Delaware Smarter Balanced Assessments

2017–2018 Technical Report

Addendum to the 2017–18 Smarter Balanced Technical Report

Submitted to
Delaware Department of Education
by American Institutes for Research

TABLE OF CONTENTS

1. OVERVIEW 1
2. TEST ADMINISTRATION 3
2.1 Testing Windows 3
2.2 Test Options and Administrative Roles 3
2.2.1 Administrative Roles 4
2.2.2 Online Test Administration 6
2.2.3 Paper-Pencil Test Administration 7
2.2.4 Braille Test Administration 7
2.3 Training and Information for Test Coordinators and Administrators 8
2.3.1 Practice and Training Site 9
2.3.2 Manuals and User Guides 10
2.3.3 Training Modules 11
2.4 Test Security 11
2.4.1 DeSSA Test Security Manual 12
2.4.2 Student-Level Testing Confidentiality 12
2.4.3 System Security 13
2.4.4 Security of the Testing Environment 14
2.4.5 Test Security Violations 15
2.4.6 Monitoring Test Administration 15
2.5 Student Participation 16
2.5.1 Homeschooled Students 16
2.5.2 Student Exemptions 16
2.6 Online Testing Features and Accommodations 16
2.6.1 Online Universal Tools for All Students 17
2.6.2 Designated Supports and Accommodations 20
2.7 Data Forensics Program 32

2.7.1 Data Forensics Report *32*
2.7.2 Changes in Student Performance *32*
2.7.3 Item Response Time *33*
2.7.4 Inconsistent Item Response Pattern *33*
2.8 Prevention and Recovery of Disruptions in Test Delivery System 34
2.8.1 High-Level System Architecture *35*
2.8.2 Automated Backup and Recovery *36*
2.8.3 Other Disruption Prevention and Recovery *37*
3. SUMMARY OF 2017–2018 OPERATIONAL TEST ADMINISTRATION 38
3.1 Student Population 38
3.2 Summary of Overall Student Performance 38
3.3 Test-Taking Time 49
3.4 Distribution of Student Ability and Item Difficulty Distribution 51
4. VALIDITY 54
4.1 Evidence on Test Content 54
4.2 Evidence on Internal Structure 58
5. RELIABILITY 61
5.1 Marginal Reliability 61
5.2 Standard Error Curves 62
5.3 Reliability of Achievement Classification 65
5.4 Reliability for Subgroups 70
5.5 Reliability for Claim Scores 71
6. SCORING 73
6.1 Estimating Student Ability Using Maximum Likelihood Estimation 73
6.2 Rules for Transforming Theta to Vertical Scale Scores 74
6.3 Lowest/Highest Obtainable Scores (LOSS/HOSS) 75
6.4 Scoring All Correct and All Incorrect Cases 75
6.5 Rules for Calculating Strengths and Weaknesses for Claim Scores 75
6.6 Target Scores 76
6.7 Handscoring 77

6.7.1 Reader Selection *77*
6.7.2 Reader Training *78*
6.7.3 Reader Statistics *79*
6.7.4 Reader Monitoring and Retraining *80*
6.7.5 Reader Validity Checks *81*
6.7.6 Reader Dismissal *81*
6.7.7 Reader Agreement *81*
7. REPORTING AND INTERPRETING SCORES 84
7.1 Online Reporting System for Students and Educators 84
7.1.1 Types of Online Score Reports *84*
7.1.2 Online Reporting System *86*
7.2 Paper Family Score Reports 99
7.3 Interpretation of Reported Scores 101
7.3.1 Scale Score *101*
7.3.2 Standard Error of Measurement *101*
7.3.3 Achievement Level *101*
7.3.4 Performance Category for Claims *102*
7.3.5 Performance Category for Targets *102*
7.3.6 Aggregated Score *102*
8. QUALITY CONTROL PROCEDURES 104
8.1 Adaptive Test Configuration 104
8.1.1 Platform Review *104*
8.1.2 User Acceptance Testing and Final Review *105*
8.2 Quality Assurance in Document Processing 105
8.3 Quality Assurance in Data Preparation 105
8.4 Quality Assurance in Handscoring 105
8.4.1 Double Scoring Rates, Agreement Rates, Validity Sets, and Ongoing Read-Behinds *105*
8.4.2 Handscoring QA Monitoring Reports *106*
8.4.3 Monitoring by State Department of Education *106*
8.4.4 Identifying, Evaluating, and Informing the State on Alert Responses *106*
8.5 Quality Assurance in Test Scoring 107

8.5.1 Score Report Quality Check *108*
REFERENCES 110
APPENDICES 111

LIST OF TABLES

Table 1. 2017–2018 Testing Windows 3

Table 2. Test Options in 2017–2018 3

Table 3. Number of Students who Took Paper-Pencil Tests in 2017–2018 Summative Test Administration 7

Table 4. Smarter Balanced Assessment Training Requirements 9

Table 5. Manuals and User Guides 10

Table 6. Smarter Balanced-Developed Training Modules 11

Table 7. Universal Tools, Designated Supports, and Accommodations in 2017–2018 25

Table 8. Students with Embedded and Non-Embedded Accommodations in ELA/Lit 26

Table 9. Students with Embedded Designated Supports in ELA/Lit 27

Table 10. Students with Non-Embedded Designated Supports in ELA/Lit 28

Table 11. Students with Embedded and Non-Embedded Accommodations in Mathematics 29

Table 12. Students with Embedded Designated Supports in Mathematics 30

Table 13. Students with Non-Embedded Designated Supports in Mathematics 31

Table 14. Number of Students in Summative ELA/Lit Assessment 38

Table 15. Number of Students in Summative Mathematics Assessment 38

Table 16. ELA/Lit Percentage of Students in Achievement Levels for Overall and by Subgroup (Grades 3–5) 40

Table 17. ELA/Lit Percentage of Students in Achievement Levels for Overall and by Subgroup (Grades 6–8) 41

Table 18. Mathematics Percentage of Students in Achievement Levels for Overall and by Subgroup (Grades 3–5) 42

Table 19. Mathematics Percentage of Students in Achievement Levels for Overall and by Subgroup (Grades 6–8) 43

Table 20. ELA/Lit Percentage of Students in Performance Categories by Claim 48

Table 21. Mathematics Percentage of Students in Performance Categories by Claim 49

Table 22. ELA/Lit Test-Taking Time 50

Table 23. Mathematics Test-Taking Time 51

Table 24. ELA/Lit Percentage of Delivered Tests Meeting Blueprint Requirements for Each Claim and the Number of Passages Administered 55

Table 25. ELA/Lit Percentage of Delivered Tests Meeting Blueprint Requirements for Depth of Knowledge and Item Type 55

Table 26. Mathematics Percentage of Delivered Tests Meeting Blueprint Requirements for Each Claim and Targets (Grades 3–5) 56

Table 27. Mathematics Percentage of Delivered Tests Meeting Blueprint Requirements for Each Claim and Targets (Grades 6–8) 57

Table 28. Average and Range of the Number of Unique Targets Assessed Within Each Claim Across All Delivered Tests 58

Table 29. Correlations Among Claim Scores for ELA/Lit 59

Table 30. Correlations Among Claim Scores for Mathematics 60

Table 31. Marginal Reliability for ELA/Lit and Mathematics 62

Table 32. Average Conditional Standard Error of Measurement by Achievement Level 65

Table 33. Average Conditional Standard Error of Measurement at Each Achievement Level Cut and Difference of the SEMs Between Two Cuts 65

Table 34. Classification Accuracy and Consistency by Achievement Level 69

Table 35. ELA/Lit Marginal Reliability Coefficients for Overall and by Subgroup 70

Table 36. Mathematics Marginal Reliability Coefficients for Overall and by Subgroup 70

Table 37. ELA/Lit Marginal Reliability Coefficients for Claim Scores 71

Table 38. Mathematics Marginal Reliability Coefficients for Claim Scores 72

Table 39. Vertical Scaling Constants on the Reporting Metric 74

Table 40. Cut Scores in Scale Scores 75

Table 41. ELA/Lit Reader Agreements for Short-Answer Items 82

Table 42. ELA/Lit Reader Agreements for Full-Write Items 82

Table 43. Mathematics Reader Agreements 83

Table 44. Types of Online Score Reports by Level of Aggregation 85

Table 45. Types of Subgroups 85

Table 46. Overview of Quality Assurance Reports 108

LIST OF FIGURES

Figure 1. ELA/Lit % Proficient Across Years 44
Figure 2. Mathematics % Proficient Across Years 45
Figure 3. ELA/Lit Average Scale Score Across Years 46
Figure 4. Mathematics Average Scale Score Across Years 47
Figure 5. Student Ability–Item Difficulty Distribution for ELA/Lit 52
Figure 6. Student Ability–Item Difficulty Distribution for Mathematics 53
Figure 7. Conditional Standard Error of Measurement for ELA/Lit 63
Figure 8. Conditional Standard Error of Measurement for Mathematics 64

LIST OF EXHIBITS

Exhibit 1. Home Page: State Level 86
Exhibit 2. Home Page: District Level 87
Exhibit 3. Subject Detail Page for ELA/Lit by Gender: District Level 88
Exhibit 4. Claim Detail Page for Mathematics by ELL: District Level 89
Exhibit 5. Target Detail Page for ELA/L: School Level 90
Exhibit 6. Target Detail Page for ELA/L: Roster Level 91
Exhibit 7. Target Detail Page for Mathematics: School Level 92
Exhibit 8. Target Detail Page for Mathematics: Roster Level 93
Exhibit 9. Trend Report for ELA/L: District Level 94
Exhibit 10. Student Detail Page for ELA/Lit 96
Exhibit 11. Student Detail Page for Mathematics 97
Exhibit 12. Participation Rate Report at District Level 98
Exhibit 13. Sample Paper Family Score Report for Grade 4 ELA/Lit 99
Exhibit 14. Sample Paper Family Score Report for Grade 4 Mathematics 100

LIST OF APPENDICES

Appendix A Summary of the 2017–2018 Interim Assessments
Appendix B Student Performance Across Four Years for All Students and by Subgroup
Appendix C Classification Accuracy and Consistency Indexes by Subgroup

1. OVERVIEW

The Smarter Balanced Assessment Consortium (SBAC) developed a next-generation assessment system. The assessments are designed to measure the Common Core State Standards (CCSS) in English language arts/literacy (ELA/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 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 define the knowledge and skills that students need to succeed in college and careers after graduating from high school. These standards include rigorous content and application of knowledge through higher-order skills and align with college and workforce expectations.

Since the adoption of the CCSS in 2010, the Delaware Department of Education fully implemented the CCSS in all grade levels in SY 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 SY 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.

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:** An online adaptive test that provides an individualized assessment for each student
- **Performance Task:** 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.

This report provides a technical summary of the 2017–2018 summative assessments in ELA/lit and mathematics administered in grades 3–8 under the Delaware Smarter Balanced assessments. The report includes eight chapters: overview, test administration, summary of 2017–2018 operational test administration, validity, reliability, scoring, reporting and interpreting scores, and quality control procedures. The data included in this report are based on Delaware data for the summative assessments only. For the interim assessments, the number of students who took ICAs and IABs and their performance are provided in Appendix A.

While this report includes information on all aspects of the technical quality of the Smarter Balanced test administration in Delaware, it is an addendum to the Smarter Balanced technical report. The information on item and test development, item content review, field-test administration, item data review, item calibrations, content alignment study, standard setting, and other validity information is included in the Smarter Balanced technical report.

Smarter Balanced produces a technical report for the Smarter Balanced assessments, including all aspects of the technical qualities for the Smarter Balanced assessments described in the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], and 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. The Smarter Balanced technical report includes information using the data at the consortium level, combining data from the consortium states.

2. TEST ADMINISTRATION

2.1 TESTING WINDOWS

The 2017–2018 Delaware Smarter Balanced assessment testing window spanned approximately three months for grades 3–8 for the online summative assessments and spanned the full school year for the interim assessments. The paper-pencil, fixed-form summative assessments were administered during 15 days of the online summative testing window. Table 1 shows the schedule for the 2017–2018 Smarter Balanced assessments.

Table 1. 2017–2018 Testing Windows

Tests	Grades	Start Date	End Date	Mode
Summative Assessments	3–8	03/07/2018	05/31/2018	Online Adaptive
	3–8	04/25/2018	05/11/2018	Paper Fixed-Form
Interim Comprehensive Assessments	3–8	08/28/2017	07/17/2018	Online Fixed-Form
Interim Assessment Blocks	3–8	08/28/2017	07/17/2018	Online Fixed-Form

2.2 TEST OPTIONS AND ADMINISTRATIVE ROLES

Smarter Balanced English language arts/literacy (ELA/lit) and mathematics assessments are administered primarily online. To ensure that all eligible students in tested grades were given the opportunity to take the Smarter Balanced assessments, a number of assessment options were available for the 2017–2018 administration to accommodate students' special needs. Table 2 lists the testing options that were offered in 2017–2018. Testing options are selected by content area. Once an option is selected, it applies to all tests of each content area.

Table 2. Test Options in 2017–2018

Assessment	Test 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 day(s) of testing.

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 (DTC)

DTCs are responsible for coordinating testing in their district. 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 properly administer the Smarter Balanced assessments
- 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 DDOE via KACE/DOE Help Desk (the DeSSA request system)

District Accommodations Manager (DAM)

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

- Complete District Accommodations Manager 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 (STC)

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 prior to testing.
- Ensure that all TAs understand and follow the protocols for student relocation.
- Ensure that all students in Department of Services for Children, 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 (TA)

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 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, and 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. Prior to 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. 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 student identifier (SSID), 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 feature(s) (see Section 2.6 for a list of accommodations). Students can begin testing only when the TA confirms the settings.

Once the assessment is started, students must answer all of 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 submitting the assessment. During an active CAT session, if a student reviews and changes the response to a previously answered item, all of the following items to which the student already responded remain the same. No new items are assigned to this student because of changing one or more than one response. For example, a student paused for 10 minutes after completing item 10. After the pause, the student went back to item 5 and changed the 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 2017–2018 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. Additional orders may be entered into TIDE by the DTC after the initial order is received by the school district. 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 document and score the test.

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

Table 3. Number of Students who Took Paper-Pencil Tests
in 2017–2018 Summative Test Administration

Subject	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Total
ELA/Lit	23	31	33	8	10	12	117
Mathematics	24	31	32	8	11	10	116

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 2017–2018 test administration, Smarter Balanced added the Braille Hybrid Adaptive Test (Braille HAT) for mathematics. The Braille HAT consists of a fixed-form segment, a computer-adaptive segment, and a fixed-form PT. The fixed-form segment includes items with tactile graphics which can be embossed at the testing location or received as a package of pre-embossed materials through the

DDOE. All items on the Braille HAT can be presented to the students using a Refreshable Braille Display (RBD).

The braille interface is described below 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 a RBD, a 40-cell RBD is recommended. The summative ELA/lit is presented to the student with items in either contracted or un-contracted literary braille (for items containing only text) and via a braille embosser (for items with tactile or spatial components that cannot be read by an RBD).

Before administering the online summative assessments using the braille interface, 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 4 and read the applicable manuals relevant to their roles. Table 4 presents the training requirements based on roles.

Table 4. Smarter Balanced Assessment Training Requirements

Role	Required Training	Course Number	Components of the Required Training	Estimated Time to Complete
All Roles	DeSSA Entry Training	24246	• Test Security • DeSSA Overview • TA Interface • Student Interface	• 30 min • 30 min • 15 min • 30 min
All Roles	Optional Training: —Introduction to TIDE	25493	• TIDE Training	• 40 min
Smarter Balanced Summative Test Administrator	Smarter Balanced Summative TA Training	26660	• Smarter Balanced Summative TA Training	• 30 min
District Test Coordinator (DTC), School Test Coordinators (STC)	DeSSA District and School Test Coordinator Training	26661	• TIDE Training • ORS Training • Smarter Balanced Interim TA Training • THSS Training	• 30 min • 35 min • 30 min • 30 min
Smarter Balanced Interim Test Administrator	Smarter Balanced Interim TA Training	26401	• Smarter Balanced Interim TA Training • THSS Training • AVA Training • AIRWays Training	• 30 min • 30 min • 5 min • 30 min
Staff Performing Accommodations Data Entry	District and School Accommodations Manager Training	24250	• District and School Accommodations Manager Training	• 25 min
Special Education Staff/ Coordinator, English Language Learners Staff/Coordinator, General Education with Supports Staff/Coordinator	Accessibility Coordinator Training	26484	• DeSSA Overview • Accessibility	• 30 min • 50 min
Secretaries, Administrative Support	Security Training	26402	• Security module only	• 30 min
TAs who are giving paper-pencil assessment only* (if TA is giving online and paper-pencil assessments, take these and the online requirements)	DeSSA Paper-Pencil TA Training for Smarter Balanced	26662	• Paper-Pencil TA Training • Security Training • DeSSA Overview	• 20 min • 30 min • 30 min
Students and Educators	Optional Training – Student Training	24484	• Let's Talk Universal Tools • What is a CAT? • Student Interface	• 30 min • 20 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 2017, separate training sites were opened for TAs and students. TAs can practice administering an assessment by doing 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 are able to practice with a variety of question types and levels of difficulty (approximately 30 items each in mathematics and ELA/lit), as well as 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 bands (grades 3–5 and 6–8), with each test containing five to 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 5 are available on the DeSSA portal at http://de.portal.airast.org.

Table 5. Manuals and User Guides

Resource	Description
Test Information Distribution Engine User Guide	The Test Information Distribution Engine (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.
Test Administrator User Guide	The *Test Administrator (TA) User Guide* supports individuals using the test delivery system applications to manage testing for students participating in the summative assessment. This resource provides information about the test delivery system, 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 test administration manual (TAM) provides the necessary information regarding policies and procedures for the Smarter Balanced English language arts/literacy 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/literacy 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.

Resource	Description
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/literacy and mathematics interim 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, as well as how each system works. All modules are provided as PowerPoint presentations; two modules include narration. Table 6 lists the training modules.

Table 6. 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 test delivery system (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 test delivery system, 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 computer-adaptive test, or 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 test delivery system 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 2017–2018 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 15, 2017. 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 prior to 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 only in accordance with their user rights.

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 only be accessed by authorized individuals with an appropriate need-to-know status.

All students, including homeschooled students, must be enrolled or registered at their testing schools in order to take the online, paper-pencil, or braille assessments. Student enrollment information, including demographic data, is generated using a DDOE file and uploaded nightly via a secure file transfer site to the online test delivery system 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) is not altered in any way, that the data source is known, and that any service can only be performed by a specific, designated user.

A hierarchy of control: As described in Section 2.2.1, DTCs, STCs, and TAs have well-defined roles and levels of access to the online test delivery system.

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, it must be collected by the TA and returned 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 in order 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, student[s] 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 2016–2017 test administration of the Smarter Balanced assessments was intended to improve test administration and monitoring for the 2017–2018 test administration. The Office of Assessment at the 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 prior to 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 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 (SWD) and ELLs, the process for identifying those who need accommodations, and the selection and provision of the appropriate accommodation(s) 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: (a) testing without accommodation(s) and supports; (b) testing without accommodation(s) but with supports; (c) testing with accommodation(s) but without supports; and (d) testing with accommodation(s) 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 online test delivery system, whereas non-embedded resources are provided outside of that system.

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

All of 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 test delivery system for a student who may be distracted by the ability to access a specific tool during a test session.

For additional information about the availability of designated supports and accommodations, refer to the Smarter Balanced *Usability, Accessibility, and Accommodations Guidelines* at https://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 test delivery system. Universal tools are available to all students based on their preference and selection and have been preset in TIDE. In the 2017–2018 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* at 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 (for example, for 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: 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: 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: 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 test administrator 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 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 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 test administrator and/or interpreter.

Refocus: The student's attention can be refocused on the test with use of intermittent prompts, including verbal, picture symbol, signed, cued speech, or physical. 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. Graph paper is required beginning in sixth grade and can be used on all mathematics assessments. A student can use an assistive technology device for scratch paper as long as the device is consistent with the child's IEP and acceptable to the member.

> **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 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 performance task testing session. Once the student completes the performance task, 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 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 morning).

Additional Non-Embedded Universal Tool options include *modified lighting*, *specialized equipment or furniture,* 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 supports and accommodations.

Embedded Designated Supports

Color choices (computer): Enable students to adjust screen background or font color based on student needs or preferences. This may include reversing the colors for the entire interface or choosing the color of font and background. Black on white, reverse contrast, black on rose, medium gray on light gray, and yellow on blue were offered for the online assessments. The test administrator 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.

Mouse pointer: This embedded support 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 prior to testing.

Permissive mode: Permissive mode must be selected if accommodations requiring additional software are to be used (i.e. Speech to Text software, ZoomText [magnification] software, or other software to support Alternate Response accommodations).

Streamlined mode: An alternate, more linear display of item and stimuli. 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.

Translated test directions (for mathematics): Translation of test directions is a language support available before beginning the actual test items. Students can see test directions in another language. As an embedded designated support, translated test directions are automatically a part of the stacked translation designated support. Students who have limited English language skills can use the translated directions support. This support should only be used for students who are proficient readers in the other language and not proficient in English.

Translations (glossaries) for mathematics: Translated glossaries are provided for selected construct-irrelevant terms for mathematics. Translations for these terms appear on the computer screen when students click on them. Students can also select the audio icon next to the glossary term and listen to the audio recording of the glossary. This is available for the following languages and dialects: Arabic, Cantonese, Ilokono, Korean, Mandarin, Punjabi, Russian, Tagalog, Ukrainian, Vietnamese, and Spanish.

Translations (Spanish stacked) for mathematics: Stacked translations are a language support available for some students; they provide the full translation above the original item in English for each test item.

Zoom: 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 the Test Information Distribution Engine (TIDE), or set by the test administrator prior to 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: Disabling of any universal accessibility tools that might be distracting or which students do not need to use, or are unable to use. Tools must be turned off one by one by the TA 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: An exemption from the ELA/lit tests. Students are eligible if, as of the final date of the testing window, the student has 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.

Interpreter – native language: Provide a native language translator to translate test questions (including multiple-choice options) into 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. This support must be approved by DDOE.

Interpret/translate orally—directions only: Provide native language/visual communication translator to translate directions only into the native language. The native language/test administrator must be proficient in the native language.

Magnification: The size of specific areas of the screen (e.g., text, formulas, tables, graphics, and navigation buttons) may be adjusted by the student 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/or mouse pointer designated support.

Noise buffer: These include ear mufflers, white noise machines, and/or 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.

Read-aloud items (for mathematics items and ELA/lit 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.

Read-aloud 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.

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 human who records what they dictate verbatim. 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.

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, 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 (DOE approved): Support or accommodation not listed in these guidelines by Smarter Balanced. Available by application only.

WhisperPhone®: A school-provided tool which students may use to read the test to themselves.

Embedded Accommodations

American Sign Language (ASL): For ELA/lit listening items and mathematics items. ASL human signer and the signed test content are viewed on the same screen. Students may view portions of the ASL video as often as needed.

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

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

Print on request: Paper copies of either passages/stimuli and/or items are printed for students. The student may request one or more test questions to 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 (TA must approve each print request).

