled items, as part of the initial validation process using available data from previous summative administration. Table 54 shows the demographic variables and categories considered. MI received separate datafiles containing (1) hand-score data and (2) student demographic data associated with responses.

Table 54. Demographic Variables and Categories

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

For each new item being modeled, analysis was performed on a subgroup if the number of observations (i.e., human-machine 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, excluded from automated scoring, and hand-scored if any subgroup comparison associated with the item was flagged.

Secondary Validation

All models associated with items that passed initial validation were subject to a secondary validation at the start of the spring 2024 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 hand-scored. Table 55 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 55. 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 56 presents the secondary validation results. Of the 595 items with models subject to secondary validation, models associated with 454 of the items (76.3%) passed all secondary evaluation criteria.

Table 56. 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	12	13	44	12	3	44
4	13	16	42	13	6	40
5	13	10	50	13	5	47
6	32	10	41	19	5	40
7	45	17	15	27	9	15
8	49	14	24	31	9	22
11	80	17	38	46	10	38
Total	244	97	254	161	47	246

Live Training and Validation

Additionally, in April-May 2024 when operational scoring was underway, a live training and validation effort was undertaken for those hand-scored items lacking validated models from prior efforts but having sufficient 2024 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 2024 operational responses offered potential to improve model performance. All models associated with these items were thus trained using either exclusively 2024 responses (when a minimum of 1,400 2024 responses/item existed) or 2024 responses supplemented with 2023 responses. In either case, the validation sets consisted exclusively of 2024 responses. Because live validation involved operational data, it was unnecessary to conduct a secondary validation.

Table 57 summarizes the results of the live training and validation. Of the 356 items associated with models that underwent live training and validation, models associated with 211 of the items (59.3%) passed all evaluation criteria. While this pass rate is considerably lower than the pass rates during secondary (76.3%) 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 57. 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	25	9	1	16	4
4	3	24	9	3	19	1
5	1	25	33	1	14	19
6	24	16	10	15	10	4
7	28	20	7	18	12	4
8	26	25	9	17	6	7
11	36	21	4	24	12	4
Total	119	156	81	79	89	43

Following initial validation, secondary validation, and live training and validation, a total of 665 items, comprised of 240 ELA/L SA, 136 essay, and 289 mathematics SA, were scored using a hybrid process, described next.

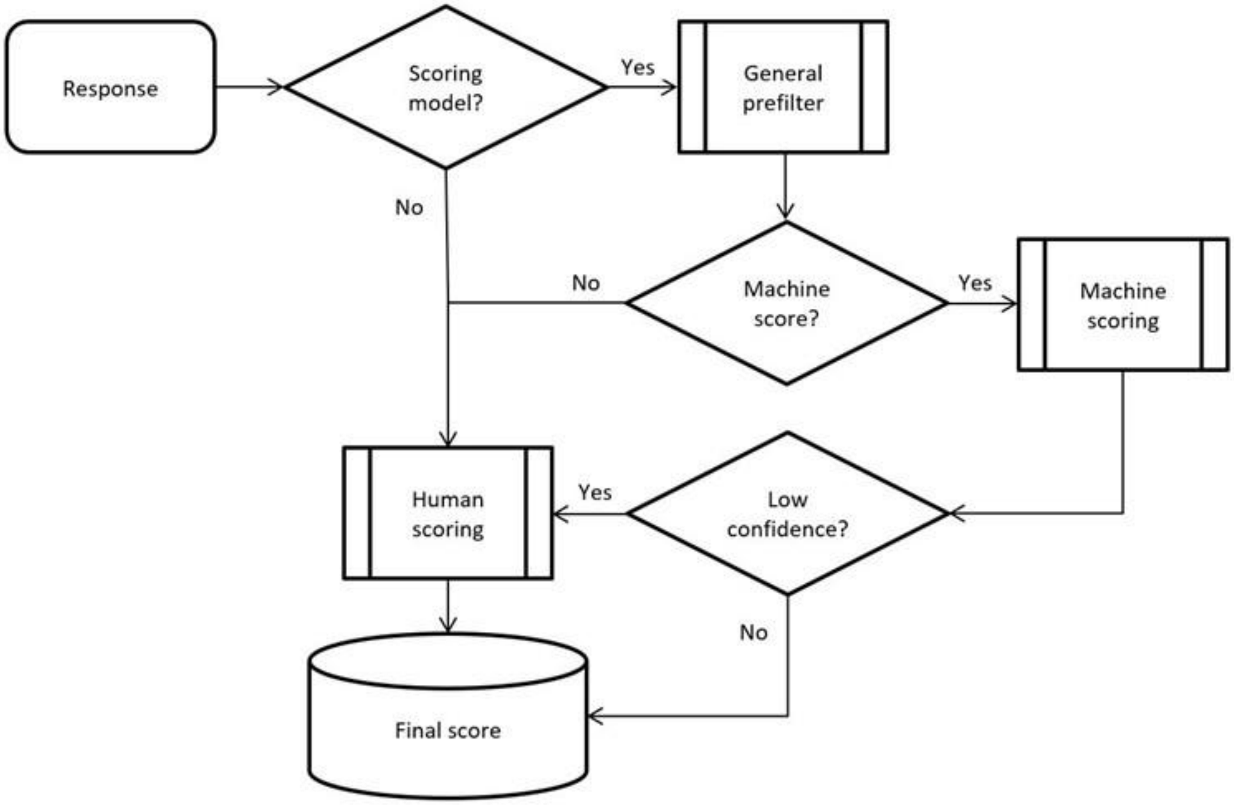
6.7.3 Automated Scoring Processes

Hybrid Scoring Process

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

Figure 14 shows the response routing rules under the hybrid scoring process. In the hybrid model, responses with associated scoring models were first pre-processed for automated scoring and to filter alert responses and certain non-scorable cases (e.g., insufficient text to score or high proportion of copied prompt text). Flags were used to indicate condition codes as defined in the hand-scoring criteria (see Table 58 and Table 59). 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 were extracted, the scoring model(s) applied, and responses assigned a score and measure of score confidence. Low-confidence responses straddle the lines between score point values on a rubric and are difficult to score accurately because they exhibit characteristics of multiple score points Higher-confidence responses received the engine score as the score of record, while lower-confidence responses were routed directly to expert raters, who assigned the score of record. Note that the expert rater pool was dynamic, and raters were added or removed several times each day based on their current performance. Overall, approximately 15% of responses to engine-scored items were flagged as low confidence and scored by expert raters.

Figure 14. Response Routing Rules



Upon receipt and validation of each response, MI routed responses for those items eligible for automated scoring to PEG and the remainder of the responses to the VSC hand-scoring system.

Table 58. Flags Currently Established

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

FLAG	USAGE DESCRIPTION	*SCORABLE
660	Language Non-English Condition Code (L): Response holds strong general resemblance to those marked 'Non-English' by human readers, but is nonetheless PEG scorable (and, so scores are provided). *PEG Configuration*: Item agnostic; but for 2021 onwards, applicable to ELA/L items only.	YES
670	Off-Topic: Applicable to ELA/L essays only and is item specific in the PEG environment.	YES
680	Off-Mode: Applicable to ELA/L essays only and is item specific in the PEG environment.	YES
900	Timeout (i.e., unable to complete essay score prediction within time limits). Not client configurable.	NO
950	System error processing essay (i.e., internal PEG error). Not client configurable.	NO

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

Table 59. Model Setting

TYPE	ASSOCIATED FLAG(S)	DESCRIPTION	VALUES
Minimum Words	200	Triggers if there are fewer than the associated value of word-tokens in a response. The flag may also appear regardless of setting if the response is blank.	0-15
Alert	500 501-599	Current setting (PREDC...1) is for the standard alert scan.	Standard settings in place
Plagiarism	620	Prompt and source material text is included in model configuration.	50% of prompt and source material characters triggers flag

Scoring Infrastructure

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

Quality Assurance

MI’s hybrid scoring approach included numerous quality assurance steps. First, models were trained using exclusively scores assigned by expert raters and the associated responses. Second, each automated scoring model was subjected to an evaluation process, as described in the model validation section. This involved evaluating the quality of the human-scored training data, as well as comparing the performance of the engine to the performance of expert raters. Third, for models trained using responses from prior administrations, the generalizability of each model to the 2023-24 operational responses was confirmed via a secondary validation. Finally, quality was further assured during scoring by routing a minimum of 15% of the responses that were most different from the training responses to expert raters and assigning the human score.

“Alert” Procedures

MI implemented a 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.

6.7.4 PEG-Human Agreement

This section summarizes the human-machine agreement for all items scored using a hybrid process in spring 2024, including (1) items passing initial model validation, (2) items passing secondary validation, and (3) items passing live validation.

Tables 60 through 62 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. For the PEG-scored items, the human-machine agreement was computed based on the combined data across all states with hybrid scoring in the 2023–2024 summative assessment.

Table 60. Human-Machine Agreement for ELA/L Short-Answer Items on Initial and Secondary Validation Samples, by Grade

Grade	Initial Validation				Secondary Validation			
	Number of Items	% Exact	%(Exact+ Adjacent)	QWK	Number of Items	% Exact	%(Exact+ Adjacent)	QWK
3	12	79.6	99.6	0.81	12	82.3	99.5	0.77
4	13	80.1	99.8	0.84	13	80.9	99.8	0.80
5	13	75.4	99.6	0.81	13	77.4	99.8	0.78
6	19	78.7	99.5	0.81	19	79.1	99.6	0.77
7	27	76.3	99.4	0.79	27	76.4	99.4	0.75
8	31	76.2	99.5	0.78	31	75.8	99.4	0.75
11	46	77.2	99.5	0.79	46	76.1	99.5	0.77

Table 61. Human-Machine Agreement for ELA/L Essay Items on Initial and Secondary Validation Samples, by Grade

Grade	Trait	Initial Validation				Secondary Validation			
		Number of Items	% Exact	%(Exact + Adjacent)	QWK	Number of Items	% Exact	% (Exact + Adjacent)	QWK
3	Conventions	3	71.6	99.7	0.72	3	72.5	99.5	0.70
3	Evid/Elab	3	77.9	99.2	0.82	3	78.2	99.7	0.77
3	Org/Purp	3	75.0	99.7	0.8	3	79.1	99.6	0.78
4	Conventions	6	69.2	99.0	0.74	6	69.7	99.3	0.74
4	Evid/Elab	6	73.6	99.5	0.84	6	73.5	99.1	0.79
4	Org/Purp	6	72.2	99.2	0.82	6	74.2	99.2	0.79
5	Conventions	5	72.5	99.6	0.71	5	73.0	99.6	0.72
5	Evid/Elab	5	73.0	99.0	0.82	5	72.6	99.6	0.80
5	Org/Purp	5	72.2	99.6	0.83	5	72.7	99.6	0.80
6	Conventions	5	75.5	99.0	0.72	5	73.5	99.5	0.74
6	Evid/Elab	5	71.4	98.7	0.78	5	76.2	99.6	0.78
6	Org/Purp	5	69.8	98.9	0.78	5	76.2	99.6	0.78
7	Conventions	9	76.1	99.7	0.70	9	75.5	99.8	0.74
7	Evid/Elab	9	75.6	99.7	0.83	9	81.7	99.8	0.84
7	Org/Purp	9	75.6	99.6	0.84	9	81.6	99.9	0.84
8	Conventions	9	77.0	99.1	0.71	9	76.1	99.7	0.74
8	Evid/Elab	9	73.7	99.1	0.82	9	76.9	99.6	0.80
8	Org/Purp	9	75.1	99.7	0.84	9	77.2	99.6	0.80
11	Conventions	10	79.1	99.7	0.75	10	77.1	99.6	0.73
11	Evid/Elab	10	76.5	99.7	0.86	10	75.6	99.9	0.84
11	Org/Purp	10	76.4	99.7	0.86	10	75.8	99.9	0.83

Table 62. Human-Machine Agreement for Mathematics Items on Initial and Secondary Validation Samples, by Grade

Grade	**Score Point Range**	**Initial Validation**				**Secondary Validation**			
		Number of Items	**% Exact**	**%(Exact+ Adjacent)**	**QWK**	**Number of Items**	**% Exact**	**%(Exact+ Adjacent)**	**QWK**[a]
3	0-1	10	94.2	100	0.86	10	94.1	100.0	NA
4	0-1	7	91.0	100	0.79	7	92.3	100.0	NA
5	0-1	7	92.6	100	0.81	7	93.5	100.0	NA
6	0-1	8	96.6	100	0.81	8	95.8	100.0	NA
7	0-1	7	96.9	100	0.85	7	96.8	100.0	NA
8	0-1	5	90.2	100	0.75	5	90.5	100.0	NA
11	0-1	16	95.6	100	0.87	16	94.2	100.0	NA
3	0-2	28	90.8	99.3	0.91	28	90.6	99.4	0.89
4	0-2	29	91.0	99.7	0.91	29	91.6	99.7	0.89
5	0-2	38	88.3	99.6	0.88	38	87.9	99.5	0.84
6	0-2	32	88.9	99.6	0.86	32	89.1	99.5	0.84
7	0-2	8	87.0	99.4	0.80	8	88.9	99.9	0.8
8	0-2	16	89.1	99.8	0.89	16	90.3	99.7	0.86
11	0-2	17	89.1	99.4	0.88	17	88.1	99.4	0.87
3	0-3	6	91.1	99.8	0.96	6	92.5	99.9	0.96
4	0-3	4	87.9	99.8	0.94	4	86.8	99.6	0.93
5	0-3	2	90.8	98.4	0.94	2	89.4	98.3	0.90
8	0-3	1	78.2	98.0	0.88	1	86.1	98.4	0.92
11	0-3	5	85.5	99.0	0.89	5	83.7	99.0	0.88

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

Tables 63 through 65 present the human-machine 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 2023-24 operational data were used to build the models.

Table 63. Human-Machine Agreement for ELA/L Short-Answer Items on Live Validation Sample, by Grade

Grade	**Live Validation**			
	Number of Items	**% Exact**	**%(Exact+ Adjacent)**	**QWK**
3	1	73.8	99.3	0.66
4	3	79.7	99.7	0.81
5	1	70.4	98.0	0.73
6	15	77.6	99.5	0.73
7	18	78.5	99.7	0.74
8	17	76.1	99.6	0.74
11	24	76.5	99.6	0.77

Table 64. Human-Machine Agreement for ELA/L Essay Items on Live Validation Sample, by Grade

Grade	Trait	Live Validation			
		Number of Items	% Exact	%(Exact+ Adjacent)	QWK
3	Conventions	16	70.5	99.6	0.71
3	Evid/Elab	16	73.4	98.8	0.77
3	Org/Purp	16	72.8	99.0	0.77
4	Conventions	19	69.4	99.2	0.73
4	Evid/Elab	19	72.2	98.9	0.78
4	Org/Purp	19	73.0	99.2	0.79
5	Conventions	14	70.8	99.5	0.70
5	Evid/Elab	14	70.1	99.0	0.78
5	Org/Purp	14	70.2	99.1	0.79
6	Conventions	10	73.2	99.4	0.72
6	Evid/Elab	10	73.6	99.3	0.79
6	Org/Purp	10	74.0	99.4	0.79
7	Conventions	12	71.5	99.6	0.72
7	Evid/Elab	12	74.6	99.4	0.80
7	Org/Purp	12	74.8	99.4	0.81
8	Conventions	6	76.7	99.6	0.72
8	Evid/Elab	6	76.9	99.8	0.84
8	Org/Purp	6	74.8	99.8	0.83
11	Conventions	12	75.8	99.5	0.73
11	Evid/Elab	12	76.0	99.7	0.84
11	Org/Purp	12	76.2	99.8	0.84

Table 65. Human-Machine Agreement for Mathematics Items on Live Validation Samples, by Grade

Grade	Score Point Range	Live Validation			
		Number of Items	% Exact	%(Exact+ Adjacent)	QWK[a]
3	0-1	3	94.4	100.0	NA
4	0-1	1	88.7	100.0	NA
5	0-1	4	95.4	100.0	NA
6	0-1	1	91.4	100.0	NA
7	0-1	1	100	100.0	NA
8	0-1	3	87.8	100.0	NA
3	0-2	1	100	100.0	1.00
5	0-2	14	84.1	99.4	0.82
6	0-2	3	87.3	99.2	0.81
7	0-2	3	90.1	99.1	0.88
8	0-2	3	92.3	100.0	0.92
11	0-2	3	97.6	100.0	0.98
5	0-3	1	88.3	98.7	0.91
8	0-3	1	72.2	97.0	0.89
11	0-3	1	90.2	98.8	0.89

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

6.7.5 Recommendations

The 2023 administrations highlighted the importance of expanding automated monitoring and implementing further interventions to maximize score quality. Building on this, the 2024 administration successfully broadened the additional rater validation stage—originally introduced in 2023 for brief write and research rater qualification—to encompass all ELA/L item types. Furthermore, validity-based measures of scoring accuracy were refined in 2024 to include a comparison of mean score differences between the distributions of benchmark and rater-assigned scores in addition to the previously utilized agreement (QWK). This enhancement provided a more nuanced and sensitive measure of rater quality, ensuring that scoring accuracy is maintained at a high standard.

Despite these improvements, the primary challenge faced during the spring 2024 administration was related to rater productivity, with raters not meeting the expected number of working hours projected from 2023. This issue became particularly evident in April and May, leading to bottlenecks, especially in the scoring of full write and brief write responses, which are time-consuming to train for and score accurately. In response, additional raters were recruited, and pay incentives were offered in key production bottleneck areas. However, some responses still experienced delays in scoring. To address these challenges for the 2025 administration, it is recommended to develop a core pool of full-time raters, establish a minimum work commitment for part-time raters, and collect a measure of rater quality earlier, ideally during qualification. Additionally, surveying raters on their availability and work preferences, as well as enhancing the rater management system, will be crucial steps in improving rater productivity and maintaining the quality and timeliness of scoring.

Furthermore, a review of the scoring outcomes revealed that while the mean QWK values for inter-rater agreement generally met expectations, there were concerns regarding the relatively low minimum QWKs observed for some ELA/L short-answer items, as indicated by the minimum QWK values in Table 49. These low QWK values suggest variability in rater agreement for certain items, which could undermine the overall reliability of the scoring process. To address this issue, it is recommended that additional targeted training and calibration sessions be conducted for raters assigned to items with historically low QWK values. This could include additional focused trainings on interpreting and applying scoring rubrics for those items, the development of supplemental materials, as well as implementing more frequent monitoring and feedback loops during the scoring process.

7. REPORTING AND INTERPRETING SCORES

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

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

7.1 CENTRALIZED REPORTING SYSTEM

The CRS is designed to help educators, families, and students answer questions about how well students have performed on the English language arts/literacy (ELA/L) and mathematics ISAT assessments. The CRS provides all stakeholders with timely, relevant score reports. The CRS is designed to provide score reports that are understandable to all stakeholders. Available reports use plain, non-technical language to facilitate review by parents/families 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 to help readers 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 66 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 for navigating the online reporting system can be found in the *Centralized Reporting System User Guide*, embedded within the CRS via a Help button.

Table 66. Types of Online Score Reports by Level of Aggregation

Level of Aggregation	Types of Online Score Reports
State District School Teacher Roster	• Number of students tested and percentage of students Proficient (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 by subgroup) • Student growth in scale score and achievement level over time • On-demand student roster report
Student	• Total scale score and standard error of measurement • Achievement level on overall score with achievement-level descriptors • Average scale scores and standard errors of average scale scores for student's school, district, and state • Student growth in scale score and achievement level over time • 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 67 presents the types of subgroup and subgroup categories provided in the CRS.

