Delaware Smarter Balanced Assessments

2018–2019 Technical Report

Submitted to
Delaware Department of Education
by the American Institutes for Research

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EXHIBITS

LIST OF APPENDICES

1. OVERVIEW

This report provides a technical summary of the 2018–2019 Delaware administration of Smarter Balanced summative assessments in English language arts/literacy (ELA/lit) and mathematics at grades 3–8. The report includes eight chapters: Overview, Test Administration, Summary of 2018–2019 Operational Test Administration, Validity, Reliability, Scoring, Reporting and Interpreting Scores, and Quality Control Procedures. For the interim assessments, the number of students who took ICAs and IABs and their performance are provided in Appendix A. The data included in this report are based on the Delaware Smarter Balanced test scores in ELA/lit and mathematics.

While this report includes information on all aspects of the technical quality of the Smarter Balanced test administration in Delaware, the information on test design, the process for item and test development, alignment study, standard setting, and other information about the technical characteristics can be found in the Smarter Balanced technical documentations. 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.

1.1. SMARTER BALANCED ASSESSMENTS IN DELAWARE

The Smarter Balanced Assessment Consortium (SBAC) developed a next-generation assessment system. The assessments are designed to measure the Common Core State Standards (CCSS) in ELA/lit and mathematics for grades 3–8 and 11 and to provide valid, reliable, and fair test scores about student academic achievement. Delaware was among the 18 member states (plus the U.S. Virgin Islands) leading the development of assessments in ELA/lit and mathematics. The system includes both summative assessments for accountability purposes, as well as optional interim assessments that provide meaningful feedback and actionable data that teachers and educators can use to help students succeed. Smarter Balanced, a state-led enterprise, is intended to provide leadership and resources to improve teaching and learning by creating and maintaining a suite of summative and interim assessments and tools aligned to the CCSS in ELA/lit and mathematics.

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

- **Computer-Adaptive Test.** The CAT is an online adaptive test that provides an individualized assessment for each student.
- **Performance Task.** A *performance task* is a task that challenges students to apply their knowledge and skills to respond to real-world problems. Performance tasks can best be described as collections of questions and activities that are coherently connected to a single theme or scenario. They are used to better measure capacities such as depth of understanding, research skills, and complex

analysis, none of which can be adequately assessed with selected- or constructed-response items. Some performance task items can be scored by the computer, but most are handscored.

Optional interim assessments allow teachers to check student progress throughout the year and give them information that they can use to improve instruction and learning. These tools are used at the discretion of schools and districts, and teachers can employ them to check students' progress in mastering specific concepts at strategic points during the school year. The interim assessments are available as fixed-form tests and consist of the following features:

- **Interim Comprehensive Assessments (ICAs)** test the same content and report scores on the same scale as the summative assessments.
- **Interim Assessment Blocks (IABs)** focus on specific sets of related concepts and provide more detailed information about student learning.

The Delaware State Board of Education formally adopted the CCSS in ELA/lit and mathematics on August 19, 2010 (State Board meeting minutes, 2010). Delaware CCSS defines the knowledge and skills that students need to succeed in college and careers after graduating from high school. These standards include rigorous content and application of knowledge through higher-order skills and align with college and workforce expectations.

Since the adoption of the CCSS in 2010, the Delaware Department of Education fully implemented the CCSS in all grade levels in school year 2013–2014. The new Delaware statewide assessments in ELA/lit and mathematics aligned with the CCSS were administered for the first time in spring 2015 to students in grades 3–8 and 11 in all public schools. In school year 2015–2016, Delaware adopted the SAT to replace the Smarter Balanced grade 11 assessments for high school students. The American Institutes for Research (AIR) delivered and scored the Smarter Balanced assessments and produced score reports. Measurement Incorporated (MI) scored the handscored items.

1.2. CHANGES IN 2018–2019 SUMMATIVE ASSESSMENTS

In 2018, Smarter Balanced updated test blueprints of the ELA/lit summative assessment, shortening the test length by three to four items. The updated test blueprints were implemented in the 2018–2019 test administration.

The purpose of the changes in the test blueprints was to reduce testing time burden by removing short answer items while keeping the claim and target coverage specified in the initial test blueprint and ensuring the same test score reliability as in previous years.

In the CAT component, the requirement for short answer items in claim 1 reading and claim 2 writing were removed from grades 3–5 assessments while no such changes in grades 6–8 assessments. Four items were reduced in claim 2 writing, and two items were added in claim 4 research in all grades. In the PT component, one to two research items were removed and thus PT component consisted of one research item and one full-write item. Overall the total test length (CAT and PT combined) was shortened by three to four items in ELA/lit summative assessments across grades. Table 1 summarizes the changes in the test blueprints in ELA/lit.

Table 1. Changes in ELA/Lit Test Blueprints

Component	Claim	2017–2018 BP	2018–2019 BP	Changes in 2018–2019 BP
CAT	Claim 1 Reading *	14–19	14–19	*Grades 3–5*: Removed zero-to-one short answer item requirement *Grades 6–8*: No change
	Claim 2 Writing	10	6	*Grades 3–5*: Removed one brief-write item requirement and reduced the item requirement by four items *Grades 6–8, 11*: Reduced the item requirement by four items
	Claim 3 Listening	8–9	8–9	*All grades*: no change
	Claim 4 Research	6	8	*All grades*: Added two items
PT	Claim 4 Research	2–3	1	*All grades*: Kept one DOK 3 item, with preference for machine-scored item
	Claim 2 Full-Write	1	1	*All grades*: no change

* Required items for claim 1 reading are 14–16 in grades 3–5, 14–19 in grades 6–7, 16–19 in grade 8, and 15–16 in grade 11.

1.3. IMPACT OF CHANGES IN 2018–2019 ELA/LIT TEST BLUEPRINTS

The impacts of changes in the ELA/lit test blueprints on testing in Delaware are presented in Tables 2 and 3. As expected, the overall testing time was reduced for all grades, more impact in grades 3–5 than in grades 6–8 because of the removal of short answer items in claims 1 and 2 in grades 3–5. The decrease in overall testing time for grades 3–5 was estimated to be 35–40 minutes on the average, and 43–54 minutes at the 80th percentile. The decrease in overall testing time for grades 6–8 was estimated to be 10–33 minutes on the average and 8–37 minutes at the 80th percentile.

Table 2. Changes in Testing Time

Grade	Overall			CAT			PT		
	Mean	Median	80th	Mean	Median	80th	Mean	Median	80th
2017–2018 Testing Time									
3	5:04	4:26	6:54	2:26	2:09	3:09	2:38	2:11	3:49
4	5:27	4:53	7:21	2:39	2:23	3:28	2:48	2:23	3:58
5	5:17	4:49	6:57	2:35	2:23	3:23	2:41	2:22	3:41
6	4:49	4:23	6:15	2:26	2:16	3:09	2:22	2:01	3:18
7	4:08	3:50	5:19	2:07	2:00	2:43	2:00	1:48	2:48
8	3:58	3:39	5:10	2:01	1:52	2:35	1:57	1:43	2:44
2018–2019 Testing Time									
3	4:25	3:56	6:00	2:07	1:52	2:46	2:18	1:57	3:21
4	4:47	4:16	6:28	2:16	2:01	2:57	2:31	2:09	3:37
5	4:42	4:15	6:14	2:14	2:04	2:53	2:27	2:09	3:26
6	4:16	3:55	5:38	2:15	2:06	2:54	2:01	1:45	2:52
7	3:48	3:30	4:57	1:56	1:48	2:30	1:52	1:38	2:34
8	3:48	3:31	5:02	1:57	1:49	2:31	1:52	1:36	2:37
Decrease in Testing Time									
3	0:39	0:30	0:54	0:19	0:17	0:23	0:20	0:14	0:28
4	0:40	0:37	0:53	0:23	0:22	0:31	0:17	0:14	0:21
5	0:35	0:34	0:43	0:21	0:19	0:30	0:14	0:13	0:15
6	0:33	0:28	0:37	0:11	0:10	0:15	0:21	0:16	0:26
7	0:20	0:20	0:22	0:11	0:12	0:13	0:08	0:10	0:14
8	0:10	0:08	0:08	0:04	0:03	0:04	0:05	0:07	0:07

The test score reliabilities are estimated to be similar for the overall scores and for claims 1, 3, and 4 scores. The reliability for claim 2 writing scores decreased slightly because of the reduction of number of items from 11 to 7 items in combined CAT and PT tests.

Table 3. Changes in Test Score Reliabilities

Grade	Total Score	Claim 1 Reading	Claim 2 Writing	Claim 3 Listening	Claim 4 Research
2017–2018 Administration					
3	0.92	0.77	0.80	0.60	0.71
4	0.92	0.74	0.79	0.60	0.70
5	0.92	0.76	0.79	0.61	0.75
6	0.93	0.77	0.81	0.63	0.70
7	0.92	0.79	0.80	0.59	0.69
8	0.93	0.78	0.80	0.62	0.71
2018–2019 Administration					
3	0.92	0.76	0.74	0.61	0.73
4	0.92	0.76	0.72	0.62	0.72
5	0.92	0.75	0.72	0.63	0.76
6	0.92	0.77	0.74	0.64	0.72
7	0.92	0.79	0.74	0.59	0.73
8	0.92	0.76	0.74	0.62	0.74

2. TEST ADMINISTRATION

2.1 TESTING WINDOWS

The 2018–2019 Delaware Smarter Balanced assessment testing window spanned approximately three months for grades 3–8 for the online administration of summative assessments and extended over the full school year for the interim assessments. The paper-pencil, fixed-form summative assessments were administered over one month during the online summative testing window. Table 4 shows the schedule for the 2018–2019 Smarter Balanced assessments.

Table 4. 2018–2019 Testing Windows

Tests	Grades	Start Date	End Date	Mode
Summative Assessments	3–8	3/4/2019	5/31/2019	Online Adaptive
	3–8	4/1/2019	4/30/2019	Paper Fixed-Form
Interim Comprehensive Assessments	3–8	8/27/2018	7/18/2019	Online Fixed-Form
Interim Assessment Blocks	3–8	8/27/2018	7/18/2019	Online Fixed-Form

2.2 TEST OPTIONS AND ADMINISTRATIVE ROLES

Smarter Balanced English language arts/literacy (ELA/lit) and mathematics assessments are primarily administered online. To ensure that all eligible students in tested grades were given the opportunity to take the assessments, a number of options were available for the 2018–2019 administration to accommodate students with special needs. Table 5 lists the testing options that were offered, which might applied to one or both content areas. .

Table 5. 2018–2019 Testing Options

Assessment	Testing Options	Test Mode
Summative Assessments	English	Online
	Braille	Online
	Spanish (Mathematics Only)	Online
	Paper-Pencil, Fixed-Form	Paper-Pencil
	Braille Hybrid Adaptive Form	Paper-Pencil
Interim Assessments	English	Online
	Braille	Online
	Spanish (Mathematics Only)	Online

To ensure standardized administration conditions, test administrators (TAs) must follow the procedures outlined in the *Smarter Balanced ELA/Lit and Mathematics Online Summative Test Administration Manual* (TAM). TAs must review the TAM 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 the standardized conditions. Make-up procedures should be established for any students who are absent on the testing days.

2.2.1 Administrative Roles

The key personnel involved with test administration are district test coordinators (DTCs),district accommodations managers (DAMs), school test coordinators (STCs), and test administrators (TAs). The main responsibilities of these key personnel are described below. More detailed descriptions can be found in the TAM, provided online at the Delaware System of Student Assessments (DeSSA) portal, http://de.portal.airast.org.

District Test Coordinator

DTCs are responsible for coordinating testing in their districts. They ensure that STCs and TAs in their districts are appropriately trained and aware of policies and procedures. DTCs also ensure that their STCs are trained in the reporting system.

DTC responsibilities include the following:

- Oversee all test administration-related activities in the district.
- Complete all required DeSSA trainings
- Complete all required DeSSA security forms
- Finalize testing schedules and requirements with STCs
- Ensure that all STCs and TAs are trained to administer the Smarter Balanced assessments properly
- Ensure that all STCs and TAs understand and follow the protocols in the event that a student moves to a new district and/or school
- Ensure that all STCs and TAs are appropriately trained regarding the test security policies and procedures
- Ensure that all STCs and TAs have completed DeSSA security forms
- Create and manage appeals through the Test Information Distribution Engine (TIDE)
- Review and submit incidents, exemptions, security incidents, and data reviews to the Delaware Department of Education (DDOE) via the KACE/DOE help desk (the DeSSA request system)

District Accommodations Manager

DAMs are responsible for ensuring that student accommodations are correctly entered into TIDE. DAM responsibilities include the following:

- Complete the DAM training
- Update the accessibility features in TIDE
- Report or submit security issues, data reviews, unique accommodations, and exemption requests during the testing window via KACE/DOE help desk

School Test Coordinator

STCs coordinate the administration of the Smarter Balanced assessments and ensure that testing operates smoothly and properly at the school level. STC responsibilities include the following:

- Oversee all test administration-related activities in the school
- Complete the STC training
- Complete required security forms for reporting incidents
- Ensure that all TAs complete Smarter Balanced assessment training modules
- Ensure that the DeSSA secure browser has been installed and works properly for test administration
- Develop the test schedule
- Review student records on the Delaware Student Information System (DELSIS) and TIDE applications before testing
- Ensure that all TAs understand and follow the protocols for student relocation
- Ensure that all students in the Department of Services for Children, the Youth and Their Families (DSCYF), Delaware Adolescent Program, Inc. (DAPI), or the Consortium Discipline Alternative Program (CDAP) have a homeschool record
- Ensure that accommodations have been reviewed and updated in TIDE
- Report or submit security issues, incidents, data reviews, unique accommodations, and exemptions via the KACE/DOE help desk

Test Administrator

TAs are qualified personnel who administer the Smarter Balanced assessments. The pool of TAs may include the following authorized personnel:

- Delaware-certified educators (teachers, administrators, or guidance counselors).
- Paraprofessionals, if closely supervised by a Delaware-certified educator.
- Translators. If they are not Delaware-certified educators, they must be closely supervised by a Delaware-certified educator.
- Substitute teachers. If they are not Delaware-certified educators, they must be closely supervised by a Delaware-certified educator.

If there is a severe shortage of staff, a test can be administered by the following:

- Student-teachers acting as TAs, if closely supervised by a Delaware-certified educator
- Student-teachers and school support staff acting as proctors

TAs responsibilities include the following:

- Complete the Smarter Balanced training
- Review necessary manuals and user guides

- Review student information for accuracy before testing to ensure that each student receives the right testing materials and/or is tested with the appropriate accommodations and supports
- Report any errors in student information to the KACE/DOE help desk for corrections
- Prepare the testing environment, ensuring that students have the necessary equipment and materials as appropriate (e.g., scratch paper, pencils, rulers, etc.)
- Administer the Smarter Balanced assessments
- Report all potential test security incidents and irregularities to the STC and/or DTC by following the security procedures
- Securely dispose of all testing materials including print-on-demand documents, scratch paper, and performance task (PT) materials

2.2.2 Online Test Administration

Within the state's testing window, each school needs to set testing schedules to use the testing rooms and facilities efficiently, allow multiple sessions for students to complete the test, and minimize the interruptions of classroom instruction.

STCs oversee all aspects of testing at their school level and serve as the main point of contact, while TAs administer the online assessments only. TAs are trained in the online testing requirements and the mechanics of starting, pausing, and ending a test session. Training materials for test administration are available online. All school personnel who serve as TAs must complete the required DeSSA training courses listed on the DeSSA portal at http://de.portal.airast.org. Before testing, DAMs are responsible for ensuring that student accommodations are correctly entered into TIDE.

To start a test session, the TA must first log in to the TA Interface of the online test delivery system (TDS). A test session ID is generated when the test session is created. The TA reads the *Directions for Administration* in the *Smarter Balanced ELA/Literacy and Mathematics Online Test Administration Manual* to students and guides them through the login process. Students who are taking the assessment need to enter their Statewide Student Identifier (SSID), their first name, and the test session ID into the Student Interface using computers provided by the school. The TA then verifies that the student is taking the appropriate assessment with the appropriate accessibility features. (See Section 2.6, Online Testing Features and Accommodations, for a list of accommodations.) Students can begin testing only when the TA confirms the settings.

Once the assessment has started, students must answer all the test questions presented on one page before proceeding to the next page. Skipping questions 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 the session has not been paused for more than 20 minutes before the assessment was submitted. During an active CAT session, if a student reviews and changes the response to a previously answered item, all the following items to which the student already responded remain the same. No new items are assigned to this student because he or she changed one or more responses. For example, assume a student paused for 10 minutes after completing item 10. After the pause, the student went back to item 5 and changed the response. If the response change in item 5 changed the item score from wrong to right, the student's overall score would improve; however, there would be no change in items 6–10.

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

The summative assessment may be started in one test session and completed in a different session. The CAT must be completed within 45 calendar days of the start date, or the assessment will expire. The PT must be completed within 20 calendar days of the start date.

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

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

2.2.3 Paper-Pencil Test Administration

The paper-pencil version of the Smarter Balanced ELA/lit and mathematics assessments is provided as an accommodation for students who cannot access a computer and students with blindness or visual impairment. Although the online braille form was available, only the paper-pencil braille test was used in Delaware in the 2018–2019 administration.

The non-embedded support for the paper-pencil version must be set by the deadline in TIDE to ensure the on-time delivery of the paper-pencil test booklets with the initial shipment. To receive the braille paper-pencil materials, the request for the non-embedded accommodation for braille (paper-pencil version) must also be set in TIDE by the deadline. The list of requests is extracted from TIDE for DDOE approval. After the request is approved, the testing contractor ships the corresponding test booklets to the school district. The DTC can enter additional orders into TIDE after the school district receives the initial order. Additional orders for paper-pencil test materials must be approved by DDOE if the request exceeds 50 test booklets or if the request is for one or more braille test booklets.

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

After the student completes the assessment, the DTC returns the test booklets to the testing contractor to scan the response documents and score the test.

The total number of students who took paper-pencil tests is presented in Table 6.

Table 6. Number of Students who Took Paper-Pencil Tests
in 2018–2019 Summative Test Administration

Subject	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Total
ELA/Lit	20	29	33	3	1	1	87
Mathematics	20	28	33	3	1	1	86

2.2.4 Braille Test Administration

The adaptive braille test was available with the same test blueprint in English in both ELA/lit and mathematics. In the 2018–2019 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 DDOE. All items on the Braille HAT can be presented to the students using a Refreshable Braille Display (RBD).

The braille interface is described as follows in several formats:

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

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

2.3 Training and Information for Test Coordinators and Administrators

All DTCs, DAMs, STCs, TAs, and school administrative staff who will be involved in Smarter Balanced administration must complete the Smarter Balanced Test Administrator Training Modules. Modules include security, test administration, and other information related to the administration of Smarter Balanced assessments. Successful completion of training is required before the administration of Smarter Balanced assessments. More detailed information can be found in the *Smarter ELA/Literacy and Mathematics Online Summative Test Administration Manual*, provided at the DeSSA portal at http://de.portal.airast.org.

Before administering a Smarter Balanced assessment, all individuals participating in, or otherwise associated with, any test administration must complete the training requirements in Table 7 and read the applicable manuals relevant to their roles. Table 7 presents the training requirements based on roles.

Table 7. Smarter Balanced Assessment Training Requirements

Role	Required Training	Components of the Required Training	Estimated Time to Complete
All Roles	Optional Training: Introduction to TIDE	• TIDE Training	• 40 min.
Smarter Balanced Summative Test Administrator	Smarter Balanced Summative TA Training	• Smarter Balanced Summative TA Training	• 30 min.
Smarter Balanced Interim Test Administrator	Smarter Balanced Interim TA Training	• DeSSA Overview • Smarter Balanced Interim TA Training • AVA Training • AIRWays Training	• 30 min. • 5 min. • 30 min.
Staff Performing Accommodations Data Entry	District and School Accommodations Manager Training	• DeSSA Accommodations Overview • DeSSA ELA/Mathematics Accommodations/ Supports Entry	• 50 min.
Special Education Staff/Coordinator, English Language Learners Staff/Coordinator, General Education with Supports Staff/Coordinator	Accessibility Coordinator Training	• DeSSA Overview • Accessibility	• 30 min. • 50 min.
All Building Staff	Security Training	• Security Module Only	• 30 min.
TAs Who Are Administering the Paper-Pencil Assessment Only* (if the TA is giving online and paper-pencil assessments, take these and the online requirements)	DeSSA Paper-Pencil TA Training for Smarter Balanced	• Paper-Pencil TA Training • Security Training • DeSSA Overview	• 20 min. • 30 min. • 30 min.

* Paper-pencil TAs must also take the TA Training for the relevant test.

2.3.1 Practice and Training Site

In August 2018, separate training sites were opened for TAs and students. TAs can practice administering an assessment by performing tasks such as starting and ending a test session on the TA Training Site. Students can take an online practice test on the Student Practice and Training Site. The Smarter Balanced assessment practice tests mirror the corresponding summative assessments. Each test provides students with a grade-specific testing experience, and students can practice with a variety of question types and levels of difficulty (approximately 30 items each in mathematics and ELA/lit) and practice the PT.

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

A student can log in directly to the practice and training test site as a guest without a TA-generated test session ID number, or the student can log in through a training test session created by the TA in the TA

Training Site. The student training test includes all item types in the operational item pool, including multiple-choice, grid, and natural-language items.

2.3.2 Manuals and User Guides

The manuals and user guides in Table 8 are available on the DeSSA portal at http://de.portal.airast.org.

Table 8. Manuals and User Guides

Resource	Description
Test Information Distribution Engine (TIDE) User Guide	TIDE is the system used to manage student information and user accounts for online testing. The *TIDE User Guide* provides a step-by-step approach to using the enhanced user management system.
Online Reporting System User Guide	The Online Reporting System (ORS) is the system used to view student performance and participation data. The *ORS User Guide* provides information on how to use the ORS to create reports.
AIRWays Reporting User Guide	This *AIRWays Reporting User Guide* provides information on how authorized users may use AIRWays Reporting to view a variety of student performance reports for the Smarter Balanced interim assessments.
Test Administrator (TA) User Guide	The *TA User Guide* supports individuals using TDS applications to manage testing for students participating in the summative assessment. This resource provides information about the TDS, the TA Interface, and the Student Interface.
Accessibility Guidelines for Delaware System of Student Assessments (DeSSA)	This document provides information about identifying and documenting students who are eligible to receive designated supports and accommodations on Smarter Balanced and other DeSSA assessments. The document also provides information on determining which assessments are appropriate for students and lists the designated supports and accommodations permitted on each assessment and in each content area. Finally, it explains the procedures for documenting supports and accommodations, including the necessary forms and deadlines.
Smarter ELA/Literacy and Mathematics Online Summative Test Administration Manual	This TAM provides the necessary information regarding policies and procedures for the Smarter Balanced ELA/lit and mathematics online summative assessments.
Smarter Summative ELA/Literacy Assessment Paper-Pencil Test Administration Manual	This TAM provides an overview of the Smarter Balanced summative ELA/lit assessment paper-pencil test administration and supplements the online summative TAM.
Smarter Summative Mathematics Assessment Paper-Pencil Test Administration Manual	This TAM provides an overview of the Smarter Balanced summative mathematics assessment paper-pencil test administration and supplements the online summative TAM.
Smarter ELA/Literacy and Mathematics Interim Comprehensive Assessment and Interim Assessment Blocks Test Administration Manual	This TAM provides the necessary information regarding policies and procedures for the Smarter Balanced ELA/lit and mathematics interim

Resource	Description
	comprehensive assessment and interim assessment blocks.
Technology Specifications Manual for Online Testing	This manual provides technology staff with the technical specifications for online testing, including information on Internet and network requirements, general hardware and software requirements, secure browser installation, and supporting the text-to-speech accommodation.
DeSSA Test Security Manual	The *DeSSA Test Security Manual* provides information regarding test security policies for all DeSSA tests. School personnel, including TAs, should review this document carefully.
Secure Browser Installation Manual	This manual provides instructions for installing the secure browser on supported operating systems and is organized by operating system. This document is a supplement to the *Technical Specifications Manual for Online Testing*.
Smarter Braille Requirements and Testing Manual	The *Smarter Braille Requirements and Testing Manual* provides information about supported hardware and software requirements and how to configure JAWS. Information about administering a test to a student requiring braille and navigating a test with JAWS is also included.

2.3.3 Training Modules

The following training modules were created to help users in the field understand the overall Smarter Balanced assessments and how each system works. All modules are provided as PowerPoint presentations; two modules include narration. Table 9 lists the training modules.

Table 9. Smarter Balanced-Developed Training Modules

Module Name	Primary Audience	Objective
Let's Talk Universal Tools	• Students • TAs • Teachers	This presentation provides an overview of the embedded universal tools available to students when using the TDS for the online Smarter Balanced assessment.
Student Interface for Online Testing	• Students • DTCs and STCs • TAs • Teachers	This presentation provides information on how students log in and navigate the TDS, including information on layout and functionality of the test tools.
What Is a CAT (Computer-Adaptive Test)?	• DTCs and STCs • Teachers	This presentation, produced by Smarter Balanced, introduces TAs and students to the concept of a CAT.

2.4 TEST SECURITY

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

2.4.1 DeSSA Test Security Manual

Test security is critically important to protecting intellectual properties, reducing test fraud and theft, and maintaining the integrity of the state assessments. Test integrity is paramount, as it ensures the validity and reliability of test scores and ensures fairness in testing for all Delaware students. The *Test Security Manual* provided online at the DeSSA portal (http://de.portal.airast.org) sets forth test security policies, procedures, and responsibilities for DeSSA assessments. This manual is intended to be used for training those who administer the state assessments.

In preparation for the 2018–2019 school year, each district, school, and charter school adopted and enforced a plan to set procedures for test security and submitted its Test Security Plan to the state by October 16, 2018. All unethical or inappropriate practices and behaviors in the process of test preparation, test administration, and scoring must be reported in writing. Additionally, all personnel associated with assessment administration must read and sign the Test Security and Non-Disclosure Agreement as documentation.

The *Test Security Manual* provides examples for appropriate practices in assessment administration. Any test security violations—such as missing test materials, unauthorized access to test materials, test misadministration, and any other deviations from acceptable security requirements—must be documented and reported to the Office of Assessment at the Delaware Department of Education.

Title 14 (Education, Subchapter IV, State Assessment Security and Violations, of the Delaware Code) outlines the rules and regulations that ensure the security of assessment administration and collection, as well as the reporting of assessment data. Title 14, Subchapter IV, is located in its entirety in Appendix A of the *Test Security Manual*.

The *Test Security Manual* defines security incidents during testing in three levels: impropriety, irregularity, and breach. *Impropriety* refers to an unusual circumstance that has a low impact on an individual or a group of students, with a low risk of potentially affecting student performance on the test; an impropriety can be corrected and contained at the local level. *Irregularity* refers to an unusual circumstance that may potentially affect student performance on the test; an irregularity can be corrected and contained at the local level but must be submitted in the online appeal system for resolution. *Breach* refers to an event that poses a threat to the validity of the assessment (e.g., exposure of secure test materials). A breach has external implications and may result in a decision to remove certain test items from field operation.

The manual specifically indicates test security in the administration of the Smarter Balanced assessments in ELA/lit and mathematics. For example, scratch paper and any materials developed during the classroom activities must be securely disposed of before the administration of a PT. Unless needed as a print-on-demand or braille accommodation, no copies may be made of any test items, stimuli, reading passages, PT materials, writing prompts, or any secure test materials. The electronic policy clearly prohibits the use of cell phones and other electronic devices in the testing area.

2.4.2 Student-Level Testing Confidentiality

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

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

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

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

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

Test materials and score reports should not be exposed to identify student names with test scores, and these should be accessed by only authorized individuals with an appropriate need-to-know status.

All students, including homeschooled students, must be enrolled or registered at their testing schools to take the online, paper-pencil, or braille assessments. Student enrollment information, including demographic data, is generated using a DDOE file and uploaded nightly via a secure file transfer site to the TDS during the testing window.

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

After a test session, only staff with the administrative roles of DTC, STC, or teacher can view their students' scores. TAs do not have access to student scores.

2.4.3 System Security

The objective of system security is to ensure that all data are protected and accessed appropriately by the right user groups. It is about protecting data and maintaining data and system integrity as intended, including ensuring that all personal information is secured, that transferred data (whether sent or received) are not altered in any way, that the data source is known, and that any service can be performed only by a specific, designated user.

A Hierarchy of Control. As described in Section 2.2.1, Administrative Roles, DTCs, STCs, and TAs have well-defined roles and levels of access to the TDS.

Password Protection. All access points by different roles—at the state level, district level, school principal level, and school staff level—require a password to log in to the system. Newly added STCs, TAs, and teachers require access to all DeSSA applications via the DeSSA Single Sign-On System.

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

2.4.4 Security of the Testing Environment

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

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

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

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

Room Preparation. The room should be prepared before the start of the test session. Any information displayed on bulletin boards, chalkboards, or charts that students might use to help answer test questions should be removed or covered. This rule applies to rubrics, vocabulary charts, student work, posters, graphs, content area strategies charts, etc. The cell phones of both testing personnel and students must be turned off and stored in the testing room out of sight. It is recommended that students' cell phones be left in their lockers during the testing sessions. If a student enters the testing room with a cell phone, the TA must

collect it and return it to the student only once testing is completed. TAs are encouraged to minimize access to the testing rooms by posting signs in halls and entrances to promote optimum testing conditions; they should also post "TESTING—DO NOT DISTURB" signs on the doors of testing rooms.

Seating Arrangements. TAs should provide adequate spacing between students' seats. Students should be seated so that they will not be tempted to look at the answers of others. Because the online CAT is adaptive, it is unlikely that students will see the same test questions as other students; however, through appropriate seating arrangements, students should be discouraged from communicating. For the PTs, different forms are distributed throughout a classroom so that students receive different PTs.

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

For the paper-pencil versions, specific instructions on how to package and secure the test booklets to be returned to the testing contractor's office are provided in the *Paper-Pencil Test Administration Manual,* located on the portal at http://de.portal.airast.org.

2.4.5 Test Security Violations

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

> **Impropriety.** This is a test security incident that has a low impact on the individual or group of students who are testing and has a low risk of potentially affecting student performance on the test, test security, or test validity (for example, students leaving the testing room without authorization).
>
> **Irregularity.** This is a test security incident that impacts an individual or group of students who are testing and may potentially affect student performance on the test, test security, or test validity. These circumstances can be contained at the local level (for example, disruption during the test session, such as a fire drill).
>
> **Breach.** This is a test security incident that poses a threat to the validity of the test. Breaches require immediate attention and escalation to the state agency. Examples may include such situations as exposure of secure materials or a repeatable security/system risk. These circumstances have external implications (for example, administrators modifying student answers or students sharing test items through social media).

District and school personnel must document all test security incidents. DTCs are responsible for reporting test security incidents to the state via the KACE/DOE help desk within 24 hours. Throughout testing, test security incidents are reported in accordance with the guidelines in the *DeSSA Test Security Manual* at the DeSSA portal at http://de.portal.airast.org.

2.4.6 Monitoring Test Administration

The observation of the 2017–2018 test administration of the Smarter Balanced assessments was intended to improve test administration and monitoring for the 2018–2019 test administration. The Office of

Assessment at the Delaware Department of Education scheduled on-site visits (upon agreement with schools) during the testing window, and all observers followed the procedure for the on-site visits without interfering with test activities.

The Observation and Discussion Form provides each observer with a general checklist for the appropriate test practices and standardized test conditions. The observation includes six elements: (1) computer sign-on and start-up process, (2) security, (3) test environment and administration procedures, (4) test atmosphere, (5) calculator use in mathematics, and (6) accommodations.

2.5 STUDENT PARTICIPATION

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

2.5.1 Homeschooled Students

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

2.5.2 Student Exemptions

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

- Students with significant cognitive disabilities who meet the criteria for the ELA/lit alternate assessment based on alternate achievement standards (approximately 1% or less of the student population)
- Students with significant cognitive disabilities who meet the criteria for the mathematics alternate assessment based on alternate achievement standards (approximately 1% or less of the student population)
- English language learners (ELLs) who enrolled in a U.S. school within the 12 months before the beginning of the testing window have a one-time exemption. These students may instead participate in their state's English language proficiency assessment consistent with state and federal policy. Students who are participating in the Interim Comprehensive Assessments or Interim Assessment Blocks may also have an exemption from completing the ELA/lit assessment.

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

2.6 ONLINE TESTING FEATURES AND ACCOMMODATIONS

The Smarter Balanced Assessment Consortium's *Usability, Accessibility, and Accommodations Guidelines* are intended for school-level personnel and decision-making teams, including individual education program (IEP) and Section 504 Plan teams, as they prepare for and implement the Smarter Balanced assessments. The *Guidelines* provide information for classroom teachers, English language development educators, special education teachers, and instructional assistants to use in selecting and administering universal tools, designated supports, and accommodations for those students who need

them. The *Guidelines* are also intended for assessment staff and administrators who oversee the decisions that are made in instruction and assessment.

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

Following the Smarter Balanced guidelines, the Accessibility Guidelines for Delaware System of Student Assessments on the DeSSA portal at http://de.portal.airast.org contain the Delaware policies governing the provision and documentation of test supports and available accommodations for students participating in the DeSSA Smarter Balanced assessments. The Delaware Guidelines clearly describe the process for the inclusion of Students with Disabilities (SWDs) and ELLs, the process for identifying those who need accommodations, and the selection and provision of the appropriate accommodations and related supports. This document also provides test users with the state policy for "General Education Students Receiving Supports" who are eligible to receive supports (e.g., text-to-speech on items), not accommodations, on the Smarter Balanced ELA/lit and mathematics assessments. The two types of accessibility features are classified as embedded features provided directly through the online test environment (e.g., text-to-speech, Spanish-English stacked) and non-embedded features that must be provided by the school (e.g., translator, enhanced lighting).

The administration of Smarter Balanced assessments is classified into four general categories in Delaware: (1) testing without accommodations and supports, (2) testing without accommodations but with supports, (3) testing with accommodations but without supports, and (4) testing with accommodations and supports.

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

State-level users, DAs, and DAMs can set embedded and non-embedded designated supports and accommodations based on their specific user role. Designated supports and accommodations must be set in TIDE before starting a test session.

All the embedded and non-embedded universal tools will be activated for use by all students during a test session. One or more of the preselected universal tools can be deactivated by a TA in the TA Interface of the TDS for a student who may be distracted by the ability to access a specific tool during a test session.

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

2.6.1 Online Universal Tools for All Students

Universal tools are access features of an assessment or exam that are digitally delivered (i.e., embedded) or separately delivered (i.e., non-embedded) components of the TDS. Universal tools are available to all students based on their preference and selection and have been preset in TIDE. In the 2018–2019 test administration, the following features (universal tools) were available for all students to access. For specific

information on how to access and use these features, refer to the *Test Administrator User Guide* on the DeSSA portal at http://de.portal.airast.org.

Embedded Universal Tools

Breaks. The number of items per session can be flexibly defined based on the student's need. Breaks of more than 20 minutes will prevent the student from returning to items that have been already attempted. (An exception is the PT.) There is no limit on the number of breaks that a student may be given. The use of this universal tool may result in the student needing additional overall time to complete the assessment. See pause rules in the *Smarter ELA/Literacy and Mathematics Online Summative Test Administration Manual* for details about the length of time a student may pause and still be able to review items previously answered.

Calculator. An embedded, on-screen, digital calculator can be accessed for calculator-allowed items when students click the calculator button. This tool is available only with the specific items for which the Smarter Balanced item specifications indicate that it would be appropriate. When the embedded calculator, as presented for all students, is not appropriate for a student (e.g., a student who is blind), the student may use the calculator offered with assistive technology devices, such as a talking calculator or a braille calculator (for calculator-allowed items only).

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

English Dictionary. An English dictionary may be available for the full-write portion of an ELA/lit PT. The use of this universal tool may result in the student needing additional overall time to complete the assessment.

English Glossary. Grade- and context-appropriate definitions of specific construct-irrelevant terms are shown in English on the screen via a pop-up window. The student can access the embedded glossary by clicking any of the pre-selected terms. The use of this accommodation may result in the student needing additional overall time to complete the assessment.

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

Global Notes. Global notes is a notepad available for ELA/lit PTs in which students complete the full-write portion of an ELA/lit PT. The student clicks the notepad icon for the notepad to appear. During the ELA/lit PTs, the notes are retained from segment to segment so that the student may go back to the notes even though he or she cannot go back to specific items in the previous segment.

Highlighter. This is a digital tool for marking desired text, item questions, item answers, or parts of these with a color. Highlighted text remains available throughout each test segment.

Keyboard Navigation. Navigation throughout a test can be accomplished by using a keyboard.

Mark for Review. This tool allows students to flag items for future review during the assessment. Markings are not saved when the student moves on to the next segment or after a break of more than 20 minutes.

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

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

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

Writing Tools. Selected writing tools (e.g., bold, italic, bullets, undo/redo) are available for all student-generated responses. (Also see *spell check*.)

Zoom. This is a tool for making text or other graphics in a window or frame appear larger on the screen. The default font size for most tests is 12 points, and the default size for grades 3 and 4 is 14 points. The student can enlarge text and graphics by clicking the Zoom In button. The student can click the Zoom Out button to return to the default or a smaller print size. When using the zoom feature, the student only changes the size of text and graphics on the current screen for the displayed item. To increase the default print size of the entire test, the print size must be set for the student in TIDE or set by the TA before the start of the test. The use of this universal tool may result in the student needing additional overall time to complete the assessment.

Non-Embedded Universal Tools

Assistive Listening Device. Students may use amplification assistive technology (e.g., headphones, FM system, noise buffers, white noise machines) to increase the volume provided in the assessment platform for the ELA/lit and mathematics PTs. Use of this resource likely requires a separate setting. If the device has additional features that may compromise the validity of the test (e.g., Internet access), the additional functionality must be deactivated to maintain test security.

