

South Dakota Assessments 2023–2024 Technical Report



south dakota
DEPARTMENT OF EDUCATION
Learning. Leadership. Service.

**Submitted to
South Dakota Department of Education
by Cambium Assessment, Inc.**

TABLE OF CONTENTS

1. OVERVIEW.....	1
1.1. South Dakota Assessments.....	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 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. Online Training.....	8
2.3.2. District Trainings.....	11
2.4. Test Security.....	11
2.4.1. Student-Level Testing Confidentiality.....	11
2.4.2. System Security.....	12
2.4.3. Security of the Testing Environment.....	13
2.4.4. Test Security Violations.....	14
2.5. Student Participation.....	15
2.5.1. Exempt Students.....	15
2.6. Online Testing Features and Testing Accommodations.....	15
2.6.1. Online Universal Tools for All Students.....	16
2.6.2. Designated Supports and Accommodations.....	18
2.7. Testing Time.....	29
2.8. Data Forensics Program.....	31
2.8.1. Changes in Student Performance.....	32
2.8.2. Test-Taking Time.....	32
2.8.3. Inconsistent Item Response Pattern (Person Fit).....	33
2.8.4. Item Response Change.....	33
2.9. Prevention and Recovery of Disruptions in Test Delivery System.....	34

2.9.1. <i>High-Level System Architecture</i>	34
2.9.2. <i>Automated Backup and Recovery</i>	36
2.9.3. <i>Other Disruption Prevention and Recovery</i>	36
3. SUMMARY OF 2023–2024 OPERATIONAL TEST ADMINISTRATION	37
3.1. Student Population.....	37
3.2. Summary of Overall Student Performance.....	38
3.3. Distribution of Student Ability and Item Difficulty	50
4. VALIDITY	57
4.1. Evidence on Test Content.....	57
4.2. Evidence on Internal Structure	63
5. RELIABILITY	66
5.1. Marginal Reliability.....	66
5.2. Standard Error Curves	67
5.3. Reliability of Achievement Classification.....	70
5.4. Reliability for Subgroups	75
5.5. Reliability for Claim Scores	78
6. SCORING	81
6.1. Estimating Student Ability Using Maximum Likelihood Estimation	81
6.2. Rules for Transforming Theta to Vertical Scale Scores	82
6.3. Lowest/Highest Obtainable Scores (LOSS/HOSS)	83
6.4. Scoring All Correct and All Incorrect Cases	84
6.5. Rules for Calculating Strengths and Weaknesses for Claim Scores.....	84
6.6. Target Scores	84
6.6.1. <i>Target Scores Relative to Student’s Overall Estimated Ability</i>	84
6.6.2. <i>Target Scores Relative to Proficiency Standard (Level 3 Cut)</i>	86
6.7. Handscoring.....	87
6.7.1. <i>Rater Selection</i>	87
6.7.2. <i>Rater Training, Qualification, and Scoring</i>	88
6.7.3. <i>Rater Monitoring, Feedback, and Evaluation</i>	91
6.7.4. <i>Rater Agreement</i>	93

6.8. Automated Scoring	95
6.8.1. Project Essay Grade	95
6.8.2. Model Training and Validation.....	96
6.8.3. Automated Scoring Processes	101
6.8.4. PEG-Human Agreement.....	104
6.8.5. Recommendations	108
7. REPORTING AND INTERPRETING SCORES	109
7.1. Reporting System	109
7.1.1. Dashboard.....	111
7.1.2. Aggregate Score Reports: Overall Performance	112
7.1.3. Aggregate Score Reports: Claim and Target Performance	114
7.1.4. Roster Performance Report.....	114
7.1.5. Trend Report	115
7.1.6. Individual Student Report	116
7.2. Interpretation of Reported Scores.....	119
7.2.1. Scale Score.....	119
7.2.2. Conditional Standard Error of Measurement	120
7.2.3. Achievement Level.....	120
7.2.4. Performance Category for Claims	120
7.2.5. Performance Category for Targets	120
7.2.6. Aggregated Scale Score	121
7.3. Appropriate Uses of Test Result.....	121
8. QUALITY CONTROL PROCEDURE.....	123
8.1. Adaptive Test Configuration	123
8.1.1. Platform Review	123
8.1.2. User Acceptance Testing and Final Review.....	124
8.2. Quality Assurance in Document Processing.....	124
8.3. Quality Assurance in Data Preparation	124
8.4. Quality Assurance in Online Test Delivery System	124
8.4.1. Score Report Quality Check.....	125

REFERENCES 128

LIST OF TABLES

Table 1. 2023–2024 Testing Windows	3
Table 2. Summary of Tests and Testing Options in 2023–2024.....	3
Table 3. Number of Students Who Took Paper-Pencil Tests in the 2023–2024 Summative Test Administration.....	7
Table 4. 2023–2024 Universal Tools, Designated Supports, and Accommodations	23
Table 5. ELA/L Total Students with Allowed Embedded and Non-Embedded Accommodations	24
Table 6. ELA/L Total Students with Allowed Embedded Designated Supports	25
Table 7. ELA/L Total Students with Allowed Non-Embedded Designated Supports.....	26
Table 8. Mathematics Total Students with Allowed Embedded and Non-Embedded Accommodations...	27
Table 9. Mathematics Total Students with Allowed Embedded Designated Supports.....	27
Table 10. Mathematics Total Students with Allowed Non-Embedded Designated Supports.....	28
Table 11. ELA/L Testing Time.....	30
Table 12. Mathematics Testing Time	31
Table 13. Participation Rates by Percentage in ELA/L Summative Assessment	37
Table 14. Participation Rates by Percentage in Mathematics Summative Assessment	37
Table 15. Number of Students in ELA/L Summative Assessment.....	38
Table 16. Number of Students in Mathematics Summative Assessment.....	38
Table 17. Descriptive Statistics and Percentage of Students in Achievement Levels for Overall and by Subgroup: ELA/L (Grades 3–5)	39
Table 18. Descriptive Statistics and Percentage of Students in Achievement Levels for Overall and by Subgroup: ELA/L (Grades 6–8)	40
Table 19. Descriptive Statistics and Percentage of Students in Achievement Levels for Overall and by Subgroup: ELA/L (Grade 11).....	41
Table 20. Descriptive Statistics and Percentage of Students in Achievement Levels for Overall and by Subgroup: Mathematics (Grades 3–5).....	42
Table 21. Descriptive Statistics and Percentage of Students in Achievement Levels for Overall and by Subgroup: Mathematics (Grades 6–8).....	43
Table 22. Descriptive Statistics and Percentage of Students in Achievement Levels for Overall and by Subgroup: Mathematics (Grade 11)	44
Table 23. ELA/L Percentage of Students in Performance Categories by Claim	49
Table 24. Mathematics Percentage of Students in Performance Categories by Claim	50

Table 25. Percentage of ELA/L CAT Delivered Tests Meeting Blueprint Requirements for Each Claim and the Number of Passages Administered (Grades 3–5)	58
Table 26. Percentage of ELA/L CAT Delivered Tests Meeting Blueprint Requirements for Each Claim and the Number of Passages Administered (Grades 6–8, 11)	59
Table 27. Mathematics Percentage of CAT Delivered Tests Meeting Blueprint Requirements for Claims and Targets (Grades 3–5)	60
Table 28. Mathematics Percentage of CAT Delivered Tests Meeting Blueprint Requirements for Claims and Targets (Grades 6–8)	61
Table 29. Mathematics Percentage of CAT Delivered Tests Meeting Blueprint Requirements for Claims and Targets (Grade 11)	62
Table 30. Average and Range of the Number of Unique Targets Assessed within Each Claim Across All Delivered CAT Tests	63
Table 31. Correlations Among Claim Scores for ELA/L	64
Table 32. Correlations among Claim Scores for Mathematics	65
Table 33. Marginal Reliability for ELA/L and Mathematics	67
Table 34. Average Conditional Standard Error of Measurement by Achievement Levels	70
Table 35. Average Conditional Standard Error of Measurement at Each Achievement Level Cut and Difference of the Standard Errors of Measurement Between Two Cuts	70
Table 36. Classification Accuracy and Consistency	74
Table 37. Marginal Reliability Coefficients for Overall and by Subgroup: ELA/L (Grades 3–4)	75
Table 38. Marginal Reliability Coefficients for Overall and by Subgroup: ELA/L (Grades 5–6)	75
Table 39. Marginal Reliability Coefficients for Overall and by Subgroup: ELA/L (Grades 7–8)	76
Table 40. Marginal Reliability Coefficients for Overall and by Subgroup: ELA/L (Grade 11)	76
Table 41. Marginal Reliability Coefficients for Overall and by Subgroup: Mathematics (Grades 3–4)	77
Table 42. Marginal Reliability Coefficients for Overall and by Subgroup: Mathematics (Grades 5–6)	77
Table 43. Marginal Reliability Coefficients for Overall and by Subgroup: Mathematics (Grades 7–8)	78
Table 44. Marginal Reliability Coefficients for Overall and by Subgroup: Mathematics (Grade 11)	78
Table 45. Marginal Reliability Coefficients for Claim Scores in ELA/L	79
Table 46. Marginal Reliability Coefficients for Claim Scores in Mathematics	80
Table 47. Vertical Scaling Constants on the Reporting Metric	82
Table 48. Cut Scores in Scale Scores	83
Table 49. Lowest and Highest Obtainable Scores	83

Table 50. Number of Handscored Items in 2023–2024 Smarter Balanced Summative Item Pool, by Grade and Subject	87
Table 51. Inter-Rater Agreement for ELA/L Short-Answer Items	93
Table 52. Inter-Rater Agreement for ELA/L Essay Items	94
Table 53. Inter-Rater Agreement for Mathematics Items	95
Table 54. Number of Items Eligible for Automated Scoring, by Grade and Subject Area.....	96
Table 55. Initial Model Evaluation Criteria	98
Table 56. Demographic Variables and Categories.....	99
Table 57. Secondary Validation Criteria.....	99
Table 58. Summary of Secondary Validation Results, by Grade and Subject Area	100
Table 59. Summary of Live Training and Validation Results, by Grade and Subject Area	100
Table 60. Flags Currently Established	102
Table 61. Model Setting.....	103
Table 62. Human-Machine Agreement for ELA/L Short-Answer Items on Initial and Secondary Validation Samples, by Grade	104
Table 63. Human-Machine Agreement for ELA/L Essay Items on Initial and Secondary Validation Samples, by Grade.....	105
Table 64. Human-Machine Agreement for Mathematics Items on Initial and Secondary Validation Samples, by Grade.....	106
Table 65. Human-Machine Agreement for ELA/L Short-Answer Items on Live Validation Sample, by Grade	106
Table 66. Human-Machine Agreement for ELA/L Essay Items on Live Validation Sample, by Grade..	107
Table 67. Human-Machine Agreement for Mathematics Items on Live Validation Samples, by Grade .	107
Table 68. Types of Online Score Reports by Level of Aggregation.....	110
Table 69. Types of Subgroups	110
Table 70. Overview of Quality Assurance Reports	125

LIST OF FIGURES

Figure 1. ELA/L Percent Proficient Across Years.....	45
Figure 2. Mathematics Percent Proficient Across Years	46
Figure 3. ELA/L Average Scale Score Across Years	47
Figure 4. Mathematics Average Scale Score Across Years.....	48
Figure 5. Student Ability—Item Difficulty Distribution for ELA/L	51
Figure 6. Student Ability—Item Difficulty Distribution by Claim: ELA/L (Grades 3–5)	52
Figure 7. Student Ability—Item Difficulty Distribution by Claim: ELA/L (Grades 6–8, 11)	53
Figure 8. Student Ability—Item Difficulty Distribution for Mathematics	54
Figure 9. Student Ability—Item Difficulty Distribution by Claim: Mathematics (Grades 3–5).....	55
Figure 10. Student Ability—Item Difficulty Distribution by Claim: Mathematics (Grades 6–8, 11).....	56
Figure 11. Conditional Standard Error of Measurement for ELA/L.....	68
Figure 12. Conditional Standard Error of Measurement for Mathematics	69
Figure 13. PEG Architecture.....	96
Figure 14. Response Routing Rules.....	102

LIST OF EXHIBITS

Exhibit 1. Dashboard: State Level	111
Exhibit 2. Dashboard: District Level	112
Exhibit 3. Detailed Dashboard: District Level.....	112
Exhibit 4. Overall Performance Summary Results for Grade 8 ELA/L: District Level	113
Exhibit 5. Overall Performance Results for Grade 8 ELA/L by Gender: District Level	113
Exhibit 6. Claim- and Target-Level Results for Grade 5 Mathematics: District Level	114
Exhibit 7. Roster Performance Report for Grade 5 Mathematics	115
Exhibit 8. Trend Report for ELA/L: Student Level.....	115
Exhibit 9. Individual Student Report for ELA/L	117

1. OVERVIEW

This report provides a technical summary of the 2023–2024 administration of the South Dakota English-language arts/literacy (ELA/L) Assessment and the South Dakota Mathematics Assessment in grades 3–8 and 11. This report includes eight chapters, including: Overview, Testing Administration, Summary of 2023–2024 Operational Test Administration, Validity, Reliability, Scoring, Reporting and Interpreting Scores, and Quality Control Procedure. For the interim assessments, the number of students who took Interim Comprehensive Assessments (ICAs) and Interim Assessment Blocks (IABs) and their performance are provided in Appendix A, Summary of the 2023–2024 Interim Assessments. The data included in this report are based on South Dakota data for the summative assessment in ELA/L and mathematics.

While this report includes information on all aspects of the technical quality of the test administration in South Dakota, the information on item and test development, item content review, field-test administration, item data review, item calibrations, content-alignment study, standard setting, and other validity information can be found in the overall Smarter Balanced technical report. The Smarter Balanced technical report includes all aspects of the technical qualities of the assessments described in the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014) and the requirements of the U.S. Department of Education *Peer Review of State Assessment Systems Non-Regulatory Guidance for States* (U.S. Department of Education, 2015).

1.1. SOUTH DAKOTA ASSESSMENTS

The Smarter Balanced Assessment Consortium (SBAC) has developed a next-generation assessment system in English-language arts/literacy (ELA/L) and mathematics in grades 3–8 and 11. At the time of development, South Dakota was one of 18 member states (plus the U.S. Virgin Islands) leading the development of assessments in ELA/L and mathematics.

The South Dakota English-language arts and mathematics content standards define the knowledge and skills students need to succeed in college and careers after graduating high school. The standards are tailored specifically to meet the needs of students in South Dakota, ensuring a readiness for the workforce, military service, university or technical school coursework. The standards are (1) research and evidence based, (2) aligned with college and work expectations, (3) rigorous, and (4) benchmarked in their design and content.

The South Dakota assessments are designed to measure students' comprehension and proficiency in South Dakota ELA/L standards and South Dakota mathematics standards for grades 3–8 and 11 and to provide valid, reliable, and fair test scores about student academic achievement. The assessment system includes both summative and interim assessments. The summative assessment is administered every spring to all students in tested grades, while the interim assessments are optional for districts to use. The assessments use computer adaptive testing technologies to provide meaningful feedback and actionable data that teachers can use to help students succeed.

The South Dakota assessments consist of the end-of-year summative assessment designed for accountability purposes and the optional interim assessments designed to support teaching and learning throughout the year. The summative assessments determine student achievement and track student progress toward college and career readiness in ELA/L and mathematics. The summative assessments consist of a computer-adaptive test (CAT) and a performance task (PT).

- **Computer-Adaptive Test (CAT).** The CAT is an online adaptive test that provides an individualized assessment for each student.
- **Performance Task (PT).** A PT is a task that challenges students to apply their knowledge and skills to respond to real-world problems. PTs can best be described as collections of questions and activities coherently connected to a single theme or scenario. They are used to better measure capacities such as depth of understanding, research skills, and complex analysis, which cannot be adequately assessed with selected- or constructed-response items. The computer can score some performance task items, but most are handscored.

The optional interim assessments allow teachers to monitor student progress throughout the year and provide information that they can use to improve instruction and learning. These tools are used at the discretion of schools and districts, and teachers can employ them to evaluate students' mastery of specific concepts at strategic points during the school year. There are three types of interim assessments available as fixed-form tests:

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

In spring 2015, the new statewide assessments in ELA/L and mathematics were administered for the first time to students in grades 3–8 and 11 in all South Dakota public elementary and secondary schools. American Institutes for Research delivered the assessment until the 2019–20 school year. Starting with the 2020–21 school year, Cambium Assessment, Inc. (CAI) delivered and scored the South Dakota assessments and produced score reports. Measurement Incorporated (MI) scored the handscored items.

2. TEST ADMINISTRATION

2.1. TESTING WINDOWS

The 2023–2024 South Dakota English language arts/literacy (ELA/L) and mathematics assessments testing window provided by the South Dakota Department of Education (SDDOE) spanned approximately two months for the online summative assessments and approximately seven to nine months prior to summative assessments for the interim assessments. The paper-pencil fixed-form summative assessments were administered over a one-month period during the online summative window. Table 1 shows the testing windows for both the online and paper-pencil summative and interim assessments.

Table 1. 2023–2024 Testing Windows

Tests	Grade	Start Date	End Date	Mode
Interim Comprehensive Assessments	3–8, 11	8/18/2023	3/1/2024	Online Fixed-Form
Interim Assessment Blocks	3–8, 11	8/18/2023	5/3/2024	Online Fixed-Form
Summative Assessments	3–8, 11	3/25/2024	5/3/2024	Online Computer Adaptive
	3–8, 11	4/1/2024	4/19/2024	Paper Fixed-Form

2.2. TEST OPTIONS AND ADMINISTRATIVE ROLES

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

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

Assessments	Test Options	Test Mode
Summative Assessments	English	Online
	Braille	Online
	Spanish (mathematics only)	Online
	Fixed-Form (standard)	Paper
	Fixed-Form (braille)	Paper
Interim Assessments	Fixed-Form (large print)	Paper
	English	Online
	Spanish (mathematics only)	Online
	Braille	Online

To ensure standardized administration conditions, teachers (TEs) and proctors (PRs) follow procedures outlined in the *Online, Summative, Test Administration Manual* (TAM). TEs and PRs must review the TAM before testing begins to ensure that the testing room is prepared appropriately (e.g., removing certain classroom posters, arranging desks). Make-up procedures are established for any students who are absent on testing day(s). Relying on the TAM for guidance, TEs and PRs read aloud the boxed directions verbatim to students, ensuring uniform administration procedures and testing conditions.

2.2.1. Administrative Roles

The key personnel involved with test administration are Assessment Coordinators (AC), District Administrators (DAs), School Coordinators (SCs), Proctors (PRs), Teachers (TEs), and Paraprofessionals (PARAs). The main responsibilities of these key personnel are described in this section. Detailed descriptions can be found in the TAM provided online at <https://sd.portal.cambiumast.com/resources/>.

Assessment Coordinator

The AC is authorized to add users to the Test Information Distribution Engine (TIDE) and to assign them any role except that of a DA. If assigned, an AC can modify student records within their district in TIDE (including accommodations, designated supports, and interim test eligibility) or submit appeals. Their primary responsibility is coordinating the administration of the South Dakota assessments in the district.

ACs are responsible for the following:

- Enter student test settings in TIDE.
- Identify students who may require designated supports and test accommodations and ensure that procedures for testing these students follow state policies.
- Monitor testing progress during the testing window and ensure that all eligible students participate in the testing process.

District Administrator

The DA's role is assigned by the South Dakota Department of Education (SDDOE) to district-level personnel who need access to the system, mainly to access district-level data, but isn't the Assessment Coordinator.

School Coordinator

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

SCs are also responsible for the following:

- Establish a testing schedule with ACs, TEs, and PRs based on test administration windows.
- Work with technology staff to ensure timely computer setup and installations.
- Work with TEs and PRs to review student information in TIDE to ensure that student information and test settings for designated supports and accommodations are correctly applied.
- Create, update, or import TE and PR accounts into TIDE.
- Enter student test settings in TIDE.
- Identify students who may require designated supports and test accommodations and ensure that procedures for testing these students follow state policies.
- Attend all district trainings and read all South Dakota policies and test administration documents.

- Ensure that all TEs and PRs attend state, district, or school trainings and review online training modules posted on the Gateway.
- Establish secure and separate testing rooms when needed.
- Monitor the secure administration of assessments.
- Monitor testing progress during the testing window and ensure that all eligible students participate in the testing process.
- Investigate and report all testing improprieties, irregularities, and breaches reported by TEs and PRs in the school.
- Attend to secure material before, during, and after the testing window, in accordance with state policies.

Teacher and Proctor

TEs and PRs are responsible for administering the South Dakota assessments. They must be certified staff.

TEs can view student results when they are made available. This role may be assigned to teachers who do not administer an assessment but need access to student results.

The PR's role does not allow access to student results. The role is designed for PRs, such as technology staff, who administer tests but should not have access to student results.

TEs/PRs have responsibilities that include the following:

- Complete South Dakota assessment administration training.
- Read all state policies and test administration documents before administering any South Dakota assessment.
- View student information before testing to ensure that a student receives the proper assessment with the appropriate supports and accommodations. TEs and PRs also report any potential demographic data or test support errors to SCs and ACs, as appropriate.
- Administer the South Dakota ELA/L and mathematics assessments.
- Report all potential assessment security incidents to the SCs and ACs in a manner consistent with state and district policies.

Paraprofessional

A PARA is a district-managed, non-testing user who may assist a TE or PR to administer the South Dakota assessments, but a PARA cannot administer the South Dakota assessments themselves. Prior to assisting the administration of a South Dakota assessment, PARAs must sign a Non-Disclosure Agreement. PARAs are mainly responsible for assisting TE/PRs in the administration of South Dakota assessments. The PARA role does not allow access to student results.

2.2.2. Online Administration

Within the state's testing window, schools can set testing schedules in intervals (e.g., multiple sessions) rather than in one long period, minimizing the interruption of classroom instruction and efficiently using their facilities. With online testing, schools do not need to handle test booklets and address the storage and security problems inherent in large shipments of materials to a school site.

SCs oversee all aspects of testing at their schools and serve as the main point of contact; TEs and PRs administer the online assessments only. TEs and PRs are trained in the online administration requirements and the mechanics of starting, pausing, and ending a testing session. Training materials for the assessment administration are available online. All school personnel who act as TEs and PRs are encouraged to complete the online PR Certification Course before testing begins.

To start a test session, the PR or TE must access the PR Interface of the online test delivery system (TDS) using his or her own computer. A session ID is generated when the test session is created. Students who are taking the assessment with the TE or PR need to enter their Statewide Student Identification (SSID) number, first name, and session ID into the Student Interface using computers provided by the school. The TE or PR then verifies that the students are taking the appropriate assessments with the appropriate accessibility features. (See Section 2.6, Online Testing Features and Testing Accommodations, for a list of accommodations.) Students can begin testing only when the TE or PR confirms the test tool settings. The TE or PR reads Section 10, Day of Test Administration, in the *Online, Summative, Test Administration Manual* to the students and walks them through the login process.

Once a student begins an assessment, he or she must answer all test questions on the current page before proceeding to the next page. Skipping questions is not permitted. For the online computer-adaptive test (CAT), students can review and edit previously answered items as long as these items are in the same test session and segment and the session has not been paused for more than 20 minutes. During an active CAT session, if a student changes his or her response to a previously answered item, the responses given earlier to all subsequent items remain the same. No new items are assigned to this student for changing his or her answers. The following provides an example of how this works.

A student pauses for 10 minutes after completing Item 10. The student then goes back to Item 5 and changes his or her answer. If the response change in Item 5 changes the item score from incorrect to correct, the student's overall score will improve; however, the answers to Items 6–10 will not change. For performance tasks (PTs), there is no pause rule, but the same rules that apply to the CAT for reviewing and changing assessment responses also apply to PTs.

When proctoring the summative assessment, the assessment can be started in one test or component (e.g., CAT) and completed in a different component. It is recommended that students take the CAT and PT items on separate days to minimize student fatigue. For each content area, it is also recommended that students begin with the CAT items followed by the PT. The CAT assessment must be completed within 45 calendar

days of the start date; after 45 days, the assessment opportunity will expire. For a PT, the assessment must be completed within 10 calendar days of the start date; after 10 days, the assessment will expire and may not be reopened.

During a test session for one or more students, TEs and PRs may pause the test for a break. It is up to the TEs or PRs to determine an appropriate stopping point; however, for the English language arts/literacy (ELA/L) and mathematics CAT, the assessment cannot be paused for more than 20 minutes to ensure the integrity of the test scores and security of the assessment items. When an assessment has been paused for more than 20 minutes, students must start a new test session and pick up where they left off. Previous responses will no longer be available for editing.

The TE or PR will always remain in the room during a test session to monitor student testing. Once the test session ends, the TE or PR will ensure that each student has successfully logged out of the system and will collect any handouts or scratch paper that students used during the assessment. Notes and handouts will be securely shredded immediately following the testing sessions.