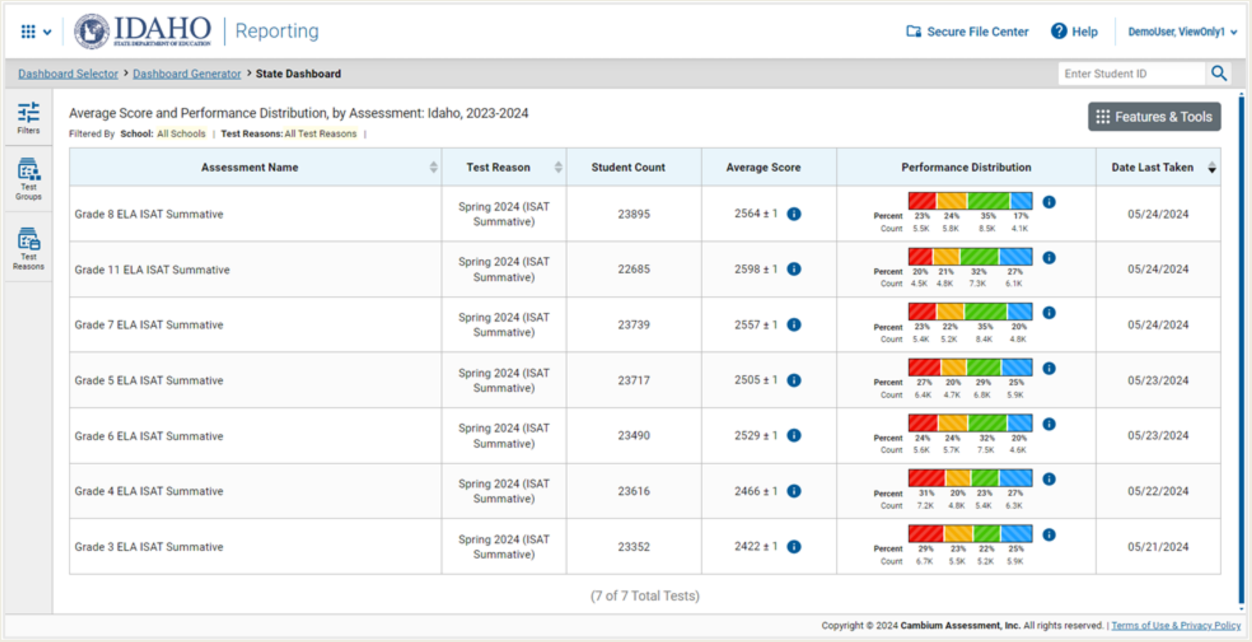
Table 67. Types of Subgroups with Subgroup Categories

Subgroup	Subgroup Category
Gender	Female Male
Special Education Status	Yes No
EL Status	Yes No
EL Category	L1, LE, EW, X1, X2, X3, X4, FL, SO
Section 504 Status	Yes No
Race/Ethnicity	American Indian/Alaskan Native Black or African American Asian Hispanic or Latino Native Hawaiian or Other Pacific Islander White

7.1.1 Dashboard

The CRS provides a state dashboard for authorized state-level users to track student performance for all tests in all grades 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 an example dashboard page at the state level.

Exhibit 1. Dashboard: State Level



Dashboard Selector > Dashboard Generator > State Dashboard

Average Score and Performance Distribution, by Assessment: Idaho, 2023-2024

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

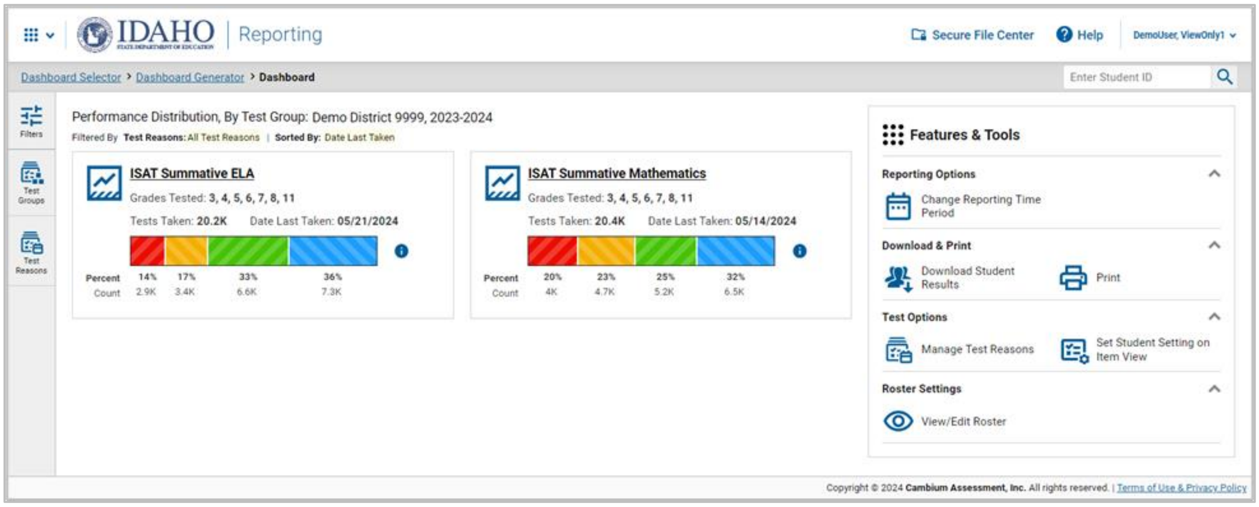
Assessment Name	Test Reason	Student Count	Average Score	Performance Distribution	Date Last Taken
Grade 8 ELA ISAT Summative	Spring 2024 (ISAT Summative)	23895	2564 ± 1	Percent 23% 24% 35% 17% Count 5.5K 5.8K 8.5K 4.1K	05/24/2024
Grade 11 ELA ISAT Summative	Spring 2024 (ISAT Summative)	22685	2598 ± 1	Percent 20% 21% 32% 27% Count 4.5K 4.8K 7.3K 6.1K	05/24/2024
Grade 7 ELA ISAT Summative	Spring 2024 (ISAT Summative)	23739	2557 ± 1	Percent 23% 22% 35% 20% Count 5.4K 5.2K 8.4K 4.8K	05/24/2024
Grade 5 ELA ISAT Summative	Spring 2024 (ISAT Summative)	23717	2505 ± 1	Percent 27% 20% 29% 25% Count 6.4K 4.7K 6.8K 5.9K	05/23/2024
Grade 6 ELA ISAT Summative	Spring 2024 (ISAT Summative)	23490	2529 ± 1	Percent 24% 24% 32% 20% Count 5.6K 5.7K 7.5K 4.6K	05/23/2024
Grade 4 ELA ISAT Summative	Spring 2024 (ISAT Summative)	23616	2466 ± 1	Percent 31% 20% 23% 27% Count 7.2K 4.8K 5.4K 6.3K	05/22/2024
Grade 3 ELA ISAT Summative	Spring 2024 (ISAT Summative)	23352	2422 ± 1	Percent 29% 23% 22% 25% Count 6.7K 5.5K 5.2K 5.9K	05/21/2024

(7 of 7 Total Tests)

Copyright © 2024 **Cambium Assessment, Inc.** All rights reserved. | Terms of Use & Privacy Policy

Upon logging into the CRS, each authorized user, regardless of role (e.g., district, school, or teacher), will see a dashboard page displaying the overall test results for all tests students have taken grouped by test family (e.g., ISAT Summative Mathematics). The dashboard summarizes students' performance by test family for both ELA/L and mathematics across all grades, including (1) the grades of the students who have tested, (2) the number of tests taken, (3) the test date last taken, and (4) the percentage and counts of students at each achievement level. District personnel see district summaries, school personnel see school summaries, and teachers see summaries of their students. Exhibit 2 presents an example dashboard page at the district level.

Exhibit 2. Dashboard: District Level



IDAHO | Reporting

Secure File Center | Help | DemoUser, ViewOnly1

Dashboard Selector > Dashboard Generator > Dashboard

Enter Student ID

Filters | Test Groups | Test Reasons

Performance Distribution, By Test Group: Demo District 9999, 2023-2024

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

ISAT Summative ELA

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

Tests Taken: **20.2K** Date Last Taken: **05/21/2024**

Percent	14%	17%	33%	36%
Count	2.9K	3.4K	6.6K	7.3K

ISAT Summative Mathematics

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

Tests Taken: **20.4K** Date Last Taken: **05/14/2024**

Percent	20%	23%	25%	32%
Count	4K	4.7K	5.2K	6.5K

Features & Tools

Reporting Options
- Change Reporting Time Period

Download & Print
- Download Student Results
- Print

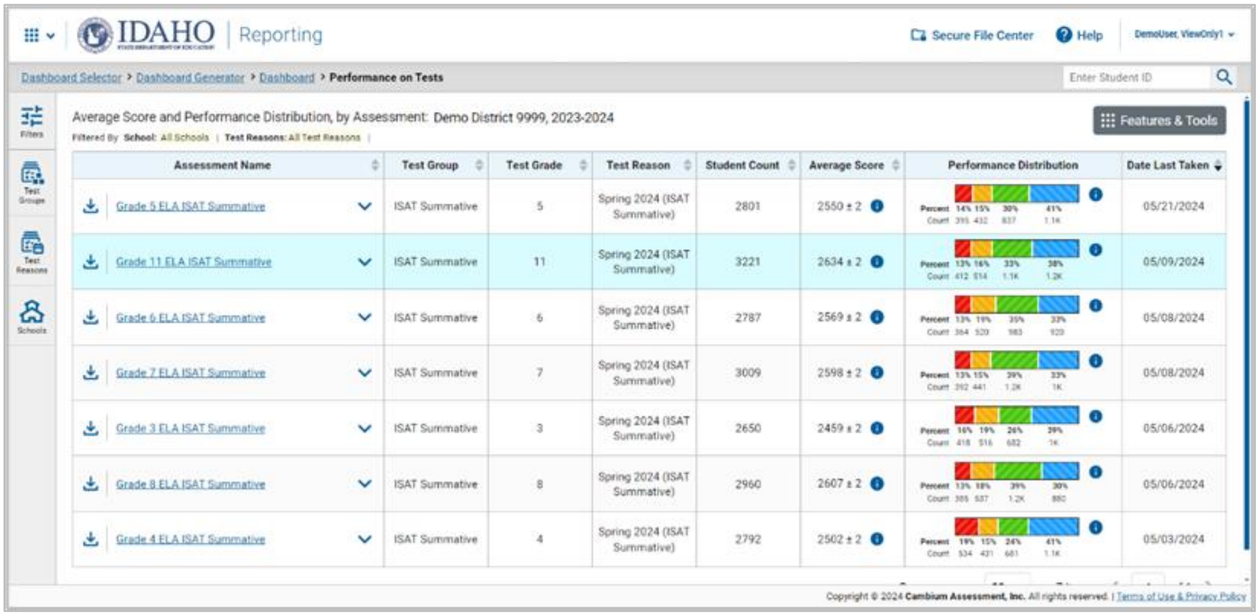
Test Options
- Manage Test Reasons
- Set Student Setting on Item View

Roster Settings
- View/Edit Roster

Copyright © 2024 **Cambium Assessment, Inc.** All rights reserved. | Terms of Use & Privacy Policy

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

Exhibit 3. Detailed Dashboard: District Level



IDAHO | Reporting

Secure File Center | Help | DemoUser, ViewOnly1

Dashboard Selector > Dashboard Generator > Dashboard > Performance on Tests

Enter Student ID

Filters | Test Groups | Test Reasons | Schools

Average Score and Performance Distribution, by Assessment: Demo District 9999, 2023-2024

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

Features & Tools

Assessment Name	Test Group	Test Grade	Test Reason	Student Count	Average Score	Performance Distribution	Date Last Taken
Grade 5 ELA ISAT Summative	ISAT Summative	5	Spring 2024 (ISAT Summative)	2801	2550 ± 2	Percent 14% 15% 30% 41% Count 395 432 837 1.1K	05/21/2024
Grade 11 ELA ISAT Summative	ISAT Summative	11	Spring 2024 (ISAT Summative)	3221	2634 ± 2	Percent 13% 16% 33% 38% Count 412 514 1.1K 1.2K	05/09/2024
Grade 6 ELA ISAT Summative	ISAT Summative	6	Spring 2024 (ISAT Summative)	2787	2569 ± 2	Percent 13% 19% 35% 33% Count 364 520 983 920	05/08/2024
Grade 7 ELA ISAT Summative	ISAT Summative	7	Spring 2024 (ISAT Summative)	3009	2598 ± 2	Percent 13% 15% 39% 33% Count 392 441 1.2K 1K	05/08/2024
Grade 3 ELA ISAT Summative	ISAT Summative	3	Spring 2024 (ISAT Summative)	2650	2459 ± 2	Percent 16% 19% 26% 39% Count 418 516 682 1K	05/06/2024
Grade 8 ELA ISAT Summative	ISAT Summative	8	Spring 2024 (ISAT Summative)	2960	2607 ± 2	Percent 13% 18% 39% 30% Count 385 537 1.2K 880	05/06/2024
Grade 4 ELA ISAT Summative	ISAT Summative	4	Spring 2024 (ISAT Summative)	2792	2502 ± 2	Percent 19% 15% 24% 41% Count 534 431 681 1.1K	05/03/2024

Copyright © 2024 **Cambium Assessment, Inc.** All rights reserved. | Terms of Use & Privacy Policy

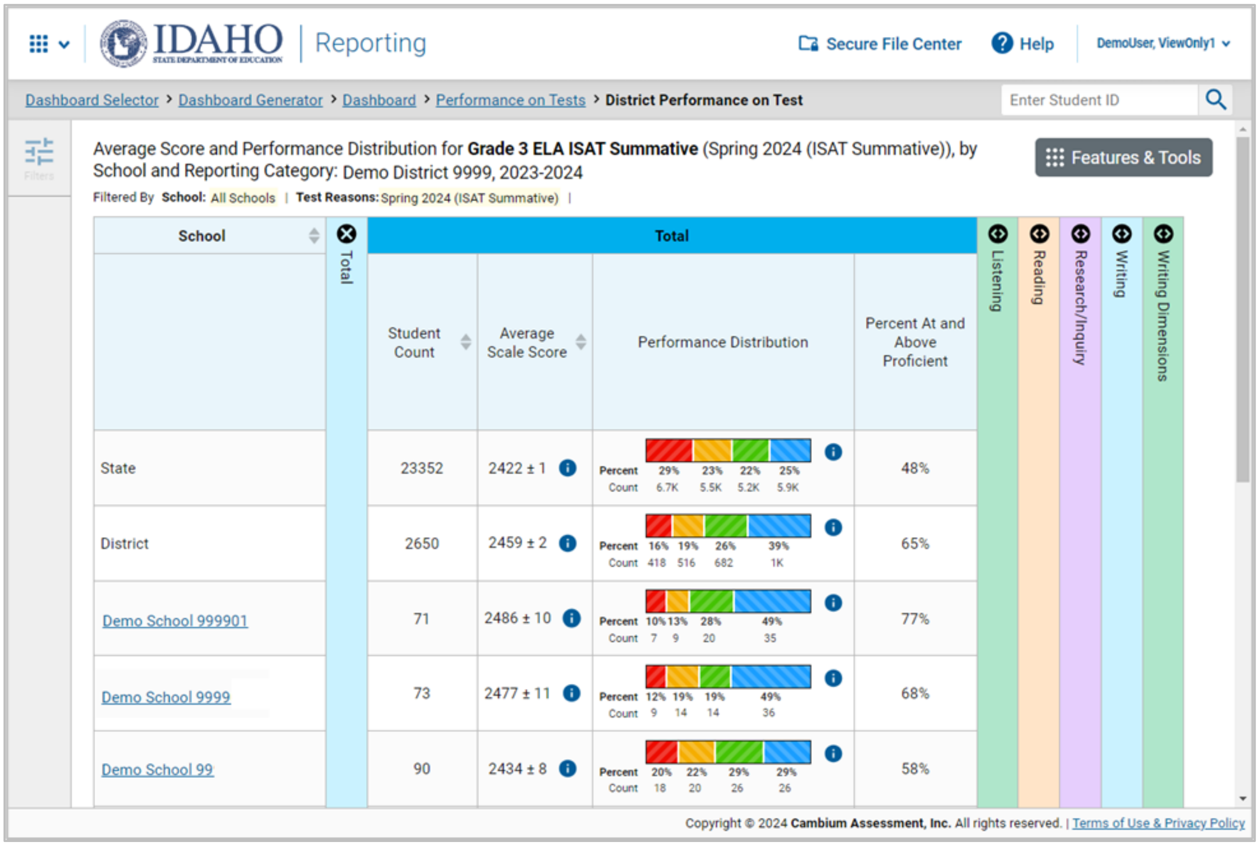
7.1.2 Aggregate Score Reports: Overall Performance

When users select a specific assessment name (e.g., Grade 3 ELA/L ISAT Summative) from the detailed dashboard, they will see a summary of student performance on the chosen assessment for a selected aggregate unit (e.g., district, school, roster). On each aggregate report, the summary report presents the summary results for the selected aggregate unit, the summary results for the state, and the summary for the aggregate unit both above and below the selected aggregate. For example, if a district is selected, the summary results of the state and individual schools within the district are provided as well as the district summary results so that district performance can be compared with the other aggregate levels.

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

Exhibit 4 presents an example overall performance summary result for grade 3 ELA/L at the district level. Exhibit 5 presents an example summary by gender at the district level.

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


IDAHO
STATE DEPARTMENT OF EDUCATION
Reporting
Secure File Center
Help
DemoUser, ViewOnly1
Dashboard Selector > Dashboard Generator > Dashboard > Performance on Tests > District Performance on Test
Enter Student ID
Filters
Features & Tools

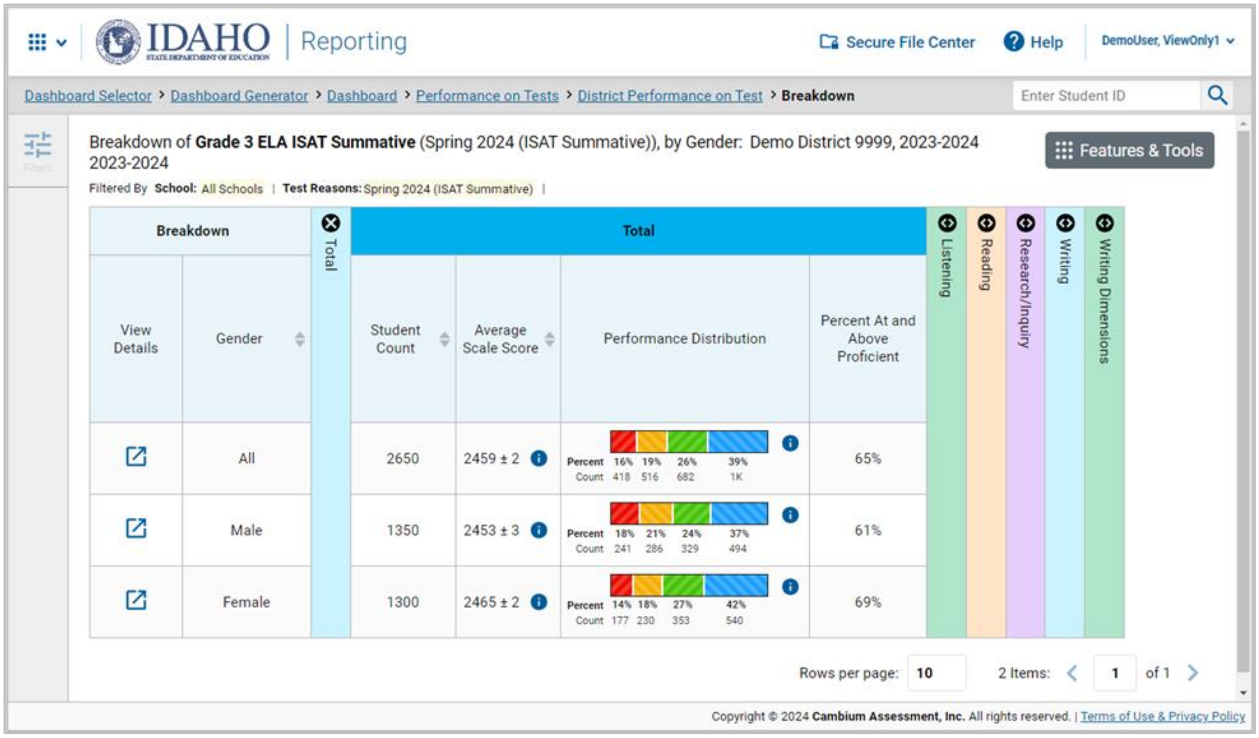
Average Score and Performance Distribution for **Grade 3 ELA ISAT Summative** (Spring 2024 (ISAT Summative)), by School and Reporting Category: Demo District 9999, 2023-2024

Filtered By **School:** All Schools | **Test Reasons:** Spring 2024 (ISAT Summative) |

School	Total					Listening	Reading	Research/Inquiry	Writing	Writing Dimensions
		Student Count	Average Scale Score	Performance Distribution	Percent At and Above Proficient					
State		23352	2422 ± 1	Percent 29% 23% 22% 25% Count 6.7K 5.5K 5.2K 5.9K	48%					
District		2650	2459 ± 2	Percent 16% 19% 26% 39% Count 418 516 682 1K	65%					
Demo School 999901		71	2486 ± 10	Percent 10% 13% 28% 49% Count 7 9 20 35	77%					
Demo School 9999		73	2477 ± 11	Percent 12% 19% 19% 49% Count 9 14 14 36	68%					
Demo School 99		90	2434 ± 8	Percent 20% 22% 29% 29% Count 18 20 26 26	58%					