Breaks. All students may take breaks, including *frequent breaks*, as needed. The term *frequent breaks* refers to multiple, planned, short breaks during testing based on a specific student's needs (for example, the student becomes fatigued easily). During each break, the testing clock is stopped.

English Dictionary. An English dictionary can be provided for the full-write portion of an ELA/lit PT. The use of this universal tool may result in the student needing additional overall time to complete the assessment.

Familiar TA. The student knows the TA and/or interpreter.

Refocus. The student's attention can be refocused on the test with use of intermittent verbal, picture symbol, signed, cued speech, or physical prompts. Refocus should not in any way cue a student to return to a previous item or indicate that the student may have made an error. This would be considered a test security violation. The use of this universal tool may result in the student needing additional overall time to complete the assessment.

Scratch/Blank/Grid Paper. Scratch paper to make notes, write computations, or record responses may be made available. Only plain paper or lined paper is appropriate for ELA/lit. Graph paper is required

beginning in grade 6 and can be used on all mathematics assessments. A stu'dent can use an assistive technology device for scratch paper as long as the device is consistent with the child's IEP and acceptable to the DDOE.

> **CAT.** All scratch paper must be collected and securely destroyed at the end of each CAT assessment session to maintain test security. All notes on whiteboards or assistive technology devices must be erased at the end of each CAT session.
>
> **Performance Tasks.** For mathematics and ELA/lit PTs, if a student needs to take the PT in more than one session, scratch paper, whiteboards, and/or assistive technology devices must be collected at the end of each session, securely stored, and made available to the student at the next PT testing session. Once the student completes the PT, scratch paper must be collected and securely destroyed, and whiteboards and notes on assistive technology devices should be erased to maintain test security.

Small Group. A small group is a subset of a larger testing group assessed in a separate location. There is no specific number defined for a small group, but a group of two to eight students is typical. Separately testing a single student is also permissible. Small groups may be appropriate for a human read-aloud, translated test administration, or WhisperPhone®, or to reduce distractors for some students. If a small group is selected for a non-embedded universal tool, it is not necessary to also select a separate setting as a non-embedded designated support.

Thesaurus. A thesaurus provides synonyms of terms while a student interacts with text included in the assessment and may be available for the full-write portion of an ELA/lit PT. The use of this universal tool may cause the student to need additional overall time to complete the assessment.

Time of Day. A student should be tested during the time of day that is best for the student (e.g., only in the morning).

Additional non-embedded universal tool options include modified lighting, specialized equipment or furniture, *Whiteboards/Assistive Devices,* and specified area or seating.

2.6.2 Designated Supports and Accommodations

Designated supports for the Smarter Balanced assessments are those features that are available for use by any student for whom the need has been indicated by an educator (or team of educators with parent/guardian and student). Scores achieved by students using designated supports will be included for federal accountability purposes. It is recommended that a consistent process be used to determine these supports for individual students. All educators making these decisions should be trained on the process and should understand the range of designated supports available. Smarter Balanced members have identified digitally embedded and non-embedded designated supports for students for whom an adult or team has indicated a need for the support.

Accommodations are modifications in testing conditions and/or presentation of the test to facilitate access for students with special needs in order to demonstrate what they know and can do. Accommodations must be familiar to the student and used in the classroom to support instruction. Consortium-approved accommodations do not compromise the learning expectations, construct, grade-level standard, or intended outcome of the assessments.

Below are brief descriptions of embedded and non-embedded designated supports and accommodations.

Embedded Designated Supports

Color Choices/Contrast. These enable students to adjust computer screen background or text font color based on student needs or preferences. This may include reversing the colors for the entire interface or choosing the color of font and background. Black on white, reverse contrast, black on rose, medium gray on light gray, and yellow on blue were offered for the online assessments. The TA must set this feature in the TA Interface.

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 masking.

Glossaries. The glossaries support allows students to view a glossary for certain words in the test content.

Languages. Languages set the language presentation for the test content.

Mouse Pointer. mouse pointer is an embedded support that allows the mouse pointer to be set to a larger size or to a different color during registration. These settings cannot be changed during test administration. A TA sets the size and color of the mouse pointer before testing.

Permissive Mode. Permissive mode must be selected if accommodations requiring additional software are to be used (e.g., speech-to-text software, ZoomText [magnification] software, or other software to support alternate response accommodations).

Streamlined Mode. Streamlined mode is an alternate, more linear display of item and stimuli. It is needed for the language feature for braille or Spanish and with a zoom level of 5 and above.

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

Zoom. Zoom is a tool for making text or other graphics in a window or frame appear larger on the screen. To increase the default print size of the entire test (from 1X up to 20X), the print size must be set for the student in TIDE or set by the TA before the start of the test. Zoom levels of 5X or greater must be used with streamlined mode.

Non-Embedded Designated Supports

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/lit PT.

Color Contrast (printed). Test content of online items may be printed (using print on request) with different colors.

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

Disable Universal Tools. Any universal accessibility tools that might be distracting or that students do not need to use, or are unable to use, can be disabled. The TA must turn off tools one by one at the time of test administration. Tools that can be switched off include highlighting, strikethrough, expandable passages, mark for review, and global notes.

ELL First Year Exemption. The ELL first year exemption is an exemption from the ELA/lit tests. Students are eligible if, as of the final date of the testing window, they have been enrolled in U.S. schools for less than one year.

Human Read-Aloud Items/Stimuli (for ELA/lit PT passages). Text is read aloud to the student by a trained and qualified human reader who follows the administration guidelines provided in the *Smarter Online Summative Test Administration Manual*. All or portions of the content may be read aloud.

Human Read-Aloud Items (for mathematics items and ELA/lit PT items, but not for reading passages). Text is read aloud to the student by a trained and qualified human reader who follows the administration guidelines provided in the *Smarter ELA/Literacy and Mathematics Online Summative Test Administration Manual*. All or portions of the content may be read aloud. In each grade, 0.5—4% of students used this designated support.**Human Reader in Spanish (for mathematics tests).** Spanish text is read aloud to the student by a trained and qualified human reader who follows the administration guidelines provided in the *Smarter Balanced Test Administration Manual*. All or portions of the content may be read aloud.

Interpreter—Native Language. The test taker is provided with a native language translator to translate test questions (including multiple-choice options) into his or her native language. The instructor may determine that the translator must translate all items or only items requested by the student. The native language translator must be proficient in the native language. DDOE must approve this support.

Interpret/Translate Orally—Directions Only. The test taker is provided with a native language/visual communication translator to translate directions into his or her native language only. The native language translator/TA must be proficient in the native language.

Magnification. The student may adjust the size of specific areas of the screen (e.g., text, formulas, tables, graphics, and navigation buttons) with an assistive technology device. Magnification allows increasing the size to a level not allowed by the universal zoom tool, color contrast designated support, and mouse pointer designated support.

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

Noise Buffer. These include ear mufflers, white noise machines, and other equipment to reduce external sounds.

Paper-Pencil Test. The test is presented in a fixed-form, paper-pencil format. This support is to be used only when print-on-demand is not practical due to the student's testing location or access needs. This support includes the use of a handheld calculator in the case of mathematics.

Scribe—All Items Except Writing Items on ELA/Lit PTs (for ELA/lit non-writing items and mathematics items). For this type of scribe, students may not have a scribe during writing items. Students dictate their responses to a scribe who records verbatim what they dictate. The scribe must be trained and qualified and must follow the administration guidelines provided in the *Smarter ELA/Literacy and Mathematics Online Summative Test Administration Manual*.

Separate Setting in School. The test location is altered so that the student is tested in an in-school setting different from that made available for most students.

Separate Setting not in School. The test location is altered so that the student is tested in a non-school setting different from that made available for most students.

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

Translated Test Directions in Print. This is a PDF file of directions translated into each of the languages currently supported (except Spanish, as it is already an embedded support). This is available for the following languages and dialects: Arabic, Cantonese, Ilokono, Korean, Mandarin, Punjabi, Russian, Tagalog, Ukrainian, and Vietnamese. A bilingual adult can read this file to the student.

Translations (glossaries) for Mathematics Paper-Pencil Tests. Translated glossaries are a language support provided for selected construct-irrelevant terms for mathematics. Glossary terms are listed by item and include the English term and its translated equivalent.

Unique Accommodation (DDOE approved). This is support or accommodations not listed in these guidelines by Smarter Balanced. This is available by application only.

WhisperPhone®. The WhisperPhone® is a school-provided tool students may use to read the test to themselves.

Embedded Accommodations

American Sign Language video (ASL). This is for ELA/lit listening items and mathematics items. An American Sign Language (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.

Closed Captioning. Printed text appears on the computer screen as audio materials are presented.

Emboss (passages/stimuli and items). It turns on embossing for students testing in braille.

Emboss Request Type. It sets test content to be embossed automatically or only at the student's request.

Print on Request. Paper copies of either passages/stimuli or items are printed for students. A student may request that one or more test questions be printed electronically from the online system to review on paper. All printed test material must be shredded at the end of the test session. (The TA must approve each print request.)

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

Non-Embedded Accommodations

100s Number Table (grades 4 and above mathematics tests). A paper-pencil table listing of numbers 1–100 is available from Smarter Balanced for reference.

Abacus. This tool may be used in place of scratch paper for students who typically use an abacus. Some students with visual impairments who typically use an abacus may use one in place of scratch paper.

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

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

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

Human Read Aloud (for ELA/Lit Passages). Text is read aloud to the student via an external screen reader or by a trained and qualified human reader who follows the administration guidelines provided in the *Smarter ELA/Literacy and Mathematics Online Summative Test Administration Manual*. All or portions of the content may be read aloud. Members can refer to the *Accessibility Guidelines for the Delaware System of Student Assessments* when deciding if this accommodation is appropriate for a student.

Human Interpreter—Visual Communication. An adult with the necessary qualifications provides translation/interpretation of the mathematics test using cued speech or signed English to a student with disabilities. Reading passages may not be translated through visual communication. This support must be approved by the DDOE.

Multiplication Table (grades 4 and above mathematics tests). A paper-pencil, single-digit (1–9) multiplication table will be available from Smarter Balanced for reference.

Physical Assistance from a TA. Students can use physical assistance from a TA, such as direct assistance with turning pages, recording answers for the paper-pencil test (scribing), or navigating in electronic format.

Scribe (for ELA/lit writing items). Students dictate their responses to a TA who records verbatim what they dictate. The scribe must be trained and qualified and must follow the administration guidelines provided in the *Smarter ELA/Literacy and Mathematics Online Summative Test Administration Manual*.

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

Word Prediction. Word prediction 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. Students may use their own assistive technology devices.

Table 10 lists universal tools, designated supports, and accommodations that were offered in the 2018–2019 administration. Tables 11–16 provide the number of students who were offered the accommodations and/or designated supports. In general, most of the designated supports and accommodations were used by less than 2% of students across subjects and grades, with a few exceptions. Among the designated supports, Text-to-Speech was the most frequently used in both subjects, ranging from 12% to 26% of students. Among the accommodations, Print-on-Demand was the most frequently used in ELA/lit (3% –5%) while Multiplication Table was the most frequently used in in mathematics (2% –12%).

Table 10. Universal Tools, Designated Supports, and Accommodations in 2018–2019

	Universal Tools	Designated Supports	Accommodations
Embedded	Breaks Calculator[1] Digital Notepad English Dictionary[2] English Glossary Expandable Passages Global Notes Highlighter Keyboard Navigation Mark for Review Mathematics Tools[3] Spell Check Strikethrough Writing Tools[4] Zoom	Color Contrast (Computer) Glossaries Language Masking Mouse Pointer Permissive Mode Streamlined Mode Text-to-Speech[5] Zoom	American Sign Language[6] Closed Captioning[7] Emboss (passages/stimuli and items) Emboss Request Type Print-on-Request Type of Refreshable Braille
Non-Embedded	Assistive Listening Device Breaks English Dictionary[2] Familiar TA Modified Lighting Refocus Scratch/Blank/Grid Paper Small Group Specialized Equipment or Furniture Specified Area or Seating Thesaurus[2] Time of Day	Bilingual Dictionary[2] Color Contrast (Printed) Color Overlay Disable Universal Tools ELL First Year Exemption Human Read-Aloud Passages for PT[8] Interpreter—Native Language[9] Interpret/Translate Orally—Directions Only Magnification Medical Device Noise Buffers Paper-Pencil Test Read-Aloud Items[10] Scribe[11] Separate Setting in School Separate Setting Not in School/Homebound Simplify Directions in English Translated Test Directions Translations (Glossary)[12] Unique Accommodation[8] WhisperPhone®	100s Number Table[13] Abacus Alternate Response [14] Braille (Paper-Pencil Version) Calculator[1] Human Read-Aloud Passages[15] Interpreter—Visual Communication[8] Multiplication Table[13] Physical Assistance from a TA Scribe for SwD Speech-to-Text Word Prediction

Note: Items shown are available for ELA/lit and mathematics unless otherwise noted.
[1] For calculator-allowed items only in grades 6–8
[2] For ELA/lit performance task full-writes
[3] Includes embedded ruler, embedded protractor
[4] Includes bold, italic, underline, indent, cut, paste, spell check, bullets, undo/redo
[5] For mathematics test
[6] For ELA/lit listening items and mathematics items
[7] For ELA/lit listening items
[8] For ELA/lit performance task passages
[9] Must be approved by DDOE

[10] For ELA/lit items (not ELA/lit reading passages) and mathematics items
[11] For ELA/lit non-writing items and mathematics items
[12] For mathematics items on paper-pencil test
[13] For mathematics items beginning in grade 4
[14] Includes adapted keyboards, large keyboard, StickyKeys, MouseKeys, FilterKeys, adapted mouse, touch screen, head wand, and switches
[15] For ELA/lit CAT reading passages, all grades—must be approved by DDOE

Table 11. Students with Embedded and Non-Embedded Accommodations in ELA/Lit

Accommodations	Grade					
	3	**4**	**5**	**6**	**7**	**8**
Embedded Accommodations						
American Sign Language	3	6	5	7	4	3
Closed Captioning	9	17	17	14	11	10
Print-on-Request: Items	2		1		3	
Print-on-Request: Passages	43	26	25	9	12	10
Print-on-Request: Passages and Items	334	462	527	435	404	392
Print-on-Request: Stimuli	36	3	13	2	4	2
Non-Embedded Accommodations						
Alternate Response						1
Braille (Paper-Pencil Version)						1
Human Read-Aloud Passages	20	30	14	15	10	5
Physical Assistance from a TA	37	39	38	2	2	2
Scribe for SwD	128	142	109	44	28	20
Speech-to-Text	17	16	15	26	35	34
Word Prediction	7	4	6	22	13	8

Table 12. Students with Embedded Designated Supports in ELA/Lit

Designated Supports	Subgroup	Grade					
		3	**4**	**5**	**6**	**7**	**8**
Color Choices/Contrast	Overall	9	11	20	22	20	15
	ELL			1	1	1	1
	Special Ed	9	10	17	21	16	13
Masking	Overall	120	225	164	226	198	228
	ELL	36	74	49	39	33	28
	Special Ed	82	124	93	211	171	202
Mouse Pointer	Overall	1	1	1	1	1	1
	ELL						
	Special Ed	1	1	1	1	1	1
Permissive Mode	Overall	20	13	17	63	64	52
	ELL		1	1	13	13	8
	Special Ed	19	12	17	57	53	44
Streamlined Mode	Overall	3	11	4	8	6	5
	ELL	1	2	1	3	1	1
	Special Ed	3	9	3	6	5	4
Text-to-Speech: Items	Overall	2,600	2,512	2,534	1,533	1,360	1,207
	ELL	1,121	1,033	887	407	292	255
	Special Ed	1,068	1,159	1,300	1,189	1,088	985
Text-to-Speech: Stimuli & Items	Overall	2,645	2,550	2,554	1,599	1,404	1,263
	ELL	1,136	1,049	894	426	309	278
	Special Ed	1,102	1,195	1,320	1,242	1,124	1,018
Zoom	Overall	42	70	33	19	12	8
	ELL	15	19	14	1	3	
	Special Ed	11	17	12	13	8	5

Table 13. Students with Non-Embedded Designated Supports in ELA/Lit

Designated Supports	Subgroup	Grade					
		3	4	5	6	7	8
Color Contrast	Overall	1		1			3
	ELL						1
	Special Ed	1		1			
Color Overlay	Overall	4	4	8	11	1	1
	ELL	1		2	2		
	Special Ed	4	3	7	5	1	1
Disable Universal Tools	Overall	1	1		1		
	ELL						
	Special Ed		1		1		
ELL First Year Exemption	Overall	2	1	1	1	3	2
	ELL	2	1	1	1	3	2
	Special Ed				1		
Human Read-Aloud Items	Overall	439	468	412	167	140	83
	ELL	112	115	105	25	24	14
	Special Ed	264	306	290	147	113	72
Interpret/Translate Orally—Directions Only	Overall	12	14	11	17	9	10
	ELL	11	10	9	11	8	9
	Special Ed	5	6	7	6	2	1
Magnification	Overall	6	4	12	4	7	
	ELL	2		3	1		
	Special Ed	6	2	2	3	5	
Medical Device	Overall	6	7	6	7	6	8
	ELL	1	1				
	Special Ed		1	1		1	
Noise Buffers	Overall	62	122	107	50	50	34
	ELL	7	20	12	1	1	3
	Special Ed	51	65	80	43	41	28
Paper-Pencil Test	Overall	1	3	2			
	ELL						
	Special Ed	1	3	2			
Scribe Items (Non-Writing)	Overall	14	22	15	5	3	2
	ELL	6	5	5	2	1	
	Special Ed	8	15	13	4	2	1
Separate Setting in School	Overall	451	484	453	286	288	284
	ELL	101	104	89	49	21	24
	Special Ed	331	379	381	254	243	261
Separate Setting not in School/Homebound	Overall	1		1	1	2	1
	ELL						
	Special Ed	1		1	1	2	1
Simplified Test Directions	Overall	427	401	288	292	274	239
	ELL	316	246	171	141	147	99
	Special Ed	127	165	149	199	181	173
Translated Test Directions	Overall	2	3	1	7	3	
	ELL	2	3		7	3	
	Special Ed	2		1	2	1	

Designated Supports	Subgroup	Grade					
		3	4	5	6	7	8
Unique Accommodation	Overall	13	14	8	7	4	8
	ELL	1	2				
	Special Ed	5	7	2			1
WhisperPhone®	Overall	222	167	96	23	7	
	ELL	82	42	42	2		
	Special Ed	107	89	61	20	7	

Table 14. Students with Embedded and Non-Embedded Accommodations in Mathematics

Accommodations	Grade					
	3	4	5	6	7	8
Embedded Accommodations						
American Sign Language	3	6	5	7	4	4
Emboss: Stimuli and Items	1					
Emboss Request Type: Auto	1					
Print-on-Request: Stimuli and Items	359	454	521	439	409	397
Non-Embedded Accommodations						
100s Number Table	442	707	591	317	196	185
Abacus	2	3	2	1		
Alternate Response	1		2			1
Braille (Paper-Pencil Version)						1
Calculator	13	57	70	243	259	275
Interpreter—Visual Communication			1			
Multiplication Table	200	1,010	1,297	1,227	987	865
Physical Assistance from a TA	37	32	37	3	3	2
Scribe for SwD	108	132	100	39	21	18
Speech-to-Text	18	11	13	23	31	30
Word Prediction	3	2	4	15	16	6

Table 15. Students with Embedded Designated Supports in Mathematics

Designated Supports	Subgroup	Grade					
		3	**4**	**5**	**6**	**7**	**8**
Color Choices/Contrast	Overall	18	20	15	20	15	14
	ELL			1	1	1	2
	Special Ed	18	19	12	19	10	12
Glossaries: Spanish	Overall	120	170	105	124	109	90
	ELL	119	166	105	124	109	89
	Special Ed	11	6	11	18	19	21
Glossaries: Other Languages	Overall	4	14	7	9	10	8
	ELL	4	14	7	9	10	8
	Special Ed						
Language: Braille	Overall	1					
	ELL						
	Special Ed	1					
Language: Spanish	Overall	81	92	70	83	82	83
	ELL	78	90	69	83	82	82
	Special Ed	3	4	4	6	5	8
Masking	Overall	121	220	161	225	198	227
	ELL	36	76	50	39	32	28
	Special Ed	81	121	91	207	173	200
Mouse Pointer	Overall	1	1	1	1	1	
	ELL						
	Special Ed	1	1	1	1	1	
Permissive Mode	Overall	17	14	15	67	63	58
	ELL		1	1	15	12	11
	Special Ed	16	13	15	59	52	48
Streamlined Mode	Overall	15	19	4	27	26	20
	ELL	12	11	1	23	20	17
	Special Ed	4	8	3	5	6	3
Text-to-Speech: Stimuli & Items	Overall	2,678	2,570	2,590	1,591	1,422	1,284
	ELL	1,163	1,063	903	423	338	299
	Special Ed	1,103	1,187	1,341	1,246	1,115	1,020
Zoom	Overall	43	70	31	19	13	8
	ELL	15	20	14	1	3	
	Special Ed	10	17	12	13	8	5

Table 16. Students with Non-Embedded Designated Supports in Mathematics

Designated Supports	Subgroup	Grade					
		3	4	5	6	7	8
Color Contrast	Overall			1			2
	ELL						
	Special Ed			1			
Color Overlay	Overall	4	4	8	8	1	1
	ELL	1		2	1		
	Special Ed	4	3	7	2	1	1
Disable Universal Tools	Overall				2		
	ELL						
	Special Ed				1		
Human Read-Aloud Stimuli and Items	Overall	429	459	401	159	89	43
	ELL	119	134	113	28	16	14
	Special Ed	245	276	275	131	71	28
Human Read Aloud in Spanish	Overall	29	24	4	14	7	5
	ELL	26	23	4	14	5	3
	Special Ed	2		1	3	2	1
Interpreter—Native Language	Overall	14	20	9	7	6	9
	ELL	14	20	9	7	6	8
	Special Ed	1			1		
Interpret/Translate Orally—Directions Only	Overall	14	15	14	28	21	20
	ELL	14	13	13	23	21	19
	Special Ed	4	4	6	5	2	1
Magnification	Overall	6	4	11	5	5	1
	ELL	1		2	1		
	Special Ed	5	2	2	4	3	1
Medical Device	Overall	7	7	7	8	6	8
	ELL	1	1				
	Special Ed		1	1	1	1	
Noise Buffers	Overall	61	120	106	52	48	33
	ELL	7	22	12	2		3
	Special Ed	50	62	79	45	40	27
Paper-Pencil Test	Overall	3	3	2		1	
	ELL	1					
	Special Ed	1	3	2		1	
Scribe	Overall	16	27	14	5	2	2
	ELL	6	8	4	2		
	Special Ed	8	14	13	3	1	1
Separate Setting in School	Overall	446	480	450	291	278	289
	ELL	106	112	95	50	20	25
	Special Ed	324	365	379	250	243	265
Separate Setting Not in School/Homebound	Overall	1			1	1	
	ELL				1		
	Special Ed	1			1	1	
Simplified Test Directions	Overall	442	411	295	288	291	235
	ELL	332	261	179	146	162	102
	Special Ed	121	159	150	191	183	164

Designated Supports	Subgroup	Grade					
		3	4	5	6	7	8
Translated Test Directions	Overall	13	11	2	12	10	3
	ELL	13	10	2	12	10	3
	Special Ed	1	3		2	1	
Translations (Glossaries): Paper-Pencil	Overall	2	6	3	5		
	ELL	2	6	2	4		
	Special Ed		1	1	2		
Unique Accommodation	Overall	172	10	7	8	4	8
	ELL	27	1				
	Special Ed	156	2	1			1
WhisperPhone	Overall	206	165	69	16	9	
	ELL	72	44	26		1	
	Special Ed	99	83	58	13	8	

2.7 DATA FORENSICS PROGRAM

2.7.1 Data Forensics Report

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

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

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

2.7.2 Changes in Student Performance

Score changes between years are examined using a regression model with the current-year score regressed on the test score from the previous year using the number of days between test-end days in two years to control the effect of instruction time. Between-year comparisons are reported between the 2018–2019 and 2017–2018 school years.

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

The residuals for individual students are also aggregated for a testing session, TA, and school. The system flags any unusual changes in an aggregate performance between administrations or years based on the average residuals in an aggregate unit (e.g., testing session, TA, and school). For each aggregate unit, a critical *t* value is computed and flagged when $|t|$ 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* = standard deviation of residuals in an aggregate unit; *n* = number of students in an aggregate unit (e.g., testing session, TA, or school), σ^2 is the MSE from the regression, and $\widehat{e_i}$ 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),

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

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

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

2.7.3 Item Response Time

In the online environment, item response time is captured as the item page time (the time that a student spends on each item page) in milliseconds. For discrete items, each item appears on the screen one item at a time, whereas stimulus-based items appear on the screen together. The page time is the time spent on one item for discrete items and the time spent on all items associated with a stimulus for stimulus-based items. For each student, the total time taken to complete the test is computed by adding up the page time for all items and item groups (stimulus-based items).

The expectation is that the total test-taking time will be shorter than the average time if students may have a prior knowledge of items. 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 is far less than that required of students statewide. If students already know the answers to the questions, the test-taking time will be much shorter than the test-taking time for those who 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 mean and the standard deviation of test-taking time are computed across all students. Individual students and relative aggregated units were flagged if the test-taking time was greater than |3| standard deviations of the state average. The state average and standard deviation was computed based on all students when the analysis was performed. The QA report includes a list of the flagged aggregate units.

2.7.4 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 inppropriate prior knowledge of some test items (or is provided with answer keys during theadministration), he or she will respond correctly to those items at a higher probability than his or her estimated ability based on all items. In this case, the person-fit index will be unexpectedly larger for the student.

The person-fit index is based on all item responses in a test. An unlikely response(s) to a single test question or entire test questions may not result in a flagged person-fit index. It should be noted that not all unlikely response patterns indicate cheating, as in the case of the probility of guessing for selected-response items. . Therefore, the evidence of person-fit index should be evaluated along with other violations of test security to determine possible testing irregularities. The number of flagged students is summarized for every testing session, TA, and school.

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

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

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

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

2.8 PREVENTION AND RECOVERY OF DISRUPTIONS IN TEST DELIVERY SYSTEM

AIR is continuously improving our ability to protect our systems from interruptions. AIR's TDS is designed to ensure that student responses are captured accurately and stored on more than one server in case of a failure. Our architecture, described 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 is transferred to a different data center each night.

AIR has developed a unique monitoring system that is very sensitive to changes in server performance. Most monitoring systems provide warnings when something is going wrong. Ours does, too, but it also provides warnings when any given server is performing differently from its performance over the 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 us to detect potential problems, investigate them, and

mitigate them before a failure. On multiple occasions, this has enabled us to make adjustments and replace equipment before any problems occurred.

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

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

2.8.1 High-Level System Architecture

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

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

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

Student Machine

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

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

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

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

Test Delivery Satellites

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

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

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

Hub

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

Demographic and History Servers

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

Quality Assurance System

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

Database of Record

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

2.8.2 Automated Backup and Recovery

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

2.8.3 Other Disruption Prevention and Recovery

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

- The system's hosting provider has redundant power generators that can continue to operate for up to 60 hours without refueling. With the multiple refueling contracts that are in place, these generators can operate indefinitely.
- The hosting provider has multiple redundancies in the flow of information to and from the system's data centers by partnering with nine different network providers. Each fiber carrier must enter the data center at separate physical points, protecting the data center from a complete service failure caused by an unlikely network cable cut.
- On the network level are redundant firewalls and load balancers throughout the environment.
- The system uses redundant power and switching within all server cabinets.
- Data are protected by nightly backups. A full weekly backup and incremental nightly backups protect data. Should a catastrophic event occur, AIR is able to reconstruct real-time data using the data retained on the TDS satellites and hubs.
- The server backup agents send alerts to notify system administration staff in the event of a backup error, at which time they will inspect the error to determine whether the backup was successful or if they 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 2018–2019 OPERATIONAL TEST ADMINISTRATION

3.1 STUDENT POPULATION

All students enrolled in grades 3–8 in all public elementary and secondary schools are required to participate in the Smarter Balanced ELA/lit and mathematics assessments. Tables 17 and 18 present the demographic composition of Delaware students who meet attemptedness requirements for scoring and reporting of the Smarter Balanced assessments.

Table 17. Number of Students in Summative ELA/Lit Assessment

Group	G3	G4	G5	G6	G7	G8
All Students	10,234	10,468	10,827	10,572	10,540	10,207
Female	4,995	5,148	5,282	5,281	5,299	5,085
Male	5,239	5,320	5,545	5,291	5,241	5,122
African American	3,107	3,193	3,309	3,249	3,169	3,198
AmerIndian/Alaskan	23	40	28	45	37	51
Asian	420	417	392	375	376	384
Hispanic	1,924	1,996	2,021	1,863	1,880	1,784
Pacific Islander	9	18	12	12	11	11
White	4,254	4,312	4,548	4,573	4,629	4,389
Multi-Racial	497	492	517	455	438	390
ELL	1,750	1,651	1,264	752	534	451
Special Education	1,555	1,707	1,802	1,697	1,623	1,525
CD 504	398	448	555	547	570	559
Title I	1,002	1,059	1,050	1,019	1,191	1,274

Note. AmerIndian/Alaskan= American Indian/Alaskan Native; Pacific Islander =Native Hawaiian/Pacific Islander

Table 18. Number of Students in Summative Mathematics Assessment

Group	G3	G4	G5	G6	G7	G8
All Students	10,287	10,522	10,852	10,607	10,572	10,232
Female	5,011	5,172	5,295	5,291	5,308	5,102
Male	5,276	5,350	5,557	5,316	5,264	5,130
African American	3,109	3,196	3,312	3,243	3,160	3,197
AmerIndian/Alaskan	24	40	28	45	38	51
Asian	424	432	398	379	378	388
Hispanic	1,974	2,035	2,044	1,900	1,923	1,807
Pacific Islander	9	18	12	12	11	11
White	4,253	4,313	4,542	4,574	4,624	4,389
Multi-Racial	494	488	516	454	438	389
ELL	1,814	1,722	1,308	811	603	493
Special Education	1,549	1,700	1,801	1,698	1,618	1,522
CD 504	398	446	555	547	565	559
Title I	1,007	1,066	1,051	1,018	1,187	1,272

3.2 SUMMARY OF OVERALL STUDENT PERFORMANCE

Tables 19–22 summarize the 2018–2019 summative test results for all students and by subgroup, including the mean 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 show the percentage of proficient students in five years with cohort comparisons. Figures 3 and 4 show the average scale scores in five years for all students by grade and test. The mean and the standard deviation of scale scores, as well as the percentage of proficient students for each test administration by subgroup, are provided in Appendix B.

Table 19. ELA/Lit Percentage of Students in Achievement Levels for Overall and by Subgroup (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	10,234	2429.72	89.67	26	24	23	27	50
Female	4,995	2439.13	87.35	21	25	24	30	54
Male	5,239	2420.75	90.93	30	23	23	24	47
African American	3,107	2396.75	84.95	38	26	20	15	35
AmerIndian/Alaskan	23	2425.39	73.14	22	30	26	22	48
Asian	420	2485.99	86.52	10	15	22	53	75
Hispanic	1,924	2403.77	82.31	34	29	22	15	37
Pacific Islander	9*							
White	4,254	2458.92	84.07	15	20	26	39	65
Multi-Racial	497	2439.04	86.95	20	25	27	28	55
ELL	1,750	2394.44	77.39	37	31	21	11	33
Special Education	1,555	2346.47	73.62	64	23	10	4	13
CD 504	398	2425.50	79.95	25	30	24	21	45
Title I	1,002	2437.43	83.29	21	27	25	27	52
Grade 4								
All Students	10,468	2476.12	94.60	27	20	25	28	53
Female	5,148	2485.99	91.24	23	20	26	32	58
Male	5,320	2466.56	96.80	31	19	24	25	49
African American	3,193	2437.86	87.90	40	24	22	14	36
AmerIndian/Alaskan	40	2472.55	92.72	20	28	28	25	53
Asian	417	2546.97	92.13	7	10	24	59	83
Hispanic	1,996	2454.23	87.72	35	21	24	20	44
Pacific Islander	18	2518.57	81.02	17	0	33	50	83
White	4,312	2506.80	88.57	15	17	28	40	68
Multi-Racial	492	2482.91	90.40	24	19	28	29	57
ELL	1,651	2442.09	80.33	38	23	26	13	39
Special Education	1,707	2385.76	77.69	67	19	10	4	14
CD 504	448	2473.19	85.30	25	25	25	24	50
Title I	1,059	2488.46	81.65	18	24	30	29	59
Grade 5								
All Students	10,827	2514.25	95.15	23	20	32	25	57
Female	5,282	2525.89	91.00	18	20	33	28	62
Male	5,545	2503.15	97.67	27	20	31	22	53
African American	3,309	2474.30	90.00	36	25	27	12	39
AmerIndian/Alaskan	28	2501.32	84.49	18	29	39	14	54
Asian	392	2585.20	87.81	7	9	27	56	83
Hispanic	2,021	2494.38	86.02	26	25	33	15	48
Pacific Islander	12	2525.91	103.62	17	17	25	42	67
White	4,548	2545.37	89.03	13	16	35	37	71
Multi-Racial	517	2520.38	92.61	20	23	32	25	57
ELL	1,264	2456.63	74.34	40	32	25	3	28
Special Education	1,802	2416.98	79.93	64	21	13	2	15
CD 504	555	2514.01	82.80	19	23	38	21	58
Title I	1,050	2525.12	86.19	16	22	38	25	62

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

Table 20. ELA/Lit Percentage of Students in Achievement Levels for Overall and by Subgroup (Grades 6–8)

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
Grade 6								
All Students	10,572	2528.86	97.58	23	25	34	18	52
Female	5,281	2541.36	94.09	19	25	35	21	57
Male	5,291	2516.39	99.40	28	25	32	15	47
African American	3,249	2491.17	92.97	36	28	28	8	36
AmerIndian/Alaskan	45	2527.03	101.40	20	31	31	18	49
Asian	375	2599.24	89.82	6	15	35	44	79
Hispanic	1,863	2504.52	93.12	31	28	31	10	41
Pacific Islander	12	2587.68	79.19	8	8	58	25	83
White	4,573	2558.99	90.34	13	22	39	27	65
Multi-Racial	455	2535.55	91.43	18	31	34	17	51
ELL	752	2442.77	76.88	57	31	11	1	11
Special Education	1,697	2427.76	80.76	65	24	10	1	11
CD 504	547	2523.30	84.68	23	27	37	14	50
Title I	1,019	2542.81	88.80	15	25	41	18	59
Grade 7								
All Students	10,540	2555.38	102.66	23	23	36	19	55
Female	5,299	2572.40	98.69	18	21	38	23	61
Male	5,241	2538.17	103.74	28	24	34	14	48
African American	3,169	2514.73	97.96	35	27	30	8	38
AmerIndian/Alaskan	37	2550.26	107.03	27	16	41	16	57
Asian	376	2637.25	96.22	7	8	33	51	85
Hispanic	1,880	2532.32	95.71	28	27	33	11	45
Pacific Islander	11	2525.30	131.52	45	9	18	27	45
White	4,629	2585.74	95.44	13	19	41	26	68
Multi-Racial	438	2558.52	98.23	21	26	36	17	53
ELL	534	2455.35	81.87	60	26	13	0	14
Special Education	1,623	2446.91	81.88	65	24	10	1	11
CD 504	570	2555.51	95.04	20	28	36	16	52
Title I	1,191	2567.39	93.53	17	23	40	19	60
Grade 8								
All Students	10,207	2566.18	103.51	23	25	35	17	52
Female	5,085	2582.53	99.25	18	24	38	21	58
Male	5,122	2549.95	105.09	28	26	32	14	46
African American	3,198	2528.18	98.17	35	29	28	8	36
AmerIndian/Alaskan	51	2557.33	104.96	25	27	31	16	47
Asian	384	2643.28	97.96	8	10	36	46	82
Hispanic	1,784	2539.85	97.04	30	29	32	9	41
Pacific Islander	11	2602.57	89.98	0	45	27	27	55
White	4,389	2597.23	96.60	13	21	41	25	65
Multi-Racial	390	2572.94	102.40	19	27	34	20	54
ELL	451	2457.65	80.92	65	27	7	1	8
Special Education	1,525	2458.29	82.79	64	25	10	1	11
CD 504	559	2562.52	93.75	21	30	34	14	49
Title I	1,274	2568.49	94.73	19	27	40	14	54