2.2.3. Paper-Pencil Test Administration

The paper-pencil versions of the South Dakota ELA/L and mathematics assessments are provided as an alternative test administration method for students who cannot access a computer or for students with blindness or visual impairments. In South Dakota paper-pencil tests are offered in the standard, non-accommodated format, large print, and braille formats.

In any district where students are eligible to take the paper-pencil version of the test, the AC must submit a request on their behalf to the SDDOE Office of Assessment for appropriate testing materials. If the request is approved, the testing contractor ships the appropriate test booklets to the district. For ELA/L, the field (i.e., schools, districts) also receives a listening script that contains secure information needed to administer the listening session.

Separate test booklets are used for the ELA/L and mathematics assessments. Items from the CAT and the PT components are combined into one, fixed-form test booklet, including two sessions for CAT and one session for PT in both content areas. The TE or PR can break up the assessment components into separate sessions as needed.

After the student has completed the assessments, the AC returns the test booklets, answer booklets, and listening script to the testing vendor. The testing vendor scans the answer document and scores the test, including the handscored items.

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

Subject	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 11	Total
ELA/L	17	16	14	17	21	16		101
Mathematics	19	18	14	18	21	17		107

2.2.4. Braille Test Administration

An adaptive braille format of the South Dakota ELA/L and mathematics assessments is available in English. In the 2017–2018 assessment 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 SDDOE. All items on the Braille HAT are presented to the students using a Refreshable Braille Display (RBD).

The braille interface assessment is described in the following paragraphs:

- 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 Braille code via a braille embosser through the adaptive online summative assessment and a fixed-form PT.
- Students taking the summative ELA/L assessment can emboss both reading passages and items as they progress through the assessment. If a student has an RBD, a 40-cell RBD is recommended. The summative ELA/L is presented to the student with items in either contracted or non-contracted Literary Braille (for items containing only text) and via a braille embosser (for items with tactile or spatial components that an RBD cannot read).

Before administering the online summative assessments using the braille interface, TEs and PRs ensure that the technical requirements are met. The student's computer, the TE/PR's computer, and any supporting braille technologies used in conjunction with the braille interface are verified.

2.3. TRAINING AND INFORMATION FOR TEST COORDINATORS AND ADMINISTRATORS

ACs, DAs, and SCs oversee all aspects of testing at their schools and serve as the main points of contact, while TEs and PRs administer the online assessments. The online PR Certification Course, PowerPoint presentations, user guides, manuals, and regional trainings are used to train ACs and SCs in the online testing requirements and the mechanics of starting, pausing, and ending a test session. Training materials for the administration are online at <https://sd.portal.cambiumast.com/resources>. District and School Test Coordinators are responsible for training TEs and PRs.

2.3.1. Online Training

Multiple online training opportunities are offered to key staff through the Internet.

Proctor Certification Course

All school personnel who serve as TEs and PRs are strongly encouraged to complete an online Proctor (PR) Certification Course before administering the secure and valid assessments. This web-based course is 30–45 minutes long and covers information on testing policies and the steps for administering a test session in the online system. The course is interactive, requiring participants to start test sessions under different scenarios. Throughout the training, and at the end of the course, participants answer multiple-choice questions about the information provided. A certification of completion is provided to TEs and PRs upon successful completion of the course. The certification is tracked in TIDE and should be kept on file at the associated school.

System Tutorials

The following presentations are offered to explain how the assessment system works. Each of these presentations lasts approximately 60 minutes. The slides are available on the portal at <https://sd.portal.cambiumast.com/resources>.

Assessment Viewing Application (AVA). This application allows district- and school-level users to view the interim assessments (ICAs, IABs, and FIABs) for administrative or instructional purposes. The tutorial provides an overview of the AVA for the South Dakota interim assessments.

Reporting System. The reporting system enables district- and school-level users to handscore interim assessments or view their associated reports. The webinar provides an overview of the reporting system for handscoring and reporting the South Dakota interim assessments. In addition, slide notes and an additional presentation are provided as resources.

Student Interface Overview. This tutorial provides an overview of the online student interface in the test delivery system (TDS).

Technology Requirements for Online Testing. This tutorial provides an overview of the technology requirements needed on all computers and devices used for online testing, information on secure browser installation, and voice packs for text-to-speech.

Test Delivery System (TDS) Training. This tutorial prepares ACs, SCs, TEs, and PRs for the assessments by providing an overview of the PR Interface and TDS, including how to start and monitor a test session using the PR Interface.

Test Information Distribution Engine (TIDE). This tutorial provides an overview of how to navigate the TIDE system, including how to register users, enroll students, manage, and edit users/students, and process/view test invalidations.

Testing with Braille. This tutorial provides an overview of the information needed to administer an online braille test in TDS. This also includes information about the specific hardware and software requirements needed to support online braille testing.

Test Design Modules

The following training modules are designed to explain the overall test design of the South Dakota assessments:

Accessibility and Accommodations. This module covers accessibility options, including designated supports and accommodations for students taking the South Dakota assessments. It focuses on students with disabilities, students with a Section 504 Plan, and students identified as English language learners. It also provides additional information for general education students.

Universal Tools for Online Testing. This module covers embedded universal tools and designated supports for students taking the South Dakota assessments.

What is a CAT? This module describes what a computer-adaptive test is and how it works when taking ELA/L and mathematics online assessments.

What Is a Performance Task? This module presents information on performance tasks and how they work when taking ELA/L and mathematics online assessments.

All four of these training modules are available on the portal (<https://sd.portal.cambiumast.com/resources>).

Practice and Training Test Site

Separate training and practice sites are available for TEs, PRs, and students. TEs and PRs can practice administering assessments and starting and ending test sessions on the PR training site, and students can practice taking online assessments on the student practice and training site. The South Dakota assessment practice tests mirror the corresponding summative assessments for ELA/L and mathematics. Each test provides students with a grade-specific testing experience, including a variety of question types and difficulty levels (approximately 30 items each in ELA/L and mathematics) as well as a performance task.

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

A student can log directly into the practice and training test site as a “Guest” without a PR-generated test session ID, or the student can log in through a training test session created by the TE or PR in the PR training site. Items in the student training test include all item types that are in the operational item pool, including multiple-choice items, grid items, and natural language items.

The practice test is available on the South Dakota portal at <https://sd.portal.cambiumast.com>.

Manuals and User Guides

All manuals and user guides pertaining to the 2023–2024 test administration can be found on the South Dakota portal at <https://sd.portal.cambiumast.com/>. ACs, DAs, and SCs can use these manuals and user guides to train TEs and PRs regarding test administration policies and procedures.

The *Assistive Technology Manual* provides an overview of the embedded and non-embedded assistive technology tools that can be used to help students with special accessibility needs complete online tests in TDS.

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

The *Online, Summative, Test Administration Manual* provides information for TEs and PRs administering the South Dakota online summative assessments in ELA/L and mathematics. It includes screen captures and step-by-step instructions on how to administer the online tests.

The *Paper, Summative, Test Administration Manual* provides information for TEs and PRs administering the South Dakota paper summative assessments in ELA/L and mathematics.

The *Secure Browser Requirements* provide instructions for downloading and installing the Secure Browser on supported operating systems used for online assessments. It also includes the technical specifications for online testing, including information on Internet and network requirements, general hardware and software requirements, and the text-to-speech function.

The *Quick Guide For Setting Up Your Online Technology System* document outlines the basic technology requirements for administering an online assessment, including operating system requirements and supported web browsers.

The *Reporting User Guide* provides instructions and support for users viewing the 2023–2024 interim assessment performance reports and handscoring in the reporting system.

The *Test Delivery System User Guide* is designed to help users navigate TDS, including the Student Interface and the PR Interface, and help TEs and PRs manage and administer online testing for students.

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

The *Tools, Supports, and Accommodations Guidelines (TSA)* provide information for school-level personnel and decision-making teams, particularly Individualized Education Program (IEP) teams, to use when selecting and administering universal tools, designated supports, and accommodations for students who need them.

All manuals, user guides, video tutorials, and quick guides are available on the South Dakota portal at <https://sd.portal.cambiumast.com/resources>.

2.3.2. District Trainings

The SDDOE provided in-person regional trainings and state-wide virtual trainings available to all districts during the 2023–2024 school year. The trainings took place February through March 2024 and topics included overviews of proctoring the South Dakota Math and ELA Assessments and the TIDE, TDS, Reporting, and DEI systems.

2.4. TEST SECURITY

All test items, test materials, and student-level testing information are secured 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 testing system also protect test security. This section describes system security, student confidentiality, and policies on testing impropriety.

2.4.1. Student-Level Testing Confidentiality

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

Confirming that the right students are accessing appropriate test content involves three elements:

- *Test eligibility*, which refers to the assignment of a test for a specific student

- *Test accommodation*, which refers to the assignment of a test setting to specific students based on their needs
- *Test session*, which refers to the authentication process of a TE/PR creating and managing a test session, the TE/PR reviewing and approving a test (and its settings) for every student, and the student signing in to take the test

The public disclosure of student information or test results is prohibited by FERPA. Examples of prohibited practices include

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

Test materials and score reports that identify student names with test scores must not be sent to anyone other than authorized individuals with an appropriate need to know. If information about an individual test must be sent via email or fax, only the SSID number is included, not the student's name.

All students must be enrolled or registered at their testing schools to take the online, paper-pencil, or braille assessments. Student enrollment information, including demographic data, is generated at the district level, and uploaded directly into TIDE during the testing period.

Students log in to the online assessment using their legal first name, SSID number, and a test session ID. Only students can log in to an online test session. TEs, PRs, 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, TEs or PRs are required to affix the student Pre-ID label to the student's answer document.

After a test session, only staff with the administrative roles of AC, DA, SC, or TE can view their students' scores. PRs and PARAs do not have access to student scores.

2.4.2. System Security

The objective of system security is to ensure that all data are protected and accessed appropriately by the right user group. 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 be performed only by a specific, designated user.

A Hierarchy of Control. As described in Section 2.2, Test Options and Administrative Roles, SCs, PRs, and TEs have well-defined roles and access to the testing system. When the TIDE window opens, the SDDOE creates a verified list of ACs that is uploaded into TIDE. ACs are then responsible for selecting and entering SC information into TIDE, and ACs or SCs are responsible for entering PR and TE information into TIDE. Throughout the year, the ACs, DAs, and SCs are also expected to delete the information of any staff members in TIDE who have transferred, resigned, or no longer serve as educators in the designated school.

Password Protection. Access points for each system role—at the state, district, and school levels—require a password to log in to the system. Newly added users receive separate passwords through the email address assigned by the school.

Secure Browser. A key role of the Technology Coordinator (TC) 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 it prevents students from searching for answers on the Internet or communicating with other students through the school’s Internet connection. The assessments can be accessed only through the CAI Secure Browser and not by other Internet browsers.

Take a Test App. The TC may also choose to set up Windows 10 computers for testing with the native Take a Test application. Developed by Microsoft, the Take a Test app enforces a locked-down, secure testing environment identical to CAI’s Secure Browser. Users of the Take a Test app do not need to install the CAI Secure Browser on the testing machine. This application is configurable based on user needs. South Dakota had approved the application for student testing in the winter of 2018 but had stopped support after its low usage in winter 2019.

2.4.3. Security of the Testing Environment

The ACs, SCs, TEs, and PRs work together to determine appropriate testing schedules based on the number of computers available, the number of students in each tested grade, and the average length of time needed to complete each assessment. PARAs will assist TEs and PRs in the administration of the assessments, as needed.

Testing personnel are reminded in the online training, face-to-face 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.

TEs and PRs must establish procedures to maintain a quiet environment during each test session, recognizing that some students may finish more quickly than others. When students are allowed to leave the testing room upon test completion, TEs or PRs must explain the procedures for leaving and where they are expected to report once they leave. Students must not disrupt others while leaving the testing environment. If students are expected to remain in the testing room until the end of the session, TEs or PRs are encouraged to prepare some quiet work for students to do after they finish the assessment. The work must be on a different subject than in the assessment the students just completed.

If a student needs to leave the room for a brief time, the TEs or PRs must pause the student’s assessment. For the CAT, if the pause lasts longer than 20 minutes, the student can continue with the rest of the assessment in a new test session, but the system will not allow the student to return to the answers he or she provided 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

out of sight in the testing room. Rooms should have minimized access by others; TEs and PRs are encouraged to post signs in halls and entrances to promote optimum testing conditions. A sign indicating “TESTING—DO NOT DISTURB” should be affixed to testing room doors.

Seating Arrangements

When arranging testing room seating, TEs and PRs should provide adequate spacing between students’ seats. Students should be seated so 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, appropriate seating arrangements should still discourage them from communicating. For the performance tasks, different forms are spiraled within a classroom so that students do not receive the same form as their neighbors.

After the Test

At the end of a test session, TEs or PRs 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 at the end of each test session.

For the paper-pencil assessment versions, the *Paper-Pencil Test Administration Manual* for mathematics or ELA/L provides specific instructions on how to package and secure the test booklets so that they can be properly returned to the testing contractor’s office.

2.4.4. Test Security Violations

Anyone who administers or proctors a South Dakota assessment is responsible for understanding the assessment security procedures and prohibited practices. Prohibited practices, as detailed in the *Online, Summative Test Administration Manual*, fall into three groups:

Impropriety. A test security incident that has a minor impact on the individual or group of students who are testing and has a low risk of potentially affecting student performance on the test, test security, or test validity (for example, students leaving the testing room without authorization).

Irregularity. A test security incident that affects an individual or group of students who are testing and may affect student performance on the test, test security, or test validity. These circumstances, such as a fire drill or other disruption, can be contained at the local level.

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

District and school personnel must document and submit all test security incidents using the TIDE Testing Irregularity form to the SDDOE Office of Assessment. The forms are housed in TIDE and are the record for all test security incidents.

2.5. STUDENT PARTICIPATION

All students (including retained students) currently enrolled in grades 3–8 and 11 at public schools in South Dakota must participate in the South Dakota assessments (Note: some non-public and BIE schools utilize South Dakota assessments for their required standardized testing.). Students are tested in the enrolled grade assessment, with the exception of some grade 10 and 12 students. Out-of-grade-level testing must be approved by the state.

2.5.1. Exempt Students

The following students are exempt from participating in the South Dakota assessments:

- A student who has a significant medical emergency, with the approval of the SDDOE Office of Assessment
- A Limited English Proficiency (LEP) student who has moved to the country within the school year (ELA/L exemption only)

2.6. ONLINE TESTING FEATURES AND TESTING ACCOMMODATIONS

South Dakota’s *Tools, Supports, and Accommodations Guidelines (TSA)* are modified from the Smarter Balanced Assessment Consortium’s *Usability, Accessibility, and Accommodations Guidelines*. The *TSA* is intended for school-level personnel and decision-making teams, including IEP and Section 504 Plan teams, as they prepare for and implement the South Dakota assessments. The *TSA* provides 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 students who need them. The *TSA* is also intended for assessment staff and administrators who oversee decisions made in instruction and assessment.

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

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

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

All embedded and non-embedded universal tools will be activated for all students to use during a test session. A TE or PR can deactivate SDDOE-approved universal tools in the testing system’s PR Interface 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 South Dakota’s *TSA* for complete information. This document is located at <https://sd.portal.cambiumast.com/resources>.

2.6.1. Online Universal Tools for All Students

Universal tools are features of an assessment or exam that are embedded or non-embedded components of the test administration system. Universal tools are available to all students based on their preference and selection, and they have been preset in the Test Information Distribution System (TIDE). In the 2023–2024 test administration, the following universal tools were available for all students. For specific information on how to access and use these features, refer to the *Test Delivery System User Guide*, found at <https://sd.portal.cambiumast.com>.

Embedded Universal Tools

Breaks (Pauses). A student may pause the assessment and return to the test question he or she was working on. However, if an assessment is paused for more than 20 minutes, students cannot return to previous test questions if they have navigated to the next segment (for a CAT or PT) or advanced to the next page (for a CAT).

Calculator. Students in grades 6-8, 11 can access an embedded on-screen digital calculator by clicking the calculator icon. This tool is available only with the specific items that the South Dakota item specifications indicate are appropriate.

Digital Notepad. This tool is used for making notes about an item. The digital notepad is item specific and is available through the end of 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 embedded English dictionary is available for the full-write portion of an ELA/L performance task.

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

Expandable Passages and/or Items. 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/L performance tasks in which students complete a full-write. The student clicks the notepad icon for the notepad to appear. During the ELA/L performance tasks, the notes are retained from segment to segment so that the student may return to the notes even though he or she cannot return to specific items in the previous segment.

Highlighter. The student can highlight passages or sections of passages and test questions.

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

Line Reader. The student uses a universal onscreen tool to assist in reading by raising and lowering the tool for each line of text on the screen.

Mark for Review. A student can mark a question and return to it later. However, for the CAT, if the assessment is paused for more than 20 minutes, students cannot return to marked test questions.

Mathematics Tools. These digital tools (e.g., embedded ruler, protractor) are used for measurement and are available only with mathematics items for which the South Dakota item specifications deem them appropriate.

Spellcheck. This is a writing tool for checking the spelling of words in student-generated responses. Spellcheck gives an indication only that a word is misspelled; it does not provide the correct spelling. This tool is available only with the specific items for which the South Dakota item specifications indicate that it would be appropriate. It 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. A student may use the strikethrough function to cross out response options.

Thesaurus. An embedded on-screen thesaurus is available for the full-write portion of an ELA/L performance task. A thesaurus contains synonyms of terms used in the assessment. A full-write is the second part of a performance task. The use of this universal tool may result in the student needing additional overall time to complete the assessment.

Tutorials. The student can watch a short video demonstrating how to respond to a particular question type.

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

Zoom. The student can zoom in on test questions, text, and graphics up to 3X.

Non-Embedded Universal Tools

Breaks. Breaks may be given at predetermined intervals or after the completion of sections of the assessment for students taking a paper-pencil test. Sometimes, students can take breaks if needed to reduce cognitive fatigue when they experience heavy assessment demands. The use of this universal tool may result in the student needing additional overall time to complete the assessment.

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

Scratch Paper. Scratch paper to make notes, write computations, or record responses may be made available. Only plain or lined paper is appropriate for ELA/L. Plain graph paper (no coordinate plane or other graphics) is highly recommended for grade 6 and above and can be used on all mathematics assessments. A whiteboard with a marker may be used as scratch paper. If the construct being measured is not affected, students may use assistive technology devices including low-tech assistive technologies such as Math Window, which are permitted to help students make notes. The assistive technology device must be consistent with the student's IEP or Section 504 Plan. Access to the Internet must be disabled on assistive technology devices.

- *Scratch Paper for the CAT (Computer-Adaptive Test).* 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 also be erased.

- *Scratch Paper for the Performance Tasks.* For mathematics and ELA/L, if a student needs to take the assessment in more than one session, scratch paper, whiteboards, and/or assistive technology devices may 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, the scratch paper must be collected and securely destroyed, and whiteboards and notes on assistive technology devices must be erased to maintain test security.

Thesaurus. A thesaurus provides synonyms of terms. While a student interacts with text included in the assessment, a thesaurus is made available for a full-write. A full-write is the second part of a performance task. The use of this universal tool may result in the student needing additional overall time to complete the assessment.

2.6.2. Designated Supports and Accommodations

Designated supports for the South Dakota 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 a parent or guardian and the 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 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 changes in procedures or materials that increase equitable access during the South Dakota assessments. Assessment accommodations generate valid assessment results for students who need them; they allow these students to show what they know and can do. Accommodations are available for students with documented IEPs or Section 504 Plans. Consortium-approved accommodations do not compromise the learning expectations, construct, grade-level standard, or intended outcome of the assessments.

Embedded Designated Supports

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

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

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

Mouse Pointer. Students may be given a mouse pointer of various colors or sizes. Pointer colors include black, green, yellow, red, and white.

Streamline. This accommodation provides a streamlined interface of the test in an alternative, simplified format in which the items appear below the stimuli. When Streamlined Mode is turned on, it disables the ability to use the split-screen feature.

Text-to-Speech. Text-to-speech is allowed for all mathematics stimuli and items. Text-to-speech is allowed for ELA/L PT in three categories: items only, stimuli only, and both stimuli and items. It is also available for ELA/L CAT items. To have text-to-speech for ELA/L stimuli, the student needs an accommodation (see the Embedded Accommodations section for ELA/L CAT reading passages). Items refer to the actual questions being asked to the student and include response options or choices. Stimuli is anything that leads to the question. For example, stimuli may be a description of something related to the test items, could include a diagram, or could be a short passage to help establish the premise of the items that follow it. Text is read aloud to the student via embedded text-to-speech technology. The student can control the speed and raise or lower the volume of the voice via a volume control. This support is also available in Spanish for mathematics items.

Translated Test Directions (for mathematics). Translation of test directions is a language support available before students begin the actual test items. Students can see test directions in another language. As an embedded designated support, translated test directions are automatically part of the stacked translation designated support.

Translations (Glossaries) (for mathematics). Translated glossaries are a language support and are provided for selected construct-irrelevant terms for mathematics. Translations for these terms appear on the computer screen when students click them. The following language glossaries were offered in SY 2023–2024: Arabic, Burmese, Cantonese, Filipino/Tagalog, Hmong, Korean, Mandarin, Punjabi, Russian, Somali, Spanish, Ukrainian, and Vietnamese.

Translations (Dual Language) (for mathematics). Dual language translations are a linguistic support available for some students. These provide the full translation of each English test item and stimulus.

Turn Off Any Universal Tools. Teachers can disable any universal tools that might be distracting or that students do not need or are unable to use. South Dakota has an approved subset of universal tools that can be tuned off.

Zoom. To increase the default print size for the entire test, the print size must be set for the student in TIDE or set by a TA before the start of the test.

Non-Embedded Designated Supports

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

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

Color Contrast. Test content of online items may be printed on different-colored paper.

Color Overlays. Color transparencies are placed over a paper-based assessment.

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

Magnification. Students may adjust the size of specific areas of the screen (e.g., text, formulas, tables, graphics, and navigation buttons) with an assistive technology device. Magnification allows increasing the size to a level not provided for by the zoom universal tool.

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

Noise Buffers. Noise buffers include ear mufflers, white noise, and other equipment to reduce environmental noises.

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

Read Aloud. The read-aloud function is available for mathematics and ELA/L items, except passages. Text is read aloud to the student by a trained and qualified human reader who follows the administration guidelines provided in the *Online, Summative, Test Administration Manual* and the *Guidelines for Read Aloud, Test Reader*. All or portions of the content may be read aloud, except ELA/L passages.

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 *Online, Summative, Test Administration Manual* and in the *Guidelines for Read Aloud, Test Reader*. All or portions of the content may be read aloud.

Separate Setting. The student is tested in a setting different from that which is available for most students.

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

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

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

Translations (Glossaries) (for mathematics paper-pencil tests). Translated glossaries are a language support provided for selected construct-irrelevant terms for mathematics. Glossary terms are listed by item and include the English term and its translated equivalent.

Embedded Accommodations

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

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

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

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

Permissive Mode. Permissive Mode is required by the contractor for use of assistive technology devices. Use of an assistive technology device may require Permissive Mode to be set in TIDE (i.e., alternate response options, amplification devices, speech-to-text, etc.).

Speech to Text. This allows student to dictate their verbal response into English. This tool is only available on constructed response items.

Text-to-Speech (for ELA/L CAT reading passages or reading items and passages). For the CAT portion of the ELA/L assessment, text-to-speech is available for reading passages or both reading items and passages. The reading passages refer to the text that is on the left-hand side of the screen and items refer to the text that is on the right-hand side of the screen. The 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 is appropriate for a very small number of students and is available to those whose need is documented in an IEP or Section 504 Plan.

Word Prediction. Word prediction allows students to begin writing a word and choose from a list of words that have been predicted from word frequency and syntax rules.

Non-Embedded Accommodations

100s Number Table. A paper-based list of all the digits from 1 to 100 in table format will be available from Smarter Balanced for reference and was approved for use on the South Dakota mathematics assessment.

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

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

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

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

Large Print. A large print paper form of the test that is provided to the student with a visual impairment.

Multiplication Table. A paper-based single digit (1–12) multiplication table is available from Smarter Balanced and approved for use on the South Dakota mathematics assessment.

Print-on-Demand. Paper copies of passages, stimuli, and/or items are printed for students. For those students needing a paper copy of a passage or stimulus, permission to request printing must first be set in TIDE. For those students needing a paper copy of one or more items, the SDDOE must be contacted by the SC or DC to review the student’s case before setting the accommodation for the student.