Copyright © 2024 **Cambium Assessment, Inc.** All rights reserved. | Terms of Use & Privacy Policy

Exhibit 5. Overall Performance Summary Results for Grade 3 ELA/L by Gender: District Level


IDAHO STATE DEPARTMENT OF EDUCATION
Reporting
Secure File Center
Help
DemoUser, ViewOnly1
Dashboard Selector > Dashboard Generator > Dashboard > Performance on Tests > District Performance on Test > Breakdown
Enter Student ID
Filters
Features & Tools
Listening
Reading
Research/Inquiry
Writing
Writing Dimensions

Breakdown of **Grade 3 ELA ISAT Summative** (Spring 2024 (ISAT Summative)), by Gender: Demo District 9999, 2023-2024 2023-2024

Filtered By **School:** All Schools | **Test Reasons:** Spring 2024 (ISAT Summative) |

Breakdown		Total	Total			
View Details	Gender		Student Count	Average Scale Score	Performance Distribution	Percent At and Above Proficient
	All		2650	2459 ± 2	Percent 16% 19% 26% 39% Count 418 516 682 1K	65%
	Male		1350	2453 ± 3	Percent 18% 21% 24% 37% Count 241 286 329 494	61%
	Female		1300	2465 ± 2	Percent 14% 18% 27% 42% Count 177 230 353 540	69%

Rows per page: 10 2 Items: < 1 of 1 >

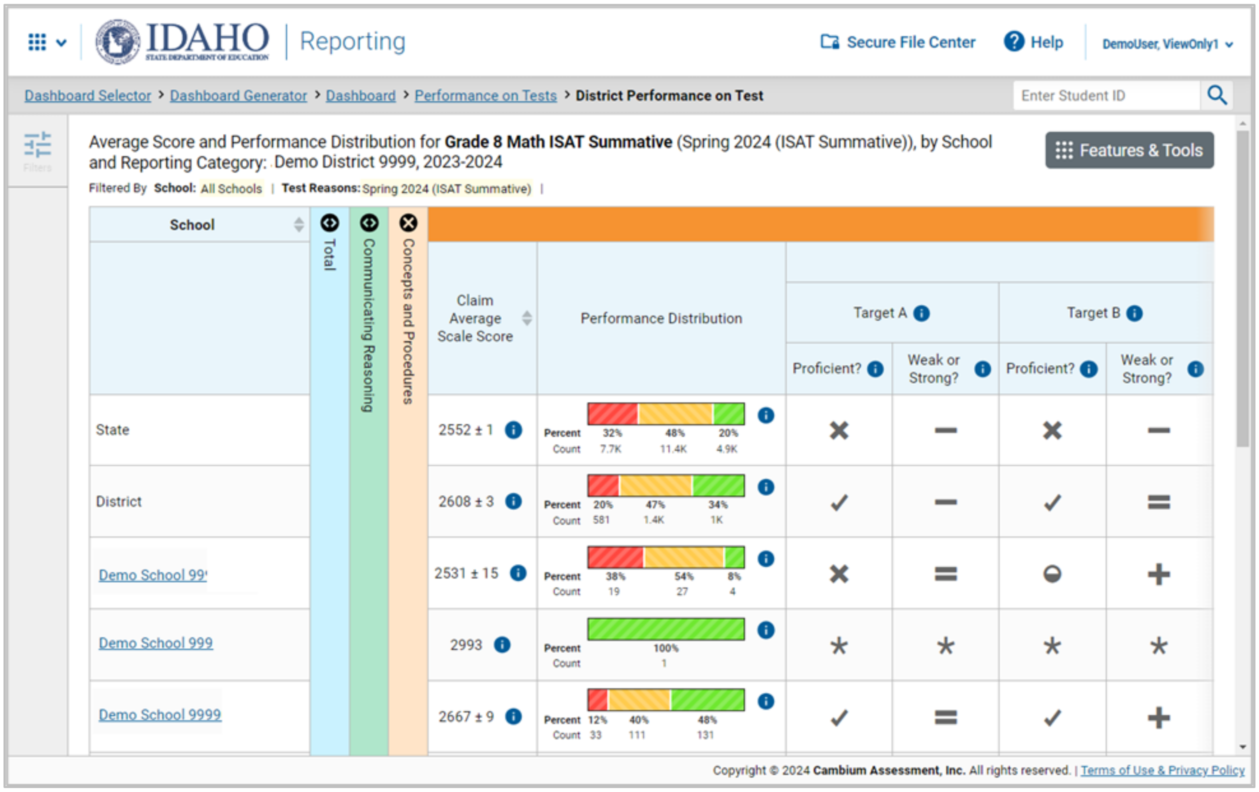
Copyright © 2024 **Cambium Assessment, Inc.** All rights reserved. | Terms of Use & Privacy Policy

7.1.3 Aggregate Score Reports: Claim and Target Performance

On the same report page, detailed summaries on aggregated claim and target results are also available. The claim and target results can be accessed by clicking a claim (e.g., listening, reading) on the right side of the page.. 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 (check mark), less than (x mark), or not different from (half-filled circle) the proficiency standard for the selected test. The "Weak or Strong?" measure presents whether the group's performance on each target is lower than (minus sign), higher than (plus sign), or not different from (equal sign) the group's overall performance. If there is insufficient information in the "Proficient?" measure or "Weak or Strong?" measure, this is indicated with a star sign (*).

Like the overall performance summary results, the summary report presents results for the selected aggregate unit, 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 an example of claim- and target-level results for grade 8 mathematics at the district level.

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



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

Average Score and Performance Distribution for **Grade 8 Math ISAT Summative** (Spring 2024 (ISAT Summative)), by School and Reporting Category: Demo District 9999, 2023-2024

Filtered By **School:** All Schools | **Test Reasons:** Spring 2024 (ISAT Summative) |

Total | Communicating Reasoning | Concepts and Procedures

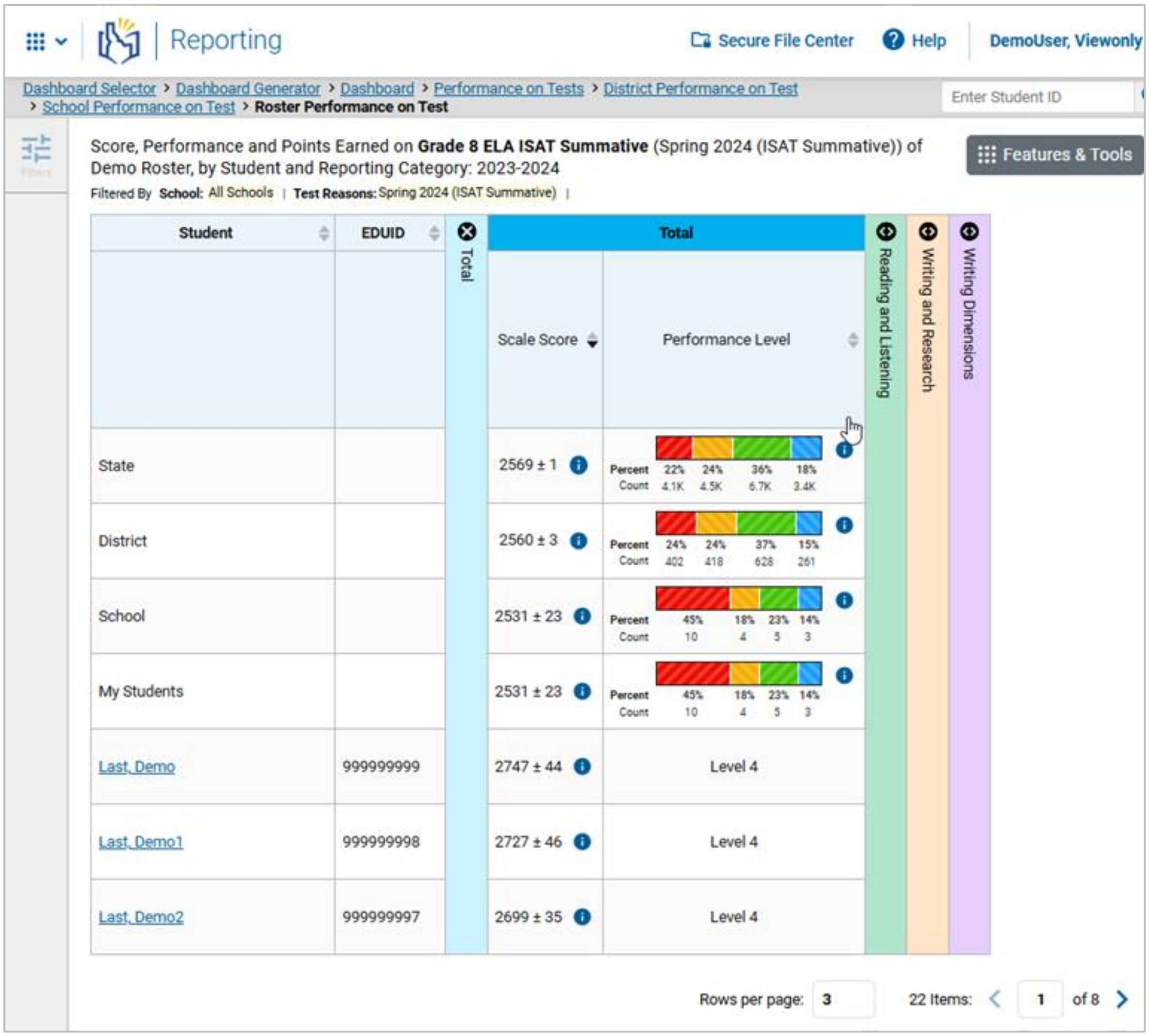
School	Claim Average Scale Score	Performance Distribution	Target A		Target B	
			Proficient?	Weak or Strong?	Proficient?	Weak or Strong?
State	2552 ± 1	Percent 32% 48% 20% Count 7.7K 11.4K 4.9K	✕	−	✕	−
District	2608 ± 3	Percent 20% 47% 34% Count 581 1.4K 1K	✓	−	✓	=
Demo School 99'	2531 ± 15	Percent 38% 54% 8% Count 19 27 4	✕	=	◒	+
Demo School 999	2993	Percent 100% Count 1	*	*	*	*
Demo School 9999	2667 ± 9	Percent 12% 40% 48% Count 33 111 131	✓	=	✓	+

Copyright © 2024 **Cambium Assessment, Inc.** All rights reserved. | Terms of Use & Privacy Policy

7.1.4 Roster Performance Report

Roster performance reports provide users with performance data for a group of students belonging to a system-defined or user-defined class. The report includes (1) the students' overall subject scale scores with standard error of measurement, (2) the achievement level and (3) for ELA/L only, writing dimensions scores. In the roster report, each student's performance can be compared with state, district, and school levels. Exhibit 7 shows a sample roster performance report for grade 8 ELA/L.

Exhibit 7. Roster Performance Report for Grade 8 ELA/L


Reporting
Secure File Center
Help
DemoUser, Viewonly
Dashboard Selector › Dashboard Generator › Dashboard › Performance on Tests › District Performance on Test
› School Performance on Test › Roster Performance on Test
Enter Student ID
Filters
Features & Tools

Score, Performance and Points Earned on **Grade 8 ELA ISAT Summative** (Spring 2024 (ISAT Summative)) of Demo Roster, by Student and Reporting Category: 2023-2024

Filtered By **School:** All Schools | **Test Reasons:** Spring 2024 (ISAT Summative) |

Student	EDUID	Total	Total		Reading and Listening	Writing and Research	Writing Dimensions
			Scale Score	Performance Level			
State			2569 ± 1	Percent 22% 24% 36% 18% Count 4.1K 4.5K 6.7K 3.4K			
District			2560 ± 3	Percent 24% 24% 37% 15% Count 402 418 628 261			
School			2531 ± 23	Percent 45% 18% 23% 14% Count 10 4 5 3			
My Students			2531 ± 23	Percent 45% 18% 23% 14% Count 10 4 5 3			
Last, Demo	999999999		2747 ± 44	Level 4			
Last, Demo1	999999998		2727 ± 46	Level 4			
Last, Demo2	999999997		2699 ± 35	Level 4			

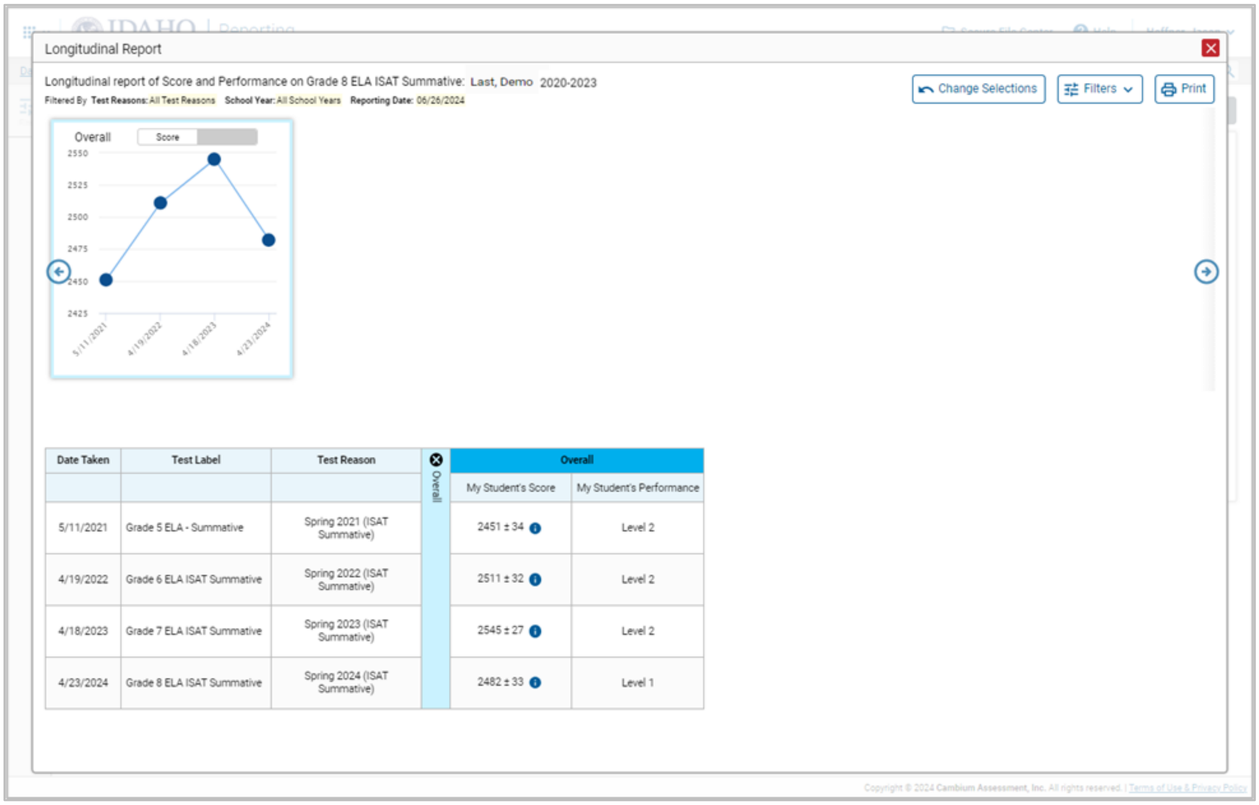
Rows per page: 3   22 Items: ‹ 1 of 8 ›

7.1.5 Trend Report

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

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 ISAT Summative: Last, Demo 2020-2023

Filtered By Test Reasons: All Test Reasons School Year: All School Years Reporting Date: 06/26/2024

Date Taken	Test Label	Test Reason	Overall	
			My Student's Score	My Student's Performance
5/11/2021	Grade 5 ELA - Summative	Spring 2021 (ISAT Summative)	2451 ± 34	Level 2
4/19/2022	Grade 6 ELA ISAT Summative	Spring 2022 (ISAT Summative)	2511 ± 32	Level 2
4/18/2023	Grade 7 ELA ISAT Summative	Spring 2023 (ISAT Summative)	2545 ± 27	Level 2
4/23/2024	Grade 8 ELA ISAT Summative	Spring 2024 (ISAT Summative)	2482 ± 33	Level 1

7.1.6 Individual Student Report

An individual student report can be generated and exported as a PDF file. The individual student report shows the student's overall performance on the test with detailed information on multiple pages. In each subject area, the individual student report provides the scale score and conditional standard error of measurement (CSEM) for overall test; (2) achievement level for overall test; (3) average scale scores for the student's state, district, and school; (4) student performance and performance level description for individual reporting categories; (5) writing scores and performance descriptors in each dimension for ELA/L only; and (6) trend of student performance over time.

Specifically, the student's name, scale score with the CSEM, and achievement level 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 CSEM using a "±" sign. CSEM 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 for 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.

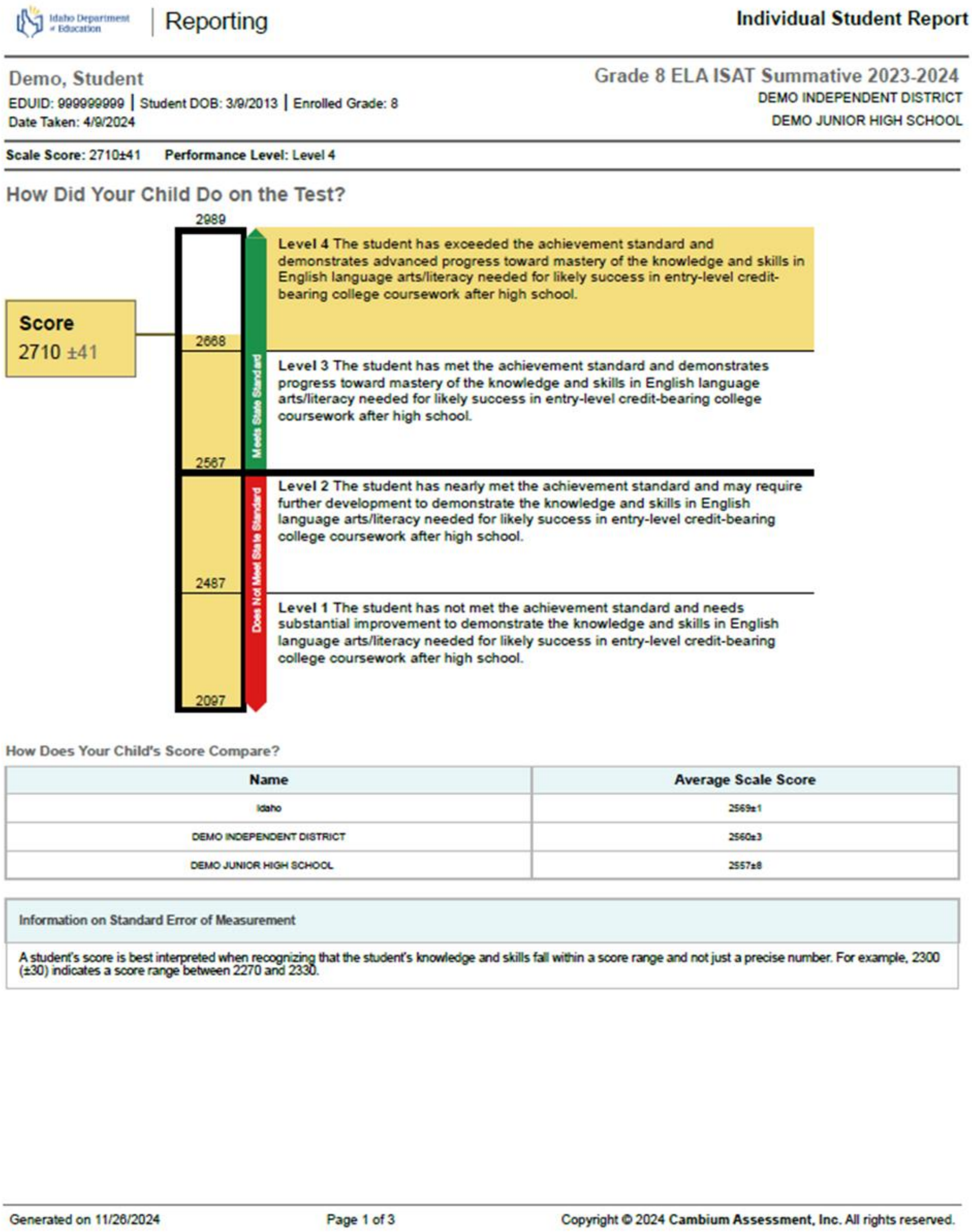
Underneath, average scale scores and standard errors of the average scale scores for the student's state, district, and school are displayed so the student's achievement can be compared with the above aggregate levels. It should be noted that the "±" next to the student's scale score is the SEM of the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

The next page of reports shows the student's performance across different reporting categories, along with descriptions, at the top of the page. Below this, the student's performance in the different writing dimensions is displayed with detailed descriptions for ELA/L only.