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

Table 21. Mathematics Percentage of Students in Achievement Levels for Overall and by Subgroup (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	10,287	2439.79	84.71	25	22	29	24	53
Female	5,011	2440.11	81.55	24	23	29	24	53
Male	5,276	2439.49	87.62	25	22	28	25	53
African American	3,109	2403.78	78.95	39	27	23	11	35
AmerIndian/Alaskan	24	2428.04	102.12	38	13	21	29	50
Asian	424	2512.60	83.95	7	8	27	57	84
Hispanic	1,974	2419.02	77.83	31	27	27	15	42
Pacific Islander	9*							
White	4,253	2467.54	77.35	13	19	34	34	68
Multi-Racial	494	2448.68	81.62	21	23	28	28	56
ELL	1,814	2415.21	75.85	33	26	28	13	41
Special Education	1,549	2359.46	78.94	63	21	12	4	16
CD 504	398	2439.55	75.83	24	25	31	21	51
Title I	1,007	2455.31	76.80	16	23	32	28	61
Grade 4								
All Students	10,522	2484.14	85.81	19	30	29	22	51
Female	5,172	2483.09	80.62	18	32	29	21	50
Male	5,350	2485.16	90.55	20	28	28	24	52
African American	3,196	2446.52	78.23	31	38	23	9	31
AmerIndian/Alaskan	40	2476.13	87.38	18	33	33	18	50
Asian	432	2559.05	93.63	4	13	27	56	83
Hispanic	2,035	2465.09	77.67	25	34	27	14	41
Pacific Islander	18	2516.43	89.82	17	22	33	28	61
White	4,313	2512.79	78.91	10	25	33	32	66
Multi-Racial	488	2489.83	82.69	18	27	30	25	55
ELL	1,722	2457.81	76.29	27	35	26	11	38
Special Education	1,700	2402.14	76.92	55	30	11	3	14
CD 504	446	2486.08	76.27	17	30	34	19	53
Title I	1,066	2500.07	72.67	10	29	36	25	61
Grade 5								
All Students	10,852	2510.75	93.24	28	27	20	24	44
Female	5,295	2511.57	88.75	27	29	21	23	44
Male	5,557	2509.97	97.34	29	26	19	26	45
African American	3,312	2466.02	84.86	45	31	14	10	24
AmerIndian/Alaskan	28	2492.88	88.96	46	18	14	21	36
Asian	398	2596.29	83.53	5	12	21	62	83
Hispanic	2,044	2493.62	84.05	32	33	19	16	35
Pacific Islander	12	2533.81	110.03	25	8	25	42	67
White	4,542	2542.87	86.64	16	24	24	36	60
Multi-Racial	516	2517.47	88.13	25	31	21	23	45
ELL	1,308	2464.79	75.40	45	34	15	7	22
Special Education	1,801	2419.16	78.19	70	21	6	3	9
CD 504	555	2509.18	79.26	23	37	21	18	39
Title I	1,051	2528.26	81.79	18	28	26	28	53

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

Table 22. Mathematics Percentage of Students in Achievement Levels for Overall and by Subgroup (Grades 6–8)

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
Grade 6								
All Students	10,607	2514.54	107.18	33	30	19	18	38
Female	5,291	2517.99	103.26	31	31	20	18	38
Male	5,316	2511.11	110.84	34	29	19	18	37
African American	3,243	2467.40	100.28	50	30	14	7	20
AmerIndian/Alaskan	45	2524.93	112.43	22	40	18	20	38
Asian	379	2620.16	102.55	7	19	18	55	74
Hispanic	1,900	2489.10	100.24	41	32	17	10	27
Pacific Islander	12	2556.29	110.42	17	25	25	33	58
White	4,574	2549.33	96.36	20	29	24	27	51
Multi-Racial	454	2516.96	102.79	34	29	19	18	37
ELL	811	2433.10	90.79	64	28	5	2	8
Special Education	1,698	2405.70	92.07	78	17	4	1	5
CD 504	547	2516.04	95.11	31	33	22	15	36
Title I	1,018	2536.33	93.33	22	33	24	21	45
Grade 7								
All Students	10,572	2536.23	111.83	32	27	22	19	41
Female	5,308	2540.47	109.32	30	28	23	19	42
Male	5,264	2531.95	114.16	33	27	21	19	40
African American	3,160	2487.09	101.64	48	30	15	7	22
AmerIndian/Alaskan	38	2523.46	105.42	39	29	18	13	32
Asian	378	2651.39	116.01	7	14	21	57	78
Hispanic	1,923	2511.32	104.97	39	29	19	13	31
Pacific Islander	11	2496.66	152.32	55	18	0	27	27
White	4,624	2571.23	102.00	19	26	27	27	55
Multi-Racial	438	2533.29	107.45	34	28	20	18	38
ELL	603	2437.08	97.35	70	22	6	2	9
Special Education	1,618	2417.31	89.72	78	18	3	1	5
CD 504	565	2537.76	101.22	29	33	22	17	38
Title I	1,187	2551.30	100.75	25	26	27	21	48
Grade 8								
All Students	10,232	2546.44	119.07	37	25	18	20	38
Female	5,102	2553.28	113.77	34	27	20	20	40
Male	5,130	2539.64	123.75	40	24	17	19	36
African American	3,197	2496.98	106.34	53	27	13	8	20
AmerIndian/Alaskan	51	2554.82	110.31	39	25	16	20	35
Asian	388	2658.26	136.12	12	13	19	56	75
Hispanic	1,807	2517.76	107.58	45	27	17	11	27
Pacific Islander	11	2579.17	81.86	18	27	36	18	55
White	4,389	2583.96	111.88	24	24	23	29	52
Multi-Racial	389	2549.17	112.21	34	29	19	18	38
ELL	493	2443.99	98.36	77	15	5	3	8
Special Education	1,522	2423.68	88.61	83	14	2	1	4
CD 504	559	2545.35	105.12	36	30	20	15	35
Title I	1,272	2549.39	113.75	33	29	20	18	38

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

Figure 1. ELA/Lit Percent Proficient Across Years

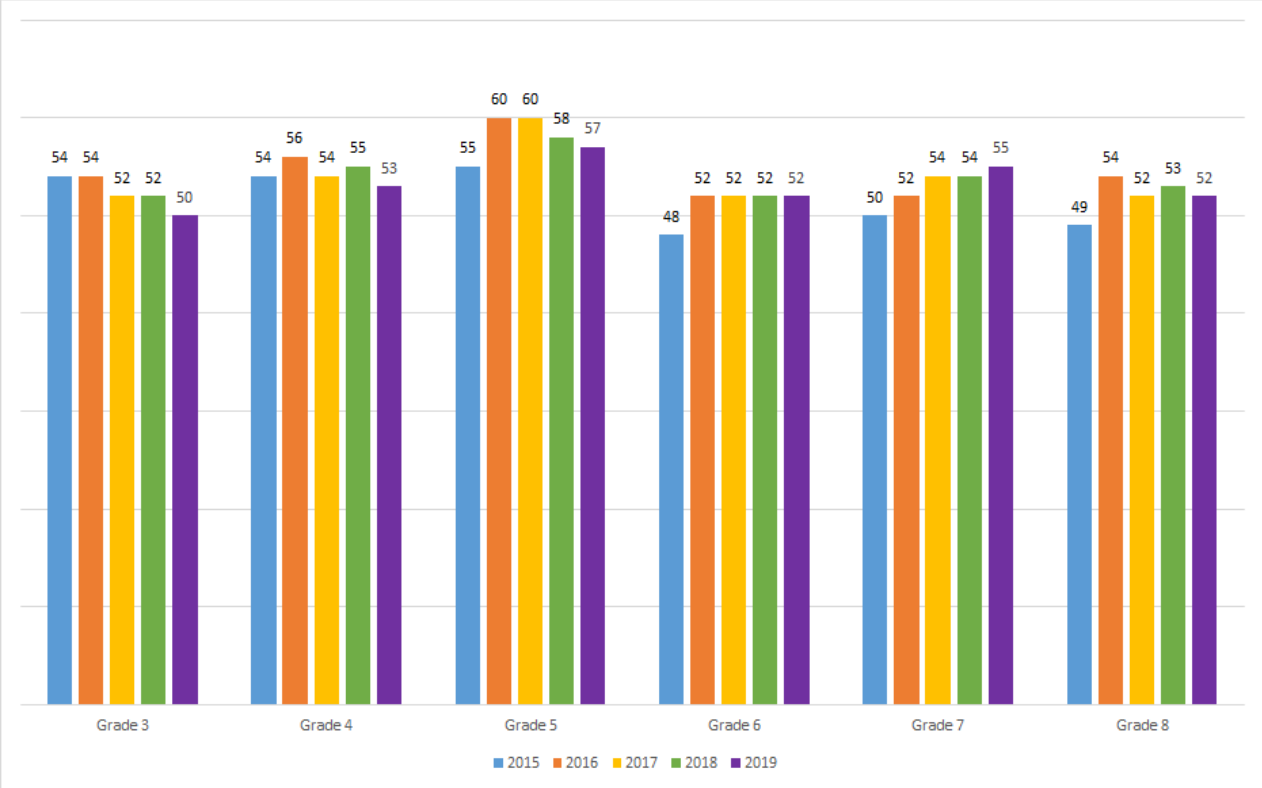


Figure 2. Mathematics Percent Proficient Across Years

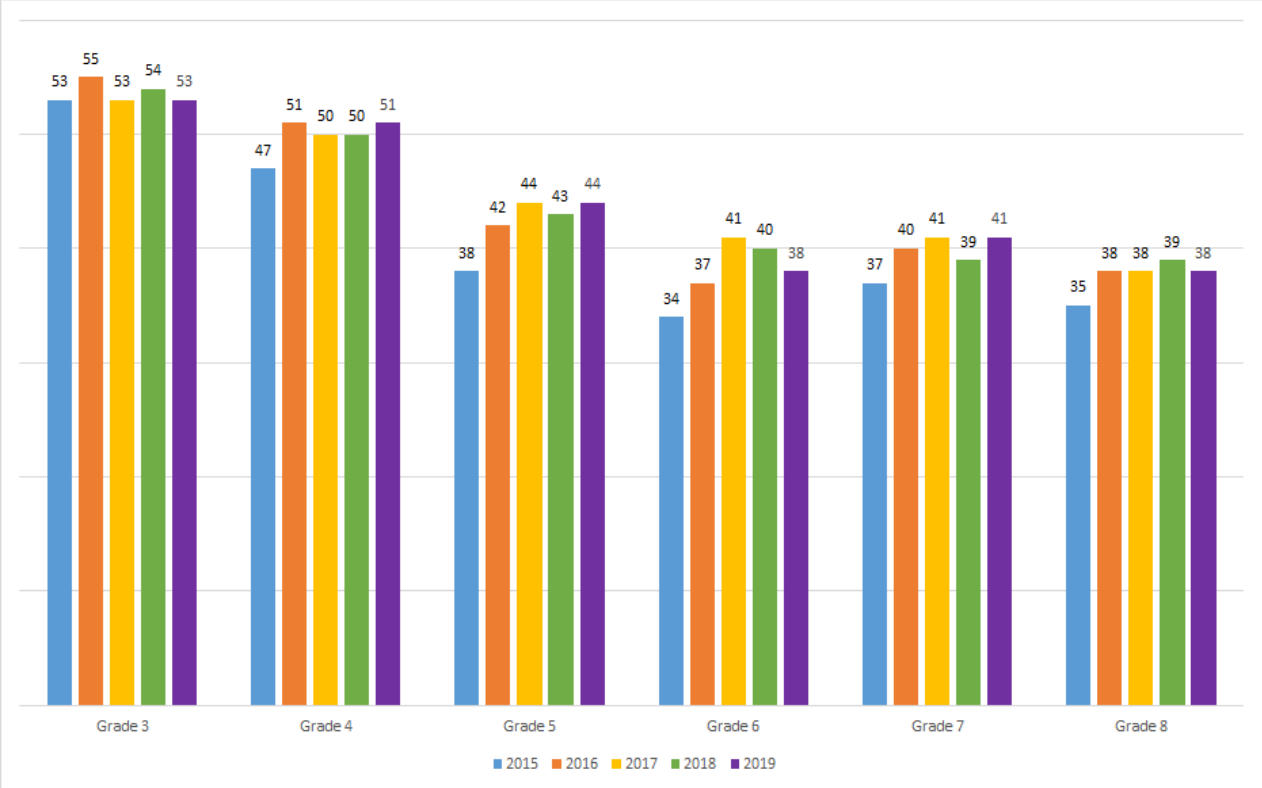


Figure 3. ELA/Lit Average Scale Score Across Years

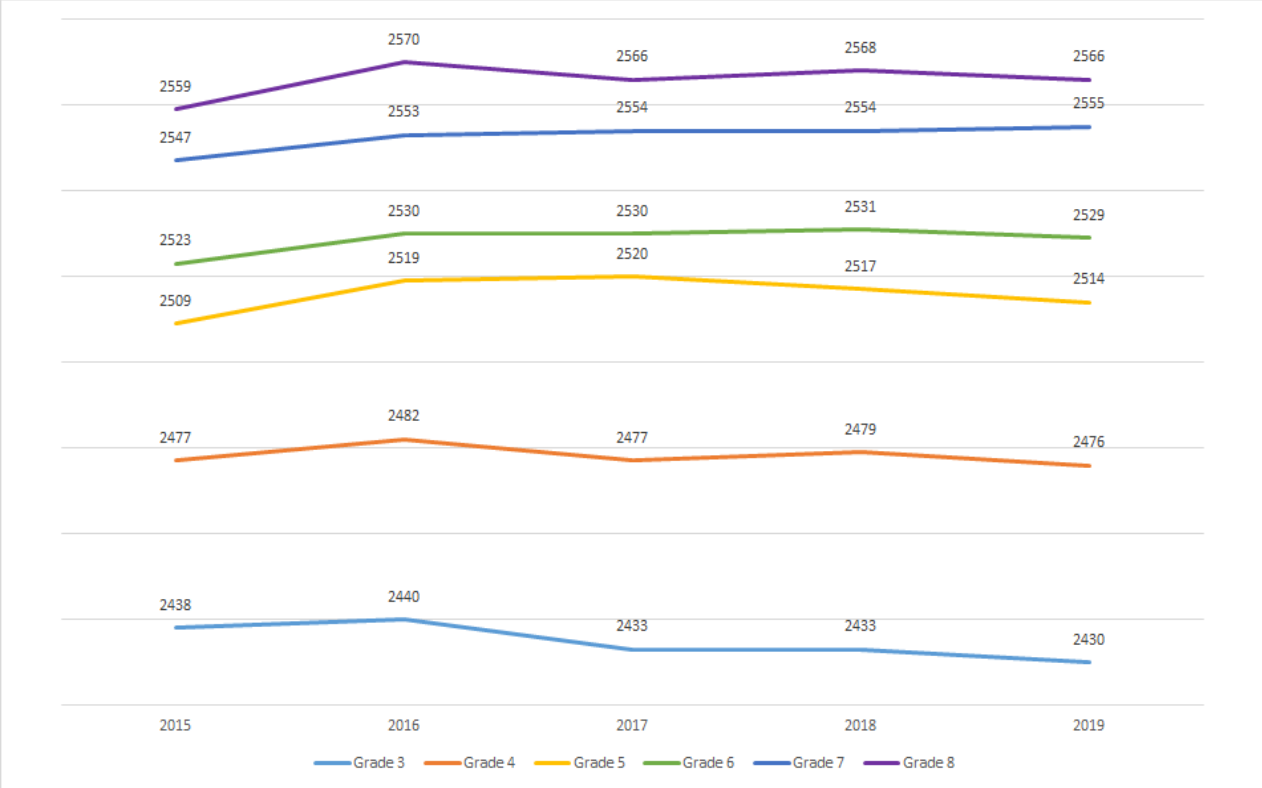
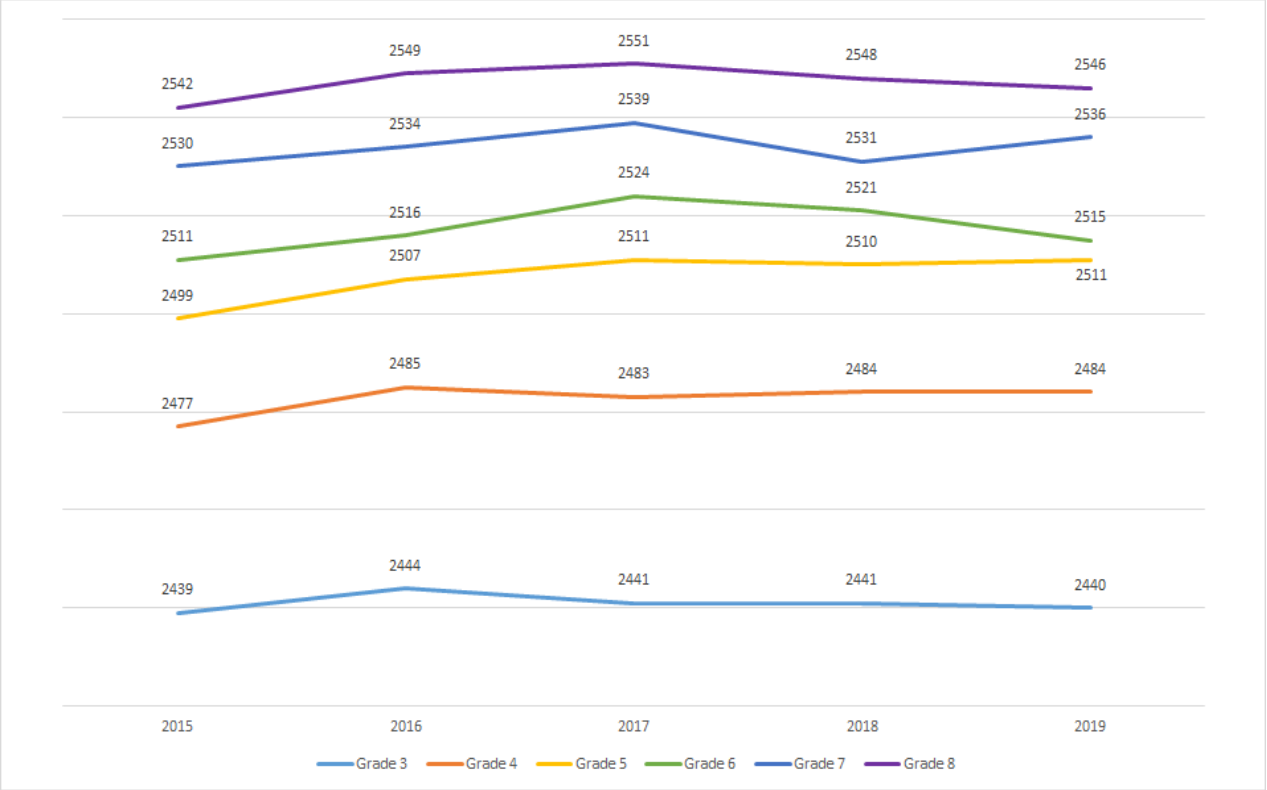


Figure 4. Mathematics Average Scale Score Across Years



Given the small number of items, the precision of claim-level scores is insufficient for reporting purpose. Instead, student performance on each claim is reported using three categories, (1) Below Standard, (2) At/Near Standard, or (3) Above Standard (see Section 6.5, Rules for Calculating Strengths and Weaknesses for Claim Scores) when taking the Standard Error of Measurement (SEM) into account. Tables 23 and 24 present the distribution of performance categories for each claim by grade and test. There are four claims in ELA/lit and three claims in mathematics by combining claims 2 and 4.

Table 23. ELA/Lit Percentage of Students in Performance Categories by Claim

Grade	Performance Category	Claim 1: Reading	Claim 2: Writing	Claim 3: Listening	Claim 4: Research
3	Below	25	28	16	25
	At/Near	48	51	62	50
	Above	27	22	22	25
4	Below	24	24	16	24
	At/Near	48	53	62	51
	Above	27	23	23	25
5	Below	22	19	18	22
	At/Near	47	55	62	47
	Above	31	26	20	31
6	Below	30	25	18	21
	At/Near	46	54	62	52
	Above	25	21	20	27
7	Below	28	22	19	21
	At/Near	45	51	66	50
	Above	26	27	15	29
8	Below	28	23	17	24
	At/Near	44	53	64	49
	Above	28	23	18	27

Table 24. Mathematics Percentage of Students in Performance Categories by Claim

Grade	Performance Category	Claim 1: Concepts and Procedures	Claims 2 and 4: Problem Solving and Modeling and Data Analysis	Claim 3: Communicating Reasoning
3	Below	31	24	22
	At/Near	32	47	47
	Above	37	30	32
4	Below	32	27	25
	At/Near	33	47	46
	Above	35	26	29
5	Below	38	30	29
	At/Near	31	47	48
	Above	31	24	23
6	Below	42	38	35
	At/Near	34	43	46
	Above	24	19	19
7	Below	41	32	24
	At/Near	33	46	56
	Above	26	22	21
8	Below	42	34	30
	At/Near	33	43	50
	Above	24	22	19

3.3 DISTRIBUTION OF STUDENT ABILITY AND ITEM DIFFICULTY

Figures 5–10 display the empirical distribution of the 2018-2019 Delaware student scale scores on summative assessments and the distribution of item difficulty parameters by grade and test and at the claim level. For overall, the student ability distribution is shifted to the left in all grades and subjects, a pattern more pronounced in mathematics for upper grades, indicating that the pool includes more difficult items than the ability of students in tested population per grade. This indicates that the pool includes enough difficult items to accurately measure high-performing students but needs additional easy items to better measure low-performing students. At the reporting category, the student ability distribution is shifted to the left in claims 1 (reading) and 4 (research) in ELA/lit. In mathematics, the student ability distribution is shifted to the left for all claims except for claim 1 in lower grades. The Smarter Balanced Assessment Consortium plans to add more 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/Lit

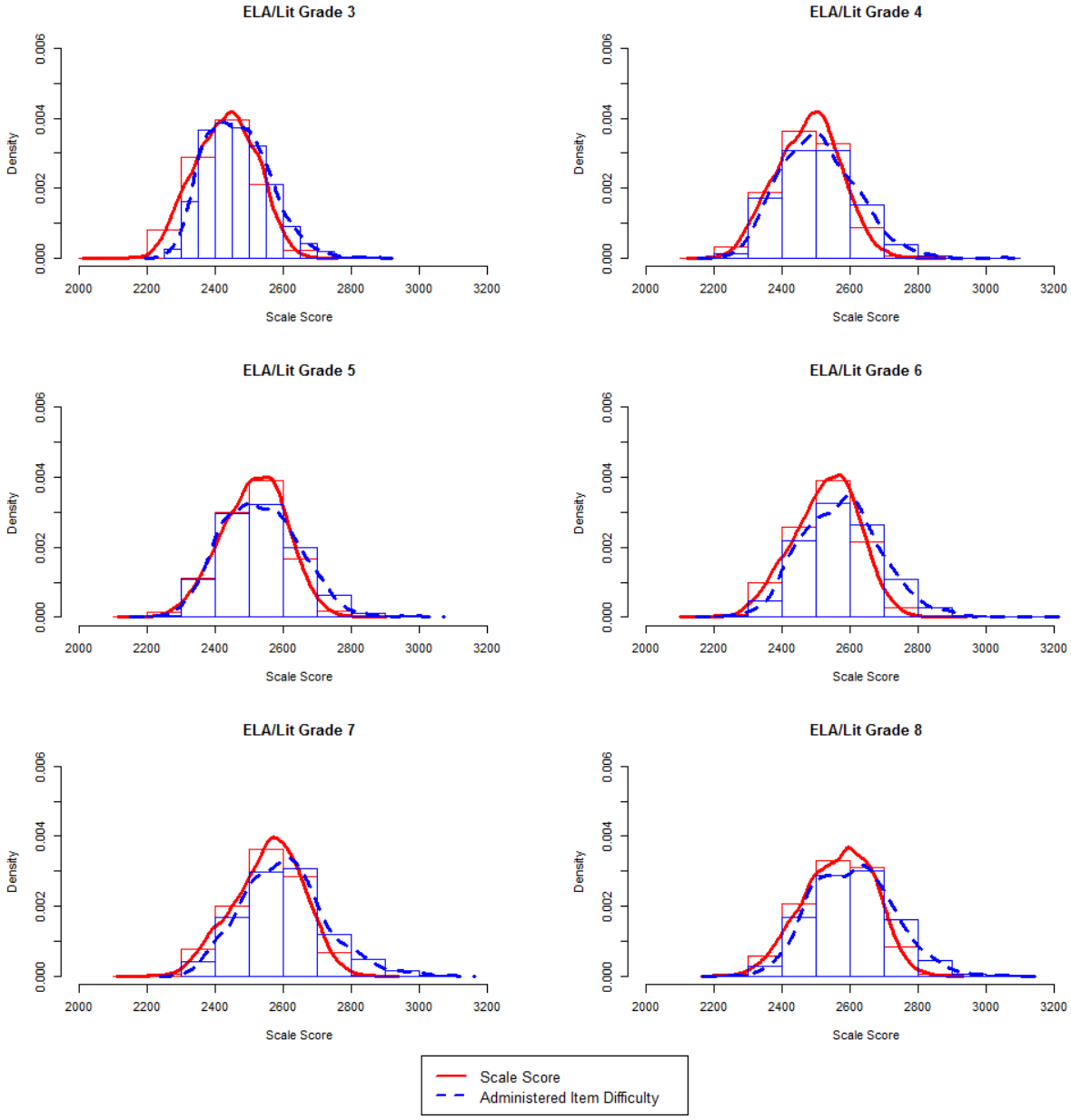


Figure 6. Student Ability–Item Difficulty Distribution by Claim: ELA/Lit (Grades 3–5)

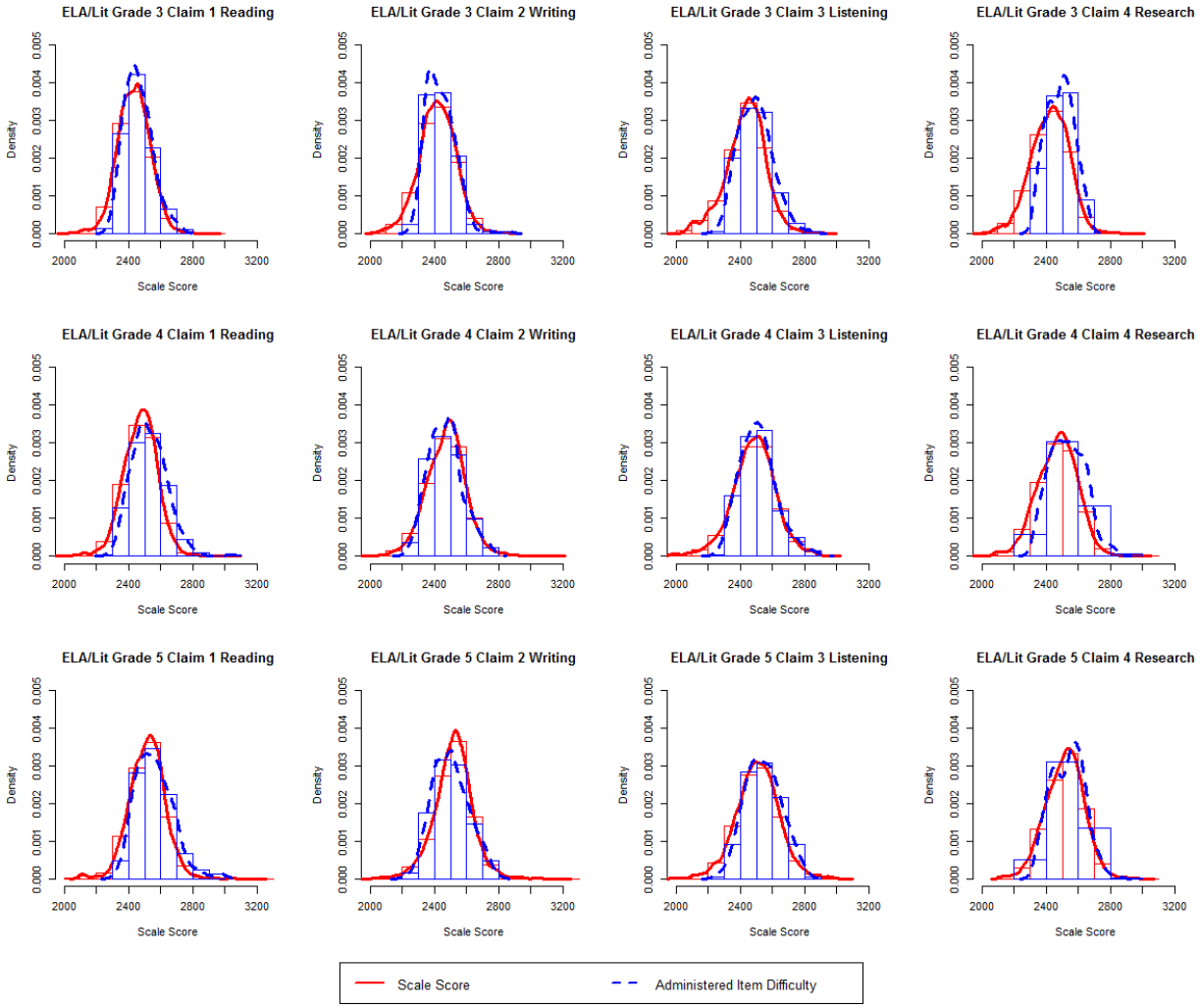


Figure 7. Student Ability–Item Difficulty Distribution by Claim: ELA/Lit (Grades 6–8)

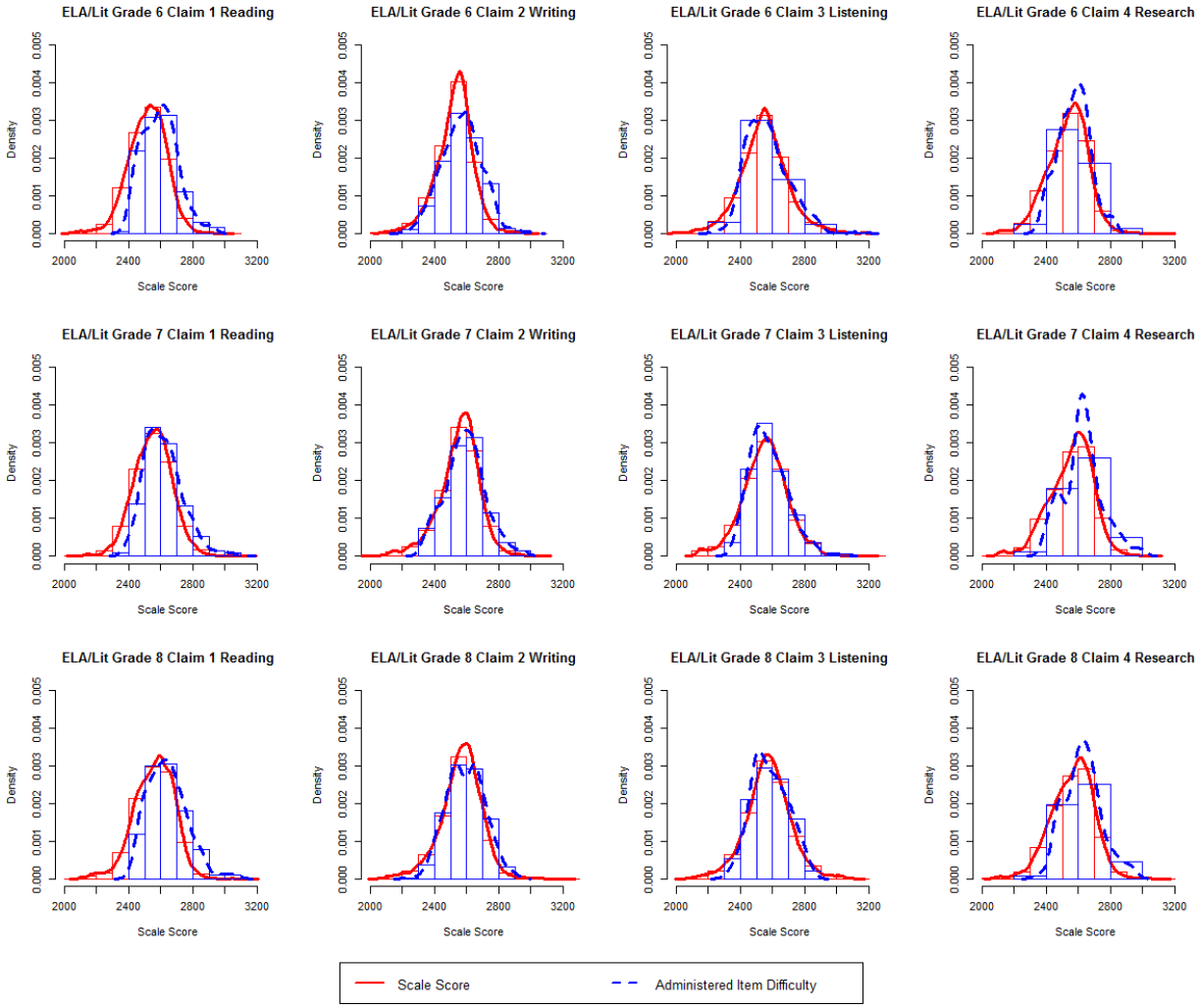


Figure 8. Student Ability–Item Difficulty Distribution for Mathematics

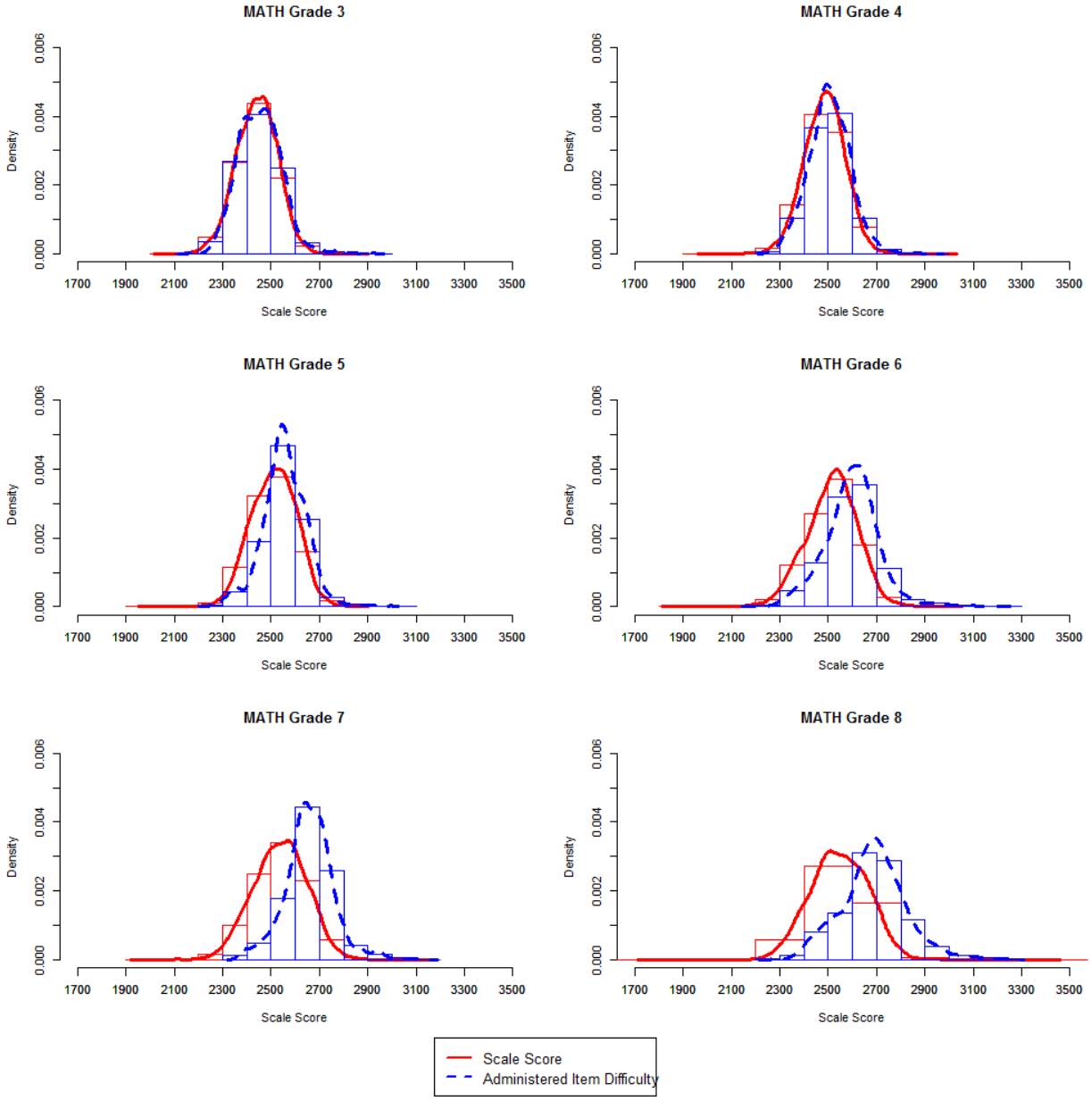


Figure 9. Student Ability–Item Difficulty Distribution by Claim: Mathematics (Grades 3–5)