Read Aloud (for ELA/L reading passages). Text is read aloud to the student via an external screen reader or by a trained and qualified human reader who follows the administration guidelines provided in Appendix

D in the *Online, Summative, Test Administration Manual* and in the *Tools, Supports, and Accommodations Guidelines*. All or portions of the content may be read aloud. Members can refer to the *Guidelines for Choosing the Read Aloud Accommodation* when deciding if this accommodation is appropriate for a student.

Scribe. Students dictate their responses to a human, who then records verbatim what they dictate. The scribe must be trained and qualified and must follow the administration guidelines provided in the *Online, Summative, Test Administration Manual* and in the *Tools, Supports, and Accommodations Guidelines*.

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., opening application programs, pulling down menus, saving work). Voice recognition software generally can recognize speech up to 160 words per minute. Students may use their own assistive technology devices.

Word Prediction. Word prediction allows students to begin writing a word and choose from a list of words that have been predicted from word frequency and syntax rules. Word prediction is delivered via a non-embedded software program. Students may use their own assistive technology devices.

Table 4 presents a list of the universal tools, designated supports, and accommodations offered in the 2023–2024 administration. Tables 5–10 provide the number of students who utilized any of the offered accommodations and designated supports.

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

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

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

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

² For ELA/L performance task full-writes

³ For ELA/L performance tasks

⁴ Includes embedded ruler, embedded protractor

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

⁶ For mathematics items

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

⁸ For mathematics items

⁹ For ELA/L listening items and mathematics items

¹⁰ For ELA/L listening passages

¹¹ For ELA/L reading passages: must be set in TIDE by state-level user and must be set before test begins

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

¹³ For mathematics items on the paper-pencil test

¹⁴ Includes adapted keyboards, large keyboard, StickyKeys, MouseKeys, FilterKeys, adapted mouse, touch screen, head wand, and switches

¹⁵ For paper-pencil assessments

¹⁶ For ELA/L reading passages, all grades

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

Accommodations	Grade						
	3	4	5	6	7	8	11
Embedded Accommodations							
American Sign Language	1	3	1	1	3	0	1
Closed Captioning	0	4	3	4	4	1	5
Permissive Mode	1	1	1	1	2	0	0
Speech-to-Text	30	50	67	44	37	28	18
Text-to-Speech: Reading Passages and Items	115	134	111	116	84	99	74
Word Prediction	4	5	12	7	3	3	3
Non-Embedded Accommodations							
Alternate Response Options	0	2	0	0	2	0	0
Print-on-Demand: Passages and Items	0	1	0	0	1	0	0
Read Aloud Passages	4	6	5	2	4	1	5
Scribe	23	30	28	16	15	14	9

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

Designated Supports	Subgroup	Grade						
		3	4	5	6	7	8	11
Color Contrast	Overall	8	9	10	14	12	5	1
	LEP	1	0	0	0	0	0	0
	IDEA	3	3	3	4	7	0	0
Masking	Overall	15	7	6	21	10	14	0
	LEP	2	0	0	2	1	1	0
	IDEA	12	6	5	1	2	3	0
Mouse Pointer	Overall	3	1	1	1	0	1	1
	LEP	0	0	0	0	0	0	0
	IDEA	1	1	1	1	0	1	1
Streamline	Overall	4	5	0	10	10	9	1
	LEP	1	0	0	0	0	0	0
	IDEA	4	5	0	5	8	9	1
Text-to-Speech: CAT Items	Overall	1,884	1,847	1,689	1,324	1,246	1,248	733
	LEP	399	435	311	269	266	289	219
	IDEA	1,040	1,092	1,081	817	728	723	346
Text-to-Speech: PT Items	Overall	310	283	271	225	215	197	87
	LEP	24	28	13	23	22	22	15
	IDEA	247	237	237	192	183	161	68
Text-to-Speech: PT Passages	Overall	4	2	5	2	1	4	0
	LEP	2	1	3	1	0	1	0
	IDEA	1	2	1	1	0	1	0
Text-to-Speech: PT Passages and Items	Overall	1,702	1,698	1,519	1,209	1,114	1,132	719
	LEP	400	418	303	251	246	271	208
	IDEA	902	974	933	723	623	640	347
Zoom	Overall	6	2	4	4	1	6	3
	LEP	0	0	1	1	0	0	0
	IDEA	5	1	3	2	1	4	1

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

Designated Supports	Subgroup	Grade						
		3	4	5	6	7	8	11
Amplification	Overall	2	3	2	2	4	4	2
	LEP	0	0	0	0	0	0	0
	IDEA	2	3	1	0	1	2	1
Bilingual Dictionary	Overall	2	4	7	15	11	19	10
	LEP	2	4	6	15	11	19	10
	IDEA	0	0	0	0	0	3	0
Color Contrast	Overall	1	3	3	1	2	4	0
	LEP	1	0	0	1	0	0	0
	IDEA	0	3	3	1	2	3	0
Color Overlay	Overall	1	2	3	0	1	1	0
	LEP	0	0	1	0	0	0	0
	IDEA	1	2	3	0	1	0	0
Magnification	Overall	2	4	5	2	4	3	4
	LEP	1	0	2	1	0	0	0
	IDEA	1	4	4	2	4	3	2
Medical Supports	Overall	3	1	4	6	3	4	16
	LEP	0	0	1	0	0	0	0
	IDEA	0	0	2	1	0	0	1
Noise Buffers	Overall	12	5	12	8	5	5	6
	LEP	1	0	2	1	2	0	0
	IDEA	12	5	9	5	2	5	5
Read Aloud: Items	Overall	51	63	63	40	45	33	26
	LEP	4	3	10	2	2	4	0
	IDEA	48	59	54	38	40	30	24
Read Aloud: Passages	Overall	30	44	48	33	30	25	20
	LEP	2	1	7	1	1	3	0
	IDEA	27	41	41	32	26	23	18
Separate Setting	Overall	1,074	1,199	1,103	861	840	802	566
	LEP	229	255	182	76	83	86	32
	IDEA	865	950	934	744	700	649	463
Simplified Test Directions	Overall	473	532	427	342	370	378	260
	LEP	222	259	166	135	137	148	148
	IDEA	255	289	286	217	231	227	143
Translated Test Directions	Overall	97	95	68	57	59	59	95
	LEP	97	95	68	56	59	59	94
	IDEA	13	12	13	11	6	8	15
Translated Test Directions in ASL	Overall	1	0	0	0	0	0	0
	LEP	0	0	0	0	0	0	0
	IDEA	1	0	0	0	0	0	0

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

Accommodations	Grade						
	3	4	5	6	7	8	11
Embedded Accommodations							
American Sign Language	3	3	1	1	3	1	1
Permissive Mode	1	1	1	2	2	0	0
Speech-to-Text	23	45	51	38	25	24	12
Word Prediction	2	5	11	6	2	0	3
Non-Embedded Accommodations							
100s Number Table	131	171	100	44	30	17	0
Abacus	2	0	0	0	0	0	0
Alternate Response Options	0	2	0	0	2	0	0
Multiplication Table	196	334	404	318	263	266	54
Print-on-Demand: Stimuli and Items	0	1	0	1	1	0	0
Scribe	22	27	23	13	14	13	9

Table 9. Mathematics Total Students with Allowed Embedded Designated Supports

Designated Supports	Subgroup	Grade						
		3	4	5	6	7	8	11
Color Contrast	Overall	8	9	10	14	12	5	2
	LEP	1	0	0	0	0	0	0
	IDEA	3	3	3	4	7	0	1
Illustration Glossaries	Overall	226	259	174	168	170	168	145
	LEP	218	245	162	155	153	162	145
	IDEA	26	38	35	26	18	19	22
Masking	Overall	16	7	7	20	9	14	0
	LEP	2	0	0	2	1	1	0
	IDEA	12	6	6	0	1	3	0
Mouse Pointer	Overall	3	1	1	1	0	1	1
	LEP	0	0	0	0	0	0	0
	IDEA	1	1	1	1	0	1	1
Streamline	Overall	5	4	0	9	9	10	1
	LEP	1	0	0	0	0	0	0
	IDEA	5	4	0	5	8	9	1
Text-to-Speech: Items	Overall	112	133	91	72	69	48	27
	LEP	3	10	5	4	2	3	3
	IDEA	103	114	76	55	59	46	22
Text-to-Speech: Stimuli	Overall	3	1	5	4	2	4	1
	LEP	0	0	3	1	0	0	0
	IDEA	3	1	2	3	2	4	1
Text-to-Speech: Stimuli and Items	Overall	1,968	1,919	1,749	1,419	1,311	1,329	790
	LEP	483	495	354	313	312	328	229
	IDEA	1,044	1,102	1,090	865	747	759	394
Translations (Glossaries): Spanish	Overall	77	82	57	73	84	74	71
	LEP	76	82	57	70	80	72	71
	IDEA	4	7	7	8	5	6	6

Designated Supports	Subgroup	Grade						
		3	4	5	6	7	8	11
Translations (Glossaries): Other Languages	Overall	10	7	13	12	13	19	4
	LEP	9	7	13	12	13	19	4
	IDEA	1	0	0	1	1	2	2
Translations (Dual Language): Spanish	Overall	12	23	20	18	25	22	29
	LEP	12	23	19	15	23	20	29
	IDEA	0	0	0	0	0	1	0
Zoom	Overall	6	2	4	4	1	6	3
	LEP	0	0	1	1	0	0	0
	IDEA	5	1	3	2	1	4	1

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

Designated Supports	Subgroup	Grade						
		3	4	5	6	7	8	11
Amplification	Overall	2	3	2	2	4	4	2
	LEP	0	0	0	0	0	0	0
	IDEA	2	3	1	0	1	2	1
Color Contrast	Overall	1	3	2	1	1	4	0
	LEP	1	0	0	1	0	0	0
	IDEA	0	3	2	1	1	3	0
Color Overlay	Overall	1	2	3	0	1	1	0
	LEP	0	0	1	0	0	0	0
	IDEA	1	2	3	0	1	0	0
Illustration Glossaries	Overall	1	0	0	0	0	0	0
	LEP	1	0	0	0	0	0	0
	IDEA	0	0	0	0	0	0	0
Magnification	Overall	2	4	4	1	3	3	4
	LEP	1	0	1	1	0	0	0
	IDEA	1	4	3	1	3	3	2
Medical Supports	Overall	3	1	4	6	3	4	15
	LEP	0	0	1	0	0	0	0
	IDEA	0	0	2	1	0	0	1
Noise Buffers	Overall	13	5	12	8	5	4	6
	LEP	2	0	2	1	2	0	0
	IDEA	12	5	9	5	2	4	5
Read Aloud: Items	Overall	46	56	54	39	45	33	24
	LEP	3	2	6	2	2	3	0
	IDEA	44	52	47	37	40	32	23
Read Aloud: Stimuli	Overall	33	50	44	37	39	32	22
	LEP	2	2	3	1	0	4	0
	IDEA	30	47	39	36	36	30	20
Read Aloud in Spanish: Items	Overall	3	6	8	4	4	3	3
	LEP	3	4	6	2	4	3	2
	IDEA	0	2	1	1	0	0	1

Designated Supports	Subgroup	Grade						
		3	4	5	6	7	8	11
Read Aloud in Spanish: Stimuli	Overall	3	7	6	3	3	4	3
	LEP	3	5	5	2	3	2	1
	IDEA	0	2	1	0	0	1	1
Separate Setting	Overall	1,100	1,236	1,121	881	863	826	567
	LEP	259	294	207	95	99	105	34
	IDEA	857	947	924	743	705	652	462
Simplified Test Directions	Overall	499	562	454	365	402	408	271
	LEP	250	293	195	162	168	171	157
	IDEA	250	283	284	212	227	229	143
Translated Test Directions	Overall	126	132	97	73	83	74	99
	LEP	124	132	96	72	81	74	99
	IDEA	13	13	13	11	7	8	13
Translated Test Directions in ASL	Overall	1	0	0	0	0	0	0
	LEP	0	0	0	0	0	0	0
	IDEA	1	0	0	0	0	0	0
Translation (Glossaries): Spanish	Overall	15	16	12	19	19	19	19
	LEP	15	16	12	16	18	18	19
	IDEA	1	0	0	0	0	1	0
Translation (Glossaries): Other Languages	Overall	1	2	4	2	1	1	0
	LEP	1	2	4	2	1	1	0
	IDEA	0	0	0	0	0	0	0

2.7. TESTING TIME

The online environment allows item response time to be captured as the item page time (the time each item page is presented) 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 South Dakota assessments are not timed, and an individual student may need more or less time overall. The length of a test session is determined by TEs or PRs who are knowledgeable about the class periods in the school's instructional schedule and the timing needs associated with the assessments. Students should be allowed extra time if they need it, but TEs or PRs must use their best professional judgment when allowing students extra time.

Tables 11 and 12 present the average test-taking time and the testing time at percentiles for the overall test, the CAT component, and the PT component.

Table 11. ELA/L Testing Time

Grade	Average Testing Time (hh:mm)	SD* of Testing Time (hh:mm)	Median Testing Time (hh:mm)	Testing Time in Percentiles (hh:mm)				
				75th	80th	85th	90th	95th
Overall Test								
3	3:04	1:41	2:46	3:49	4:08	4:32	5:08	6:09
4	3:06	1:32	2:50	3:53	4:12	4:33	5:04	5:55
5	3:05	1:27	2:51	3:47	4:03	4:24	4:51	5:41
6	2:53	1:24	2:39	3:32	3:47	4:07	4:33	5:18
7	2:46	1:16	2:34	3:26	3:42	4:00	4:24	5:01
8	2:36	1:13	2:25	3:13	3:26	3:44	4:08	4:47
11	2:10	0:56	2:06	2:38	2:46	2:56	3:12	3:42
CAT Component								
3	1:25	0:42	1:18	1:44	1:52	2:01	2:15	2:39
4	1:26	0:39	1:21	1:44	1:51	2:00	2:13	2:36
5	1:26	0:37	1:21	1:44	1:51	2:00	2:12	2:31
6	1:30	0:39	1:25	1:48	1:55	2:04	2:16	2:37
7	1:25	0:36	1:21	1:42	1:49	1:57	2:08	2:27
8	1:20	0:34	1:16	1:37	1:43	1:50	2:00	2:17
11	1:12	0:29	1:10	1:28	1:32	1:38	1:46	1:59
PT Component								
3	1:39	1:14	1:23	2:10	2:24	2:42	3:06	3:55
4	1:40	1:08	1:26	2:11	2:24	2:41	3:03	3:48
5	1:38	1:04	1:26	2:07	2:19	2:35	2:58	3:36
6	1:23	0:57	1:11	1:46	1:57	2:10	2:31	3:07
7	1:21	0:51	1:11	1:45	1:57	2:10	2:28	2:58
8	1:16	0:49	1:06	1:38	1:49	2:02	2:19	2:48
11	0:58	0:34	0:54	1:14	1:20	1:27	1:37	1:55

Note. SD: standard deviation

Table 12. Mathematics Testing Time

Grade	Average Testing Time (hh:mm)	SD* of Testing Time (hh:mm)	Median Testing Time (hh:mm)	Testing Time in Percentiles (hh:mm)				
				75th	80th	85th	90th	95th
Overall Test								
3	1:43	1:00	1:29	2:05	2:17	2:31	2:51	3:29
4	1:45	0:53	1:34	2:08	2:19	2:34	2:52	3:23
5	1:56	1:00	1:45	2:22	2:32	2:47	3:06	3:42
6	1:51	0:54	1:42	2:12	2:21	2:33	2:51	3:22
7	1:36	0:43	1:29	1:57	2:05	2:16	2:29	2:52
8	1:36	0:45	1:30	1:56	2:04	2:15	2:29	2:54
11	1:21	0:35	1:17	1:41	1:48	1:55	2:05	2:21
CAT Component								
3	1:07	0:40	0:58	1:21	1:29	1:39	1:54	2:20
4	1:11	0:37	1:03	1:26	1:34	1:44	1:58	2:20
5	1:13	0:38	1:07	1:29	1:36	1:45	1:57	2:21
6	1:13	0:34	1:07	1:27	1:33	1:41	1:52	2:11
7	1:13	0:33	1:07	1:29	1:35	1:43	1:53	2:10
8	1:11	0:34	1:06	1:25	1:31	1:39	1:49	2:09
11	0:57	0:25	0:54	1:11	1:16	1:22	1:29	1:41
PT Component								
3	0:35	0:28	0:28	0:44	0:49	0:56	1:07	1:25
4	0:34	0:24	0:29	0:43	0:47	0:53	1:02	1:18
5	0:43	0:32	0:35	0:54	1:00	1:09	1:20	1:39
6	0:38	0:28	0:33	0:47	0:51	0:57	1:06	1:23
7	0:23	0:16	0:20	0:29	0:32	0:37	0:43	0:52
8	0:26	0:17	0:22	0:32	0:35	0:39	0:45	0:56
11	0:23	0:15	0:21	0:31	0:34	0:38	0:42	0:51

Note. SD: standard deviation

2.8. DATA FORENSICS PROGRAM

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

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

Analyses are performed at the student level and summarized for each aggregate unit, including testing session, PR, and school. Flagging criteria used for these analyses are described below and are configurable by an authorized user. When the aggregate unit size is small, the aggregate unit is flagged if the percentage of flagged students is greater than 50% in the analysis. The default small aggregate unit size is five or fewer students, but this value is configurable. For each aggregate unit, small groups are identified based on the number of tests included in the aggregate unit from that analysis. Thus, a small unit identified in one analysis

may not be a small unit in another analysis. The QA reports are provided to state clients to review and ensure the test integrity after the testing window closes.

2.8.1. Changes in Student Performance

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

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

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

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

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

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

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

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

2.8.2. Test-Taking Time

The summative assessments are not timed, and thus individual test-taking times may vary across students. However, unusual test-taking times such as excessively shorter or longer test-taking times may indicate irregularities in test administration. An example of an unusual test-taking time is a test record for an individual who scores very well on the test, even though the average time spent is far less than that required of students statewide. If students already know the answers to the questions, the test-taking time may be much shorter than those who have no prior knowledge of the item content. Conversely, if a TE or PR helps students by coaching them to change their responses during the test, the testing time could be longer than expected.

The state average testing time and standard deviation are computed based on all students available when the analysis was performed. Students and aggregate units are flagged if the test-taking time is different from the state average by three standard deviations or more, although the flagging criteria can be adjusted by an authorized user.

2.8.3. Inconsistent Item Response Pattern (Person Fit)

In item response theory (IRT) models, person-fit measurement is used to identify test takers whose response patterns are improbable given an IRT model. If a test has psychometric integrity, little irregularity will be seen in the item responses of the individual who responds to the items fairly and honestly.

If a test taker has prior knowledge of some test items (or is provided answers during the exam), he or she will respond correctly to those items at a higher probability than indicated by his or her ability as estimated across all items. In this case, the person-fit index will be large for the student. However, if a student has prior knowledge of the entire test content, this will not be detected based on the person-fit index, although the item response time index might flag such a student.

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

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

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

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

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

2.8.4. Item Response Change

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

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

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

2.9. PREVENTION AND RECOVERY OF DISRUPTIONS IN TEST DELIVERY SYSTEM

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

A unique monitoring system has been developed by CAI that is very sensitive to changes in server performance. Most monitoring systems provide warnings when something goes wrong. CAI's monitoring system does, too, but it also provides warnings when any given server is performing differently from its performance over the prior few hours or differently than the other servers performing the same jobs. Subtle changes in performance often precede actual failure by hours or days, allowing CAI to detect potential problems, investigate them, and mitigate them *before* a failure. On multiple occasions, this has enabled CAI to make adjustments and replace equipment before any problems occurred.

Clients are also alerted within minutes of any disruption because of CAI's escalation procedure. This emergency alert system notifies CAI's executive and technical staff by text message; staff then immediately communicate with one another to understand the problem.

The next section describes CAI's system architecture and how it recovers from device failures, Internet interruptions, and other problems.

2.9.1. High-Level System Architecture

CAI's high-level system architecture provides the redundancy, robustness, and reliability required by a large-scale, high-stakes testing program. CAI's general approach, which has been adopted by South Dakota and Smarter Balanced as standard policy, is pragmatic and well supported by the system architecture.

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

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

Student Machines

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

Responses are saved asynchronously, with a background process on the student machine waiting for confirmation of successfully stored data on the server. If confirmation is not received within the designated time (usually 30–90 seconds), the system will prevent the student from doing any more work until connectivity is restored. The student is offered the choice of asking the system to try again or pausing the test and returning later. 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 restored silently, the student is prevented from testing and given the option of logging out or retrying the save.
- If the system fails completely, upon logging back into the system, the student returns to the item he or she was viewing when the failure occurred.

In short, data integrity is preserved by confirmed saves to system 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 a redundant array of independent disks (RAID) system to mitigate the risk of disk failure. Each response is stored on multiple independent disks.

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

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

Hub

Hub servers are redundant clusters of database servers with RAID drive systems. Hub servers continuously gather data from the test delivery satellites and their mini-hubs and store these data as described here. 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 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. Any anomalies (such as unscored or missing items, unexpected test lengths, or other unlikely issues) are flagged, and an immediate notification goes out to CAI's psychometricians and project team.

Database of Record

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

2.9.2. Automated Backup and Recovery

Every system is backed up nightly. Industry-standard backup and recovery procedures are in place to ensure 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.

2.9.3. Other Disruption Prevention and Recovery

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

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

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

3. SUMMARY OF 2023–2024 OPERATIONAL TEST ADMINISTRATION

3.1. STUDENT POPULATION

All students enrolled in grades 3–8 and 11 in all public elementary and secondary schools must participate in the South Dakota English language arts/literacy (ELA/L) and mathematics assessments. School districts are responsible for maintaining current student enrollment records in Infinite Campus. This enrollment data flows through a series of connected systems. The data is first shared with SD-STARS (South Dakota's Longitudinal Data System), which processes this information to generate a student file each night. SD-STARS then delivers this nightly file to Cambium, whose TIDE system updates daily with the new enrollment information. These consistently updated enrollment files serve an important purpose: they allow for accurate calculation of test participation rates by showing what percentage of enrolled students have attempted the test. Tables 13 and 14 present the participation rates, i.e., the percentage of students who attempted the test by subgroups. Tables 15 and 16 present the number of South Dakota students who meet attemptedness requirements for scoring and reporting the results of the South Dakota summative assessments.

Table 13. Participation Rates by Percentage in ELA/L Summative Assessment

Group	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 11
All Students	97.7	97.8	97.8	97.8	97.7	97.5	97.4
Female	98.1	98.1	98.2	98.1	97.9	97.9	97.5
Male	97.4	97.5	97.4	97.5	97.5	97.2	97.4
African American	93.2	94.3	95.9	96.4	94.8	95.6	94.2
AmerIndian/Alaskan	96.5	97.2	96.6	94.3	95.2	92.8	94.2
Asian	94.4	95.1	95.2	96.6	91.8	96.3	96.4
Hispanic	94.3	94.1	94.2	94.4	93.1	94.0	95.3
Pacific Islander	100.0	100.0	94.1	94.1	95.0	93.3	93.8
White	98.6	98.6	98.6	98.8	98.8	98.7	98.2
Multi-Racial	98.2	98.7	97.7	98.8	99.2	97.8	97.7
LEP	91.0	89.4	88.5	88.0	85.9	89.1	90.4
IDEA	94.0	94.5	93.3	92.9	92.8	91.6	89.4
Section 504 Plan	97.9	98.6	99.2	98.6	98.7	97.3	98.3

Note. AmerIndian/Alaskan = American Indian/Alaskan Native; LEP = limited English proficiency status

Table 14. Participation Rates by Percentage in Mathematics Summative Assessment

Group	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 11
All Students	98.3	98.4	98.3	98.3	98.3	97.9	97.6
Female	98.7	98.7	98.8	98.5	98.4	98.3	97.7
Male	98.0	98.2	97.8	98.0	98.1	97.6	97.5
African American	98.2	96.8	97.3	98.4	97.7	98.2	94.8
AmerIndian/Alaskan	96.3	97.1	96.4	95.2	95.3	93.3	94.3
Asian	96.6	98.4	100.0	98.5	95.3	98.1	95.8
Hispanic	98.5	99.1	98.3	97.9	98.0	97.7	97.2
Pacific Islander	100.0	100.0	94.1	100.0	100.0	93.3	93.8
White	98.7	98.6	98.7	98.8	98.8	98.7	98.2
Multi-Racial	97.9	98.9	97.7	98.9	99.2	97.7	97.7
LEP	99.0	98.5	98.5	97.6	98.4	97.9	94.5
IDEA	94.2	94.3	93.2	93.1	92.9	91.8	89.6
Section 504 Plan	97.9	98.3	99.2	98.6	98.7	97.3	98.3

Table 15. Number of Students in ELA/L Summative Assessment

Group	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 11
All Students	10,372	10,650	10,607	10,464	10,511	10,575	9,876
Female	4,982	5,270	5,173	5,115	5,200	5,095	4,746
Male	5,390	5,380	5,434	5,349	5,311	5,480	5,130
African American	315	348	325	353	327	325	303
AmerIndian/Alaskan	1,151	1,229	1,167	1,164	1,161	1,092	857
Asian	168	173	178	197	157	155	163
Hispanic	832	895	869	853	828	846	744
Pacific Islander	12	9	16	16	19	14	15
White	7,239	7,383	7,382	7,251	7,375	7,513	7,351
Multi-Racial	655	613	670	630	644	630	443
LEP	698	667	478	409	425	430	333
IDEA	2,106	2,097	1,920	1,593	1,433	1,379	893
Section 504 Plan	325	401	387	425	447	522	526

Table 16. Number of Students in Mathematics Summative Assessment

Group	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 11
All Students	10,434	10,717	10,662	10,520	10,574	10,621	9,893
Female	5,012	5,298	5,204	5,138	5,227	5,118	4,754
Male	5,422	5,419	5,458	5,382	5,347	5,503	5,139
African American	332	357	330	360	337	334	307
AmerIndian/Alaskan	1,149	1,230	1,167	1,174	1,163	1,096	857
Asian	172	179	187	201	163	157	162
Hispanic	869	941	907	884	871	880	756
Pacific Islander	12	9	16	17	20	14	15
White	7,248	7,388	7,386	7,252	7,376	7,511	7,354
Multi-Racial	652	613	669	632	644	629	442
LEP	757	735	531	453	486	471	347
IDEA	2,106	2,084	1,917	1,598	1,429	1,379	893
Section 504 Plan	324	401	389	426	447	521	529

3.2. SUMMARY OF OVERALL STUDENT PERFORMANCE

Tables 17–22 summarize the 2023–2024 summative test results for all students and by subgroup, including the average and standard deviation of scale scores, the percentage of students in each achievement level, and the percentage of proficient students. Figures 1 and 2 present the percentage of proficient students in the past five test administrations for all students (cohort comparisons). Figures 3 and 4 present the average scale scores over the past five years for all students. In Figures 1–4, the 2019–2020 overall student performance is not included because the testing was cancelled due to the COVID-19 pandemic. Appendix B provides the average and standard deviations of scale scores and the percentage of proficient students by subgroup for each test administration across four years.