The last page provides the trend of student performance over time. Student scale scores and achievement levels over time are graphed, showing how the student's scale scores changed over time and whether the student met the standards each year.

Exhibit 9 presents an example of an individual student report for grade 8 ELA/L.

Exhibit 9. Individual Student Report for Grade 8 ELA/L



Individual Student Report

Demo, Student

EDUID: 999999999 | Student DOB: 3/9/2013 | Enrolled Grade: 8
Date Taken: 4/9/2024

Grade 8 ELA ISAT Summative 2023-2024
DEMO INDEPENDENT DISTRICT
DEMO JUNIOR HIGH SCHOOL

Scale Score: 2710±41 **Performance Level:** Level 4

How Did Your Child Do on the Test?

Score
2710 ±41

Level 4 The student has exceeded the achievement standard and demonstrates advanced 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.

Level 3 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.

Level 2 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 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?

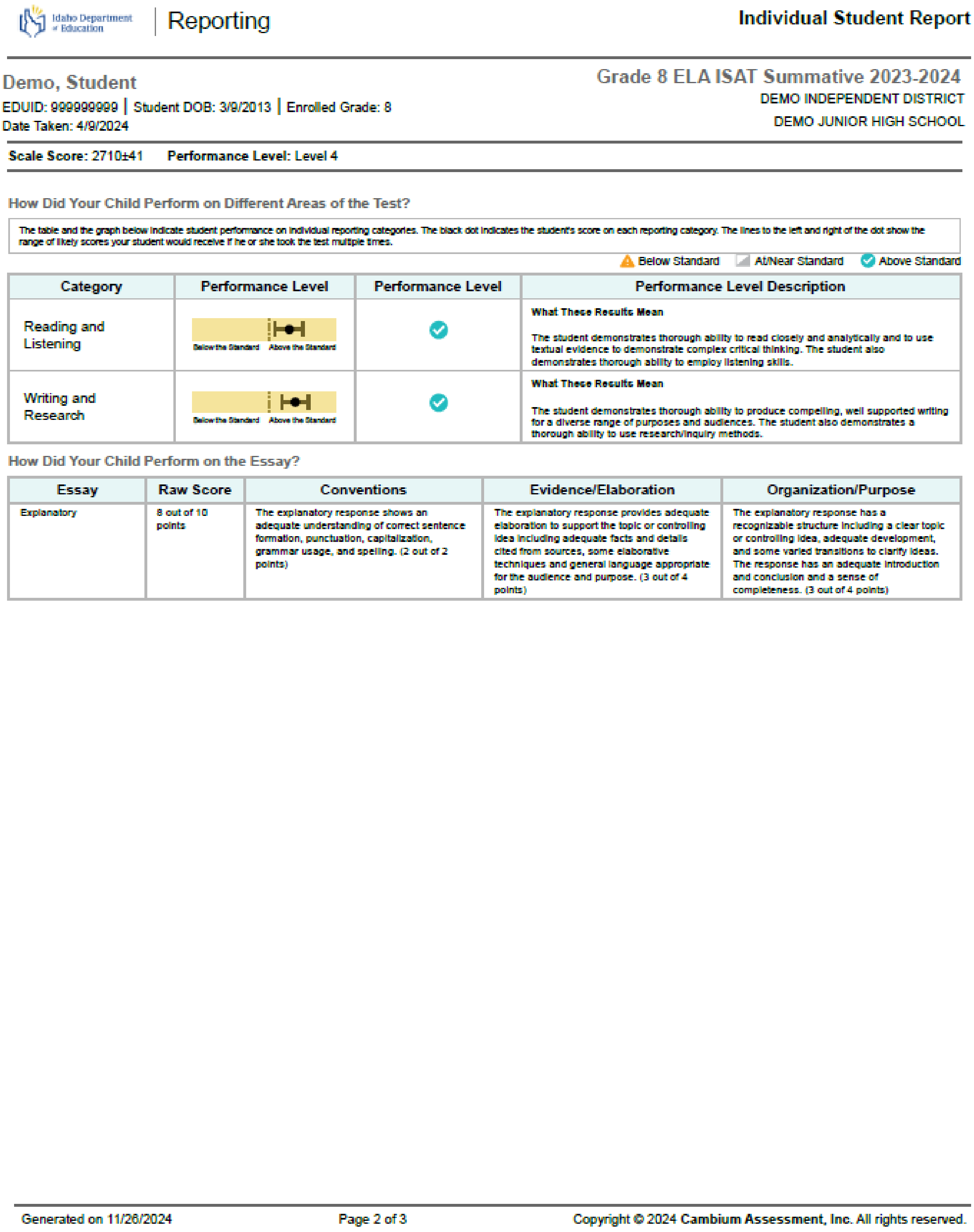
Name	Average Scale Score
Idaho	2569±1
DEMO INDEPENDENT DISTRICT	2560±3
DEMO JUNIOR HIGH SCHOOL	2557±8

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 (±30) indicates a score range between 2270 and 2330.

Generated on 11/26/2024 Page 1 of 3 Copyright © 2024 **Cambium Assessment, Inc.** All rights reserved.

Exhibit 9. Individual Student Report for Grade 8 ELA/L (Continued)



Idaho Department of Education | Reporting

Individual Student Report

Demo, Student

EDUID: 999999999 | Student DOB: 3/9/2013 | Enrolled Grade: 8
Date Taken: 4/9/2024

Grade 8 ELA ISAT Summative 2023-2024
DEMO INDEPENDENT DISTRICT
DEMO JUNIOR HIGH SCHOOL

Scale Score: 2710±41 **Performance Level:** Level 4

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

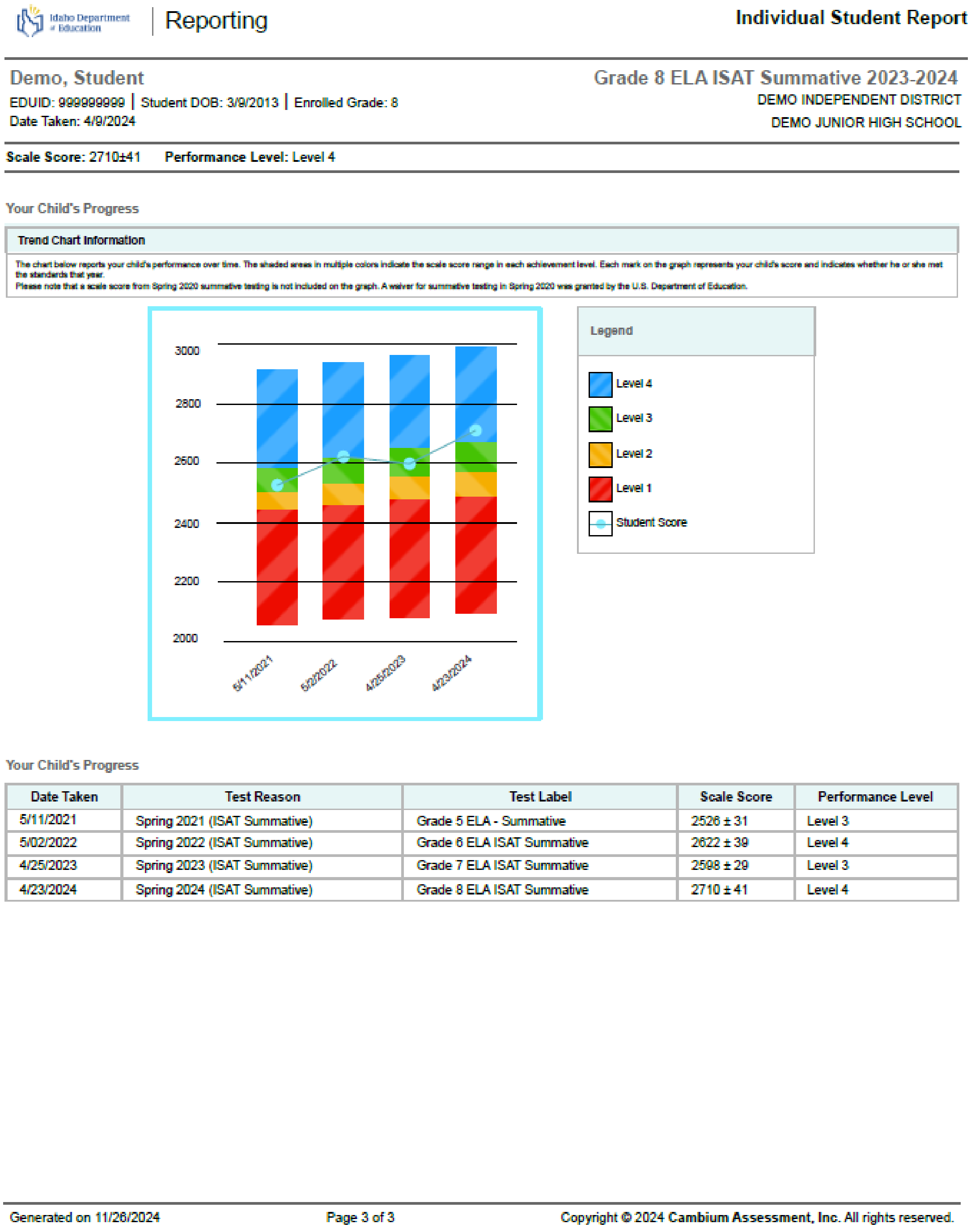
Category	Performance Level	Performance Level	Performance Level Description
Reading and Listening	Below the Standard Above the Standard	✓	**What These Results Mean** The student demonstrates thorough ability to read closely and analytically and to use textual evidence to demonstrate complex critical thinking. The student also demonstrates thorough ability to employ listening skills.
Writing and Research	Below the Standard Above the Standard	✓	**What These Results Mean** The student demonstrates thorough ability to produce compelling, well supported writing for a diverse range of purposes and audiences. The student also demonstrates a thorough ability to use research/inquiry methods.

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 11/26/2024 Page 2 of 3 Copyright © 2024 **Cambium Assessment, Inc.** All rights reserved.

Exhibit 9. Individual Student Report for Grade 8 ELA/L (Continued)



Idaho Department of Education | Reporting

Individual Student Report

Demo, Student

EDUID: 999999999 | Student DOB: 3/9/2013 | Enrolled Grade: 8
Date Taken: 4/9/2024

Grade 8 ELA ISAT Summative 2023-2024
DEMO INDEPENDENT DISTRICT
DEMO JUNIOR HIGH SCHOOL

Scale Score: 2710±41 **Performance Level:** Level 4

Your Child's Progress

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.
Please note that a scale score from Spring 2020 summative testing is not included on the graph. A waiver for summative testing in Spring 2020 was granted by the U.S. Department of Education.

Legend

🟦 Level 4
🟩 Level 3
🟧 Level 2
🟥 Level 1
● Student Score

Your Child's Progress

Date Taken	Test Reason	Test Label	Scale Score	Performance Level
5/11/2021	Spring 2021 (ISAT Summative)	Grade 5 ELA - Summative	2526 ± 31	Level 3
5/02/2022	Spring 2022 (ISAT Summative)	Grade 6 ELA ISAT Summative	2622 ± 39	Level 4
4/25/2023	Spring 2023 (ISAT Summative)	Grade 7 ELA ISAT Summative	2598 ± 29	Level 3
4/23/2024	Spring 2024 (ISAT Summative)	Grade 8 ELA ISAT Summative	2710 ± 41	Level 4

Generated on 11/26/2024 Page 3 of 3 Copyright © 2024 **Cambium Assessment, Inc.** All rights reserved.

7.2 Interpretation of Reported Scores

A student's performance on a test is reported as a scale score and with an achievement level. The following section provides more details on how to interpret these values.

7.2.1 Scale Score

A scale score is a numeric value 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. The scale score is a transformed score derived from the student's theta score, which is estimated based on a mathematical model. Lower scale scores indicate that the student has not demonstrated sufficient knowledge and skills in the relevant subject areas, as measured by the test. Conversely, higher scale scores indicate that the student has demonstrated proficient knowledge and skills in the relevant subject areas, as measured by the test. Scale scores can be used to compare student performance to the established proficiency thresholds as well as to measure student growth over time. Interpretation of scale scores is more meaningful when the scale scores are used along with achievement levels and Achievement Level Descriptors (ALDs).

7.2.2 Conditional Standard Error of Measurement

A scale score is an estimate of the true score. The standard error of measurement (SEM) represents the estimate precision of an estimated scale score. It reflects the range in which the test estimates the student's true ability to be. For example, a student who receives a test score of 2500 with a SEM of 35 is estimated to have a "true" performance on the test somewhere between 2465 and 2535. The SEM is included after the "±" next to the student's scale score. The SEM will vary across students, depending on a student's ability and the characteristics of the administered items, yielding a conditional SEM (CSEM). The CSEM is conditional on the specific items included in the test and the student's response to each item, which is why two students may have the same estimated scale score but different CSEM values. A student's scale score is best interpreted in conjunction with the CSEM. The scale score and CSEM indicate the scale range within which the student's knowledge and skills are expected to be.

7.2.3 Achievement Level

Achievement levels are proficiency categories on a test that students fall into based on their scale scores. They provide a broader description of a student's performance than scale scores. For the ELA/L and mathematics ISAT assessments, scale scores are categorized into four achievement levels (i.e., Level 1, Level 2, Level 3, and Level 4) based on grade-specific proficiency cut scores. Achievement levels can be interpreted based on the provided Achievement Level Descriptors (ALDs). For example, the grade 6 ELA/L ALD for Level 3 states, "The student has met the achievement standard and demonstrates mastery of the knowledge and skills of grade level state standards in English Language Arts." Generally, student performance in Achievement Levels 3 and 4 is considered to demonstrate proficiency at the current grade level and on track to demonstrating the knowledge and skills necessary for college and career readiness. More information on achievement levels and ALDs are available on the Smarter Balanced website at https://validity.smarterbalanced.org/scoring/.

7.2.4 Performance Category for Claims

Student performance on each claim and individual reporting category is reported in three categories: (1) Below Standard, (2) At/Near Standard, and (3) Above Standard. Unlike the achievement level for the overall test, student performance on each claim is evaluated with respect to the “Meets Standard” achievement standard. For students performing at “Below Standard” or “Above Standard,” this can be interpreted to mean that their performance is clearly below or above the “Meets Standard” cut score for a specific claim. For students performing at “At/Near Standard,” this can be interpreted to mean that their performance does not provide enough information to tell whether they reached the “Meets Standard” mark for the specific claim.

7.2.5 Performance Category for Targets

Teachers and educators sometimes need more detailed reports on student performance for instructional needs. The target report provides information on student performance about relative strength and weakness scores for each target within a claim. The strengths and weaknesses reports are generated for aggregate units of roster/classroom, school, and district and provide information about how a group of students in a class, school, or district performed on each target, either relative to the proficiency standard (i.e., "Proficient?" target measure) or relative to their overall performance on the test (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 score). At the aggregate level, when observed performance within a target is greater than the proficiency cut score, the reporting unit shows a relative strength in that target compared to the proficiency standard. Conversely, when observed performance within a target is below the proficiency cut score, the reporting unit shows a relative weakness in that target.

For the “Weak or Strong?” target measure, students’ observed performance on items within the reporting element is compared with the expected performance based on the overall ability estimate. At the aggregate level, when the observed performance within a target is greater than the expected performance, the reporting unit (e.g., roster, teacher, school, district) shows a relative strength in that target. Conversely, when observed performance within a target is below the level expected based on overall achievement, the reporting unit shows a relative weakness in that target.

Although performance categories for targets provide some evidence to help address students’ strengths and weaknesses, they should not be over interpreted because student performance on some targets may be based on relatively few items, especially for a small group.

7.2.6 Aggregated Score

Student scale scores are aggregated at the roster/classroom, school, district, 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 the overall test are reported at the aggregate level to represent how well a group of students performs.

7.3 APPROPRIATE USES OF TEST RESULTS

Assessment results provide information about student achievement in a subject area. They measure what a student knows and is able to do and estimate whether the student is on track to demonstrate the knowledge and skills necessary for college and career readiness. Assessment results can be used to identify the relative strengths and weaknesses of students in particular content areas. For example, performance categories for different content target levels can be used to identify relative strengths and weaknesses for a group of students.

The information about student achievement provided by summative and interim assessments is a useful tool for teachers and administrators looking to improve teaching methods and increase student learning. Aggregate test results at the classroom and school levels provide information about curriculum and instruction effectiveness. For example, a group of students may perform very well in the overall test, but it is possible that they would not perform well in some targets. In this case, teachers and schools can identify the strengths and weaknesses of their students through the group performance by targets and promote instruction on specific content areas. Furthermore, by narrowing down the student performance result by subgroup, teachers and schools can determine what strategies may need to be implemented to improve teaching and student learning, particularly for students from a disadvantaged subgroup. For example, teachers may view student assessment results by EL status and might observe that EL students struggle with literary response and analysis in reading. Teachers could then provide additional instruction for these students to enhance their achievement in a specific area.

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 students in other schools, districts, and states overall. 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 from multiple years. In the ISAT assessments, the scale scores across grades are on the same scale because the scores are vertically linked across grades.

While assessment results provide valuable information to understand student performance, these scores and reports should be interpreted in context. It is important to note that scale scores reported are estimates of true scores and therefore do not represent a precise measure of 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 student 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 about student learning. Finally, when student performance is compared across groups, users need to consider the group size. The smaller the group size, the larger the measurement error related to these aggregate data, thus requiring interpretation with more caution.

8. QUALITY CONTROL PROCEDURE

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

8.1 ADAPTIVE TEST CONFIGURATION

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

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

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

8.1.1 Platform Review

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

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

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

8.1.2 User Acceptance Testing and Final Review

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

8.2 QUALITY ASSURANCE IN DOCUMENT PROCESSING

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

8.3 QUALITY ASSURANCE IN DATA PREPARATION

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

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

8.4 QUALITY ASSURANCE IN ONLINE TEST DELIVERY SYSTEM

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

A series of QA reports can also be generated at any time during the online assessment window, such as blueprint match rate, item exposure rate, and item statistics, for early detection of any unexpected issues. Any deviations from the expected outcome are flagged, investigated, and resolved. In addition to these statistics, a cheating analysis report is produced to flag any unlikely patterns of behavior in a testing session, as discussed in Section 2.8, Data Forensic Program.

For example, an item statistics analysis report allows psychometricians to ensure that items are performing as intended and serves as an empirical key check through the operational testing window. The item statistics analysis report is used to monitor the performance of test items throughout the testing window and serves as a key check for the early detection of potential problems with item scoring, including incorrect designation of a keyed response or other scoring errors, as well as potential breaches of test security that may be indicated by changes in the difficulty of test items. This report generates classical item analysis indicators 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 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 blueprint, and items are performing as anticipated.

Table 68 presents an overview of the QA reports.

Table 68. Overview of Quality Assurance Reports

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

8.4.1 Score Report Quality Check

Two types of score reports are produced in the ISAT summative assessments: 1) online reports and 2) printed reports (family reports).

8.4.1.1 Online Report Quality Assurance

The systems automatically assign scores on 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 centralized location for all student scores and responses, ensuring that there is only one place where the official record is stored. Only after scores have passed the QA checks and are uploaded to the DOR are they passed to the Centralized Reporting System (CRS), which is responsible for presenting individual-level results and calculating and presenting aggregate results. Absolutely no score is reported in the CRS until it passes all the QA system's validation checks. All of the previously mentioned processes take milliseconds to complete so that within less than one second after CAI receives hand-scores and they pass QA validation checks, the composite score will be available in the CRS.

8.4.1.2 Paper Report Quality Assurance

Statistical Programming

The family reports contain custom programming and require rigorous QA processes to ensure 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 the agreed-on procedures. Custom programming is implemented independently by two statistical programming teams working from the specifications. The scripts are released for production when the output from both teams matches 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 mainly made up of calls to various macros, including macros that verify the data and conversion tables and the macros that perform the many complicated calculations. This program is developed and tested using artificial data generated to test both typical and extreme cases. Additionally, the program goes through a rigorous code review by a senior statistician.