Text-to-speech (for ELA/lit reading passages): Text is read aloud to the student via embedded text-to-speech technology. The student can control the speed and raise or lower the volume of the voice via a volume control. This accommodation may only be activated by DOE.

Non-Embedded Accommodations

100s Number Table (grades 4 and above mathematics tests): A paper-pencil table listing of numbers from 1–100 will be 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.

Calculator (for grades 6–8 mathematics tests): 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.

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 test administrator: Using physical assistance, such as direct assistance with turning pages, recording answers for the paper-pencil test (scribing), or navigating in electronic format from a test administrator.

Scribe (for ELA/lit writing items): Students dictate their responses to a human who records what they dictate verbatim. The scribe must be trained and qualified and must follow the administration guidelines provided in the *Smarter 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.

Table 7 presents a list of universal tools, designated supports, and accommodations that were offered in the 2017–2018 administration. Tables 8 through 13 provide the number of students who were offered the accommodations and/or designated supports.

Table 7. Universal Tools, Designated Supports, and Accommodations in 2017–2018

	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) Masking Mouse Pointer Permissive Mode Streamlined Mode Text-to-Speech[5] Translated Test Directions[6] Translations (Glossary)[6] Translations (Stacked) [7] Zoom	American Sign Language[8] Braille Braille Transcript Closed Captioning[9] Print-on-Request Text-to-Speech[10]
Non-embedded	Assistive Listening Device Breaks English Dictionary[2] Familiar TA Modified Lighting Refocus Scratch/Blank/Grid Paper/Whiteboards 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[11] Interpreter – Native Language[12] Interpret/Translate Orally – Directions Only Magnification Noise Buffers Paper/Pencil Test Read Aloud Items[13] Scribe[14] Separate Setting in School Separate Setting Not in School Simplify Directions in English Translated Test Directions Translations (Glossary)[15] Unique Accommodation[11] WhisperPhone®	100s Number Table[16] Abacus Alternate Response Options[17] Braille (Paper-Pencil Version) Calculator[1] Human Read Aloud Passages[18] Interpreter—Visual Communication[11] Multiplication Table[16] Physical Assistance from a TA Scribe Speech-to-Text

*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 ELA/lit PT stimuli, ELA/lit PT and CAT items (not ELA/lit CAT reading passages), and mathematics stimuli and items: must be set in TIDE before test begins
[6] For mathematics items
[7] For mathematics test
[8] For ELA/lit listening items and mathematics items

[9] For ELA/lit listening items
[10] For ELA/lit CAT reading passages; must be set in TIDE by state-level user
[11] For ELA/lit performance task passages
[12] Must be approved by DDOE
[13] For ELA/lit items (not ELA/lit reading passages) and mathematics items
[14] For ELA/lit non-writing items and mathematics items
[15] For mathematics items on the paper-pencil test
[16] For mathematics items beginning in grade 4
[17] Includes adapted keyboards, large keyboard, StickyKeys, MouseKeys, FilterKeys, adapted mouse, touch screen, head wand, and switches
[18] For ELA/lit CAT reading passages, all grades—must be approved by DDOE

Table 8. Students with Embedded and Non-Embedded Accommodations in ELA/Lit

Accommodations	Grade					
	3	**4**	**5**	**6**	**7**	**8**
Embedded Accommodations						
American Sign Language	5	5	10	3	4	2
Braille					1	
Closed Captioning	14	17	18	13	30	40
Print-on-Request: Items	3		3	1	3	2
Print-on-Request: Passages	25	36	19	10	6	15
Print-on-Request: Passages & Items	378	503	463	461	425	380
Print-on-Request: Stimuli	2	1		2		2
Text-to-Speech: Passages		1	1			
Text-to-Speech: Passages & Items	25	22	33	31	19	18
Non-Embedded Accommodations						
Alternate Response Options			1		1	
Human Read Aloud Passages	23	16	23	13	4	3
Physical Assistance from a TA	51	51	20	11	3	2
Scribe Items (Writing)	107	120	98	38	12	12
Speech-to-Text	8	13	18	29	16	31

Table 9. Students with Embedded Designated Supports in ELA/Lit

Designated Supports	Subgroup	Grade					
		3	**4**	**5**	**6**	**7**	**8**
Color Contrast	Overall	5	17	11	23	27	18
	ELL		8	2	3	2	1
	Special Ed	4	7	3	21	25	14
Masking	Overall	152	224	320	321	312	258
	ELL	46	55	104	62	66	54
	Special Ed	97	132	175	219	218	207
Mouse Pointer	Overall		1	2		2	1
	ELL						
	Special Ed			2		2	1
Permissive Mode	Overall	4	6	18	47	39	22
	ELL	2		1	22	30	15
	Special Ed	4	6	17	25	15	9
Streamlined Mode	Overall	4	18	25	15	34	18
	ELL	2	6	5	6	15	10
	Special Ed	3	12	19	9	24	9
Text-to-Speech: Items	Overall	2,524	2,718	2,597	2,041	1,756	1,315
	ELL	1,105	1,071	679	382	315	279
	Special Ed	996	1,176	1,199	1,188	1,142	983
Text-to-Speech: Stimuli & Items	Overall	2,557	2,739	2,612	2,039	1,746	1,360
	ELL	1,115	1,076	688	391	315	278
	Special Ed	1,019	1,198	1,235	1,228	1,150	1,034
Zoom	Overall	26	61	55	50	20	9
	ELL	10	27	18	14	2	1
	Special Ed	9	16	25	31	7	5

Table 10. Students with Non-Embedded Designated Supports in ELA/Lit

Designated Supports	Subgroup	Grade					
		3	**4**	**5**	**6**	**7**	**8**
Bilingual Dictionary	Overall	16	51	64	61	67	96
	ELL	16	51	63	61	67	93
	Special Ed	2	6	5	8	8	11
Color Contrast	Overall		1	6		6	1
	ELL					2	1
	Special Ed		1			5	1
Color Overlay	Overall	5	5	19	5	3	2
	ELL			9	1		
	Special Ed	2	5	16	4	3	1
Disable Universal Tools	Overall	1	2	3		1	
	ELL		1			1	
	Special Ed	1	1	2			
ELL 1st Year Exemption	Overall	4	6	8	16	5	10
	ELL	4	5	7	15	5	9
	Special Ed						
Human Read Aloud Stimuli & Items	Overall	539	606	441	224	135	102
	ELL	171	206	106	28	12	7
	Special Ed	297	349	295	189	116	97
Interpreter—Native Language	Overall	2	4	7	1	2	
	ELL	1	1	1			
	Special Ed		4	4	1	1	
Interpret/Translate Orally - Directions Only	Overall	43	12	15	9	7	10
	ELL	41	10	9	8	5	10
	Special Ed	2	2	5	1	2	
Magnification	Overall	19	16	26	6	7	11
	ELL		1	1			1
	Special Ed	4	2	8	5	4	9
Noise Buffers	Overall	89	137	162	97	65	38
	ELL	7	25	17	7	4	1
	Special Ed	62	76	122	68	46	27
Paper-Pencil Test	Overall	8			5	11	8
	ELL	1				1	
	Special Ed	7			3	9	4
Read Aloud Items	Overall	573	610	487	280	197	171
	ELL	177	202	107	32	18	15
	Special Ed	328	343	315	229	160	157
Scribe Items (Non-Writing)	Overall	22	25	14	13	2	4
	ELL	10	6	3	2		1
	Special Ed	12	18	11	9	1	2
Separate Setting in School	Overall	514	621	535	404	390	330
	ELL	100	118	88	50	33	40
	Special Ed	430	508	458	346	343	286
Separate Setting Not in School/Homebound	Overall	2		2		4	4
	ELL						
	Special Ed			1		2	1

Designated Supports	Subgroup	Grade					
		3	4	5	6	7	8
Simplified Test Directions	Overall	548	554	333	218	179	181
	ELL	430	395	207	118	86	91
	Special Ed	172	213	155	116	107	94
Translated Test Directions	Overall		1	9	11	17	14
	ELL		1	9	11	16	13
	Special Ed						
WhisperPhone®	Overall	200	262	145	48	22	10
	ELL	61	77	33	10	15	9
	Special Ed	117	173	115	37	5	1

Table 11. Students with Embedded and Non-Embedded Accommodations in Mathematics

Accommodations	Grade					
	3	4	5	6	7	8
Embedded Accommodations						
American Sign Language	5	5	10	2	4	3
Print-on-Request: Stimuli & Items	360	475	453	462	416	388
Non-Embedded Accommodations						
100s Number Table	393	610	519	203	105	64
Abacus	1	4		1	1	2
Alternate Response Options	1	1	1	1	1	
Calculator	34	82	59	162	238	322
Interpreter—Visual Communication	18	12	11	11	15	11
Multiplication Table	98	1,089	1,206	1,153	981	848
Physical Assistance from a TA	51	54	20	14	5	1
Scribe Items (Writing)	102	117	86	32	10	7
Speech-to-Text	9	12	16	27	17	21

Table 12. Students with Embedded Designated Supports in Mathematics

Designated Supports	Subgroup	Grade					
		3	**4**	**5**	**6**	**7**	**8**
Color Contrast	Overall	5	18	10	22	31	18
	ELL		9	2	4	2	1
	Special Ed	4	8	4	20	29	14
Masking	Overall	145	217	306	325	308	252
	ELL	46	55	106	68	70	56
	Special Ed	94	133	165	209	212	203
Mouse Pointer	Overall		1	1			
	ELL						
	Special Ed			1			
Permissive Mode	Overall	5	5	19	47	41	27
	ELL	2		2	23	32	20
	Special Ed	5	5	18	23	15	10
Streamlined Mode	Overall	9	26	36	48	74	59
	ELL	7	15	15	39	56	51
	Special Ed	1	10	17	9	23	10
Text-to-Speech: Stimuli & Items	Overall	2,573	2,787	2,671	2,101	1,778	1,397
	ELL	1,125	1,102	697	412	331	285
	Special Ed	1,002	1,218	1,260	1,234	1,150	1,034
Translation (Glossary): Spanish	Overall	140	209	107	97	88	87
	ELL	137	206	102	92	86	84
	Special Ed	5	20	12	13	11	7
Translation (Glossary): Other Languages	Overall	15	16	18	31	16	24
	ELL	13	16	15	14	17	25
	Special Ed	1			1		
Zoom	Overall	27	63	57	50	18	8
	ELL	12	29	19	14	2	1
	Special Ed	8	17	26	30	6	4

Table 13. Students with Non-Embedded Designated Supports in Mathematics

Designated Supports	Subgroup	Grade					
		3	4	5	6	7	8
Color Contrast	Overall		1			6	2
	ELL					2	1
	Special Ed		1			5	2
Color Overlay	Overall	5	5	8	5	2	2
	ELL			1	1		
	Special Ed	2	5	5	4	2	1
Disable Universal Tools	Overall	1	2	6		3	
	ELL		1			3	
	Special Ed	1	1	3			
Human Read Aloud Stimuli & Items	Overall	542	622	472	201	110	82
	ELL	184	223	133	38	20	15
	Special Ed	290	359	299	163	85	72
Human Read Aloud in Spanish	Overall	22	20	21	17	8	13
	ELL	20	19	20	17	8	13
	Special Ed	2	1				2
Interpreter—Native Language	Overall	16	2	5	2	1	1
	ELL	2	1	1			
	Special Ed	14	2	3	1		1
Interpret/Translate Orally—Directions Only	Overall	52	15	19	16	15	20
	ELL	50	14	13	15	13	20
	Special Ed	2	2	5	1	2	2
Magnification	Overall	18	14	25	6	9	7
	ELL		1	1			1
	Special Ed	4	3	8	5	5	5
Noise Buffers	Overall	88	138	166	92	62	37
	ELL	7	25	19	5	3	1
	Special Ed	61	75	124	65	43	26
Paper-Pencil Test	Overall	6			6	8	8
	ELL	1			1	1	
	Special Ed	5			4	6	4
Scribe Items	Overall	17	27	21	11	1	1
	ELL	10	12	9	2		
	Special Ed	7	16	13	8	1	1
Separate Setting in School	Overall	515	611	531	400	382	327
	ELL	114	124	95	55	35	43
	Special Ed	429	492	451	336	331	280
Separate Setting Not in School/Homebound	Overall	2	5	4		4	6
	ELL						
	Special Ed		5	2		2	3
Simplified Test Directions	Overall	565	572	358	244	199	203
	ELL	452	413	234	146	112	115
	Special Ed	166	213	158	114	103	93
Translated Test Directions	Overall	6	6	17	23	26	21
	ELL	6	6	17	23	25	21
	Special Ed					1	1
Translations (Glossaries): Paper-Pencil	Overall		17		2	3	11
	ELL		16		2	3	10
	Special Ed						
Unique Accommodation	Overall				1		
	ELL						
	Special Ed				1		

Designated Supports	Subgroup	Grade					
		3	4	5	6	7	8
WhisperPhone®	Overall	170	244	106	49	24	13
	ELL	52	72	11	15	16	12
	Special Ed	107	166	96	34	6	1

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 test delivery system (TDS) captures all of 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 are 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 2017–2018 and 2016–2017 school years.

Commented [ZL1]: Did you use the testing window of the state or actually testing date from individual students? In addition, this analysis is an across-grade comparison on vertical scale.

Commented [YB2R1]: The instruction time is based on the number of days between test-end days in two years, between two grades. This is computed for all completed tests.

Formatted: Highlight

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 predicted value. Unusual residuals are determined based on the studentized *t* residuals. An unusual increased or decreased changing score is flagged when studentized *t* residuals are greater than |3|.

The residuals for individual~~of~~ students are also aggregated for a testing session, TA, and school. The system flags any unusual changes in an aggregate performance between administrations and/or years based on the average studentized *t* 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* was greater than |3|,

$$t = \frac{Average\ residuals}{\sqrt{\frac{s^2}{n} + \frac{\sum_{i=1}^{n} var(\hat{e}_i)}{n^2}}},$$

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), and $\hat{e}_i$ is the residual for *i*th student.

The total variance of 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, page 456),

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

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

Commented [ZL3]: Reference for the approach?

Commented [YB4R3]: Billingsley, 1995, page 456

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

The average and the standard deviation of test-taking time are computed across all students. Students and aggregate 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 prior knowledge of some test items (or is provided answers during the exam), he or she will respond correctly to those items at a higher probability than indicated by his or her ability as estimated across all items. In this case, the person-fit index will be large for the student. We note, however, that if a student has prior knowledge of the entire test content, this will not be detected based on the person-fit index, though the item response time index might flag such a student.

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

The person-fit index is computed using a standardized log-likelihood statistic. Following Drasgow, Levine, and Williams (1985), 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|,

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

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

2.8 Prevention and Recovery of Disruptions in Test Delivery System

AIR is continuously improving our ability to protect our systems from interruptions. AIR's test delivery system 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 below, 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 are able to respond robustly to such inevitable failures. Thus, AIR's test delivery system (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 below. Fault tolerance and automated recovery are built into every component of the system.

Student Machine

Commented [ZL5]: Student Machine, meaning?

Commented [YB6R5]: The computer that each student used to take the test.

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 during testing.

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

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

In short, data integrity is preserved by confirmed saves to 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 below), with backup copies remaining on the drive arrays of the disabled satellite.

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

Hub

Hub servers are redundant clusters of database servers with RAID drive systems. Hub servers continuously gather data from the test delivery satellites and their mini-hubs and store that data as described earlier. This real-time backup copy remains on the hub until the hub receives 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 quality assurance (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 system failure.

2.8.3 Other Disruption Prevention and Recovery

We have designed our system to be extremely fault-tolerant. The system can withstand failure of any component with little to no interruption of service. One way that we achieve this robustness is through redundancy. Key redundant systems are as follows:

- Our 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.
- Our hosting provider has multiple redundancies in the flow of information to and from our 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, we have redundant firewalls and load balancers throughout the environment.
- We use redundant power and switching within all of our server cabinets.
- Data are protected by nightly backups. We complete a full weekly backup and incremental backups nightly. 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 will need to rerun the backup.

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

3. SUMMARY OF 2017–2018 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 14 and 15 present the demographic composition of Delaware students who meet attemptedness requirements for scoring and reporting of the Smarter Balanced assessments.

Table 14. Number of Students in Summative ELA/Lit Assessment

Group	G3	G4	G5	G6	G7	G8
All Students	10,467	10,658	10,579	10,425	10,219	10,106
Female	5,160	5,210	5,275	5,222	5,070	4,955
Male	5,307	5,448	5,304	5,203	5,149	5,151
African American	3,174	3,252	3,216	3,087	3,160	3,219
American Indian/Alaskan	43	37	40	36	43	47
Asian	420	384	384	370	381	368
Hispanic	1,952	2,000	1,872	1,854	1,770	1,641
Pacific Islander	22	13	11	13	11	14
White	4,373	4,496	4,575	4,647	4,457	4,520
Multi-Racial	482	476	481	418	397	297
ELL	1,727	1,608	886	492	423	374
Special Education	1,447	1,610	1,612	1,574	1,510	1,437
CD 504	342	417	493	510	506	534
Title I	1,092	1,054	1,066	1,214	1,312	1,527

Table 15. Number of Students in Summative Mathematics Assessment

Group	G3	G4	G5	G6	G7	G8
All Students	10,517	10,689	10,633	10,446	10,231	10,117
Female	5,184	5,227	5,304	5,236	5,071	4,951
Male	5,333	5,462	5,329	5,210	5,160	5,166
African American	3,181	3,246	3,219	3,071	3,151	3,210
American Indian/Alaskan	43	37	40	35	44	48
Asian	427	396	397	374	388	373
Hispanic	1,982	2,023	1,909	1,888	1,809	1,674
Pacific Islander	22	13	11	14	11	15
White	4,378	4,499	4,574	4,646	4,436	4,506
Multi-Racial	483	475	483	418	392	291
ELL	1,790	1,663	952	543	477	427
Special Education	1,441	1,626	1,619	1,557	1,511	1,422
CD 504	343	420	496	510	500	537
Title I	1,096	1,061	1,070	1,212	1,314	1,524

3.2 SUMMARY OF OVERALL STUDENT PERFORMANCE

Tables 16–19 present a summary of the 2017–2018 summative test results for all students and by subgroup, including the average and the standard deviation of scale scores, the percentage of students in each

achievement level, and the percentage of proficient students. Figures 1–2 show the percentage of proficient students in four years for all students (cohort comparisons). Figures 3–4 show the average scale scores in four years for all students. The average and the standard deviation of scale scores, as well as the percentage of proficient students for each test administration, are provided in Appendix B.

Table 16. 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	10,467	2433.24	87.16	23	25	24	28	52
Female	5,160	2441.46	84.11	19	25	25	30	56
Male	5,307	2425.25	89.32	27	25	23	26	48
African American	3,174	2400.53	81.07	35	29	21	15	36
AmerIndian/Alaskan	43	2424.05	89.06	23	26	21	30	51
Asian	420	2498.23	82.05	6	15	21	58	79
Hispanic	1,952	2406.53	80.62	32	30	22	16	38
Pacific Islander	22	2446.17	77.35	18	18	27	36	64
White	4,373	2461.97	81.57	13	21	27	40	67
Multi-Racial	482	2439.87	84.20	20	24	25	30	55
ELL	1,727	2401.39	76.73	33	31	22	13	36
Special Education	1,447	2349.23	72.70	61	26	9	4	12
CD 504	342	2430.65	76.38	20	29	28	23	51
Title I	1,092	2448.55	80.14	15	27	29	30	59
Grade 4								
All Students	10,658	2479.25	92.34	25	20	25	30	55
Female	5,210	2488.58	89.91	21	20	26	33	58
Male	5,448	2470.33	93.74	28	20	25	27	52
African American	3,252	2443.65	88.76	37	24	23	16	39
AmerIndian/Alaskan	37	2471.96	88.28	27	22	22	30	51
Asian	384	2543.84	84.26	8	9	24	59	83
Hispanic	2,000	2455.85	84.90	32	24	26	18	44
Pacific Islander	13	2512.44	108.23	15	8	23	54	77
White	4,496	2509.15	85.57	15	17	26	43	69
Multi-Racial	476	2485.91	90.82	21	21	26	32	58
ELL	1,608	2442.83	80.15	37	26	25	13	38
Special Education	1,610	2389.25	82.36	64	19	12	5	17
CD 504	417	2469.87	85.24	25	23	28	25	53
Title I	1,054	2492.57	80.51	17	22	29	32	61
Grade 5								
All Students	10,579	2516.58	92.05	21	20	33	25	58
Female	5,275	2528.18	88.95	17	20	34	28	63
Male	5,304	2505.05	93.62	26	21	32	21	54
African American	3,216	2479.14	87.13	34	25	29	12	41
AmerIndian/Alaskan	40	2523.52	91.30	20	20	30	30	60
Asian	384	2587.59	84.88	6	9	27	58	86
Hispanic	1,872	2494.65	84.96	26	25	33	16	48
Pacific Islander	11	2540.37	112.29	18	0	45	36	82
White	4,575	2544.71	86.13	12	16	36	35	71
Multi-Racial	481	2526.99	85.92	15	24	34	27	61
ELL	886	2447.35	78.39	46	31	19	5	23
Special Education	1,612	2419.85	77.21	63	23	12	2	14
CD 504	493	2507.95	81.08	19	26	38	17	55
Title I	1,066	2526.69	82.79	14	23	38	24	63

Table 17. 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	10,425	2531.16	95.72	22	26	33	19	52
Female	5,222	2545.73	93.11	17	24	36	23	59
Male	5,203	2516.54	96.08	27	28	31	15	46
African American	3,087	2496.65	89.74	33	30	29	9	37
AmerIndian/Alaskan	36	2505.53	105.59	31	22	33	14	47
Asian	370	2606.60	86.28	7	9	34	49	83
Hispanic	1,854	2503.83	90.48	29	31	30	9	40
Pacific Islander	13	2475.94	134.19	54	8	23	15	38
White	4,647	2559.09	89.84	13	22	38	27	65
Multi-Racial	418	2533.96	95.17	18	29	34	18	52
ELL	492	2420.01	76.68	69	25	5	1	6
Special Education	1,574	2426.55	78.02	65	26	8	1	9
CD 504	510	2527.01	80.33	18	33	36	13	50
Title I	1,214	2537.59	88.53	18	27	37	19	55
Grade 7								
All Students	10,219	2553.50	98.61	22	24	37	17	54
Female	5,070	2569.11	94.17	17	22	40	20	61
Male	5,149	2538.13	100.45	27	25	35	13	48
African American	3,160	2515.25	94.22	35	28	30	7	38
AmerIndian/Alaskan	43	2578.20	80.64	16	23	35	26	60
Asian	381	2626.91	91.83	8	8	40	44	83
Hispanic	1,770	2526.64	93.16	28	29	34	8	42
Pacific Islander	11	2579.50	53.88	0	27	55	18	73
White	4,457	2584.12	90.73	12	19	44	24	68
Multi-Racial	397	2560.25	95.11	20	24	38	19	56
ELL	423	2440.76	78.72	67	26	7	0	7
Special Education	1,510	2445.58	81.97	65	25	10	1	10
CD 504	506	2553.31	87.23	20	27	39	14	53
Title I	1,312	2557.86	88.67	18	27	40	15	55
Grade 8								
All Students	10,106	2568.47	99.33	22	25	37	16	53
Female	4,955	2586.63	95.04	16	24	40	21	60
Male	5,151	2551.00	100.24	28	26	34	12	46
African American	3,219	2531.77	94.85	33	30	30	7	37
AmerIndian/Alaskan	47	2565.10	92.82	17	32	38	13	51
Asian	368	2647.59	96.96	7	10	38	45	84
Hispanic	1,641	2541.00	94.40	29	29	35	8	43
Pacific Islander	14	2569.38	115.45	29	21	29	21	50
White	4,520	2597.71	91.20	13	21	43	23	66
Multi-Racial	297	2575.54	96.79	19	29	33	19	52
ELL	374	2453.36	79.45	68	22	8	1	9
Special Education	1,437	2463.72	78.24	64	26	10	1	10
CD 504	534	2559.75	90.76	21	31	36	12	48
Title I	1,527	2570.15	94.47	21	25	40	14	54