Table 17. Descriptive Statistics and Percentage of Students in Achievement Levels for Overall and by Subgroup: ELA/L (Grades 3–5)

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
Grade 3								
All Students	10,372	2420.56	88.35	28	25	24	23	47
Female	4,982	2427.03	86.97	25	24	26	24	50
Male	5,390	2414.58	89.19	31	26	22	21	44
African American	315	2378.95	82.89	46	24	20	10	30
AmerIndian/Alaskan	1,151	2355.81	77.68	56	26	12	6	17
Asian	168	2427.03	93.34	29	18	27	26	53
Hispanic	832	2385.86	87.29	44	26	17	13	30
Pacific Islander	12	2345.40	72.94	58	25	17	0	17
White	7,239	2437.20	83.48	21	25	27	27	54
Multi-Racial	655	2414.26	88.24	30	25	26	19	45
LEP	698	2370.13	77.28	50	28	16	6	22
IDEA	2,106	2364.85	84.75	56	22	14	8	22
Section 504 Plan	325	2409.00	82.01	31	30	20	19	39
Grade 4								
All Students	10,650	2460.57	95.37	33	20	23	24	47
Female	5,270	2466.87	93.87	30	20	24	25	49
Male	5,380	2454.40	96.43	35	20	23	22	44
African American	348	2426.45	93.20	49	19	17	14	32
AmerIndian/Alaskan	1,229	2388.45	84.53	65	17	12	6	18
Asian	173	2467.43	91.50	31	20	23	27	50
Hispanic	895	2422.11	91.14	49	20	19	12	31
Pacific Islander	9*							
White	7,383	2479.64	90.38	24	21	26	29	55
Multi-Racial	613	2449.90	89.78	38	20	24	19	43
LEP	667	2387.69	72.58	65	22	10	2	12
IDEA	2,097	2395.09	90.33	63	16	12	8	20
Section 504 Plan	401	2450.98	86.83	38	21	24	17	41
Grade 5								
All Students	10,607	2496.80	97.12	29	21	30	20	50
Female	5,173	2503.91	96.11	26	21	30	23	53
Male	5,434	2490.02	97.60	31	22	29	18	47
African American	325	2457.42	97.82	47	18	24	10	34
AmerIndian/Alaskan	1,167	2423.59	90.44	60	19	17	5	21
Asian	178	2500.05	111.28	28	19	33	20	53
Hispanic	869	2461.22	97.75	41	24	24	11	35
Pacific Islander	16	2448.13	79.99	56	19	19	6	25
White	7,382	2515.39	90.30	21	22	33	25	57
Multi-Racial	670	2484.97	95.43	33	23	28	16	44
LEP	478	2408.47	73.53	64	26	9	1	10
IDEA	1,920	2418.75	90.60	62	20	13	5	18
Section 504 Plan	387	2489.33	97.07	32	22	27	19	45

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

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

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

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
Grade 6								
All Students	10,464	2520.97	94.32	25	27	33	15	48
Female	5,115	2532.51	91.84	21	27	34	18	53
Male	5,349	2509.93	95.35	29	27	32	13	44
African American	353	2487.21	87.25	37	27	30	6	36
AmerIndian/Alaskan	1,164	2449.16	87.88	55	27	15	4	19
Asian	197	2545.93	90.51	15	25	40	20	60
Hispanic	853	2490.09	91.51	36	33	23	9	31
Pacific Islander	16	2500.00	115.80	38	6	44	13	56
White	7,251	2538.09	88.97	18	26	37	19	56
Multi-Racial	630	2510.03	93.91	28	27	33	12	45
LEP	409	2434.80	72.00	62	29	9	0	9
IDEA	1,593	2432.45	86.17	64	23	10	3	13
Section 504 Plan	425	2518.55	91.91	25	28	34	13	47
Grade 7								
All Students	10,511	2547.18	97.05	23	26	36	15	51
Female	5,200	2559.63	94.79	20	25	38	18	56
Male	5,311	2535.00	97.69	27	27	34	12	46
African American	327	2500.89	92.87	41	28	26	5	31
AmerIndian/Alaskan	1,161	2473.78	88.12	52	29	17	2	19
Asian	157	2563.27	99.00	20	22	39	18	57
Hispanic	828	2509.78	101.27	37	27	28	8	36
Pacific Islander	19	2541.50	90.48	26	21	47	5	53
White	7,375	2565.23	91.01	17	25	40	18	58
Multi-Racial	644	2540.72	91.16	26	25	37	12	49
LEP	425	2450.47	79.56	62	27	11	0	11
IDEA	1,433	2451.25	85.80	62	26	10	2	12
Section 504 Plan	447	2539.37	91.70	25	28	36	11	47
Grade 8								
All Students	10,575	2557.86	99.73	24	26	36	14	49
Female	5,095	2575.62	95.80	18	26	40	17	57
Male	5,480	2541.35	100.48	30	27	32	11	43
African American	325	2517.57	97.39	38	29	26	6	33
AmerIndian/Alaskan	1,092	2481.51	90.53	53	27	17	2	19
Asian	155	2578.96	103.05	19	28	34	19	54
Hispanic	846	2518.24	101.19	37	28	28	7	34
Pacific Islander	14	2562.08	97.17	21	29	36	14	50
White	7,513	2575.44	93.38	18	26	40	16	56
Multi-Racial	630	2549.31	102.21	28	25	35	12	47
LEP	430	2461.37	77.53	61	30	9	0	9
IDEA	1,379	2457.77	85.66	66	24	10	1	11
Section 504 Plan	522	2549.87	99.58	27	28	33	13	46

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

Table 19. Descriptive Statistics and Percentage of Students in Achievement Levels for Overall and by Subgroup: ELA/L (Grade 11)

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
Grade 11								
All Students	9,876	2606.39	110.90	17	21	36	27	62
Female	4,746	2624.09	104.13	12	19	38	31	69
Male	5,130	2590.02	114.42	21	23	33	23	56
African American	303	2556.35	105.06	28	30	31	12	42
AmerIndian/Alaskan	857	2530.04	104.33	39	29	24	8	32
Asian	163	2627.92	107.91	10	21	36	32	68
Hispanic	744	2556.49	117.64	31	23	32	14	46
Pacific Islander	15	2558.29	84.83	13	47	33	7	40
White	7,351	2622.39	105.47	12	19	38	31	69
Multi-Racial	443	2600.34	108.03	17	25	33	25	58
LEP	333	2460.88	85.38	67	23	9	1	10
IDEA	893	2481.36	93.96	56	29	13	2	15
Section 504 Plan	526	2606.91	109.35	16	21	37	26	63

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

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

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
Grade 3								
All Students	10,434	2437.87	84.86	24	22	30	23	54
Female	5,012	2432.19	81.27	25	24	30	21	51
Male	5,422	2443.12	87.72	22	21	30	26	56
African American	332	2380.54	83.58	48	23	24	5	30
AmerIndian/Alaskan	1,149	2369.44	78.07	55	24	16	5	21
Asian	172	2435.70	84.91	23	26	26	25	51
Hispanic	869	2396.11	85.62	42	26	21	11	32
Pacific Islander	12	2363.23	70.47	58	33	0	8	8
White	7,248	2457.42	76.77	15	21	34	29	63
Multi-Racial	652	2427.84	83.66	27	26	29	18	47
LEP	757	2379.67	82.31	48	27	18	6	24
IDEA	2,106	2386.38	89.70	47	24	19	10	29
Section 504 Plan	324	2425.46	86.70	29	22	29	19	48
Grade 4								
All Students	10,717	2475.25	87.19	23	30	27	21	48
Female	5,298	2468.95	82.42	24	32	28	17	44
Male	5,419	2481.40	91.20	21	28	27	24	51
African American	357	2427.98	87.09	40	36	17	6	24
AmerIndian/Alaskan	1,230	2399.77	76.19	58	28	10	3	14
Asian	179	2475.19	88.79	23	29	26	22	48
Hispanic	941	2427.86	82.09	42	33	18	7	25
Pacific Islander	9*							
White	7,388	2497.35	79.13	13	29	31	27	58
Multi-Racial	613	2461.66	79.84	26	34	28	13	41
LEP	735	2404.94	73.68	53	34	10	3	13
IDEA	2,084	2416.17	89.58	49	29	14	8	22
Section 504 Plan	401	2466.53	88.25	24	30	30	16	46
Grade 5								
All Students	10,662	2500.71	92.74	31	28	20	21	41
Female	5,204	2493.96	88.94	33	30	19	18	37
Male	5,458	2507.15	95.78	29	26	21	24	45
African American	330	2455.85	92.34	50	26	15	9	24
AmerIndian/Alaskan	1,167	2419.63	82.17	68	21	8	3	11
Asian	187	2508.54	100.18	34	21	20	25	45
Hispanic	907	2460.77	89.33	47	30	14	10	24
Pacific Islander	16	2413.88	85.69	75	13	6	6	13
White	7,386	2521.77	84.62	21	29	23	27	50
Multi-Racial	669	2485.84	91.40	36	31	16	16	33
LEP	531	2419.46	72.07	68	25	6	1	7
IDEA	1,917	2427.64	90.69	65	20	8	7	14
Section 504 Plan	389	2497.18	89.78	30	29	22	18	41

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

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

Table 21. Descriptive Statistics and Percentage of Students in Achievement Levels for Overall and by Subgroup: Mathematics (Grades 6–8)

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
Grade 6								
All Students	10,520	2521.68	103.65	28	30	22	20	42
Female	5,138	2520.64	99.46	28	32	22	18	41
Male	5,382	2522.68	107.50	29	28	23	21	43
African American	360	2465.95	107.15	46	34	15	5	19
AmerIndian/Alaskan	1,174	2426.57	98.58	65	24	8	3	10
Asian	201	2542.84	103.02	26	25	20	28	49
Hispanic	884	2478.85	100.60	43	33	16	8	24
Pacific Islander	17	2479.86	128.09	41	24	24	12	35
White	7,252	2545.76	92.60	19	30	26	25	51
Multi-Racial	632	2508.19	99.17	32	33	20	15	35
LEP	453	2419.36	91.84	69	26	4	1	6
IDEA	1,598	2424.82	106.16	67	21	8	4	12
Section 504 Plan	426	2525.40	98.71	26	34	20	20	40
Grade 7								
All Students	10,574	2536.57	107.98	31	28	22	19	41
Female	5,227	2534.05	105.56	31	28	23	18	41
Male	5,347	2539.03	110.25	30	28	22	21	42
African American	337	2465.07	103.56	56	26	12	6	18
AmerIndian/Alaskan	1,163	2439.85	88.89	69	22	6	2	8
Asian	163	2543.78	111.84	31	28	21	19	40
Hispanic	871	2478.31	108.07	54	25	11	10	21
Pacific Islander	20	2484.83	115.22	50	10	35	5	40
White	7,376	2562.72	97.57	20	29	26	24	50
Multi-Racial	644	2527.68	105.19	33	30	22	15	37
LEP	486	2420.92	85.27	77	19	4	1	5
IDEA	1,429	2434.51	95.32	72	18	7	3	10
Section 504 Plan	447	2533.58	100.42	29	34	19	18	36
Grade 8								
All Students	10,621	2551.95	117.12	33	26	21	20	40
Female	5,118	2553.36	112.27	32	27	22	18	41
Male	5,503	2550.64	121.45	35	25	19	21	40
African American	334	2491.83	110.25	55	27	10	8	18
AmerIndian/Alaskan	1,096	2449.12	99.67	74	17	6	3	9
Asian	157	2578.38	140.39	31	22	22	26	48
Hispanic	880	2492.61	110.66	55	26	12	7	20
Pacific Islander	14	2510.12	120.06	43	29	14	14	29
White	7,511	2577.66	107.51	24	28	25	24	49
Multi-Racial	629	2533.41	117.72	38	30	17	15	32
LEP	471	2435.47	92.14	78	17	4	1	5
IDEA	1,379	2436.36	104.14	76	16	6	3	9
Section 504 Plan	521	2548.75	112.19	37	26	19	18	36

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

Table 22. Descriptive Statistics and Percentage of Students in Achievement Levels for Overall and by Subgroup: Mathematics (Grade 11)

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
Grade 11								
All Students	9,893	2580.11	116.91	38	27	22	13	35
Female	4,754	2579.77	110.31	37	29	23	11	34
Male	5,139	2580.43	122.70	39	26	21	14	35
African American	307	2519.31	106.63	61	23	12	4	16
AmerIndian/Alaskan	857	2479.29	93.01	77	17	6	1	7
Asian	162	2610.78	115.70	27	32	22	19	41
Hispanic	756	2525.64	109.94	58	22	15	4	19
Pacific Islander	15	2526.14	111.23	67	27	0	7	7
White	7,354	2600.48	111.70	30	29	25	15	41
Multi-Racial	442	2562.67	113.13	41	32	18	9	27
LEP	347	2451.20	79.50	89	9	1	1	2
IDEA	893	2452.42	92.49	86	11	3	1	4
Section 504 Plan	529	2581.69	121.14	41	24	19	16	35