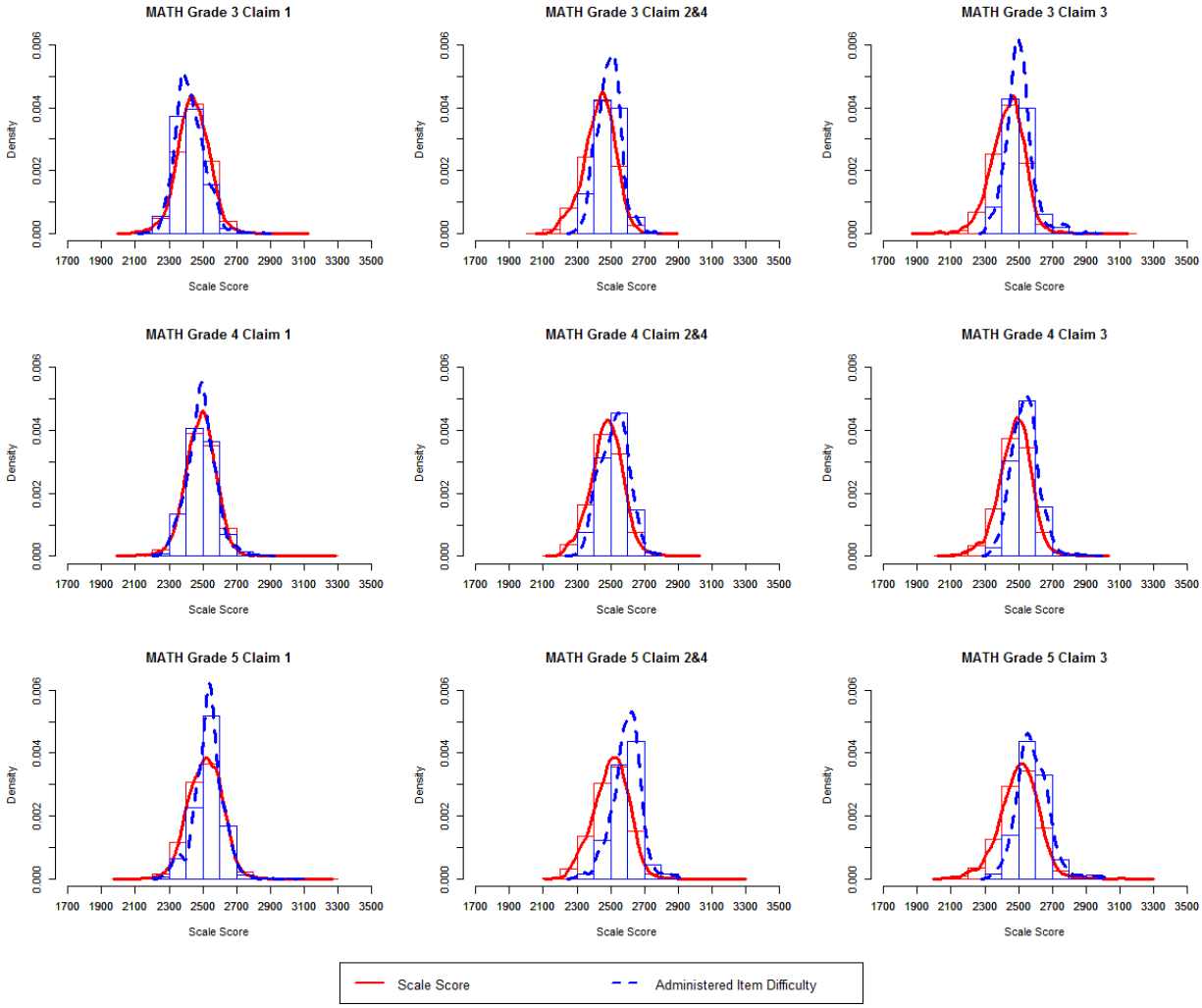
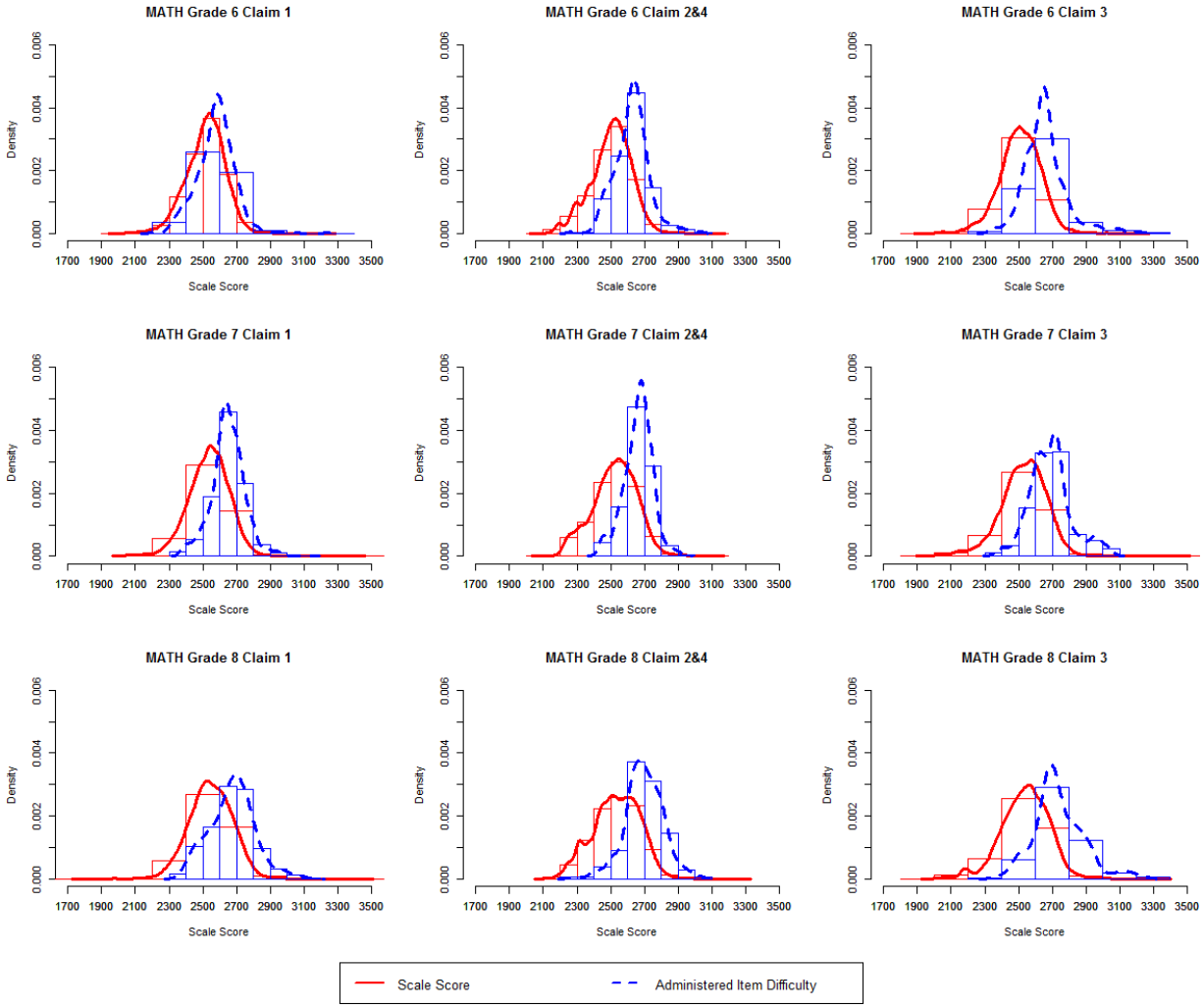


Figure 10. Student Ability–Item Difficulty Distribution by Claim: Mathematics (Grades 6–8)



3.4 TEST-TAKING TIME

The Smarter Balanced assessments are not timed. The time spent on each item may vary among individual students, which may provide useful information about student testing behaviors and motivation, for example. Since the length of a test session could be monitored by TAs who are knowledgeable about their schools and their students, additional time for students who need it would be arranged.

Tables 25 and 26 present an average testing time and the testing time at percentiles for the overall test, the CAT component, and the PT component.

Table 25. ELA/Lit Test-Taking Time

Grade	Average Testing Time (hh:mm)	Median Testing Time (hh:mm)	Testing Time in Percentiles (hh:mm)				
			75th	80th	85th	90th	95th
Overall Test							
3	4:25	3:56	5:31	6:00	6:38	7:33	9:10
4	4:47	4:16	5:57	6:28	7:12	8:05	9:43
5	4:42	4:15	5:46	6:14	6:49	7:38	8:57
6	4:16	3:55	5:15	5:38	6:08	6:51	8:16
7	3:48	3:30	4:36	4:57	5:23	6:01	6:59
8	3:48	3:31	4:40	5:02	5:27	6:06	7:03
CAT Component							
3	2:07	1:52	2:33	2:46	3:04	3:28	4:16
4	2:16	2:01	2:43	2:57	3:15	3:39	4:28
5	2:14	2:04	2:41	2:53	3:08	3:29	4:07
6	2:15	2:06	2:43	2:54	3:09	3:32	4:05
7	1:56	1:48	2:20	2:30	2:42	2:59	3:27
8	1:57	1:49	2:21	2:31	2:43	3:01	3:30
PT Component							
3	2:18	1:57	3:03	3:21	3:46	4:22	5:23
4	2:31	2:09	3:16	3:37	4:03	4:40	5:50
5	2:27	2:09	3:08	3:26	3:52	4:23	5:20
6	2:01	1:45	2:36	2:52	3:13	3:41	4:33
7	1:52	1:38	2:21	2:34	2:53	3:18	4:02
8	1:52	1:36	2:22	2:37	2:57	3:21	4:06

Table 26. Mathematics Test-Taking Time

Grade	Average Testing Time (hh:mm)	Median Testing Time (hh:mm)	Testing Time in Percentiles (hh:mm)				
			75th	80th	85th	90th	95th
Overall Test							
3	2:40	2:18	3:14	3:34	4:03	4:42	5:40
4	2:54	2:34	3:36	3:54	4:17	4:50	5:49
5	3:18	2:56	4:02	4:23	4:51	5:31	6:39
6	2:51	2:35	3:25	3:40	3:59	4:27	5:25
7	2:18	2:06	2:49	3:01	3:17	3:39	4:16
8	2:35	2:25	3:09	3:23	3:40	4:02	4:41
CAT Component							
3	1:45	1:30	2:06	2:19	2:37	3:05	3:48
4	1:59	1:44	2:27	2:41	2:57	3:23	4:07
5	1:56	1:44	2:20	2:31	2:46	3:10	3:52
6	1:52	1:42	2:16	2:26	2:38	2:56	3:31
7	1:42	1:33	2:05	2:14	2:27	2:43	3:09
8	1:53	1:46	2:19	2:30	2:41	2:58	3:28
PT Component							
3	0:55	0:44	1:10	1:18	1:28	1:44	2:14
4	0:55	0:46	1:09	1:17	1:26	1:39	2:02
5	1:23	1:09	1:45	1:55	2:11	2:32	3:14
6	0:59	0:49	1:12	1:19	1:28	1:42	2:10
7	0:36	0:30	0:46	0:51	0:58	1:07	1:24
8	0:42	0:36	0:53	0:59	1:06	1:16	1:34

4. VALIDITY

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

Validity evidence provided in this chapter is as follows:

- Test Content
- Internal Structure

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

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

4.1 EVIDENCE ON TEST CONTENT

The Smarter Balanced summative assessment includes two components: the computer-adaptive test (CAT) and the performance task (PT). For the CAT, each student receives a different set of items adapted to his or her ability. For the PT, each student is administered a fixed-form test. The content coverage in all PT forms is the same.

In the adaptive item-selection algorithm, item selection takes place in two discrete stages: blueprint satisfaction and match-to-ability. The Smarter Balanced blueprints (Smarter Balanced Assessment Consortium, 2015) specify a range of items to be administered in each claim, content domain/standard, and target. Moreover, blueprints constrain the DOK and item and passage types. For DOK constraints, the Smarter Balanced blueprint specifies the minimum number of items, not the maximum. In blueprints, all content blueprint elements are configured to obtain a strictly enforced range of items administered. The algorithm also seeks to satisfy target-level constraints, but these ranges are not strictly enforced. In ELA/lit, the blueprints also specify the number of passages in reading (claim 1) and listening (claim 3) claims.

Tables 27 and 28 present the percentages of tests aligned with the ELA/L test blueprint constraints for items in claims, targets and DOK, and passages in claims 1 and 3. For the passage constraints, four passages in claim 1 reading and three to four passages in claim 3 listening are required. The composition of four reading passages in claim 1 is two literary-text passages (one long and one short passage) and two informational-text passages (one long and one short passage) in grades 3–5 and one literary-text passage (long passage) and three information-text passages (one long and two short passages) in grades 6–8.

All Ela/Lit tests met the blueprint requirements, except for claims 1 and 2 targets, which administered a few items more or less than the item requirement. The violations in claim 1 reading targets (e.g., target 9 and target sets of 9 and 11) appeared in most grades due to the uneven distribution of items across targets

and DOKs within and across passages. The violations in claim 2 writing targets are appeared in grade 6 due to the uneven distribution of items across targets and DOKs.

Tables 29 and 30 provide the percentages of tests aligned with the test blueprint constraints for the mathematics CAT for claims, DOK, and target constraints. In mathematics, the tests met all blueprint requirements, except for grade 6. In grade 6, the violation was in the claim 1 for target sets of B and G, each administered fewer or more items than required.

Table 27. ELA/Lit Percentage of Delivered Tests Meeting Blueprint Requirements for Each Claim and the Number of Passages Administered (Grades 3-5)

Claim	Content Category/Target	Required Itemsin G3–5	%BP Match for Item Requirements			%BP Match for Passage Requirement
			G3	G4	G5	G3–5
1	**Literary Text**	7–8	100	100	100	100
	Target 2: Central Ideas	1–2	100	100	100	
	Target 4: Reasoning and Evaluation	1–2	100	100	100	
	Targets 1, 3, 5, 6, and 7	3–6	100	100	100	
	Informational Text	7–8	100	100	100	100
	Target 9: Central Ideas	1–2	83	98	99	
	Target 11: Reasoning and Evaluation	1–2	100	100	100	
	Targets 8, 10, 12, 13, and 14	3–6	100	100	100	
	DOK 2	≥ 7	100	100	100	
	DOK 3 or 4	≥ 2	100	100	100	
2	**Writing**	6	100	100	100	
	Target 1, 3, or 6: Organization/Purpose	1	100	100	100	
	Target 1, 3, or 6: Evidence/Elaboration	1	100	100	100	
	Target 8: Language and Vocabulary Use	1	100	100	100	
	Target 9: Edit/Clarify	3	100	100	100	
	DOK 2 or higher	≥ 2	100	100	100	
3	**Listening**	8–9	100	100	100	100
	Target 4: Listen/Interpret	8–9	100	100	100	
	DOK 2 or higher	≥ 3	100	100	100	
4	**Research**	8	100	100	100	
	Target 2: Interpret and Integrate Information	2–3	100	100	100	
	Target 3: Analyze Information/Sources	2–3	100	100	100	
	Target 4: Use Evidence	2–3	100	100	100	

Table 28. ELA/Lit Percentage of Delivered Tests Meeting Blueprint Requirements for Each Claim and the Number of Passages Administered (Grades 6-8)

Claim	Content Category/Target	Required Items in G6–8	%BP Match for Item Requirements			%BP Match for Passage Requirement
			G6	G7	G8	G6–8
1	**Literary Text**	4–7	100	100	100	100
	Target 2: Central Ideas	1	99	100	100	
	Target 4: Reasoning and Evaluation	1	100	100	100	
	Targets 1, 3, 5, 6, and 7	2–5	100	100	100	
	Target 2 or 4 short text	0–1	100	100	100	
	Informational Text	10–12*	100	100	100	100
	Targets 9 and 11	2–5	100	98	96	
	Targets 8, 10, 12, 13, and 14	7–10	100	98	96	
	Target 9 or 11 short text	0–1	100	100	100	
	DOK 1	≤ 5	100	100	100	
	DOK 3 or 4	≥ 2	100	100	100	
2	**Writing**	6	100	100	100	
	Target 1, 3, or 6: Organization/Purpose	1	100	100	100	
	Target 1, 3, or 6: Evidence/Elaboration	1	100	100	100	
	Target 8: Language and Vocabulary Use	1	94	100	100	
	Target 9: Edit/Clarify	3	94	100	100	
	DOK 2	≥ 2	100	100	100	
	DOK 3 or 4	1	100	100	100	
	Brief Write	1	100	100	100	
3	**Listening**	8–9	100	100	100	100
	Target 4: Listen/Interpret	8–9	100	100	100	
	DOK 2 or higher	≥ 3	100	100	100	
4	**Research**	8	100	100	100	
	Target 2: Interpret and Integrate Information	2–3	100	100	100	
	Target 3: Analyze Information/Sources	2–3	100	100	100	
	Target 4: Use Evidence	2–3	100	100	100	

* Required items for Informational Text are 10–12 in grades 6 and 7 and 12 for grade 8.

Table 29. Mathematics Percentage of Delivered Tests Meeting Blueprint Requirements for Each Claim and Targets (Grades 3–5)

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

Table 30. Mathematics Percentage of Delivered Tests Meeting Blueprint Requirements for Each Claim and Targets (Grades 6–8)

Claim	Content / Target	Grade 6		Grade 7		Grade 8	
		Required Items	**% BP Match**	**Required Items**	**% BP Match**	**Required Items**	**% BP Match**
1	Overall	16–20	100	16–20	100	16–20	100
	DOK 2 or higher	≥ 7	100	≥ 7	100	≥ 7	100
	Priority Cluster	12–15	100				
	Targets E, F	5–6	100				
	Target A	3–4	100				
	Targets G, B	2	99				
	Target D	2	100				
	Supporting Cluster	4–5	100				
	Targets C, H, I, J	4–5	100				
	Priority Cluster			12–15	100		
	Targets A, D			8–9	100		
	Targets B, C			5–6	100		
	Supporting Cluster			4–5	100		
	Targets E, F			2–3	100		
	Targets G, H, I			1–2	100		
	Priority Cluster					12–15	100
	Targets C, D					5–6	100
	Targets B, E, G					5–6	100
	Targets F, H					2–3	100
	Supporting Cluster					4–5	100
	Targets A, I, J					4–5	100
2 and 4	Overall	6	100	6	100	6	100
	DOK 3 or higher	≥ 2	100	≥ 2	100	≥ 2	100
	2. Target A	2	100	2	100	2	100
	2. Targets B, C, D	1	100	1	100	1	100
	4. Targets A, D	1	100	1	100	1	100
	4. Targets B, E	1	100	1	100	1	100
	4. Targets C, F	1	100	1	100	1	100
3–Calc	Overall	7	100	8	100	8	100
	DOK 3 or higher	≥ 2	100	≥ 2	100	≥ 2	100
	Targets A, D	2-3	100	3	100	3	100
	Targets B, E	2-3	100	3	100	3	100
	Targets C, F, G	1–2	100	2	100	2	100
3–No Calc	Overall	1	100				

Table 31 summarizes the target coverage by claim that includes the average and the range of the number of unique targets administered in each delivered CAT test. Since the test blueprint is not required to cover all targets in each test, it is expected that the number of targets covered slightly varies across individual tests. Although the target coverage varies somewhat across individual tests, all targets are covered at an aggregate level across all tests combined.

Table 31. Average and Range of the Number of Unique Targets Assessed Within Each Claim Across All Delivered Tests

Grade	Total Targets in BP				Mean				Range (Minimum–Maximum)			
	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4
ELA/Lit												
3	14	5	1	3	10.13	4.00	1.00	3.00	8–13	4–4	1–1	3–3
4	14	5	1	3	10.76	4.00	1.00	3.00	8–14	4–4	1–1	3–3
5	14	5	1	3	11.40	4.00	1.00	3.00	9–14	4–4	1–1	3–3
6	14	5	1	3	10.27	4.00	1.00	3.00	8–11	4–5	1–1	3–3
7	14	5	1	3	10.74	4.00	1.00	3.00	8–11	4–4	1–1	3–3
8	14	5	1	3	10.86	4.00	1.00	3.00	8–11	4–4	1–1	3–3
Mathematics												
3	11	4	6	6	10.93	2.00	5.70	3.00	9–11	2–2	4–6	3–4
4	12	4	6	6	10.00	2.00	5.43	3.00	9–11	2–2	3–6	3–3
5	11	4	6	6	9.00	2.00	5.24	3.00	9–9	2–2	3–6	3–3
6	10	4	7	6	9.99	2.00	4.53	3.00	8–10	2–2	3–7	3–3
7	9	4	7	6	8.00	2.00	4.64	3.00	8–8	2–2	3–6	3–3
8	10	4	7	6	10.00	2.00	4.82	3.00	10–10	2–2	3–6	3–4

The adaptive testing algorithm assembles a test form unique to each individual student, targeting the student's level of ability and meeting the test blueprints. These test forms are not statistically parallel (e.g., the same test difficulty); however, they are parallel in test construct that support the comparability in test scores and their interpretations. Since each test form measures the same content based on the blueprints, 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 Smarter Balanced assessments assumes a single underlying latent trait in student ability estimates on both summative and interim assessments, 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 mixture of items from each claim. Assessing the relationship between these different claim scores is a measure of internal validity according to the the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014). A high correlations among claim scores is evidence that the Smarter Balanced assessment measures a single underlying ability and the claim scores are related to each other.

The correlations among claim scores, both observed (below diagonal) and corrected for attenuation (above diagonal, disattenuated correlation), are presented in Tables 32 and 33. The correction for attenuation

indicates what the correlation would be if claim scores could be measured with perfect reliability, corrected (adjusted) for measurement error estimates.

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

When corrected for attenuation (i.e., disttenutated correlation, above diagonal), the correlations among claim scores are higher than observed correlations. The disattenuated correlations are quite high in both subjects, showing evidence of unidimensional tests. The correction for attenuation is large particularly in mathematics because the marginal reliabilities of claim 2 and 4 and claim 3 scores are low. The low reliabilities are due to the larger standard errors for low performing students and a shortage of easy items in the item pool.

Because the reliability for claim scores is low, students' performance in the claim scores is reported in three broad performance categories, Below Standard, At/Near Standard, and Above Standard. The distribution of performance categories for each claim is provided in Tables 23 and 24 in Section 3.2, Summary of Overall Student Performance. Scale scores are not reported for claims.

Table 32. Correlations Among Claim Scores for ELA/Lit

Grade	Claim	Observed and Disattenuated Correlation			
		Claim 1	**Claim 2**	**Claim 3**	**Claim 4**
3	Claim 1: Reading		0.88	0.94	0.91
	Claim 2: Writing	0.66		0.87	0.89
	Claim 3: Listening	0.64	0.58		0.90
	Claim 4: Research	0.68	0.65	0.60	
4	Claim 1: Reading		0.86	0.91	0.91
	Claim 2: Writing	0.64		0.82	0.85
	Claim 3: Listening	0.62	0.55		0.88
	Claim 4: Research	0.67	0.62	0.59	
5	Claim 1: Reading		0.88	0.91	0.92
	Claim 2: Writing	0.64		0.85	0.88
	Claim 3: Listening	0.62	0.57		0.91
	Claim 4: Research	0.70	0.65	0.63	
6	Claim 1: Reading		0.88	0.92	0.92
	Claim 2: Writing	0.66		0.88	0.88
	Claim 3: Listening	0.64	0.60		0.92
	Claim 4: Research	0.68	0.64	0.62	
7	Claim 1: Reading		0.88	0.92	0.92
	Claim 2: Writing	0.67		0.87	0.88
	Claim 3: Listening	0.63	0.58		0.91
	Claim 4: Research	0.70	0.65	0.60	
8	Claim 1: Reading		0.90	0.95	0.93
	Claim 2: Writing	0.68		0.91	0.90
	Claim 3: Listening	0.65	0.61		0.93
	Claim 4: Research	0.70	0.67	0.63	

Table 33. Correlations Among Claim Scores for Mathematics

Grade	Claims	Observed and Disattenuated Correlation		
		Claim 1	**Claims 2 and 4**	**Claim 3**
3	Claim 1		0.96	0.93
	Claims 2 and 4	0.77		0.98
	Claim 3	0.78	0.73	
4	Claim 1		0.96	0.97
	Claims 2 and 4	0.80		1
	Claim 3	0.80	0.77	
5	Claim 1		0.99	0.95
	Claims 2 and 4	0.79		1
	Claim 3	0.77	0.74	
6	Claim 1		1	0.96
	Claims 2 and 4	0.82		1
	Claim 3	0.77	0.75	
7	Claim 1		1	0.94
	Claims 2 and 4	0.79		1
	Claim 3	0.73	0.69	
8	Claim 1		1	0.96
	Claims 2 and 4	0.80		1
	Claim 3	0.76	0.71	

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

5. RELIABILITY

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

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

5.1 MARGINAL RELIABILITY

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 conditional SEM of the scale score for student i; and σ^2 is the variance of the scale score. The higher the reliability coefficient, the greater the precision of the test.

Another way to examine test reliability is with CSEM. In IRT, SEM is estimated as a function of test information provided by a given set of items that makes up the test. In 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 34 presents descriptive statistics, marginal reliability coefficients and the average CSEM for the total scale scores by test and grade.

Table 34. Marginal Reliability for ELA/Lit and Mathematics

Grade	N	Number of Items Specified in Test Blueprint		Marginal Reliability	Scale Score Mean	Scale Score SD	Average CSEM
		Min.	Max.				
ELA/Lit							
3	10,234	38	41	0.92	2429.72	89.67	25.63
4	10,468	38	41	0.92	2476.12	94.60	27.52
5	10,827	38	41	0.92	2514.25	95.15	27.09
6	10,572	38	41	0.92	2528.86	97.58	27.72
7	10,540	38	41	0.92	2555.38	102.66	28.99
8	10,207	40	41	0.92	2566.18	103.51	29.16
Mathematics							
3	10,287	39	40	0.95	2439.79	84.71	19.32
4	10,522	37	40	0.95	2484.14	85.81	19.94
5	10,852	38	40	0.94	2510.75	93.24	22.56
6	10,607	38	39	0.94	2514.54	107.18	27.09
7	10,572	38	40	0.93	2536.23	111.83	28.51
8	10,232	38	40	0.93	2546.44	119.07	30.62

5.2 STANDARD ERROR CURVES

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

Given that classifying students into achievement levels, especially into proficient or not proficient levels based on the Level 3 cut, 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 alorthim that focused on maximizing information over blueprint requirements would produce CSEM curves that are more flat. The Smarter Balenced assessments focus on increasing precision where it is most needed, ability scores near and in between the cut scores. It is worth noting that larger standard errors are observed at the lower ends of the score distribution, relative to the higher ends. This occurs because the item pools currently have a shortage of very easy items that are better targeted toward these lower-achieving students. Content experts use this information to consider how to further target and populate item pools.

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

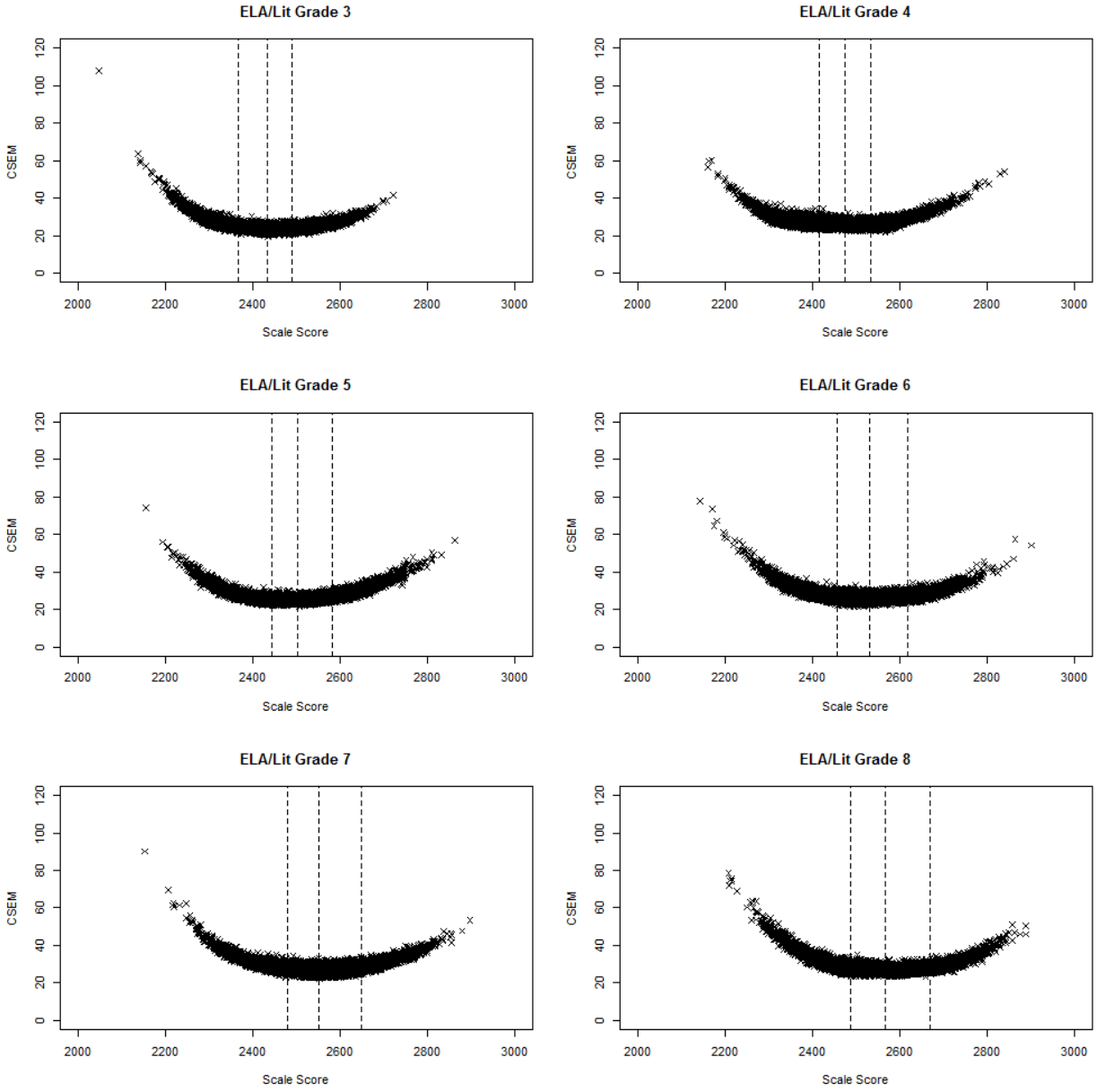
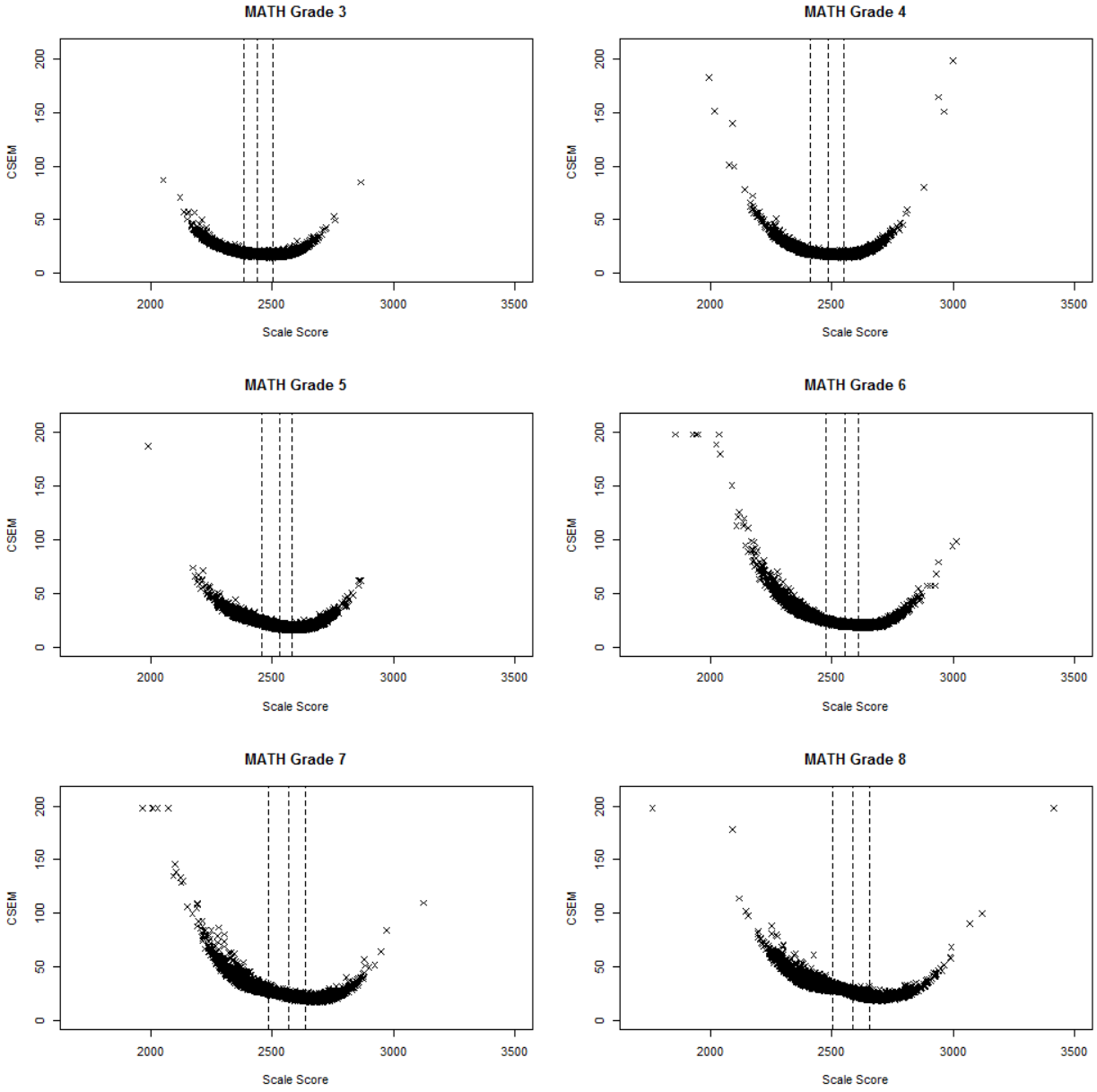


Figure 12. Conditional Standard Error of Measurement for Mathematics



The CSEMs presented in the figures are summarized in Tables 35 and 36 for ELA/lit and mathematics at each grade. Table 35 provides the average CSEM for all scores and by achievement level. Table 36 presents the average conditional SEMs at each cut score and the difference in average conditional SEMs between two cut scores. As shown in Figures 11 and 12 above, the largest average CSEM is at Level 1 in both ELA/lit and mathematics. The average CSEMs at all cut scores are similar in ELA/lit, but larger in Level 2 cut scores in mathematics.

Table 35. Average Conditional Standard Error of Measurement by Achievement Level

Grade	Level 1	Level 2	Level 3	Level 4	Average CSEM
ELA/Lit					
3	28.8	24.1	23.7	25.4	25.6
4	29.0	26.4	26.0	28.2	27.5
5	28.0	24.8	25.4	29.9	27.1
6	30.8	25.7	26.1	29.0	27.7
7	31.8	26.9	27.2	31.1	29.0
8	32.9	27.4	27.1	30.4	29.2
Mathematics					
3	22.9	18.2	17.2	18.7	19.3
4	25.4	18.3	17.1	20.1	19.9
5	28.7	20.8	18.3	19.5	22.6
6	35.7	22.4	20.4	21.9	27.1
7	37.4	25.3	21.8	21.9	28.5
8	37.4	28.9	24.2	23.1	30.6

Table 36. Average Conditional Standard Error of Measurement at Each Achievement Level Cut and Difference of the Standard Error of Measurements Between Two Cuts

Grade	L2 Cut	L3 Cut	L4 Cut	|L2–L3|	|L3–L4|	|L2–L4|
ELA/Lit						
3	24.8	23.6	23.7	1.2	0.0	1.2
4	26.3	26.1	26.1	0.2	0.0	0.3
5	24.8	25.0	26.8	0.1	1.8	2.0
6	25.7	25.6	27.0	0.1	1.4	1.3
7	27.0	27.1	28.6	0.1	1.5	1.6
8	27.4	27.0	27.9	0.4	1.0	0.6
Mathematics						
3	19.0	17.7	17.1	1.3	0.6	1.9
4	19.5	17.4	17.0	2.1	0.4	2.5
5	23.3	18.8	18.3	4.5	0.6	5.0
6	24.3	21.1	19.9	3.2	1.2	4.5
7	27.5	23.4	20.5	4.1	2.8	6.9
8	30.9	26.3	22.2	4.7	4.0	8.7

5.3 RELIABILITY OF ACHIEVEMENT CLASSIFICATION

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

Classification accuracy refers to the agreement between the classifications based on the form actually taken and the classifications that would be made on the basis of the test takers' true scores, if their true scores could somehow be known. Classification consistency refers to the agreement between the classifications based on the form (adaptively administered items) actually taken and the classifications that would be made on the basis of an alternate form (another set of adaptively administered items given the same ability), that is, the percentages of students who are consistently classified in the same achievement levels on two equivalent test forms.

In reality, the true ability is unknown and students do not take an alternate, equivalent form; therefore, the classification accuracy and the classification consistency are estimated based on students' item scores, the item parameters, and the assumed underlying latent ability distribution as described below. The true score is an expected value of the test score with a measurement error.

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

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 $\hat{\theta}_i \sim N\left(\theta_i, se^2(\hat{\theta}_i)\right)$, we can estimate the above probabilities directly using the likelihood function.

The likelihood function of theta, given a student's item scores, represents the likelihood of the student's ability at that theta value. Integrating the likelihood values over the range of theta at and above the cut point (with proper normalization) represents the probability of the student's latent ability or the true score being at or above that cut point. If a student with estimated theta is below the cut point, a probability of 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, we can define various classification probabilities.

The probability of the *i*th student being classified at achievement level l ($l = 1,2,\cdots,L$) based on the cut scores cut_{l-1} and cut_l, given the student's item scores $\mathbf{z}_i = (z_{i1},\cdots,z_{iJ})$ and item parameters $\mathbf{b} = (\mathbf{b}_1,\cdots,\mathbf{b}_J)$ 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; $\mathbf{b}_j = (a_j, b_j, c_j)$ if the *j*th item is a dichotomous item, and $\mathbf{b}_j = (a_j, b_{j1}, \ldots, b_{jK_i})$ if the *j*th item is a polytomous item; a_j is the item's discrimination parameter (for Rasch model, $a_j = 1$), c_j is the guessing parameter (for Rasch and 2PL models, $c_j = 0$), and *D* is 1.7 for non-Rasch models and 1 for Rasch model.

Classification Accuracy

Using p_{il}, we can construct an $L \times L$ table as

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

where $n_{alm} = \sum_{pl_i=l} p_{im}$. n_{alm} is the expected 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 $L \times L$ table, the row represents the observed level and the column represents the expected level.

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

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

and the overall classification accuracy is estimated by

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

where N is the total number of students.

Classification Consistency

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

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

where $n_{clm} = \sum_{i=1}^{N} p_{il} p_{im}$. p_{il} , and p_{im} are the probabilities of the *i*th student being classified at achievement level l and *m*, respectively, based on observed scores and hypothetical scores from 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

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

The analysis of the classification index is performed based on overall scale scores. Table 37 provides the results of classification accuracy and consistency both overall and by achievement level.

The overall classification index ranged from 78% to 84% for accuracy and from 70% to 77% 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 intervals used to compute the classification probability to classify students into L1 [$-\infty$, L2 cut] or L4 [L4 cut, ∞] being wider than the intervals used in L2 [L2 cut, L3 cut] and L3 [L3 cut, L4 cut]. The misclassification probability tends to be higher for narrower intervals.