Table 18. 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	10,517	2441.20	83.11	23	23	30	24	54
Female	5,184	2440.22	78.78	22	25	30	23	53
Male	5,333	2442.16	87.10	24	22	29	25	54
African American	3,181	2406.44	78.26	36	28	26	11	37
AmerIndian/Alaskan	43	2434.74	68.62	23	28	33	16	49
Asian	427	2515.35	84.04	6	9	25	60	85
Hispanic	1,982	2419.68	72.87	30	29	29	13	42
Pacific Islander	22	2456.00	78.74	18	14	41	27	68
White	4,378	2468.56	76.70	12	19	33	35	68
Multi-Racial	483	2444.75	80.78	20	24	31	25	55
ELL	1,790	2420.28	74.59	30	27	29	14	43
Special Education	1,441	2359.24	81.79	61	22	13	4	17
CD 504	343	2438.76	71.49	20	29	30	21	51
Title I	1,096	2456.98	75.26	16	22	35	27	62
Grade 4								
All Students	10,689	2484.42	82.61	18	31	28	22	50
Female	5,227	2482.85	78.32	18	34	29	20	49
Male	5,462	2485.91	86.50	19	29	28	24	52
African American	3,246	2448.97	76.14	30	38	22	10	32
AmerIndian/Alaskan	37	2484.95	90.31	16	32	24	27	51
Asian	396	2556.20	83.92	5	13	27	56	83
Hispanic	2,023	2465.70	75.32	23	37	27	13	40
Pacific Islander	13	2474.56	138.72	15	23	38	23	62
White	4,499	2511.60	76.07	9	25	34	32	65
Multi-Racial	475	2489.25	81.66	17	30	29	23	52
ELL	1,663	2458.52	73.53	25	38	25	11	37
Special Education	1,626	2407.12	76.82	54	31	12	4	15
CD 504	420	2475.14	72.53	17	38	30	15	45
Title I	1,061	2505.50	71.59	10	28	34	28	63
Grade 5								
All Students	10,633	2510.37	89.99	28	29	20	23	43
Female	5,304	2510.39	86.70	27	31	20	22	42
Male	5,329	2510.35	93.16	28	27	20	24	44
African American	3,219	2469.30	82.30	45	31	15	9	24
AmerIndian/Alaskan	40	2512.40	110.59	28	33	8	33	40
Asian	397	2595.06	87.86	6	16	19	59	78
Hispanic	1,909	2488.45	80.65	34	33	19	14	33
Pacific Islander	11	2543.31	102.17	18	18	27	36	64
White	4,574	2540.52	82.80	16	26	25	34	59
Multi-Racial	483	2514.61	85.56	23	36	18	22	41
ELL	952	2456.11	77.06	51	32	12	6	18
Special Education	1,619	2421.70	74.68	70	21	7	2	9
CD 504	496	2507.80	78.80	26	35	21	18	39
Title I	1,070	2526.77	83.89	19	31	23	27	50

Table 19. 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	10,446	2520.99	104.76	29	31	21	19	40
Female	5,236	2527.53	100.49	27	31	23	19	42
Male	5,210	2514.42	108.49	31	31	20	18	38
African American	3,071	2477.37	100.49	44	32	16	7	24
AmerIndian/Alaskan	35	2509.23	93.44	31	34	23	11	34
Asian	374	2615.86	106.16	8	18	22	52	74
Hispanic	1,888	2495.70	94.26	37	35	18	10	28
Pacific Islander	14	2472.18	122.51	50	21	7	21	29
White	4,646	2552.87	96.22	17	29	26	27	53
Multi-Racial	418	2519.02	97.72	29	33	22	17	39
ELL	543	2416.10	88.43	72	23	4	1	5
Special Education	1,557	2405.80	95.49	76	20	3	1	4
CD 504	510	2518.82	89.29	30	35	21	14	35
Title I	1,212	2534.06	89.20	22	33	26	19	45
Grade 7								
All Students	10,231	2531.37	108.32	32	29	22	17	39
Female	5,071	2534.20	105.31	30	30	22	18	39
Male	5,160	2528.58	111.14	33	28	22	17	39
African American	3,151	2485.77	99.44	47	32	14	6	21
AmerIndian/Alaskan	44	2536.47	103.57	32	30	16	23	39
Asian	388	2631.78	110.14	10	11	28	51	79
Hispanic	1,809	2503.67	102.32	40	31	19	9	28
Pacific Islander	11	2555.03	79.57	18	36	36	9	45
White	4,436	2565.71	99.20	20	27	28	25	53
Multi-Racial	392	2536.42	100.98	30	31	22	17	40
ELL	477	2422.92	103.55	74	18	6	2	8
Special Education	1,511	2416.13	87.59	79	16	4	1	5
CD 504	500	2530.91	95.83	30	35	21	14	35
Title I	1,314	2537.66	97.78	27	32	25	16	41
Grade 8								
All Students	10,117	2548.26	117.93	36	25	19	20	39
Female	4,951	2555.14	112.61	33	26	21	20	41
Male	5,166	2541.68	122.47	39	24	18	20	37
African American	3,210	2499.31	110.21	53	24	15	8	23
AmerIndian/Alaskan	48	2541.41	110.49	35	27	17	21	38
Asian	373	2662.72	124.96	11	13	19	57	76
Hispanic	1,674	2517.93	105.99	46	26	16	11	27
Pacific Islander	15	2543.67	134.43	40	20	20	20	40
White	4,506	2584.47	108.30	22	26	24	28	52
Multi-Racial	291	2556.88	108.35	34	26	18	21	39
ELL	427	2447.88	95.64	77	15	5	3	8
Special Education	1,422	2426.24	93.26	81	15	3	2	4
CD 504	537	2539.01	102.53	39	30	16	15	31
Title I	1,524	2549.89	110.53	34	27	20	18	39

Figure 1. ELA/Lit % Proficient Across Years

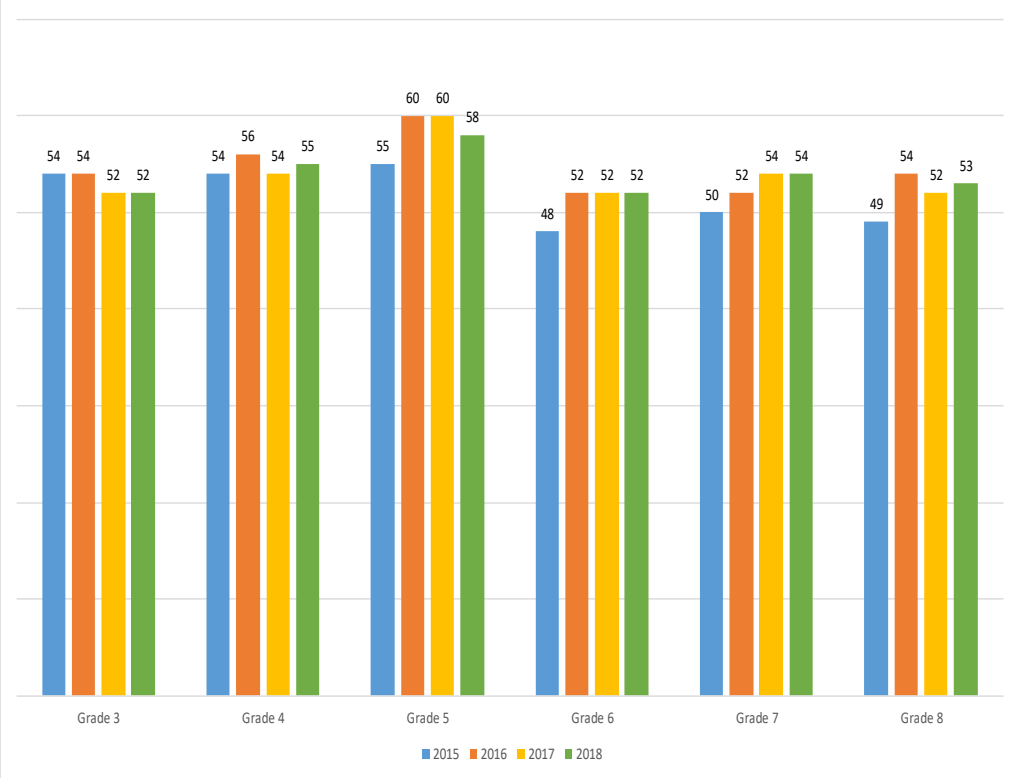


Figure 2. Mathematics % Proficient Across Years

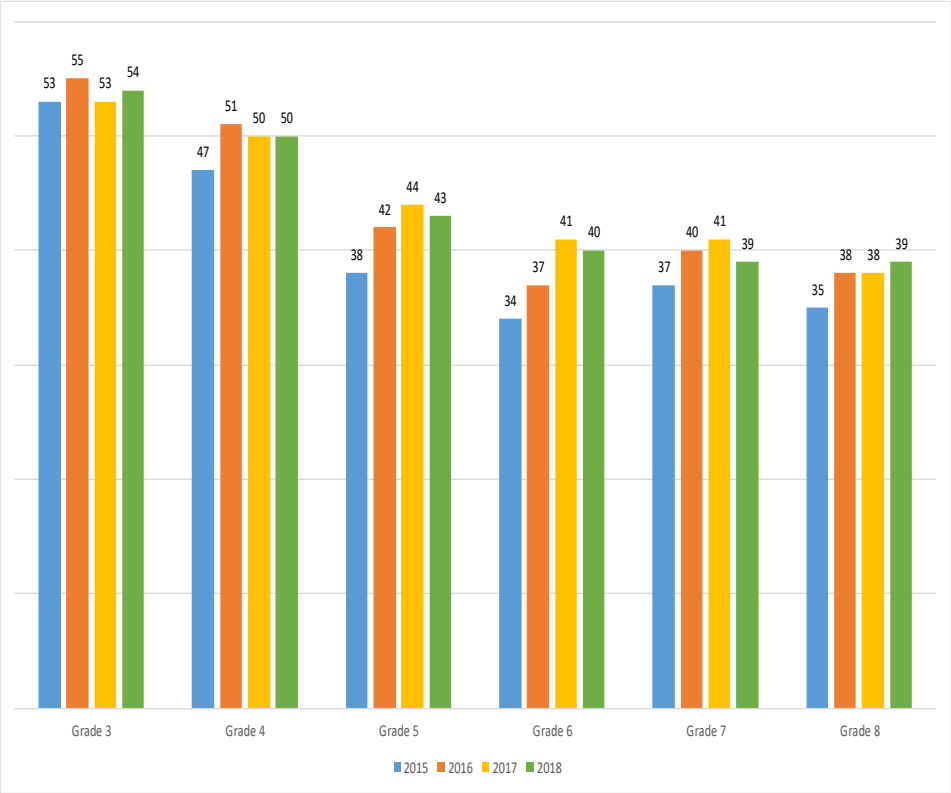


Figure 3. ELA/Lit Average Scale Score Across Years

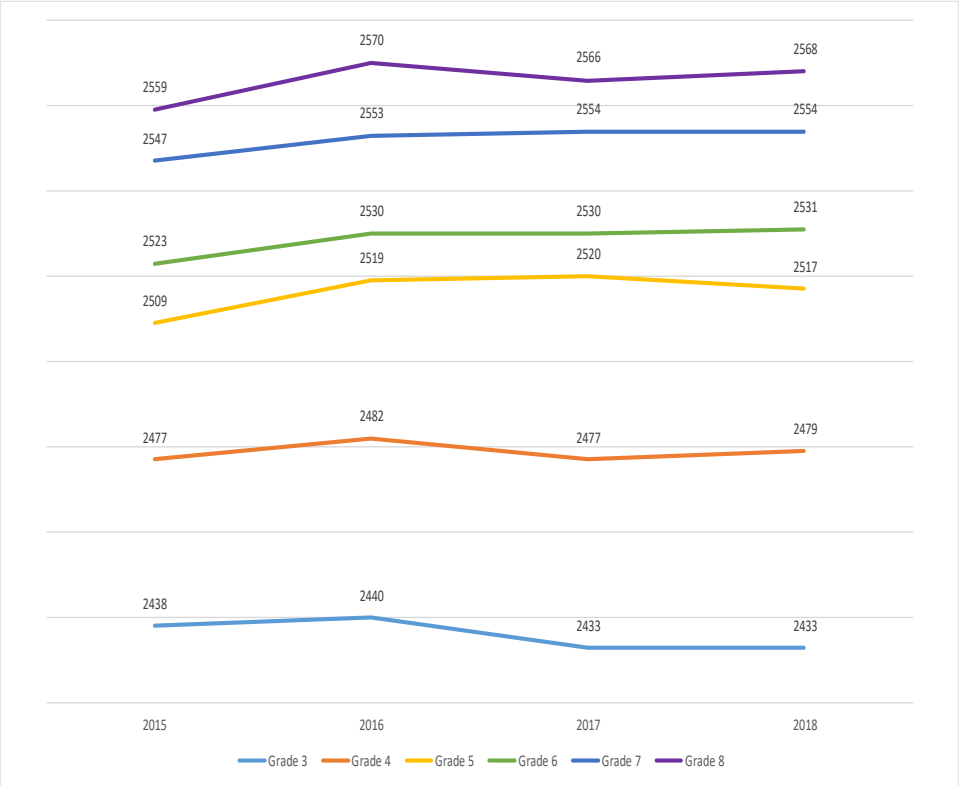
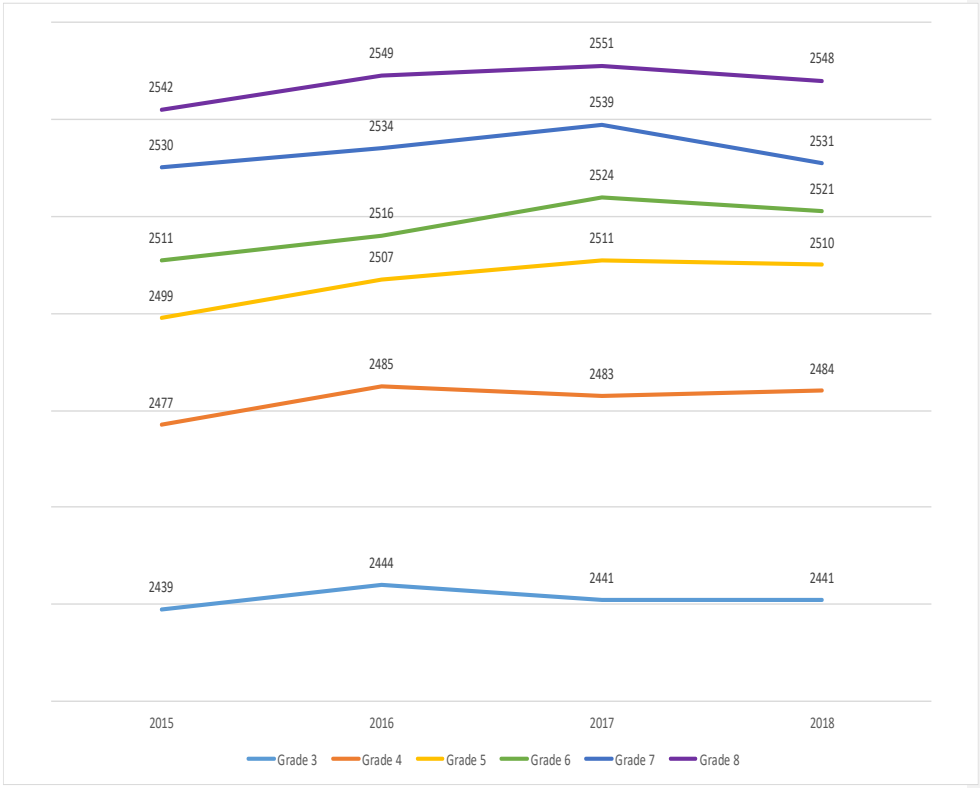


Figure 4. Mathematics Average Scale Score Across Years



Because the precision of scores in each claim is not sufficient to report scores, given a small number of items, the scores on each claim are reported using one of the three performance categories, taking into account the SEM of the claim score: (1) Below standard, (2) At/Near standard, or (3) Above standard (see Section 6.5 for the rules). Tables 20 and 21 present the distribution of performance categories for each claim. The number of claims is four in ELA/lit, and three in mathematics, combining claims 2 and 4.

Table 20. 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	26	28	15	20
	At/Near	47	46	63	51
	Above	27	26	22	29
4	Below	22	26	14	19
	At/Near	50	48	64	53
	Above	28	26	23	29
5	Below	22	21	17	18
	At/Near	48	50	63	48
	Above	30	29	20	34
6	Below	29	28	17	18
	At/Near	45	48	64	52
	Above	26	25	19	30
7	Below	27	23	20	18
	At/Near	46	48	66	52
	Above	27	29	14	30
8	Below	27	27	17	19
	At/Near	45	49	64	51
	Above	28	24	20	30

Table 21. Mathematics Percentage of Students in Performance Categories by Claim

Grade	Performance Category	Claim 1: Concepts and Procedures	Claims 2 & 4: Problem Solving & Modeling and Data Analysis	Claim 3: Communicating Reasoning
3	Below	30	25	19
	At/Near	34	44	47
	Above	37	30	34
4	Below	32	26	24
	At/Near	34	48	47
	Above	34	26	29
5	Below	38	29	28
	At/Near	33	48	49
	Above	30	23	22
6	Below	40	35	33
	At/Near	35	45	46
	Above	25	19	21
7	Below	42	32	25
	At/Near	33	48	57
	Above	25	20	19
8	Below	41	25	30
	At/Near	34	51	50
	Above	25	23	20

3.3 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.

In the test delivery system (TDS), item response time is captured as the item page time (the time that a student spend on each item page) in milliseconds. Discrete items appear on the screen one item at a time, and items associated with a stimulus appear on the screen together with the page time measured as the total time spent on all associated items. In this case, the page time for each item is the average time for all the items associated with the stimulus. For each student, the total testing time for the test was the sum of the page time for all items.

Tables 22 and 23 present an average testing time and the testing time at percentiles for the overall test, the CAT component, and the PT component.

Table 22. 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	5:04	4:26	6:18	6:54	7:41	8:43	10:33
4	5:27	4:53	6:47	7:21	8:04	9:00	10:38
5	5:17	4:49	6:28	6:57	7:37	8:25	9:48
6	4:49	4:23	5:50	6:15	6:50	7:39	9:21
7	4:08	3:50	5:00	5:19	5:46	6:26	7:33
8	3:58	3:39	4:50	5:10	5:38	6:15	7:16
CAT Component							
3	2:26	2:09	2:56	3:09	3:28	4:00	4:56
4	2:39	2:23	3:13	3:28	3:48	4:14	5:03
5	2:35	2:23	3:09	3:23	3:39	4:02	4:41
6	2:26	2:16	2:57	3:09	3:23	3:44	4:21
7	2:07	2:00	2:33	2:43	2:55	3:13	3:44
8	2:01	1:52	2:25	2:35	2:48	3:05	3:36
PT Component							
3	2:38	2:11	3:26	3:49	4:22	5:05	6:15
4	2:48	2:23	3:35	3:58	4:27	5:05	6:17
5	2:41	2:22	3:22	3:41	4:08	4:43	5:43
6	2:22	2:01	3:00	3:18	3:42	4:17	5:26
7	2:00	1:48	2:33	2:48	3:06	3:33	4:20
8	1:57	1:43	2:30	2:44	3:03	3:29	4:12

Table 23. 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	2:47	2:25	3:21	3:39	4:05	4:43	5:48
4	2:50	2:30	3:32	3:49	4:13	4:46	5:42
5	3:23	2:58	4:07	4:28	4:55	5:38	6:53
6	2:57	2:40	3:32	3:47	4:07	4:36	5:40
7	2:20	2:09	2:49	3:01	3:17	3:38	4:17
8	2:43	2:30	3:20	3:34	3:54	4:19	5:01
CAT Component							
3	1:51	1:35	2:13	2:25	2:43	3:09	3:56
4	1:56	1:41	2:24	2:36	2:52	3:14	3:54
5	1:54	1:41	2:19	2:31	2:45	3:07	3:44
6	1:53	1:43	2:17	2:27	2:40	2:57	3:31
7	1:43	1:35	2:06	2:15	2:26	2:41	3:10
8	1:57	1:49	2:24	2:35	2:48	3:06	3:36
PT Component							
3	0:56	0:46	1:10	1:18	1:28	1:43	2:09
4	0:55	0:45	1:08	1:16	1:26	1:40	2:07
5	1:29	1:13	1:49	2:00	2:18	2:46	3:30
6	1:03	0:52	1:17	1:25	1:36	1:53	2:27
7	0:36	0:31	0:46	0:51	0:57	1:06	1:24
8	0:45	0:39	0:58	1:04	1:11	1:22	1:41

3.4 Distribution of Student Ability and Item Difficulty

Figures 5 and 6 display the empirical distribution of the Delaware student scale scores in the 2017–2018 administration and the distribution of the administered summative item difficulty parameters. The student ability distribution is shifted to the left in all grades and subjects, a pattern more pronounced in the mathematics upper grades, indicating that the pool includes more difficult items than the ability of students in the tested population. The pool includes difficult items to accurately measure high-performing students but needs additional easy items to better measure low-performing students. The Smarter Balanced Assessment Consortium plans to add additional easy items to the pool and to augment the pool in proportion to the test blueprint constraints (e.g., content, Depth of Knowledge [DOK], item type, item difficulties) to better measure low-performing students.