Display Programming

The paper report development process uses graphical programming, which takes place in a Xerox-developed programming language called Variable Data Intelligent PostScript Printware (VIPP) and allows virtually infinite control of the visual appearance of the reports. After designers at CAI create backgrounds, CAI's VIPP programmers write code that indicates where to place all variable information (data, graphics, and text) on the reports. The VIPP code is tested using both artificial and real data. CAI's data generation utilities can read the output layout specifications and generate artificial data for direct input into the VIPP programs. This allows the testing of these programs to begin before the statistical programming is complete. In later stages, artificial data are generated according to the input layout and are run through the psychometric process and the score reporting statistical programs, and the output is formatted as VIPP input. This 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 the CAI building. All reports containing actual data are stored in a locked storage area. Before the reports are printed, CAI provides a live data file and individual student reports with sample districts for Department staff review. CAI will work closely with the Department to resolve questions and correct any problems. The reports will not be delivered unless the Department approves the sample reports and data file.

REFERENCES

American Educational Research Association (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.

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

Guo, F. (2006). Expected classification accuracy using the latent distribution. *Practical Assessment, Research & Evaluation, 11*(6).

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

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

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

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

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

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

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

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

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

Idaho Standards Achievement Tests in English Language Arts and Mathematics 2023–2024 Technical Report Appendices

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

LIST OF APPENDICES

Appendix A: Summary of the 2023–2024 Interim Assessments

For the ISAT ELA/L and mathematics interim assessments, four types of interim assessments are available as fixed-form tests: Interim Comprehensive Assessment (ICA), shortened Interim Comprehensive Assessment (SICA), Interim Assessment Block (IAB), and Focused Interim Assessment Block (FIAB). In each grade and subject, one ICA and one SICA are available along with multiple IABs and FIABs. Idaho created the shortened Interim Comprehensive Assessment (SICA) by dropping the PT component and short answer items in the non-PT component from the standard ICA.

Idaho administered both the standard ICAs and the SICAs. Most students took either an ICA or a SICA once, but some students took them multiple times. Tables A-1 and A-2 present the total number of students who took ICAs and SICAs in ELA/L and mathematics by the number of attempts. Total number of tests indicates the total tests taken by the total number of students, counting multiple attempts as multiple tests. For example, if a student took an ICA twice, the number of tests for this student is counted as two. Tables A-3 and A-4 summarize student performance on ICAs and SICAs for all tests taken in ELA/L and mathematics, 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 and SICAs for ELA/L

Grade	Number of Students by Number of Attempts							Total Number of Tests Taken
	Once	Twice	Three Times	Four Times	Five Times	Six Times	Total Number of Students	
ICA								
3	613	159	2	2	0	0	776	945
4	644	117	2	0	0	1	764	890
5	595	135	1	0	0	0	731	868
6	542	219	2	0	0	0	763	986
7	914	180	0	0	1	0	1,095	1,279
8	646	194	0	0	0	0	840	1,034
9	405	155	0	0	0	0	560	715
10	593	207	0	0	0	0	800	1,007
11	593	189	0	0	0	0	782	971
SICA								
3	3,825	1,467	144	9	0	0	5,445	7,227
4	3,947	1,503	155	12	71	0	5,688	7,821
5	4,060	1,429	164	19	1	0	5,673	7,491
6	4,237	2,714	324	6	0	0	7,281	10,661
7	3,874	2,528	323	6	1	0	6,732	9,928
8	4,293	2,604	282	4	0	0	7,183	10,363
9	4,174	1,399	321	17	1	0	5,912	8,008
10	3,711	1,289	383	2	0	0	5,385	7,446
11	2,684	1,259	26	5	4	0	3,978	5,320

Table A-2. Number of Students Who Took ICAs and SICAs for Mathematics

Grade	Number of Students by Number of Attempts							Total Number of Tests Taken
	Once	Twice	Three Times	Four Times	Five Times	Six Times	Total Number of Students	
				ICA				
3	459	139	1	2	0	0	601	748
4	427	118	0	1	0	0	546	667
5	275	130	1	0	0	0	406	538
6	246	154	2	0	0	0	402	560
7	442	201	2	0	0	0	645	850
8	610	209	0	0	0	0	819	1,028
9	403	200	0	0	0	0	603	803
10	441	223	1	0	0	0	665	890
11	425	191	0	0	0	0	616	807
				SICA				
3	3,832	1,483	163	5	0	0	5,483	7,307
4	3,563	1,722	107	13	67	2	5,474	7,727
5	3,711	1,594	217	3	0	0	5,525	7,562
6	4,398	3,041	409	6	0	0	7,854	11,731
7	4,840	2,735	276	5	0	0	7,856	11,158
8	4,303	2,848	359	3	0	0	7,513	11,088
9	3,419	1,435	349	7	2	0	5,212	7,374
10	3,531	1,398	347	9	0	0	5,285	7,404
11	2,625	1,302	19	1	0	0	3,947	5,290

Table A-3. Percentage of Tests in Achievement Levels for ELA/L

Grade	Total Number of Tests Taken	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
				ICA				
3	945	2403.7	91.9	36	25	21	17	38
4	890	2432.8	97.1	43	22	19	16	35
5	868	2446.3	120.8	47	15	26	12	38
6	986	2509.0	92.5	29	26	34	11	46
7	1,279	2515.6	106.2	36	26	28	11	38
8	1,034	2560.1	108.5	27	21	35	16	51
9	715	2573.5	109.2	21	27	34	19	53
10	1,007	2587.3	108.0	19	25	35	21	57
11	971	2614.8	108.7	14	22	34	29	64
				SICA				
3	7,227	2375.5	83.9	51	25	14	10	24
4	7,821	2407.3	97.9	56	18	15	11	26
5	7,491	2457.8	104.3	45	20	23	12	35
6	10,661	2485.2	108.7	43	23	23	11	34
7	9,928	2503.4	113.4	44	22	24	11	34
8	10,363	2530.4	117.3	38	23	27	12	39
9	8,008	2538.2	118.1	36	26	25	14	39
10	7,446	2550.0	123.0	33	26	25	15	41
11	5,320	2566.2	122.0	29	27	26	18	44

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

Table A-4. Percentage of Tests in Achievement Levels for Mathematics

Grade	Total Number of Tests Taken	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
				ICA				
3	748	2423.2	88.8	33	25	26	17	42
4	667	2478.5	87.9	20	33	27	20	47
5	538	2491.9	99.4	33	31	17	18	36
6	560	2519.9	105.5	30	30	21	19	40
7	850	2527.1	101.4	32	33	22	13	35
8	1,028	2533.6	98.4	39	33	17	11	29
9	803	2531.2	109.2	40	32	20	8	28
10	890	2550.5	124.6	39	30	21	10	31
11	807	2584.0	116.2	33	30	26	11	37
				SICA				
3	7,307	2379.1	73.7	51	28	16	5	21
4	7,727	2432.4	85.8	39	35	19	8	27
5	7,562	2458.2	94.9	50	29	12	9	21
6	11,731	2467.2	107.8	49	30	14	7	21
7	11,158	2497.5	108.4	45	29	17	9	26
8	11,088	2499.2	109.5	55	25	12	8	20
9	7,374	2487.6	110.0	60	25	12	3	15
10	7,404	2507.2	123.9	57	25	13	5	18
11	5,290	2528.2	124.1	55	25	13	6	20

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

For ELA/L, there were six IABs and nine FIABs in each grade, for a total of 15 assessment blocks in each grade. For mathematics, there were three to five IABs and seven to 10 FIABs in each grade, for a total of 10 to 15 assessment blocks in each grade.

Students were allowed to take as many IABs and FIABs as they wanted, and to take the same assessment block multiple times. Table A-5 shows the total number of students who took at least one assessment block and the number of students by the number of distinct assessment blocks taken. For example, in grade 3 ELA/L, a total of 8,528 students took at least one assessment block. Among 8,528 students, 3,073 students took one assessment block, 2,441 students took two distinct assessment blocks, and so on. Tables A-6 to A-13 disaggregate the number of students in Table A-5 by each individual assessment block. For example, among the 3,073 students who took one distinct assessment block only in grade 3, 123 students took the Brief Writes IAB, 274 students took the Editing FIAB, and so on.

Tables A-14 to A-19 summarize student performance on each individual assessment block for all tests taken, including the percentage of tests in each performance category. The total number of tests indicates the total number of assessment blocks taken by all students, counting multiple attempts as multiple tests. For example, if a student took the same assessment block twice, the number of tests for this student is counted as two.

Table A-5. Number of Students Who Took Distinct Assessment Blocks (Grades 3–8, 11)

Grade	**Total Students with At Least One Block**	**Number of Distinct Assessment Blocks Taken**														
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
ELA/L																
3	8,528	3,073	2,441	1,236	828	368	234	104	92	70	29	34	1	4	5	9
4	8,937	3,932	2,374	1,295	604	249	249	117	56	29	32					
5	7,856	3,651	2,115	1,087	471	242	88	60	65	53	14	2	4	1	3	
6	6,417	2,644	1,669	992	477	327	132	117	52	7						
7	5,439	2,093	1,920	896	183	106	14	38	83	41	42	23				
8	5,286	2,983	1,296	781	27	118	70	3	4	4						
11	9,604	5,542	3,106	797	74	18	16	26	18	7						
Mathematics																
3	9,720	3,879	2,757	1,421	679	421	161	93	123	56	33	95	2			
4	10,546	4,032	3,218	1,807	881	287	81	50	62	81	26	7	14			
5	9,971	3,853	3,113	1,711	769	266	69	78	75	37						
6	7,782	2,888	2,434	1,626	485	134	69	30	64	43	1	8				
7	6,240	2,659	2,273	985	156	37	48	38	42	2						
8	6,539	2,771	2,106	1,261	203	45	53	94		6						
11	7,358	2,937	2,795	969	397	173	56	9	5	3	14					

Table A-6: ELA/L Number of Students Who Took Distinct Assessment Blocks by Block Labels (Grades 3–4)

Grade	**Block**	**Number of Distinct Assessment Blocks Taken**														
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
3	Brief Writes	123	153	146	130	23	79	50	49	29	16	12	1	4	5	9
	Editing (FIAB)	274	328	401	494	293	153	56	89	69	26	34	1	4	5	9
	Language and Vocabulary Use (FIAB)	596	808	709	624	316	199	93	83	69	28	34	1	4	5	9
	Listen/Interpret (FIAB)	304	574	510	632	332	207	84	90	65	28	34	1	4	5	9
	Read Informational Texts	495	969	408	272	104	141	62	51	25	11	12	1	4	5	9
	Read Literary Texts	547	754	419	222	114	150	60	53	23	8	10		2	3	9
	Research	65	259	150	82	58	70	37	51	20	23	34	1	4	5	9
	Research: Analyze Information (FIAB)	62	90	138	196	123	76	55	37	50	20	25	1	4	4	9
	Research: Interpret and Integrate (FIAB)	50	86	62	91	83	61	39	39	49	19	25	1	4	5	9
	Research: Use Evidence (FIAB)	17	137	71	142	95	83	45	38	46	18	23	1	4	5	9
	Revision	80	126	103	40	14	56	35	13	21	22	34	1	4	5	9
	Write and Revise Informational Texts (FIAB)	109	214	186	122	66	32	27	40	49	23	31	1	3	4	9
	Write and Revise Narratives (FIAB)	3	64	72	84	57	8	19	28	49	21	31		3	5	9
	Write and Revise Opinion Texts (FIAB)	64	143	175	81	91	72	37	19	39	15	23			4	9
	Performance Task	284	177	158	100	71	17	29	56	27	12	12	1	4	5	9
4	Brief Writes	395	165	188	71	28	43	12	34	6	12					
	Editing (FIAB)	245	318	211	214	160	113	85	27	29	32					
	Language and Vocabulary Use (FIAB)	433	578	485	357	198	186	114	48	24	32					
	Listen/Interpret (FIAB)	388	537	650	399	147	191	97	54	24	32					
	Read Informational Texts	1,079	1,118	502	265	127	144	25	21	14	28					
	Read Literary Texts	776	716	408	268	110	150	25	24	13	28					
	Research	222	413	399	100	77	102	8	21	14	28					
	Research: Analyze Information (FIAB)	13	69	84	147	89	70	78	54	29	32					
	Research: Interpret and Integrate (FIAB)	12	80	53	47	22	68	80	35	23	24					
	Research: Use Evidence (FIAB)	40	218	261	200	97	125	83	43	29	32					
	Revision	152	245	243	99	51	122	55	28	11	28					
	Write and Revise Informational Texts (FIAB)	31	49	98	110	18	36	40	14	15	4					
	Write and Revise Narratives (FIAB)	3	53	92	60	28	45	35	18	15	4					
	Write and Revise Opinion Texts (FIAB)	18	17	81	19	33	33	37	15	15	4					
	Performance Task	125	172	130	60	60	66	45	12							

Table A-7: ELA/L Number of Students Who Took Distinct Assessment Blocks by Block Labels (Grades 5–6)

Grade	**Block**	**Number of Distinct Assessment Blocks Taken**														
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
5	Brief Writes	127	107	212	108	55	7	4	1		1					
	Editing (FIAB)	226	220	228	81	81	48	33	41	52	14	2	3	1	3	
	Language and Vocabulary Use (FIAB)	521	556	348	193	131	81	56	63	52	14	2	4	1	3	
	Listen/Interpret (FIAB)	188	492	292	175	134	49	42	59	51	12	1	2	1	3	
	Read Informational Texts	1,168	1,047	476	255	135	54	55	52	43	14	2	4	1	3	
	Read Literary Texts	703	834	636	302	127	60	46	40	32	14	2	3	1	3	
	Research	91	413	81	89	100	21	4	10	37	11	2	4	1	3	
	Research: Analyze Information (FIAB)	91	86	170	63	29	22	13	28	39	12		4	1	3	
	Research: Interpret and Integrate (FIAB)	27	19	107	152	69	48	21	31	39	12		3	1	3	
	Research: Use Evidence (FIAB)	157	98	43	46	75	45	20	38	26	9	1	3	1	3	
	Revision	53	81	167	55	104	7	13	22	36	11	2	3	1	3	
	Write and Revise Informational Texts (FIAB)	5	41	121	87	14	21	48	55	27	5	2	4	1	3	
	Write and Revise Narratives (FIAB)	102	17	37	68	13	19	18	37	27	4	2	4		3	
	Write and Revise Opinion Texts (FIAB)	2	54	27	32	27	23	42	41	14	3	2	3	1	3	
	Performance Task	190	165	316	178	116	23	5	2	2	4	2	4	1	3	
6	Brief Writes	78	157	90	3	3	16	58								
	Editing (FIAB)	54	303	294	264	144	81	100	45	7						
	Language and Vocabulary Use (FIAB)	144	506	546	330	181	76	91	25	7						
	Listen/Interpret (FIAB)	30	68	82	135	146	73	96	31	7						
	Read Informational Texts	191	699	490	169	167	61	97	21							
	Read Literary Texts	243	196	212	92	182	60	39	21							
	Research	616	226	98	41	38										
	Research: Analyze and Integrate Information (FIAB)	243	211	337	121	131	110	37	51	7						
	Research: Evaluate Information and Sources (FIAB)	113	287	174	244	138	79	51	52	7						
	Research: Use Evidence (FIAB)	120	38	45	117	82	48	37	51	7						
	Revision	214	120	94	172	102	5	76								
	Write and Revise Argumentative Texts (FIAB)	18	38	158	59	44	57	26	36	7						
	Write and Revise Explanatory Texts (FIAB)	275	175	29	51	33	28	35	52	7						
	Write and Revise Narratives (FIAB)	82	285	244	79	108	75	13	31	7						
	Performance Task	223	29	83	31	136	23	63								

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

Grade	**Block**	**Number of Distinct Assessment Blocks Taken**														
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
7	Brief Writes	46	155	28	50	85	4									
	Editing (FIAB)	209	403	305	79	11	8	36	74	17						
	Language and Vocabulary Use (FIAB)	431	314	295	141	102	11	37	76	35	40	23				
	Listen/Interpret (FIAB)	20	149	304	48	17	4	33	74	17						
	Read Informational Texts	328	903	318	154	92	2	24	62							
	Read Literary Texts	342	571	216	153	93	7	24	68	21	42	23				
	Research	180	594	219	7	5	6	1	8	24	42	23				
	Research: Analyze and Integrate Information (FIAB)	1	4	20	8	7	6	32	19	40	42	23				
	Research: Evaluate Information and Sources (FIAB)	7	17	311	22	15	3	14	19	41	42	23				
	Research: Use Evidence (FIAB)	190	311	195	2	3	4	12	19	40	42	23				
	Revision	4	111	153	30	7	4	24	71	22	42	23				
	Write and Revise Argumentative Texts (FIAB)	7	144	66	2	2	5	14	19	41	42	23				
	Write and Revise Explanatory Texts (FIAB)	84	49	4	10	4	8	7	20	29	41	23				
	Write and Revise Narratives (FIAB)	122	110	245	9	5	5	5	63	8	3	23				
	Performance Task	122	5	9	17	82	7	3	72	34	42	23				
8	Brief Writes	104	122	39	14	34	69	1								
	Edit/Revise	245	174	457	6	78	3									
	Editing (FIAB)	42	38	197	6	70	3	1	4	4						
	Language and Vocabulary Use (FIAB)	534	138	189	21	114	70	2	2	4						
	Listen/Interpret (FIAB)	45	201	123	8	97	67	3	4	4						
	Read Informational Texts	718	726	418	1	10	2	1								
	Read Literary Texts	426	52	312		4	1									
	Research	177	269	227	3	2	2									
	Research: Analyze and Integrate Information (FIAB)	224	176	68	9	31	65	2	3	4						
	Research: Evaluate Information and Sources (FIAB)	32	309	117	4	13	2	1	4	4						
	Research: Use Evidence (FIAB)	314	263	178	18	98	65	3	4	4						
	Write and Revise Argumentative Texts (FIAB)			1				2	4	4						
	Write and Revise Explanatory Texts (FIAB)	5	81	7	3	3	2	2	4	4						
	Write and Revise Narratives (FIAB)	5	38	4	14	35	67	3	3	4						
	Performance Task	112	5	6	1	1	2									

Table A-9: ELA/L Number of Students Who Took Distinct Assessment Blocks by Block Labels (Grade 11)

Grade	**Block**	**Number of Distinct Assessment Blocks Taken**														
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
11	Brief Writes	117	954	379						1						
	Editing (FIAB)	198	263	137	74	17	14	26	18	7						
	Language and Vocabulary Use (FIAB)	646	787	389	68	17	13	25	18	7						
	Listen/Interpret (FIAB)	330	918	142	7	11	12	25	17	7						
	Read Informational Texts	641	293	195	1	1	3	11	7	1						
	Read Literary Texts	519	625	301	2	1	4	11	7	1						
	Research	979	380	300	3				1							
	Research: Analyze and Integrate Information (FIAB)	1,706	973	179	21	16	14	17	17	7						
	Research: Evaluate Information and Sources (FIAB)	250	862	78	15	16	15	26	18	7						
	Research: Use Evidence (FIAB)	64	51	31	48	2	2	4	10	6						
	Revision	31	12	179												
	Write and Revise Argumentative Texts (FIAB)	53	47	42	51	6	12	24	18	7						
	Write and Revise Explanatory Texts (FIAB)	2	14	19	1		5	10	10	6						
	Write and Revise Narratives (FIAB)	3	8	6	3	1		1	1	6						
	Performance Task	3	25	14	2	2	2	2	2							

Table A-10: Mathematics Number of Students Who Took Distinct Assessment Blocks by Block Labels (Grades 3–4)