The accuracy of classifications is higher than the consistency of classifications at all achievement levels. The consistency of classification rates can be lower because the consistency is based on two tests with measurement errors, but the accuracy is based on one test with a measurement error and the true score. The classification indexes by subgroup are provided in Appendix C.

Table 37. Classification Accuracy and Consistency by Achievement Level

Grade	Achievement Level	ELA/Lit		Mathematics	
		% Accuracy	% Consistency	% Accuracy	% Consistency
3	Overall	79	71	83	76
	L1	89	83	90	85
	L2	70	60	73	63
	L3	68	57	79	71
	L4	88	83	90	85
4	Overall	78	70	84	77
	L1	89	83	89	83
	L2	63	51	80	73
	L3	65	55	79	71
	L4	88	81	90	84
5	Overall	79	71	83	76
	L1	89	83	90	85
	L2	67	55	78	69
	L3	74	66	71	61
	L4	86	79	91	86
6	Overall	80	72	83	76
	L1	89	83	92	87
	L2	72	62	77	69
	L3	76	69	72	61
	L4	85	76	89	84
7	Overall	80	72	83	76
	L1	90	84	91	86
	L2	70	59	75	67
	L3	78	71	74	65
	L4	85	76	90	85
8	Overall	80	73	82	75
	L1	89	82	90	86
	L2	72	62	71	62
	L3	79	72	71	60
	L4	83	75	90	85

5.4 RELIABILITY FOR SUBGROUPS

The reliability of test scores is also computed by subgroup. Tables 38 and 39 present the marginal reliability coefficients by the subgroup. The reliability coefficients are similar across subgroups but somewhat lower for English language learners (ELL) and special education subgroups, a large percentage of whom received Level 1 with large SEMs as shown in Tables 19-22 .

Table 38. ELA/Lit Marginal Reliability Coefficients for Overall and by Subgroup

Subgroup	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
All Students	0.92	0.92	0.92	0.92	0.92	0.92
Female	0.92	0.91	0.91	0.92	0.91	0.92
Male	0.92	0.92	0.92	0.92	0.92	0.92
African American	0.90	0.90	0.91	0.91	0.91	0.91
AmerIndian/Alaskan	0.89	0.91	0.90	0.93	0.93	0.92
Asian	0.91	0.90	0.89	0.90	0.90	0.91
Hispanic	0.90	0.90	0.91	0.91	0.91	0.91
Pacific Islander	0.94	0.89	0.93	0.88	0.95	0.91
White	0.91	0.90	0.91	0.91	0.91	0.91
Multi-Racial	0.91	0.91	0.91	0.91	0.91	0.92
ELL	0.89	0.89	0.87	0.86	0.86	0.85
Special Education	0.86	0.86	0.88	0.86	0.86	0.85
CD 504	0.90	0.90	0.90	0.90	0.91	0.91
Title I	0.91	0.89	0.90	0.91	0.91	0.91

Table 39. Mathematics Marginal Reliability Coefficients for Overall and by Subgroup

Subgroup	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
All Students	0.95	0.95	0.94	0.94	0.93	0.93
Female	0.95	0.94	0.94	0.94	0.93	0.93
Male	0.95	0.95	0.95	0.94	0.94	0.94
African American	0.94	0.93	0.92	0.91	0.90	0.91
AmerIndian/Alaskan	0.96	0.94	0.93	0.94	0.93	0.93
Asian	0.94	0.93	0.94	0.94	0.95	0.95
Hispanic	0.94	0.93	0.93	0.92	0.92	0.91
Pacific Islander	0.97	0.95	0.96	0.94	0.96	0.89
White	0.94	0.94	0.94	0.94	0.94	0.94
Multi-Racial	0.95	0.94	0.94	0.93	0.94	0.92
ELL	0.93	0.93	0.90	0.87	0.84	0.84
Special Education	0.92	0.91	0.87	0.84	0.81	0.81
CD 504	0.94	0.94	0.93	0.93	0.93	0.92
Title I	0.94	0.93	0.93	0.93	0.93	0.93

5.5 RELIABILITY FOR CLAIM SCORES

The descriptive statistics, marginal reliability coefficients, and the 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. Because the precision of scores in claims is insufficient to report given a small number of items per claim, three performance categories, taking into account the SEM of the claim score, are reported as (1) Below Standard, (2) At/Near Standard, or (3) Above Standard. Tables 40 and 41 present the marginal reliability coefficients and descriptive statistics by claim in ELA/lit and mathematics, respectively.

Table 40. ELA/Lit Marginal Reliability Coefficients for Claim Scores

Grade	Claim	Number of Items Specified in Test Blueprint		Marginal Reliability	Scale Score Mean	Scale Score SD	Average CSEM
		Min.	Max.				
3	Claim 1: Reading	14	16	0.76	2433.49	101.14	49.77
	Claim 2: Writing	7	7	0.74	2418.47	113.90	58.59
	Claim 3: Listening	8	9	0.61	2435.05	124.44	77.95
	Claim 4: Research	9	9	0.73	2423.05	116.98	61.21
4	Claim 1: Reading	14	16	0.76	2473.07	105.94	51.78
	Claim 2: Writing	7	7	0.72	2473.06	122.45	64.49
	Claim 3: Listening	8	9	0.62	2485.13	136.45	83.61
	Claim 4: Research	9	9	0.72	2470.20	125.19	65.68
5	Claim 1: Reading	14	16	0.75	2515.62	112.32	56.17
	Claim 2: Writing	7	7	0.72	2516.20	120.99	64.05
	Claim 3: Listening	8	9	0.63	2506.74	134.93	82.06
	Claim 4: Research	9	9	0.76	2515.17	118.57	57.49
6	Claim 1: Reading	14	16	0.77	2517.26	117.50	56.30
	Claim 2: Writing	7	7	0.74	2524.95	114.07	57.91
	Claim 3: Listening	8	9	0.64	2547.80	148.52	89.58
	Claim 4: Research	9	9	0.72	2533.87	124.96	66.18
7	Claim 1: Reading	14	16	0.79	2551.74	118.69	55.00
	Claim 2: Writing	7	7	0.74	2554.50	127.31	64.42
	Claim 3: Listening	8	9	0.59	2555.24	140.82	89.85
	Claim 4: Research	9	9	0.73	2557.19	134.83	69.88
8	Claim 1: Reading	16	16	0.76	2558.92	122.87	59.93
	Claim 2: Writing	7	7	0.74	2566.40	127.15	64.85
	Claim 3: Listening	8	9	0.62	2574.46	138.96	85.93
	Claim 4: Research	9	9	0.74	2562.83	130.86	67.20

Table 41. Mathematics Marginal Reliability Coefficients for Claim Scores

Grade	Claims	Number of Items Specified in Test Blueprint		Marginal Reliability	Scale Score Mean	Scale Score SD	Average CSEM
		Min.	Max.				
3	Claim 1	20	20	0.91	2443.37	95.57	29.15
	Claims 2 and 4	8	11	0.72	2433.45	94.98	50.09
	Claim 3	9	11	0.77	2436.12	99.66	47.66
4	Claim 1	20	20	0.90	2486.96	93.51	28.97
	Claims 2 and 4	8	10	0.77	2477.25	96.43	46.22
	Claim 3	9	10	0.76	2480.67	99.89	49.10
5	Claim 1	20	20	0.90	2513.81	101.49	32.30
	Claims 2 and 4	8	10	0.70	2501.97	103.20	56.36
	Claim 3	9	10	0.74	2505.26	116.41	59.83
6	Claim 1	19	19	0.88	2516.90	115.77	39.50
	Claims 2 and 4	9	10	0.72	2503.00	120.76	64.14
	Claim 3	9	11	0.73	2510.30	124.84	64.58
7	Claim 1	20	20	0.89	2535.71	120.24	39.86
	Claims 2 and 4	9	10	0.67	2525.01	131.07	75.01
	Claim 3	9	10	0.68	2530.12	138.95	78.58
8	Claim 1	20	20	0.88	2547.28	127.87	43.58
	Claims 2 and 4	8	10	0.70	2533.56	139.05	75.80
	Claim 3	9	10	0.70	2539.28	141.19	77.77

Legend: Claim 1: Concepts and Procedures; Claims 2 and 4: Problem Solving and 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 handscoring procedure.

6.1 ESTIMATING STUDENT ABILITY USING MAXIMUM LIKELIHOOD ESTIMATION

The Smarter Balanced tests are scored using maximum likelihood estimation (MLE). The likelihood function for generating the MLEs is based on a mixture of 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 $p_{ij}(z_{ij} \mid \theta_j, a_i, b_{i,1}, \dots, b_{i,m_i})$ takes either the form of a two-parameter logistic (2PL) model for items with one point or the form based on the generalized partial credit model (GPCM) for items with two or more points.

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

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

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:

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

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

For the CAT component, the algorithm allows previously answered items to be changed; however, it does not allow items to be skipped. Item selection requires iteratively updating 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 number of items a student answers correctly and the difficulty of the items presented are used to statistically transform theta scores to scale scores so that scores from different sets of items can be meaningfully compared. The scale scores represent a linear transformation of the ability estimates (theta scores) using the formula, $SS = a * \theta + b$. The scaling constants a and b are provided by the Smarter Balanced Assessment Consortium. Table 42 presents the scaling constants for each subject for the theta-to-scale score linear transformation. Scale scores are rounded to an integer.

Table 42. Vertical Scaling Constants on the Reporting Metric

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

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

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

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

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

Table 43. Cut Scores in Scale Scores

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

6.3 LOWEST/HIGHEST OBTAINABLE SCORES (LOSS/HOSS)

In the 2014–2015 administration, Delaware applied the Smarter Balanced LOSS/HOSS to truncate extreme student ability estimates in both theta and scale score metrics. Starting with the 2015–2016 administration, Delaware removed the LOSS and HOSS in the summative tests while kept 2014-15 LOSS/HOSS in the interim test.

6.4 SCORING ALL CORRECT AND ALL INCORRECT CASES

In the 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 and the lowest obtainable scores were assigned in the 2014–2015 administration. Since the 2015–2016 administration, all incorrect and correct cases were scored by either adding 0.5 to or subtracting 0.5 from an item score with the smallest item discrimination parameter among the administered operational items (CAT and PT) for a student.

6.5 RULES FOR CALCULATING STRENGTHS AND WEAKNESSES FOR CLAIM SCORES

In ELA/lit, claim scores are computed for each claim. In mathematics, claim scores are computed for claim 1, claims 2 and 4 combined, and claim 3. For each claim, three performance categories relative strengths and weaknesses are produced.

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

For summative tests, the specific rules are as follows:

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

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

6.6 TARGET SCORES

The target-level reports are not appropriate to produce for a fixed-form test because the number of items included per target (i.e., benchmark) is too small to produce a reliable score at the target level. A typical fixed-form test includes only one or two items per target. Even when aggregated, these data narrowly reflect the target 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 group level, such as class, school, and district level. With an adequate item pool, a class of 20 students might respond to 10 or 15 different items measuring any given targets. Target scores are computed for attempted tests based on the responded items. Target scores are computed for each of the four claims in ELA/lit and for only claim 1 in mathematics. A target performance provides information on strengths and weaknesses on the target for a group of students, not for individual students.

For Delaware, target scores are computed relative to the proficiency standard (level 3 cut).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

6.7 HANDSCORING

AIR provides the automated electronic scoring for select-response items and Measurement Incorporated (MI) provides all handscoring for the Smarter Balanced summative assessments on constructed-response items, such as short-answer (SA) items and full-write items in ELA/lit and SA items in mathematics. The general procedures for hand-scoring are specified by Smarter Balanced Assessmewnt Consortium (SBAC). Outlined in the following sections provides details about the hand-scoring process by MI for the 2018-2019 Delaware administration.

6.7.1 Rater Selection

MI maintains a large pool of raters at each scoring center, as well as distributive raters who work remotely. MI's recruiting team first recruits qualified raters who have experience scoring the Smarter Balanced assessment. Rater accuracy parameters are used to focus recruitment efforts for experienced Smarter Balanced raters in order to recruit the most objectively accurate raters. Once recruited, experienced raters are assigned to the content area and grade bands in which they are most experienced. These experienced, demonstrably accurate raters make up the majority of the total rater pool. To supplement this core pool, MI contacts other raters in their database who have experience successfully scoring other large-scale assessments. These raters are assigned to the grade level, subject area, and item type for which they are most qualified based on their performance on similar projects. Returning staff are selected based on experience and performance, as well as attendance, punctuality, and cooperation with work procedures and MI policies. MI maintains evaluations and performance data for all staff who work on each scoring project in order to determine employment eligibility for future projects. Finally, MI targets recruitment of new raters for site-based and remote scoring as needed, in order to continue to identify talent across the country that will best fulfill the handscoring requirements. For new raters, MI's recruiting team reviews applications, including prospective raters' resumes, references, proof of degree, and recognition of rater requirements, before offering employment.

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

MI is an equal opportunity employer that actively recruits minority staff. Historically, MI's temporary staff on major projects averages about 51% female, 49% male, 76% Caucasian, and 24% minority.

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

6.7.2 Rater Training

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

Once hired, raters are placed into a scoring group that corresponds to the subject/grade that they are deemed best suited to score (based on work history, results of the placement assessments, and performance on past scoring projects). Raters are trained on a specific item type (i.e., brief writes, reading, research, full-writes, or mathematics). Within each group, raters are divided into teams consisting of one team leader and 10–15 raters. Each team leader and rater is assigned a unique number for easy identification of their scoring work throughout the scoring session. The number of items an individual rater scores is minimized so that the rater becomes highly experienced in scoring responses to a given set of items.

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

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

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

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

Like team leaders, raters must demonstrate their ability to score accurately by attaining the qualifying agreement percentage established by Smarter Balanced before they may score actual student responses.

Any raters unable to meet the qualifying standards are not permitted to score that item. Raters who reach the qualifying standard on some items but not others will only score the items on which they have successfully qualified. All raters understand this stipulation when they are hired.

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

Training design varies slightly depending on Smarter Balanced item type:

- **Full-Writes.** Raters train and qualify on baseline sets for each grade and writing purpose (e.g., Grade 3 Narrative, Grade 6 Argumentative, etc.), then take qualifying sets for each item in that grade and purpose.
- **Brief Writes, Reading, and Research.** Raters train and qualify on a baseline set within a specific grade band and target.
- **Mathematics.** Raters train on baseline items, which qualify the raters for that item as well as any items associated with it; for items with no associated items, training is for the specific item.

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

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

MI also implements a series of automated score verifications to ensure the accuracy of scores. For example, MI conducts a blank check that resets scores when a condition code of "blank" is assigned to a response that has one or more characters in the response string (e.g., a response made up of spaces or tabs). In this case, the score is recorded only after three independent raters have assigned a condition code of "blank" to a response that appears blank but includes characters in the response string. A similar check is run when a score or condition code other than "blank" is assigned to a response that includes 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 provides an additional score verification. In addition to automatically resetting and rescoring these responses, the rater information is captured in a report and reviewed by scoring directors, as one of many tools used to determine re-training needs.

6.7.3 Rater Statistics

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

In addition to extensive experience in the preparation of training materials and employing management and staff with unparalleled expertise in the field of handscored educational assessments, MI constantly monitors the quality of each rater's work throughout every project. Rater status reports are used to monitor raters' scoring habits during the Smarter Balanced handscoring project.

MI has developed and operates a comprehensive system for collecting and analyzing scoring data. After the raters' scores are submitted into the VSC handscoring system, the data are uploaded into the scoring data report servers located at MI's corporate headquarters in Durham, North Carolina.

More than 20 reports are available and can be customized to meet the information needs of the client and MI's scoring department. These reports provide the following data:

- Rater ID and team
- Number of responses scored
- Number of responses assigned each score point (1–4 or other)
- Percentage of responses scored that day in exact agreement with a second rater
- Percentage of responses scored that day within one point of agreement with a second rater
- Number and percentage of responses receiving adjacent scores at each line (0/1, 1/2, 2/3, etc.)
- Number and percentage of responses receiving nonadjacent scores at each line
- Number of correctly assigned scores on the validity responses

Updated real-time reports are available that show both daily and cumulative (project-to-date) data. These reports are available for access by the handscoring project monitors at each MI scoring center via a secure website, and the handscoring project monitors provide updated reports to the scoring directors several times per day. MI further used dynamic threshold reports, which, based on inputted criteria, immediately identify potential scoring performance issues. These reports allow scoring leadership to pinpoint areas of concern and to take corrective action with great efficiency. MI scoring directors are experienced in examining these reports and using the information to determine a need for re-training of individual raters or the group as a whole. If a rater is consistently scoring high or low, this can be easily determined along with the specific score points with which they may be having difficulty. The scoring directors share such information with the team leaders and direct all re-training efforts.

6.7.4 Rater Monitoring and Re-Training

Team leaders spot-check (i.e., read behind) each rater's scoring to ensure that he or she is on target and conduct one-on-one re-training sessions addressing any problems found. At the beginning of the project, team leaders read behind every rater every day; they become more selective about the frequency and number

of read-behinds as raters become more proficient at scoring. The daily rater reliability reports and validity/calibration results are used to identify raters who need more frequent monitoring.

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

Each student response is scored holistically by a trained and qualified rater using the scoring criteria developed and approved by Smarter Balanced, with a second read conducted on 15% of responses for each item for reliability purposes. Responses are randomly selected for second reads and scored by raters who are not aware of the score assigned by the first rater or even that the response has been read before. MI's QA/reliability procedures allow the handscoring staff to identify struggling raters very early and begin re-training at once. While re-training these raters, MI also monitors their scoring intensively to ensure that all responses are scored accurately. In fact, MI's monitoring is also used as a re-training method. MI shows raters responses that the raters have scored incorrectly, explains the correct scores, and has the raters change the scores.

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

6.7.5 Validity Checks

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

MI staff and all clients have access to real-time validity reports that include the response identification number, the scores assigned by the raters, and the "true" scores. A daily and project-to-date summary of the percentages of correct scores and low/high considerations at each score point is also provided. Re-training may be conducted with the raters using the validity data as a guide for how to focus the re-training. Validity results are not used in isolation but as one piece of evidence along with the second read and read-behind agreement to make decisions about re-training and dismissing raters.

MI has amassed a large, longitudinal dataset of rater performance data from years of Smarter Balanced handscoring. In spring 2019 we launched an enhanced accuracy monitoring system drawing on these data. This system used validity responses, calibrated to fit a unidimensional item response theory (IRT) model for each content area/item type. Calibrating validity responses allows us to prioritize them (using correlations and fit statistics) so that those responses that provide the greatest information about rater accuracy are distributed to raters first. MI runs nightly analyses to evaluate performance nightly during scoring. Empirically determined cutpoints are used to classify raters into performance tiers based on recent validity and inter-rater reliability (IRR). A rater with unacceptable performance initially receives feedback

and additional monitoring in the form of increased read-behinds. If performance does not improve quickly, the rater is assigned an assessment composed of validity responses, the results of which determine whether the rater may continue to score.

6.7.6 Rater Dismissal

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

6.7.7 Rater Agreement

The inter-rater reliability (IRR) is computed based on scorable responses (numeric scores) by two independent raters only, excluding non-scorable responses (e.g., off-topic, off-purpose, or foreign-language responses) that are determined by scoring leadershipin Delaware.

Student essay on the full-writing is scored in three dimensions: convention (0–2 rubric), evidence/elaboration (1–4 rubric), and organization/purpose (1–4 rubric). The short answer (SA) items in ELA/Lit are scored with the 0–2 rubric. The mathematics SA items are scored using 0–1, 0–2, or 0–3 rubrics.

Tables 44–46 summarize the IRR based on a sample size is greater than 50, including the average percentage of exact agreement, minimum and maximum percentages of exact agreement, combined percentage of exact and adjacent agreement, and quadratic weighted Kappa (QWK).

Table 44. ELA/Lit Rater Agreements for Short-Answer Items

Grade	# of Items	% Exact			% (Exact+ Adjacent)	QWK
		Average	Min.	Max.		
3	12	80	74	89	100	0.80
4	15	79	68	86	100	0.79
5	16	73	61	84	100	0.75
6	25	74	60	88	100	0.71
7	36	71	55	89	100	0.69
8	30	74	61	88	100	0.73

Note. Adjacent scores are two scores assigned by two raters with one score-point difference of each other.

Table 45. ELA/Lit Rater Agreements for Full-Write Items

Grade	Dimensions	# of Items	% Exact			% (Exact+ Adjacent)	QWK
			Average	Min.	Max.		
3	Conventions	12	71	63	78	99	0.63
	Evid/Elab	12	70	62	81	99	0.67
	Org/Purp	12	70	59	81	99	0.68
4	Conventions	16	64	48	77	99	0.60
	Evid/Elab	16	66	59	72	99	0.64
	Org/Purp	16	66	61	75	100	0.67
5	Conventions	24	68	58	80	100	0.51
	Evid/Elab	24	63	51	75	99	0.65
	Org/Purp	24	64	54	74	99	0.66
6	Conventions	19	72	61	83	98	0.57
	Evid/Elab	19	65	56	75	99	0.66
	Org/Purp	19	65	54	75	99	0.67
7	Conventions	22	71	51	82	99	0.56
	Evid/Elab	22	68	43	81	99	0.70
	Org/Purp	22	68	45	82	99	0.70
8	Conventions	22	75	62	89	99	0.60
	Evid/Elab	22	70	55	79	99	0.73
	Org/Purp	22	70	54	81	99	0.74

Legend: Evid/Elab = Evidence/Elaboration, and Org/Purp = Organization/Purpose

Table 46. Mathematics Rater Agreements

Grade	Score Points	# of Items	% Exact			% (Exact+ Adjacent)	QWK
			Average	Min.	Max.		
3	1	10	92	86	97	100	0.83
3	2	28	90	77	99	100	0.91
3	3	5	89	85	94	100	0.95
4	1	11	85	78	96	100	0.64
4	2	40	89	75	99	100	0.89
4	3	4	82	80	83	100	0.91
5	1	5	92	88	97	100	0.68
5	2	50	88	74	98	100	0.87
5	3	7	84	80	90	100	0.88
6	1	8	99	98	100	100	0.89
6	2	41	91	80	99	100	0.89
7	1	8	96	92	99	100	0.72
7	2	25	90	81	95	100	0.87
7	3	1	76	76	76	100	0.83
8	1	15	93	80	100	100	0.81
8	2	26	90	84	99	100	0.88

7. REPORTING AND INTERPRETING SCORES

The Online Reporting System (ORS) generates a set of online score reports that includes the information describing student performance for students, parents, educators, and other stakeholders. The online score reports are produced immediately after students complete a test and the test is handscored. Because the score reports on student performance are updated each time that the students' completed tests are handscored, authorized users (e.g., school principals, teachers) can have quickly available information on students' performance on the tests and use the information to improve student learning. In addition to the individual student score report, the ORS also produces aggregate score reports by class, school, district, and state. The timely accessibility of aggregate score reports could help users monitor student 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 describes the types of scores reported in the ORS and the ways to interpret and use these scores.

7.1 ONLINE REPORTING SYSTEM FOR STUDENTS AND EDUCATORS

7.1.1 Types of Online Score Reports

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

Once authorized users log in to the ORS and select "Score Reports," the online score reports are presented hierarchically. The ORS starts by presenting summaries on student performance by subject and grade at a selected aggregate level. To view student performance for a specific aggregate unit, users can select the specific aggregate unit from a drop-down list of aggregate units (e.g., schools within a district or teachers within a school) to select. For more detailed student assessment results for a school, a teacher, or a roster, users can select the subject and grade on the online score reports. Additionally, when authorized state-level users log in to the ORS and select "State at a Glance," the ORS generates a summary of student performance data for a test across the entire state.

Generally, the ORS provides two categories of online score reports: (1) aggregate score reports, and (2) student score reports. Table 47 summarizes the types of online score reports available at the aggregate level and the individual student level. Detailed information about the online score reports and instructions on how to navigate the online score reporting system can be found in the *Online Reporting System User Guide*, located via a help button on the ORS.

Table 47. 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 with Level 3 or 4 (for overall students and by subgroup) • Average scale score and standard error of average scale score (for overall students and by subgroup) • Percentage of students at each achievement level on the overall test and by claims (for overall students and by subgroup) • Performance category level in each target (for overall students) • Participation rate (for overall students)[1] • On-demand student roster report
Student	• Total scale score and SEM • Achievement level on overall and claim scores with achievement-level descriptors • Average scale scores and standard errors of average scale scores for student's school, district, and state • Student growth in scale score and achievement level over time • Writing performance descriptors and scores by dimensions

1 Participation rate reports are provided at the state, district, and school level.

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 48 presents the types of subgroups and subgroup categories provided in ORS.

Table 48. Types of Subgroups

Subgroup	Subgroup Category
Gender	Male Female
CD504	CD504 Not CD504
ELL	ELL Not ELL
Special Education	Special Education Not Special Education
Title I	Title I Not Title I
Ethnicity	African American American Indian/Alaskan Native Asian Hispanic Native Hawaiian/Pacific Islander White Multi-Racial

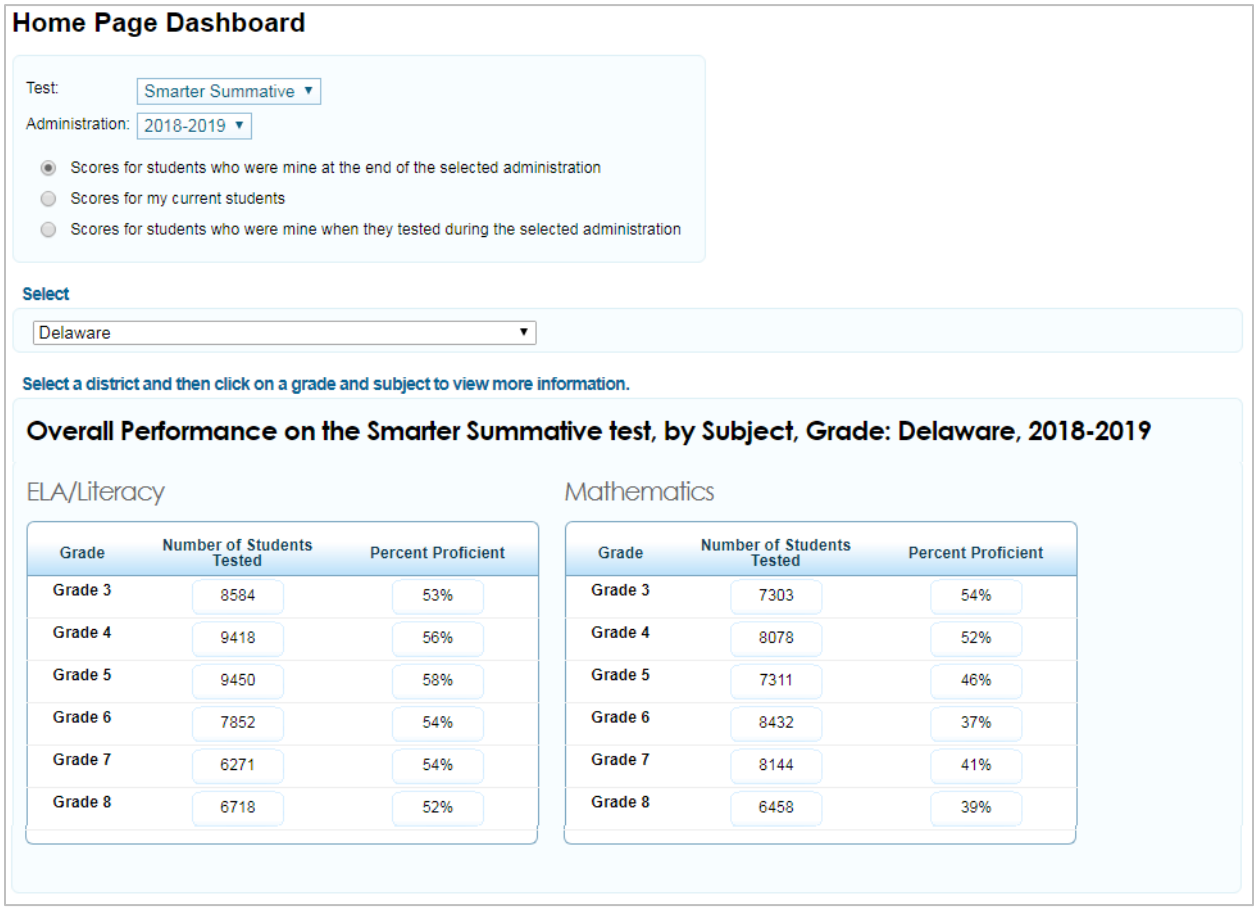
7.1.2 Online Reporting System

7.1.2.1 Home Page

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

The home page summarizes student performance, including: (1) number of students tested, and (2) percentage of students at Level 3 or above. Exhibits 1 and 2 present a sample of home pages at the state level and the district level, respectively.

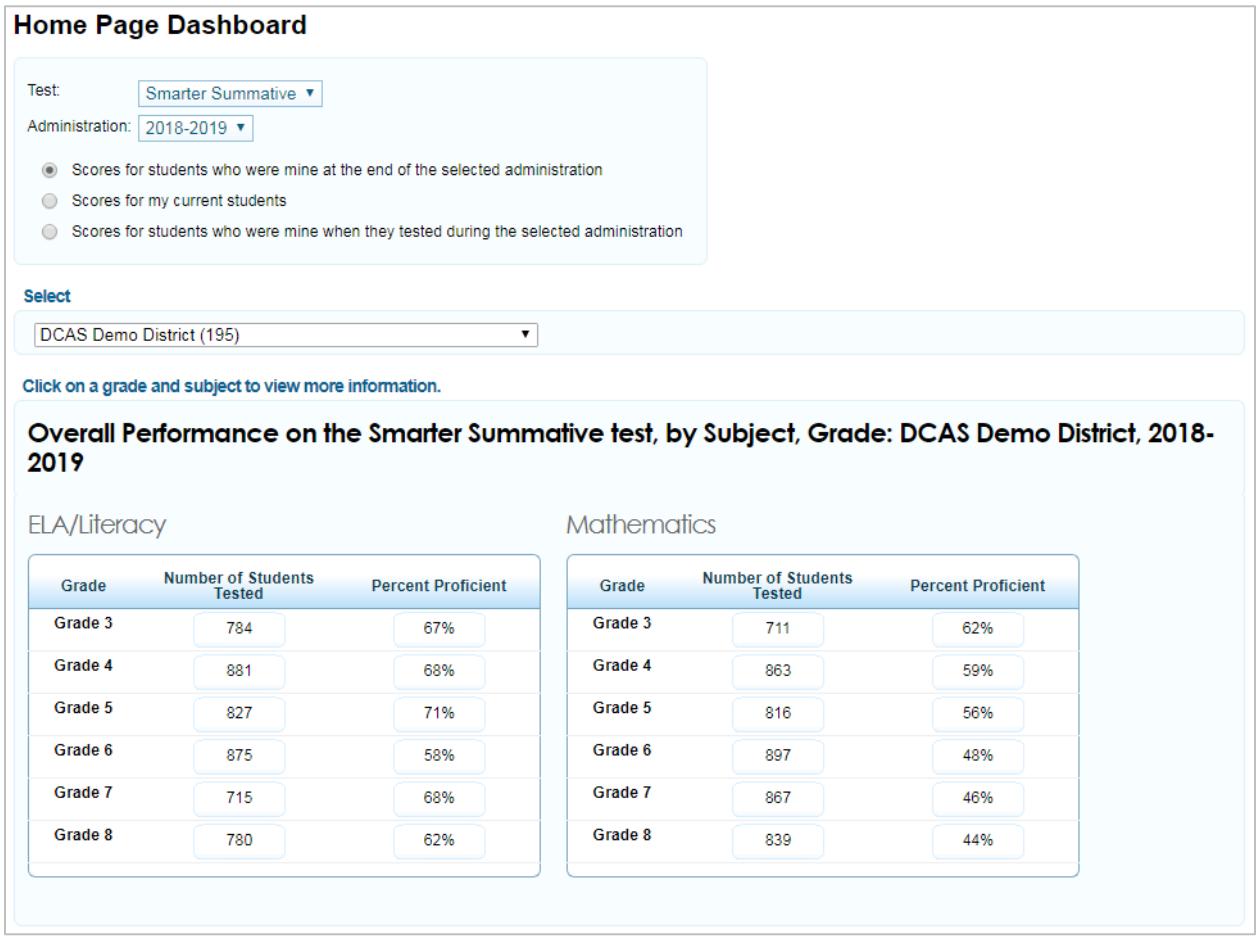
Exhibit 1. Home Page: State Level



Grade	Number of Students Tested	Percent Proficient
Grade 3	8584	53%
Grade 4	9418	56%
Grade 5	9450	58%
Grade 6	7852	54%
Grade 7	6271	54%
Grade 8	6718	52%

Grade	Number of Students Tested	Percent Proficient
Grade 3	7303	54%
Grade 4	8078	52%
Grade 5	7311	46%
Grade 6	8432	37%
Grade 7	8144	41%
Grade 8	6458	39%

Exhibit 2. Home Page: District Level

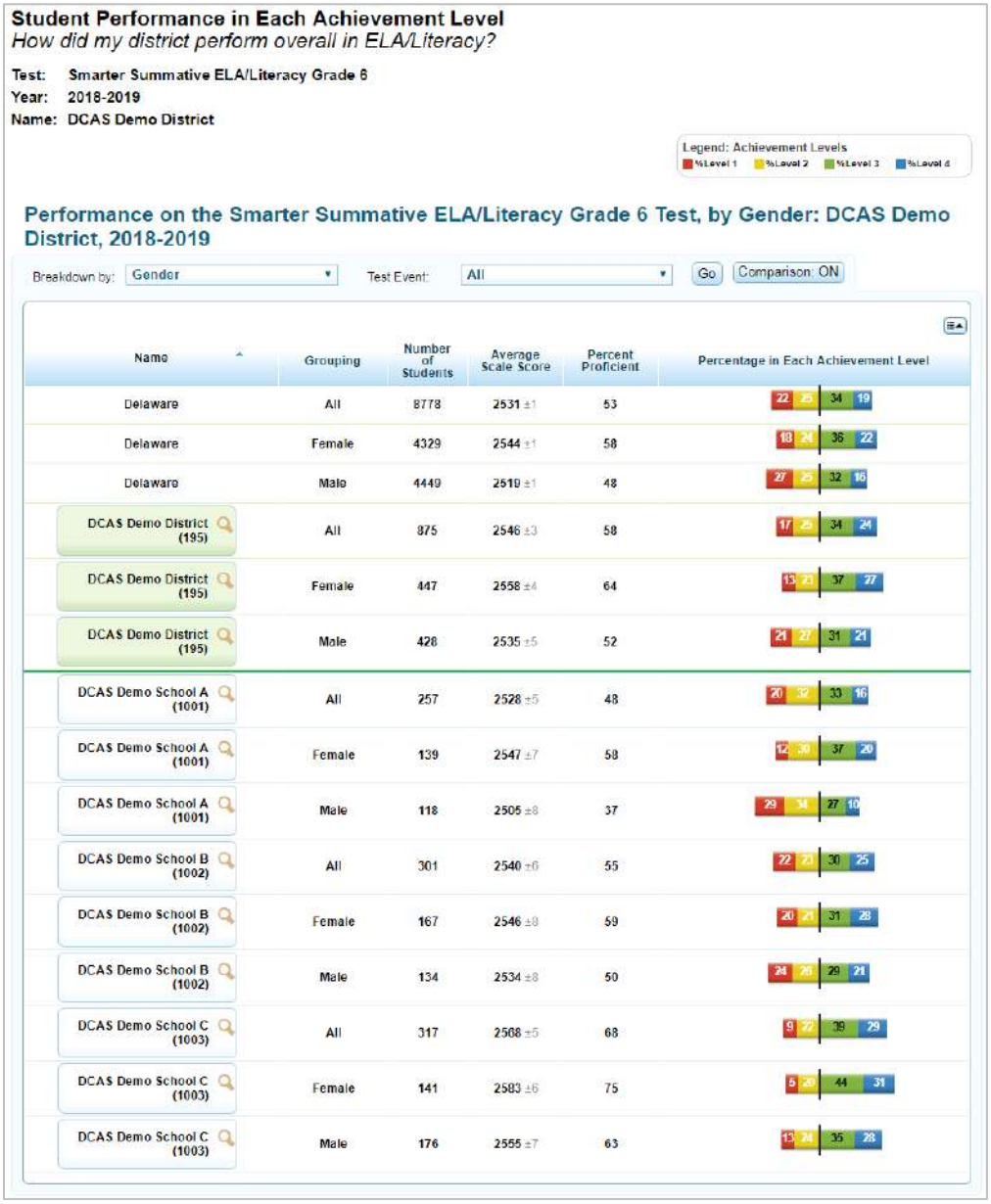


7.1.2.2 Subject Detail Page

More detailed summaries of student performance for each grade in a subject area for a selected aggregate level are presented when users select a grade within a subject on the home page. On each aggregate report, the summary report presents the summary results for the selected aggregate unit as well as the summary results for the state and the aggregate unit above the selected aggregate. For example, if a school is selected on the subject detail page, the summary results of the state, the district, and the school are provided above the school summary results, as well, so that school performance can be compared with the above aggregate levels.

The subject detail page provides the aggregate summaries on a specific-subject area, including (1) number of students, (2) average scale score and standard error of the average scale score, (3) percentage proficient, and (4) percentage of students in each achievement level. The summaries are also presented for overall students and by subgroup. Exhibit 3 presents an example of subject detail pages for ELA/lit at the district level when a user selects a subgroup of gender.

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



Student Performance in Each Achievement Level
How did my district perform overall in ELA/Literacy?