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

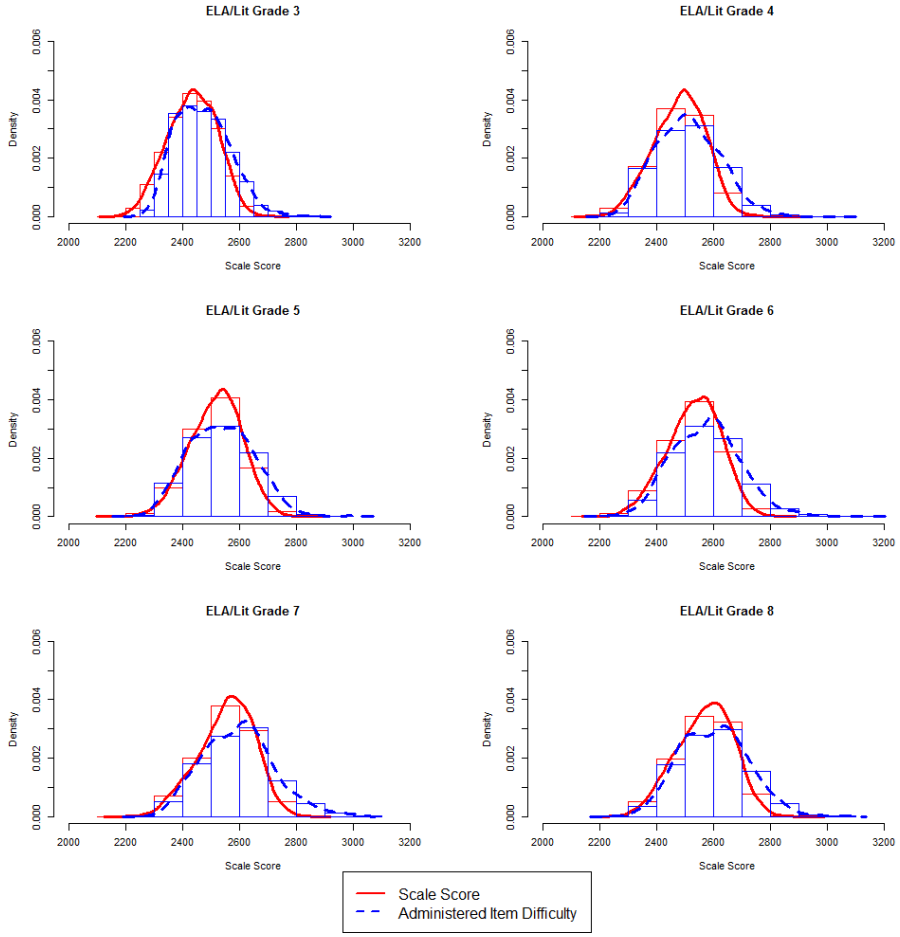
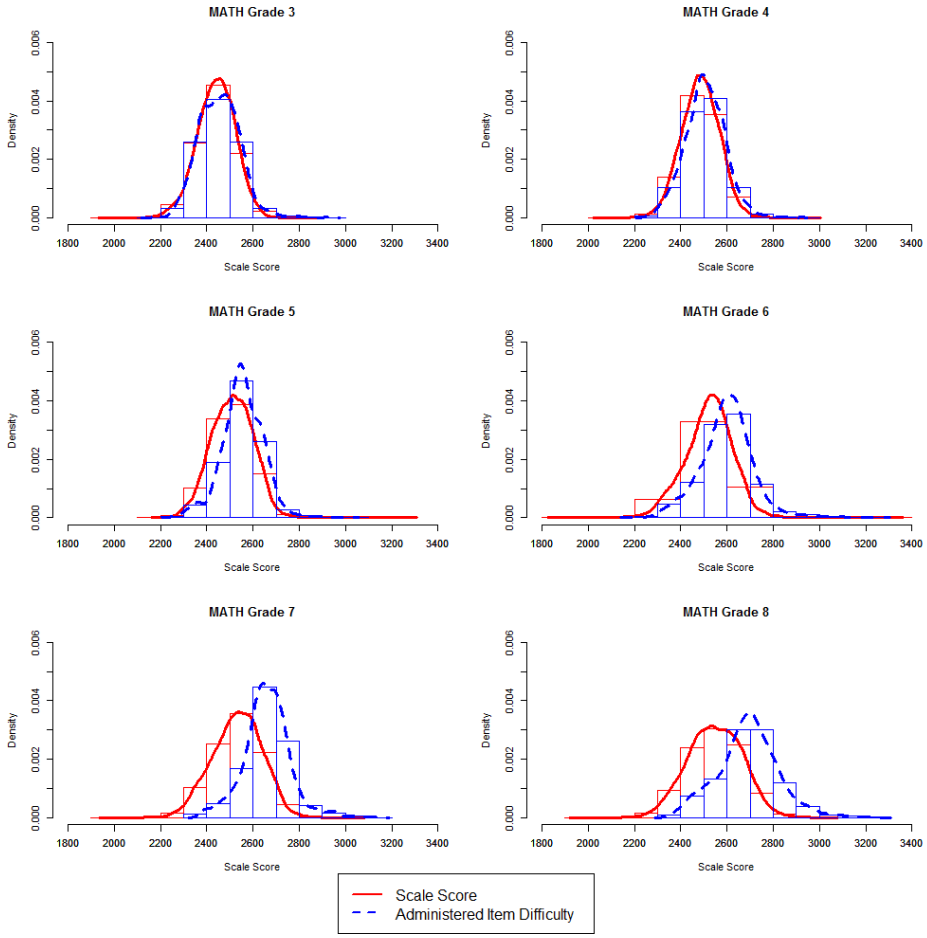


Figure 6. Student Ability–Item Difficulty Distribution for Mathematics



4. VALIDITY

According to the *Standards for Educational and Psychological Testing* (AERA, APA, and 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 depends on the assessments meeting the relevant standards of validity.

Validity evidence provided in this chapter is as follows:

- Test Content
- Internal Structure

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

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

4.1 EVIDENCE ON TEST CONTENT

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

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. 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 24 and 25 present the percentages of tests aligned with the test blueprint constraints for ELA/lit CAT. Table 24 provides the blueprint match rates for item and passage requirements for each claim. All tests met the requirements for items and passages. For DOK constraints, the Smarter Balanced blueprint specifies the minimum number of items, not the maximum. Table 25 presents the percentages of tests that satisfied the DOK and item-type constraints for each claim. All tests met the requirements.

Tables 26–27 provide the percentages of tests aligned with the test blueprint constraints for the mathematics CAT, the blueprint match rates for claims, DOK, and target constraints. In mathematics, the tests met all blueprint requirements, except for grades 6 and 8. In grade 6, the violation was in the claim 1 for target sets of E and F and target sets of B and G, and claim 3 calculator segment for target sets of A and D, each administered fewer or more items than required. In grade 8, the violation was in claim 1 for target sets of C and D and target sets of B, E, and G, each administered one item more or one item fewer than required.

Table 24. ELA/Lit Percentage of Delivered Tests Meeting Blueprint Requirements for Each Claim and the Number of Passages Administered

Grade	Claim	Min	Max	% BP Match for Item Requirement	% BP Match for Passage Requirement
3	1-IT	7	8	100	100
	1-LT	7	8	100	100
	2-W	10	10	100	
	3-L	8	8	100	100
	4-CR	6	6	100	
4	1-IT	7	8	100	100
	1-LT	7	8	100	100
	2-W	10	10	100	
	3-L	8	8	100	100
	4-CR	6	6	100	
5	1-IT	7	8	100	100
	1-LT	7	8	100	100
	2-W	10	10	100	
	3-L	8	9	100	100
	4-CR	6	6	100	
6	1-IT	10	12	100	100
	1-LT	4	4	100	100
	2-W	10	10	100	
	3-L	8	9	100	100
	4-CR	6	6	100	
7	1-IT	10	12	100	100
	1-LT	4	4	100	100
	2-W	10	10	100	
	3-L	8	9	100	100
	4-CR	6	6	100	
8	1-IT	12	12	100	100
	1-LT	4	4	100	100
	2-W	10	10	100	
	3-L	8	9	100	100
	4-CR	6	6	100	

Legend: 1-IT: Reading with Informational Text; 1-LT: Reading with Literary Text; 2-W: Writing; 3-L: Listening; and 4-CR: Research

Table 25. ELA/Lit Percentage of Delivered Tests Meeting Blueprint Requirements for Depth of Knowledge and Item Type

DOK and Item Type Constraints	Required Items	% BP Match					
		G3	G4	G5	G6	G7	G8
Claim 1 DOK2	≥ 7	100	100	100	100	100	100
Claim 1 DOK3 or higher	≥ 2	100	100	100	100	100	100
Claim 1 Short Answer in Target 2 or 4	0-1	100	100	100	100	100	100
Claim 1 Short Answer in Target 9 or 11	0-1	100	100	100	100	100	100
Claim 2 DOK2	≥ 4	100	100	100	100	100	100
Claim 2 DOK3 or higher	≥ 1	100	100	100	100	100	100
Claim 2 Brief Write	1	100	100	100	100	100	100
Claim 3 DOK2 or higher	≥ 3	100	100	100	100	100	100

Table 26. 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&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 27. 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	99				
	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	85
	Targets B, E, G					5–6	85
	Targets F, H					2–3	100
	Supporting Cluster					4–5	100
	Targets A, I, J					4–5	100
2&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	3	99	3	100	3	100
	Targets B, E	2-3	100	3	100	3	100
	Targets C, F, G	2	100	2	100	2	100
3–No Calc	Overall	1	100				

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

Table 28. Average and Range of the Number of Unique Targets Assessed Within Each Claim Across All Delivered 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	10.5	5.0	1.0	3.0	8–14	4–5	1–1	3–3
4	14	5	1	3	10.6	4.9	1.0	3.0	8–13	4–5	1–1	3–3
5	14	5	1	3	10.8	4.9	1.0	3.0	8–14	3–5	1–1	3–3
6	14	5	1	3	10.6	5.0	1.0	3.0	9–11	4–5	1–1	3–3
7	14	5	1	3	10.5	4.8	1.0	3.0	8–11	3–5	1–1	3–3
8	14	5	1	3	10.2	4.7	1.0	3.0	8–11	3–5	1–1	3–3
Mathematics												
3	11	4	6	6	10.8	2.0	5.8	3.0	8–11	2–2	3–6	3–4
4	12	4	6	6	10.0	2.0	5.5	3.0	9–10	2–2	3–6	3–3
5	11	4	6	6	9.0	2.0	5.2	3.0	9–9	2–2	3–6	3–3
6	10	4	7	6	9.9	2.0	4.8	3.0	8–10	2–2	3–7	2–3
7	9	4	7	6	8.0	2.0	4.7	3.0	8–8	2–2	3–6	3–3
8	10	4	7	6	10.0	2.0	4.8	3.0	10–10	2–2	3–6	3–4

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

4.2 Evidence on Internal Structure

The measurement and reporting model used in the Smarter Balanced assessments assumes a single underlying latent trait, with achievement reported as a total score as well as scores for each claim measured. The evidence on the internal structure is examined based on the correlations among claim scores.

The correlations among claim scores, both observed (below diagonal) and corrected for attenuation (above diagonal), are presented in Tables 29 and 30. The correction for attenuation indicates what the correlation would be if claim scores could be measured with perfect reliability, corrected (adjusted) for measurement error estimates.

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

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

Because the reliability for claim scores are low, the performance of all the claim scores is reported in three performance categories. The distribution of performance categories for each claim is provided in Tables 20 and 21, Section 3.2. Scale scores are not reported for claims.

Table 29. Correlations Among Claim Scores for ELA/Lit

Grade	Claim	Observed & Disattenuated Correlation			
		Claim 1	**Claim 2**	**Claim 3**	**Claim 4**
3	Claim 1: Reading		0.89	0.92	0.90
	Claim 2: Writing	0.70		0.88	0.90
	Claim 3: Listening	0.62	0.61		0.88
	Claim 4: Research	0.66	0.68	0.57	
4	Claim 1: Reading		0.91	0.92	0.91
	Claim 2: Writing	0.70		0.87	0.89
	Claim 3: Listening	0.61	0.60		0.90
	Claim 4: Research	0.66	0.66	0.58	
5	Claim 1: Reading		0.90	0.90	0.93
	Claim 2: Writing	0.70		0.84	0.91
	Claim 3: Listening	0.61	0.59		0.89
	Claim 4: Research	0.7	0.70	0.60	
6	Claim 1: Reading		0.89	0.91	0.93
	Claim 2: Writing	0.71		0.90	0.91
	Claim 3: Listening	0.63	0.65		0.91
	Claim 4: Research	0.68	0.69	0.6	
7	Claim 1: Reading		0.90	0.91	0.94
	Claim 2: Writing	0.71		0.88	0.93
	Claim 3: Listening	0.62	0.60		0.92
	Claim 4: Research	0.70	0.69	0.59	
8	Claim 1: Reading		0.91	0.94	0.93
	Claim 2: Writing	0.72		0.91	0.93
	Claim 3: Listening	0.65	0.64		0.92
	Claim 4: Research	0.69	0.70	0.61	

Table 30. Correlations Among Claim Scores for Mathematics

Grade	Claims	Observed & Disattenuated Correlation		
		Claim 1	**Claim 2&4**	**Claim 3**
3	Claim 1		0.96	0.92
	Claim 2 & 4	0.80		0.96
	Claim 3	0.77	0.74	
4	Claim 1		0.96	0.97
	Claim 2 & 4	0.80		0.98
	Claim 3	0.80	0.75	
5	Claim 1		1	0.95
	Claim 2 & 4	0.78		1
	Claim 3	0.76	0.73	
6	Claim 1		1	0.95
	Claim 2 & 4	0.81		1
	Claim 3	0.78	0.76	
7	Claim 1		1	0.96
	Claim 2 & 4	0.80		1
	Claim 3	0.74	0.69	
8	Claim 1		1	0.96
	Claim 2 & 4	0.78		1
	Claim 3	0.76	0.70	

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

5. RELIABILITY

Reliability refers to the consistency in test scores. Reliability is evaluated in terms of the standard errors 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. Within the item response theory (IRT) framework, measurement error varies conditioning on ability. The amount of precision in estimating achievement can be determined by the test information, which describes the amount of information provided by the test at each score point along the ability continuum. Test information is a value that is the inverse of the measurement error of the test; the larger the measurement error, the less test information is being provided. In computer-adaptive testing, because selected items vary among students, the measurement error can vary for the same ability, depending on the selected items for each student.

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

5.1 MARGINAL RELIABILITY

For reliability, the marginal reliability was computed for the scale scores, taking into account the varying measurement errors across the ability range. Marginal reliability is a measure of the overall reliability of an assessment based on the average conditional SEM, 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 standard error of measurement 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 SEM. In IRT, SEM is estimated as a function of test information provided by a given set of items that makes up the test. In computer-adaptive testing, items administered vary among all students, so the SEM also can vary among students, which yields conditional SEM. The average conditional SEM 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 conditional SEM, the greater the accuracy of test scores.

Table 31 presents the marginal reliability coefficients and the average conditional SEM for the total scale scores.

Table 31. 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	10,467	41	44	0.92	2433.24	87.16	24.06
4	10,658	41	44	0.92	2479.25	92.34	26.51
5	10,579	41	45	0.92	2516.58	92.05	25.62
6	10,425	41	45	0.93	2531.16	95.72	25.79
7	10,219	41	45	0.92	2553.50	98.61	27.23
8	10,106	43	45	0.93	2568.47	99.33	27.16
Mathematics							
3	10,517	39	40	0.95	2441.20	83.11	18.82
4	10,689	37	40	0.95	2484.42	82.61	19.36
5	10,633	38	40	0.94	2510.37	89.99	22.29
6	10,446	39	39	0.94	2520.99	104.76	26.10
7	10,231	38	40	0.93	2531.37	108.32	28.38
8	10,117	38	39	0.93	2548.26	117.93	30.95

5.2 STANDARD ERROR CURVES

Figures 7 and 8 present plots of the conditional SEM of scale scores across the range of ability. The vertical lines indicate the cut scores for Level 2, Level 3, and Level 4. The item selection algorithm matched items to each student's ability and to the test blueprints with the same precision across the range of abilities.

Overall, the standard error curves suggest that students are measured with a high degree of precision, given that the standard errors are consistently low. However, 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 7. Conditional Standard Error of Measurement for ELA/Lit

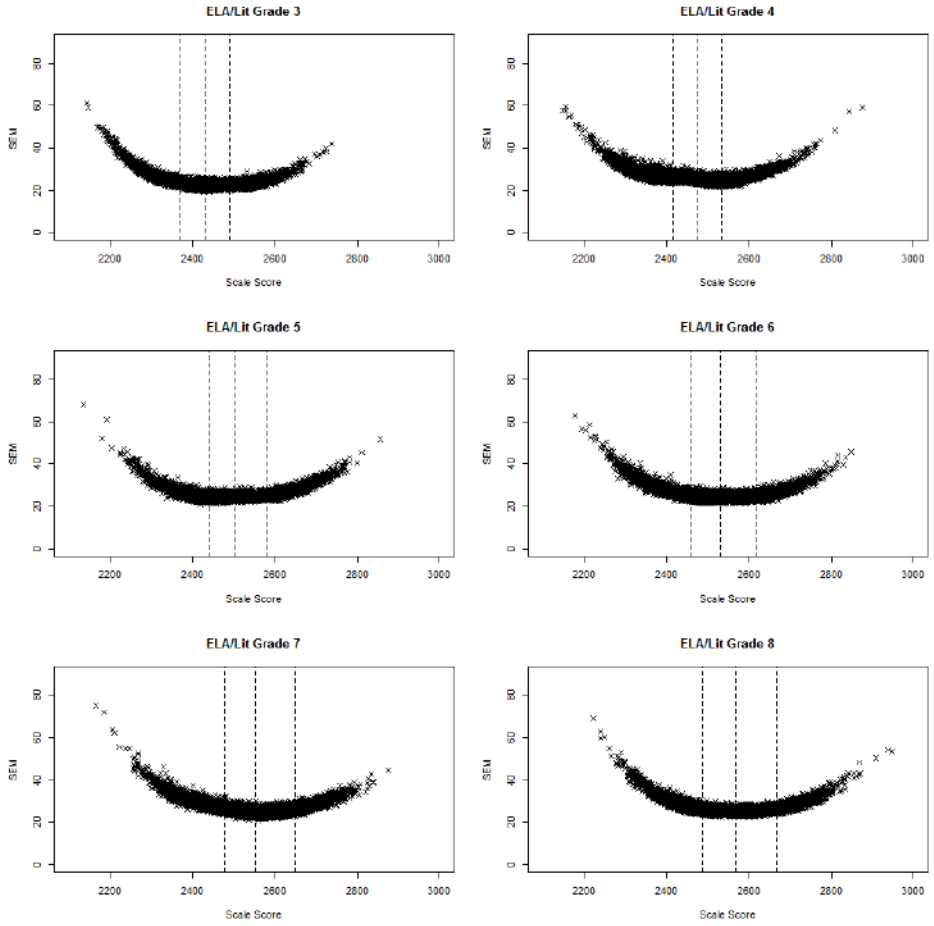
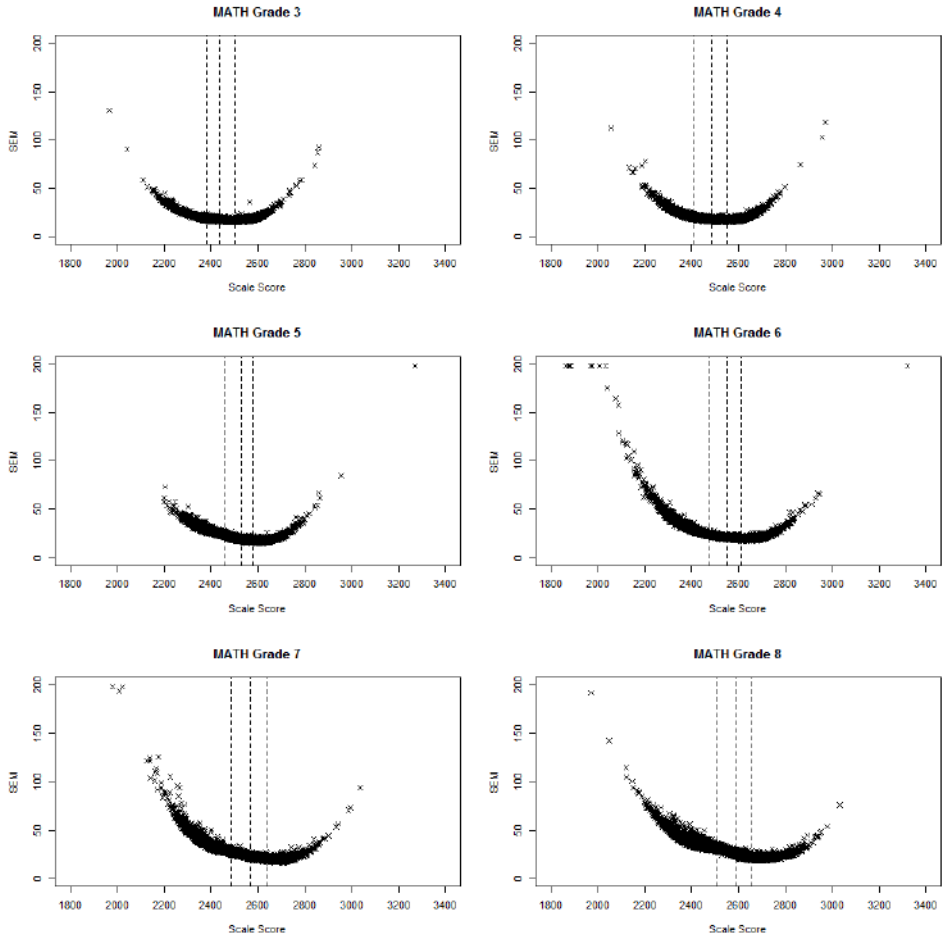


Figure 8. Conditional Standard Error of Measurement for Mathematics



The SEMs presented in the figures are summarized in Tables 32 and 33. Table 32 provides the average conditional SEM for all scores and for scores in each achievement level. Table 33 presents the average conditional SEMs at each cut score and the difference in average conditional SEMs between two cut scores. As shown in Figures 7 and 8, the greatest average conditional SEM is in Level 1 in both ELA/lit and mathematics. Average conditional SEMs at all cut scores are similar in ELA/lit, but larger in Level 2 cut scores in mathematics.

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

Grade	Level 1	Level 2	Level 3	Level 4	Average CSEM
ELA/Lit					
3	27.4	22.7	22.4	23.7	24.1
4	28.8	25.8	24.9	26.3	26.5
5	26.7	24.2	24.5	27.3	25.6
6	28.8	24.3	24.4	26.6	25.8
7	30.8	25.7	25.4	28.4	27.2
8	30.1	25.7	25.5	28.7	27.2
Mathematics					
3	22.6	17.9	16.8	18.2	18.8
4	24.1	18.3	17.1	19.1	19.4
5	28.2	20.7	18.1	19.3	22.3
6	35.0	22.0	20.1	21.7	26.1
7	37.3	25.2	21.7	20.9	28.4
8	38.8	29.3	23.6	22.0	31.0

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

Grade	L2 Cut	L3 Cut	L4 Cut	|L2–L3|	|L3–L4|	|L2–L4|
ELA/Lit						
3	23.4	22.2	22.7	1.2	0.5	0.7
4	26.1	25.6	24.3	0.5	1.3	1.8
5	24.4	24.4	24.7	0.0	0.3	0.3
6	24.9	24.2	24.7	0.6	0.5	0.2
7	26.3	25.1	26.2	1.2	1.0	0.1
8	26.1	25.3	26.5	0.8	1.3	0.4
Mathematics						
3	18.8	17.3	16.4	1.5	0.9	2.4
4	19.5	17.6	16.7	1.9	0.9	2.8
5	23.2	18.6	17.6	4.6	1.0	5.6
6	24.0	20.9	19.7	3.0	1.3	4.3
7	27.5	23.2	19.8	4.3	3.4	7.7
8	32.2	26.1	22.0	6.1	4.1	10.2

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.