Grade	**Block**	**Number of Distinct Assessment Blocks Taken**														
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
3	Four Operations (FIAB)	315	518	474	327	234	61	76	106	53	32	95	2			
	Geometry (FIAB)	14	153	74	69	103	89	74	104	49	32	95	2			
	Linear and Area Measurement (FIAB)	30	147	260	141	200	56	61	92	55	29	95	2			
	Measurement and Data	53	365	177	118	117	73	33	75	22	31	95	2			
	Multiplication and Division (FIAB)	251	490	418	302	172	61	62	69	50	32	95	2			
	Multiply and Divide within 100 (FIAB)	620	736	699	410	325	143	73	92	51	33	95	2			
	Number and Operations - Fractions (FIAB)	520	567	504	289	260	122	57	83	52	32	95	2			
	Number and Operations in Base Ten (FIAB)	915	1,119	824	561	306	103	84	116	52	33	95	2			
	Operations and Algebraic Thinking	976	963	374	140	121	86	37	78	24	31	95	2			
	Properties of Multiplication and Division (FIAB)	18	240	258	227	170	94	32	84	52	27	95	2			
	Time, Volume, and Mass (FIAB)	98	205	193	122	83	68	62	85	43	18	95	2			
	Performance Task	69	11	8	10	14	10			1			2			
4	Build Fractions from Unit Fractions (FIAB)	284	172	282	274	169	58	23	32	80	26	4	9			
	Factors and Multiples (FIAB)	201	415	397	498	228	71	44	57	76	26	6	14			
	Four Operations (FIAB)	157	382	261	301	193	51	44	58	81	25	7	14			
	Fraction Equivalence and Ordering (FIAB)	358	285	440	370	206	56	25	49	77	21	3	5			
	Fractions and Decimal Notation (FIAB)	22	43	138	68	43	37	25	53	58	19	7	14			
	Generate and Analyze Patterns (FIAB)	20	104	85	72	20	28	35	51	63	26	6	14			
	Geometry (FIAB)	21	45	247	91	74	23	17	38	53	7	7	14			
	Measurement and Data	81	124	232	202	1	2	5	12	16	5	7	14			
	Multi-Digit Arithmetic (FIAB)	61	268	302	313	131	43	39	44	66	24	7	14			
	Number and Operations - Fractions (FIAB)	608	1,162	779	119	23	2	4	11	20	5	6	14			
	Number and Operations in Base Ten	1,558	1,966	1,046	426	92	35	21	17	31	26	5	14			
	Operations and Algebraic Thinking	254	681	624	230	44	22	26	20	31	26	6	14			
	Place Value and Multi-Digit Whole Numbers (FIAB)	385	778	582	553	200	58	40	52	77	24	6	14			
	Performance Task	22	11	6	7	11		2	2							

Table A-11: Mathematics Number of Students Who Took Distinct Assessment Blocks by Block Labels (Grades 5–6)

Grade	**Block**	**Number of Distinct Assessment Blocks Taken**														
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
5	Add and Subtract with Equivalent Fractions (FIAB)	498	1,078	860	559	247	65	73	71	33						
	Convert Measurements (FIAB)	31	128	57	63	46	17	22	7	15						
	Geometry (FIAB)	13	84	134	210	66	35	26	20	33						
	Measurement and Data	45	143	187	131	26	6	31	50	31						
	Number and Operations - Fractions (FIAB)	899	1,629	868	405	114	28	57	72	37						
	Number and Operations in Base Ten	1,404	1,482	742	261	94	33	53	69	30						
	Numerical Expressions (FIAB)	20	83	88	188	139	35	65	69	30						
	Operations and Algebraic Thinking	55	367	484	202	64	11	35	69	37						
	Operations with Whole Numbers and Decimals (FIAB)	456	480	614	400	164	61	60	64	31						
	Place Value System (FIAB)	260	392	670	268	185	60	62	75	37						
	Volume Concepts (FIAB)	148	359	428	387	185	62	62	33	19						
	Performance Task	24	1	1	2		1		1							
6	Algebraic Expressions (FIAB)	337	235	313	247	56	59	21	51	43	1	8				
	Dependent and Independent Variables (FIAB)	135	169	270	47	60	62	23	63	43	1	8				
	Divide Fractions by Fractions (FIAB)	540	1,034	1,035	425	80	65	21	61	36	1	8				
	Expressions and Equations	60	355	126	113	54	5	2	5	8	1	8				
	Geometry (FIAB)	405	578	830	147	56	11	24	62	43	1	8				
	Multi-Digit Numbers, Factors, and Multiples (FIAB)	362	583	535	318	98	58	23	17	36	1	8				
	One-Variable Expressions and Equations (FIAB)	40	379	87	112	81	65	28	62	43	1	8				
	Rational Number System II (FIAB)	50	528	379	131	50	17	20	64	43	1	8				
	Ratios and Proportional Relationships (FIAB)	691	667	803	277	81	59	25	62	43	1	8				
	Statistics and Probability (FIAB)	19	28	146	76	49	9	21	61	42	1	8				
	The Number System	237	278	333	44	4	1	2	4	7		8				
	Performance Task 1	3	21	20	3	1	3									
	Performance Task 2	9	13	1												

Note: There are two performance task IABs offered in Grade 6. Performance Task 1 is called Cell Phone Plan, and Performance Task 2 is called Feeding the Giraffe.

Table A-12: Mathematics Number of Students Who Took Distinct Assessment Blocks by Block Labels (Grades 7–8)

Grade	**Block**	**Number of Distinct Assessment Blocks Taken**														
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
7	Algebraic Expressions and Equations (FIAB)	278	906	679	121	26	42	38	42	2						
	Angles, Areas, and Volume (FIAB)	120	12	33	37	19	36	38	42	2						
	Equivalent Expressions (FIAB)	178	476	534	84	31	41	37	42	2						
	Expressions and Equations	338	181	143	15	8	5			2						
	Geometric Figures (FIAB)	25	180	114	42	9	28	24	42	2						
	Geometry	9	120	14	1			1		2						
	Ratios and Proportional Relationships (FIAB)	1,245	1,057	685	130	33	43	32	42	2						
	Statistics and Probability (FIAB)	25	10	68	47	25	47	35	42	2						
	The Number System (FIAB)	428	1,595	664	143	21	29	35	42	2						
	Performance Task	13	9	21	4	13	17	26	42							
8	Analyze and Solve Linear Equations (FIAB)	672	678	579	91	40	52	94		6						
	Congruence and Similarity (FIAB)	299	461	850	126	21	53	94		6						
	Expressions and Equations I	197	368	25	103	6	12			6						
	Expressions and Equations II (FIAB)	78	394	205	39	39	41	94		6						
	Functions (FIAB)	368	775	728	100	39	41	94		6						
	Geometry	8	155	30	55	5	12			6						
	Proportional Relationships, Lines, and Linear Equations (FIAB)	1,058	971	1,149	140	37	41	94		6						
	The Number System (FIAB)	73	312	120	65	31	39	94		6						
	Volume of Cylinders, Cones, and Spheres (FIAB)	17	81	82	93	7	27	94		6						
	Performance Task	1	17	15												

Table A-13: Mathematics Number of Students Who Took Distinct Assessment Blocks by Block Labels (Grade 11)

Grade	**Block**	**Number of Distinct Assessment Blocks Taken**														
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
11	Algebra and Functions I	462	921	236	215	126	48									
	Algebra and Functions II	32	151	112	125	68	48									
	Create Equations: Linear and Exponential (FIAB)	134	41	110	58	1	7	8	5	3	14					
	Create Equations: Quadratic (FIAB)	24	118	242	100	34	2		3	3	14					
	Equations and Reasoning (FIAB)	181	404	265	51	20	32	8	4	3	14					
	Geometry Congruence	17	141	217	60	34										
	Geometry Measurement and Modeling	2	57	94	60	32										
	Geometry and Right Triangle Trigonometry (FIAB)	542	729	294	97	73	3	1	3	2	14					
	Interpreting Functions (FIAB)	193	398	180	232	82	27	9	4	1	14					
	Number and Quantity (FIAB)	224	383	182	64	48	28	9	3	3	14					
	Seeing Structure in Expressions/Polynomial Expressions (FIAB)	263	999	356	192	57	32	9	5	3	14					
	Solve Equations and Inequalities: Linear and Exponential (FIAB)	549	618	374	187	93	54	9	5	3	14					
	Solve Equations and Inequalities: Quadratic (FIAB)	259	410	107	47	86	31	9	5	3	14					
	Statistics and Probability (FIAB)	15	120	15	24		1	1	3	3	14					
	Performance Task	40	100	123	76	111	23									

Table A-14: ELA/L Percentage of Tests in Performance Categories by Assessment Block Labels (Grades 3–5)

Grade	**Block**	**Total Number of Tests Taken**	**% Below**	**% At/Near**	**% Above**
3	Brief Writes	838	0	88	12
	Editing (FIAB)	2,319	32	48	20
	Language and Vocabulary Use (FIAB)	3,710	34	46	20
	Listen/Interpret (FIAB)	3,030	29	52	19
	Read Informational Texts	2,743	24	62	14
	Read Literary Texts	2,588	28	47	25
	Research	869	25	45	30
	Research: Analyze Information (FIAB)	910	25	52	23
	Research: Interpret and Integrate (FIAB)	647	24	48	28
	Research: Use Evidence (FIAB)	762	20	60	19
	Revision	618	25	54	21
	Write and Revise Informational Texts (FIAB)	921	22	60	17
	Write and Revise Narratives (FIAB)	458	34	55	11
	Write and Revise Opinion Texts (FIAB)	780	24	58	18
	Performance Task	984	0	89	11
4	Brief Writes	966	29	67	4
	Editing (FIAB)	1,485	31	50	19
	Language and Vocabulary Use (FIAB)	2,825	23	54	23
	Listen/Interpret (FIAB)	2,618	28	55	18
	Read Informational Texts	3,529	24	55	21
	Read Literary Texts	2,729	33	51	17
	Research	1,488	29	47	24
	Research: Analyze Information (FIAB)	667	43	44	13
	Research: Interpret and Integrate (FIAB)	447	28	47	25
	Research: Use Evidence (FIAB)	1,457	35	46	19
	Revision	1,119	39	49	11
	Write and Revise Informational Texts (FIAB)	418	37	51	12
	Write and Revise Narratives (FIAB)	356	43	49	8
	Write and Revise Opinion Texts (FIAB)	272	40	49	11
	Performance Task	676	26	70	4
5	Brief Writes	697	19	67	14
	Editing (FIAB)	1,062	22	45	33
	Language and Vocabulary Use (FIAB)	2,383	26	53	21
	Listen/Interpret (FIAB)	1,648	18	53	29
	Read Informational Texts	3,703	14	61	26
	Read Literary Texts	3,155	20	51	29
	Research	908	29	51	20
	Research: Analyze Information (FIAB)	563	26	47	27
	Research: Interpret and Integrate (FIAB)	570	28	40	32
	Research: Use Evidence (FIAB)	591	28	50	22
	Revision	613	33	43	24
	Write and Revise Informational Texts (FIAB)	461	29	55	16
	Write and Revise Narratives (FIAB)	353	33	54	13
	Write and Revise Opinion Texts (FIAB)	293	45	51	4
	Performance Task	1,124	36	49	15

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

Table A-15: ELA/L Percentage of Tests in Performance Categories by Assessment Block Labels (Grades 6–8)

Grade	**Block**	**Total Number of Tests Taken**	**% Below**	**% At/Near**	**% Above**
6	Brief Writes	415	13	68	19
	Editing (FIAB)	1,489	30	56	14
	Language and Vocabulary Use (FIAB)	1,981	32	50	18
	Listen/Interpret (FIAB)	670	29	49	22
	Read Informational Texts	2,051	21	57	22
	Read Literary Texts	1,084	23	54	22
	Research	1,020	18	53	29
	Research: Analyze and Integrate Information (FIAB)	1,262	10	71	19
	Research: Evaluate Information and Sources (FIAB)	1,146	22	53	25
	Research: Use Evidence (FIAB)	554	19	61	19
	Revision	785	33	56	11
	Write and Revise Argumentative Texts (FIAB)	519	36	51	13
	Write and Revise Explanatory Texts (FIAB)	686	43	49	8
	Write and Revise Narratives (FIAB)	1,067	22	60	18
	Performance Task	660	25	69	6
7	Brief Writes	374	22	67	11
	Editing (FIAB)	1,174	14	71	16
	Language and Vocabulary Use (FIAB)	1,657	30	50	21
	Listen/Interpret (FIAB)	706	23	59	18
	Read Informational Texts	1,895	24	51	26
	Read Literary Texts	1,640	26	52	22
	Research	1,132	21	60	19
	Research: Analyze and Integrate Information (FIAB)	219	19	69	12
	Research: Evaluate Information and Sources (FIAB)	535	30	48	22
	Research: Use Evidence (FIAB)	855	15	54	32
	Revision	496	33	52	15
	Write and Revise Argumentative Texts (FIAB)	381	18	67	15
	Write and Revise Explanatory Texts (FIAB)	294	26	62	13
	Write and Revise Narratives (FIAB)	611	20	70	10
	Performance Task	497	26	60	15
8	Brief Writes	495	18	73	8
	Edit/Revise	1,166	22	53	25
	Editing (FIAB)	366	26	47	27
	Language and Vocabulary Use (FIAB)	1,244	22	53	24
	Listen/Interpret (FIAB)	555	28	53	19
	Read Informational Texts	2,106	18	48	34
	Read Literary Texts	994	23	48	30
	Research	690	28	50	23
	Research: Analyze and Integrate Information (FIAB)	594	27	50	23
	Research: Evaluate Information and Sources (FIAB)	487	25	54	21
	Research: Use Evidence (FIAB)	955	19	57	23
	Write and Revise Argumentative Texts (FIAB)	11	18	73	9
	Write and Revise Explanatory Texts (FIAB)	111	37	51	12
	Write and Revise Narratives (FIAB)	175	18	70	12
	Performance Task	188	30	66	4

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

Table A-16: ELA/L Percentage of Tests in Performance Categories by Assessment Block Labels (Grade 11)

Grade	Block	Total Number of Tests Taken	% Below	% At/Near	% Above
11	Brief Writes	2,117	75	21	4
	Editing (FIAB)	756	26	48	26
	Language and Vocabulary Use (FIAB)	1,989	29	51	20
	Listen/Interpret (FIAB)	1,471	23	55	22
	Read Informational Texts	1,232	24	42	34
	Read Literary Texts	1,603	23	51	27
	Research	2,212	18	49	32
	Research: Analyze and Integrate Information (FIAB)	2,958	22	49	29
	Research: Evaluate Information and Sources (FIAB)	1,289	23	51	27
	Research: Use Evidence (FIAB)	218	18	41	40
	Revision	222	47	44	9
	Write and Revise Argumentative Texts (FIAB)	269	36	42	22
	Write and Revise Explanatory Texts (FIAB)	67	69	30	1
	Write and Revise Narratives (FIAB)	29	59	38	3
	Performance Task	52	50	48	2

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

Table A-17: Mathematics Percentage of Tests in Performance Categories by Assessment Block Labels (Grades 3–5)

Grade	Block	Total Number of Tests Taken	% Below	% At/Near	% Above
3	Four Operations (FIAB)	2,470	33	47	20
	Geometry (FIAB)	919	33	50	16
	Linear and Area Measurement (FIAB)	1,280	24	49	28
	Measurement and Data	1,333	34	38	28
	Multiplication and Division (FIAB)	2,159	26	46	29
	Multiply and Divide within 100 (FIAB)	3,665	32	32	36
	Number and Operations - Fractions (FIAB)	2,849	24	47	30
	Number and Operations in Base Ten (FIAB)	4,547	33	39	27
	Operations and Algebraic Thinking	3,217	34	48	18
	Properties of Multiplication and Division (FIAB)	1,433	24	41	35
	Time, Volume, and Mass (FIAB)	1,189	30	33	37
	Performance Task	125	8	65	27
4	Build Fractions from Unit Fractions (FIAB)	1,553	27	44	28
	Factors and Multiples (FIAB)	2,113	35	45	20
	Four Operations (FIAB)	1,672	42	35	23
	Fraction Equivalence and Ordering (FIAB)	2,122	44	32	24
	Fractions and Decimal Notation (FIAB)	531	22	51	27
	Generate and Analyze Patterns (FIAB)	527	25	60	15
	Geometry (FIAB)	643	12	69	19
	Measurement and Data	812	32	52	16
	Multi-Digit Arithmetic (FIAB)	1,438	32	51	17
	Number and Operations - Fractions (FIAB)	3,121	37	41	22
	Number and Operations in Base Ten	6,056	29	47	24
	Operations and Algebraic Thinking	2,290	39	47	14
	Place Value and Multi-Digit Whole Numbers (FIAB)	3,026	22	44	34
	Performance Task	61	0	79	21
5	Add and Subtract with Equivalent Fractions (FIAB)	4,229	41	31	28
	Convert Measurements (FIAB)	468	38	40	21
	Geometry (FIAB)	652	35	47	17
	Measurement and Data	697	41	35	24
	Number and Operations - Fractions (FIAB)	5,135	49	36	15
	Number and Operations in Base Ten	4,921	33	47	20
	Numerical Expressions (FIAB)	904	25	33	42
	Operations and Algebraic Thinking	1,434	35	47	18
	Operations with Whole Numbers and Decimals (FIAB)	2,649	33	45	22
	Place Value System (FIAB)	2,456	33	36	31
	Volume Concepts (FIAB)	1,905	22	45	34
	Performance Task	30	0	80	20

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

Table A-18: Mathematics Percentage of Tests in Performance Categories by Assessment Block Labels (Grades 6–8)

Grade	Block	Total Number of Tests Taken	% Below	% At/Near	% Above
6	Algebraic Expressions (FIAB)	1,445	24	46	30
	Dependent and Independent Variables (FIAB)	930	37	51	12
	Divide Fractions by Fractions (FIAB)	3,769	23	39	38
	Expressions and Equations	938	34	39	27
	Geometry (FIAB)	2,179	24	50	26
	Multi-Digit Numbers, Factors, and Multiples (FIAB)	2,303	32	40	27
	One-Variable Expressions and Equations (FIAB)	958	26	30	44
	Rational Number System II (FIAB)	1,481	16	52	32
	Ratios and Proportional Relationships (FIAB)	2,976	38	37	25
	Statistics and Probability (FIAB)	551	31	60	10
	The Number System	1,034	30	48	22
	Performance Task 1	51	0	100	0
	Performance Task 2	23	0	96	4
7	Algebraic Expressions and Equations (FIAB)	2,336	31	47	22
	Angles, Areas, and Volume (FIAB)	340	16	55	29
	Equivalent Expressions (FIAB)	1,731	22	44	34
	Expressions and Equations	791	31	41	27
	Geometric Figures (FIAB)	507	20	45	35
	Geometry	147	6	75	19
	Ratios and Proportional Relationships (FIAB)	3,753	22	56	22
	Statistics and Probability (FIAB)	306	43	46	11
	The Number System (FIAB)	3,074	27	55	18
	Performance Task	156	28	66	6
8	Analyze and Solve Linear Equations (FIAB)	2,739	28	53	19
	Congruence and Similarity (FIAB)	2,135	26	47	27
	Expressions and Equations I	801	26	49	24
	Expressions and Equations II (FIAB)	1,034	37	48	15
	Functions (FIAB)	2,452	44	44	12
	Geometry	271	17	55	27
	Proportional Relationships, Lines, and Linear Equations (FIAB)	3,715	17	56	27
	The Number System (FIAB)	991	35	39	26
	Volume of Cylinders, Cones, and Spheres (FIAB)	407	0	75	25
	Performance Task	33	55	45	0

Notes: 1. The percentage of each achievement level may not add up to 100% due to rounding.
2. There are two performance task IABs offered in Grade 6. Performance Task 1 is called Cell Phone Plan, and Performance Task 2 is called Feeding the Giraffe.