Test: Smarter Summative ELA/Literacy Grade 6
Year: 2018-2019
Name: DCAS Demo District

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

Performance on the Smarter Summative ELA/Literacy Grade 6 Test, by Gender: DCAS Demo District, 2018-2019

Breakdown by: Gender | Test Event: All | Go | Comparison: ON

Name	Grouping	Number of Students	Average Scale Score	Percent Proficient	Percentage in Each Achievement Level
Delaware	All	8778	2531 ±1	53	22 25 34 19
Delaware	Female	4329	2544 ±1	58	18 24 36 22
Delaware	Male	4449	2519 ±1	48	27 25 32 16
DCAS Demo District (195)	All	875	2546 ±3	58	17 25 34 24
DCAS Demo District (195)	Female	447	2558 ±4	64	13 23 37 27
DCAS Demo District (195)	Male	428	2535 ±5	52	21 27 31 21
DCAS Demo School A (1001)	All	257	2528 ±5	48	20 32 33 16
DCAS Demo School A (1001)	Female	139	2547 ±7	58	12 30 37 20
DCAS Demo School A (1001)	Male	118	2505 ±8	37	29 34 27 10
DCAS Demo School B (1002)	All	301	2540 ±6	55	22 23 30 25
DCAS Demo School B (1002)	Female	167	2546 ±8	59	20 21 31 28
DCAS Demo School B (1002)	Male	134	2534 ±8	50	24 26 29 21
DCAS Demo School C (1003)	All	317	2568 ±5	68	9 23 39 29
DCAS Demo School C (1003)	Female	141	2583 ±6	75	5 20 44 31
DCAS Demo School C (1003)	Male	176	2555 ±7	63	13 24 35 28

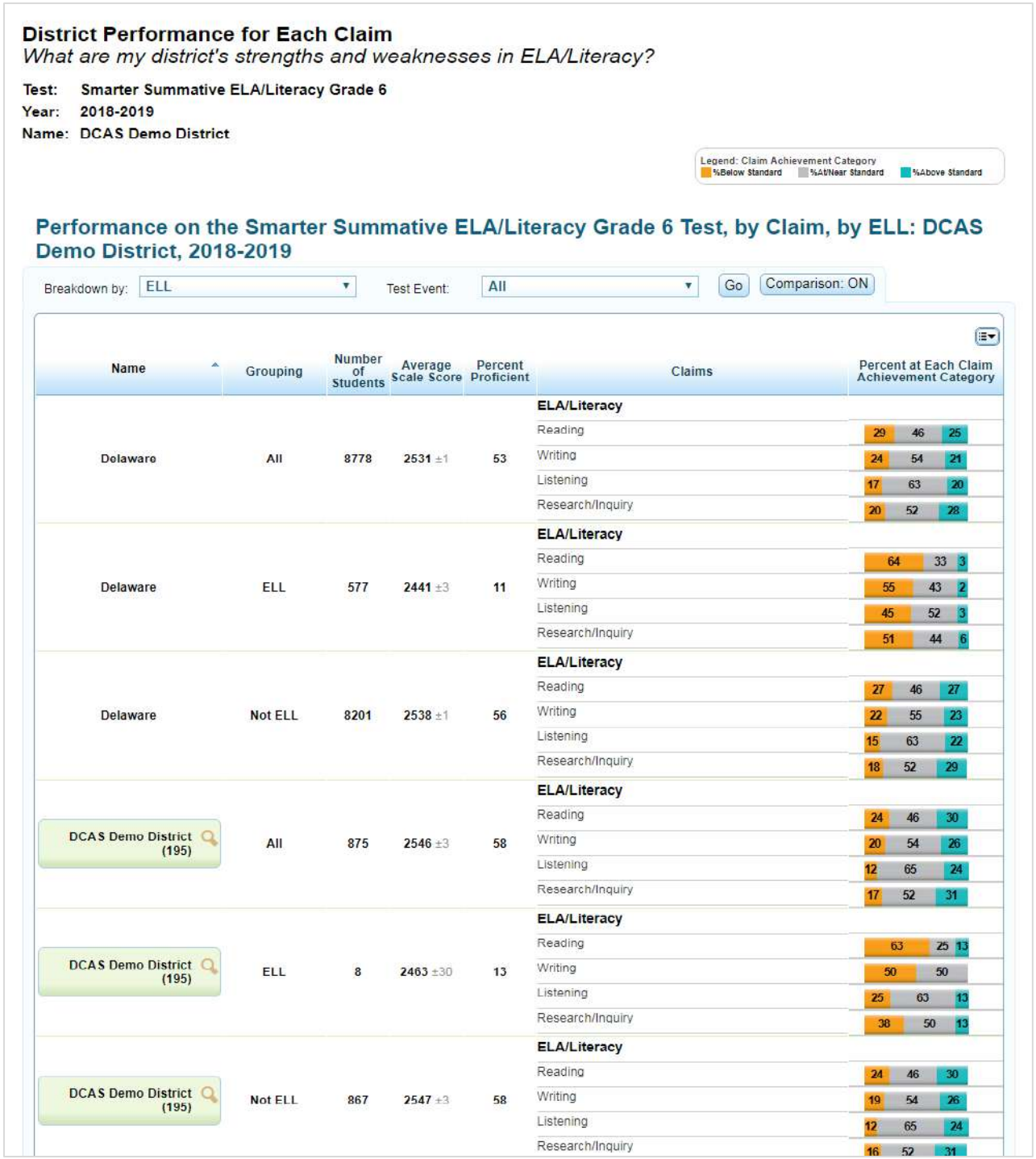
7.1.2.3 Claim Detail Page

The claim detail page provides the aggregate summaries on student performance in each claim for a particular grade and subject. The aggregate summaries on the claim detail page include: (1) number of students, (2) average scale score and standard error of the average scale score, (3) percentage proficient, and (4) percentage of students in each claim performance category.

As with the subject detail page, the summary report presents the summary results for the selected aggregate unit, as well as the summary results for the state and aggregate unit above the selected aggregate. Also, the summaries on claim-level performance can be presented for overall students and by subgroup. Exhibit 4

presents an example of a claim detail page for ELA/lit at a district level when users select a subgroup of ELL.

Exhibit 4. Claim Detail Page for ELA/Lit by ELL: District Level



District Performance for Each Claim

What are my district's strengths and weaknesses in ELA/Literacy?

Test: Smarter Summative ELA/Literacy Grade 6
Year: 2018-2019
Name: DCAS Demo District

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

Performance on the Smarter Summative ELA/Literacy Grade 6 Test, by Claim, by ELL: DCAS Demo District, 2018-2019

Breakdown by: ELL | Test Event: All | Go | Comparison: ON

Name	Grouping	Number of Students	Average Scale Score	Percent Proficient	Claims	Percent at Each Claim Achievement Category
Delaware	All	8778	2531 ±1	53	**ELA/Literacy**	
					Reading	29 46 25
					Writing	24 54 21
					Listening	17 63 20
					Research/Inquiry	20 52 28
Delaware	ELL	577	2441 ±3	11	**ELA/Literacy**	
					Reading	64 33 3
					Writing	55 43 2
					Listening	45 52 3
					Research/Inquiry	51 44 6
Delaware	Not ELL	8201	2538 ±1	56	**ELA/Literacy**	
					Reading	27 46 27
					Writing	22 55 23
					Listening	15 63 22
					Research/Inquiry	18 52 29
DCAS Demo District (195)	All	875	2546 ±3	58	**ELA/Literacy**	
					Reading	24 46 30
					Writing	20 54 26
					Listening	12 65 24
					Research/Inquiry	17 52 31
DCAS Demo District (195)	ELL	8	2463 ±30	13	**ELA/Literacy**	
					Reading	63 25 13
					Writing	50 50
					Listening	25 63 13
					Research/Inquiry	38 50 13
DCAS Demo District (195)	Not ELL	867	2547 ±3	58	**ELA/Literacy**	
					Reading	24 46 30
					Writing	19 54 26
					Listening	12 65 24
					Research/Inquiry	16 52 31

7.1.2.4 Target Detail Page

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

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

Institution Performance on Each Target for the ELA/Literacy Test
What are my institution's strengths and weaknesses in the ELA/Literacy Target?

Test: Smarter Summative ELA/Literacy Grade 6
Year: 2018-2019
Name: DCAS Demo School C

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

Average Scale Scores on the Smarter Summative ELA/Literacy Grade 6 Test: DCAS Demo School C and Comparison Groups, 2018-2019

Name	Average Scale Score
Delaware	2531 ±1
DCAS Demo District (195)	2546 ±3
DCAS Demo School C (1003)	2568 ±5

Performance on the Smarter Summative ELA/Literacy Grade 6 Test, by Target: DCAS Demo School C, 2018-2019

Target	Performance Relative to Proficiency
Reading	
Literary Text	
Target 1 (Literary Text) KEY DETAILS: Given an inference or conclusion, use explicit details and implicit information from the text to support the inference or conclusion provided.	+
Target 2 (Literary Text) CENTRAL IDEAS: Determine a theme or central idea from details in the text, or provide a summary distinct from personal opinions or judgment.	+
Target 3 (Literary Text) WORD MEANINGS: Determine intended or precise meanings of words, including academic/tier 2 words, domain-specific (tier 3) words, and words with multiple meanings, based on context, word relationships (e.g., connotations, denotations), word structure (e.g., common Greek or Latin roots, affixes), or use of reference materials (e.g., dictionary) with primary focus on determining meaning based on context and the academic (tier 2) vocabulary common to complex texts in all disciplines.	=
Target 4 (Literary Text) REASONING & EVIDENCE: Make an inference or draw a conclusion about a text OR make inferences or draw conclusions in order to compare texts (e.g., character development, plot, point of view, themes, topics) and use supporting evidence as justification/explanation.	=
Target 5 (Literary Text) ANALYSIS WITHIN OR ACROSS TEXTS: Describe and explain relationships among literary elements (e.g., plot, character, resolution) within or across texts or explain how the author develops the narrator or speakers' point of view within or across texts.	+
Target 6 (Literary Text) TEXT STRUCTURES & FEATURES: Analyze text structures and the impact of those choices on meaning or presentation.	−
Target 7 (Literary Text) LANGUAGE USE: Interpret and analyze figurative language use (e.g., figurative, connotative meanings) or demonstrate understanding of nuances in word meanings used in context and the impact of those word choices on meaning and tone.	★
Informational Text	
Target 8 (Informational Text) KEY DETAILS: Given an inference or conclusion, use explicit details and implicit information from the text to support the inference or conclusion provided.	+
Target 9 (Informational Text) CENTRAL IDEAS: Determine a central idea and the key details that support it, or provide a summary of the text distinct from personal opinions or judgement.	+
Target 10 (Informational Text) WORD MEANINGS: Determine intended meanings of words including academic/tier 2 words, domain-specific (tier 3) words, and words with multiple meanings, based on context, word relationships (e.g., connotations, denotations), word structure (e.g., common Greek or Latin roots, affixes), or use of reference materials (e.g., dictionary) with primary focus on determining meaning based on context and the academic (tier 2) vocabulary common to complex texts in all disciplines.	+
Target 11 (Informational Text) REASONING & EVIDENCE: Make an inference or draw a conclusion about a text OR make inferences or draw conclusions in order to compare texts (e.g., how a key individual, event, or idea is introduced, illustrated, and elaborated in a text; author's point of view/purpose; use of media or formats; trace and evaluate the argument and specific claims) and use supporting evidence as justification/explanation.	+
Target 12 (Informational Text) ANALYSIS WITHIN OR ACROSS TEXTS: Analyze how information is presented within or across texts (e.g. individuals, events, or ideas) or determine how information within or across texts reveals author's point of view or purpose.	+
Target 13 (Informational Text) TEXT STRUCTURES OR TEXT FEATURES: Relate knowledge of text structures (e.g. sentence, paragraph) or text features to analyze or integrate the impact of those choices on meaning or presentation.	=
Target 14 (Informational Text) LANGUAGE USE: Interpret understanding of figurative language, word relationships, nuances of words and phrases, or figures of speech (e.g., personification) used in context and the impact of those word choices on meaning.	+

Exhibit 6. Target Detail Page for ELA/Lit: Teacher Level

Student Performance on Each Target for the ELA/Literacy Test
What are my students' strengths and weaknesses in the ELA/Literacy Target?

Test: Smarter Summative ELA/Literacy Grade 6
Year: 2018-2019
Name: Demo, Teacher A

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

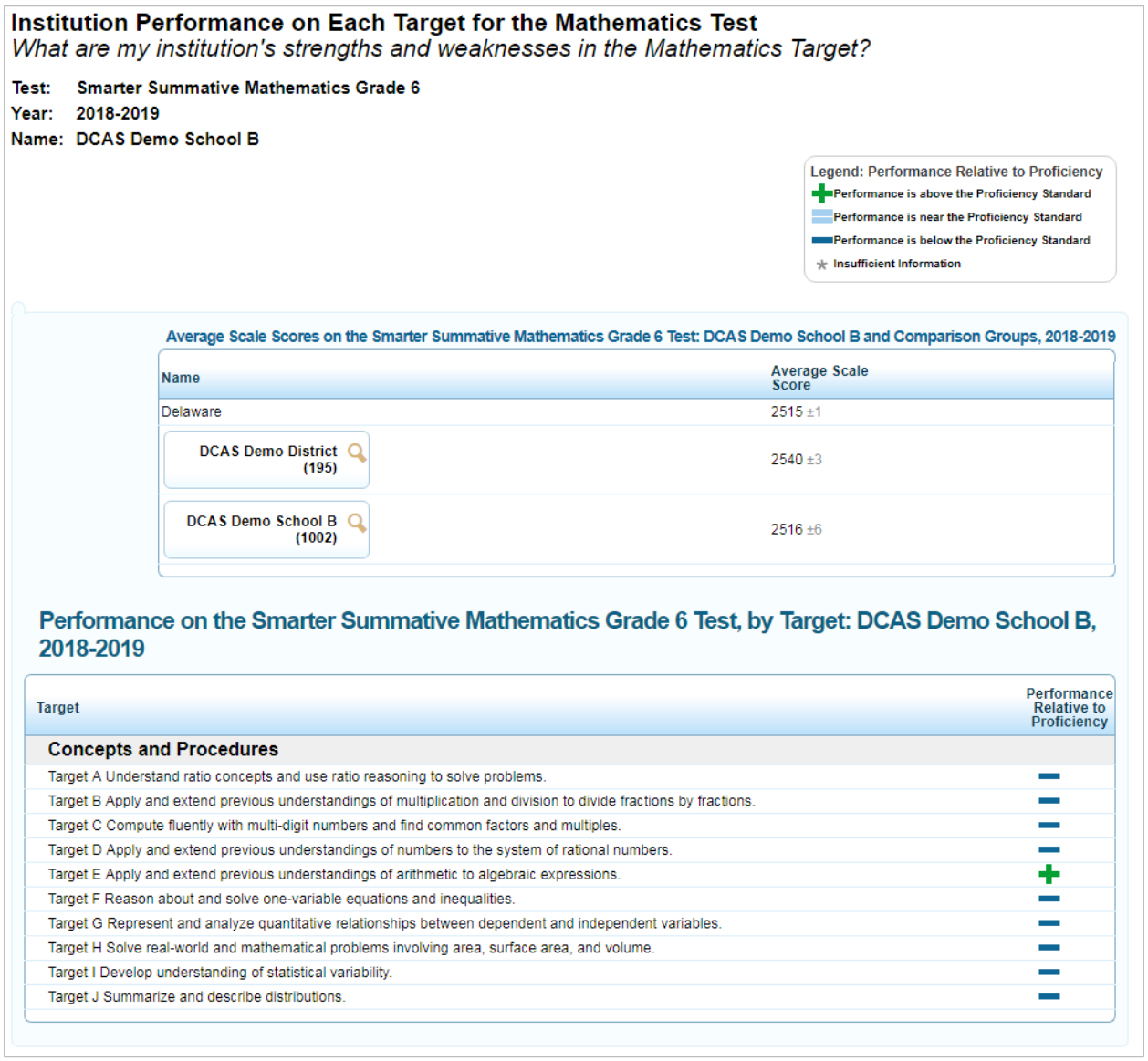
Average Scale Scores on the Smarter Summative ELA/Literacy Grade 6 Test: Demo, Teacher A and Comparison Groups, 2018-2019

Name	Average Scale Score
Delaware	2531 ±1
DCAS Demo District (195)	2546 ±3
DCAS Demo School C (1003)	2568 ±5
Demo, Teacher A	2571 ±5

Performance on the Smarter Summative ELA/Literacy Grade 6 Test, by Target: Demo, Teacher A, 2018-2019

Target	Performance Relative to Proficiency
Reading	
Literary Text	
Target 1 (Literary Text) KEY DETAILS: Given an inference or conclusion, use explicit details and implicit information from the text to support the inference or conclusion provided.	+
Target 2 (Literary Text) CENTRAL IDEAS: Determine a theme or central idea from details in the text, or provide a summary distinct from personal opinions or judgment.	=
Target 3 (Literary Text) WORD MEANINGS: Determine intended or precise meanings of words, including academic/tier 2 words, domain-specific (tier 3) words, and words with multiple meanings, based on context, word relationships (e.g., connotations, denotations), word structure (e.g., common Greek or Latin roots, affixes), or use of reference materials (e.g., dictionary) with primary focus on determining meaning based on context and the academic (tier 2) vocabulary common to complex texts in all disciplines.	=
Target 4 (Literary Text) REASONING & EVIDENCE: Make an inference or draw a conclusion about a text OR make inferences or draw conclusions in order to compare texts (e.g., character development, plot, point of view, themes, topics) and use supporting evidence as justification/explanation.	=
Target 5 (Literary Text) ANALYSIS WITHIN OR ACROSS TEXTS: Describe and explain relationships among literary elements (e.g., plot, character, resolution) within or across texts or explain how the author develops the narrator or speakers' point of view within or across texts.	+
Target 6 (Literary Text) TEXT STRUCTURES & FEATURES: Analyze text structures and the impact of those choices on meaning or presentation.	=
Target 7 (Literary Text) LANGUAGE USE: Interpret and analyze figurative language use (e.g., figurative, connotative meanings) or demonstrate understanding of nuances in word meanings used in context and the impact of those word choices on meaning and tone.	★
Informational Text	
Target 8 (Informational Text) KEY DETAILS: Given an inference or conclusion, use explicit details and implicit information from the text to support the inference or conclusion provided.	+
Target 9 (Informational Text) CENTRAL IDEAS: Determine a central idea and the key details that support it, or provide a summary of the text distinct from personal opinions or judgement.	+
Target 10 (Informational Text) WORD MEANINGS: Determine intended meanings of words including academic/tier 2 words, domain-specific (tier 3) words, and words with multiple meanings, based on context, word relationships (e.g., connotations, denotations), word structure (e.g., common Greek or Latin roots, affixes), or use of reference materials (e.g., dictionary) with primary focus on determining meaning based on context and the academic (tier 2) vocabulary common to complex texts in all disciplines.	+
Target 11 (Informational Text) REASONING & EVIDENCE: Make an inference or draw a conclusion about a text OR make inferences or draw conclusions in order to compare texts (e.g., how a key individual, event, or idea is introduced, illustrated, and elaborated in a text; author's point of view/purpose; use of media or formats; trace and evaluate the argument and specific claims) and use supporting evidence as justification/explanation.	+
Target 12 (Informational Text) ANALYSIS WITHIN OR ACROSS TEXTS: Analyze how information is presented within or across texts (e.g. individuals, events, or ideas) or determine how information within or across texts reveals author's point of view or purpose.	+
Target 13 (Informational Text) TEXT STRUCTURES OR TEXT FEATURES: Relate knowledge of text structures (e.g. sentence, paragraph) or text features to analyze or integrate the impact of those choices on meaning or presentation.	=
Target 14 (Informational Text) LANGUAGE USE: Interpret understanding of figurative language, word relationships, nuances of words and phrases, or figures of speech (e.g., personification) used in context and the impact of those word choices on meaning.	+

Exhibit 7. Target Detail Page for Mathematics: School Level



Institution Performance on Each Target for the Mathematics Test

What are my institution's strengths and weaknesses in the Mathematics Target?

Test: Smarter Summative Mathematics Grade 6
Year: 2018-2019
Name: DCAS Demo School B

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

Average Scale Scores on the Smarter Summative Mathematics Grade 6 Test: DCAS Demo School B and Comparison Groups, 2018-2019

Name	Average Scale Score
Delaware	2515 ±1
DCAS Demo District (195)	2540 ±3
DCAS Demo School B (1002)	2516 ±6

Performance on the Smarter Summative Mathematics Grade 6 Test, by Target: DCAS Demo School B, 2018-2019

Target	Performance Relative to Proficiency
Concepts and Procedures	
Target A Understand ratio concepts and use ratio reasoning to solve problems.	−
Target B Apply and extend previous understandings of multiplication and division to divide fractions by fractions.	−
Target C Compute fluently with multi-digit numbers and find common factors and multiples.	−
Target D Apply and extend previous understandings of numbers to the system of rational numbers.	−
Target E Apply and extend previous understandings of arithmetic to algebraic expressions.	+
Target F Reason about and solve one-variable equations and inequalities.	−
Target G Represent and analyze quantitative relationships between dependent and independent variables.	−
Target H Solve real-world and mathematical problems involving area, surface area, and volume.	−
Target I Develop understanding of statistical variability.	−
Target J Summarize and describe distributions.	−

Exhibit 8. Target Detail Page for Mathematics: Teacher Level

Student Performance on Each Target for the Mathematics Test

What are my students' strengths and weaknesses in the Mathematics Target?

Test: Smarter Summative Mathematics Grade 6
Year: 2018-2019
Name: Demo, Teacher A

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

Average Scale Scores on the Smarter Summative Mathematics Grade 6 Test: Demo, Teacher A and Comparison Groups, 2018-2019

Name	Average Scale Score
Delaware	2515 ±1
DCAS Demo District (195)	2540 ±3
DCAS Demo School B (1002)	2516 ±6
Demo, Teacher A	2508 ±8

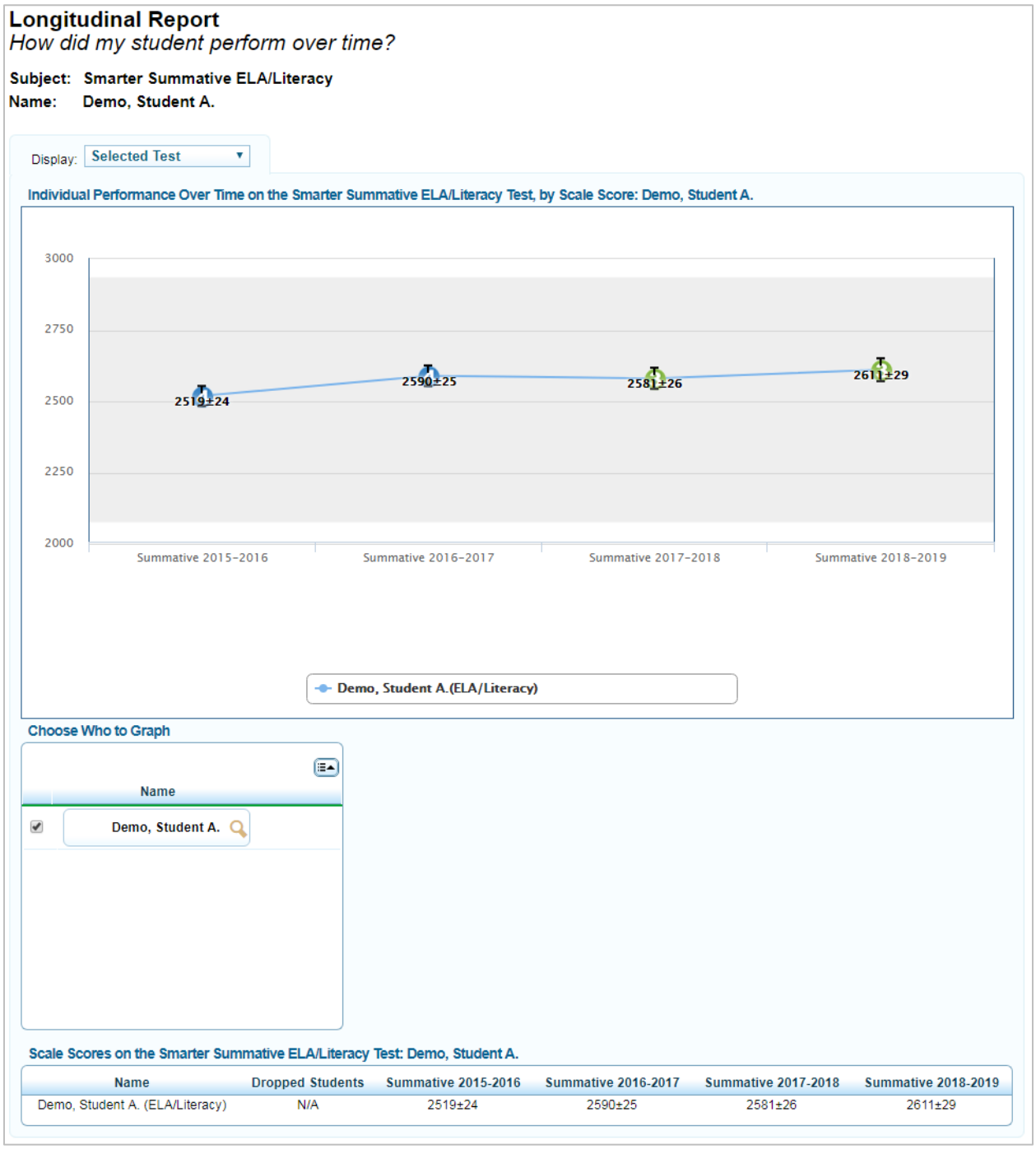
Performance on the Smarter Summative Mathematics Grade 6 Test, by Target: Demo, Teacher A, 2018-2019

Target	Performance Relative to Proficiency
Concepts and Procedures	
Target A Understand ratio concepts and use ratio reasoning to solve problems.	
Target B Apply and extend previous understandings of multiplication and division to divide fractions by fractions.	
Target C Compute fluently with multi-digit numbers and find common factors and multiples.	
Target D Apply and extend previous understandings of numbers to the system of rational numbers.	
Target E Apply and extend previous understandings of arithmetic to algebraic expressions.	
Target F Reason about and solve one-variable equations and inequalities.	
Target G Represent and analyze quantitative relationships between dependent and independent variables.	
Target H Solve real-world and mathematical problems involving area, surface area, and volume.	
Target I Develop understanding of statistical variability.	
Target J Summarize and describe distributions.	

7.1.2.5 Trend Report Page

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

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



7.1.2.6 Student Detail Page

When a student completes a test and items are handscored, an online score report appears on the student detail page in the ORS, which shows the student performance on the test. In each subject area, the student detail page provides: (1) scale score and SEM; (2) achievement level for overall test; (3) performance category in each claim; (4) average scale scores at the state, and the corresponding district, school, classroom teacher, and associated standard errors of the average scale scores; and (5) student performance growth over time.

Exhibits 10 and 11 present examples of student detail pages for ELA/lit and mathematics.

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

Exhibit 10. Student Detail Page for ELA/Lit

Individual Student Report
How did my student perform on the ELA/Literacy test?

Test: Smarter Summative ELA/Literacy Grade 6
Year: 2018-2019
Name: Demo, Student A.

Overall Performance on the Smarter Summative ELA/Literacy Grade 6 Test: Demo, Student A., 2018-2019

Name	SSID	Scale Score	Achievement Level
Demo, Student A.	195091	2611 ±29	Level 3

Scale Score and Performance on the Smarter Summative ELA/Literacy Grade 6 Test: Demo, Student A., 2018-2019

2837
2616
2631
2457
2079

Demo, Student A. Scored 2611±29

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.

Average Scale Scores on the Smarter Summative ELA/Literacy Grade 6 Test: DCAS Demo School C and Comparison Groups, 2018-2019

Name	Average Scale Score
Delaware	2631 ±1
DCAS Demo District (195)	2546 ±3
DCAS Demo School C (1003)	2568 ±5

Legend: Achievement Levels

1 Level 1 2 Level 2 3 Level 3 4 Level 4

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

Performance Over Time on the Smarter Summative ELA/Literacy Test: Demo, Student A.

3000
2750
2500
2250
2000

2519±24
2590±25
2581±26
2611±29

Summative 2015-2016
Summative 2016-2017
Summative 2017-2018
Summative 2018-2019

Demo, Student A (ELA/Literacy)

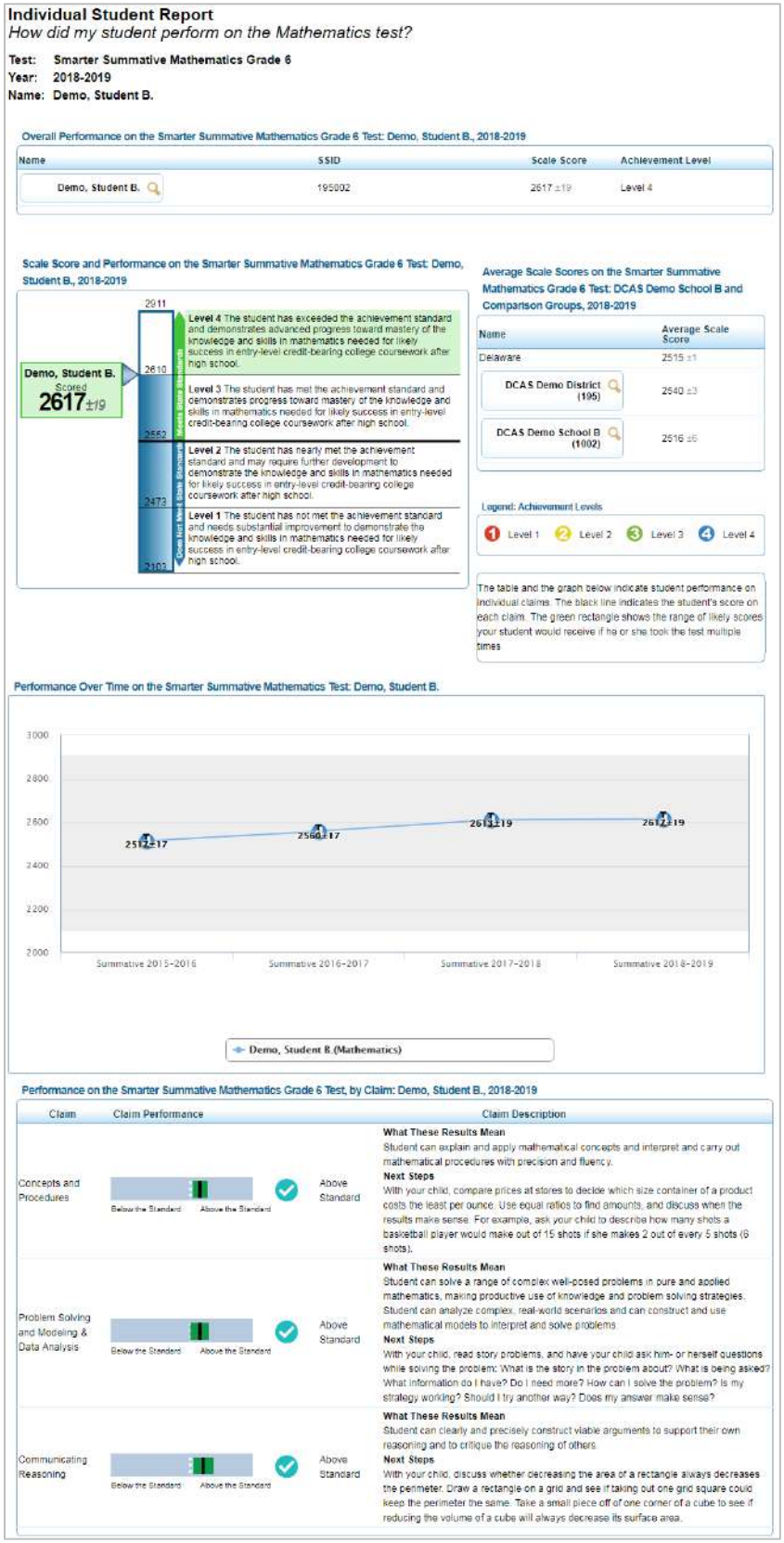
Performance on the Smarter Summative ELA/Literacy Grade 6 Test, by Claim: Demo, Student A., 2018-2019

Claim	Claim Performance	Claim Description
Reading	Below the Standard / Above the Standard — Above Standard	**What These Results Mean** Student can read closely and analytically to comprehend a range of increasingly complex literary and informational texts. **Next Steps** Have your child explain how text parts (characters, events, words, paragraphs) work together to create meaning in different types of texts. Compare texts about the same topic, and discuss different interpretations.
Writing	Below the Standard / Above the Standard — Above Standard	**What These Results Mean** Student can produce effective and well-grounded writing for a range of purposes and audiences. **Next Steps** Help your child write argumentative essays (address opposing views), or informational essays (explain a point). The essays should be organized, quote sources as support, and include specific language about the topic.
Listening	Below the Standard / Above the Standard — At/Near Standard	**What These Results Mean** Student may be able to employ effective listening skills for a range of purposes and audiences. **Next Steps** Have your child listen to or watch a report. Have him or her talk about the main topic, outline the supporting evidence in the report, and point out specific points where more evidence is needed.
Research/Inquiry	Below the Standard / Above the Standard — At/Near Standard	**What These Results Mean** Student may be able to engage in research and inquiry to investigate topics, and to analyze, integrate, and present information. **Next Steps** Have your child conduct a short research project using reliable sources, and be sure to state when reporting someone else's ideas.

Writing Performance on the Smarter Summative ELA/Literacy Grade 6 Test, Based on the Smarter Balanced Performance Task Writing Rubric: Demo, Student A., 2018-2019

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

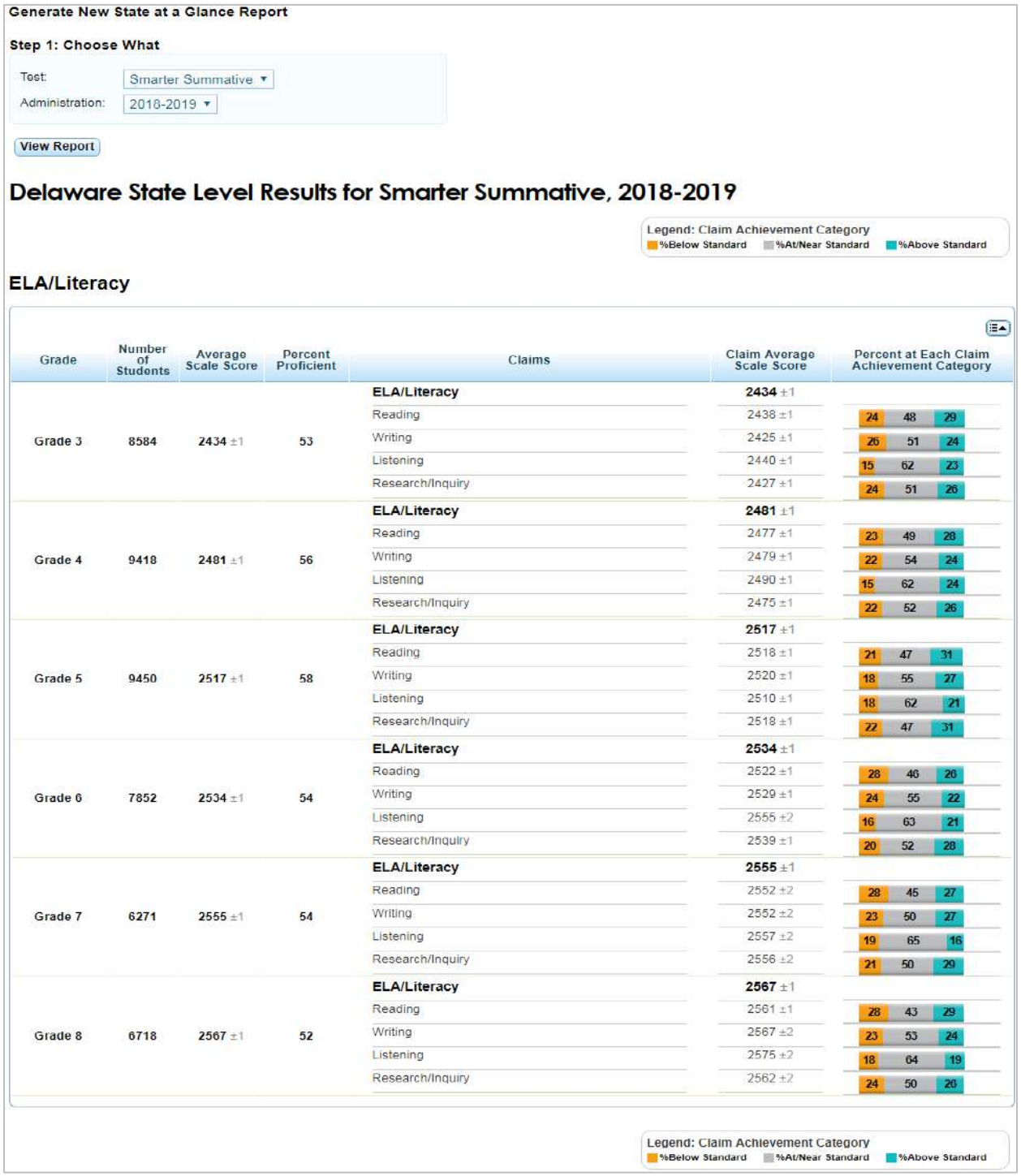
Exhibit 11. Student Detail Page for Mathematics



7.1.2.8 State-Level Summary

The ORS provides the "State at a Glance" page for authorized state-level users to track student performance for a test across the entire state. Users can specify the test and administration year to display in the report. Exhibit 12 presents a sample of state-level summary for ELA/lit.