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

The classification index can be examined in terms of the classification accuracy and the classification consistency. 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 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 1 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 a $L \times L$ 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 above table, the row represents the observed level and the column represents the expected level.

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

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

and the overall classification accuracy is estimated by

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

where N is the total number of students.

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 34 provides the percentage 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 34. Classification Accuracy and Consistency by Achievement Level

Grade	Achievement Level	ELA/Lit		Mathematics	
		% Accuracy	% Consistency	% Accuracy	% Consistency
3	Overall	80	72	83	76
	L1	89	82	90	84
	L2	72	62	74	65
	L3	69	58	79	72
	L4	88	83	89	85
4	Overall	78	70	84	77
	L1	89	82	89	82
	L2	64	52	80	73
	L3	67	57	79	71
	L4	88	82	90	85
5	Overall	80	72	83	75
	L1	89	82	89	84
	L2	68	56	77	69
	L3	76	68	72	61
	L4	86	80	90	85
6	Overall	81	73	83	76
	L1	90	83	92	86
	L2	74	64	78	70
	L3	78	71	72	62
	L4	85	78	90	83
7	Overall	80	73	83	76
	L1	90	84	91	86
	L2	72	61	76	67
	L3	79	72	75	65
	L4	84	75	89	83
8	Overall	81	74	82	75
	L1	89	83	90	85
	L2	74	64	72	62
	L3	80	74	71	61
	L4	84	75	90	84

5.4 RELIABILITY FOR SUBGROUPS

The reliability of test scores is also computed by subgroup. Tables 35 and 36 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.

Table 35. 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	0.93	0.92	0.93
Female	0.92	0.91	0.92	0.92	0.92	0.92
Male	0.93	0.92	0.92	0.93	0.93	0.93
African American	0.91	0.91	0.91	0.92	0.91	0.92
AmerIndian/Alaskan	0.92	0.91	0.92	0.94	0.89	0.92
Asian	0.92	0.90	0.90	0.91	0.91	0.92
Hispanic	0.91	0.90	0.91	0.92	0.91	0.92
Pacific Islander	0.91	0.94	0.94	0.95	0.77	0.94
White	0.92	0.91	0.91	0.92	0.91	0.91
Multi-Racial	0.92	0.92	0.91	0.93	0.92	0.92
ELL	0.90	0.89	0.89	0.86	0.86	0.86
Special Education	0.87	0.88	0.88	0.87	0.87	0.86
CD 504	0.90	0.91	0.90	0.90	0.91	0.91
Title I	0.91	0.90	0.91	0.92	0.91	0.92

Table 36. 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	0.94	0.94	0.93	0.94	0.93	0.93
Male	0.95	0.95	0.94	0.94	0.93	0.93
African American	0.94	0.93	0.92	0.92	0.90	0.90
AmerIndian/Alaskan	0.93	0.95	0.95	0.93	0.93	0.92
Asian	0.94	0.94	0.94	0.95	0.95	0.95
Hispanic	0.93	0.93	0.92	0.92	0.91	0.91
Pacific Islander	0.95	0.94	0.96	0.95	0.91	0.94
White	0.94	0.94	0.94	0.94	0.93	0.93
Multi-Racial	0.95	0.95	0.93	0.93	0.93	0.92
ELL	0.94	0.93	0.89	0.85	0.84	0.83
Special Education	0.93	0.91	0.87	0.85	0.81	0.82
CD 504	0.94	0.93	0.92	0.92	0.92	0.91
Title I	0.94	0.93	0.94	0.93	0.92	0.92

5.5 RELIABILITY FOR CLAIM SCORES

The marginal reliability coefficients and the measurement errors are also computed for claim scores. 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 scores given a small number of items, the scores on each claim are reported using one of the three performance categories, taking into account the SEM of the claim score: (1) Below standard, (2) At/Near standard, or (3) Above standard. Tables 37 and 38 present the marginal reliability coefficients for each claim score in ELA/lit and mathematics, respectively.

Table 37. 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	0.77	2434.93	100.57	48.53
	Claim 2: Writing	11	11	0.80	2424.15	99.88	44.66
	Claim 3: Listening	8	8	0.60	2436.90	123.84	78.59
	Claim 4: Research	8	9	0.71	2431.41	120.51	65.33
4	Claim 1: Reading	14	16	0.74	2476.54	104.03	52.72
	Claim 2: Writing	11	11	0.79	2472.59	107.98	49.70
	Claim 3: Listening	8	8	0.60	2489.55	137.50	87.15
	Claim 4: Research	7	9	0.70	2477.08	121.09	66.09
5	Claim 1: Reading	14	16	0.76	2516.55	107.53	53.10
	Claim 2: Writing	11	11	0.79	2512.37	105.42	48.33
	Claim 3: Listening	8	9	0.61	2512.48	132.28	82.20
	Claim 4: Research	8	9	0.75	2521.42	116.93	59.03
6	Claim 1: Reading	14	16	0.77	2521.40	118.28	56.36
	Claim 2: Writing	11	11	0.81	2526.99	104.25	45.03
	Claim 3: Listening	8	9	0.63	2549.18	145.26	88.76
	Claim 4: Research	8	9	0.70	2532.14	123.37	67.37
7	Claim 1: Reading	14	16	0.79	2552.82	115.63	53.41
	Claim 2: Writing	11	11	0.80	2552.82	110.45	49.39
	Claim 3: Listening	8	9	0.59	2541.43	138.02	88.34
	Claim 4: Research	8	9	0.69	2552.54	132.47	73.23
8	Claim 1: Reading	16	16	0.78	2563.98	114.61	54.05
	Claim 2: Writing	11	11	0.80	2562.44	114.31	51.05
	Claim 3: Listening	8	9	0.62	2580.15	138.00	84.85
	Claim 4: Research	8	9	0.71	2569.44	128.62	69.48

Table 38. 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	0.90	2443.68	91.54	28.28
	Claims 2 & 4	8	11	0.78	2433.12	96.21	45.50
	Claim 3	9	11	0.77	2440.51	97.20	46.61
4	Claim 1	20	20	0.90	2485.94	88.71	27.78
	Claims 2 & 4	8	10	0.76	2478.70	93.83	45.94
	Claim 3	9	10	0.76	2481.50	98.59	48.46
5	Claim 1	20	20	0.89	2513.89	98.04	31.92
	Claims 2 & 4	8	10	0.68	2500.81	100.04	56.83
	Claim 3	9	10	0.72	2502.87	113.31	59.44
6	Claim 1	19	19	0.88	2523.37	114.06	38.87
	Claims 2 & 4	9	10	0.73	2509.03	118.35	61.87
	Claim 3	10	11	0.76	2518.07	122.30	59.65
7	Claim 1	20	20	0.89	2533.13	115.37	38.77
	Claims 2 & 4	10	10	0.63	2515.06	126.96	77.35
	Claim 3	8	10	0.66	2522.50	137.40	79.98
8	Claim 1	20	20	0.88	2548.65	127.25	43.77
	Claims 2 & 4	8	10	0.63	2538.59	135.85	82.30
	Claim 3	9	10	0.71	2537.02	143.13	77.40

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

6. SCORING

The Smarter Balanced Assessment Consortium provided the vertically-scaled item parameters by linking across all grades using common items in adjacent grades. All scores are estimated based on these item parameters. Each student received an overall scale score, an overall achievement level, and a performance category for each claim. This section describes the rules used in generating scores, as well as the 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 $m_i = 1$,

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

in the case of items with two or more points,

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

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 item(s) for both complete and incomplete tests. The upper bound of the SE is set to 2.5 on theta metric. Any value larger than 2.5 is truncated at 2.5 on theta metric.

The algorithm allows previously answered items to be changed; however, it does not allow items to be skipped. Item selection requires iteratively updating the estimate of the overall and 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 39 presents the scaling constants for each subject for the theta-to-scale score linear transformation. Scale scores are rounded to an integer.

Table 39. 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 40 provides three achievement standards for each grade and content area.

Table 40. 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, the LOSS and HOSS truncation rule was removed.

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 fewer) 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 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 target. Target scores are computed for attempted tests based on the responded items. Target scores are computed in each claim (four claims) for ELA/lit and only in claim 1 for mathematics. A target performance provides information on strengths and weaknesses on the target for a group of students, e.g., a class, a school, or a district, but 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, and Measurement Incorporated (MI) provides all handscoring for the Delaware Smarter Balanced summative assessments. In ELA/lit, short-answer (SA) items and full-write items are scored by human readers; this is also referred to as "handscoring." In mathematics, SA items and other constructed-response items are handscored. The procedure for scoring these items is provided by Smarter Balanced.

Outlined below is the scoring process MI follows. This procedure is used to score responses to all constructed-response or written composition items.

6.7.1 Reader Selection

MI maintains a large pool of readers at each scoring center, as well as distributive readers who work remotely from their homes. MI routinely maintains supervisors' evaluations and performance data for each person who works on each scoring project in order to determine employment eligibility for future projects. 2017–2018 was the fourth consecutive year that MI scored operational Smarter Balanced assessments, and the majority of readers recruited to score the 2017–2018 summative assessment had previous experience scoring Smarter Balanced assessments.

MI procedures for selecting new readers are very thorough. After advertising and receiving applications, MI staff review the applications and schedule interviews for qualified applicants (i.e., those with a four-year college degree). Each qualified applicant must pass an interview by experienced MI staff and provide references. MI then reviews all the information about an applicant before offering employment.

In selecting team leaders, MI management staff and scoring directors 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 readers 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, readers, 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 Reader Training

All readers hired for Smarter Balanced assessment handscoring are trained using the rubric(s), anchor sets, and training/qualifying sets provided by Smarter Balanced. These sets were created during the original field-test scoring in 2014 and approved by Smarter Balanced. The same anchor sets are used each year. Additionally, MI conducts an annual review of the reader 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, readers 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). Readers are trained on a specific item type (i.e., brief writes, reading, research, full-writes, and/or mathematics). Within each group, readers are divided into teams consisting of one team leader and 10–15 readers. Each team leader and reader is assigned a unique number for easy identification of their scoring work throughout the scoring session. The number of items an individual reader scores is minimized so that the reader 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. Readers 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 readers.

After the contracts and nondisclosure forms are signed and the scoring director completes his or her introductory remarks, training begins. Reader 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 readers 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 reader monitoring.

After completing the first training set, readers 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, readers 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 readers unable to meet the qualifying standards are not permitted to score that item. Readers who reach the qualifying standard on some items but not others will only score the items on which they have successfully qualified. All readers understand this stipulation when they are hired.

Training is carefully orchestrated so that readers 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 accurately score all responses. In addition to completing all of 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: Readers 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: Readers train and qualify on a baseline set within a specific grade band and target.
- Mathematics: Readers train on baseline items, which qualify the readers for that item as well as any items associated with it; for items with no associated items, training is for the specific item.

Reader 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. Readers generally work 6.5 hours per day, excluding breaks. Evening shift readers work 3.75 hours, excluding breaks.

Multiple strategies are used to minimize rater bias. First, readers 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 readers have no knowledge of student characteristics. Second, all readers 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 readers' 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, readers are monitored and any instances of readers making scoring decisions based on anything except the criteria are discussed. Readers are further monitored, and if any continue to exhibit bias after receiving a reasonable amount of feedback, they are dismissed.

6.7.3 Reader 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 reader.

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 reader's work throughout every project. Reader status reports are used to monitor readers' scoring habits during the Smarter Balanced handscoring project.

MI has developed and operates a comprehensive system for collecting and analyzing scoring data. After the readers' 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, NC.

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:

- Reader 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 reader
- Percentage of responses scored that day within one point of agreement with a second reader
- 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 utilized 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 retraining of individual readers or the group as a whole. It can easily be determined if a reader is consistently scoring high or low, and the specific score points with which they may be having difficulty. The scoring directors share such information with the team leaders and direct all retraining efforts.

6.7.4 Reader Monitoring and Retraining

Team leaders spot-check (i.e., read-behind) each reader's scoring to ensure that he or she is on target and conduct one-on-one retraining sessions addressing any problems found. At the beginning of the project, team leaders read behind every reader every day; they become more selective about the frequency and number of read-behinds as readers become more proficient at scoring. The daily reader reliability reports and validity/calibration results are used to identify readers who need more frequent monitoring.

Retraining is an ongoing process once scoring is underway. Daily analysis of the reader status reports enables management personnel to identify individual or group retraining 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 retraining procedures include room-wide discussions led by the scoring director, team discussions conducted by team leaders, and one-on-one discussions with individual readers. It is standard practice to conduct morning room-wide retraining at MI each day, with a more extensive retraining on Monday mornings in order to re-anchor the readers after a weekend away from scoring.

Each student response is scored holistically by a trained and qualified reader 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 readers who are not aware of the score assigned by the first reader or even that the response has been read before. MI's QA/reliability procedures allow the handscoring staff to identify struggling readers very early and begin retraining at once. While retraining these readers, MI also monitors their scoring intensively to ensure that all responses are scored accurately. In fact, MI's monitoring is also used as a retraining method. MI shows readers responses that the readers have scored incorrectly, explains the correct scores, and has the readers change the scores.

During scoring, readers 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 Reader Validity Checks

Approved responses are loaded into the VSC system as validity responses. 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. These responses are imbedded into live scoring on an ongoing basis to be scored by the readers. A validity report is generated that includes the response identification number, the score(s) assigned by the readers, 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. If it is determined that a validity response and/or item is performing poorly, scoring management reviews the validity responses to ensure that the true scores have been entered correctly. If so, then retraining may be conducted with the readers using the validity data as a guide for how to focus the retraining. If the true scores have been entered incorrectly, then the database is updated to show the correct true scores. 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 retraining and dismissing readers.

6.7.6 Reader Dismissal

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

6.7.7 Reader Agreement

The inter-reader reliability (IRR) is computed-based on scorable responses (numeric scores) that are scored by two independent readers only, excluding non-scorable responses (e.g., off topic, off purpose, or foreign language responses) that are scored by scoring leadership, not by two independent readers. The inter-reader reliability is based on the readers who scored student responses in Delaware.

In ELA/lit, writing essay item responses (full-writes) are scored in three dimensions: convention (0–2 rubric), evidence/elaboration (0–4 rubric), and organization/purpose (0–4 rubric). The short-answer items are scored in 0–2 rubric. In mathematics, the maximum score points for handscored items range from 1–3.

Tables 41–43 provide a summary of the inter-reader reliability based on items with a sample size greater than 50. The inter-reader reliability is presented with average of % exact agreement, minimum and maximum % exact agreements, combined % exact and % adjacent agreement, and quadratic weighted Kappa (QWK).

Table 41. ELA/Lit Reader Agreements for Short-Answer Items

Grade	# of Items	% Exact			% (Exact+ Adjacent)	QWK
		Average	Min	Max		
3	35	80	63	95	100	0.74
4	53	78	62	95	100	0.75
5	48	76	59	92	100	0.74
6	43	76	62	94	100	0.71
7	49	75	60	95	100	0.73
8	49	73	54	92	100	0.70

Table 42. ELA/Lit Reader Agreements for Full-Write Items

Grade	Dimensions	# of Items	% Exact			% (Exact+ Adjacent)	QWK
			Average	Min	Max		
3	Conventions	13	72	65	83	100	0.61
	Evid/Elab	13	72	63	78	99	0.71
	Org/Purp	13	72	62	80	99	0.72
4	Conventions	18	70	57	81	100	0.63
	Evid/Elab	18	67	59	79	98	0.69
	Org/Purp	18	69	59	80	99	0.70
5	Conventions	20	69	53	81	100	0.54
	Evid/Elab	20	65	55	79	99	0.69
	Org/Purp	20	65	55	79	99	0.69
6	Conventions	14	68	59	79	98	0.57
	Evid/Elab	14	65	55	74	98	0.68
	Org/Purp	14	66	55	74	98	0.68
7	Conventions	19	72	61	85	99	0.60
	Evid/Elab	19	71	61	81	99	0.70
	Org/Purp	19	73	64	81	99	0.72
8	Conventions	20	77	62	93	99	0.55
	Evid/Elab	20	71	56	81	99	0.74
	Org/Purp	20	71	57	80	100	0.75

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

Table 43. Mathematics Reader Agreements

Grade	Score Points	# of Items	% Exact			% (Exact+ Adjacent)	QWK
			Average	Min	Max		
3	1	12	94	91	97	100	0.86
3	2	26	92	80	99	100	0.93
3	3	4	97	95	98	100	0.97
4	1	8	83	73	95	100	0.59
4	2	36	90	78	100	100	0.88
4	3	4	90	88	95	100	0.94
5	1	4	90	87	95	100	0.50
5	2	41	90	77	98	100	0.87
5	3	8	90	84	100	98	0.85
6	1	13	97	86	99	100	0.91
6	2	32	90	82	100	100	0.90
7	1	8	97	94	99	100	0.78
7	2	24	89	77	97	100	0.82
7	3	1	77	77	77	97	0.85
8	1	14	92	84	98	100	0.82
8	2	25	90	81	100	100	0.89

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. Additionally, the ORS provides participation data that helps monitor student participation rates.

This section contains a description of the types of scores reported in the ORS and a description of 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 to be 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.

Generally, the ORS provides two categories of online score reports: (1) aggregate score reports, and (2) student score reports. Table 44 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 44. 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 standard error of measurement • 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 45 presents the types of subgroups and subgroup categories provided in ORS.

Table 45. 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

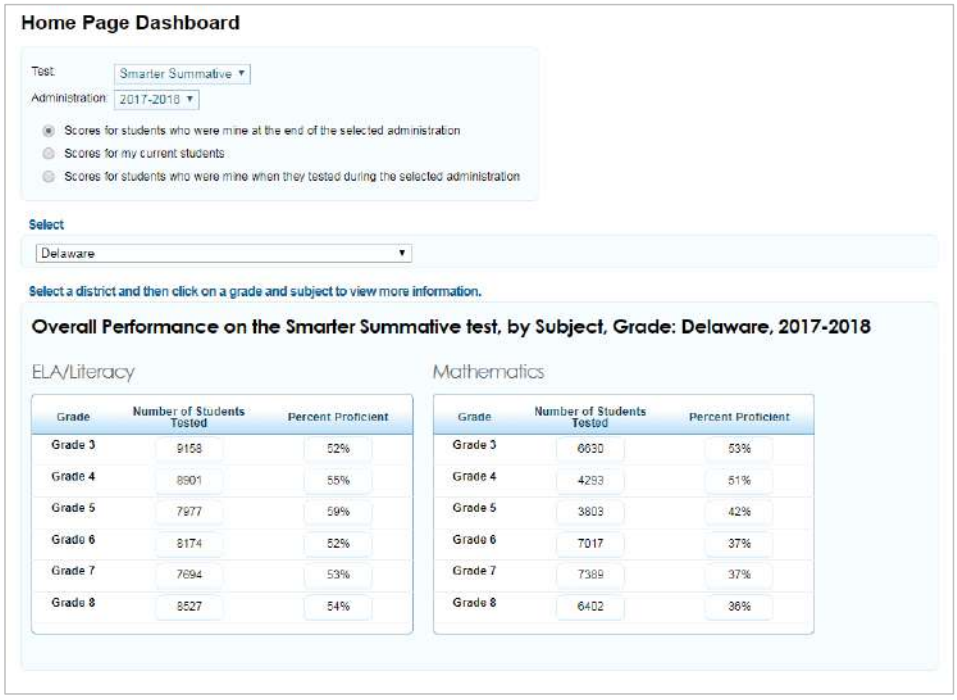
7.1.2 Online Reporting System

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.