Table A-19: Mathematics Percentage of Tests in Performance Categories by Assessment Block Labels (Grade 11)

Grade	**Block**	**Total Number of Tests Taken**	**% Below**	**% At/Near**	**% Above**
11	Algebra and Functions I	2,223	57	36	7
	Algebra and Functions II	536	30	57	13
	Create Equations: Linear and Exponential (FIAB)	387	45	30	24
	Create Equations: Quadratic (FIAB)	540	4	75	21
	Equations and Reasoning (FIAB)	983	53	24	23
	Geometry Congruence	498	9	71	19
	Geometry Measurement and Modeling	245	18	71	11
	Geometry and Right Triangle Trigonometry (FIAB)	1,864	29	36	35
	Interpreting Functions (FIAB)	1,150	29	48	23
	Number and Quantity (FIAB)	1,026	38	46	15
	Seeing Structure in Expressions/Polynomial Expressions (FIAB)	1,954	54	33	12
	Solve Equations and Inequalities: Linear and Exponential (FIAB)	2,190	51	36	13
	Solve Equations and Inequalities: Quadratic (FIAB)	974	14	59	27
	Statistics and Probability (FIAB)	257	28	55	17
	Performance Task	473	0	95	5

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

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

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

Group	2020-2021				2021-2022				2022-2023				2023-2024			
	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD
Grade 3																
All Students	22,355	48	2421.2	92.4	23,134	49	2425.4	95.8	23,266	45	2416.2	91.8	23,374	48	2421.5	97.5
Female	10,958	51	2427.6	91.1	11,391	53	2432.3	95.2	11,293	48	2423.3	90.6	11,507	50	2427.7	96.5
Male	11,397	45	2415.1	93.2	11,743	46	2418.7	95.9	11,973	42	2409.5	92.5	11,867	45	2415.5	98.2
American Indian/Alaska Native	223	22	2361.3	91.2	231	29	2378.4	86.1	246	21	2367.2	89.6	207	24	2369.4	90.1
Asian	255	60	2450.3	97.0	281	62	2457.5	107.8	251	59	2443.1	91.8	251	63	2454.0	98.5
Black or African American	257	28	2375.5	86.9	272	29	2375.4	99.6	258	21	2368.2	85.2	261	33	2377.0	98.5
Hispanic or Latino	4,034	30	2383.2	87.0	4,105	31	2386.8	89.2	4,420	28	2382.1	85.2	4,464	32	2385.7	93.8
Pacific Islander	211	48	2420.8	84.7	191	42	2410.9	95.0	354	40	2408.7	94.2	261	44	2415.4	93.4
White	17,375	52	2431.1	90.9	18,054	54	2435.2	94.4	17,737	49	2425.8	91.0	17,666	52	2431.8	95.9
EL	1,999	22	2366.4	83.6	1,981	22	2366.5	84.4	1,854	18	2359.4	80.0	2,006	22	2361.2	90.8
Special Education	2,506	19	2350.6	90.6	2,747	19	2351.1	92.0	2,966	17	2345.7	87.9	2,912	17	2343.0	91.9
Section 504 Plan	517	41	2405.4	89.2	550	40	2409.3	87.7	606	41	2412.3	88.5	719	43	2410.6	95.4
Grade 4																
All Students	22,904	50	2467.6	95.4	23,169	52	2472.9	97.6	23,457	48	2462.0	98.8	23,631	49	2465.5	102.7
Female	11,274	52	2475.2	94.8	11,369	54	2479.3	96.4	11,573	50	2468.0	98.1	11,477	52	2472.8	101.1
Male	11,630	47	2460.2	95.5	11,800	50	2466.7	98.4	11,884	46	2456.1	99.1	12,154	47	2458.6	103.7
American Indian/Alaska Native	255	24	2404.2	91.9	217	22	2398.8	91.7	231	27	2408.5	90.2	238	27	2419.3	93.8
Asian	258	59	2486.8	106.6	266	67	2513.2	100.2	291	59	2487.1	103.2	249	66	2506.6	97.7
Black or African American	297	31	2412.9	107.3	264	33	2423.9	99.4	271	26	2409.4	92.3	276	28	2414.9	101.4
Hispanic or Latino	4,274	31	2427.3	90.3	4,159	33	2431.6	93.9	4,332	30	2421.9	92.8	4,564	33	2426.6	98.1
Pacific Islander	169	50	2458.5	96.4	177	47	2472.9	87.0	236	40	2449.5	91.7	215	49	2460.6	104.3
White	17,651	55	2479.0	92.9	18,086	57	2483.4	95.3	18,096	53	2472.8	97.4	17,979	54	2476.5	101.0
EL	2,088	22	2404.5	88.0	1,992	25	2413.1	92.0	1,905	22	2402.1	89.7	2,092	24	2400.6	97.3
Special Education	2,724	16	2382.4	93.9	2,693	19	2385.1	100.6	2,967	16	2375.0	91.4	3,077	16	2373.1	98.2
Section 504 Plan	640	45	2456.2	93.2	664	46	2460.7	87.8	756	42	2449.5	94.5	903	45	2462.0	94.2

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

Group	2020-2021				2021-2022				2022-2023				2023-2024			
	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD
Grade 5																
All Students	23,270	55	2510.3	99.0	23,556	57	2513.7	101.5	23,398	51	2500.5	102.6	23,742	53	2505.2	106.6
Female	11,388	60	2521.1	97.0	11,623	60	2523.7	99.1	11,449	54	2508.4	100.4	11,701	56	2513.8	105.2
Male	11,882	51	2500.0	99.8	11,933	53	2503.9	102.9	11,949	49	2493.0	104.1	12,041	50	2496.8	107.3
American Indian/Alaska Native	270	30	2455.0	99.7	260	34	2458.9	106.5	207	25	2437.7	99.6	241	31	2452.1	103.3
Asian	249	70	2542.0	101.2	266	70	2539.2	105.1	267	65	2537.7	108.9	298	69	2537.0	121.1
Black or African American	292	38	2463.3	109.5	296	41	2467.7	103.6	236	32	2443.6	103.9	269	30	2441.4	104.2
Hispanic or Latino	4,283	37	2469.0	95.0	4,340	40	2473.7	98.7	4,282	33	2456.9	96.9	4,426	35	2460.2	100.1
Pacific Islander	156	50	2500.8	99.5	184	57	2513.1	94.1	233	49	2489.1	90.2	193	50	2492.1	112.3
White	18,020	60	2521.4	96.6	18,210	61	2524.4	99.2	18,173	56	2511.9	100.6	18,229	58	2517.5	104.4
EL	2,013	28	2446.9	97.8	1,960	29	2449.0	96.2	1,843	26	2440.2	99.3	2,125	26	2436.5	102.4
Special Education	2,693	16	2411.6	93.2	2,790	18	2414.1	99.2	2,775	14	2400.5	94.9	2,998	14	2399.5	95.9
Section 504 Plan	760	47	2493.4	92.2	808	53	2509.7	97.8	899	48	2493.1	95.9	1,080	49	2496.8	98.4
Grade 6																
All Students	23,669	52	2529.8	95.5	23,902	53	2532.3	97.8	23,619	49	2524.4	100.1	23,513	52	2529.1	100.4
Female	11,438	57	2542.8	92.9	11,700	58	2544.7	95.1	11,641	53	2534.5	97.8	11,436	57	2540.9	98.2
Male	12,231	47	2517.6	96.3	12,202	48	2520.5	98.8	11,978	46	2514.5	101.4	12,077	47	2517.9	101.2
American Indian/Alaska Native	247	29	2478.7	94.0	276	30	2481.7	95.8	253	25	2465.2	96.5	191	32	2474.3	98.2
Asian	224	63	2557.7	93.7	250	69	2572.2	98.6	267	64	2555.1	107.7	263	68	2572.3	104.2
Black or African American	271	29	2479.6	96.8	273	36	2486.7	105.6	281	24	2466.2	97.4	251	33	2477.0	99.6
Hispanic or Latino	4,348	35	2491.4	92.3	4,360	34	2490.8	94.1	4,439	31	2482.7	95.9	4,435	33	2486.4	97.4
Pacific Islander	156	47	2516.1	100.7	203	51	2530.5	101.8	217	49	2517.9	103.4	213	54	2538.0	95.4
White	18,423	57	2540.0	93.4	18,540	57	2543.0	95.3	18,162	54	2535.9	97.7	18,077	57	2540.4	97.8
EL	1,716	25	2466.2	94.2	1,842	26	2468.6	96.4	1,840	23	2460.3	96.1	2,140	26	2467.7	100.2
Special Education	2,539	12	2421.5	87.9	2,698	11	2425.0	87.2	2,728	11	2419.3	88.2	2,697	10	2416.6	86.5
Section 504 Plan	929	44	2514.0	87.6	927	43	2512.1	87.6	1,080	40	2510.3	90.9	1,259	44	2516.1	92.5

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

Group	2020-2021				2021-2022				2022-2023				2023-2024			
	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD
Grade 7																
All Students	24,515	58	2561.4	99.0	24,273	58	2563.6	102.2	23,920	53	2551.0	103.4	23,766	56	2557.0	107.8
Female	12,029	64	2575.0	94.5	11,726	63	2578.1	96.6	11,666	58	2563.1	99.4	11,710	61	2571.2	104.0
Male	12,486	53	2548.4	101.5	12,547	53	2550.1	105.4	12,254	48	2539.4	105.8	12,056	51	2543.2	109.5
American Indian/Alaska Native	277	36	2511.1	99.0	251	35	2512.6	92.8	247	32	2504.7	104.0	243	36	2508.8	104.8
Asian	245	69	2589.6	106.2	222	70	2590.6	104.5	254	69	2593.0	106.2	261	70	2596.9	114.1
Black or African American	329	35	2501.7	115.6	272	34	2503.9	109.8	268	31	2488.5	110.5	281	36	2499.6	117.1
Hispanic or Latino	4,568	40	2521.1	95.5	4,473	41	2523.9	99.7	4,432	35	2508.9	99.1	4,649	37	2511.9	108.2
Pacific Islander	174	62	2573.1	88.4	208	48	2543.7	102.1	207	53	2550.0	104.8	208	50	2550.7	108.4
White	18,922	63	2572.5	96.3	18,847	62	2574.5	99.8	18,512	58	2562.0	101.1	18,040	61	2569.8	103.7
EL	1,884	29	2494.4	98.7	1,608	30	2494.2	101.5	1,824	27	2485.5	101.9	2,199	28	2485.0	111.0
Special Education	2,580	13	2442.6	95.6	2,500	13	2445.4	94.6	2,652	10	2432.4	89.5	2,623	12	2437.4	98.3
Section 504 Plan	1,025	47	2542.1	89.0	1,163	49	2544.0	96.5	1,121	46	2536.3	95.6	1,375	47	2544.1	97.1
Grade 8																
All Students	24,361	55	2574.0	102.5	24,842	54	2570.4	103.4	24,284	51	2562.7	103.3	23,923	53	2564.3	109.8
Female	11,780	62	2591.0	97.0	12,207	60	2586.8	98.8	11,715	57	2578.2	99.4	11,623	59	2580.3	105.6
Male	12,581	49	2558.2	105.0	12,635	48	2554.5	105.3	12,569	45	2548.3	104.8	12,300	47	2549.2	111.5
American Indian/Alaska Native	251	29	2514.8	102.4	264	30	2515.1	102.9	248	28	2504.0	96.7	232	32	2520.9	109.1
Asian	295	69	2614.1	103.5	248	67	2609.8	107.3	215	60	2591.0	103.4	266	69	2602.4	120.3
Black or African American	298	36	2512.3	118.9	330	32	2507.8	109.4	261	27	2495.0	104.9	279	32	2505.3	123.6
Hispanic or Latino	4,529	38	2533.1	100.3	4,654	36	2527.5	97.2	4,536	33	2521.7	97.9	4,570	36	2521.5	107.2
Pacific Islander	203	54	2576.4	96.5	207	56	2571.9	104.1	220	45	2548.3	101.8	194	55	2571.8	99.5
White	18,785	60	2585.0	99.6	19,139	59	2582.1	101.3	18,804	56	2574.2	101.5	18,305	57	2576.0	106.9
EL	1,716	26	2501.7	102.6	1,783	25	2499.2	99.0	1,675	24	2495.1	97.8	2,231	29	2498.5	114.3
Special Education	2,521	10	2453.5	91.9	2,432	10	2449.2	94.4	2,498	9	2446.2	85.6	2,542	8	2435.8	98.0
Section 504 Plan	1,100	45	2554.5	94.2	1,235	48	2557.2	95.9	1,344	46	2554.2	96.8	1,423	46	2553.8	97.8

Table B-4. ELA/L Student Performance Across Years (Grade 11)

Group	2022-2023				2023-2024			
	N	**% Prof**	**Scale Score**	**SD**	**N**	**% Prof**	**Scale Score**	**SD**
All Students	16,602	45	2561.3	123.8	22,710	59	2598.4	122.7
Female	7,902	50	2580.0	118.8	11,052	65	2616.7	115.2
Male	8,700	40	2544.3	125.9	11,658	53	2581.0	126.9
American Indian/Alaska Native	240	27	2510.2	107.2	204	43	2547.3	115.8
Asian	212	69	2631.5	125.6	283	71	2636.1	135.8
Black or African American	202	21	2488.1	113.6	300	35	2514.3	130.9
Hispanic or Latino	3,241	28	2515.5	112.8	4,376	43	2553.5	117.6
Pacific Islander	121	48	2571.4	116.8	174	52	2592.4	118.2
White	12,586	50	2573.9	123.3	17,320	63	2611.3	120.2
EL	1,270	22	2493.4	116.9	1,832	32	2519.3	123.4
Special Education	1,632	9	2446.5	94.2	1,946	13	2460.2	105.3
Section 504 Plan	1,020	43	2556.2	121.4	1,457	55	2588.1	116.8

Note. 2022–2023 is the first year of administering grade 11 tests as an accountability grade in high school.

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

Group	**2020-2021**				**2021-2022**				**2022-2023**				**2023-2024**			
	N	**% Prof**	**Scale Score**	**SD**	**N**	**% Prof**	**Scale Score**	**SD**	**N**	**% Prof**	**Scale Score**	**SD**	**N**	**% Prof**	**Scale Score**	**SD**
Grade 3																
All Students	22,352	48	2426.0	86.5	23,154	51	2432.6	89.1	23,356	49	2430.2	87.7	23,524	50	2430.1	90.3
Female	10,957	45	2420.7	84.1	11,394	49	2427.5	87.5	11,329	46	2424.9	84.2	11,591	46	2423.8	87.3
Male	11,395	51	2431.2	88.4	11,760	54	2437.5	90.4	12,027	51	2435.2	90.6	11,933	53	2436.3	92.7
American Indian/Alaska Native	223	22	2367.5	84.0	229	32	2383.0	86.7	246	24	2386.1	80.2	208	27	2381.9	86.0
Asian	254	64	2469.7	92.4	294	62	2459.0	107.6	257	67	2465.1	95.6	255	65	2460.9	100.5
Black or African American	260	25	2374.4	87.4	281	28	2370.8	95.4	273	25	2372.0	92.5	274	30	2367.1	102.5
Hispanic or Latino	4,045	27	2385.3	83.5	4,135	30	2392.6	83.3	4,492	31	2395.5	81.7	4,566	32	2393.3	87.1
Pacific Islander	211	47	2421.3	88.1	191	44	2420.4	91.6	354	43	2419.3	86.9	261	43	2423.8	85.9
White	17,359	53	2436.5	83.5	18,024	56	2443.0	86.7	17,734	54	2440.2	86.2	17,671	55	2441.3	87.6
EL	2,012	20	2370.2	83.8	2,047	23	2375.6	86.0	1,964	23	2378.0	80.1	2,159	23	2372.6	87.8
Special Education	2,504	20	2356.0	97.1	2,746	24	2361.3	99.1	2,973	23	2361.1	96.6	2,913	20	2354.8	96.4
Section 504 Plan	517	41	2413.3	81.9	546	43	2420.7	84.0	606	45	2423.0	85.0	731	46	2424.8	85.5
Grade 4																
All Students	22,876	45	2470.0	88.2	23,166	49	2477.2	89.0	23,548	47	2473.3	88.9	23,806	48	2475.8	91.4
Female	11,261	42	2464.9	85.6	11,361	45	2470.9	85.5	11,623	43	2466.5	84.9	11,555	44	2469.5	86.9
Male	11,615	48	2475.0	90.4	11,805	52	2483.3	91.8	11,925	50	2479.9	92.2	12,251	51	2481.8	95.0
American Indian/Alaska Native	252	22	2412.2	89.0	217	19	2413.0	82.3	230	25	2430.5	76.8	238	24	2430.1	81.4
Asian	261	58	2501.6	107.9	272	65	2518.2	92.3	293	61	2513.4	104.5	253	68	2519.1	103.6
Black or African American	301	24	2418.7	91.2	266	24	2422.4	87.1	283	20	2407.5	90.5	295	24	2420.9	94.5
Hispanic or Latino	4,280	25	2428.6	82.9	4,174	28	2434.9	85.1	4,412	27	2431.9	82.8	4,676	29	2435.1	87.4
Pacific Islander	169	39	2460.0	86.4	176	48	2468.2	81.7	237	38	2456.9	89.3	214	40	2466.8	92.5
White	17,613	51	2481.4	85.4	18,061	54	2488.0	86.3	18,093	52	2484.5	86.5	17,982	53	2487.9	88.4
EL	2,108	18	2411.2	82.0	2,032	23	2421.5	85.5	2,019	21	2417.9	85.4	2,271	20	2416.4	87.0
Special Education	2,721	16	2391.6	93.6	2,695	18	2397.2	95.6	2,969	17	2394.0	91.8	3,084	16	2393.9	92.9
Section 504 Plan	651	39	2462.0	87.0	667	44	2471.7	82.7	759	42	2466.1	80.6	908	44	2474.9	82.7

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

Group	2020-2021				2021-2022				2022-2023				2023-2024			
	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD
Grade 5																
All Students	23,245	40	2499.3	95.0	23,532	43	2503.4	96.4	23,437	41	2500.8	95.0	23,864	41	2499.6	101.3
Female	11,373	37	2495.4	91.7	11,597	40	2499.5	93.2	11,451	37	2494.2	90.7	11,748	38	2493.6	97.6
Male	11,872	43	2503.0	97.8	11,935	45	2507.3	99.3	11,986	45	2507.2	98.5	12,116	44	2505.5	104.4
American Indian/Alaska Native	271	18	2446.8	92.6	261	23	2453.1	94.8	208	21	2452.7	88.6	242	17	2445.3	93.8
Asian	250	60	2539.2	97.5	276	56	2534.2	113.6	272	61	2552.3	105.1	308	58	2534.8	120.9
Black or African American	294	21	2445.1	102.2	304	23	2449.5	101.2	249	18	2431.6	101.6	285	15	2423.9	110.1
Hispanic or Latino	4,282	21	2454.9	89.7	4,358	25	2461.3	92.1	4,332	21	2455.1	87.7	4,520	22	2454.3	93.9
Pacific Islander	156	40	2493.2	93.6	183	45	2504.5	92.5	234	38	2492.6	85.3	193	38	2490.2	98.0
White	17,992	45	2511.1	92.3	18,150	47	2514.7	93.7	18,142	46	2512.6	92.5	18,200	47	2512.6	98.5
EL	2,031	16	2438.8	91.1	2,014	17	2440.7	93.3	1,933	19	2442.8	93.1	2,288	17	2434.3	96.9
Special Education	2,689	10	2407.8	94.1	2,779	13	2411.1	98.3	2,770	12	2409.8	91.2	2,999	10	2401.4	96.2
Section 504 Plan	766	34	2489.6	83.8	800	40	2501.2	89.4	901	38	2496.7	88.0	1,084	36	2493.3	92.1
Grade 6																
All Students	23,617	37	2511.2	105.7	23,877	41	2519.7	109.0	23,702	39	2514.3	108.2	23,631	40	2516.2	112.0
Female	11,414	35	2508.6	102.1	11,676	38	2516.5	105.7	11,677	37	2511.5	104.8	11,490	38	2512.8	108.9
Male	12,203	38	2513.6	109.0	12,201	43	2522.7	111.9	12,025	40	2517.1	111.4	12,141	42	2519.5	114.6
American Indian/Alaska Native	242	14	2450.4	99.7	272	18	2461.3	108.0	254	19	2453.8	108.4	192	15	2450.7	102.8
Asian	226	48	2533.4	110.4	253	61	2570.9	116.5	273	60	2564.6	131.1	269	59	2578.2	123.4
Black or African American	273	18	2443.7	118.0	282	23	2452.8	127.5	287	15	2437.9	112.6	259	17	2438.8	124.1
Hispanic or Latino	4,346	18	2461.3	101.7	4,379	22	2467.5	106.5	4,507	20	2462.8	105.1	4,526	21	2461.5	109.2
Pacific Islander	152	32	2496.8	101.6	199	39	2514.9	108.0	221	36	2503.3	105.9	213	41	2522.9	103.6
White	18,378	42	2524.6	102.3	18,492	46	2533.3	104.6	18,160	44	2528.6	103.7	18,059	45	2531.3	106.8
EL	1,726	12	2433.9	104.6	1,848	16	2444.0	110.1	1,934	15	2441.4	108.8	2,288	16	2441.4	115.3
Special Education	2,528	8	2395.9	111.0	2,696	8	2398.3	114.8	2,731	9	2397.8	108.0	2,689	8	2393.4	108.4
Section 504 Plan	930	29	2496.5	97.7	923	30	2505.1	92.4	1,078	30	2503.2	96.9	1,264	34	2507.8	100.1