Exhibit 12. State at a Glance ELA/Lit



Generate New State at a Glance Report

Step 1: Choose What

Test: Smarter Summative

Administration: 2018-2019

View Report

Delaware State Level Results for Smarter Summative, 2018-2019

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

ELA/Literacy

Grade	Number of Students	Average Scale Score	Percent Proficient	Claims	Claim Average Scale Score	Percent at Each Claim Achievement Category
Grade 3	8584	2434 ±1	53	**ELA/Literacy**	**2434** ±1	
				Reading	2438 ±1	24 48 29
				Writing	2425 ±1	26 51 24
				Listening	2440 ±1	15 62 23
				Research/Inquiry	2427 ±1	24 51 26
Grade 4	9418	2481 ±1	56	**ELA/Literacy**	**2481** ±1	
				Reading	2477 ±1	23 49 28
				Writing	2479 ±1	22 54 24
				Listening	2490 ±1	15 62 24
				Research/Inquiry	2475 ±1	22 52 26
Grade 5	9450	2517 ±1	58	**ELA/Literacy**	**2517** ±1	
				Reading	2518 ±1	21 47 31
				Writing	2520 ±1	18 55 27
				Listening	2510 ±1	18 62 21
				Research/Inquiry	2518 ±1	22 47 31
Grade 6	7852	2534 ±1	54	**ELA/Literacy**	**2534** ±1	
				Reading	2522 ±1	28 46 26
				Writing	2529 ±1	24 55 22
				Listening	2555 ±2	16 63 21
				Research/Inquiry	2539 ±1	20 52 28
Grade 7	6271	2555 ±1	54	**ELA/Literacy**	**2555** ±1	
				Reading	2552 ±2	28 45 27
				Writing	2552 ±2	23 50 27
				Listening	2557 ±2	19 65 16
				Research/Inquiry	2556 ±2	21 50 29
Grade 8	6718	2567 ±1	52	**ELA/Literacy**	**2567** ±1	
				Reading	2561 ±1	28 43 29
				Writing	2567 ±2	23 53 24
				Listening	2575 ±2	18 64 19
				Research/Inquiry	2562 ±2	24 50 26

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

7.2 PAPER FAMILY REPORTS

After the testing window is closed, parents whose children participated in a test receive a full-color paper score report (hereinafter family report) that includes their children's performance on ELA/lit and mathematics. The family report includes information on student performance that is provided on the student details page from the ORS. Exhibit 13 presents an example of paper family score reports.

Exhibit 13. Sample Paper Family Score Report

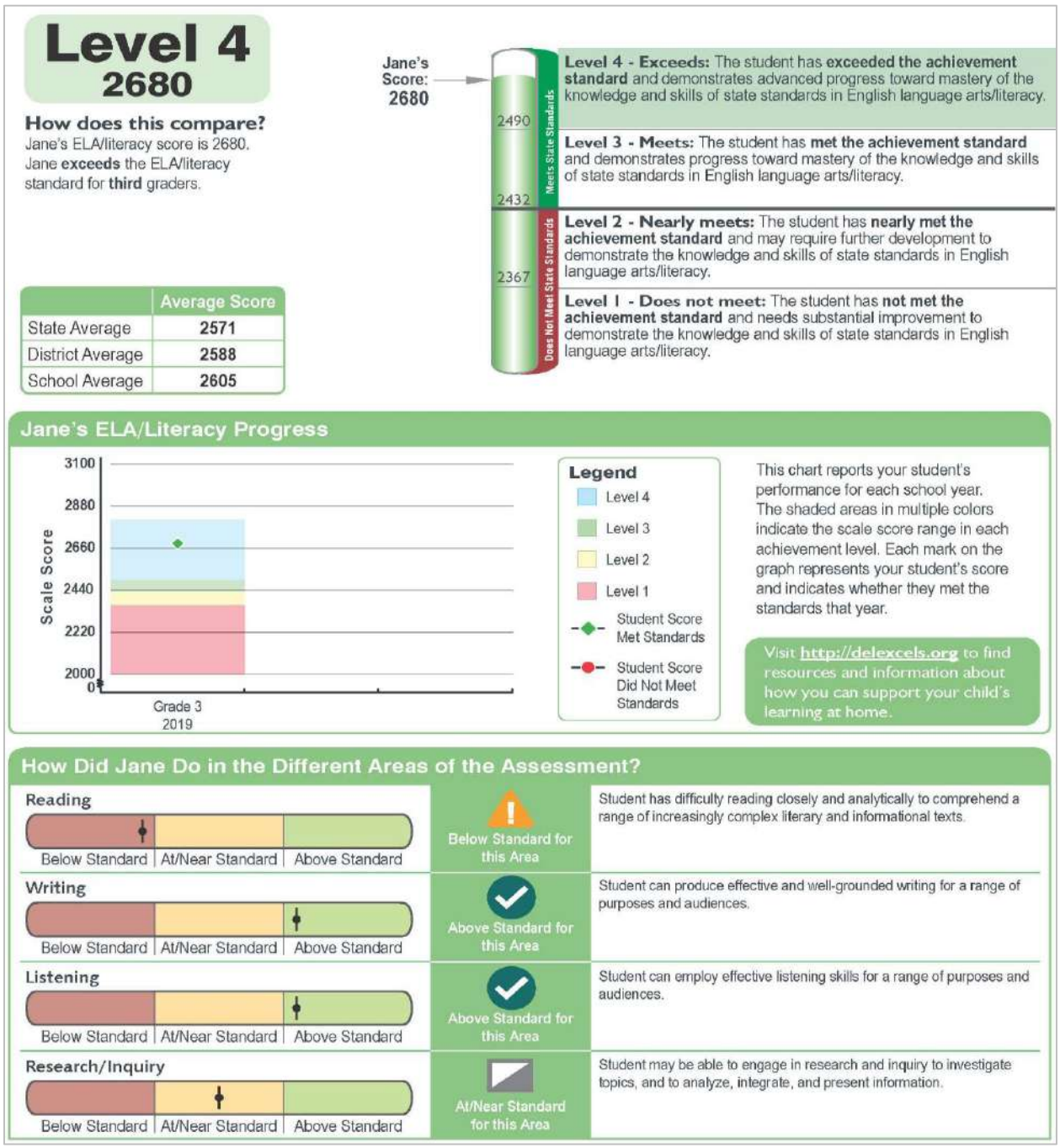
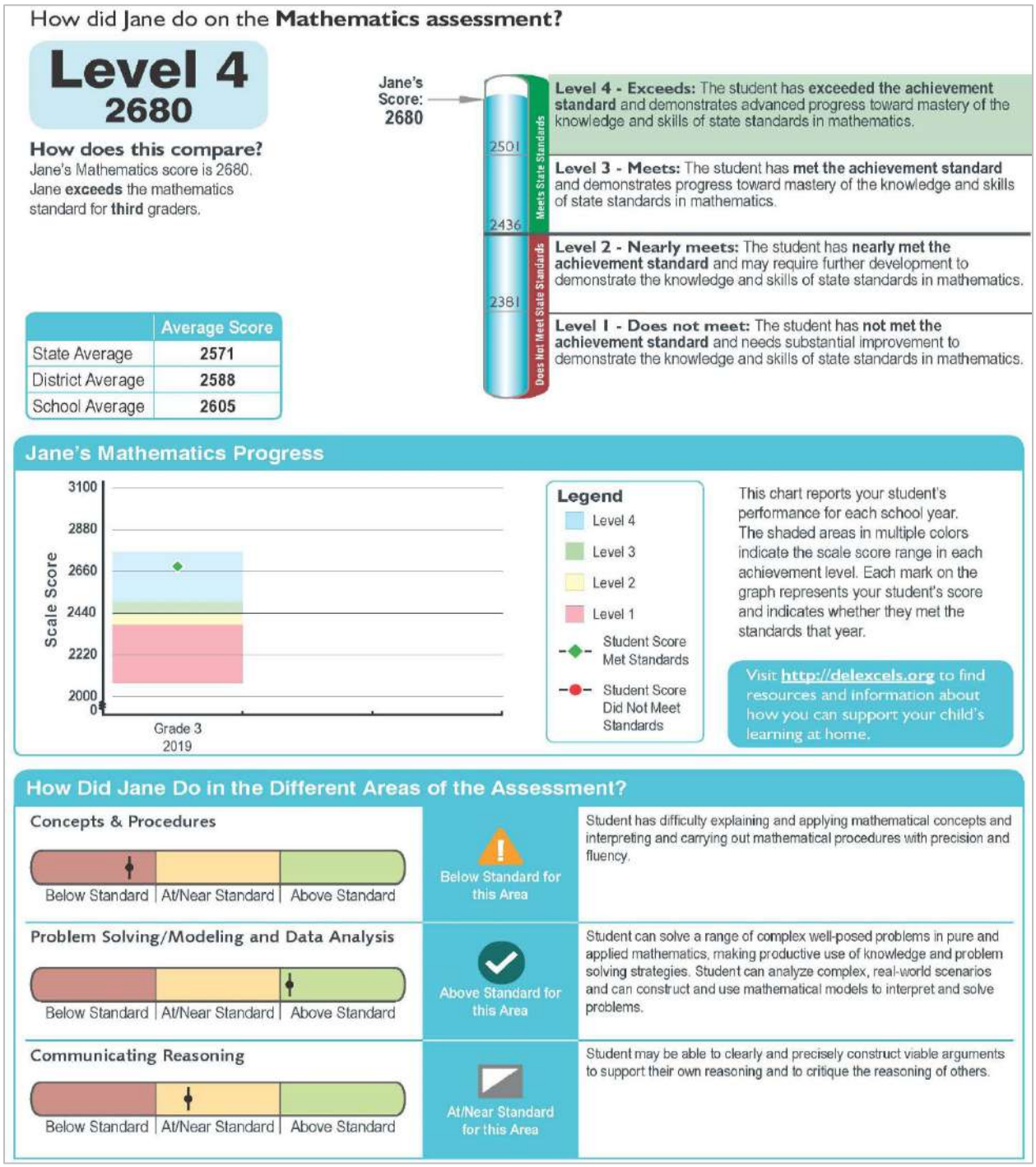


Exhibit 13. Sample Paper Family Score Report (continued)



7.3 INTERPRETATION OF REPORTED SCORES

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

7.3.1 Scale Score

A scale score is used to describe how well a student performed on a test and can be interpreted as an estimate of the student's knowledge and skills. The scale score is the transformed score from a theta score, which is estimated based on mathematical models. Low scale scores indicate that the student does not possess sufficient knowledge and skills measured by the test. Conversely, high scale scores indicate that the student has proficient knowledge and skills measured by the test. Scale scores can be used to measure student growth across school years. Interpretation of scale scores is more meaningful when the scale scores are used along with achievement levels and achievement-level descriptors.

7.3.2 Conditional Standard Error of Measurement

A scale score (observed score on any test) is an estimate of the true score. If a student takes a similar test multiple times, the resulting scale score will vary across administrations, sometimes a little higher, a little lower, or the same. The SEM represents the precision of the scale score, or the range in which the student would likely score if a similar test was administered multiple times. SEM also can be different for the same scale score, depending on how closely the administered items match the student's ability, yielding conditional SEM (CSEM). When interpreting scale scores, it is recommended to consider the range of scale scores, incorporating the CSEM of the scale score.

The ± next to the student's scale score provides information about the certainty, or confidence, of the score's interpretation. The boundaries of the score band are one CSEM above and below the student's observed scale score, representing a range of score values that is likely to contain the true score. For example, 2680 ± 10 indicates that if a student was tested again, it is likely that the student would receive a score between 2670 and 2690.

7.3.3 Achievement Level

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

7.3.4 Performance Category for Claims

Students' performance on each claim is reported in three categories: (1) Below Standard, (2) At/Near Standard, and (3) Above Standard. Unlike the achievement level for the overall test, students' performance on each of the claims is evaluated with respect to the "Meets Standard" achievement standard. For students performing at either "Below Standard" or "Above Standard," this can be interpreted to mean that students' 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 students' performance does not

provide enough information to tell whether students reached the "Meets Standard" mark for the specific claim.

7.3.5 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 report is generated for aggregate units of classroom, school, and district and provides information about how a group of students in a class, school, or district performed on the reporting target that is relative to the proficiency cut set by Smarter Balanced. At the aggregate level, when observed performance within a target is greater than the proficiency cut, the reporting unit shows a relative strength in that target. Conversely, when observed performance within a target is below the proficiency cut, the reporting unit shows a relative weakness in that target.

The performance on target is mapped into three performance categories: (1) performance is above the proficiency standard, (2) performance is near the proficiency standard, and (3) performance is below the proficiency standard. Although performance categories for targets provide some evidence to help address students' strengths and weaknesses, they should not be over-interpreted because student performance on each target is based on relatively few items, especially for a small group.

7.3.6 Aggregated Score

Student scale scores are aggregated at roster, teacher, school, district, and state levels to represent how a group of students performs on a test. When students' scale scores are aggregated, the aggregated 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 aggregated scale scores are also estimates and are subject to measures of uncertainty. In addition to the aggregated scale scores, the percentage of students in each achievement level for the overall test and by claim are reported at the aggregate level to represent how well a group of students performs on the overall test and by claim.

7.4 Appropriate Uses for Scores and Reports

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

Assessment results for student achievement on the test can be used to help teachers or schools decide how to support student learning. Aggregate score reports at the teacher and school level provide information regarding the strengths and weaknesses of their students by claim and by target and thus can be utilized to improve teaching and student learning. For example, a group of students could performe very well in the overall test, but it is possible that they would not perform as well in some claims or targets. In this case, teachers and schools can identify the strengths and weaknesses of their students through the group performance by claim or by targets and promote instruction on specific claim 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 can see student assessment results by ELL status and observe that ELL students are struggling with literary response and analysis in reading. Teachers can then provide additional instructions for these students to enhance their achievement in a specific target in a claim.

In addition, assessment results can be used to compare student 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, as well as by claim. Although all students are administered different sets of items in each computer-adaptive test, scale scores are comparable across students. Furthermore, scale scores can be used to measure the growth of individual students over time if data are available. In the Smarter Balanced assessments, the scale scores across grades are on the same scale because the scores are vertically linked across grades. Therefore, scale scores from one grade can be compared with the next grade, i.e., measuring the growth.

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

8. QUALITY CONTROL PROCEDURES

Quality assurance (QA) procedures are enforced through all stages of the test form development, administration, and scoring, and reporting of results. AIR uses a series of quality control steps to ensure the error-free production of score reports for both online and paper-pencil formats. The quality of the information produced in the test delivery system (TDS) is tested thoroughly before, during, and after the testing window opens.

8.1 ADAPTIVE TEST CONFIGURATION

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

After the adaptive testing simulations, another set of simulations for the combined tests (CAT and 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. AIR is rigorously check whether the scoring rule specified in scoring specifications were applied accurately. The scores in the simulated data file are checked independently.

8.1.1 Platform Review

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

Platform review is the process for checking every item to ensure that it is displayed appropriately on the corresponding 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 a content approval role. The UAT period provides the department with an opportunity to interact with the exact test that the students will use.

8.2 QUALITY ASSURANCE IN DOCUMENT PROCESSING

The Smarter Balanced assessments are administered primarily online; however, some students need to take the paper-pencil version of the assessments to meet their special needs. When test documents are scanned, a quality-control sample is created, consisting of 10 test cases per document type. There are normally between 500 and 600 different types of documents. All student responses and demographic grids need to be verified, including various typical errors that required editing via MI's Data Inspection, Correction, and Entry (DICE) application program. This structured testing method provided exact test parameters and a methodical way of determining that the output received from the scanners was correct. MI staff carefully compared the documents and the data file created from them to ensure further that the results from the scanner, the editing process (validation and data correction), and the transfer to the AIR database are correct.

8.3 QUALITY ASSURANCE IN DATA PREPARATION

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

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

8.4 QUALITY ASSURANCE IN HANDSCORING

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

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

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

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

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

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

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

8.4.2 Handscoring QA Monitoring Reports

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

8.4.3 Monitoring by State Department of Education

The DDOE staff also view the MI scoring activities virtually by the access to the rater trainings through the online training interface and monitor the scoring process via the Client Command Center (CCC) and reviewing the scoring data and reports during the scoring process.

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

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

This process is also used to notify each consortium state of possible instances of teacher or proctor interference or of student collusion with others. The alert procedure is habitually explained during scorer training sessions. Within the VSC system, if a scorer identifies a response which may require an alert, he or she flags or notes that response as a possible alert and transfers the image to the scoring manager. Scoring management then decides if the response should be forwarded to the client for any necessary action or follow-up.

8.5 QUALITY ASSURANCE IN TEST SCORING

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

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

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

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

Table 49 presents an overview of the QA reports.

Table 49. 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 issues
Item Exposure Rates	To monitor unlikely high exposure rates of items or passages or unusually low item pool usage (highly unused items/passages)	Early detection of any oversight in the blueprint specification
Cheating Analysis	To monitor testing irregularities	Early detection of testing irregularities

8.5.1 Score Report Quality Check

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

8.5.1.1 Online Report Quality Assurance

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

8.5.1.2 Paper Report Quality Assurance

Statistical Programming

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

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

tested and stored, changes to the macro must be approved by the director of score reporting and the director of psychometrics, as well as by the project directors for affected projects.

Each change is followed by a complete retesting with the entire collection of scenarios on which the macro was originally tested. The main statistical program is mostly made up of calls to various macros, including macros that verify the data and conversion tables and the macros that 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 VIPP and allows virtually infinite control of the visual appearance of the reports. After designers at AIR create backgrounds, our VIPP programmers write code that indicates where to place all variable information (data, graphics, and text) on the reports. The VIPP code is tested using both artificial and real data. AIR's data generation utilities can read the output layout specifications and generate artificial data for direct input into the VIPP programs. This allows the testing of these programs to begin before the statistical programming is complete. In later stages, artificial data are generated according to the input layout and are run through the psychometric process and the score reporting statistical programs, and the output is formatted as VIPP input. This enables us to test the entire system.

Programmed output goes through multiple stages of review and revision by graphics editors and the AIR Score Reporting team to ensure that design elements are accurately reproduced and data are correctly displayed. Once we receive final data and VIPP programs, the AIR Score Reporting team reviews proofs that contain actual data based on our standard quality assurance documentation. Additionally, we compare data independently calculated by AIR psychometricians with data on the reports. A large sample of reports is reviewed by several AIR staff members to make sure that all data are correctly placed on reports. This rigorous review typically is conducted over several days and takes place in a secure location in the AIR building. All reports containing actual data are stored in a locked storage area. Before the reports are printed, AIR provides a live data file and individual student reports with sample districts for DDOE staff review. AIR will work closely with the DDOE to resolve questions and correct any problems. The reports will not be delivered unless the DDOE approves the sample reports and data file.

REFERENCES

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

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

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

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

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

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

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

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

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

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

APPENDICES

Appendix A: Summary of the 2018–2019 Interim Assessments

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

Table A-1. Number of Students Who Took ICAs

Grade	Number of Students by Number of Attempts						Total Number of Tests Taken
	Once	Twice	Three Times	Four Times	Five Times	Total Number of Students	
ELA/Lit							
3	623	8	1	0	0	632	642
4	527	1	0	0	0	528	529
5	527	1	0	0	0	528	529
6	273	0	0	0	0	273	273
7	257	0	0	0	0	257	257
8	185	0	0	0	0	185	185
Mathematics							
3	567	9	2	0	0	578	591
4	575	0	0	0	0	575	575
5	487	0	0	0	0	487	487
6	290	0	0	0	0	290	290
7	309	0	0	0	0	309	309
8	243	0	0	0	0	243	243

Table A-2. ICA ELA/Lit and Mathematics Percentage of Students in Achievement Levels

Subject	Grade	Total Number of Tests Taken	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
ELA/Lit	3	642	2402.61	80.62	37	29	19	15	34
	4	529	2466.50	82.12	29	25	22	24	45
	5	529	2491.10	84.84	29	28	29	15	43
	6	273	2513.20	67.30	23	37	34	6	40
	7	257	2530.74	83.75	27	31	33	9	42
	8	185	2523.01	78.78	36	34	28	3	31
Math	3	591	2420.58	68.06	30	28	31	11	42
	4	575	2482.10	84.90	20	35	24	22	45
	5	487	2492.64	93.13	35	34	12	18	30
	6	290	2519.71	78.46	26	40	22	12	34
	7	309	2541.94	90.15	28	34	20	17	38
	8	243	2554.19	96.57	32	32	19	18	36

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

For the Interim Assessment Block assessments (IABs), there were seven to nine IABs for ELA/lit and six IABs in mathematics. Students were allowed to take as many IABs as they wanted. Table A–3 shows the total number of students who took the IABs and the number of students by the number of IABs taken. For example, in grade 3 ELA/lit, a total of 3,573 students took the IABs, and among these students, 1,481 students took one IAB, 983 students took two IABs, and so on.

Tables A–4 to A–6 disaggregated the number of students in Table A–3 by each individual block. For example, 1,481 students in grade 3 took one IAB only in ELA/lit. Among these students, six students took the Brief Writes IAB, 196 students took the Editing IAB, and so on. Tables A–7 to A–9 show the percentage of students in each performance category for all students for each IAB.

Table A-3. Number of Students Who Took IABs

Grade	Total	Number of IABs Taken								
		1	2	3	4	5	6	7	8	9
ELA/Lit										
3	3,573	1,481	983	676	215	161	53	4		
4	4,364	1,611	1,515	717	335	119	48	19		
5	5,326	1,925	1,646	736	555	282	109	68	5	
6	5,232	2,252	1,639	464	321	402	149	5		
7	5,557	2,375	2,143	722	243	74				
8	3,541	1,420	1,568	438	115					
Mathematics										
3	5,194	1,759	1,369	899	756	406	5			
4	5,857	2,316	1,764	1,320	419	38				
5	7,009	2,967	2,456	923	420	240	3			
6	7,017	2,947	2,746	667	235	380	42			
7	6,625	3,056	1,725	1,407	330	106	1			
8	6,345	2,757	2,562	601	180	236	9			

Table A-4: ELA/Lit Number of Students Who Took IABs by Block Labels (Grades 3–5)

Grade	Block	Number of IABs Taken								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
3	Brief Writes	6	33	35	3	1	3			
	Editing	196	472	317	181	104	51	4		
	Language and Vocabulary Use	285	384	420	186	159	53	4		
	Listening and Interpretation	131	169	206	136	130	40			
	Reading Informational Text	449	433	336	95	78	35	4		
	Reading Literary Text	371	378	391	78	122	53	4		
	Research	11	17	60	36	58	30	4		
	Revision	10	63	124	105	81	35	4		
	Performance Task	22	17	139	40	72	18	4		
4	Brief Writes			1	1	15	10			
	Editing	109	320	313	248	116	48	19		
	Language and Vocabulary Use	295	481	276	197	116	48	19		
	Listening and Interpretation	124	239	243	222	76	48	19		
	Reading Informational Text	714	826	591	193	72	36	19		
	Reading Literary Text	249	633	402	174	38	44	19		
	Research	53	314	235	164	95	46	19		
	Revision	53	110	56	121	67	8	18		
	Performance Task	14	107	34	20			1		
5	Brief Writes			52	4		1			
	Editing	241	278	193	410	273	97	68	5	
	Language and Vocabulary Use	587	1,043	550	465	273	108	68	5	
	Listening and Interpretation	165	339	247	395	206	78	68	5	
	Reading Informational Text	558	616	278	228	156	80	68	5	
	Reading Literary Text	210	769	389	197	118	98	68	5	
	Research	76	135	319	246	194	108	68	5	
	Revision	57	103	117	230	184	71	63	5	
	Performance Task	31	9	63	45	6	13	5	5	

Table A-5: ELA/Lit Number of Students Who Took IABs by Block Labels (Grades 6–8)

Grade	Block	Number of IABs Taken								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
6	Brief Writes	7	9	2						
	Editing	256	451	191	250	394	148	5		
	Language and Vocabulary Use	631	815	342	255	392	149	5		
	Listening and Interpretation	162	273	160	224	218	149	5		
	Reading Informational Text	160	476	134	87	55	16	5		
	Reading Literary Text	993	932	217	158	290	147	5		
	Research	19	32	156	162	362	148	5		
	Revision	24	290	190	148	299	137	5		
	Performance Task									
7	Brief Writes	4	6	10	14	38				
	Editing	750	935	298	241	74				
	Language and Vocabulary Use	274	562	294	204	36				
	Listening and Interpretation	168	242	279	236	74				
	Reading Informational Text	186	958	305	105	38				
	Reading Literary Text	769	993	529	166	74				
	Research	29	356	314	3	1				
	Revision	195	234	137	3	35				
	Performance Task									
8	Brief Writes		4							
	Editing and Revising	567	503	386	115					
	Listening and Interpretation	85	355	123	40					
	Reading Informational Text	209	909	384	115					
	Reading Literary Text	525	964	322	107					
	Research	32	401	97	83					
	Performance Task	2		2						

Table A-6: Mathematics Number of Students Who Took IABs by Block Labels

Grade	Block	Number of IABs Taken					
		1	**2**	**3**	**4**	**5**	**6**
3	Geometry	58	138	272	238	389	5
	Measurement and Data	235	305	450	679	397	5
	Number and Operations in Base Ten	331	572	649	671	406	5
	Number and Operations – Fractions	297	708	706	702	406	5
	Operational and Algebraic Thinking	833	1,015	603	706	405	5
	Performance Task	5		17	28	27	5
4	Geometry	23	239	373	277	38	
	Measurement and Data	67	131	136	219	38	
	Number and Operations in Base Ten	1,109	1,546	1,273	419	38	
	Number and Operations – Fractions	284	534	951	341	38	
	Operational and Algebraic Thinking	568	1,076	1,145	345	38	
	Performance Task	265	2	82	75		
5	Geometry	70	177	238	261	240	3
	Measurement and Data	55	264	111	362	239	3
	Number and Operations in Base Ten	1,793	2,131	886	400	240	3
	Number and Operations – Fractions	718	1,698	850	344	240	3
	Operations and Algebraic Thinking	102	621	594	280	240	3
	Performance Task	229	21	90	33	1	3
6	Expressions and Equations	71	639	352	221	379	42
	Geometry	524	1,026	405	191	377	42
	Number System	509	1,719	496	216	380	42
	Ratios and Proportional Relationships	1,827	2,008	569	222	380	42
	Statistics and Probability	12	81	152	64	364	42
	Performance Task	4	19	27	26	20	42
7	Expressions and Equations	504	539	687	316	106	1
	Geometry	190	325	616	146	106	1
	Number System	609	1,320	1,024	313	105	1
	Ratios and Proportional Relationships	1,742	1,168	1,276	314	106	1
	Statistics and Probability	5	37	238	220	103	1
	Performance Task	6	61	380	11	4	1
8	Expressions and Equations I	980	981	224	143	236	9
	Expressions and Equations II	574	1,346	505	169	236	9
	Functions	779	1,368	493	167	236	9
	Geometry	376	983	328	133	235	9
	Number System	23	184	230	100	189	9
	Performance Task	25	262	23	8	48	9

Table A-7: ELA/Lit Percentage of Students in Performance Categories by IAB Block Labels (Grades 3–5)

Grade	Block	Number Tested	% Below	% At/Near	% Above
3	Brief Writes	81	31	32	37
	Editing	1,325	29	52	18
	Language and Vocabulary Use	1,491	25	50	24
	Listening and Interpretation	812	19	57	24
	Reading Informational Text	1,430	28	51	21
	Reading Literary Text	1,397	26	42	32
	Research	216	19	46	34
	Revision	422	26	48	26
	Performance Task	312	17	43	40
4	Brief Writes	27	15	70	15
	Editing	1,173	20	56	24
	Language and Vocabulary Use	1,432	26	48	26
	Listening and Interpretation	971	19	61	19
	Reading Informational Text	2,451	16	57	27
	Reading Literary Text	1,559	25	54	21
	Research	926	20	45	35
	Revision	433	30	54	16
	Performance Task	176	12	64	24
5	Brief Writes	57	19	51	30
	Editing	1,565	20	51	29
	Language and Vocabulary Use	3,099	25	53	22
	Listening and Interpretation	1,503	20	55	25
	Reading Informational Text	1,989	7	56	37
	Reading Literary Text	1,854	22	50	29
	Research	1,151	23	47	30
	Revision	830	29	50	21
	Performance Task	177	14	54	32

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

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

Grade	Block	Number Tested	% Below	% At/Near	% Above
6	Brief Writes	18	11	22	67
	Editing	1,695	17	47	36
	Language and Vocabulary Use	2,589	26	44	30
	Listening and Interpretation	1,191	19	57	24
	Reading Informational Text	933	20	61	19
	Reading Literary Text	2,742	27	56	17
	Research	884	16	49	34
	Revision	1,093	18	59	23
	Performance Task				
7	Brief Writes	72	18	40	42
	Editing	2,298	17	69	14
	Language and Vocabulary Use	1,370	28	51	21
	Listening and Interpretation	999	19	60	21
	Reading Informational Text	1,592	29	47	24
	Reading Literary Text	2,531	30	50	20
	Research	703	19	69	13
	Revision	604	29	62	10
	Performance Task				
8	Brief Writes	4	100	0	0
	Editing and Revising	1,571	26	46	28
	Listening and Interpretation	603	23	60	16
	Reading Informational Text	1,617	15	47	38
	Reading Literary Text	1,918	31	42	27
	Research	613	25	46	30
	Performance Task	4	100	0	0

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

Table A-9: Mathematics Percentage of Students in Performance Categories by IAB Block Labels

Grade	Block	Number Tested	% Below	% At/Near	% Above
3	Geometry	1,100	24	46	31
	Measurement and Data	2,071	31	43	26
	Number and Operations in Base Ten	2,634	31	35	34
	Number and Operations—Fractions	2,824	15	45	41
	Operational and Algebraic Thinking	3,567	37	44	19
	Performance Task	82	24	33	43
4	Geometry	950	13	69	18
	Measurement and Data	591	7	47	46
	Number and Operations in Base Ten	4,385	34	48	18
	Number and Operations—Fractions	2,148	30	43	27
	Operational and Algebraic Thinking	3,172	37	47	17
	Performance Task	424	9	71	20
5	Geometry	989	20	55	25
	Measurement and Data	1,034	22	45	33
	Number and Operations in Base Ten	5,453	36	44	20
	Number and Operations—Fractions	3,853	34	44	22
	Operations and Algebraic Thinking	1,840	16	46	38
	Performance Task	377	29	51	20
6	Expressions and Equations	1,704	19	47	34
	Geometry	2,565	30	49	22
	Number System	3,362	32	49	19
	Ratios and Proportional Relationships	5,048	52	32	16
	Statistics and Probability	715	17	57	26
	Performance Task	138	25	53	22
7	Expressions and Equations	2,153	22	44	33
	Geometry	1,384	10	68	22
	Number System	3,372	28	52	20
	Ratios and Proportional Relationships	4,607	25	52	23
	Statistics and Probability	604	27	56	18
	Performance Task	463	38	52	9
8	Expressions and Equations I	2,573	32	48	20
	Expressions and Equations II	2,839	30	47	23
	Functions	3,052	35	44	21
	Geometry	2,064	26	54	20
	Number System	735	17	33	49
	Performance Task	375	22	62	16

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

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

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

Group	2014–2015				2015–2016				2016–2017				2017–2018				2018–2019			
	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD
Grade 3																				
All Students	10,231	54	2438.1	84.7	10,296	54	2439.5	85.4	10,600	52	2433.3	87.2	10,467	52	2433.2	87.2	10,234	50	2429.7	89.7
Female	5,122	59	2448.1	83.9	5,122	57	2447.5	84.6	5,171	55	2442.1	85.7	5,160	56	2441.5	84.1	4,995	54	2439.1	87.4
Male	5,109	49	2428.1	84.3	5,174	50	2431.7	85.5	5,429	48	2425.0	87.9	5,307	48	2425.3	89.3	5,239	47	2420.8	90.9
African American	3,016	39	2405.7	81.6	3,109	39	2409.3	79.7	3,206	36	2401.1	81.1	3,174	36	2400.5	81.1	3,107	35	2396.8	85.0
AmerIndian/Alaskan	38	76	2460.6	77.4	40	58	2438.8	81.9	36	53	2430.1	84.3	43	51	2424.1	89.1	23	48	2425.4	73.1
Asian	375	80	2496.6	79.2	363	80	2497.2	85.7	371	78	2494.2	80.2	420	79	2498.2	82.1	420	75	2486.0	86.5
Hispanic/Latino	1,763	41	2415.3	75.7	1,789	41	2414.9	77.0	1,997	39	2407.3	80.1	1,952	38	2406.5	80.6	1,924	37	2403.8	82.3
Pacific Islander	16	50	2426.7	107.3	13	62	2453.8	70.7	13	62	2481.8	79.4	22	64	2446.2	77.4	9*			
White	4,631	66	2462.8	80.6	4,542	66	2464.6	82.2	4,513	66	2461.6	82.8	4,373	67	2462.0	81.6	4,254	65	2458.9	84.1
Multi-Racial	392	59	2440.7	75.8	440	57	2446.6	83.7	464	57	2444.1	85.3	482	55	2439.9	84.2	497	55	2439.0	87.0
ELL	984	23	2382.5	64.5	1,249	28	2390.7	67.9	1,635	32	2397.1	77.5	1,727	36	2401.4	76.7	1,750	33	2394.4	77.4
Special Education	1,279	13	2351.3	70.0	1,334	14	2357.3	69.1	1,438	15	2354.5	72.7	1,447	12	2349.2	72.7	1,555	13	2346.5	73.6
CD 504	332	44	2424.2	73.4	319	52	2430.4	75.8	331	47	2426.7	75.6	342	51	2430.7	76.4	398	45	2425.5	80.0
Title I	1,161	54	2438.6	76.1	1,053	59	2451.2	77.0	1,035	63	2455.5	78.0	1,092	59	2448.6	80.1	1,002	52	2437.4	83.3
Grade 4																				
All Students	9,910	54	2477.4	88.0	10,268	56	2482.5	90.8	10,386	54	2477.2	92.1	10,658	55	2479.3	92.3	10,468	53	2476.1	94.6
Female	4,932	58	2486.6	86.6	5,132	61	2493.7	89.8	5,150	58	2486.9	89.6	5,210	58	2488.6	89.9	5,148	58	2486.0	91.2
Male	4,978	49	2468.3	88.4	5,136	51	2471.3	90.4	5,236	50	2467.6	93.4	5,448	52	2470.3	93.7	5,320	49	2466.6	96.8
African American	3,060	37	2444.4	82.8	3,035	41	2448.3	86.6	3,143	39	2442.8	88.4	3,252	39	2443.7	88.8	3,193	36	2437.9	87.9
AmerIndian/Alaskan	43	65	2494.1	80.1	38	61	2482.5	85.4	41	51	2478.6	81.3	37	51	2472.0	88.3	40	53	2472.6	92.7
Asian	385	81	2541.1	83.5	382	81	2550.7	88.6	383	83	2542.8	82.2	384	83	2543.8	84.3	417	83	2547.0	92.1
Hispanic	1,702	40	2452.8	78.7	1,781	43	2455.9	83.3	1,838	42	2452.0	84.3	2,000	44	2455.9	84.9	1,996	44	2454.2	87.7
Pacific Islander	15	53	2473.1	75.5	14	50	2477.0	97.7	15	80	2511.4	79.8	13	77	2512.4	108.2	18	83	2518.6	81.0
White	4,331	68	2503.9	83.7	4,611	68	2509.6	84.7	4,518	67	2505.1	86.6	4,496	69	2509.2	85.6	4,312	68	2506.8	88.6
Multi-Racial	374	57	2485.6	88.9	407	57	2481.8	87.4	448	56	2483.0	89.0	476	58	2485.9	90.8	492	57	2482.9	90.4
ELL	558	14	2399.6	69.6	641	16	2402.1	73.9	886	21	2412.5	74.6	1,608	38	2442.8	80.2	1,651	39	2442.1	80.3
Special Education	1,349	11	2380.1	71.9	1,452	13	2388.7	74.7	1,474	12	2380.8	78.4	1,610	17	2389.3	82.4	1,707	14	2385.8	77.7
CD 504	376	51	2471.7	75.4	374	49	2469.5	84.2	411	47	2467.5	86.1	417	53	2469.9	85.2	448	50	2473.2	85.3
Title I	1,274	49	2467.9	80.1	1,243	57	2484.9	78.6	1,046	58	2484.0	82.1	1,054	61	2492.6	80.5	1,059	59	2488.5	81.7