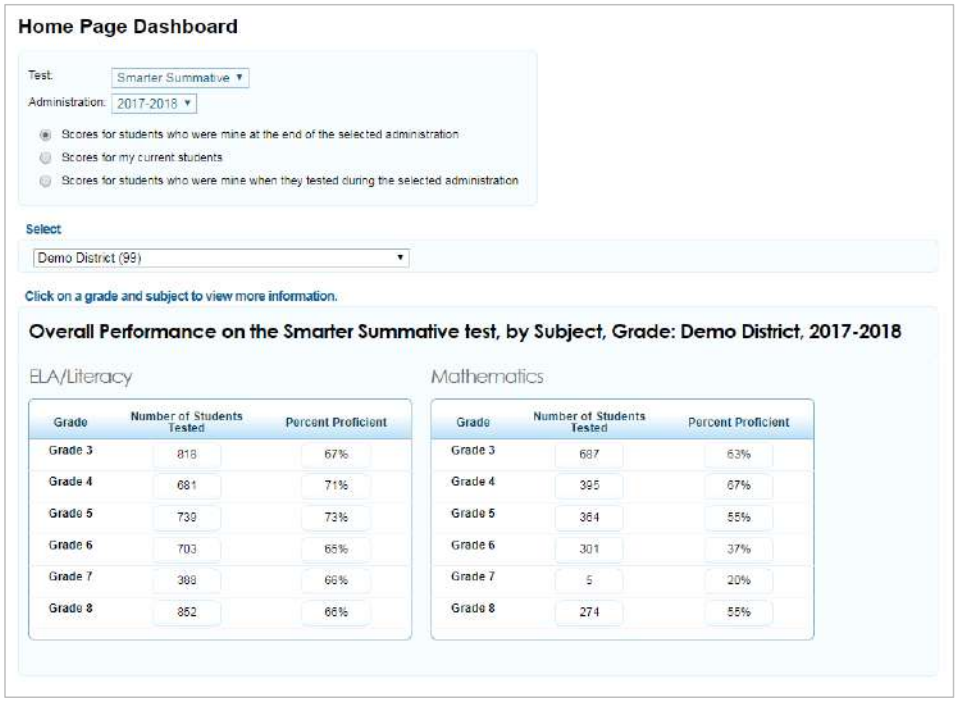
Exhibit 1. Home Page: State Level



Grade	Number of Students Tested	Percent Proficient
Grade 3	9158	52%
Grade 4	8901	55%
Grade 5	7977	59%
Grade 6	8174	52%
Grade 7	7694	53%
Grade 8	8527	54%

Grade	Number of Students Tested	Percent Proficient
Grade 3	6630	53%
Grade 4	4293	51%
Grade 5	3803	42%
Grade 6	7017	37%
Grade 7	7389	37%
Grade 8	6402	36%

Exhibit 2. Home Page: District Level

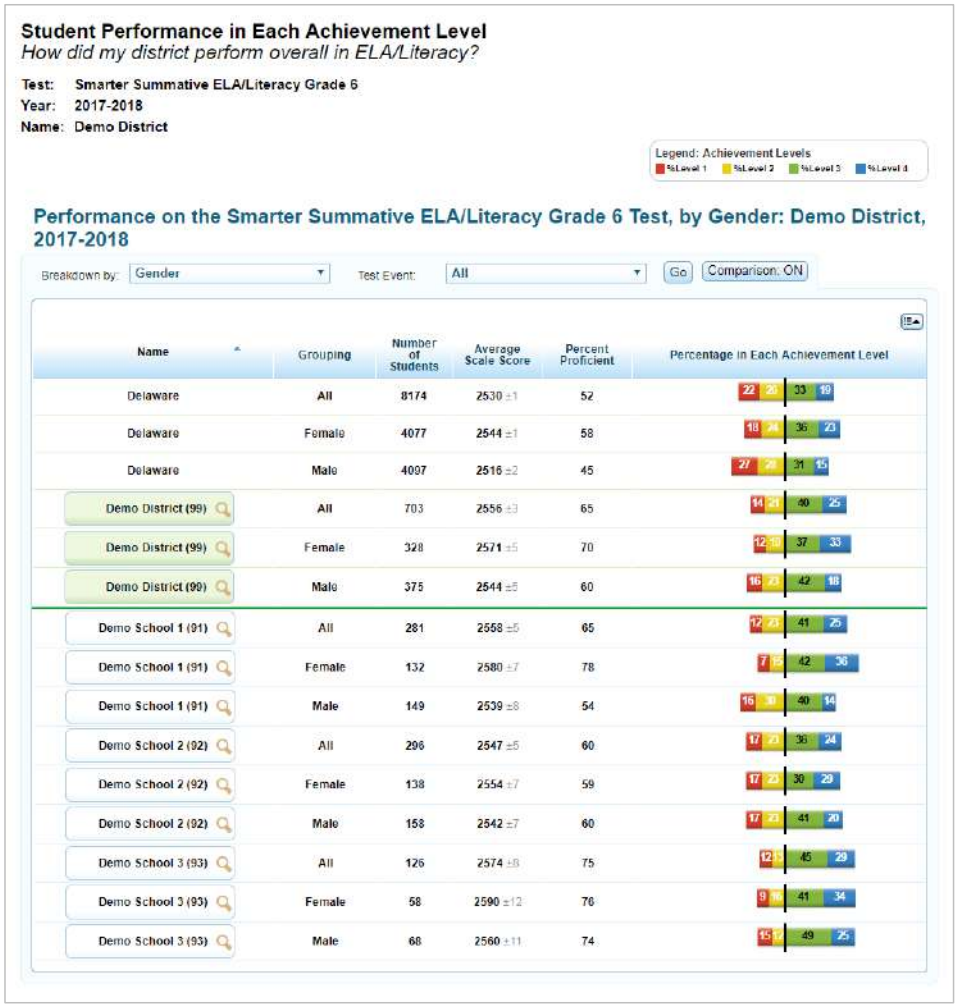


7.1.2.2 Subject Detail Page

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 roficient, 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: 2017-2018
Name: 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: Demo District, 2017-2018

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	8174	2530 ±1	52	22 26 33 19
Delaware	Female	4077	2544 ±1	58	18 24 36 23
Delaware	Male	4097	2516 ±2	45	27 28 31 15
Demo District (99)	All	703	2556 ±3	65	14 21 40 25
Demo District (99)	Female	328	2571 ±5	70	12 18 37 33
Demo District (99)	Male	375	2544 ±5	60	16 23 42 18
Demo School 1 (91)	All	281	2558 ±5	65	12 23 41 25
Demo School 1 (91)	Female	132	2580 ±7	78	7 15 42 36
Demo School 1 (91)	Male	149	2539 ±8	54	16 30 40 14
Demo School 2 (92)	All	296	2547 ±5	60	17 23 36 24
Demo School 2 (92)	Female	138	2554 ±7	59	17 23 30 29
Demo School 2 (92)	Male	158	2542 ±7	60	17 23 41 20
Demo School 3 (93)	All	126	2574 ±8	75	12 13 45 29
Demo School 3 (93)	Female	58	2590 ±12	76	9 16 41 34
Demo School 3 (93)	Male	68	2560 ±11	74	15 10 49 25

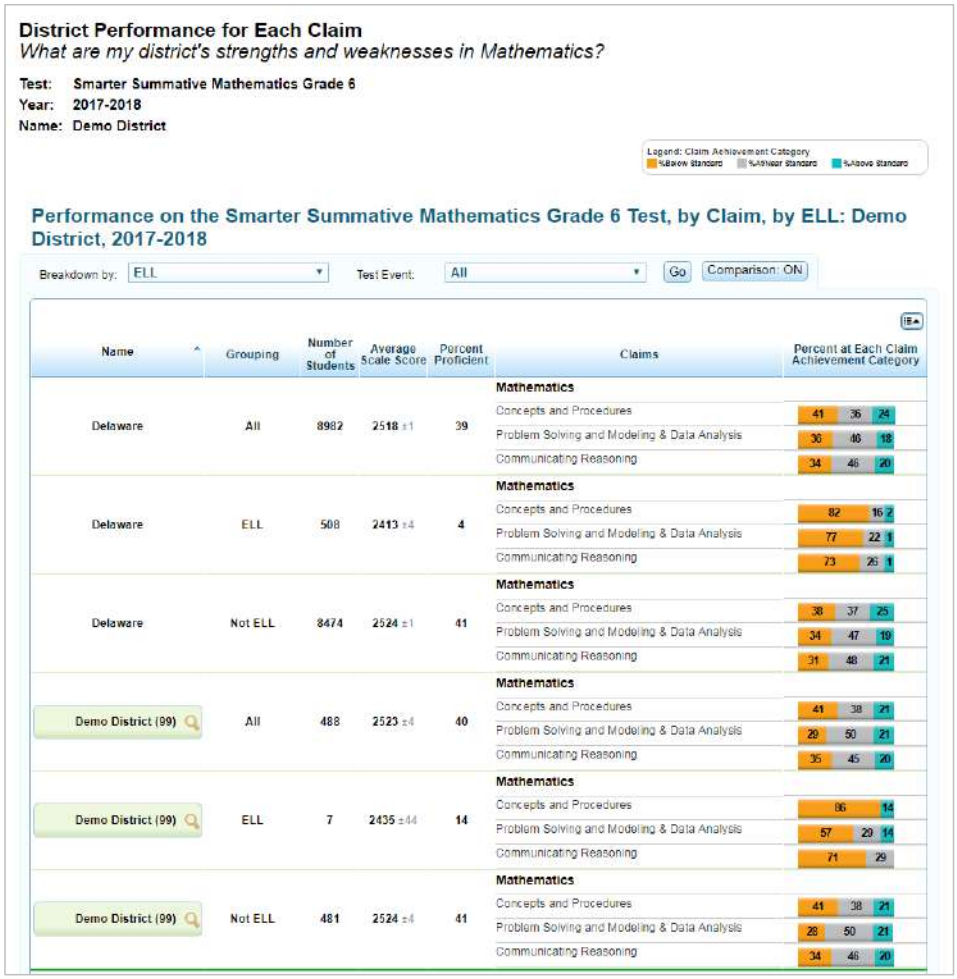
7.1.2.3 Claim Detail Page

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.

Similar to 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 pages for mathematics at a district level when users select a subgroup of ELL.

Exhibit 4. Claim Detail Page for Mathematics by ELL: District Level



District Performance for Each Claim
What are my district's strengths and weaknesses in Mathematics?

Test: Smarter Summative Mathematics Grade 6
Year: 2017-2018
Name: Demo District

Legend: Claim Achievement Category
%Below Standard | %Near Standard | %Above Standard

Performance on the Smarter Summative Mathematics Grade 6 Test, by Claim, by ELL: Demo District, 2017-2018

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	8982	2518 ±1	39	**Mathematics**	
					Concepts and Procedures	41 36 24
					Problem Solving and Modeling & Data Analysis	36 46 18
					Communicating Reasoning	34 46 20
Delaware	ELL	508	2413 ±4	4	**Mathematics**	
					Concepts and Procedures	82 16 2
					Problem Solving and Modeling & Data Analysis	77 22 1
					Communicating Reasoning	73 26 1
Delaware	Not ELL	8474	2524 ±1	41	**Mathematics**	
					Concepts and Procedures	38 37 25
					Problem Solving and Modeling & Data Analysis	34 47 19
					Communicating Reasoning	31 48 21
Demo District (99)	All	488	2523 ±4	40	**Mathematics**	
					Concepts and Procedures	41 38 21
					Problem Solving and Modeling & Data Analysis	29 50 21
					Communicating Reasoning	35 45 20
Demo District (99)	ELL	7	2435 ±44	14	**Mathematics**	
					Concepts and Procedures	86 14
					Problem Solving and Modeling & Data Analysis	57 29 14
					Communicating Reasoning	71 29
Demo District (99)	Not ELL	481	2524 ±4	41	**Mathematics**	
					Concepts and Procedures	41 38 21
					Problem Solving and Modeling & Data Analysis	28 50 21
					Communicating Reasoning	34 46 20

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 roster levels.

Exhibit 5. Target Detail Page for ELA/L: 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: 2017-2018
Name: Demo School

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

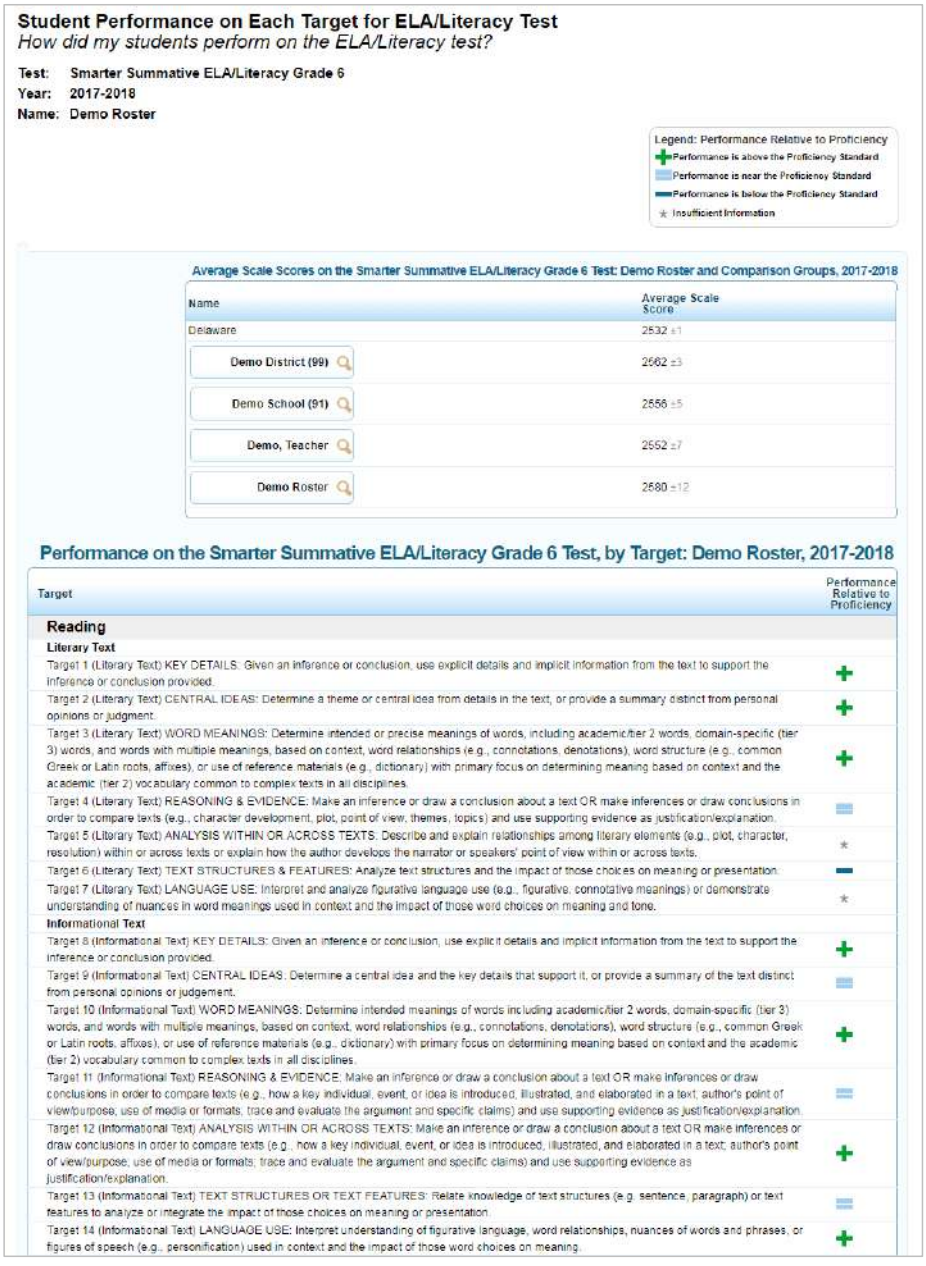
Average Scale Scores on the Smarter Summative ELA/Literacy Grade 6 Test: Demo School and Comparison Groups, 2017-2018

Name	Average Scale Score
Delaware	2532 ±1
Demo District (99)	2562 ±3
Demo School (91)	2556 ±5

Performance on the Smarter Summative ELA/Literacy Grade 6 Test, by Target: Demo School, 2017-2018

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: 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 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/L: Roster Level



Student Performance on Each Target for ELA/Literacy Test

How did my students perform on the ELA/Literacy test?

Test: Smarter Summative ELA/Literacy Grade 6
Year: 2017-2018
Name: Demo Roster

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

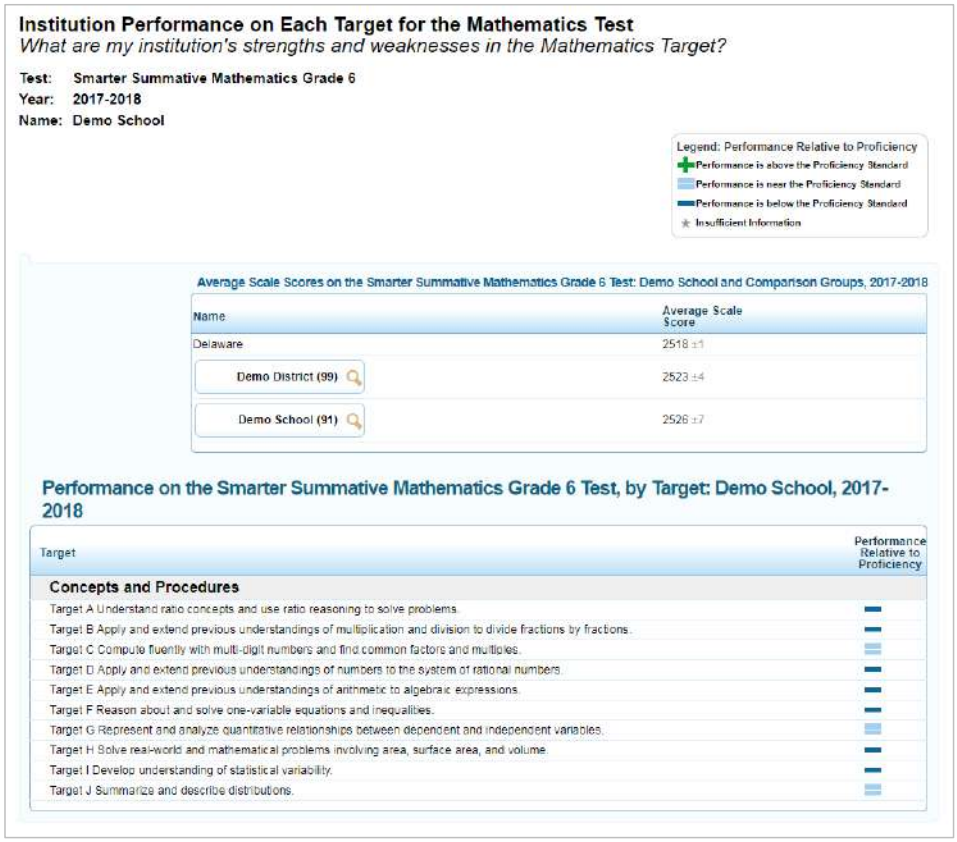
Average Scale Scores on the Smarter Summative ELA/Literacy Grade 6 Test: Demo Roster and Comparison Groups, 2017-2018

Name	Average Scale Score
Delaware	2532 ±1
Demo District (99)	2562 ±3
Demo School (91)	2556 ±5
Demo, Teacher	2552 ±7
Demo Roster	2580 ±12

Performance on the Smarter Summative ELA/Literacy Grade 6 Test, by Target: Demo Roster, 2017-2018

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: 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 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: 2017-2018
Name: Demo School

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

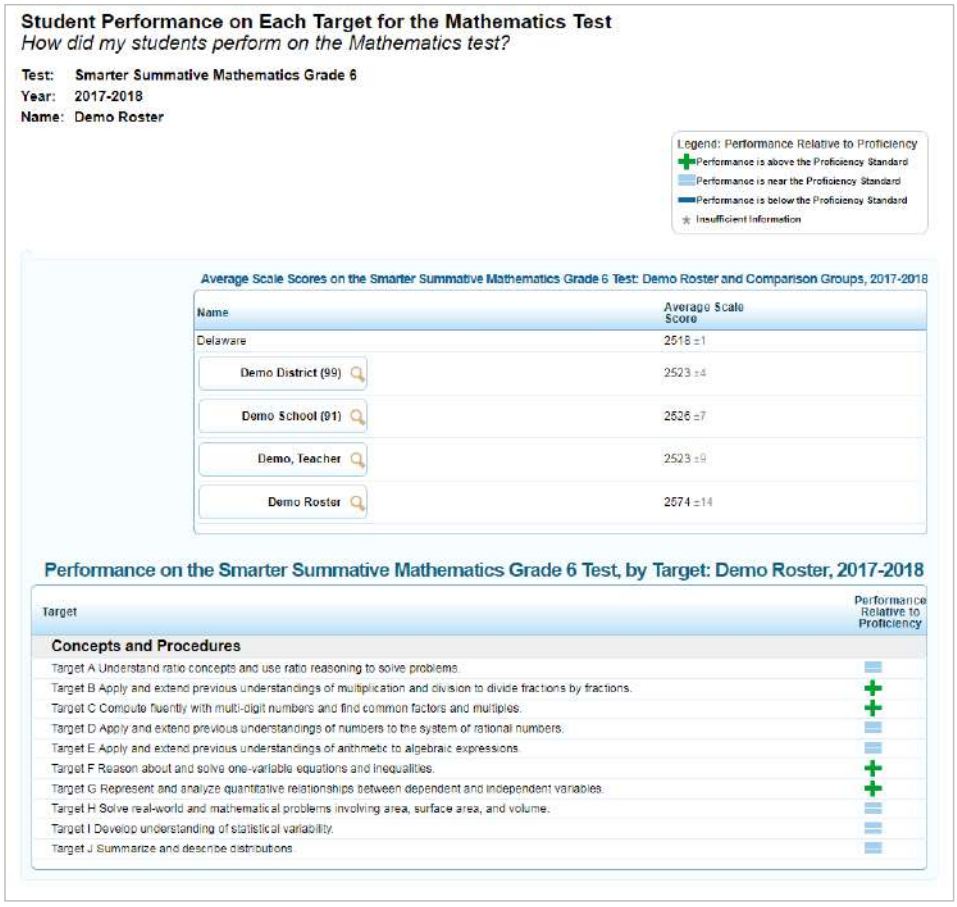
Average Scale Scores on the Smarter Summative Mathematics Grade 6 Test: Demo School and Comparison Groups, 2017-2018

Name	Average Scale Score
Delaware	2518 ±1
Demo District (99)	2523 ±4
Demo School (91)	2526 ±7

Performance on the Smarter Summative Mathematics Grade 6 Test, by Target: Demo School, 2017-2018

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: Roster Level



Student Performance on Each Target for the Mathematics Test
How did my students perform on the Mathematics test?

Test: Smarter Summative Mathematics Grade 6
Year: 2017-2018
Name: Demo Roster

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 Roster and Comparison Groups, 2017-2018

Name	Average Scale Score
Delaware	2518 ±1
Demo District (99)	2523 ±4
Demo School (91)	2526 ±7
Demo, Teacher	2523 ±9
Demo Roster	2574 ±14

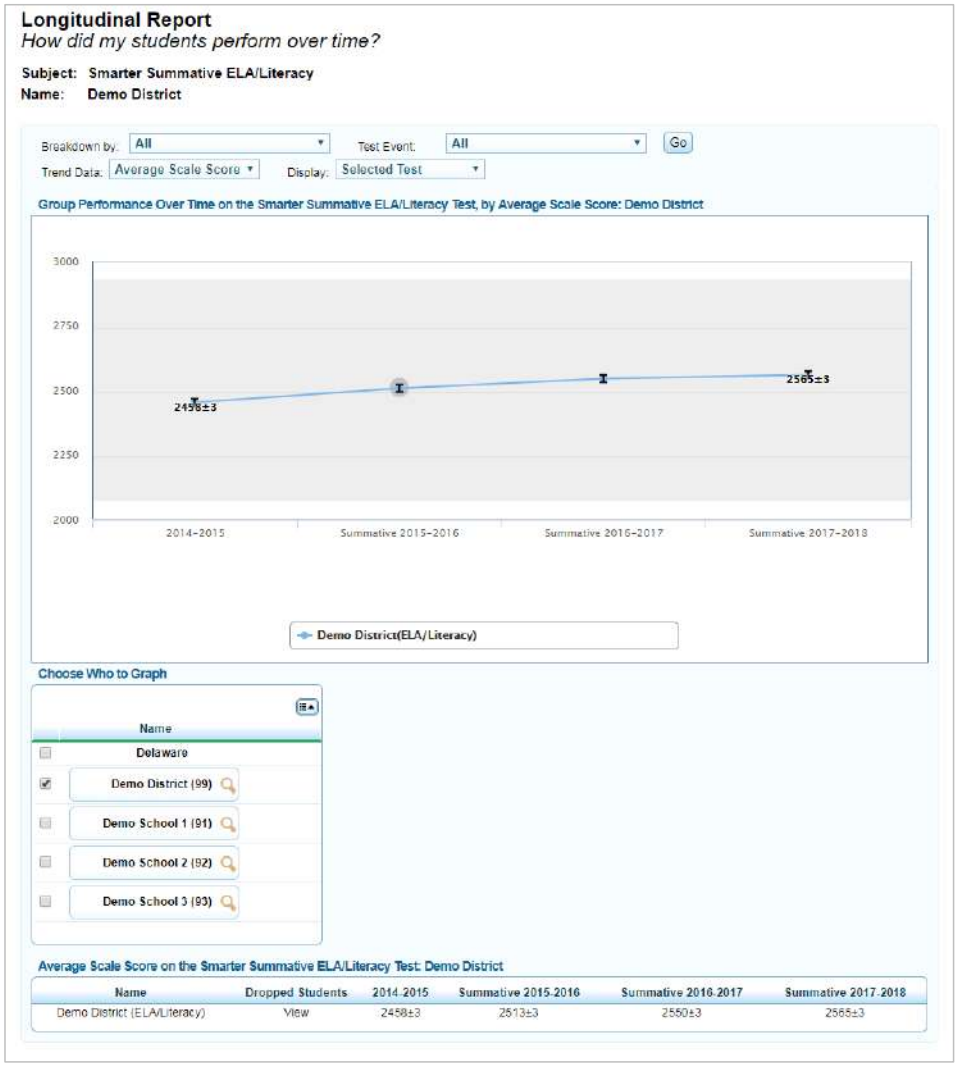
Performance on the Smarter Summative Mathematics Grade 6 Test, by Target: Demo Roster, 2017-2018

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 for an aggregate (e.g., the state, district, and school) over time. The trend 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 demographic subgroups. Exhibit 9 provides an example of trend report pages for ELA/lit at the district level.