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

Figure 1. ELA/L Percent Proficient Across Years

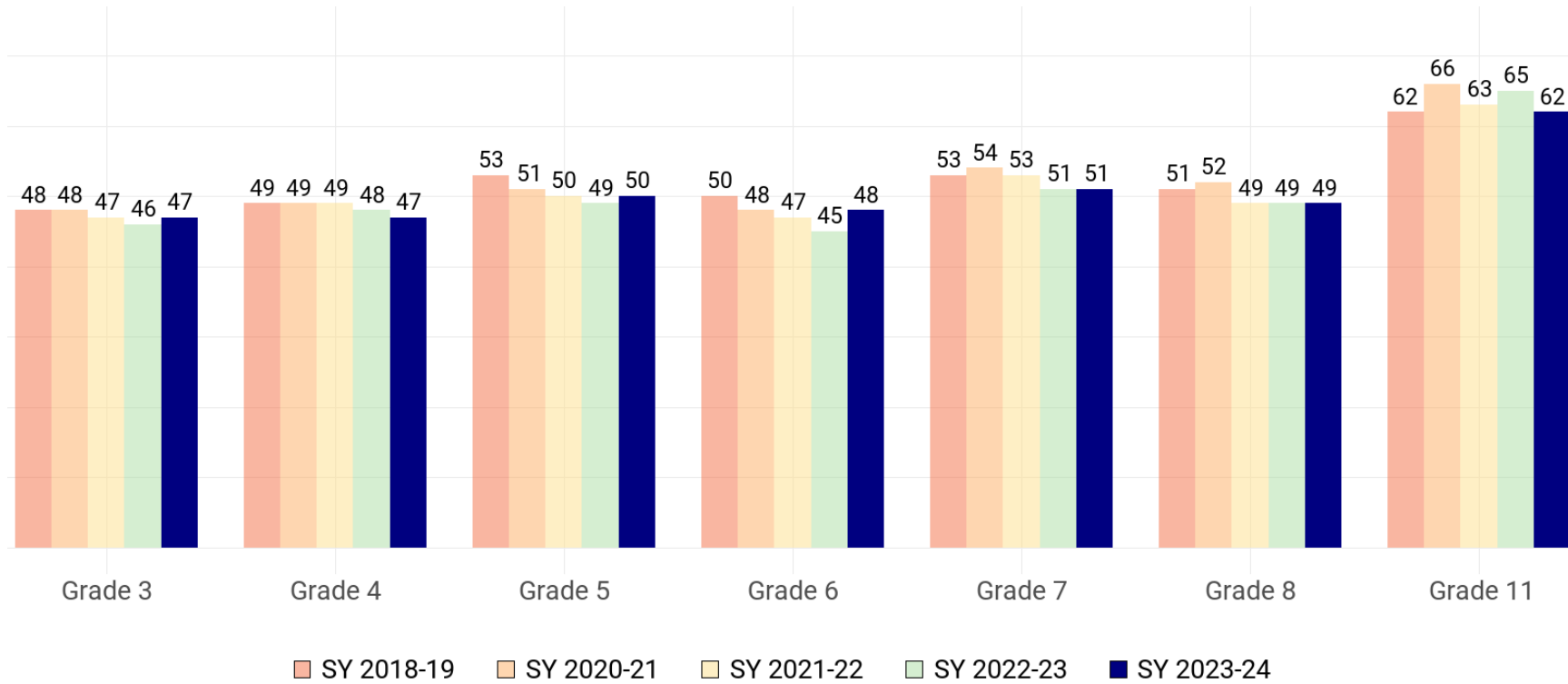


Figure 2. Mathematics Percent Proficient Across Years

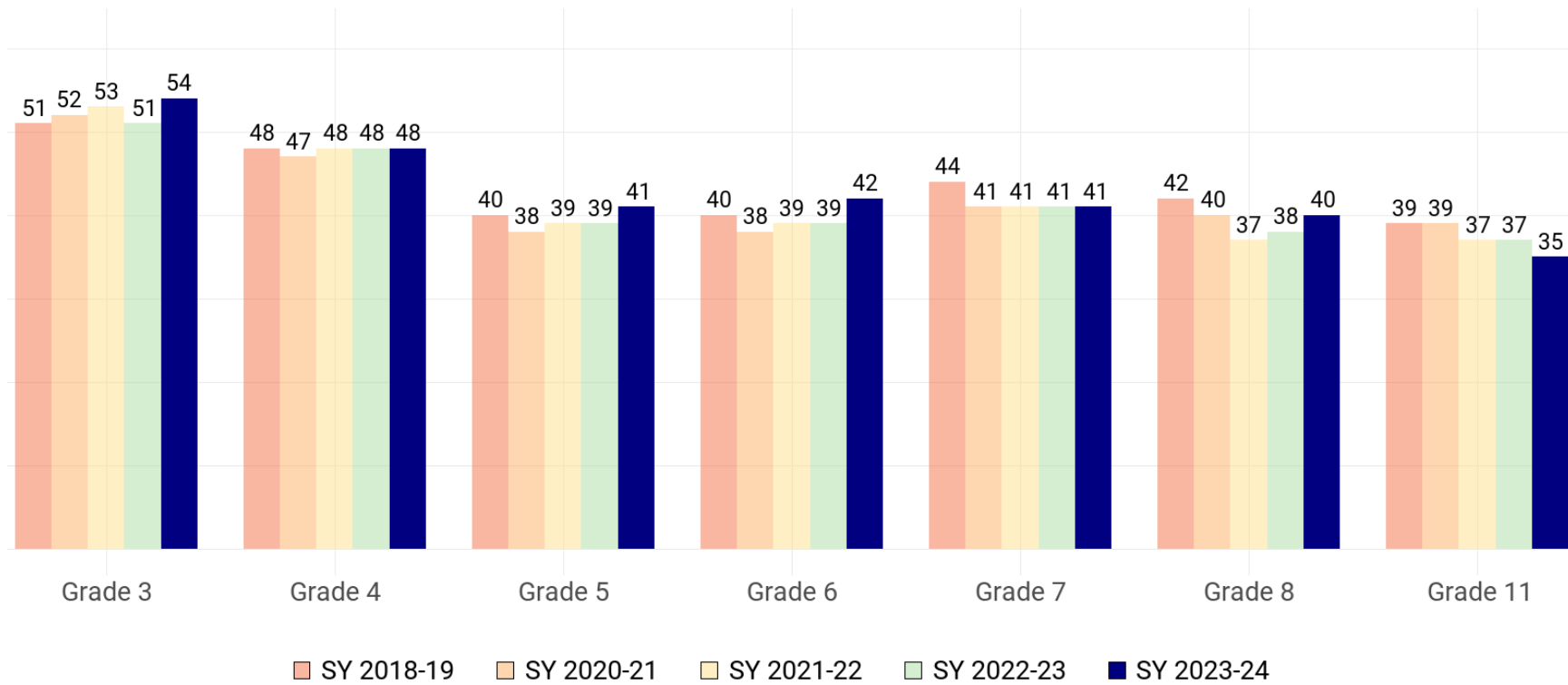


Figure 3. ELA/L Average Scale Score Across Years

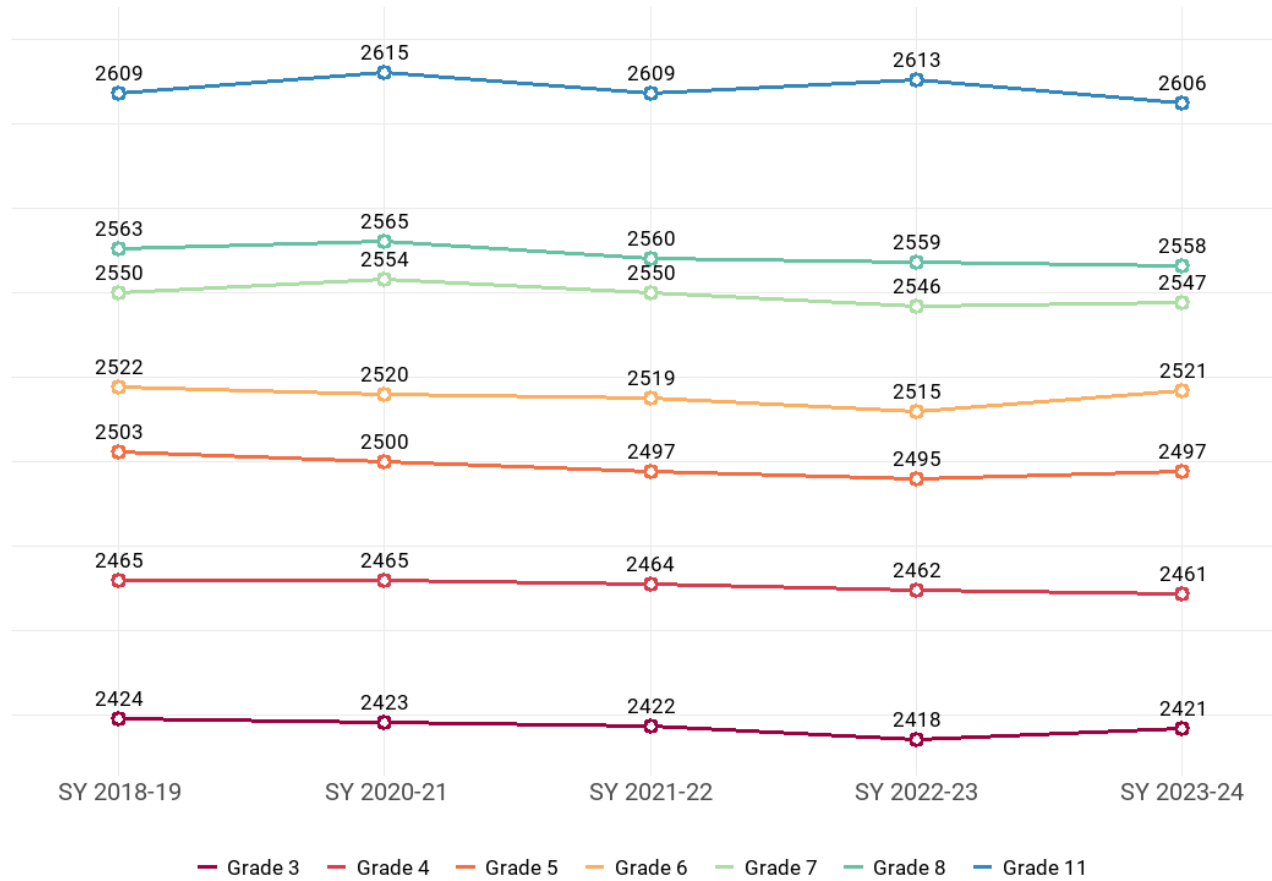
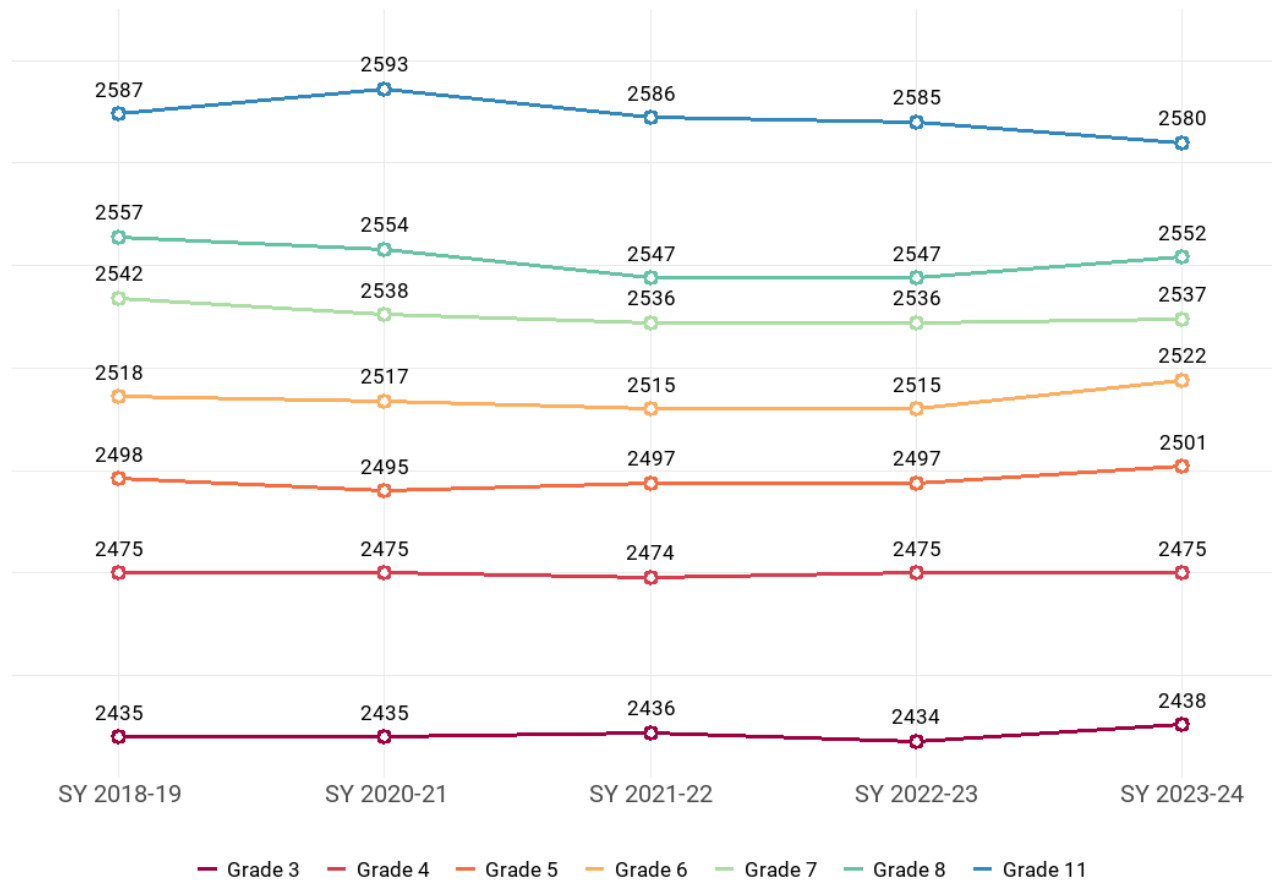


Figure 4. Mathematics Average Scale Score Across Years



Because the precision of scores in each claim is not sufficient to report scores, given the small number of items, the scores on each claim are reported using one of the three performance categories, taking into account the standard error of measurement (SEM) of the claim score: (1) Below Standard, (2) At/Near Standard, or (3) Above Standard. (Refer to Section 6.5, Rules for Calculating Strengths and Weaknesses for Claim Scores, for details on how the performance category is determined.) Tables 23 and 24 present the distribution of performance categories for each claim. The number of claims is four in ELA/L and three in mathematics, combining claims 2 and 4.

Table 23. ELA/L 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	31	27	17	28
	At/Near	47	54	66	52
	Above	22	19	17	20
4	Below	30	28	19	28
	At/Near	46	56	62	53
	Above	24	16	19	19
5	Below	27	27	21	26
	At/Near	48	53	63	50
	Above	25	20	16	24
6	Below	31	27	21	22
	At/Near	49	55	67	57
	Above	20	18	13	21
7	Below	27	23	19	23
	At/Near	50	53	68	55
	Above	23	23	14	22
8	Below	29	28	20	22
	At/Near	48	53	66	55
	Above	23	19	15	23
11	Below	21	17	16	17
	At/Near	45	51	62	52
	Above	34	32	22	32

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

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

Grade	Performance Category	Claim 1: Concepts and Procedures	Claims 2 and 4: Problem Solving and Modeling and Data Analysis	Claim 3: Communicating Reasoning
3	Below	30	24	25
	At/Near	33	46	48
	Above	38	30	28
4	Below	34	30	30
	At/Near	33	46	44
	Above	33	23	25
5	Below	39	32	33
	At/Near	34	49	50
	Above	28	19	17
6	Below	39	32	30
	At/Near	36	49	51
	Above	25	19	20
7	Below	40	31	31
	At/Near	33	47	51
	Above	27	21	18
8	Below	39	28	29
	At/Near	37	50	53
	Above	24	21	18
11	Below	48	30	36
	At/Near	31	50	49
	Above	21	20	15

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

3.3. DISTRIBUTION OF STUDENT ABILITY AND ITEM DIFFICULTY

Figures 5–10 display the empirical distribution of the South Dakota student scale scores in the 2023–2024 administration and the distribution of the administered summative item difficulty parameters for each grade for overall and by claim. Overall, the student ability distribution is shifted to the left in all grades and subjects, a pattern more pronounced in the mathematics upper grades, indicating that the pool includes more difficult items than the ability level of students in the tested population. The pool includes difficult items to measure high-performing students accurately but needs additional easy items to measure low-performing students better. At the claim level, the student ability distribution is shifted to the left particularly in claims 1 (Reading) and 4 (Research) across all grades, and in claim 3 (Listening) for grade 3 in ELA/L. In mathematics, the student ability distribution is shifted to the left for all claims except for claim 1 in lower grades. The Smarter Balanced Assessment Consortium plans to add additional easy items to the pool and augment the pool in proportion to the test blueprint constraints (e.g., content, Depth of Knowledge [DOK], item type, item difficulties) to measure low-performing students better.

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

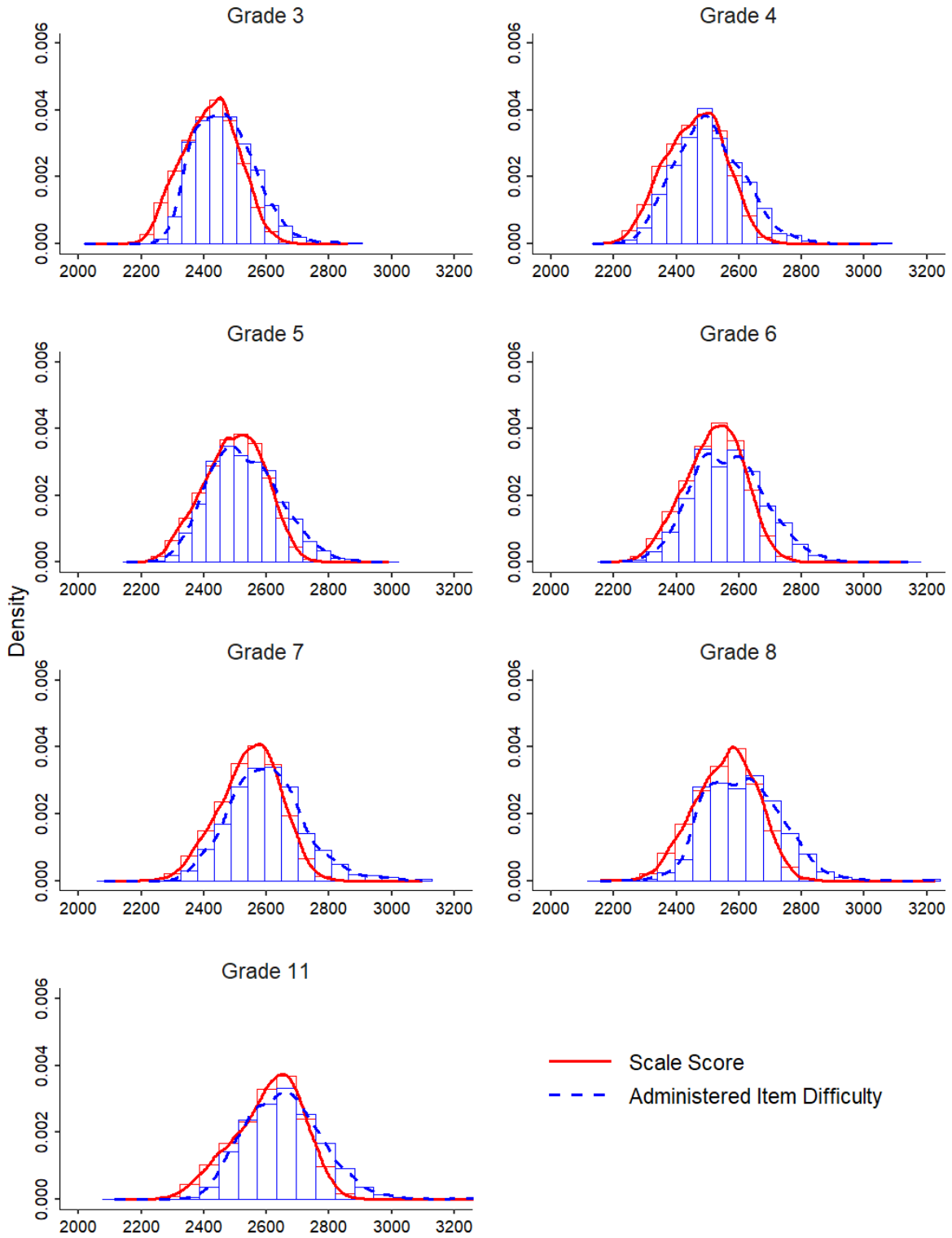


Figure 6. Student Ability—Item Difficulty Distribution by Claim: ELA/L (Grades 3–5)

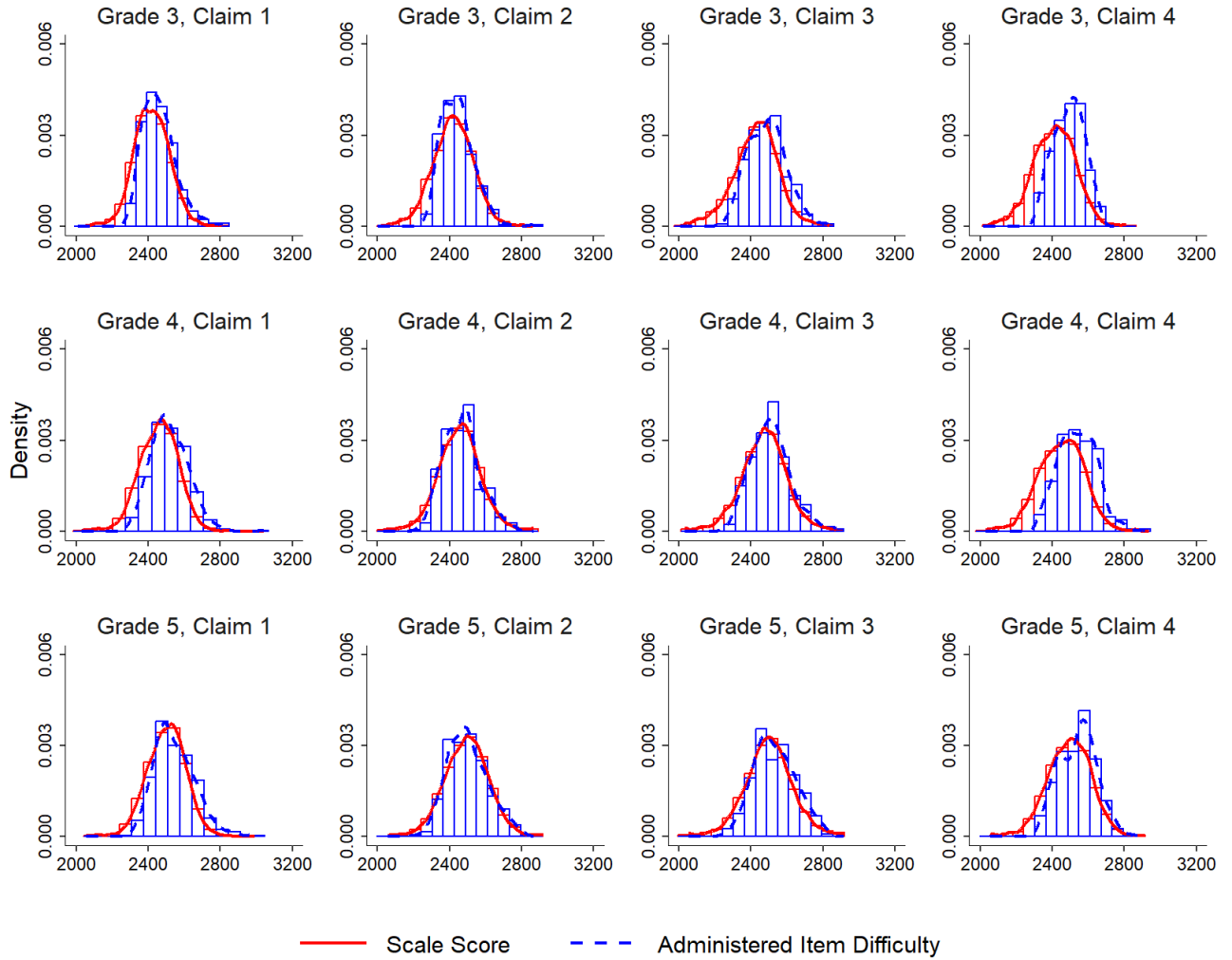


Figure 7. Student Ability—Item Difficulty Distribution by Claim: ELA/L (Grades 6–8, 11)

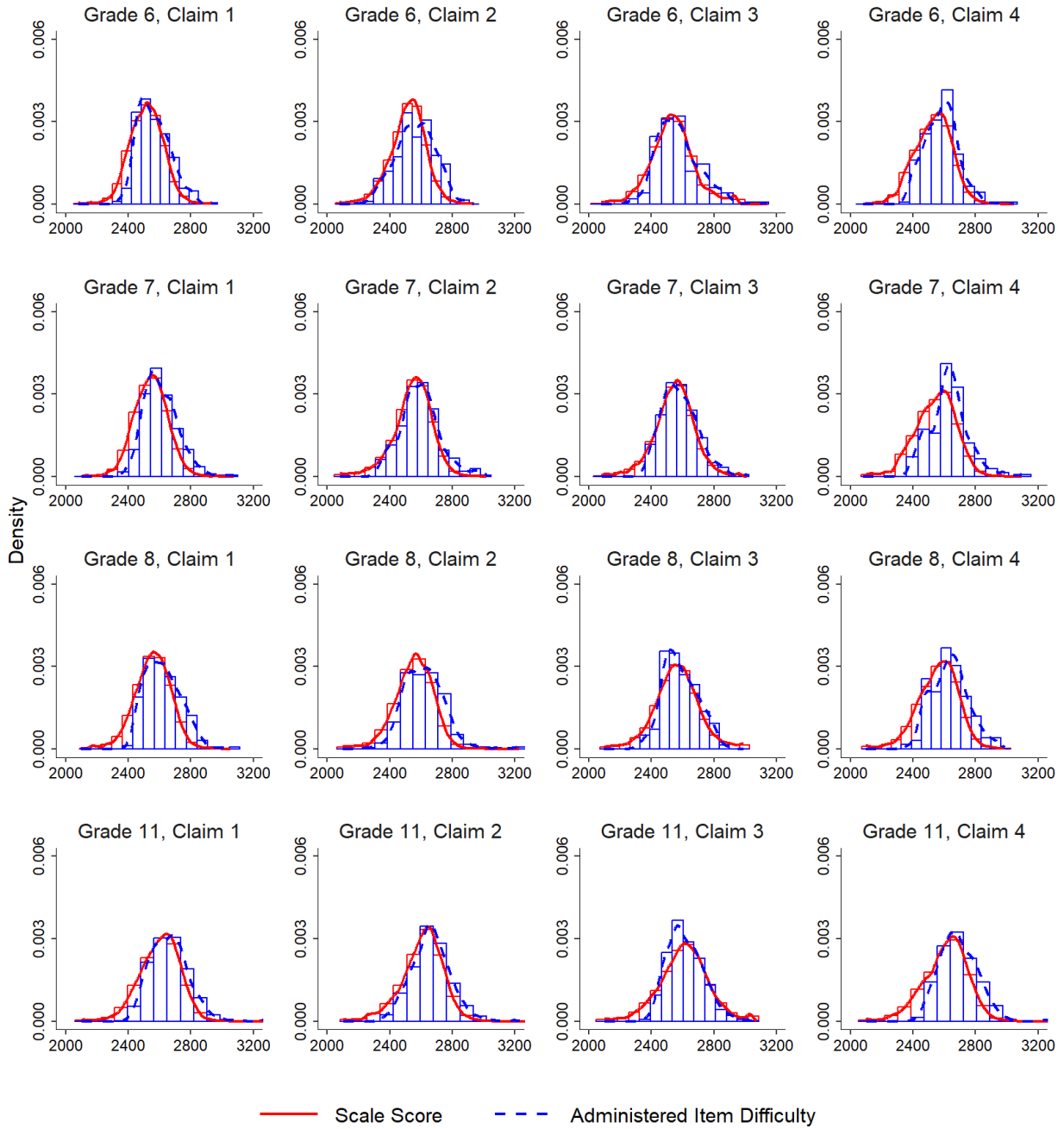


Figure 8. Student Ability—Item Difficulty Distribution for Mathematics

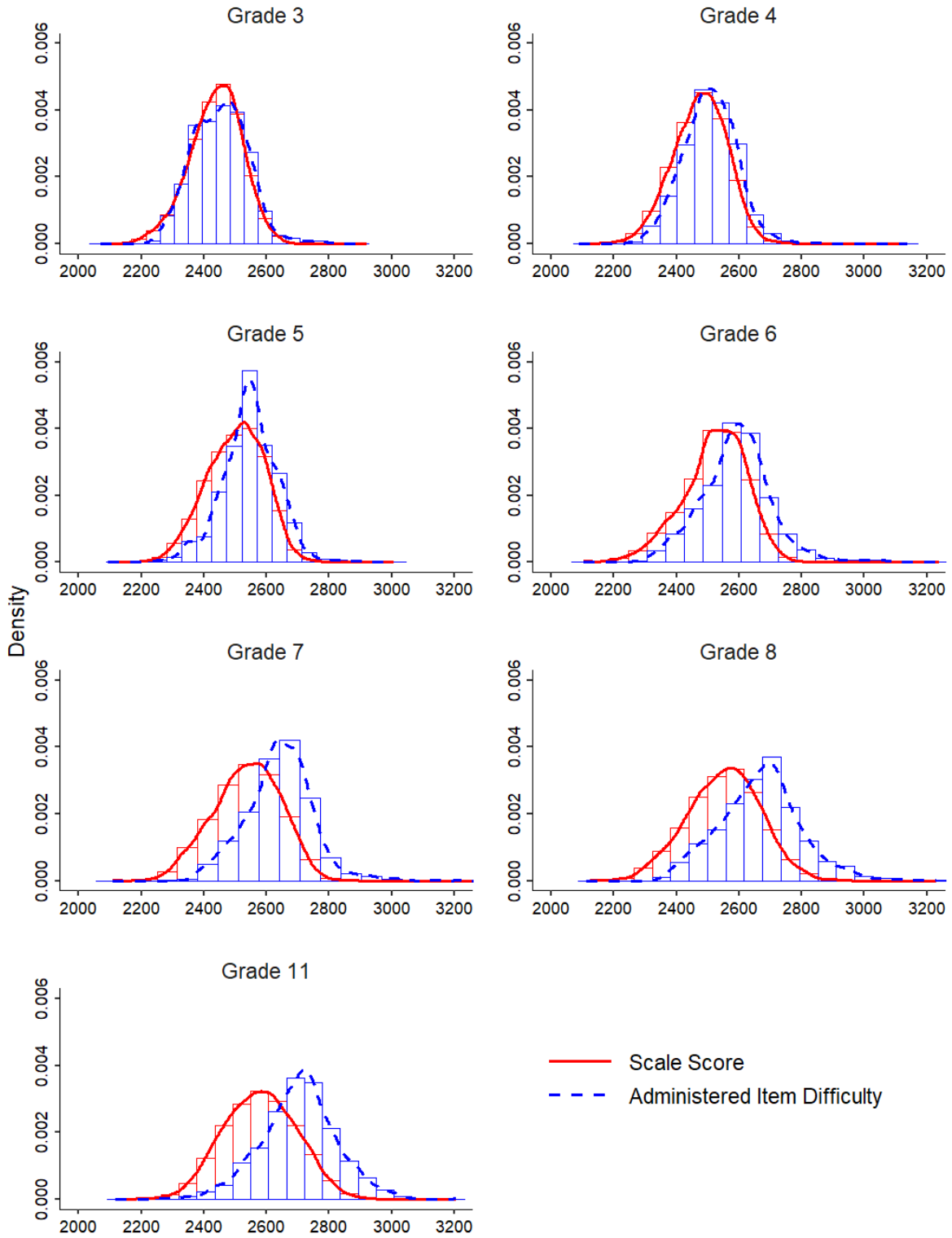


Figure 9. Student Ability—Item Difficulty Distribution by Claim: Mathematics (Grades 3–5)