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

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

Group	2014–2015				2015–2016				2016–2017				2017–2018				2018-2019			
	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD
Grade 5																				
All Students	9,922	55	2509.4	89.3	10,169	60	2519.3	90.0	10,461	60	2519.7	93.3	10,579	58	2516.6	92.1	10,827	57	2514.3	95.2
Female	4,890	61	2522.7	86.7	5,053	66	2531.1	87.0	5,230	65	2532.7	92.1	5,275	63	2528.2	89.0	5,282	62	2525.9	91.0
Male	5,032	50	2496.4	89.9	5,116	55	2507.6	91.3	5,231	55	2506.7	92.7	5,304	54	2505.1	93.6	5,545	53	2503.2	97.7
African American	3,115	39	2473.8	85.0	3,077	44	2485.0	84.9	3,077	45	2484.1	89.1	3,216	41	2479.1	87.1	3,309	39	2474.3	90.0
AmerIndian/Alaskan	41	59	2518.4	86.6	41	68	2540.1	76.2	31	61	2519.1	95.9	40	60	2523.5	91.3	28	54	2501.3	84.5
Asian	361	84	2579.1	83.6	386	85	2585.3	79.9	367	87	2591.9	86.7	384	86	2587.6	84.9	392	83	2585.2	87.8
Hispanic/Latino	1,533	44	2486.3	79.4	1,761	49	2492.9	84.0	1,824	47	2494.5	83.9	1,872	48	2494.7	85.0	2,021	48	2494.4	86.0
Pacific Islander	10	80	2534.2	75.2	12	83	2556.1	53.9	12	42	2493.1	133.1	11	82	2540.4	112.3	12	67	2525.9	103.6
White	4,585	68	2534.9	84.2	4,490	73	2546.6	84.3	4,708	72	2546.9	88.4	4,575	71	2544.7	86.1	4,548	71	2545.4	89.0
Multi-Racial	277	60	2520.8	85.2	402	64	2525.4	89.7	442	62	2522.0	87.5	481	61	2527.0	85.9	517	57	2520.4	92.6
ELL	303	9	2409.2	65.4	420	13	2418.5	75.3	440	13	2413.6	74.3	886	23	2447.4	78.4	1,264	28	2456.6	74.3
Disadvantaged	1,381	11	2408.2	70.6	1,451	15	2420.2	76.3	1,526	16	2417.7	80.3	1,612	14	2419.9	77.2	1,802	15	2417.0	79.9
Migrant	412	50	2502.1	82.6	424	53	2504.4	77.9	462	56	2510.7	80.5	493	55	2508.0	81.1	555	58	2514.0	82.8
Disability	1,621	56	2510.5	84.7	1,359	60	2519.7	81.6	1,247	64	2526.3	83.4	1,066	63	2526.7	82.8	1,050	62	2525.1	86.2
Grade 6																				
All Students	10,023	48	2522.8	92.4	9,983	52	2530.2	93.5	10,189	52	2529.7	93.4	10,425	52	2531.2	95.7	10,572	52	2528.9	97.6
Female	4,943	55	2538.9	89.1	4,923	57	2544.4	90.0	5,055	57	2542.4	91.0	5,222	59	2545.7	93.1	5,281	57	2541.4	94.1
Male	5,080	41	2507.1	92.9	5,060	46	2516.3	94.7	5,134	47	2517.1	94.0	5,203	46	2516.5	96.1	5,291	47	2516.4	99.4
African American	3,097	33	2490.4	87.3	3,135	35	2494.5	87.4	3,133	35	2493.7	87.5	3,087	37	2496.7	89.7	3,249	36	2491.2	93.0
AmerIndian/Alaskan	48	52	2536.1	81.7	43	47	2526.1	84.8	43	53	2545.4	78.4	36	47	2505.5	105.6	45	49	2527.0	101.4
Asian	352	80	2597.4	83.0	355	81	2603.0	90.7	381	82	2602.2	87.8	370	83	2606.6	86.3	375	79	2599.2	89.8
Hispanic	1,601	38	2498.7	87.3	1,549	40	2505.3	87.6	1,776	39	2502.3	86.4	1,854	40	2503.8	90.5	1,863	41	2504.5	93.1
Pacific Islander	8*				11	73	2533.8	121.2	13	54	2529.1	105.0	13	38	2475.9	134.2	12	83	2587.7	79.2
White	4,694	59	2546.3	88.4	4,615	65	2556.9	87.8	4,458	65	2558.4	87.4	4,647	65	2559.1	89.8	4,573	65	2559.0	90.3
Multi-Racial	223	52	2530.8	84.1	275	50	2536.0	91.3	385	61	2543.1	89.3	418	52	2534.0	95.2	455	51	2535.6	91.4
ELL	247	5	2409.1	72.0	298	7	2416.1	72.1	392	4	2412.5	69.7	492	6	2420.0	76.7	752	11	2442.8	76.9
Special Education	1,389	8	2422.5	75.5	1,418	9	2432.0	76.5	1,483	10	2428.5	74.3	1,574	9	2426.6	78.0	1,697	11	2427.8	80.8
CD 504	416	43	2513.5	84.1	430	47	2525.0	84.2	456	48	2523.5	82.7	510	50	2527.0	80.3	547	50	2523.3	84.7
Title I	1,814	45	2515.8	86.1	1,570	52	2531.8	86.7	1,336	49	2526.0	88.0	1,214	55	2537.6	88.5	1,019	59	2542.8	88.8

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

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

Group	2014–2015				2015–2016				2016–2017				2017–2018				2018-2019			
	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD
Grade 7																				
All Students	9,716	50	2547.1	96.0	10,049	52	2552.7	98.2	10,070	54	2553.7	97.8	10,219	54	2553.5	98.6	10,540	55	2555.4	102.7
Female	4,735	58	2564.4	92.5	4,957	59	2569.4	96.4	4,936	59	2568.0	94.2	5,070	61	2569.1	94.2	5,299	61	2572.4	98.7
Male	4,981	43	2530.7	96.4	5,092	46	2536.5	97.3	5,134	48	2540.0	99.2	5,149	48	2538.1	100.5	5,241	48	2538.2	103.7
African American	3,068	33	2509.3	89.3	3,057	35	2514.1	90.9	3,201	36	2514.9	94.9	3,160	38	2515.3	94.2	3,169	38	2514.7	98.0
AmerIndian/Alaskan	52	50	2553.6	92.6	44	66	2579.5	83.3	45	53	2558.7	87.9	43	60	2578.2	80.6	37	57	2550.3	107.0
Asian	354	81	2621.7	90.9	347	82	2633.1	94.3	358	83	2634.0	92.8	381	83	2626.9	91.8	376	85	2637.3	96.2
Hispanic/Latino	1,453	39	2521.8	90.0	1,642	41	2527.2	95.0	1,604	42	2527.4	90.3	1,770	42	2526.6	93.2	1,880	45	2532.3	95.7
Pacific Islander	8*				10	30	2536.3	101.6	13	54	2571.7	100.6	11	73	2579.5	53.9	11	45	2525.3	131.5
White	4,555	63	2574.7	90.5	4,720	65	2579.8	92.4	4,570	68	2583.8	88.6	4,457	68	2584.1	90.7	4,629	68	2585.7	95.4
Multi-Racial	226	50	2550.1	88.8	229	62	2567.0	86.8	279	51	2553.9	96.5	397	56	2560.3	95.1	438	53	2558.5	98.2
ELL	285	9	2433.3	74.1	292	5	2434.1	69.8	339	7	2435.6	76.9	423	7	2440.8	78.7	534	14	2455.4	81.9
Disadvantaged	1,328	8	2445.8	74.5	1,440	10	2449.5	77.9	1,431	11	2450.3	80.5	1,510	10	2445.6	82.0	1,623	11	2446.9	81.9
Migrant	351	44	2535.6	85.4	453	45	2542.2	88.1	488	50	2549.7	86.8	506	53	2553.3	87.2	570	52	2555.5	95.0
Disability	1,902	50	2542.8	92.1	1,778	52	2550.7	93.7	1,567	53	2550.8	92.4	1,312	55	2557.9	88.7	1,191	60	2567.4	93.5
Grade 8																				
All Students	9,546	49	2559.1	97.9	9,747	54	2569.6	98.1	10,069	52	2566.0	99.7	10,106	53	2568.5	99.3	10,207	52	2566.2	103.5
Female	4,669	56	2576.1	93.7	4,761	61	2588.0	94.2	4,942	60	2585.2	95.6	4,955	60	2586.6	95.0	5,085	58	2582.5	99.3
Male	4,877	43	2542.9	99.1	4,986	47	2552.1	98.5	5,127	45	2547.5	100.0	5,151	46	2551.0	100.2	5,122	46	2550.0	105.1
African American	3,109	33	2521.5	91.2	3,101	38	2533.3	91.2	3,096	36	2528.0	94.5	3,219	37	2531.8	94.9	3,198	36	2528.2	98.2
AmerIndian/Alaskan	38	66	2600.1	92.8	50	56	2579.1	100.8	45	67	2585.3	88.6	47	51	2565.1	92.8	51	47	2557.3	105.0
Asian	328	80	2634.7	92.0	366	80	2642.3	98.9	348	80	2646.3	98.2	368	84	2647.6	97.0	384	82	2643.3	98.0
Hispanic	1,267	38	2533.9	89.7	1,508	43	2542.7	92.7	1,646	42	2543.2	95.4	1,641	43	2541.0	94.4	1,784	41	2539.9	97.0
Pacific Islander	11	64	2597.3	97.3	9*				8*				14	50	2569.4	115.5	11	55	2602.6	90.0
White	4,574	60	2585.2	93.5	4,484	66	2597.9	92.1	4,678	64	2592.3	93.1	4,520	66	2597.7	91.2	4,389	65	2597.2	96.6
Multi-Racial	219	53	2572.6	96.6	229	51	2570.4	95.1	248	60	2578.0	95.5	297	52	2575.5	96.8	390	54	2572.9	102.4
ELL	258	7	2454.2	76.4	329	8	2450.3	77.7	322	8	2457.5	78.8	374	9	2453.4	79.5	451	8	2457.7	80.9
Special Education	1,350	10	2459.7	77.5	1,364	9	2465.4	77.4	1,432	10	2463.8	81.1	1,437	10	2463.7	78.2	1,525	11	2458.3	82.8
CD 504	404	44	2551.3	88.2	381	48	2562.9	85.2	492	48	2554.6	91.7	534	48	2559.8	90.8	559	49	2562.5	93.8
Title I	1,957	42	2545.2	94.4	1,843	54	2566.7	91.8	1,714	52	2565.5	93.4	1,527	54	2570.2	94.5	1,274	54	2568.5	94.7

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

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

Group	2014–2015				2015–2016				2016–2017				2017–2018				2018-2019			
	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD
Grade 3																				
All Students	10,268	53	2439.4	75.5	10,341	55	2444.0	78.6	10,669	53	2441.0	79.4	10,517	54	2441.2	83.1	10,287	53	2439.8	84.7
Female	5,150	53	2439.9	73.3	5,146	54	2443.4	76.8	5,203	53	2441.1	76.4	5,184	53	2440.2	78.8	5,011	53	2440.1	81.6
Male	5,118	53	2438.9	77.6	5,195	56	2444.6	80.3	5,466	54	2440.8	82.1	5,333	54	2442.2	87.1	5,276	53	2439.5	87.6
African American	3,026	36	2408.4	70.8	3,106	39	2411.8	74.4	3,216	36	2409.3	74.1	3,181	37	2406.4	78.3	3,109	35	2403.8	79.0
AmerIndian/Alaskan	38	66	2460.1	68.5	40	50	2442.3	76.1	36	44	2432.9	95.1	43	49	2434.7	68.6	24	50	2428.0	102.1
Asian	391	80	2499.6	75.3	378	87	2509.3	73.0	394	82	2503.1	77.3	427	85	2515.4	84.0	424	84	2512.6	84.0
Hispanic/Latino	1,784	41	2420.2	67.7	1,817	44	2423.8	68.9	2,031	42	2420.1	72.8	1,982	42	2419.7	72.9	1,974	42	2419.0	77.8
Pacific Islander	16	50	2442.6	84.6	13	62	2458.4	82.3	13	77	2481.5	71.7	22	68	2456.0	78.7	9*			
White	4,620	67	2462.0	71.4	4,547	68	2468.2	74.4	4,514	68	2467.0	73.8	4,378	68	2468.6	76.7	4,253	68	2467.5	77.4
Multi-Racial	393	51	2437.3	67.2	440	56	2448.0	72.4	465	56	2445.5	77.6	483	55	2444.8	80.8	494	56	2448.7	81.6
ELL	1,032	25	2395.4	63.5	1,306	35	2410.5	66.2	1,707	40	2416.1	73.6	1,790	43	2420.3	74.6	1,814	41	2415.2	75.9
Disadvantaged	1,280	14	2360.0	72.9	1,335	17	2364.6	78.1	1,441	18	2367.6	76.4	1,441	17	2359.2	81.8	1,549	16	2359.5	78.9
Migrant	333	48	2432.7	67.9	319	49	2438.6	72.1	336	50	2435.3	68.5	343	51	2438.8	71.5	398	51	2439.6	75.8
Disability	1,163	54	2440.8	62.5	1,057	61	2456.1	67.2	1,045	65	2462.2	71.9	1,096	62	2457.0	75.3	1,007	61	2455.3	76.8
Grade 4																				
All Students	9,995	47	2476.9	75.4	10,297	51	2485.1	79.4	10,442	50	2483.3	82.6	10,689	50	2484.4	82.6	10,522	51	2484.1	85.8
Female	4,970	45	2475.6	71.9	5,151	50	2485.1	76.0	5,183	49	2481.9	79.2	5,227	49	2482.9	78.3	5,172	50	2483.1	80.6
Male	5,025	48	2478.1	78.7	5,146	51	2485.0	82.7	5,259	52	2484.8	85.7	5,462	52	2485.9	86.5	5,350	52	2485.2	90.6
African American	3,063	29	2446.5	69.8	3,041	33	2451.7	72.9	3,155	32	2448.6	76.7	3,246	32	2449.0	76.1	3,196	31	2446.5	78.2
AmerIndian/Alaskan	43	56	2495.3	64.7	37	49	2489.0	61.9	41	41	2486.4	74.7	37	51	2485.0	90.3	40	50	2476.1	87.4
Asian	401	78	2539.9	73.2	391	81	2555.0	85.7	398	83	2557.9	79.4	396	83	2556.2	83.9	432	83	2559.1	93.6
Hispanic	1,736	36	2457.0	68.1	1,804	38	2462.7	70.2	1,871	37	2459.4	72.3	2,023	40	2465.7	75.3	2,035	41	2465.1	77.7
Pacific Islander	15	53	2478.0	57.1	14	57	2490.0	79.0	15	67	2513.8	84.3	13	62	2474.6	138.7	18	61	2516.4	89.8
White	4,362	60	2499.4	71.5	4,605	65	2510.1	74.6	4,514	65	2510.5	77.2	4,499	65	2511.6	76.1	4,313	66	2512.8	78.9
Multi-Racial	375	51	2484.8	71.5	405	48	2481.7	72.5	448	51	2487.2	78.5	475	52	2489.3	81.7	488	55	2489.8	82.7
ELL	613	16	2419.9	67.7	683	18	2424.9	65.8	954	22	2432.9	70.7	1,663	37	2458.5	73.5	1,722	38	2457.8	76.3
Special Education	1,355	8	2393.1	66.9	1,450	12	2405.5	68.7	1,479	13	2400.0	75.6	1,626	15	2407.1	76.8	1,700	14	2402.1	76.9
CD 504	377	40	2470.6	66.1	375	47	2478.3	78.1	416	49	2480.9	74.9	420	45	2475.1	72.5	446	53	2486.1	76.3
Title I	1,279	46	2477.8	67.2	1,247	56	2494.2	67.9	1,052	58	2498.5	73.1	1,061	63	2505.5	71.6	1,066	61	2500.1	72.7

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

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

Group	2014–2015				2015–2016				2016–2017				2017-2018				2018-2019			
	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD
Grade 5																				
All Students	10,017	38	2498.6	85.0	10,199	42	2506.8	86.8	10,519	44	2511.5	89.7	10,633	43	2510.4	90.0	10,852	44	2510.8	93.2
Female	4,935	37	2498.8	82.1	5,070	40	2505.5	84.0	5,255	44	2512.5	87.2	5,304	42	2510.4	86.7	5,295	44	2511.6	88.8
Male	5,082	39	2498.3	87.7	5,129	43	2508.0	89.5	5,264	44	2510.4	92.1	5,329	44	2510.4	93.2	5,557	45	2510.0	97.3
African American	3,148	21	2461.0	79.9	3,077	23	2468.6	78.4	3,089	26	2472.9	81.6	3,219	24	2469.3	82.3	3,312	24	2466.0	84.9
AmerIndian/Alaskan	41	34	2499.2	79.6	42	43	2518.5	79.1	31	35	2503.3	82.0	40	40	2512.4	110.6	28	36	2492.9	89.0
Asian	375	74	2573.8	82.2	395	74	2580.0	85.1	378	76	2589.1	87.9	397	78	2595.1	87.9	398	83	2596.3	83.5
Hispanic/Latino	1,565	27	2477.0	75.1	1,787	29	2483.3	79.5	1,861	31	2486.9	81.0	1,909	33	2488.5	80.7	2,044	35	2493.6	84.0
Pacific Islander	10	50	2545.3	83.3	13	54	2518.3	58.4	12	42	2486.6	108.1	11	64	2543.3	102.2	12	67	2533.8	110.0
White	4,602	50	2524.9	78.7	4,484	56	2535.2	81.3	4,706	59	2540.3	84.5	4,574	59	2540.5	82.8	4,542	60	2542.9	86.6
Multi-Racial	276	41	2505.9	77.2	401	43	2512.2	81.7	442	47	2512.3	83.7	483	41	2514.6	85.6	516	45	2517.5	88.1
ELL	346	8	2416.5	70.6	468	8	2426.1	74.1	507	7	2426.1	71.2	952	18	2456.1	77.1	1,308	22	2464.8	75.4
Disadvantaged	1,390	5	2409.4	69.8	1,449	6	2416.0	72.9	1,543	8	2420.6	74.4	1,619	9	2421.7	74.7	1,801	9	2419.2	78.2
Migrant	409	29	2493.9	77.2	423	35	2498.4	73.4	468	37	2509.1	80.7	496	39	2507.8	78.8	555	39	2509.2	79.3
Disability	1,628	38	2500.7	83.3	1,362	45	2512.4	80.0	1,254	48	2521.9	82.6	1,070	50	2526.8	83.9	1,051	53	2528.3	81.8
Grade 6																				
All Students	10,084	34	2510.5	96.3	10,004	37	2516.3	101.8	10,211	41	2523.8	103.5	10,446	40	2521.0	104.8	10,607	38	2514.5	107.2
Female	4,981	35	2515.4	92.5	4,937	37	2519.5	98.3	5,072	42	2527.4	98.6	5,236	42	2527.5	100.5	5,291	38	2518.0	103.3
Male	5,103	33	2505.8	99.7	5,067	37	2513.3	105.0	5,139	40	2520.4	108.0	5,210	38	2514.4	108.5	5,316	37	2511.1	110.8
African American	3,111	17	2470.6	87.7	3,125	21	2474.1	96.1	3,138	22	2479.6	96.2	3,071	24	2477.4	100.5	3,243	20	2467.4	100.3
AmerIndian/Alaskan	48	38	2518.8	89.9	43	28	2510.2	90.9	43	51	2554.0	85.6	35	34	2509.2	93.4	45	38	2524.9	112.4
Asian	358	69	2598.7	94.6	361	70	2606.2	114.1	389	76	2610.8	106.6	374	74	2615.9	106.2	379	74	2620.2	102.6
Hispanic	1,635	22	2486.0	90.2	1,581	24	2487.2	91.2	1,794	29	2496.3	94.1	1,888	28	2495.7	94.3	1,900	27	2489.1	100.2
Pacific Islander	8*				11	45	2535.3	156.1	13	69	2551.2	86.9	14	29	2472.2	122.5	12	58	2556.3	110.4
White	4,701	46	2538.0	90.7	4,607	50	2547.5	92.6	4,447	56	2557.2	95.8	4,646	53	2552.9	96.2	4,574	51	2549.3	96.4
Multi-Racial	223	39	2526.5	87.2	276	36	2523.9	96.2	387	44	2535.2	93.8	418	39	2519.0	97.7	454	37	2517.0	102.8
ELL	291	4	2402.4	84.4	339	4	2402.2	81.9	435	5	2412.8	84.6	543	5	2416.1	88.4	811	8	2433.1	90.8
Special Education	1,405	4	2404.9	82.6	1,414	5	2407.4	91.0	1,478	6	2410.4	92.2	1,557	4	2405.8	95.5	1,698	5	2405.7	92.0
CD 504	417	28	2506.6	83.8	429	32	2513.8	93.4	455	38	2521.4	92.7	510	35	2518.8	89.3	547	36	2516.0	95.1
Title I	1,826	30	2505.4	87.1	1,584	37	2515.5	97.4	1,339	40	2525.2	89.7	1,212	45	2534.1	89.2	1,018	45	2536.3	93.3

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

Table B-6. Mathematics Student Performance in Five Across Years (Grades 7 and 8)

Group	2014–2015				2015–2016				2016–2017				2017-2018				2018-2019			
	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD
Grade 7																				
All Students	9,754	37	2529.6	102.7	10,070	40	2534.5	106.6	10,087	41	2538.7	109.1	10,231	39	2531.4	108.3	10,572	41	2536.2	111.8
Female	4,753	39	2535.1	99.3	4,970	41	2538.6	104.8	4,943	41	2540.9	105.7	5,071	39	2534.2	105.3	5,308	42	2540.5	109.3
Male	5,001	35	2524.4	105.6	5,100	38	2530.4	108.1	5,144	41	2536.6	112.3	5,160	39	2528.6	111.1	5,264	40	2532.0	114.2
African American	3,064	19	2486.7	93.8	3,054	21	2488.0	97.5	3,199	23	2493.0	101.1	3,151	21	2485.8	99.4	3,160	22	2487.1	101.6
AmerIndian/Alaskan	52	29	2529.7	94.4	44	55	2560.0	93.1	45	36	2532.1	82.6	44	39	2536.5	103.6	38	32	2523.5	105.4
Asian	360	71	2622.9	107.9	357	77	2638.9	109.7	362	77	2640.4	117.0	388	79	2631.8	110.1	378	78	2651.4	116.0
Hispanic/Latino	1,490	26	2501.1	97.8	1,667	29	2505.7	103.8	1,636	30	2507.1	102.3	1,809	28	2503.7	102.3	1,923	31	2511.3	105.0
Pacific Islander	8*				10	40	2530.5	96.2	15	47	2550.8	126.0	11	45	2555.0	79.6	11	27	2496.7	152.3
White	4,556	50	2560.2	94.4	4,710	52	2566.2	96.6	4,552	55	2573.8	98.9	4,436	53	2565.7	99.2	4,624	55	2571.2	102.0
Multi-Racial	224	35	2533.7	97.0	228	38	2543.4	94.2	278	41	2546.2	99.1	392	40	2536.4	101.0	438	38	2533.3	107.5
ELL	334	5	2416.2	90.8	339	7	2421.8	96.4	385	6	2422.7	92.2	477	8	2422.9	103.6	603	9	2437.1	97.4
Disadvantaged	1,324	4	2419.1	86.6	1,435	6	2423.2	89.9	1,420	7	2424.4	89.7	1,511	5	2416.1	87.6	1,618	5	2417.3	89.7
Migrant	350	33	2528.2	90.6	450	36	2532.9	91.3	488	38	2540.0	91.8	500	35	2530.9	95.8	565	38	2537.8	101.2
Disability	1,912	33	2521.8	94.3	1,777	39	2534.5	96.7	1,568	42	2540.3	103.8	1,314	41	2537.7	97.8	1,187	48	2551.3	100.8
Grade 8																				
All Students	9,512	35	2541.7	112.0	9,768	38	2548.9	117.0	10,058	38	2550.5	119.7	10,117	39	2548.3	117.9	10,232	38	2546.4	119.1
Female	4,646	36	2547.3	106.6	4,765	41	2557.9	111.0	4,944	41	2560.0	114.4	4,951	41	2555.1	112.6	5,102	40	2553.3	113.8
Male	4,866	35	2536.4	116.6	5,003	35	2540.4	121.8	5,114	35	2541.3	123.9	5,166	37	2541.7	122.5	5,130	36	2539.6	123.8
African American	3,091	17	2491.4	97.3	3,097	20	2500.3	105.4	3,092	21	2498.6	107.5	3,210	23	2499.3	110.2	3,197	20	2497.0	106.3
AmerIndian/Alaskan	38	42	2560.0	120.3	50	42	2549.2	111.4	45	56	2580.2	117.7	48	38	2541.4	110.5	51	35	2554.8	110.3
Asian	329	71	2647.6	116.1	370	74	2658.6	138.8	356	72	2668.3	135.4	373	76	2662.7	125.0	388	75	2658.3	136.1
Hispanic	1,264	27	2516.4	101.0	1,530	25	2517.7	104.4	1,669	29	2526.0	109.6	1,674	27	2517.9	106.0	1,807	27	2517.8	107.6
Pacific Islander	11	36	2572.3	95.8	9*				9*				15	40	2543.7	134.4	11	55	2579.2	81.9
White	4,558	47	2574.5	106.9	4,483	51	2584.0	108.6	4,641	50	2584.1	112.3	4,506	52	2584.5	108.3	4,389	52	2584.0	111.9
Multi-Racial	221	36	2551.6	110.6	229	40	2554.4	112.5	246	38	2560.0	110.5	291	39	2556.9	108.4	389	38	2549.2	112.2
ELL	267	9	2442.1	102.0	367	9	2437.8	99.8	379	10	2452.3	102.2	427	8	2447.9	95.6	493	8	2444.0	98.4
Special Education	1,350	5	2435.1	86.3	1,364	5	2432.0	94.5	1,415	5	2432.4	91.7	1,422	4	2426.2	93.3	1,522	4	2423.7	88.6
CD 504	402	31	2540.6	99.0	382	32	2541.7	101.4	489	31	2538.4	106.9	537	31	2539.0	102.5	559	35	2545.4	105.1
Title I	1,943	30	2531.0	104.4	1,843	33	2536.9	109.6	1,714	38	2551.9	108.0	1,524	39	2549.9	110.5	1,272	38	2549.4	113.8

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

Appendix C: Classification Accuracy and Consistency Indexes by Subgroup

Table C-1. ELA/Lit Classification Accuracy and Consistency by Achievement Level (Grades 3–5)

Group	N	% Accuracy					% Consistency				
		All	L1	L2	L3	L4	All	L1	L2	L3	L4
Grade 3											
All Students	10,234	79	89	70	68	88	71	83	60	57	83
Female	4,995	79	88	71	68	89	71	81	60	57	84
Male	5,239	79	89	70	67	87	72	84	59	57	82
African American	3,107	79	89	70	68	85	71	85	59	57	77
AmerIndian/Alaskan	23	75	92*	67*	71*	75*	67	76*	59*	62*	69*
Asian	420	83	86	70	68	92	76	78	59	57	88
Hispanic	1,924	79	89	71	68	86	70	83	61	57	77
Pacific Islander	9**										
White	4,254	79	87	70	68	89	71	79	59	57	85
Multi-Racial	497	79	89	71	66	90	71	80	61	57	83
ELL	1,750	78	89	71	68	83	70	83	61	57	72
Special Education	1,555	85	93	69	67	84	79	90	58	55	71
CD 504	398	78	86	71	68	89	69	79	61	57	81
Title I	1,002	78	86	70	67	89	70	77	60	57	83
Grade 4											
All Students	10,468	78	89	63	65	88	70	83	51	55	81
Female	5,148	77	88	63	65	88	69	81	51	55	82
Male	5,320	78	90	63	65	88	71	85	51	55	81
African American	3,193	78	90	63	65	85	70	85	52	55	75
AmerIndian/Alaskan	40	77	92*	67	63	91	69	78*	58	54	82
Asian	417	83	92	64	64	92	77	79	50	55	89
Hispanic	1,996	77	89	64	65	83	68	84	51	55	75
Pacific Islander	18	78	93*	0*	67*	80*	70	94*	17*	58*	77*
White	4,312	78	88	63	65	89	70	79	51	56	83
Multi-Racial	492	77	89	64	66	88	69	82	51	57	81
ELL	1,651	76	89	64	65	81	68	84	51	56	69
Special Education	1,707	84	93	63	67	82	78	91	51	54	69
CD 504	448	76	87	64	66	89	68	79	54	55	80
Title I	1,059	75	87	63	66	87	66	77	53	56	79
Grade 5											
All Students	10,827	79	89	67	74	86	71	83	55	66	79
Female	5,282	79	88	67	75	86	71	80	55	66	80
Male	5,545	79	90	67	74	86	72	85	55	65	78
African American	3,309	79	91	67	74	83	72	85	56	65	72
AmerIndian/Alaskan	28	75	88*	64*	76	75*	66	80*	56*	66	63*
Asian	392	83	84	69	73	90	76	78	54	62	87
Hispanic	2,021	78	89	66	75	83	69	82	56	67	72
Pacific Islander	12	77	99*	64*	77*	73*	71	94*	56*	61*	75*
White	4,548	79	87	67	75	87	71	79	54	66	81
Multi-Racial	517	79	88	67	74	89	71	80	57	65	82
ELL	1,264	78	90	67	74	77	70	83	57	65	55
Special Education	1,802	84	92	66	73	81	78	90	55	63	65
CD 504	555	77	87	66	75	83	68	79	54	67	75
Title I	1,050	78	88	68	75	87	70	79	57	67	78

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

Table C-2. ELA/Lit Classification Accuracy and Consistency by Achievement Level (Grades 6–8)

Group	N	% Accuracy					% Consistency				
		All	L1	L2	L3	L4	All	L1	L2	L3	L4
Grade 6											
All Students	10,572	80	89	72	76	85	72	83	62	69	76
Female	5,281	79	88	72	76	85	71	80	62	69	77
Male	5,291	80	90	72	76	85	73	85	62	69	74
African American	3,249	80	90	72	76	81	73	85	62	68	68
AmerIndian/Alaskan	45	80	97*	70	75	85*	72	86*	65	65	75*
Asian	375	81	85	72	76	88	73	70	61	67	83
Hispanic	1,863	80	90	72	76	83	72	85	62	69	69
Pacific Islander	12	80	56*	80*	82*	82*	70	54*	53*	76*	69*
White	4,573	79	88	72	77	86	71	79	62	69	78
Multi-Racial	455	79	88	72	78	86	71	79	65	69	75
ELL	752	83	91	72	73	77*	77	88	63	59	53*
Special Education	1,697	86	93	72	75	81	80	90	61	63	60
CD 504	547	79	88	72	77	82	70	81	61	70	71
Title I	1,019	79	89	72	77	85	71	80	62	70	74
Grade 7											
All Students	10,540	80	90	70	78	85	72	84	59	71	76
Female	5,299	80	89	70	78	85	72	82	59	70	77
Male	5,241	81	91	70	78	85	73	85	60	71	75
African American	3,169	81	91	71	78	80	73	86	61	70	66
AmerIndian/Alaskan	37	78	87	69*	73	86*	71	82	53*	70	73*
Asian	376	84	88	69	78	89	77	83	52	69	87
Hispanic	1,880	79	89	70	77	82	71	83	60	70	71
Pacific Islander	11	83	95*	65*	61*	83*	78	91*	48*	57*	88*
White	4,629	80	89	70	78	86	72	81	58	71	78
Multi-Racial	438	79	88	69	78	87	71	80	60	71	76
ELL	534	84	93	70	74	90*	78	89	59	65	48*
Special Education	1,623	86	93	70	76	79	80	90	60	64	61
CD 504	570	79	90	70	77	86	71	81	61	70	77
Title I	1,191	80	90	70	78	85	71	81	59	71	76
Grade 8											
All Students	10,207	80	89	72	79	83	73	82	62	72	75
Female	5,085	80	88	73	79	83	72	80	63	72	76
Male	5,122	81	89	72	79	83	73	84	62	71	74
African American	3,198	80	90	72	78	80	73	84	62	71	68
AmerIndian/Alaskan	51	81	90	70	82	84*	73	83	62	73	76*
Asian	384	83	89	73	79	87	76	82	59	71	84
Hispanic	1,784	80	89	73	79	81	72	83	63	72	68
Pacific Islander	11	75	0*	67*	77*	88*	67	41*	65*	63*	82*
White	4,389	80	88	72	79	84	72	79	62	72	76
Multi-Racial	390	81	88	73	80	85	73	80	64	72	79
ELL	451	85	92	71	78	79*	80	89	62	63	64*
Special Education	1,525	86	92	72	77	78*	80	90	62	66	52*
CD 504	559	79	85	72	79	83	71	77	63	72	74
Title I	1,274	80	89	73	79	83	72	81	63	72	72

*The classification index is based on n < 10.

Table C-3. Mathematics Classification Accuracy and Consistency by Achievement Level (Grades 3–5)

Group	N	% Accuracy					% Consistency				
		All	L1	L2	L3	L4	All	L1	L2	L3	L4
Grade 3											
All Students	10,287	83	90	73	79	90	76	85	63	71	85
Female	5,011	83	90	73	79	90	76	84	63	71	85
Male	5,276	84	91	74	79	91	77	86	64	72	85
African American	3,109	83	91	74	79	87	76	87	64	70	79
AmerIndian/Alaskan	24	79	86*	58*	65*	89*	73	87*	47*	56*	83*
Asian	424	88	91	77	78	95	83	84	62	71	92
Hispanic	1,974	82	91	73	79	89	75	85	64	71	81
Pacific Islander	9**										
White	4,253	84	88	74	80	91	77	81	63	72	87
Multi-Racial	494	83	88	73	79	89	76	83	64	70	85
ELL	1,814	82	91	73	79	88	75	85	63	72	79
Special Education	1,549	88	94	73	79	89	82	92	63	70	77
CD 504	398	82	86	74	79	91	75	80	64	73	85
Title I	1,007	83	89	73	79	91	76	82	64	71	86
Grade 4											
All Students	10,522	84	89	80	79	90	77	83	73	71	84
Female	5,172	83	89	80	79	89	76	82	73	71	84
Male	5,350	84	90	80	79	90	77	84	72	71	85
African American	3,196	83	90	79	78	86	76	85	73	70	77
AmerIndian/Alaskan	40	86	92*	85	82	88*	80	86*	82	76	77*
Asian	432	89	92	78	79	95	84	83	69	72	93
Hispanic	2,035	83	89	80	78	87	75	83	72	71	79
Pacific Islander	18	87	79*	82*	84*	99*	80	81*	69*	77*	93*
White	4,313	84	87	81	79	91	77	79	73	72	86
Multi-Racial	488	83	87	80	79	87	76	81	71	71	83
ELL	1,722	82	89	79	78	84	75	83	72	70	77
Special Education	1,700	86	92	80	76	87	81	89	72	67	75
CD 504	446	81	84	78	77	89	74	77	69	72	82
Title I	1,066	82	87	80	78	89	75	78	72	71	83
Grade 5											
All Students	10,852	83	90	78	71	91	76	85	69	61	86
Female	5,295	82	89	78	71	90	75	84	69	61	85
Male	5,557	84	91	77	72	91	77	86	69	61	87
African American	3,312	84	92	78	71	86	77	88	69	59	80
AmerIndian/Alaskan	28	80	81	72*	71*	91*	73	79	55*	64*	89*
Asian	398	87	88	79	71	95	82	79	66	62	93
Hispanic	2,044	82	90	77	72	89	75	84	69	61	82
Pacific Islander	12	82	90*	61*	71*	87*	75	85*	47*	63*	81*
White	4,542	83	89	78	71	91	76	82	68	62	87
Multi-Racial	516	83	88	78	71	93	76	82	70	62	86
ELL	1,308	82	90	77	70	84	75	85	68	60	74
Special Education	1,801	89	94	76	71	88	84	92	66	58	79
CD 504	555	81	90	78	72	89	74	82	72	60	82
Title I	1,051	82	89	79	71	90	74	82	70	61	84

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

Table C-4. Mathematics Classification Accuracy and Consistency by Achievement Level (Grades 6–8)

Group	N	% Accuracy					% Consistency				
		All	L1	L2	L3	L4	All	L1	L2	L3	L4
Grade 6											
All Students	10,607	83	92	77	72	89	76	87	69	61	84
Female	5,291	83	91	77	72	89	76	86	70	61	83
Male	5,316	83	92	77	71	90	77	88	69	61	84
African American	3,243	85	93	76	72	86	79	89	69	61	76
AmerIndian/Alaskan	45	83	95	80	66*	89*	76	81	77	56*	82*
Asian	379	86	83	78	72	94	81	73	71	60	92
Hispanic	1,900	83	91	78	72	86	76	88	70	61	77
Pacific Islander	12	84	92*	80*	68*	94*	77	90*	66*	57*	94*
White	4,574	82	89	78	71	89	74	82	70	61	85
Multi-Racial	454	82	89	76	72	89	75	85	66	61	84
ELL	811	88	94	77	73	81	83	91	70	55	70
Special Education	1,698	91	95	78	71	80	87	94	67	57	65
CD 504	547	81	90	77	71	87	74	84	68	61	80
Title I	1,018	81	91	78	72	87	74	83	71	61	81
Grade 7											
All Students	10,572	83	91	75	74	90	76	86	67	65	85
Female	5,308	82	90	75	74	91	75	85	66	65	85
Male	5,264	83	92	76	75	90	77	87	67	65	85
African American	3,160	84	92	75	74	86	77	88	67	64	78
AmerIndian/Alaskan	38	82	88	70	79*	96*	76	87	65	66*	80*
Asian	378	88	89	74	74	96	83	80	65	66	94
Hispanic	1,923	83	91	76	73	87	76	87	67	63	81
Pacific Islander	11	86	91*	80*	0*	83*	80	90*	67*	25*	85*
White	4,624	82	89	75	74	91	75	82	67	66	86
Multi-Racial	438	82	89	74	75	90	75	85	65	64	83
ELL	603	89	94	75	75	89	84	92	66	62	77
Special Education	1,618	90	95	72	71	89	86	93	62	57	77
CD 504	565	81	89	74	76	88	73	81	66	66	82
Title I	1,187	81	88	75	74	89	74	83	65	66	83
Grade 8											
All Students	10,232	82	90	71	71	90	75	86	62	60	85
Female	5,102	81	90	72	71	90	74	85	62	60	84
Male	5,130	83	91	71	71	90	76	87	61	60	85
African American	3,197	83	92	71	71	87	77	88	62	59	79
AmerIndian/Alaskan	51	79	85	68	64*	90	72	84	58	52*	84
Asian	388	87	93	70	73	94	82	88	61	61	92
Hispanic	1,807	82	91	72	70	89	75	87	62	60	79
Pacific Islander	11	79	94*	72*	82*	70*	68	87*	62*	68*	56*
White	4,389	80	88	72	71	90	73	81	62	61	85
Multi-Racial	389	80	89	72	69	87	73	84	63	58	82
ELL	493	90	94	74	74	86	85	93	60	59	76
Special Education	1,522	91	95	70	72	87	87	94	58	52	75
CD 504	559	81	88	73	71	92	73	82	63	62	83
Title I	1,272	81	90	71	71	90	73	84	62	60	84

*The classification index is based on n < 10.