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



7.1.2.6 Student Detail Page

When a student completes a test, and the test is handscored, an online score report appears in the student detail page in the ORS. The student detail page shows individual student performance on the test. In each subject area, the student detail page provides: (1) scale score and standard error of measurement, (2) achievement level for overall test, (3) performance category in each claim, (4) average scale scores for student's state, district, school, teacher and associated standard errors of the average scale scores, and (5) student performance growth over time.

Specifically, the student's name, scale score with standard error of measurement, 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 standard error of measurement using a "±"sign. Standard error of measurement 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, which 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 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 standard error of measurement 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.

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

Exhibit 10. Student Detail Page for ELA/Lit

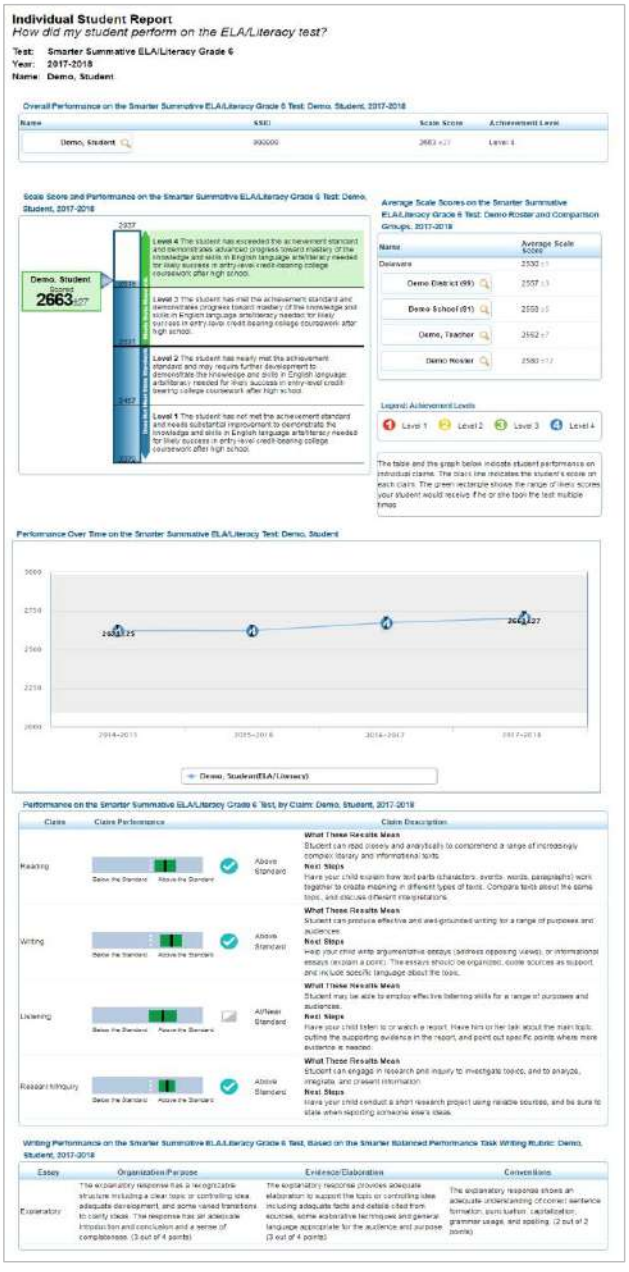
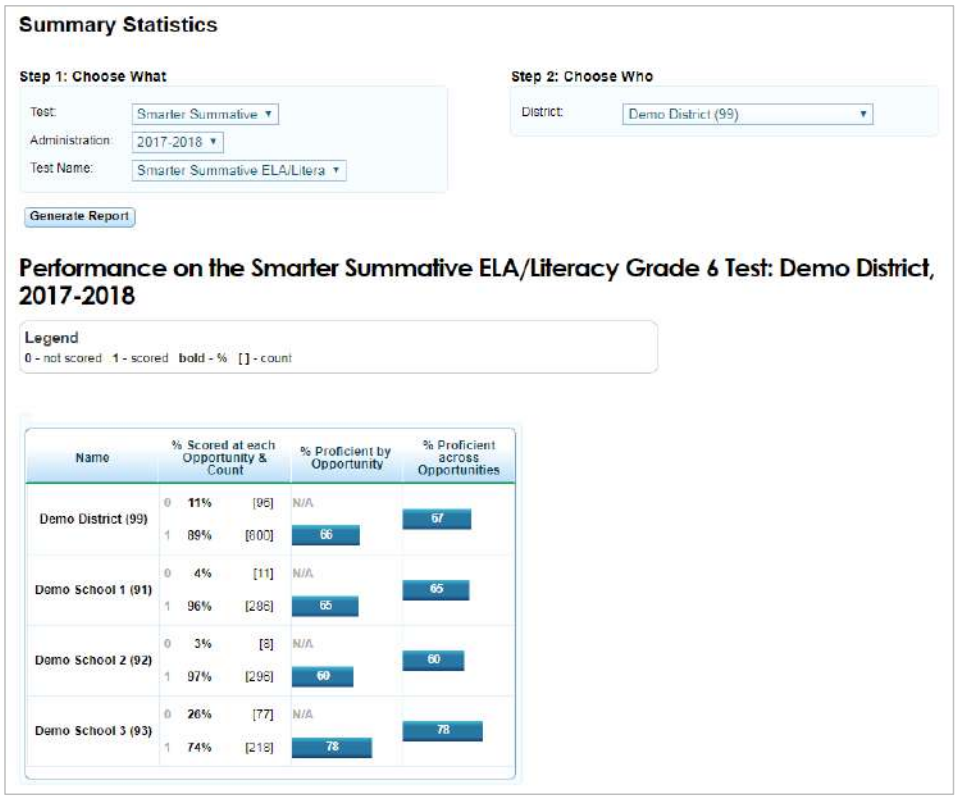


Exhibit 11. Student Detail Page for Mathematics

7.1.2.7 Participation Rate

In addition to online score reports, the ORS provides participation rate reports for districts and schools to help monitor the student participation rate. Participation data are updated each time students complete tests, and they are handscored. Included in the participation table are: (1) the number and percentage of students who are tested and not tested, and (2) the percentage proficient. Exhibit 12 presents a sample participation rate report at the district level.

Exhibit 12. Participation Rate Report at District Level



7.2 PAPER FAMILY SCORE REPORTS

After the testing window is closed, parents whose children participated in a test receive a full-color paper score report (hereinafter 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 with additional information on student performance. For example, the family report includes a progress chart that displays student's performance for each school year. The progress chart shows whether a student's performance meets the standards in each year and how much the student's performance increases. Exhibits 13 and 14 present examples of paper family score reports for grade 4 ELA/lit and mathematics.

Exhibit 13. Sample Paper Family Score Report for Grade 4 ELA/Lit

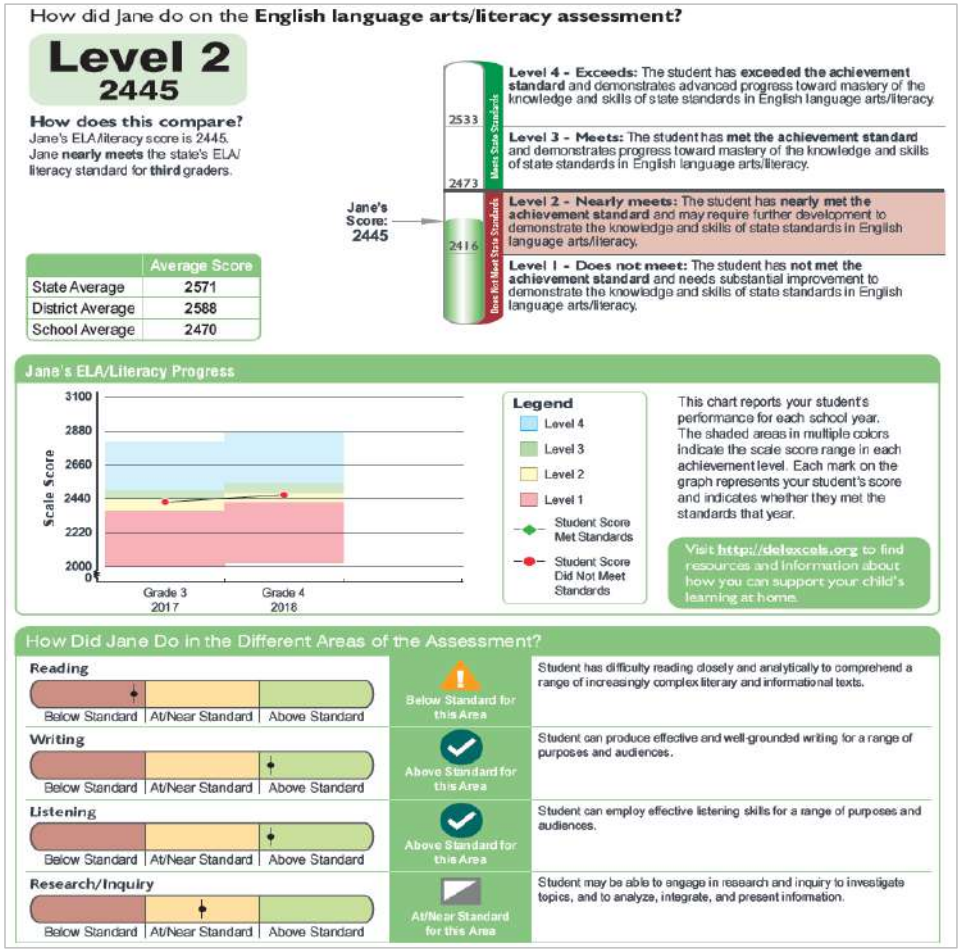
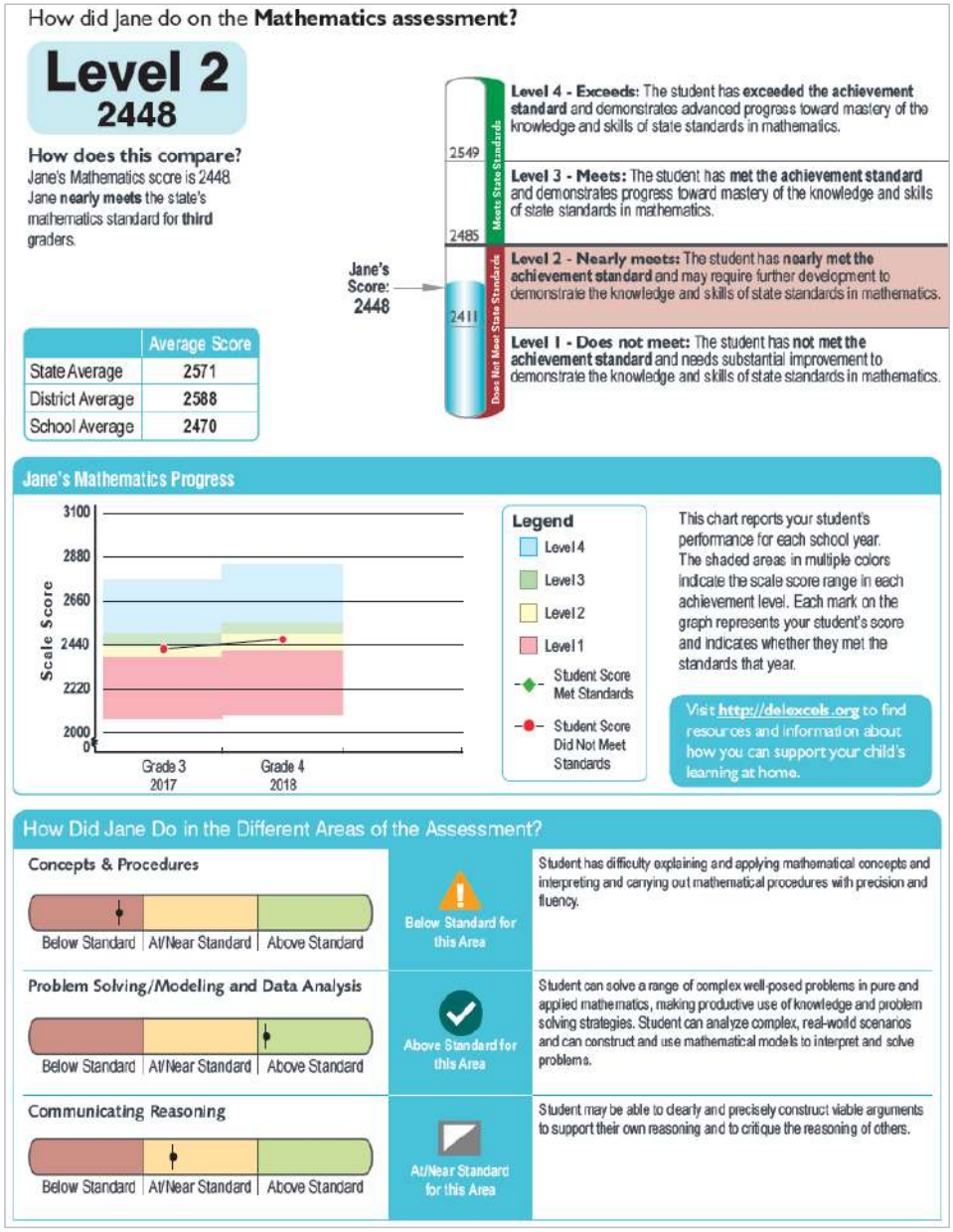


Exhibit 14. Sample Paper Family Score Report for Grade 4 Mathematics



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 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 would vary across administrations, sometimes a little higher, a little lower, or the same. The standard error of measurement (SEM) represents the precision of the scale score, or the range in which the student would likely score if a similar test was administered multiple times. When interpreting scale scores, it is recommended to consider the range of scale scores, incorporating the SEM of the scale score.

The ± next to the student's scale score provides information about the certainty, or confidence, of the score's interpretation. The boundaries of the score band are one SEM above and below the student's observed scale score, representing a range of score values that is likely to contain the true score. For example, 2680 ± 10 indicates that if a student was tested again, it is likely that the student would receive a score between 2670 and 2690. SEM can be different for the same scale score, depending on how closely the administered items match the student's ability.

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

Student performance on each claim is reported in three categories: (1) Below Standard, (2) At/Near Standard, and (3) Above Standard. Unlike the achievement level for the overall test, student performance on each 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 student 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 strangths 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 make decisions on 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 in 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 Smarter Balanced assessment 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 in both online and paper-pencil formats. The quality of the information produced in the test delivery system (TDS) is tested thoroughly before, during, and after the testing window opens.

8.1 ADAPTIVE TEST CONFIGURATION

For the CAT, 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 specification, slopes and intercepts for theta-to-scale score transformation, cut scores, and the item information (i.e., answer keys, item attributes, item parameters, and passage information). The accuracy of the information in the configuration file is independently checked and confirmed numerous times by multiple staff members before the testing window opens.

To verify the accuracy of the scoring engine, we use simulated test administrations. The simulator generates a sample of students with an ability distribution that matches that of the population (Smarter Balanced Assessment Consortium states). The ability of each simulated student is used to generate a sequence of item response scores consistent with the underlying ability distribution. These simulations provide a rigorous test of the adaptive algorithm for adaptively administered tests, as well as a check of form distributions (if administering multiple test forms) and test scores in fixed-form tests.

Simulations are generated using the production item selection and scoring engine to ensure that verification of the scoring engine is based on a wide range of student response patterns. The results of simulated test administrations are used to configure and evaluate the adequacy of the item selection algorithm used to administer the Smarter Balanced summative assessments. The purpose of the simulations is 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.

After the adaptive test simulations, another set of simulations for the combined tests (computer-adaptive test component plus a fixed-form performance task component) are performed to check scores. The simulated data are used to check whether the scoring specifications were applied accurately. The scores in the simulated data file are checked independently, following the scoring rules specified in the scoring specifications.

8.1.1 Platform Review

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

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

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

8.1.2 User Acceptance Testing and Final Review

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

8.2 Quality Assurance in Document Processing

The Smarter Balanced assessments are administered primarily online; however, a few students take paper-pencil assessments. When test documents are scanned, a quality control sample of documents consisting of 10 test cases per document type (normally between 500 and 600 documents) was created so that all possible responses and all demographic grids were verified, including various typical errors that required editing via 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 scanner(s) was correct. MI staff carefully compared the documents and the data file created from them to further ensure that the results from the scanner, the editing process (validation and data correction), and the transfer to the 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 quality assurance (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, 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 retraining and/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 retraining sessions when necessary. MI's QA procedures allow scoring staff to identify struggling scorers very quickly and to begin retraining 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. Retraining 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 retraining 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 also directly observes MI activities virtually. MI provides virtual access to the training activities through the online training interface. The DDOE monitors the scoring process through the Client Command Center (CCC) and has access to view and run specific 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 the load 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, our 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 46 presents an overview of the QA reports.

Table 46. 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 unexpected low blueprint match rates	Early detection of unexpected blueprint match issue
Item Exposure Rates	To monitor unlikely high exposure rates of items or passages or unusually low item pool usage (highly unused items/passages)	Early detection of any oversight in the blueprint specification
Cheating Analysis	To monitor testing irregularities	Early detection of testing irregularities

8.5.1 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 for online assessments are assigned by automated systems in real time. For machine-scored portions of assessments, the machine rubrics are created and reviewed along with the items, then validated and finalized during rubric validation following field testing. The review process "locks down" the item and rubric when the item is approved for web display (Web Approval). During operational testing, actual item responses are compared to expected item responses (given the item response theory [IRT] parameters), which can detect miskeyed items, item score distribution, or other scoring problems. Potential issues are automatically flagged in reports available to our psychometricians.

The handscoring processes include rigorous training, validity and reliability monitoring, and back-reading to ensure accurate scoring. Handscored items are paired to the machine-scored items by our Test Integration System (TIS). The integration is based on identifiers that are never separated from their data and are checked by our quality assurance (QA) system. The integrated scores are sent to our test-scoring system, a mature, well-tested, real-time system that applies client-specific scoring rules and assigns scores from the calibrated items, including calculating achievement-level indicators, subscale scores and other features, which then pass automatically to the reporting system and Database of Record (DoR). The scoring system is tested extensively before deployment, including hand checks of scored tests and large-scale simulations to ensure that point estimates and standard errors are correct.

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 of the QA system's validation checks. All of the above processes take milliseconds to complete so that within less than a second of handscores being received by AIR and passing 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-upon procedures. Custom programming is implemented independently by two statistical programming teams working from the specifications. Only when the output from both teams matches exactly are the scripts released for production. Quality control, however, does not stop there.

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 do 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. Prior to printing the reports, AIR provides a live data file and individual student reports with sample districts for the 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.

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 2017–2018 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 it twice. Table A–1 presents the number of students who took the ICA, and Table A–2 presents the ICA results for all students, including the average and 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 Once or Twice

Grade	ELA/Lit			Mathematics		
	Once	Twice	Total	Once	Twice	Total
3	650	1	651	655	1	656
4	377	0	377	572	4	576
5	573	0	573	564	2	566
6	125	1	126	548	1	549
7	309	7	316	538	2	540
8	235	0	235	652	1	653

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

Subject	Grade	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
ELA/Lit	3	652	2407.91	78.01	33	33	19	16	34
	4	377	2466.02	83.21	29	25	24	22	46
	5	573	2494.03	84.37	28	25	32	15	47
	6	127	2491.95	85.98	38	29	25	8	33
	7	323	2523.20	94.12	34	26	30	10	40
	8	235	2533.62	97.06	33	28	31	8	39
Math	3	657	2419.56	69.33	29	33	27	11	39
	4	580	2452.83	69.22	26	42	25	7	32
	5	568	2492.81	77.83	32	36	19	14	32
	6	550	2490.92	84.61	41	34	17	8	24
	7	542	2514.09	92.11	37	35	18	10	27
	8	654	2535.22	101.76	39	27	22	12	34

Note: Number Tested is based on the total tests, adding multiple times for the students who took the same test more than once. The percentage of each achievement level may not add up to 100% or %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 3,905 students took the IABs, and among these students, 1,763 students took one IAB, 1,303 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, 1,763 students in grade 3 ELA/lit took one IAB only. Among these students, six students took the Brief Writes IAB. 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/L										
3	3,905	1,763	1,303	481	267	53	28	10		
4	3,402	1,495	1,086	521	172	63	32	20	13	
5	3,720	1,612	1,198	462	147	147	91	9	13	41
6	4,513	1,881	1,616	279	231	262	212	32		
7	4,313	2,301	1,253	530	165	45	17	2		
8	2,411	1,504	704	193	10					
Mathematics										
3	4,953	1,683	1,461	1,049	477	275	8			
4	5,316	2,062	1,852	1,026	270	96	10			
5	5,299	2,050	1,504	1,002	402	337	4			
6	5,620	1,962	2,018	843	513	281	3			
7	5,062	2,156	1,554	938	356	58				
8	4,872	1,958	1,481	943	232	247	11			

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	6	27	8	4	1			
	Editing	217	355	244	206	39	27	10		
	Language and Vocabulary Use	361	602	376	223	36	27	10		
	Listening and Interpretation	152	344	191	159	16	27	10		
	Reading Informational Text	514	451	148	115	51	26	10		
	Reading Literary Text	379	581	213	175	50	28	9		
	Research	32	198	151	106	31	25	10		
	Revision	10	64	91	72	20	5	10		
	Performance Task	92	5	2	4	18	2	1		
4	Brief Writes			1	5			8	13	
	Editing	263	186	188	74	53	19	20	13	
	Language and Vocabulary Use	189	275	251	131	57	29	20	13	
	Listening and Interpretation	69	117	209	115	41	31	20	13	
	Reading Informational Text	582	729	345	147	41	29	12		
	Reading Literary Text	219	625	406	135	47	27	20	13	
	Research	168	166	85	55	53	16	20	13	
	Revision	3	3	21	4	21	28	20	13	
	Performance Task	2	71	57	22	2	13		13	

Grade	Block	Number of IABs Taken								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
5	Brief Writes	1	5	25	30	4	7	8	12	41
	Editing	71	302	191	95	133	86	9	11	41
	Language and Vocabulary Use	266	564	329	108	138	88	9	7	41
	Listening and Interpretation	203	72	166	69	131	91	8	13	41
	Reading Informational Text	810	582	248	70	36	21	3	13	41
	Reading Literary Text	201	591	212	82	54	89	9	13	41
	Research	9	31	63	64	118	85	8	11	41
	Revision	32	247	145	42	117	72	8	11	41
	Performance Task	19	2	7	28	4	7	1	13	41