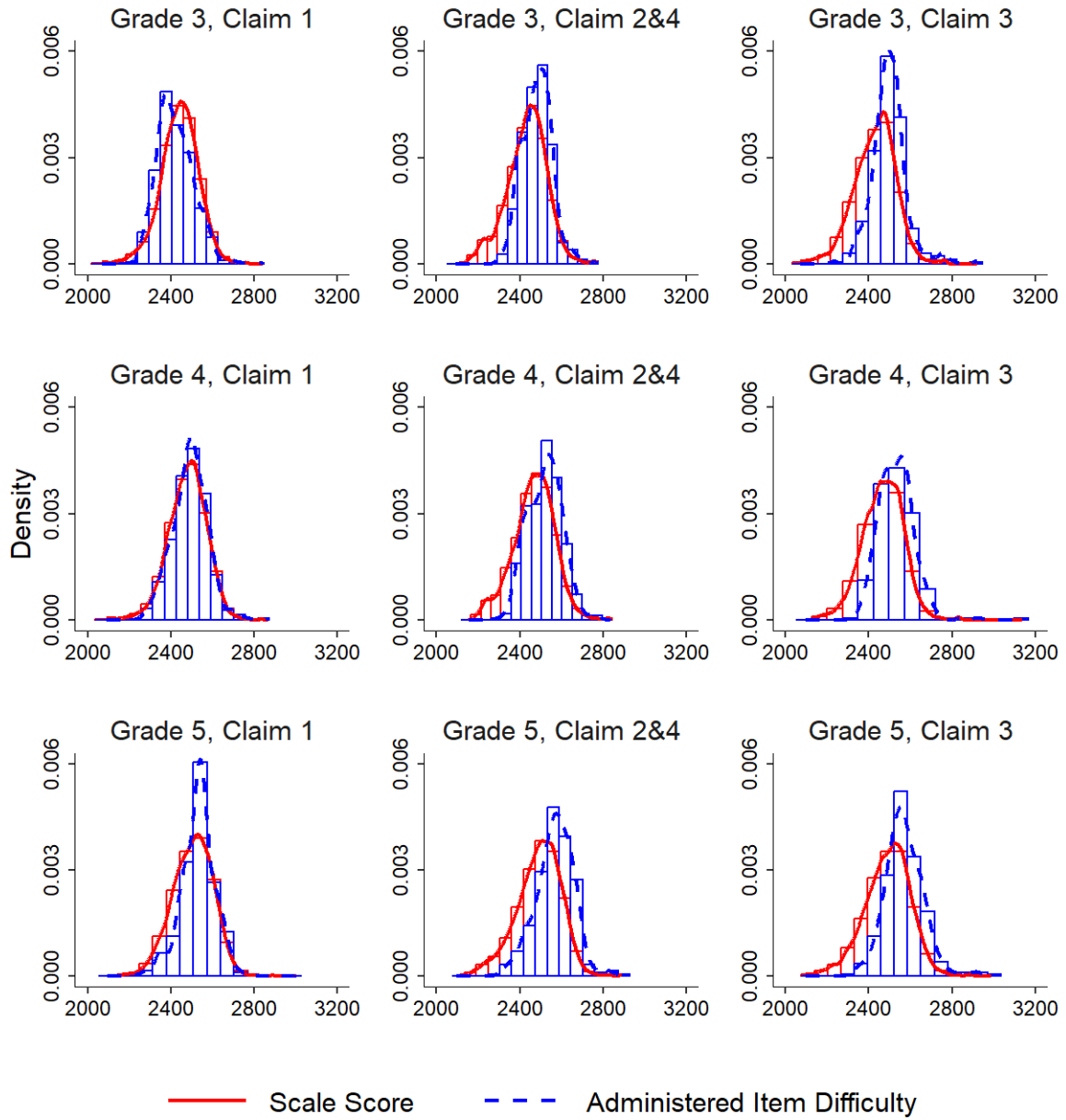
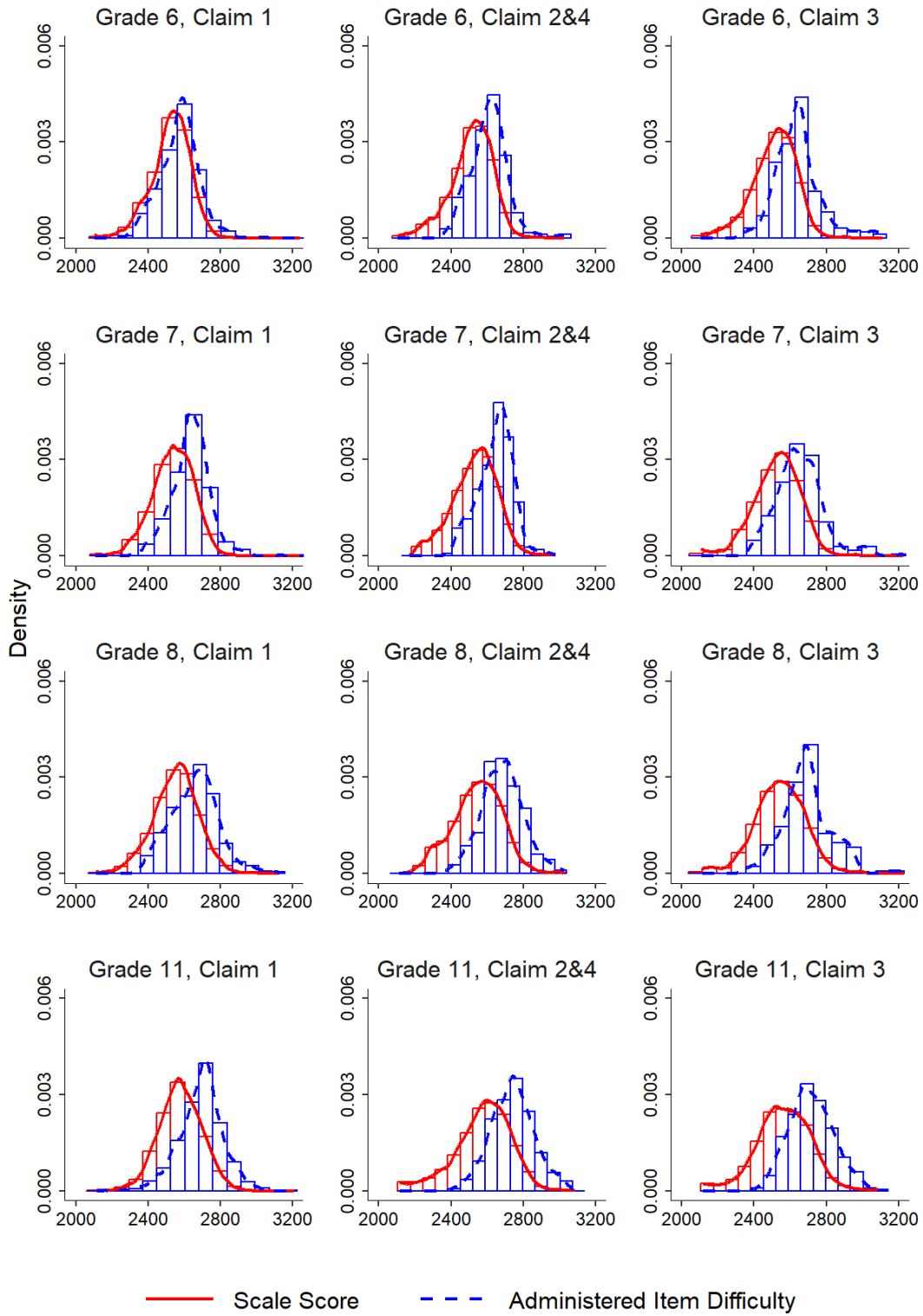


Figure 10. Student Ability—Item Difficulty Distribution by Claim: Mathematics (Grades 6–8, 11)



4. VALIDITY

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

The 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 the 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 South Dakota 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 that adapt to his or her ability. For PTs, each student is administered a fixed-form test. The content covered in all PT forms is the same. The test blueprint constraints for CAT and PT can be found at: <https://doe.sd.gov/Assessment/SD-assessments.aspx>.

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

Tables 25 and 26 present the percentages of tests aligned with the ELA/L test blueprint constraints for items in claims, targets, DOK, and the number of passage requirements. All tests met the blueprint requirements except some targets in claim 1 (Reading), which administered a few items more or less than the item requirement. The violations in the Claim 1 reading targets appeared in all grades due to the uneven distribution of items across targets and DOKs within and across the passages.

Tables 27–29 provide the percentages of tests aligned with the mathematics CAT test blueprint constraints for items in claims, DOK, and target constraints. In mathematics, all tests met all blueprint requirements, except for a few tests that had blueprint violations due to the application of pool filters limiting the item pool and one test in grade 3 that failed to meet the Claim 2/4_DOK3+ requirement because of the uneven

distribution of combinations of DOK and targets in claims 2 and 4. Pool filters, such as using an alternative language like braille or Spanish or only items with illustration or language glossaries, can result in an accommodated CAT item pool that is too limited to meet all test blueprint requirements, especially if multiple pool filters are employed on the same test.

Table 25. Percentage of ELA/L CAT Delivered Tests Meeting Blueprint Requirements for Each Claim and the Number of Passages Administered (Grades 3–5)

Claim	Content Category/Target	Required Items/Passages	% BP Match		
			Grade 3	Grade 4	Grade 5
1	Literary Text	7–8	100.00	100.00	100.00
	Target 2: Central Ideas	1–2	99.90	99.99	100.00
	Target 4: Reasoning and Evaluation	1–2	100.00	100.00	99.86
	Targets 1, 3, 5, 6, and 7	3–6	99.99	100.00	99.96
	Long Literary Text Passage	≥ 1	100.00	100.00	100.00
	Short Literary Text Passage	≤ 2	100.00	100.00	100.00
	Informational Text	7–8	100.00	100.00	100.00
	Target 9: Central Ideas	1–2	99.28	100.00	100.00
	Target 11: Reasoning and Evaluation	1–2	100.00	100.00	100.00
	Targets 8, 10, 12, 13, and 14	3–6	99.99	100.00	100.00
	Long Informational Text Passage	≥ 1	100.00	100.00	100.00
	Short Informational Text Passage	≤ 2	100.00	100.00	100.00
	DOK 2	≥ 7	100.00	100.00	100.00
	DOK 3 or 4	≥ 2	100.00	100.00	100.00
2	Writing	6	100.00	100.00	100.00
	Target 1, 3, or 6: Organization/Purpose	1	100.00	100.00	100.00
	Target 1, 3, or 6: Evidence/Elaboration	1	100.00	100.00	100.00
	Target 8: Language and Vocabulary Use	1	100.00	100.00	100.00
	Target 9: Edit/Clarify	3	100.00	100.00	100.00
	DOK 2 or Higher	≥ 2	100.00	100.00	100.00
3	Listening	8–9	100.00	100.00	100.00
	Target 4: Listen/Interpret	8–9	100.00	100.00	100.00
	DOK 2 or Higher	≥ 3	100.00	100.00	100.00
	Listening Passage	3–4	100.00	100.00	100.00
4	Research	8	100.00	100.00	100.00
	Target 2: Interpret and Integrate Information	2–3	100.00	100.00	100.00
	Target 3: Analyze Information/Sources	2–3	100.00	100.00	100.00
	Target 4: Use Evidence	2–3	100.00	100.00	100.00

Table 26. Percentage of ELA/L CAT Delivered Tests Meeting Blueprint Requirements for Each Claim and the Number of Passages Administered (Grades 6–8, 11)

Claim	Content Category/Target	Required Items/ Passages in Grades 6–8	Required Items/ Passages in Grade 11	% BP Match			
				Grade 6	Grade 7	Grade 8	Grade 11
1	Literary Text	4–7	4	100.00	100.00	100.00	100.00
	Target 2: Central Ideas	1	1	99.39	99.99	99.41	99.15
	Target 4: Reasoning and Evaluation	1	1	98.69	97.40	100.00	99.10
	Targets 1, 3, 5, 6, and 7	2–5	2	100.00	100.00	100.00	98.74
	Target 2 or 4 Short Text	0–1	0–1	100.00	100.00	100.00	100.00
	Long Literary Text Passage	≥ 1	≥ 1	100.00	100.00	100.00	100.00
	Informational Text	10–12*	11–12	100.00	100.00	100.00	100.00
	Target 9 and Target 11	2–5	2–4	99.79	98.78	100.00	98.72
	Targets 8, 10, 12, 13, and 14	7–10	7–10	99.50	98.78	100.00	99.75
	Target 9 or 11 Short Text	0–1	0–1	100.00	100.00	100.00	100.00
	Long Informational Text Passage	≥ 1	≥ 1	100.00	100.00	100.00	100.00
	Short Informational Text Passage	≤ 2	≤ 2	100.00	100.00	100.00	100.00
DOK 1	≤ 5	≤ 4	100.00	100.00	100.00	100.00	
DOK 3 or 4	≥ 2	≥ 3	100.00	100.00	100.00	100.00	
2	Writing	6	6	100.00	100.00	100.00	100.00
	Target 1, 3, and 6 (Organization/Purpose)	1	1	100.00	100.00	100.00	100.00
	Target 1, 3, and 6 (Evidence/Elaboration)	1	1	100.00	100.00	100.00	100.00
	Target 8: Language and Vocabulary Use	1	1	100.00	100.00	100.00	100.00
	Target 9: Edit/Clarify	3	3	100.00	100.00	100.00	100.00
	DOK 2	≥ 2	≥ 2	100.00	100.00	100.00	100.00
	DOK 3 or 4	1	1	100.00	100.00	100.00	100.00
	Brief Write	1	1	100.00	100.00	100.00	100.00
3	Listening	8–9	8–9	100.00	100.00	100.00	100.00
	Target 4: Listen/Interpret	8–9	8–9	100.00	100.00	100.00	100.00
	DOK 2 or Higher	≥ 3	≥ 4	100.00	100.00	100.00	100.00
	Listening Passage	3–4	3–4	100.00	100.00	100.00	100.00
4	Research	8	8	100.00	100.00	100.00	100.00
	Target 2: Analyze/Integrate Information	2–3	2–3	100.00	100.00	100.00	100.00
	Target 3: Evaluate Information/Sources	2–3	2–3	100.00	100.00	100.00	100.00
	Target 4: Use Evidence	2–3	2–3	100.00	100.00	100.00	100.00

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

Table 27. Mathematics Percentage of CAT Delivered Tests Meeting Blueprint Requirements for Claims and Targets (Grades 3–5)

Claim	Content Domain	Grade 3		Grade 4		Grade 5	
		Required Items	% BP Match	Required Items	% BP Match	Required Items	% BP Match
1	Overall	17–20	100.00	17–20	100.00	17–20	100.00
	DOK 2 or higher	≥ 7	100.00	≥ 7	100.00	≥ 7	100.00
	<i>Priority Cluster</i>	13–15	100.00				
	Targets B, C, G, I	5–6	100.00				
	Targets D, F	5–6	100.00				
	Target A	2–3	100.00				
	<i>Supporting Cluster</i>	4–5	100.00				
	Targets E, J, K	3–4	100.00				
	Target H	1	100.00				
	<i>Priority Cluster</i>			13–15	100.00		
	Targets A, E, F			8–9	100.00		
	Target G			2–3	100.00		
	Target D			1–2	100.00		
	Target H			1	100.00		
	<i>Supporting Cluster</i>			4–5	100.00		
	Targets I, K			2–3	100.00		
	Targets B, C, J			1	100.00		
	Target L			1	100.00		
	<i>Priority Cluster</i>					13–15	100.00
Targets E, I					5–6	100.00	
Target F					4–5	100.00	
Targets C, D					3–4	100.00	
<i>Supporting Cluster</i>					4–5	100.00	
Targets J, K					2–3	100.00	
Targets A, B, G, H					2	100.00	
2 and 4	Overall	6	100.00	6	100.00	6	100.00
	DOK 3 or higher	≥ 2	99.87	≥ 2	100.00	≥ 2	100.00
	2. Target A	2	100.00	2	100.00	2	100.00
	2. Targets B, C, D	1	100.00	1	100.00	1	100.00
	4. Targets A, D	1	100.00	1	100.00	1	100.00
	4. Targets B, E	1	100.00	1	100.00	1	100.00
4. Targets C, F	1	100.00	1	100.00	1	100.00	
3	Overall	8	100.00	8	100.00	8	100.00
	DOK 3 or higher	≥ 2	100.00	≥ 2	100.00	≥ 2	100.00
	Targets A, D	3	100.00	3	100.00	3	100.00
	Targets B, E	3	100.00	3	100.00	3	100.00
	Targets C, F	2	100.00	2	100.00	2	100.00

Table 28. Mathematics Percentage of CAT Delivered Tests Meeting Blueprint Requirements for Claims and Targets (Grades 6–8)

Claim	Content Domain	Grade 6		Grade 7		Grade 8	
		Required Items	% BP Match	Required Items	% BP Match	Required Items	% BP Match
1	Overall	16–20	100.00	16–20	100.00	16–20	100.00
	DOK 2 or higher	≥ 7	100.00	≥ 7	100.00	≥ 7	100.00
	<i>Priority Cluster</i>	12–15	100.00				
	Targets E, F	5–6	99.96				
	Target A	3–4	100.00				
	Targets B, G	2	99.87				
	Target D	2	99.52				
	<i>Supporting Cluster</i>	4–5	100.00				
	Targets C, H, I, J	4–5	100.00				
	<i>Priority Cluster</i>			12–15	99.83		
	Targets A, D			8–9	100.00		
	Targets B, C			5–6	99.83		
	<i>Supporting Cluster</i>			4–5	99.99		
	Targets E, F			2–3	99.99		
	Targets G, H, I			1–2	100.00		
<i>Priority Cluster</i>					12–15	99.94	
Targets C, D					5–6	99.62	
Targets B, E, G					5–6	99.63	
Targets F, H					2–3	99.96	
<i>Supporting Cluster</i>					4–5	99.94	
Targets A, I, J					4–5	99.94	
2 and 4	Overall	6	100.00	6	100.00	6	100.00
	DOK 3 or higher	≥ 2	100.00	≥ 2	99.99	≥ 2	99.99
	2. Target A	2	100.00	2	100.00	2	100.00
	2. Targets B, C, D	1	100.00	1	100.00	1	100.00
	4. Targets A, D	1	100.00	1	100.00	1	100.00
	4. Targets B, E	1	100.00	1	100.00	1	100.00
4. Targets C, F	1	100.00	1	100.00	1	100.00	
3	Overall	8	100.00	8	100.00	8	100.00
	DOK 3 or higher	≥ 2	100.00	≥ 2	100.00	≥ 2	100.00
	Targets A, D	3	100.00	3	100.00	3	100.00
	Targets B, E	3	100.00	3	100.00	3	100.00
	Targets C, F, G	2	100.00	2	100.00	2	100.00

Table 29. Mathematics Percentage of CAT Delivered Tests Meeting Blueprint Requirements for Claims and Targets (Grade 11)

Claim	Content Domain	Grade 11	
		Required Items	% BP Match
1	Overall	19–22	100.00
	DOK 2 or higher	≥ 7	100.00
	<i>Priority Cluster</i>	14–16	99.99
	Targets D, E	2	99.93
	Target F	1	100.00
	Targets G, H, I	4–5	99.91
	Target J	2	100.00
	Target K	2	99.99
	Targets L, M, N	3–4	100.00
	<i>Supporting Cluster</i>	5–6	100.00
	Target O	2	99.99
	Target P	1–2	100.00
	Targets A, B	1	99.98
	Target C	1	100.00
2 and 4	Overall	6	100.00
	DOK 3 or higher	≥ 2	100.00
	2. Target A	2	100.00
	2. Targets B, C, D	1	100.00
	4. Targets A, D	1	100.00
	4. Targets B, E	1	100.00
3	Overall	8	100.00
	DOK 3 or higher	≥ 2	100.00
	Targets A, D	3	100.00
	Targets B, E	3	100.00
	Targets C, F, G	2	100.00

Table 30 summarizes the target coverage, the average, and the range of the numbers of unique targets administered in each delivered test by claim. Because 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 30. Average and Range of the Number of Unique Targets Assessed within Each Claim Across All Delivered CAT Tests

Grade	Total Targets in BP				Average				Range (Minimum–Maximum)			
	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4
ELA/L												
3	14	5	1	3	11.5	4.0	1.0	3.0	8–14	4–4	1–1	3–3
4	14	5	1	3	11.4	4.0	1.0	3.0	8–14	4–4	1–1	3–3
5	14	5	1	3	11.8	4.0	1.0	3.0	9–14	4–4	1–1	3–3
6	14	5	1	3	10.5	4.0	1.0	3.0	9–11	4–4	1–1	3–3
7	14	5	1	3	10.8	4.0	1.0	3.0	9–11	4–4	1–1	3–3
8	14	5	1	3	10.3	4.0	1.0	3.0	8–11	4–4	1–1	3–3
11	14	5	1	3	10.6	4.0	1.0	3.0	8–11	4–4	1–1	3–3
Mathematics												
3	11	4	6	6	11.0	2.0	5.4	3.0	10–11	2–2	4–6	3–3
4	12	4	6	6	10.0	2.0	5.3	3.0	10–10	2–2	3–6	3–3
5	11	4	6	6	9.0	2.0	5.1	3.0	9–9	2–2	3–6	3–3
6	10	4	7	6	10.0	2.0	5.0	3.0	8–10	2–2	3–6	3–3
7	9	4	7	6	8.0	2.0	4.9	3.0	7–8	2–2	3–6	3–3
8	10	4	7	6	10.0	2.0	5.2	3.0	9–10	2–2	3–6	3–3
11	16	4	7	6	14.9	2.0	5.2	3.0	14–15	2–2	3–6	3–3

An adaptive testing algorithm constructs a test form unique to each student, targeting the student’s level of ability and meeting the test blueprints. Consequently, the test forms will not be statistically parallel (e.g., equal test difficulty) across individual students, but test scores from the individual tests are comparable since all test forms measure the same content, albeit with a different set of test items. Although each form is unique with respect to its items, all forms align with the same curricular expectations outlined in the test blueprints.

4.2. EVIDENCE ON INTERNAL STRUCTURE

The measurement model used in the Smarter Balanced assessments assumes a single underlying latent trait in student ability estimates, which supports the reporting of a single total ability score. During the test construction phase, the test blueprint was designed to cover multiple distinct claims under each subject. The item selection algorithm prioritizes blueprint matching to ensure each test contains an appropriate combination of items from each claim. Assessing the relationship between these different claim scores is a measure of internal validity according to the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014). The presence of high correlations among claim scores is evidence that the Smarter Balanced assessment measures a single underlying ability, and the claim scores are related to each other.

Tables 31 and 32 present the correlations among claim scores, both observed (below diagonal) and corrected for attenuation (above diagonal). The correction for attenuation indicates what the correlation would be if claim scores could be measured with perfect reliability, corrected (adjusted) for measurement error estimates.

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

is the observed correlation between x and y , r_{xx} is the reliability coefficient for x , and r_{yy} is the reliability coefficient for y .

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

Because the reliabilities for claim scores are low, the performance of each claim score is reported in three performance categories. The distribution of performance categories for each claim is provided in Tables 23 and 24, Section 3.2. Scale scores are not reported for claims.

Table 31. Correlations Among Claim Scores for ELA/L

Grade	Claim	Observed & Disattenuated Correlation			
		Claim 1	Claim 2	Claim 3	Claim 4
3	Claim 1: Reading		0.86	0.91	0.89
	Claim 2: Writing	0.64		0.85	0.86
	Claim 3: Listening	0.60	0.55		0.87
	Claim 4: Research	0.64	0.61	0.55	
4	Claim 1: Reading		0.89	0.92	0.91
	Claim 2: Writing	0.65		0.86	0.86
	Claim 3: Listening	0.63	0.56		0.89
	Claim 4: Research	0.66	0.60	0.58	
5	Claim 1: Reading		0.87	0.91	0.91
	Claim 2: Writing	0.65		0.83	0.87
	Claim 3: Listening	0.63	0.56		0.89
	Claim 4: Research	0.69	0.64	0.60	
6	Claim 1: Reading		0.87	0.91	0.92
	Claim 2: Writing	0.65		0.85	0.88
	Claim 3: Listening	0.61	0.56		0.89
	Claim 4: Research	0.66	0.62	0.56	
7	Claim 1: Reading		0.86	0.91	0.92
	Claim 2: Writing	0.65		0.84	0.87
	Claim 3: Listening	0.61	0.56		0.89
	Claim 4: Research	0.67	0.62	0.57	
8	Claim 1: Reading		0.89	0.93	0.90
	Claim 2: Writing	0.66		0.84	0.88
	Claim 3: Listening	0.62	0.55		0.88
	Claim 4: Research	0.65	0.62	0.56	
11	Claim 1: Reading		0.89	0.91	0.93
	Claim 2: Writing	0.67		0.85	0.90
	Claim 3: Listening	0.63	0.57		0.90
	Claim 4: Research	0.68	0.64	0.59	

Table 32. Correlations among Claim Scores for Mathematics

Grade	Claim	Observed & Disattenuated Correlation		
		Claim 1	Claims 2 & 4	Claim 3
3	Claim 1		0.97	0.94
	Claims 2 & 4	0.79		0.99
	Claim 3	0.77	0.73	
4	Claim 1		0.97	0.96
	Claims 2 & 4	0.81		0.98
	Claim 3	0.80	0.75	
5	Claim 1		0.98	0.96
	Claims 2 & 4	0.77		1
	Claim 3	0.76	0.71	
6	Claim 1		0.99	0.97
	Claims 2 & 4	0.80		1
	Claim 3	0.76	0.72	
7	Claim 1		1	0.96
	Claims 2 & 4	0.81		1
	Claim 3	0.77	0.72	
8	Claim 1		1	0.96
	Claims 2 & 4	0.78		1
	Claim 3	0.74	0.7	
11	Claim 1		0.96	0.94
	Claims 2 & 4	0.76		0.96
	Claim 3	0.73	0.66	

Legend.

Claim 1: Concepts and Procedures

Claims 2 & 4: Problem Solving & Modeling and Data Analysis

Claim 3: Communicating Reasoning

5. RELIABILITY

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

The reliability evidence of the South Dakota summative tests 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

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

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

Another way to examine test reliability is with the SEM. In IRT, SEM is estimated as a function of test information provided by a given set of items that make up the test. In CAT, items administered vary across all students, so the SEM can vary across students, also, which yields CSEM. The average CSEM can be computed as

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

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

Table 33 presents the marginal reliability coefficients and the average CSEM for the total scale scores.