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

Group	2020-2021				2021-2022				2022-2023				2023-2024			
	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD
Grade 7																
All Students	24,473	40	2531.5	112.3	24,259	42	2535.6	111.5	23,974	40	2532.7	110.6	23,859	42	2537.5	115.3
Female	12,011	38	2527.8	109.3	11,701	40	2532.0	107.2	11,688	37	2527.9	106.5	11,748	40	2532.0	113.3
Male	12,462	42	2535.1	115.0	12,558	43	2539.0	115.2	12,286	43	2537.2	114.1	12,111	45	2542.8	117.0
American Indian/Alaska Native	276	21	2477.5	109.1	251	21	2473.0	100.6	248	21	2475.5	103.7	243	22	2478.4	113.0
Asian	246	60	2587.8	133.6	226	50	2570.1	118.8	257	65	2601.2	128.9	265	63	2588.6	135.0
Black or African American	329	18	2460.7	118.4	283	20	2463.2	116.6	280	20	2458.8	109.1	293	19	2459.8	121.8
Hispanic or Latino	4,558	22	2478.6	108.4	4,483	23	2483.9	107.7	4,499	21	2480.0	103.6	4,714	23	2482.7	112.8
Pacific Islander	174	39	2534.0	111.3	207	32	2510.0	107.1	206	39	2525.7	115.5	209	40	2532.4	117.3
White	18,890	45	2545.5	108.1	18,809	47	2549.7	107.9	18,484	45	2546.5	107.2	18,022	48	2553.7	109.9
EL	1,895	15	2451.0	112.0	1,633	16	2454.2	110.4	1,917	17	2458.0	111.0	2,330	17	2458.1	115.9
Special Education	2,579	8	2406.6	112.4	2,497	8	2410.9	106.6	2,658	7	2408.0	98.0	2,622	8	2412.9	106.3
Section 504 Plan	1,036	31	2515.5	101.9	1,159	33	2521.4	102.4	1,121	34	2522.2	97.7	1,378	34	2527.8	102.2
Grade 8																
All Students	24,296	36	2541.5	118.7	24,845	36	2542.0	118.4	24,351	36	2540.6	119.0	24,013	39	2549.8	128.1
Female	11,744	36	2543.0	114.0	12,220	35	2541.9	114.3	11,733	35	2539.2	113.8	11,665	38	2547.9	123.3
Male	12,552	36	2540.1	123.0	12,625	37	2542.2	122.1	12,618	37	2541.9	123.6	12,348	40	2551.6	132.4
American Indian/Alaska Native	250	15	2466.2	111.2	262	17	2476.7	112.0	248	12	2462.3	104.7	228	21	2493.1	125.0
Asian	294	56	2613.3	136.3	252	54	2597.4	142.8	222	50	2583.9	135.9	271	62	2618.8	155.5
Black or African American	299	19	2467.4	131.9	335	17	2466.8	126.5	271	11	2451.2	109.1	293	20	2468.4	127.6
Hispanic or Latino	4,523	18	2487.8	111.1	4,697	17	2487.2	108.9	4,596	18	2486.0	108.1	4,652	21	2490.3	117.5
Pacific Islander	203	32	2539.2	113.9	207	33	2530.9	120.0	218	22	2512.8	110.4	195	34	2546.3	118.5
White	18,727	40	2555.5	115.3	19,092	41	2557.1	115.2	18,796	41	2556.1	116.7	18,270	45	2566.4	124.7
EL	1,712	12	2459.0	112.5	1,805	11	2460.4	109.4	1,755	12	2458.6	111.0	2,371	17	2469.0	125.1
Special Education	2,507	5	2410.6	106.9	2,419	6	2411.6	106.9	2,500	5	2412.4	95.4	2,534	5	2407.4	109.0
Section 504 Plan	1,092	26	2520.5	104.7	1,222	28	2523.4	107.2	1,341	29	2528.4	108.9	1,419	31	2535.0	110.7

Table B-8. Mathematics Student Performance Across Years (Grade 11)

Group	**2022-2023**				**2023-2024**			
	N	**% Prof**	**Scale Score**	**SD**	**N**	**% Prof**	**Scale Score**	**SD**
All Students	18,990	21	2535.8	118.6	23,022	31	2562.9	131.5
Female	9,328	20	2536.4	110.8	11,231	29	2560.4	121.7
Male	9,662	23	2535.2	125.7	11,791	33	2565.2	140.1
American Indian/Alaska Native	252	7	2474.4	99.0	204	12	2484.9	121.7
Asian	227	47	2620.3	148.5	287	51	2631.3	148.6
Black or African American	214	10	2465.8	116.2	301	11	2471.5	128.0
Hispanic or Latino	3,723	9	2485.8	102.4	4,409	14	2505.7	116.0
Pacific Islander	139	24	2546.5	111.9	174	30	2544.0	141.1
White	14,435	25	2549.4	117.9	17,587	36	2578.9	129.6
EL	1,396	10	2477.5	116.4	1,884	11	2485.2	120.0
Special Education	1,709	2	2415.5	88.9	1,942	2	2419.2	105.2
Section 504 Plan	1,153	18	2523.2	112.8	1,483	23	2545.5	123.1

Note. 2022–2023 is the first year of administering grade 11 tests as an accountability grade in high school.

Appendix C: Classification Accuracy and Consistency Indexes by Subgroup

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

Group	N	% Accuracy						% Consistency					
		All	L1	L2	L3	L4	Proficiency Cut	All	L1	L2	L3	L4	Proficiency Cut
Grade 3													
All Students	23,374	73	88	60	56	84	90	65	80	49	45	76	86
Female	11,507	73	87	60	56	84	90	64	79	49	45	76	86
Male	11,867	74	88	60	56	84	90	65	81	49	45	75	86
American Indian/Alaska Native	207	77	90	61	57	80	93	70	85	49	43	71	89
Asian	251	74	88	59	57	87	90	65	79	46	48	79	86
Black or African American	261	77	90	59	56	83	92	70	87	46	46	71	89
Hispanic or Latino	4,464	75	88	61	56	82	91	67	83	49	44	71	87
Pacific Islander	261	71	86	61	56	83	88	61	76	51	46	72	83
White	17,666	73	87	60	56	84	90	64	79	49	46	77	86
EL	2,006	77	90	61	56	80	92	70	86	49	44	67	89
Special Education	2,912	81	92	60	56	80	94	74	89	48	43	68	91
Section 504 Plan	719	73	88	60	56	83	90	64	80	49	45	74	86
Grade 4													
All Students	23,631	73	88	53	55	84	90	65	81	42	45	76	86
Female	11,477	72	87	53	55	85	90	64	79	42	45	77	85
Male	12,154	73	89	53	55	84	90	65	82	42	45	76	86
American Indian/Alaska Native	238	76	89	53	55	83	91	68	85	42	42	72	87
Asian	249	74	92	53	55	87	89	66	76	44	43	82	85
Black or African American	276	77	90	53	56	80	92	70	87	41	42	72	89
Hispanic or Latino	4,564	74	90	53	55	82	90	66	84	42	45	70	86
Pacific Islander	215	74	91	53	54	86	90	66	84	40	47	76	86
White	17,979	72	88	53	55	85	89	64	79	42	45	77	85
EL	2,092	77	91	53	55	81	91	70	87	42	44	65	88
Special Education	3,077	83	93	53	55	82	94	77	91	41	43	68	92
Section 504 Plan	903	71	86	54	55	84	89	62	78	43	44	74	84

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

Group	**N**	**% Accuracy**						**% Consistency**					
		All	**L1**	**L2**	**L3**	**L4**	**Proficiency Cut**	**All**	**L1**	**L2**	**L3**	**L4**	**Proficiency Cut**
Grade 5													
All Students	23,742	74	88	56	65	84	90	65	80	44	54	76	86
Female	11,701	74	87	56	65	84	90	65	79	44	54	77	86
Male	12,041	74	88	56	65	83	90	66	81	44	54	75	86
American Indian/Alaska Native	241	76	88	56	65	80	91	68	84	45	52	69	88
Asian	298	79	90	55	66	89	90	71	86	39	58	84	87
Black or African American	269	76	91	57	63	77	90	68	85	46	52	64	86
Hispanic or Latino	4,426	75	89	56	64	81	90	66	83	44	54	68	86
Pacific Islander	193	74	88	56	65	80	91	66	83	43	55	71	87
White	18,229	74	87	56	65	84	90	65	79	44	55	76	86
EL	2,125	78	90	56	64	81	91	70	86	44	53	69	87
Special Education	2,998	83	92	56	65	82	93	77	90	43	51	66	90
Section 504 Plan	1,080	72	87	56	64	83	89	63	78	45	55	72	85
Grade 6													
All Students	23,513	75	88	66	69	81	91	66	80	54	60	71	87
Female	11,436	75	87	66	69	82	90	65	79	54	60	73	86
Male	12,077	75	88	66	69	80	91	67	82	54	60	70	87
American Indian/Alaska Native	191	79	91	67	70	80	91	70	86	53	62	62	87
Asian	263	77	88	66	69	85	93	68	78	55	59	79	89
Black or African American	251	78	90	66	68	78	91	70	87	52	60	62	88
Hispanic or Latino	4,435	77	89	66	69	79	91	68	84	55	59	65	87
Pacific Islander	213	74	86	65	70	82	91	65	76	54	61	72	87
White	18,077	75	87	66	69	82	90	65	79	54	60	72	86
EL	2,140	79	90	66	69	80	93	71	86	55	58	65	90
Special Education	2,697	85	93	65	68	76	96	80	91	53	54	60	94
Section 504 Plan	1,259	74	87	65	68	81	90	65	78	55	59	68	85

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

Group	N	% Accuracy						% Consistency					
		All	L1	L2	L3	L4	Proficiency Cut	All	L1	L2	L3	L4	Proficiency Cut
Grade 7													
All Students	23,766	76	88	63	72	82	91	67	80	52	63	72	87
Female	11,710	75	87	63	72	83	91	66	78	52	63	73	87
Male	12,056	76	89	64	72	81	91	67	82	52	63	70	87
American Indian/Alaska Native	243	77	89	66	70	78	91	69	84	55	61	64	88
Asian	261	78	91	64	72	86	93	70	83	51	65	78	90
Black or African American	281	79	91	66	71	84	91	71	87	53	64	63	87
Hispanic or Latino	4,649	77	90	63	71	80	91	69	84	53	62	67	87
Pacific Islander	208	77	88	62	74	85	90	68	81	52	65	75	87
White	18,040	75	87	63	72	82	91	66	78	51	64	72	87
EL	2,199	79	91	63	72	81	92	72	87	52	61	67	88
Special Education	2,623	84	92	63	70	81	94	78	90	52	58	61	92
Section 504 Plan	1,375	74	85	63	72	82	90	65	77	53	63	68	86
Grade 8													
All Students	23,923	76	88	66	73	81	91	67	80	55	64	70	87
Female	11,623	76	87	67	73	82	91	66	78	55	65	71	87
Male	12,300	77	88	66	73	81	91	68	82	55	64	69	87
American Indian/Alaska Native	232	77	88	66	73	78	91	68	82	56	61	69	88
Asian	266	78	90	66	72	85	92	70	82	51	63	79	89
Black or African American	279	80	91	66	73	82	93	72	86	56	62	71	90
Hispanic or Latino	4,570	77	89	67	73	79	91	69	83	56	64	64	87
Pacific Islander	194	75	89	66	73	86	89	66	77	55	66	70	84
White	18,305	76	87	66	73	82	90	66	79	55	64	71	87
EL	2,231	80	91	67	73	82	92	72	86	56	63	67	89
Special Education	2,542	85	92	66	71	80	96	79	90	54	56	62	94
Section 504 Plan	1,423	75	85	67	73	82	90	66	76	55	65	68	86

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

Group	**N**	**% Accuracy**						**% Consistency**					
		All	**L1**	**L2**	**L3**	**L4**	**Proficiency Cut**	**All**	**L1**	**L2**	**L3**	**L4**	**Proficiency Cut**
Grade 11													
All Students	22,710	76	87	66	69	84	91	67	80	54	60	76	88
Female	11,052	76	85	66	69	84	91	67	76	54	60	77	88
Male	11,658	77	88	66	69	84	92	68	82	54	60	76	88
American Indian/Alaska Native	204	74	86	65	69	78	90	65	79	53	61	64	87
Asian	283	80	91	68	69	89	94	72	84	54	60	82	91
Black or African American	300	79	90	66	69	80	92	71	87	52	61	67	89
Hispanic or Latino	4,376	76	87	66	69	81	90	67	81	55	60	71	87
Pacific Islander	174	77	89	66	69	88	90	68	81	56	58	79	87
White	17,320	76	87	66	69	84	92	67	79	54	60	77	88
EL	1,832	78	89	66	69	82	92	70	84	55	59	70	88
Special Education	1,946	82	90	66	69	81	95	75	87	53	56	63	92
Section 504 Plan	1,457	75	88	65	69	82	90	66	79	54	60	73	86

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

Group	N	% Accuracy						% Consistency					
		All	**L1**	**L2**	**L3**	**L4**	**Proficiency Cut**	**All**	**L1**	**L2**	**L3**	**L4**	**Proficiency Cut**
Grade 3													
All Students	23,524	77	86	66	71	86	92	69	80	53	62	79	88
Female	11,591	77	86	65	71	85	91	68	80	53	62	77	88
Male	11,933	78	86	66	71	87	92	69	80	53	62	80	89
American Indian/Alaska Native	208	80	88	65	72	88	94	72	84	50	61	80	91
Asian	255	80	86	63	71	90	94	72	80	51	62	84	92
Black or African American	274	78	85	64	70	80	93	70	84	47	64	64	90
Hispanic or Latino	4,566	78	87	66	71	84	93	70	83	53	61	75	90
Pacific Islander	261	75	84	66	70	85	91	66	76	56	59	78	88
White	17,671	77	85	65	71	86	91	68	78	53	62	79	88
EL	2,159	80	87	66	72	83	94	72	85	51	60	73	92
Special Education	2,913	81	87	65	71	83	95	73	86	49	59	74	92
Section 504 Plan	731	77	85	66	72	86	92	68	79	53	62	79	89
Grade 4													
All Students	23,806	78	87	72	71	86	92	70	80	62	60	80	89
Female	11,555	78	86	72	70	85	92	69	79	63	60	77	88
Male	12,251	79	87	73	71	87	92	71	80	62	61	81	89
American Indian/Alaska Native	238	80	89	73	73	79	93	72	84	63	58	72	90
Asian	253	82	87	74	71	90	94	75	83	61	61	86	91
Black or African American	295	80	88	71	72	83	93	73	84	59	60	78	91
Hispanic or Latino	4,676	79	88	72	70	83	93	71	83	62	59	73	90
Pacific Islander	214	78	84	71	70	89	92	70	76	62	58	84	89
White	17,982	78	86	73	71	87	92	70	78	63	61	80	88
EL	2,271	81	88	72	70	88	94	73	84	61	58	76	91
Special Education	3,084	84	91	72	70	85	96	77	88	59	58	76	94
Section 504 Plan	908	78	84	73	71	87	92	69	75	64	61	79	88

Table C-6. Mathematics Classification Accuracy and Consistency by Subgroup (Grades 5 and 6)

Group	N	% Accuracy						% Consistency					
		All	L1	L2	L3	L4	Proficiency Cut	All	L1	L2	L3	L4	Proficiency Cut
Grade 5													
All Students	23,864	77	88	67	60	87	92	69	82	57	48	80	89
Female	11,748	77	88	67	59	86	92	68	82	57	47	79	88
Male	12,116	77	88	67	60	87	92	69	82	56	48	81	89
American Indian/Alaska Native	242	80	88	67	60	87	94	72	85	55	44	75	92
Asian	308	80	88	66	60	91	93	74	84	52	50	87	91
Black or African American	285	84	90	68	64	90	95	77	89	53	47	80	93
Hispanic or Latino	4,520	79	89	67	59	82	93	71	85	56	46	72	91
Pacific Islander	193	77	88	69	63	82	92	68	83	56	49	75	88
White	18,200	76	87	68	60	87	91	68	80	57	48	81	88
EL	2,288	82	90	68	59	82	95	75	87	55	46	73	93
Special Education	2,999	86	92	67	59	82	97	80	91	50	45	73	95
Section 504 Plan	1,084	76	86	68	60	86	91	67	81	56	48	78	88
Grade 6													
All Students	23,631	77	90	69	61	86	91	69	84	59	49	78	88
Female	11,490	77	89	68	61	85	91	69	83	58	49	77	88
Male	12,141	78	90	69	61	86	91	70	84	59	50	79	88
American Indian/Alaska Native	192	81	91	68	59	76	95	75	87	59	44	70	92
Asian	269	81	87	67	61	92	93	74	80	58	49	89	90
Black or African American	259	85	95	69	62	87	94	79	92	61	45	79	90
Hispanic or Latino	4,526	81	92	69	61	84	93	74	88	58	48	71	90
Pacific Islander	213	76	88	67	61	88	91	68	81	57	51	79	87
White	18,059	76	88	69	61	86	91	68	81	59	50	79	87
EL	2,288	84	93	69	61	85	94	78	90	59	47	76	91
Special Education	2,689	89	95	67	60	84	96	85	94	54	45	73	93
Section 504 Plan	1,264	77	89	69	61	85	91	68	82	59	49	75	87

Table C-7. Mathematics Classification Accuracy and Consistency by Subgroup (Grades 7 and 8)

Group	N	% Accuracy						% Consistency					
		All	L1	L2	L3	L4	Proficiency Cut	All	L1	L2	L3	L4	Proficiency Cut
Grade 7													
All Students	23,859	77	88	66	64	86	91	68	82	56	53	78	87
Female	11,748	76	88	67	64	85	91	68	82	56	52	77	87
Male	12,111	77	89	66	64	86	91	68	82	56	53	79	87
American Indian/Alaska Native	243	79	90	66	63	80	91	72	87	55	50	70	87
Asian	265	80	89	65	63	91	92	74	84	52	54	87	89
Black or African American	293	82	92	66	60	80	92	75	89	55	48	71	89
Hispanic or Latino	4,714	79	90	66	63	84	92	72	86	55	51	73	88
Pacific Islander	209	76	90	65	59	89	90	69	82	56	51	81	87
White	18,022	76	87	67	64	86	90	67	79	56	53	78	86
EL	2,330	82	91	65	64	85	93	76	88	54	51	73	89
Special Education	2,622	88	94	66	62	83	94	83	93	53	47	72	91
Section 504 Plan	1,378	75	87	67	64	85	90	66	79	58	52	75	86
Grade 8													
All Students	24,013	76	87	62	59	88	91	67	81	50	48	80	88
Female	11,665	75	87	62	59	87	91	67	80	51	48	79	87
Male	12,348	76	87	61	59	88	92	68	81	50	47	82	88
American Indian/Alaska Native	228	78	86	63	61	88	94	69	83	48	47	75	92
Asian	271	81	89	61	60	92	93	75	84	48	49	89	90
Black or African American	293	79	89	60	60	84	94	72	86	45	50	69	92
Hispanic or Latino	4,652	78	89	61	59	85	93	71	85	49	46	74	90
Pacific Islander	195	75	86	60	61	89	91	66	79	51	47	81	87
White	18,270	75	86	62	59	88	91	67	79	51	48	81	87
EL	2,371	82	90	61	59	88	95	75	87	46	45	78	93
Special Education	2,534	87	92	61	59	85	97	82	91	42	41	72	96
Section 504 Plan	1,419	74	85	60	59	86	91	65	79	50	46	76	87

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

Group	**N**	**% Accuracy**						**% Consistency**					
		All	**L1**	**L2**	**L3**	**L4**	**Proficiency Cut**	**All**	**L1**	**L2**	**L3**	**L4**	**Proficiency Cut**
Grade 11													
All Students	23,022	79	89	63	69	86	92	71	85	52	58	76	89
Female	11,231	78	89	64	69	84	92	70	84	53	57	74	89
Male	11,791	79	90	63	69	87	93	72	85	52	58	78	90
American Indian/Alaska Native	204	85	93	62	68	–	96	79	90	50	54	–	93
Asian	287	79	88	62	71	89	92	71	82	52	59	86	89
Black or African American	301	85	92	64	67	–	95	79	90	50	53	–	93
Hispanic or Latino	4,409	82	91	64	69	83	95	76	88	50	55	71	92
Pacific Islander	174	81	93	62	70	87	93	74	88	51	60	81	90
White	17,587	77	88	63	69	86	92	69	83	53	58	77	88
EL	1,884	85	92	63	69	87	96	79	90	49	55	76	94
Special Education	1,942	92	95	63	69	–	99	89	95	43	49	–	98
Section 504 Plan	1,483	79	88	63	69	86	93	71	84	52	56	76	90

Note. "–" Suppressed data due to the small sample size for the performance level, *n* < 10.