Table A-5: ELA/L 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	22	26	35	11	4	43	1		
	Editing	318	265	229	203	259	212	32		
	Language and Vocabulary Use	198	145	178	187	259	211	32		
	Listening and Interpretation	26	147	15	156	241	174	32		
	Reading Informational Text	109	766	27	45	32	162	32		
	Reading Literary Text	852	1,082	61	23	51	96	31		
	Research	150	552	75	196	255	168	32		
	Revision	206	249	217	103	209	206	32		
	Performance Task									
7	Brief Writes	19	72	27	46	2				
	Editing	204	276	461	162	42	17	2		
	Language and Vocabulary Use	399	239	422	117	43	17	2		
	Listening and Interpretation	126	135	101	60	19	15	2		
	Reading Informational Text	318	513	49	65	10	3	2		
	Reading Literary Text	1,066	744	305	106	35	17	2		
	Research	16	127	18	59	34	16	2		
	Revision	153	398	206	45	38	16	2		
	Performance Task		2	1		2	1			
8	Brief Writes	2	12	113	6					
	Editing and Revising	433	230	187	10					
	Listening and Interpretation	58	163	22	4					
	Reading Informational Text	129	456	63	2					
	Reading Literary Text	847	366	171	8					
	Research	35	181	20	4					
	Performance Task			3	6					

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	81	122	256	308	273	8
	Measurement and Data	84	311	479	413	275	8
	Number and Operations in Base Ten	378	551	693	356	275	8
	Number and Operations – Fractions	280	848	783	358	275	8
	Operational and Algebraic Thinking	857	1,087	929	455	275	8
	Performance Task	3	3	7	18	2	8
4	Geometry	33	121	94	121	78	10
	Measurement and Data	5	139	159	208	77	10
	Number and Operations in Base Ten	785	1,387	952	268	96	10
	Number and Operations – Fractions	410	623	933	250	96	10
	Operational and Algebraic Thinking	553	1,432	936	229	96	10
	Performance Task	276	2	4	4	37	10
5	Geometry	70	185	365	266	337	4
	Measurement and Data	34	171	273	282	336	4
	Number and Operations in Base Ten	794	1,009	810	325	337	4
	Number and Operations – Fractions	1,067	969	943	387	337	4
	Operations and Algebraic Thinking	85	672	600	310	337	4
	Performance Task		2	15	38	1	4
6	Expressions and Equations	352	903	351	322	278	3
	Geometry	42	181	202	469	281	3
	Number System	585	1,489	741	459	281	3
	Ratios and Proportional Relationships	909	1,344	789	511	281	3
	Statistics and Probability	13	46	121	275	269	3
	Performance Task	61	73	325	16	15	3
7	Expressions and Equations	612	913	493	334	58	
	Geometry	4	72	108	339	58	
	Number System	822	1,394	793	248	58	
	Ratios and Proportional Relationships	700	682	918	356	58	
	Statistics and Probability	12	22	117	144	58	
	Performance Task	6	25	385	3		
8	Expressions and Equations I	733	936	472	224	246	11
	Expressions and Equations II	151	589	481	177	246	11
	Functions	683	694	729	221	247	11
	Geometry	156	315	296	214	246	11
	Number System	225	320	434	82	247	11
	Performance Task	10	108	417	10	3	11

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	52	42	42	15
	Editing	1,098	38	44	18
	Language and Vocabulary Use	1,635	26	52	23
	Listening and Interpretation	899	16	60	24
	Reading Informational Text	1,315	23	53	24
	Reading Literary Text	1,435	24	43	32
	Research	553	21	51	28
	Revision	272	32	55	13
	Performance Task	124	12	65	23
4	Brief Writes	27	19	44	37
	Editing	816	15	49	35
	Language and Vocabulary Use	965	22	49	29
	Listening and Interpretation	615	12	58	30
	Reading Informational Text	1,885	12	61	27
	Reading Literary Text	1,492	24	50	26
	Research	576	21	48	31
	Revision	113	19	59	22
	Performance Task	180	14	61	26
5	Brief Writes	133	22	41	38
	Editing	939	24	46	31
	Language and Vocabulary Use	1,550	25	54	21
	Listening and Interpretation	794	28	51	21
	Reading Informational Text	1,824	8	58	34
	Reading Literary Text	1,292	17	55	28
	Research	430	29	47	24
	Revision	715	34	45	21
	Performance Task	122	28	52	20

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	142	7	32	61
	Editing	1,518	17	59	24
	Language and Vocabulary Use	1,210	19	43	37
	Listening and Interpretation	791	13	51	36
	Reading Informational Text	1,173	20	60	20
	Reading Literary Text	2,196	25	55	20
	Research	1,428	12	52	36
	Revision	1,222	24	55	21
	Performance Task	0			
7	Brief Writes	166	32	46	22
	Editing	1,164	18	69	13
	Language and Vocabulary Use	1,239	26	50	24
	Listening and Interpretation	458	23	57	20
	Reading Informational Text	960	27	48	25
	Reading Literary Text	2,275	28	53	19
	Research	272	21	66	14
	Revision	858	28	53	19
	Performance Task	6	83	0	17
8	Brief Writes	133	20	47	33
	Editing and Revising	860	34	52	14
	Listening and Interpretation	247	18	49	33
	Reading Informational Text	650	16	56	28
	Reading Literary Text	1,392	29	46	25
	Research	240	18	54	28
	Performance Task	9	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	1,048	25	47	29
	Measurement and Data	1,570	30	37	33
	Number and Operations in Base Ten	2,261	31	37	32
	Number and Operations – Fractions	2,552	18	44	38
	Operational and Algebraic Thinking	3,611	36	46	18
	Performance Task	41	12	56	32
4	Geometry	457	5	59	36
	Measurement and Data	598	15	54	30
	Number and Operations in Base Ten	3,498	32	47	21
	Number and Operations – Fractions	2,322	31	42	27
	Operational and Algebraic Thinking	3,256	39	46	15
	Performance Task	333	16	64	20
5	Geometry	1,227	25	54	21
	Measurement and Data	1,100	26	46	28
	Number and Operations in Base Ten	3,279	38	44	18
	Number and Operations – Fractions	3,707	38	44	18
	Operations and Algebraic Thinking	2,008	27	50	23
	Performance Task	60	33	50	17
6	Expressions and Equations	2,209	30	46	24
	Geometry	1,178	21	43	36
	Number System	3,558	39	43	18
	Ratios and Proportional Relationships	3,837	49	32	19
	Statistics and Probability	727	17	66	18
	Performance Task	493	26	61	12
7	Expressions and Equations	2,410	27	47	26
	Geometry	581	11	63	26
	Number System	3,315	33	51	17
	Ratios and Proportional Relationships	2,714	29	51	20
	Statistics and Probability	353	18	46	35
	Performance Task	419	38	48	14
8	Expressions and Equations I	2,622	38	45	17
	Expressions and Equations II	1,655	34	46	20
	Functions	2,585	41	39	20
	Geometry	1,238	29	44	27
	Number System	1,319	29	41	30
	Performance Task	559	33	51	16

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

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

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

Group	2014–2015				2015–2016				2016–2017				2017-2018			
	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
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
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
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
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
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
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
Pacific Islander	16	50	2426.7	107.	13	62	2453.8	70.7	13	62	2481.8	79.4	22	64	2446.2	77.4
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
Two or More Races	392	59	2440.7	75.8	440	57	2446.6	83.7	464	57	2444.1	85.3	482	55	2439.9	84.2
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
Disadvantaged	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
Migrant	332	44	2424.2	73.4	319	52	2430.4	75.8	331	47	2426.7	75.6	342	51	2430.7	76.4
Disability	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
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
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
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
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
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
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
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
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.
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
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
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
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
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
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

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

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

Group	2014–2015				2015–2016				2016–2017				2017-2018			
	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
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
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
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
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
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
Hispanic	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
Pacific Islander	10	80	2534.2	75.2	12	83	2556.1	53.9	12	42	2493.1	133.	11	82	2540.4	112.3
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
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
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
Special Education	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
CD 504	412	50	2502.1	82.6	424	53	2504.4	77.9	462	56	2510.7	80.5	493	55	2508.0	81.1
Title I	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
Grade 6																
All Students	10,02	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
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
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
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
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
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
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
Pacific Islander	8*				11	73	2533.8	121.	13	54	2529.1	105.	13	38	2475.9	134.2
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
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
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
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
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
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

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

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

Group	2014–2015				2015–2016				2016–2017				2017-2018			
	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
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
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
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
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
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
Hispanic	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
Pacific Islander	8*				10	30	2536.3	101.	13	54	2571.7	100.	11	73	2579.5	53.9
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
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
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
Special Education	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
CD 504	351	44	2535.6	85.4	453	45	2542.2	88.1	488	50	2549.7	86.8	506	53	2553.3	87.2
Title I	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
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
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
Male	4,877	43	2542.9	99.1	4,986	47	2552.1	98.5	5,127	45	2547.5	100.	5,151	46	2551.0	100.2
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
AmerIndian/Alaskan	38	66	2600.1	92.8	50	56	2579.1	100.	45	67	2585.3	88.6	47	51	2565.1	92.8
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
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
Pacific Islander	11	64	2597.3	97.3	9*				8*				14	50	2569.4	115.5
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
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
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
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
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
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

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

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

Group	2014–2015				2015–2016				2016–2017				2017-2018			
	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
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
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
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
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
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
Hispanic	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
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
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
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
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
Special Education	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
CD 504	333	48	2432.7	67.9	319	49	2438.6	72.1	336	50	2435.3	68.5	343	51	2438.8	71.5
Title I	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
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
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
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
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
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
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
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
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
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
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
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
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
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
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

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

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

Group	2014–2015				2015–2016				2016–2017				2017-2018			
	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
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
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
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
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
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
Hispanic	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
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
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
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
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
Special Education	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
CD 504	409	29	2493.9	77.2	423	35	2498.4	73.4	468	37	2509.1	80.7	496	39	2507.8	78.8
Title I	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
Grade 6																
All Students	10,084	34	2510.5	96.3	10,004	37	2516.3	101.	10,211	41	2523.8	103.5	10,446	40	2521.0	104.8
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
Male	5,103	33	2505.8	99.7	5,067	37	2513.3	105.	5,139	40	2520.4	108.0	5,210	38	2514.4	108.5
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
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
Asian	358	69	2598.7	94.6	361	70	2606.2	114.	389	76	2610.8	106.6	374	74	2615.9	106.2
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
Pacific Islander	8*				11	45	2535.3	156.	13	69	2551.2	86.9	14	29	2472.2	122.5
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
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
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
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
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
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

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

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

Group	2014–2015				2015–2016				2016–2017				2017-2018			
	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
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
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
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
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
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
Hispanic	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
Pacific Islander	8*				10	40	2530.5	96.2	15	47	2550.8	126.0	11	45	2555.0	79.6
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
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
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
Special Education	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
CD 504	350	33	2528.2	90.6	450	36	2532.9	91.3	488	38	2540.0	91.8	500	35	2530.9	95.8
Title I	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
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
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
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
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
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
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
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
Pacific Islander	11	36	2572.3	95.8	9*				9*				15	40	2543.7	134.4
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
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
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
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
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
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

***** 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	10,467	80	89	72	69	88	72	82	62	58	83
Female	5,160	79	88	72	69	88	71	81	62	59	83
Male	5,307	80	90	71	69	88	72	84	61	58	83
African American	3,174	79	90	72	69	85	71	84	62	58	77
AmerIndian/Alaskan	43	80	98	72	72*	80	73	88	65	58*	79
Asian	420	84	87	72	70	92	78	74	62	58	90
Hispanic	1,952	78	88	71	69	85	70	82	61	58	78
Pacific Islander	22	81	93*	74*	74*	84*	73	84*	65*	62*	78*
White	4,373	80	88	72	69	89	72	79	61	59	85
Multi-Racial	482	79	87	73	68	87	71	81	63	57	81
ELL	1,727	78	88	71	69	83	70	82	62	58	74
Special Education	1,447	84	92	71	68	85	78	89	62	54	75
CD 504	342	78	88	74	68	84	69	79	64	58	77
Title I	1,092	78	84	71	69	88	69	75	62	59	82
Grade 4											
All Students	10,658	78	89	64	67	88	70	82	52	57	82
Female	5,210	78	88	64	67	89	70	80	53	57	83
Male	5,448	78	90	64	67	87	70	84	52	57	81
African American	3,252	78	90	64	67	85	70	85	52	57	76
AmerIndian/Alaskan	37	79	87	68*	68*	86	70	81	55*	55*	85
Asian	384	83	88	65	68	91	76	81	49	57	88
Hispanic	2,000	77	88	65	67	85	68	83	53	57	76
Pacific Islander	13	82	97*	71*	70*	85*	75	93*	49*	55*	84*
White	4,496	79	87	64	68	89	71	78	52	57	84
Multi-Racial	476	78	87	64	68	89	69	80	53	57	84
ELL	1,608	76	88	64	67	82	68	83	53	57	71
Special Education	1,610	84	93	64	67	83	78	91	52	55	68
CD 504	417	76	90	64	68	83	68	82	53	58	76
Title I	1,054	76	87	64	67	87	68	77	53	56	81
Grade 5											
All Students	10,579	80	89	68	76	86	72	82	56	68	80
Female	5,275	80	88	68	76	87	72	80	57	67	81
Male	5,304	80	89	68	76	85	72	84	56	68	78
African American	3,216	80	90	68	76	83	72	85	57	68	72
AmerIndian/Alaskan	40	78	89*	73*	68	86	71	86*	60*	58	82
Asian	384	84	83	67	76	90	77	75	52	65	88
Hispanic	1,872	78	89	68	76	83	70	82	57	68	74
Pacific Islander	11	78	90*	-	76*	75*	71	91*	28*	70*	75*
White	4,575	80	87	68	76	87	72	79	55	68	82
Multi-Racial	481	78	84	68	76	85	70	74	58	67	79
ELL	886	80	91	67	75	80	73	85	58	64	65
Special Education	1,612	84	92	69	74	80	78	90	57	62	68
CD 504	493	79	89	70	77	84	70	78	58	70	75
Title I	1,066	78	86	69	76	86	70	75	58	69	78

*The classification index is based on 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	10,425	81	90	74	78	85	73	83	64	71	78
Female	5,222	81	89	74	78	86	73	81	64	71	79
Male	5,203	81	90	73	78	85	73	84	64	70	76
African American	3,087	81	91	74	79	81	74	85	65	71	69
AmerIndian/Alaskan	36	85	96	70*	85	86*	78	89	61*	78	78*
Asian	370	83	81	71	76	89	76	80	55	67	86
Hispanic	1,854	81	90	74	79	82	73	84	66	71	70
Pacific Islander	13	85	98*	57*	68*	82*	82	95*	42*	66*	77*
White	4,647	80	88	73	78	86	73	80	63	70	80
Multi-Racial	418	80	90	73	78	87	73	80	65	71	79
ELL	492	87	93	72	80	79*	82	91	63	61	63*
Special Education	1,574	87	93	73	77	82	81	90	64	65	61
CD 504	510	79	88	74	79	82	71	76	65	71	73
Title I	1,214	80	89	74	78	85	72	82	65	71	77
Grade 7											
All Students	10,219	80	90	72	79	84	73	84	61	72	75
Female	5,070	80	89	72	79	84	72	82	61	73	76
Male	5,149	81	90	72	79	83	73	85	62	72	73
African American	3,160	81	90	72	79	81	74	85	62	72	68
AmerIndian/Alaskan	43	78	80*	68	81	84	70	78*	57	70	80
Asian	381	83	88	70	79	89	77	84	55	73	84
Hispanic	1,770	80	90	72	79	78	72	84	63	72	66
Pacific Islander	11	75	-	72*	80*	65*	64	28*	58*	72*	57*
White	4,457	80	89	71	79	84	72	81	60	73	76
Multi-Racial	397	80	86	72	79	83	72	80	61	72	76
ELL	423	87	93	72	80	74*	81	90	63	63	49*
Special Education	1,510	86	93	71	77	75	81	91	62	65	54
CD 504	506	79	88	72	79	82	71	81	62	72	73
Title I	1,312	79	87	72	79	83	71	79	63	73	73
Grade 8											
All Students	10,106	81	89	74	80	84	74	83	64	74	75
Female	4,955	81	88	74	80	84	73	80	65	73	76
Male	5,151	82	90	74	80	83	74	85	64	74	73
African American	3,219	81	89	74	80	80	74	84	65	73	67
AmerIndian/Alaskan	47	83	94*	74	83	92*	76	78*	67	79	81*
Asian	368	84	92	73	80	89	78	86	59	73	85
Hispanic	1,641	81	90	74	80	81	74	85	64	73	67
Pacific Islander	14	80	82*	68*	78*	93*	73	76*	57*	74*	87*
White	4,520	81	88	74	80	84	73	80	63	74	76
Multi-Racial	297	81	89	75	81	85	74	80	67	72	79
ELL	374	87	92	74	75	67*	82	91	63	63	56*
Special Education	1,437	86	92	74	80	88	80	89	63	69	69
CD 504	534	80	87	73	80	83	72	80	64	73	73
Title I	1,527	80	87	74	80	83	73	82	63	74	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	10,517	83	90	74	79	89	76	84	65	72	85
Female	5,184	82	89	74	80	88	75	82	65	72	84
Male	5,333	83	90	74	79	90	77	85	64	71	86
African American	3,181	83	91	74	79	85	76	86	65	72	78
AmerIndian/Alaskan	43	82	88	69	85	89*	74	81	61	76	78*
Asian	427	88	93	75	77	94	83	86	62	68	92
Hispanic	1,982	81	89	73	79	87	74	83	64	71	79
Pacific Islander	22	85	97*	67*	85*	88*	80	89*	56*	80*	84*
White	4,378	83	88	75	80	90	77	80	65	72	86
Multi-Racial	483	83	89	76	80	88	75	82	66	71	83
ELL	1,790	82	89	74	80	87	75	84	64	72	80
Special Education	1,441	87	94	73	79	87	82	92	64	69	75
CD 504	343	82	87	75	81	89	75	79	66	73	85
Title I	1,096	83	86	75	80	91	76	80	65	73	85
Grade 4											
All Students	10,689	84	89	80	79	90	77	82	73	71	85
Female	5,227	83	88	80	79	90	76	80	73	71	84
Male	5,462	84	90	80	78	90	77	84	73	70	85
African American	3,246	83	90	80	77	87	76	84	74	68	79
AmerIndian/Alaskan	37	86	93*	81	78*	96	80	82*	78	72*	88
Asian	396	87	89	82	78	93	82	80	71	70	91
Hispanic	2,023	83	89	80	79	87	76	81	73	71	81
Pacific Islander	13	88	92*	86*	88*	88*	81	92*	75*	79*	85*
White	4,499	84	88	80	80	90	77	79	73	72	86
Multi-Racial	475	83	85	80	79	90	76	79	71	71	85
ELL	1,663	83	89	80	79	86	76	82	73	71	80
Special Education	1,626	86	92	79	76	86	80	89	71	66	78
CD 504	420	83	87	82	78	87	75	79	75	71	80
Title I	1,061	83	86	81	80	90	76	77	73	72	85
Grade 5											
All Students	10,633	83	89	77	72	90	75	84	69	61	85
Female	5,304	82	88	78	72	90	75	82	69	61	85
Male	5,329	83	90	77	72	90	76	85	68	62	86
African American	3,219	83	90	77	71	88	76	87	68	60	79
AmerIndian/Alaskan	40	81	83	78	68*	86	75	77	71	47*	87
Asian	397	87	89	77	72	93	81	76	69	61	92
Hispanic	1,909	82	89	77	71	88	74	83	69	62	80
Pacific Islander	11	82	89*	82*	63*	93*	75	80*	70*	57*	85*
White	4,574	82	88	78	72	91	75	80	69	62	86
Multi-Racial	483	81	88	76	72	91	74	80	70	59	86
ELL	952	83	90	77	71	85	77	86	68	59	76
Special Education	1,619	88	93	77	73	83	83	92	65	60	70
CD 504	496	81	88	78	72	88	73	82	71	60	81
Title I	1,070	82	87	79	72	91	75	79	71	62	86

*The classification index is based on 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	10,446	83	92	78	72	90	76	86	70	62	83
Female	5,236	82	91	78	72	89	75	85	70	63	83
Male	5,210	83	92	78	72	90	77	87	70	62	84
African American	3,071	84	93	78	72	85	77	89	70	62	75
AmerIndian/Alaskan	35	81	95	74	65*	93*	74	87	66	57*	84*
Asian	374	87	93	77	72	95	81	83	68	62	93
Hispanic	1,888	83	91	78	73	86	76	86	71	62	77
Pacific Islander	14	84	97*	71*	61*	74*	79	93*	64*	47*	79*
White	4,646	82	90	78	72	90	75	82	70	63	85
Multi-Racial	418	83	92	79	72	89	76	86	71	62	82
ELL	543	90	95	78	72	85*	86	93	69	57	67
Special Education	1,557	91	95	77	69	83	87	94	68	53	76
CD 504	510	82	90	79	72	90	75	84	71	61	83
Title I	1,212	81	91	78	72	88	74	84	70	62	81
Grade 7											
All Students	10,231	83	91	76	75	89	76	86	67	65	83
Female	5,071	82	91	76	75	88	75	85	68	65	83
Male	5,160	83	91	76	75	90	76	86	66	66	84
African American	3,151	84	92	75	74	88	77	87	68	63	77
AmerIndian/Alaskan	44	82	89	77	70*	85	75	85	71	57*	81
Asian	388	87	92	76	76	94	81	87	62	69	91
Hispanic	1,809	83	91	75	75	86	76	86	67	66	78
Pacific Islander	11	84	93*	78*	80*	100*	75	81*	71*	74*	77*
White	4,436	82	89	76	75	89	74	82	67	66	84
Multi-Racial	392	82	90	76	74	89	75	86	68	64	84
ELL	477	90	95	74	75	95*	86	94	63	64	79*
Special Education	1,511	91	95	75	70	90	87	94	64	56	76
CD 504	500	81	87	76	77	89	74	81	68	66	82
Title I	1,314	81	89	77	73	87	73	83	68	65	81
Grade 8											
All Students	10,117	82	90	72	71	90	75	85	62	61	84
Female	4,951	81	89	72	71	89	74	84	62	62	83
Male	5,166	83	91	71	71	91	76	87	61	61	86
African American	3,210	83	91	72	70	87	77	88	61	60	78
AmerIndian/Alaskan	48	80	93	72	66*	81	73	87	63	53*	79
Asian	373	87	90	71	71	95	81	83	60	61	93
Hispanic	1,674	82	90	71	72	88	75	87	61	61	80
Pacific Islander	15	88	99*	78*	77*	89*	83	97*	70*	65*	84*
White	4,506	81	88	72	72	90	73	81	62	62	85
Multi-Racial	291	80	89	70	71	87	73	84	61	59	84
ELL	427	89	93	68	73	94	84	92	55	58	84
Special Education	1,422	90	95	71	71	82	86	94	58	51	77
CD 504	537	81	89	72	70	88	73	84	62	58	82
Title I	1,524	81	90	71	72	89	74	84	62	62	83

*The classification index is based on n < 10.