Table 33. Marginal Reliability for ELA/L and Mathematics

Grade	N	Number of Items Specified in Test Blueprint	Marginal Reliability	Scale Score Mean	Scale Score SD	Average CSEM
ELA/L						
3	10,372	38–41	0.91	2420.56	88.35	26.33
4	10,650	38–41	0.91	2460.57	95.37	28.11
5	10,607	38–41	0.92	2496.80	97.12	27.74
6	10,464	38–42	0.91	2520.97	94.32	28.46
7	10,511	38–42	0.91	2547.18	97.05	29.41
8	10,575	38–42	0.91	2557.86	99.73	30.21
11	9,876	39–41	0.91	2606.39	110.90	32.60
Mathematics						
3	10,434	35–40	0.95	2437.87	84.86	19.66
4	10,717	35–40	0.95	2475.25	87.19	19.84
5	10,662	35–40	0.94	2500.71	92.74	23.40
6	10,520	34–40	0.94	2521.68	103.65	25.60
7	10,574	34–40	0.94	2536.57	107.98	27.00
8	10,621	34–40	0.93	2551.95	117.12	30.65
11	9,893	37–42	0.93	2580.11	116.91	30.80

5.2. STANDARD ERROR CURVES

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

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

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

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

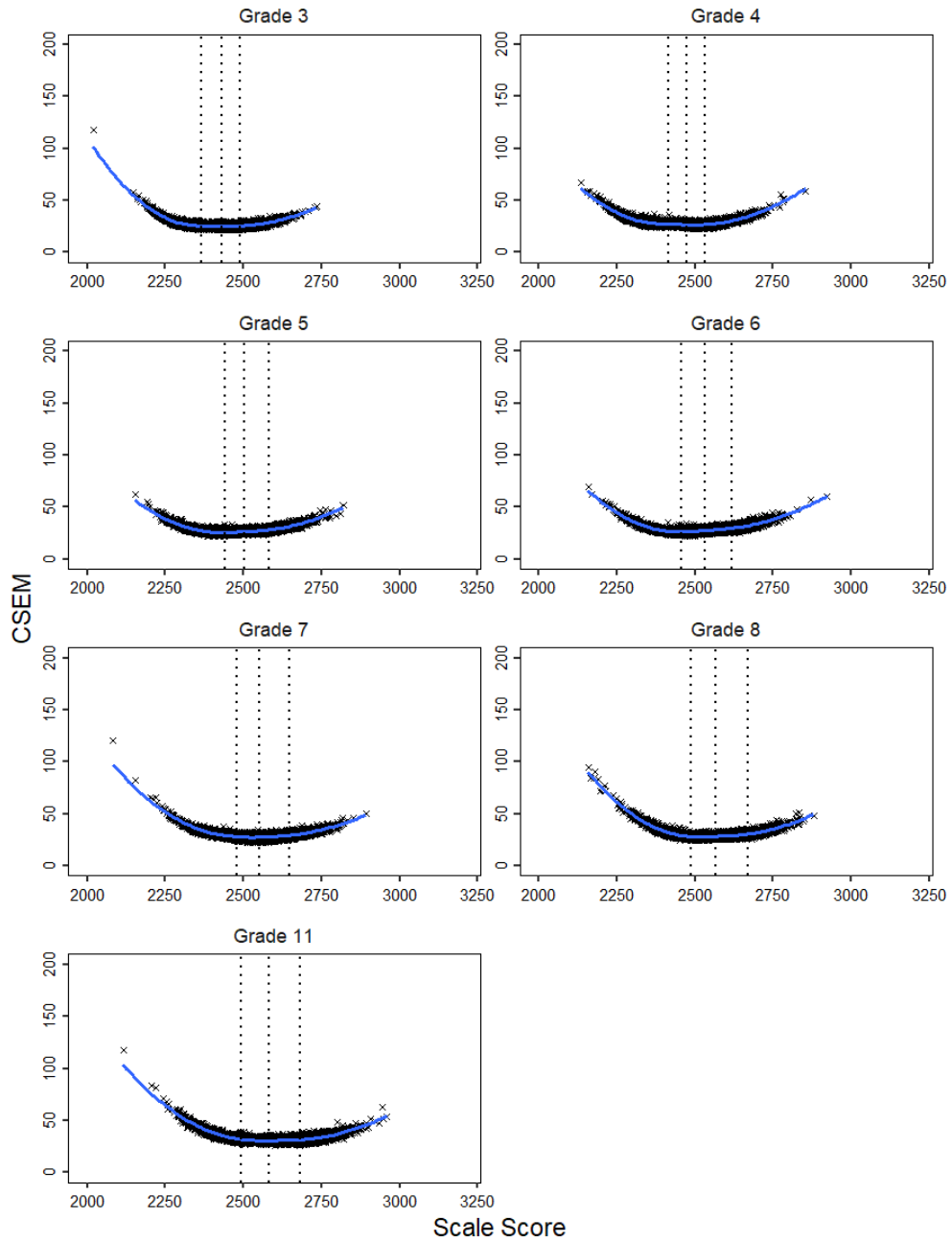
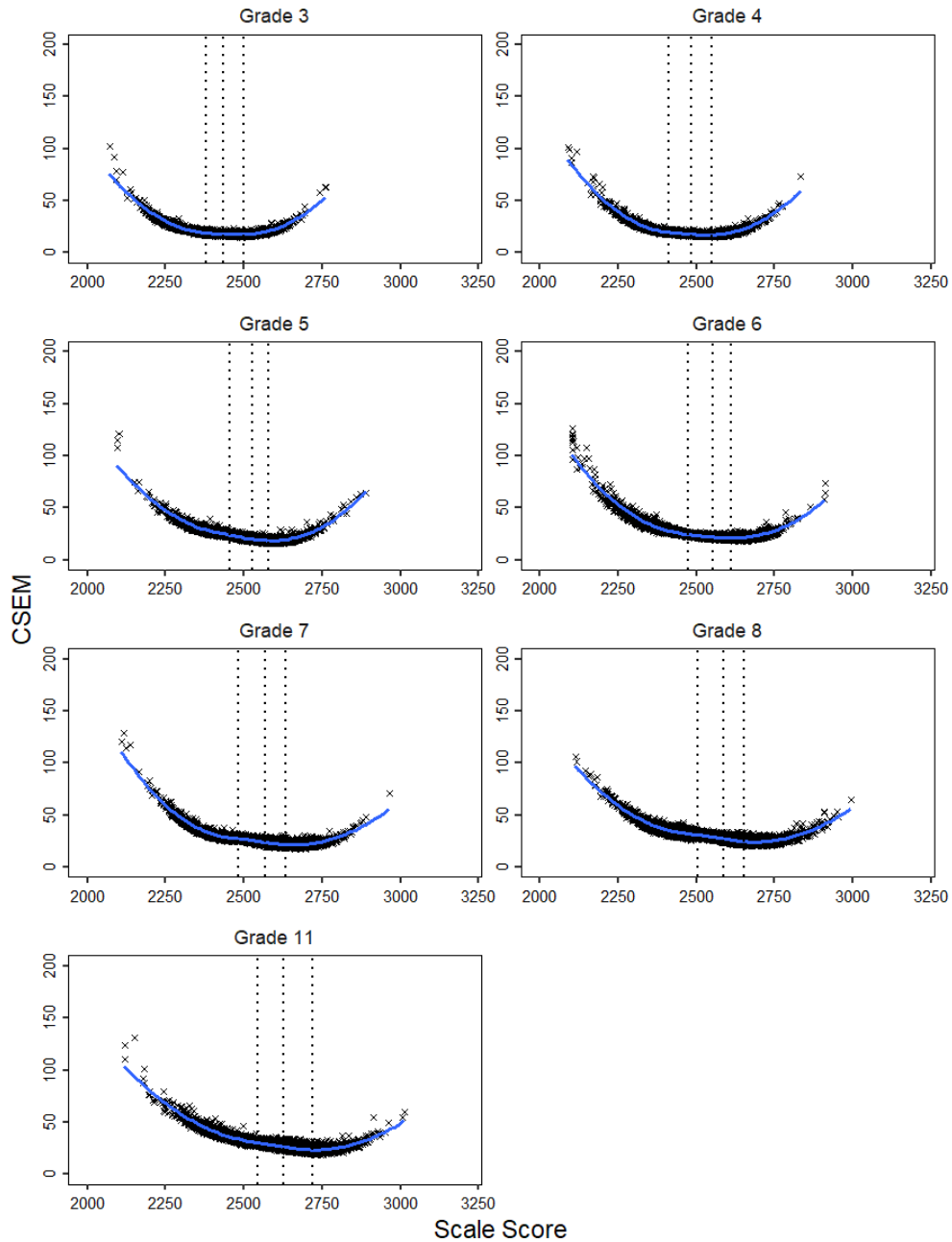


Figure 12. Conditional Standard Error of Measurement for Mathematics



The SEMs presented in Figures 11 and 12 are summarized in Tables 34 and 35. Table 34 provides the average CSEM for all scale scores and by achievement level. Table 35 presents the average CSEM at each cut score and the difference in average CSEMs between two cut scores. As shown in Figures 11 and 12, the largest average CSEM is in Level 1 for most grades in ELA/L and all grades in mathematics. The average CSEMs at all cut scores are similar in ELA/L, but larger at the Level 2 cut in mathematics.

Table 34. Average Conditional Standard Error of Measurement by Achievement Levels

Grade	Level 1	Level 2	Level 3	Level 4	Average CSEM
ELA/L					
3	27.98	25.09	25.03	26.94	26.33
4	29.41	26.93	26.41	28.87	28.11
5	27.82	25.86	26.97	30.52	27.74
6	28.49	26.59	28.36	31.62	28.46
7	32.25	27.36	28.15	31.11	29.41
8	32.87	27.97	28.94	32.58	30.21
11	36.96	30.81	30.89	33.26	32.60
Mathematics					
3	23.44	18.25	17.57	19.28	19.66
4	24.62	18.48	17.28	18.93	19.84
5	29.04	21.75	19.22	19.71	23.40
6	32.65	22.90	21.50	21.99	25.60
7	33.28	25.71	22.43	22.02	27.00
8	36.79	29.28	25.71	25.23	30.65
11	36.99	28.39	24.62	24.40	30.80

Table 35. Average Conditional Standard Error of Measurement at Each Achievement Level Cut and Difference of the Standard Errors of Measurement Between Two Cuts

Grade	L2 Cut	L3 Cut	L4 Cut	L2-L3	L3-L4	L2-L4
ELA/L						
3	25.14	24.78	25.26	0.36	0.48	0.12
4	27.77	26.11	27.05	1.66	0.94	0.72
5	25.45	26.22	27.89	0.77	1.67	2.44
6	25.98	27.24	29.45	1.26	2.21	3.47
7	27.79	27.92	29.13	0.13	1.20	1.34
8	27.38	28.30	30.59	0.92	2.29	3.21
11	31.54	30.09	31.65	1.44	1.56	0.11
Mathematics						
3	18.90	17.99	17.46	0.91	0.53	1.43
4	19.65	17.79	16.98	1.86	0.81	2.67
5	24.03	20.07	18.47	3.96	1.60	5.56
6	24.09	21.80	21.83	2.29	0.03	2.27
7	27.32	23.96	21.46	3.36	2.51	5.86
8	31.13	26.75	24.27	4.39	2.48	6.86
11	30.21	26.30	22.22	3.91	4.08	7.99

5.3. RELIABILITY OF ACHIEVEMENT CLASSIFICATION

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

For a fixed-form test, the accuracy and consistency of classifications are estimated on a single form's test scores from a single test administration based on the true-score distribution estimated by fitting a bivariate beta-binomial model or a four-parameter beta model (Huynh, 1976; Livingston & 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 on the basis of all sets of items administered across students using an IRT-based method (Guo, 2006).

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

In reality, the true ability is unknown, and students do not take an alternative, 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 later in this section. The true score is an expected value of the test score with a measurement error.

For the i th student, the student's estimated ability is $\hat{\theta}_i$ with SEM of $se(\hat{\theta}_i)$, and the estimated ability is distributed, as $\hat{\theta}_i \sim N(\theta_i, se^2(\hat{\theta}_i))$, 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

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

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

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

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

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

$$p_{i1} = P(-\infty < \theta_i < \text{cut}_1 | \mathbf{z}, \mathbf{b}) = \frac{\int_{-\infty}^{\text{cut}_1} L(\theta | \mathbf{z}, \mathbf{b}) d\theta}{\int_{-\infty}^{+\infty} L(\theta | \mathbf{z}, \mathbf{b}) d\theta}$$

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

where the likelihood function based on general IRT models is

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

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 $\mathbf{b}_j = (a_j, b_{j1}, \dots, b_{jK_j})$ if the j th item is a polytomous item; a_j is the item's discrimination parameter (for Rasch model, $a_j=1$), c_j is the guessing parameter (for Rasch and two-parameter logistic [2PL] models, $c_j=0$), and D is 1.7 for non-Rasch models and 1 for Rasch model.

Classification Accuracy

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

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

where $n_{alm} = \sum_{pl_i=l} p_{im} \cdot n_{alm}$ is the expected count 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 ($l = 1, \dots, L$) is estimated by

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

and the overall classification accuracy is estimated by

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

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

Classification Consistency

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

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

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 the equivalent test form.

The classification consistency (CC) at level l ($l = 1, \dots, L$) is estimated by

$$CC_l = \frac{n_{c ll}}{\sum_{m=1}^L n_{c lm}},$$

and the overall classification consistency is

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

As with classification accuracy, classification consistency is also considered at the proficiency level by repeating the process for overall classification consistency of achievement levels but with the four achievement levels collapsed into two proficiency categories: proficient (achievement levels 3 and 4) and not proficient (achievement levels 1 and 2).

The analysis of the classification index is performed based on overall scale scores. Table 36 provides the percentages of classification accuracy and consistency for 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 fact that the intervals used to compute the classification probabilities for students in L1 and L4 $[-\infty, L2 \text{ cut}; L4 \text{ cut}, \infty]$ are wider than the intervals used to compute the classification probabilities for students in L2 and L3 $[L2 \text{ cut}, L3 \text{ cut}; L3 \text{ cut}, L4 \text{ cut}]$. The misclassification probability tends to be higher for narrower intervals. Classification accuracy and classification consistency at the proficiency cut scores were high, ranging from 91% to 94% for accuracy and from 88% to 91% for consistency.

The accuracy of classifications is higher than the consistency of the classifications in all achievement levels. The accuracy is higher than the consistency because the accuracy is based on one test with a measurement error and the true score. In contrast, consistency is based on two tests with measurement errors. The classification indexes by subgroup are provided in Appendix C: Classification Accuracy and Consistency Index by Subgroup.

Table 36. Classification Accuracy and Consistency

Grade	Achievement Level	ELA/L		Mathematics	
		% Accuracy	% Consistency	% Accuracy	% Consistency
3	Overall	78	70	82	75
	L1	89	83	88	83
	L2	69	58	73	63
	L3	66	56	78	71
	L4	87	79	89	83
	Proficiency Cut	92	88	93	91
4	Overall	78	70	84	77
	L1	90	85	90	85
	L2	63	51	80	72
	L3	65	55	78	70
	L4	87	79	89	83
	Proficiency Cut	92	89	94	91
5	Overall	79	71	82	75
	L1	90	84	90	85
	L2	66	55	76	67
	L3	73	64	70	59
	L4	85	77	88	83
	Proficiency Cut	92	89	93	91
6	Overall	78	70	82	75
	L1	90	83	92	86
	L2	71	61	77	69
	L3	74	66	70	60
	L4	82	72	87	80
	Proficiency Cut	91	88	93	90
7	Overall	79	71	82	75
	L1	89	82	91	85
	L2	70	60	75	67
	L3	77	70	74	64
	L4	83	72	88	83
	Proficiency Cut	91	88	93	90
8	Overall	79	71	81	73
	L1	89	82	90	85
	L2	72	61	72	61
	L3	77	70	69	59
	L4	82	71	89	83
	Proficiency Cut	91	88	93	90
11	Overall	80	71	83	76
	L1	89	82	91	87
	L2	72	61	73	64
	L3	75	67	78	69
	L4	86	78	88	81
	Proficiency Cut	93	90	94	91

5.4. RELIABILITY FOR SUBGROUPS

The reliability of test scores is also computed by subgroup. Tables 37 through 44 present the marginal reliability coefficients by subgroup. The reliability coefficients are similar across subgroups but somewhat lower for American and Alaskan, Limited English Proficiency (LEP), and IDEA subgroups in some grades. A large percentage of students in these subgroups received Level 1 with large SEMs.

Table 37. Marginal Reliability Coefficients for Overall and by Subgroup: ELA/L (Grades 3–4)

Subgroup	Grade 3					Grade 4				
	N	MR	SS	SD	CSEM	N	MR	SS	SD	CSEM
All Students	10,372	0.91	2420.56	88.35	26.33	10,650	0.91	2460.57	95.37	28.11
Female	4,982	0.91	2427.03	86.97	26.28	5,270	0.91	2466.87	93.87	27.97
Male	5,390	0.91	2414.58	89.19	26.39	5,380	0.91	2454.40	96.43	28.24
African American	315	0.90	2378.95	82.89	26.68	348	0.91	2426.45	93.20	28.50
AmerIndian/Alaskan	1,151	0.88	2355.81	77.68	27.40	1,229	0.88	2388.45	84.53	29.14
Asian	168	0.92	2427.03	93.34	26.37	173	0.91	2467.43	91.50	27.86
Hispanic	832	0.91	2385.86	87.29	26.69	895	0.90	2422.11	91.14	28.44
Pacific Islander	12	0.86	2345.40	72.94	27.69	9*				
White	7,239	0.90	2437.20	83.48	26.10	7,383	0.90	2479.64	90.38	27.89
Multi-Racial	655	0.91	2414.26	88.24	26.32	613	0.90	2449.90	89.78	27.90
LEP	698	0.88	2370.13	77.28	26.73	667	0.84	2387.69	72.58	28.76
IDEA	2,106	0.90	2364.85	84.75	27.30	2,097	0.90	2395.09	90.33	29.18
Section 504 Plan	325	0.90	2409.00	82.01	26.08	401	0.90	2450.98	86.83	27.73

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

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

Table 38. Marginal Reliability Coefficients for Overall and by Subgroup: ELA/L (Grades 5–6)

Subgroup	Grade 5					Grade 6				
	N	MR	SS	SD	CSEM	N	MR	SS	SD	CSEM
All Students	10,607	0.92	2496.80	97.12	27.74	10,464	0.91	2520.97	94.32	28.46
Female	5,173	0.92	2503.91	96.11	27.86	5,115	0.90	2532.51	91.84	28.50
Male	5,434	0.92	2490.02	97.60	27.63	5,349	0.91	2509.93	95.35	28.43
African American	325	0.92	2457.42	97.82	27.52	353	0.90	2487.21	87.25	28.02
AmerIndian/Alaskan	1,167	0.91	2423.59	90.44	27.76	1,164	0.89	2449.16	87.88	28.80
Asian	178	0.93	2500.05	111.28	28.74	197	0.90	2545.93	90.51	28.75
Hispanic	869	0.92	2461.22	97.75	28.49	853	0.91	2490.09	91.51	28.10
Pacific Islander	16	0.89	2448.13	79.99	26.40	16	0.94	2500.00	115.80	28.70
White	7,382	0.91	2515.39	90.30	27.67	7,251	0.90	2538.09	88.97	28.47
Multi-Racial	670	0.92	2484.97	95.43	27.43	630	0.91	2510.03	93.91	28.36
LEP	478	0.86	2408.47	73.53	27.46	409	0.85	2434.80	72.00	27.84
IDEA	1,920	0.91	2418.75	90.60	27.88	1,593	0.89	2432.45	86.17	28.68
Section 504 Plan	387	0.92	2489.33	97.07	27.67	425	0.91	2518.55	91.91	28.27

Table 39. Marginal Reliability Coefficients for Overall and by Subgroup: ELA/L (Grades 7–8)

Subgroup	Grade 7					Grade 8				
	N	MR	SS	SD	CSEM	N	MR	SS	SD	CSEM
All Students	10,511	0.91	2547.18	97.05	29.41	10,575	0.91	2557.86	99.73	30.21
Female	5,200	0.90	2559.63	94.79	29.52	5,095	0.90	2575.62	95.80	30.02
Male	5,311	0.91	2535.00	97.69	29.29	5,480	0.91	2541.35	100.48	30.38
African American	327	0.90	2500.89	92.87	29.66	325	0.90	2517.57	97.39	30.88
AmerIndian/Alaskan	1,161	0.87	2473.78	88.12	31.15	1,092	0.88	2481.51	90.53	31.73
Asian	157	0.91	2563.27	99.00	29.40	155	0.91	2578.96	103.05	30.07
Hispanic	828	0.91	2509.78	101.27	29.89	846	0.91	2518.24	101.19	30.97
Pacific Islander	19	0.90	2541.50	90.48	28.65	14	0.91	2562.08	97.17	29.78
White	7,375	0.90	2565.23	91.01	29.10	7,513	0.90	2575.44	93.38	29.86
Multi-Racial	644	0.90	2540.72	91.16	28.88	630	0.91	2549.31	102.21	30.25
LEP	425	0.85	2450.47	79.56	30.97	430	0.83	2461.37	77.53	31.91
IDEA	1,433	0.87	2451.25	85.80	30.94	1,379	0.86	2457.77	85.66	32.53
Section 504 Plan	447	0.90	2539.37	91.70	28.92	522	0.91	2549.87	99.58	30.20

Table 40. Marginal Reliability Coefficients for Overall and by Subgroup: ELA/L (Grade 11)

Subgroup	Grade 11				
	N	MR	SS	SD	CSEM
All Students	9,876	0.91	2606.39	110.90	32.60
Female	4,746	0.90	2624.09	104.13	32.30
Male	5,130	0.92	2590.02	114.42	32.88
African American	303	0.90	2556.35	105.06	32.91
AmerIndian/Alaskan	857	0.90	2530.04	104.33	33.31
Asian	163	0.91	2627.92	107.91	32.66
Hispanic	744	0.92	2556.49	117.64	33.44
Pacific Islander	15	0.86	2558.29	84.83	31.58
White	7,351	0.91	2622.39	105.47	32.42
Multi-Racial	443	0.91	2600.34	108.03	32.57
LEP	333	0.82	2460.88	85.38	36.04
IDEA	893	0.86	2481.36	93.96	35.31
Section 504 Plan	526	0.91	2606.91	109.35	32.60

Table 41. Marginal Reliability Coefficients for Overall and by Subgroup: Mathematics (Grades 3–4)

Subgroup	Grade 3					Grade 4				
	N	MR	SS	SD	CSEM	N	MR	SS	SD	CSEM
All Students	10,434	0.95	2437.87	84.86	19.66	10,717	0.95	2475.25	87.19	19.84
Female	5,012	0.94	2432.19	81.27	19.55	5,298	0.94	2468.95	82.42	19.61
Male	5,422	0.95	2443.12	87.72	19.76	5,419	0.95	2481.40	91.20	20.05
African American	332	0.93	2380.54	83.58	21.73	357	0.93	2427.98	87.09	22.75
AmerIndian/Alaskan	1,149	0.92	2369.44	78.07	21.65	1,230	0.91	2399.77	76.19	22.51
Asian	172	0.95	2435.70	84.91	19.46	179	0.95	2475.19	88.79	20.25
Hispanic	869	0.94	2396.11	85.62	21.30	941	0.93	2427.86	82.09	21.54
Pacific Islander	12	0.92	2363.23	70.47	20.53	9*				
White	7,248	0.94	2457.42	76.77	19.00	7,388	0.94	2497.35	79.13	18.97
Multi-Racial	652	0.94	2427.84	83.66	19.82	613	0.94	2461.66	79.84	19.64
LEP	757	0.93	2379.67	82.31	22.08	735	0.90	2404.94	73.68	22.74
IDEA	2,106	0.94	2386.38	89.70	21.61	2,084	0.94	2416.17	89.58	22.49
Section 504 Plan	324	0.95	2425.46	86.70	19.98	401	0.95	2466.53	88.25	20.69

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

Table 42. Marginal Reliability Coefficients for Overall and by Subgroup: Mathematics (Grades 5–6)

Subgroup	Grade 5					Grade 6				
	N	MR	SS	SD	CSEM	N	MR	SS	SD	CSEM
All Students	10,662	0.94	2500.71	92.74	23.40	10,520	0.94	2521.68	103.65	25.60
Female	5,204	0.93	2493.96	88.94	23.50	5,138	0.94	2520.64	99.46	25.30
Male	5,458	0.94	2507.15	95.78	23.31	5,382	0.94	2522.68	107.50	25.87
African American	330	0.92	2455.85	92.34	25.99	360	0.92	2465.95	107.15	30.39
AmerIndian/Alaskan	1,167	0.89	2419.63	82.17	27.82	1,174	0.90	2426.57	98.58	31.05
Asian	187	0.95	2508.54	100.18	23.40	201	0.94	2542.84	103.02	24.25
Hispanic	907	0.92	2460.77	89.33	25.41	884	0.92	2478.85	100.60	27.89
Pacific Islander	16	0.89	2413.88	85.69	27.91	17	0.95	2479.86	128.09	29.40
White	7,386	0.93	2521.77	84.62	22.16	7,252	0.93	2545.76	92.60	24.05
Multi-Racial	669	0.93	2485.84	91.40	23.97	632	0.93	2508.19	99.17	25.52
LEP	531	0.85	2419.46	72.07	28.04	453	0.87	2419.36	91.84	32.81
IDEA	1,917	0.91	2427.64	90.69	27.61	1,598	0.91	2424.82	106.16	32.03
Section 504 Plan	389	0.93	2497.18	89.78	23.33	426	0.94	2525.40	98.71	24.73

Table 43. Marginal Reliability Coefficients for Overall and by Subgroup: Mathematics (Grades 7–8)

Subgroup	Grade 7					Grade 8				
	N	MR	SS	SD	CSEM	N	MR	SS	SD	CSEM
All Students	10,574	0.94	2536.57	107.98	27.00	10,621	0.93	2551.95	117.12	30.65
Female	5,227	0.93	2534.05	105.56	27.07	5,118	0.93	2553.36	112.27	30.33
Male	5,347	0.94	2539.03	110.25	26.93	5,503	0.94	2550.64	121.45	30.95
African American	337	0.91	2465.07	103.56	30.61	334	0.91	2491.83	110.25	33.18
AmerIndian/Alaskan	1,163	0.87	2439.85	88.89	32.38	1,096	0.86	2449.12	99.67	36.74
Asian	163	0.94	2543.78	111.84	26.40	157	0.95	2578.38	140.39	31.17
Hispanic	871	0.92	2478.31	108.07	30.69	880	0.91	2492.61	110.66	33.65
Pacific Islander	20	0.93	2484.83	115.22	30.77	14	0.93	2510.12	120.06	30.74
White	7,376	0.93	2562.72	97.57	25.33	7,511	0.93	2577.66	107.51	29.04
Multi-Racial	644	0.93	2527.68	105.19	27.54	629	0.93	2533.41	117.72	31.75
LEP	486	0.84	2420.92	85.27	34.00	471	0.84	2435.47	92.14	36.97
IDEA	1,429	0.88	2434.51	95.32	32.55	1,379	0.87	2436.36	104.14	36.97
Section 504 Plan	447	0.93	2533.58	100.42	26.50	521	0.93	2548.75	112.19	30.04

Table 44. Marginal Reliability Coefficients for Overall and by Subgroup: Mathematics (Grade 11)

Subgroup	Grade 11				
	N	MR	SS	SD	CSEM
All Students	9,893	0.93	2580.11	116.91	30.80
Female	4,754	0.92	2579.77	110.31	30.40
Male	5,139	0.94	2580.43	122.70	31.16
African American	307	0.90	2519.31	106.63	34.34
AmerIndian/Alaskan	857	0.85	2479.29	93.01	36.46
Asian	162	0.93	2610.78	115.70	29.57
Hispanic	756	0.90	2525.64	109.94	33.95
Pacific Islander	15	0.91	2526.14	111.23	32.68
White	7,354	0.93	2600.48	111.70	29.54
Multi-Racial	442	0.92	2562.67	113.13	31.51
LEP	347	0.75	2451.20	79.50	39.42
IDEA	893	0.82	2452.42	92.49	39.20
Section 504 Plan	529	0.94	2581.69	121.14	30.75

5.5. RELIABILITY FOR CLAIM SCORES

The marginal reliability coefficients and the measurement errors are also computed for the 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 not sufficient to report scores, given the 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 45 and 46 present the marginal reliability coefficients for each claim score in ELA/L and mathematics, respectively.

Table 45. Marginal Reliability Coefficients for Claim Scores in ELA/L

Grade	Claim	Number of Items Specified in Test Blueprint	Marginal Reliability	Scale Score Mean	Scale Score SD	Average CSEM
3	Claim 1: Reading	14–16	0.76	2419.45	100.59	49.55
	Claim 2: Writing	7	0.72	2418.42	113.42	60.12
	Claim 3: Listening	8–9	0.58	2428.36	124.27	80.85
	Claim 4: Research	9	0.70	2411.27	117.89	64.94
4	Claim 1: Reading	14–16	0.76	2458.55	110.41	53.73
	Claim 2: Writing	7	0.71	2455.29	122.29	66.41
	Claim 3: Listening	8–9	0.62	2471.60	128.81	79.84
	Claim 4: Research	9	0.69	2450.90	127.82	70.79
5	Claim 1: Reading	14–16	0.76	2500.14	109.64	53.21
	Claim 2: Writing	7	0.74	2493.98	124.96	64.16
	Claim 3: Listening	8–9	0.62	2499.30	133.77	82.77
	Claim 4: Research	9	0.74	2493.97	123.14	62.92
6	Claim 1: Reading	14–17	0.76	2516.50	110.33	54.20
	Claim 2: Writing	7	0.74	2515.07	116.45	59.60
	Claim 3: Listening	8–9	0.60	2535.43	141.93	90.19
	Claim 4: Research	9	0.67	2529.73	121.50	69.74
7	Claim 1: Reading	14–17	0.76	2544.05	109.27	53.82
	Claim 2: Writing	7	0.74	2546.11	126.53	64.48
	Claim 3: Listening	8–9	0.59	2548.95	130.57	83.83
	Claim 4: Research	9	0.69	2544.47	131.80	73.43
8	Claim 1: Reading	14–17	0.76	2555.91	117.02	57.17
	Claim 2: Writing	7	0.72	2553.02	124.82	66.01
	Claim 3: Listening	8–9	0.59	2563.68	144.14	92.22
	Claim 4: Research	9	0.69	2559.14	131.22	73.29
11	Claim 1: Reading	15–16	0.77	2601.55	128.81	61.57
	Claim 2: Writing	7	0.74	2607.75	134.89	68.74
	Claim 3: Listening	8–9	0.62	2608.44	157.66	97.43
	Claim 4: Research	9	0.69	2608.98	143.28	79.15

Table 46. Marginal Reliability Coefficients for Claim Scores in Mathematics

Grade	Claim	Number of Items Specified in Test Blueprint	Marginal Reliability	Scale Score Mean	Scale Score SD	Average CSEM
3	Claim 1	17–20	0.91	2441.42	92.21	27.98
	Claims 2 & 4	8–10	0.73	2433.85	97.55	50.39
	Claim 3	8–10	0.74	2428.82	101.69	51.94
4	Claim 1	17–20	0.91	2479.21	94.09	28.37
	Claims 2 & 4	8–10	0.77	2468.09	100.92	48.50
	Claim 3	8–10	0.75	2467.15	99.69	49.44
5	Claim 1	17–20	0.89	2504.60	99.67	32.64
	Claims 2 & 4	8–10	0.69	2490.60	109.13	60.54
	Claim 3	8–10	0.70	2489.28	112.91	61.58
6	Claim 1	16–20	0.90	2524.45	111.45	35.66
	Claims 2 & 4	8–10	0.73	2513.04	119.77	62.76
	Claim 3	8–10	0.67	2511.76	124.27	71.21
7	Claim 1	16–20	0.89	2537.62	115.06	37.96
	Claims 2 & 4	8–10	0.73	2531.70	125.22	65.24
	Claim 3	8–10	0.71	2526.45	129.60	69.95
8	Claim 1	16–20	0.89	2553.38	124.79	41.87
	Claims 2 & 4	8–10	0.67	2544.77	137.23	78.48
	Claim 3	8–10	0.68	2542.46	140.03	79.07
11	Claim 1	19–22	0.89	2578.20	120.24	39.83
	Claims 2 & 4	8–10	0.69	2576.20	156.76	86.98
	Claim 3	8–10	0.68	2562.96	152.45	86.63

Legend.

Claim 1: Concepts and Procedures

Claims 2 & 4: Problem Solving & Modeling and Data Analysis

Claim 3: Communicating Reasoning

6. SCORING

The Smarter Balanced Assessment Consortium provided the item parameters that are vertically scaled by linking across 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 and the hand scoring procedure.

6.1. ESTIMATING STUDENT ABILITY USING MAXIMUM LIKELIHOOD ESTIMATION

The South Dakota summative 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

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

where $\mathbf{b}_i = (b_{i,1}, \dots, b_{i,m_i})$ 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 indices the step of the item i .

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

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

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

in the case of items with two or more points,

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

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

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

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

The algorithm allows previously answered items to be changed; however, it does not allow item skipping. 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. Although the update of the ability estimates is performed at each iteration, the overall and claim scores are recalculated using all data at the end of the assessment for the final score.

6.2. RULES FOR TRANSFORMING THETA TO VERTICAL SCALE SCORES

The student's performance in each subject is summarized in an overall test score referred to as a *scale score*. The scale scores represent a linear transformation of the ability estimates (theta scores) using the formula, $SS = a * \theta + b$. The scaling constants a and b are provided by the Smarter Balanced Assessment Consortium. Table 47 presents the scaling constants for each subject for the theta-to-scale score linear transformation. Scale scores are rounded to an integer.

Table 47. Vertical Scaling Constants on the Reporting Metric

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

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

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

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

The scale scores are mapped into four achievement levels using three achievement standards (i.e., cut scores). Table 48 provides three achievement standards for each grade and content area.

Table 48. Cut Scores in Scale Scores

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

6.3. LOWEST/HIGHEST OBTAINABLE SCORES (LOSS/HOSS)

Although the observed score is measured more precisely in a computer-adaptive test than in a fixed-form test, especially for high- and low-performing students, if the item pool does not include enough easy or difficult items to measure low- and high-performing students, the standard error can be large in the low and high ends of the ability range. The Smarter Balanced Assessment Consortium decided to truncate extreme, unreliable student ability estimates. Table 49 presents the lowest obtainable score (LOT or LOSS) and the highest obtainable score (HOT or HOSS) in both theta and scale score metrics. Estimated thetas lower than LOT or higher than HOT are truncated to the LOT and HOT values and are assigned LOSS and HOSS associated with the LOT and HOT. LOT and HOT were applied to all tests and all scores (total and claim scores). The standard error for LOT and HOT is computed using the LOT, and HOT ability estimates given the administered items.

Table 49. Lowest and Highest Obtainable Scores

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

6.4. SCORING ALL CORRECT AND ALL INCORRECT CASES

In item response theory (IRT) maximum likelihood (ML) ability estimation methods, zero and perfect scores are assigned the ability of minus and plus infinity. For all correct and all incorrect cases, the highest obtainable scores (HOT and HOSS) or the lowest obtainable scores (LOT and LOSS) were assigned in the 2014–2015 administration. Since the 2015–2016 administrations, 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/L, 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, indicating relative strength and weakness, are produced.

The difference between the proficiency cut score and the claim score plus or minus 1.5 times the standard error of the claim is used to determine the relative strengths and weaknesses. For summative tests, the specific rules are as follows:

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

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 cannot be produced for a fixed-form test because the number of items included per target (i.e., benchmark) is too low to produce a reliable score at the target level. A typical fixed-form test has only one or two items per target. Even when aggregated, these data reflect the benchmark narrowly because they reflect only one or two ways of measuring the target. However, an adaptive test offers a tremendous opportunity for target-level data at the class, school, and district levels. 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, in each claim (four claims) in ELA/L, and Claim 1 only in mathematics.

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

6.6.1. Target Scores Relative to Student’s Overall Estimated Ability

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

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

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

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

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

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

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

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

For an aggregate unit, a target score is computed by averaging individual student target scores for the target, across all students in the aggregate unit.

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

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 roster, teacher, school, or district is more effective (if $\bar{\delta}_{Tg}$ is positive) or less effective (negative $\bar{\delta}_{Tg}$) in teaching a given target.

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

For target level strengths and weaknesses, the following are reported:

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

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

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

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

For items with two or more score points, using the generalized partial credit 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:

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

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

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

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

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

For an aggregate unit, a target score is computed by averaging individual student target scores for the target, across all students in the aggregate unit.

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

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.

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

For target-level strengths and weaknesses, the following are reported.

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

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

6.7. HANDSCORING

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

For handscoring items in the 2023–2024 summative operational item pool, there were a total of 497 ELA/L SA items, 186 ELA/L essay items, and 334 mathematics items. Table 50 shows the number of handscored items by grade and subject.

Table 50. Number of Handscored Items in 2023–2024 Smarter Balanced Summative Item Pool, by Grade and Subject

Grade	ELA/L		Mathematics
	Short Answer	Essay	
3	13	25	54
4	16	27	49
5	14	27	86
6	85	20	51
7	91	29	22
8	83	29	30
11	195	29	42
Total	497	186	334

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

6.7.1. Rater Selection

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

To supplement this pool, MI also recruited raters with experience successfully scoring other large-scale assessments. MI assigned those raters to the grade level, subject area, and item type for which they were most qualified based on their performance on similar projects. Returning raters were selected based on experience and performance, as well as attendance, and cooperation with work procedures and MI policies. MI maintains evaluations and performance data for all staff who work on each scoring project in order to

determine employment eligibility for future projects. Finally, MI targeted recruitment of new raters as needed, in an effort to continue to identify talent across the country that will best fulfill the handscoring requirements.

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

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

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

6.7.2. Rater Training, Qualification, and Scoring

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

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

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

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

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

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

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

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

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

Rater training time varied by grade and content area. Training for SA brief write, reading, research, and

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

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

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

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

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

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

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

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

A series of automated score verifications were implemented to further ensure the accuracy of scores. For example, a blank check was conducted, which reset scores when a condition code of "blank" was assigned to a response that had one or more characters in the response string (e.g., a response comprised of spaces or tabs). In this case, only after three independent raters had assigned a condition code of "blank" to a response that appeared blank, but which included characters in the response string, was the score recorded. A similar check was run when a score or condition code other than "blank" was assigned to a response that

included no characters in the response string. Automatic resetting of double-scored responses when two raters assign non-adjacent scores, mismatched condition codes, or a combination of a condition code and a numeric score provided an additional score verification. In addition to automatically resetting and rescored these responses, the raters' information was captured in a report and reviewed by scoring directors, one of many tools used to determine retraining needs.

6.7.3. Rater Monitoring, Feedback, and Evaluation

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

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

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

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

Scoring accuracy during handscoring was maintained by continuously assessing rater performance using validity responses. MI specifically evaluated how closely raters' scores aligned with the benchmark scores of these validity responses. Key performance measures included the agreement between rater and

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

benchmark scores, quantified using Quadratic Weighted Kappa (QWK)², and the comparison of mean score differences between the distributions of benchmark and rater-assigned scores.

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

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

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

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

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

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

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

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

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

6.7.4. Rater Agreement

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

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

Tables 51 through 53 provide a summary of the human-human IRR based on items with a sample size greater than or equal to 50. For Mathematics and ELA/L essay items, the tables show the majority of the items administered. For ELA/L SA items, relatively fewer items reached a sample size greater than or equal to 50, and thus a subset of the items administered are represented in the tables. The IRR is presented with mean of percent exact agreement, minimum and maximum percent exact agreements, combined percent exact and adjacent agreement, and the mean, minimum and maximum QWK. The average number of responses, as well as minimum and maximum number of responses to a given item are presented as well.

Table 51. Inter-Rater Agreement for ELA/L Short-Answer Items

Grade	Number of Items	Number of Responses			%Exact			% (Exact+ Adjacent)	QWK		
		Mean	Min	Max	Mean	Min	Max		Mean	Min	Max
3	1	60.0	60	60	76.7	76.7	76.7	100.0	0.73	0.73	0.73
4	4	57.5	56	59	71.7	69.5	75.9	100.0	0.75	0.73	0.77
5	3	57.3	54	59	68.6	64.8	71.2	100.0	0.70	0.68	0.72
6	13	108.5	75	147	69.0	61.9	75.9	100.0	0.63	0.50	0.72
7	13	65.8	51	115	72.0	60.0	78.4	100.0	0.67	0.52	0.80
8	15	64.8	50	137	68.2	51.9	78.0	100.0	0.63	0.42	0.79
11	14	57.2	50	81	71.4	62.7	83.6	100.0	0.66	0.34	0.80

Table 52. Inter-Rater Agreement for ELA/L Essay Items

Grade	Trait	Number of Items	Number of Responses			%Exact			% (Exact+ Adjacent)	QWK		
			Mean	Min	Max	Mean	Min	Max		Mean	Min	Max
3	Conventions	9	51.9	50	54	67.9	53.8	78.8	100.0	0.61	0.51	0.73
	Evid/Elab	9	51.9	50	54	67.0	58.0	74.1	100.0	0.65	0.45	0.75
	Org/Purp	9	51.9	50	54	66.0	58.0	75.9	100.0	0.64	0.45	0.73
4	Conventions	11	51.5	50	54	67.7	61.1	74.5	100.0	0.73	0.63	0.84
	Evid/Elab	11	51.5	50	54	67.9	58.5	84.3	100.0	0.68	0.59	0.80
	Org/Purp	11	51.5	50	54	67.7	58.0	82.4	100.0	0.68	0.56	0.79
5	Conventions	12	53.8	50	58	69.9	57.4	88.2	100.0	0.69	0.60	0.84
	Evid/Elab	12	53.8	50	58	65.0	48.1	78.4	100.0	0.71	0.58	0.83
	Org/Purp	12	53.8	50	58	65.7	53.7	76.5	100.0	0.72	0.61	0.82
6	Conventions	16	66.8	54	74	72.5	65.1	80.6	100.0	0.65	0.43	0.80
	Evid/Elab	16	66.8	54	74	67.2	48.3	75.7	100.0	0.69	0.58	0.84
	Org/Purp	16	66.8	54	74	67.2	46.7	74.3	100.0	0.69	0.59	0.85
7	Conventions	11	51.5	50	54	71.2	59.6	78.4	100.0	0.67	0.57	0.72
	Evid/Elab	11	51.5	50	54	70.7	60.8	80.8	100.0	0.73	0.59	0.82
	Org/Purp	11	51.5	50	54	71.9	60.8	80.8	100.0	0.74	0.59	0.82
8	Conventions	12	51.8	50	54	71.9	59.6	80.0	100.0	0.64	0.37	0.77
	Evid/Elab	12	51.8	50	54	70.1	58.8	79.6	100.0	0.73	0.62	0.83
	Org/Purp	12	51.8	50	54	69.5	60.0	81.5	100.0	0.73	0.65	0.82
11	Conventions	3	50.0	50	50	80.0	70.0	92.0	100.0	0.66	0.38	0.89
	Evid/Elab	3	50.0	50	50	74.7	70.0	84.0	100.0	0.79	0.78	0.79
	Org/Purp	3	50.0	50	50	74.7	68.0	84.0	100.0	0.79	0.77	0.81

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

Table 53. Inter-Rater Agreement for Mathematics Items

Grade	Score Point Range	Number of Items	Number of Responses			%Exact			% (Exact+ Adjacent)	QWK ^a		
			Mean	Min	Max	Mean	Min	Max		Mean	Min	Max
3	0-1	5	67.4	57	75	91.4	83.3	96.0	100.0	NA	NA	NA
4	0-1	4	77.8	74	82	81.7	77.2	84.1	100.0	NA	NA	NA
5	0-1	5	57.8	54	60	93.1	87.9	96.7	100.0	NA	NA	NA
6	0-1	5	83.0	69	102	97.6	96.1	98.8	100.0	NA	NA	NA
7	0-1	5	106.0	87	120	97.9	93.9	100.0	100.0	NA	NA	NA
8	0-1	8	131.6	112	142	87.6	82.8	97.3	100.0	NA	NA	NA
11	0-1	4	89.0	83	98	93.0	88.1	96.7	100.0	NA	NA	NA
3	0-2	13	69.5	54	83	90.8	78.9	95.9	100.0	0.88	0.70	0.98
4	0-2	12	78.6	69	88	91.1	77.1	100.0	100.0	0.71	0.62	1.00
5	0-2	34	57.4	53	65	89.2	68.4	98.3	100.0	0.84	0.39	0.99
6	0-2	29	91.0	81	98	87.4	74.4	100.0	100.0	0.78	0.28	1.00
7	0-2	10	124.1	113	130	91.1	80.8	95.2	100.0	0.83	0.62	0.95
8	0-2	9	133.3	114	145	92.0	83.4	98.2	100.0	0.84	0.68	0.96
11	0-2	12	96.8	84	112	93.8	79.4	100.0	100.0	0.82	0.42	1.00
3	0-3	2	73.0	66	80	91.1	86.3	97.0	100.0	0.92	0.87	0.98
5	0-3	7	56.7	53	61	85.1	79.7	92.7	100.0	0.88	0.73	0.96
7	0-3	1	126.0	126	126	94.4	94.4	94.4	100.0	0.90	0.90	0.90
8	0-3	2	129.5	117	142	80.7	78.6	82.4	100.0	0.92	0.88	0.95
11	0-3	6	104.3	101	110	88.5	79.2	92.1	100.0	0.85	0.68	0.92

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

6.8. AUTOMATED SCORING

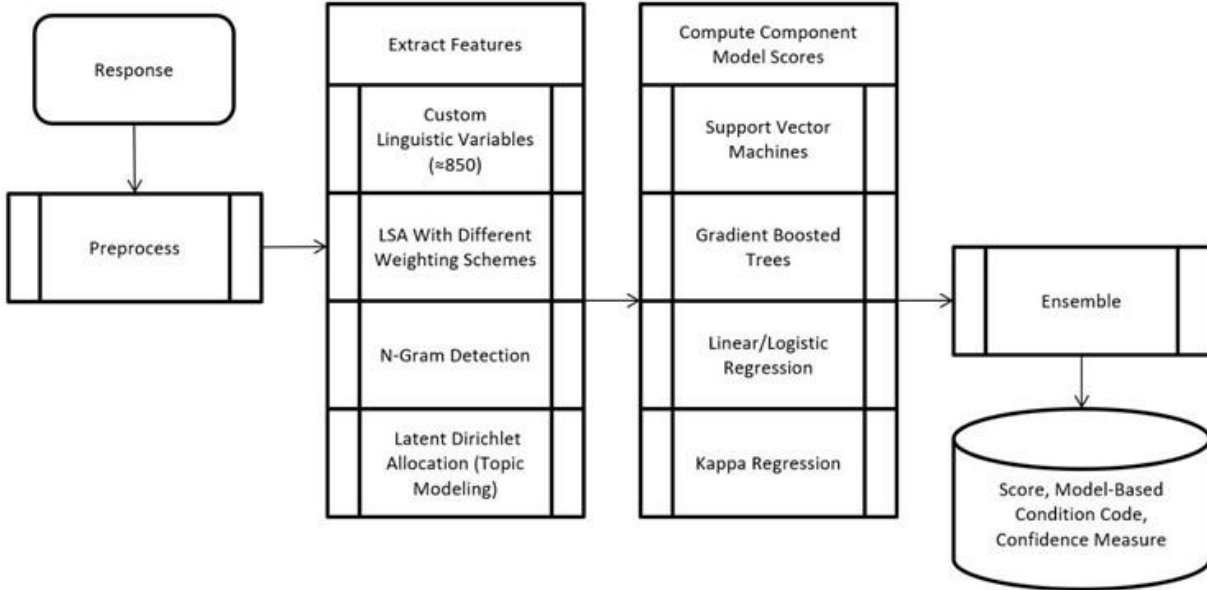
MI’s PEG automated scoring technology was used to score eligible SA and essay items in ELA/L and SA items in mathematics. This section describes PEG, the training and validation sample and process, and the automated scoring process, concluding with the human-machine (HM) agreement statistics.

6.8.1. Project Essay Grade

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

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

Figure 13. PEG Architecture



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

6.8.2. Model Training and Validation

Sample

Automated scoring models were not created for items that had an insufficient quantity of training responses. This was this case for items with low exposure to students, as dictated by the adaptive testing algorithm. Additionally, mathematics performance task items that had multiple parts with scoring dependencies were not considered for automated scoring. Table 54 shows that pretrained models existed for 595 items, thus, no additional training was conducted in preparation for the spring 2024 administration. The remainder of this section describes the process used to train and validate the 595 existing models.

Table 54. Number of Items Eligible for Automated Scoring, by Grade and Subject Area

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

Training Data

Student responses used for training and validation were sourced from the 2018–2019, 2020–2021, 2021–2022, and 2022–2023 Smarter Balanced operational test administrations. Responses were randomly sampled from available on-grade responses in the operational population. For all items, the sample included 1,500–2,000 responses, stratified by score point. The score of record used to train the engine was the score assigned to each response by an expert rater.

For each item, the sample was divided as follows:

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

Model Training

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

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

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

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

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

The approaches just described typically result in 100 models for a single item or essay trait. Ensembling is the process of selecting the “best of the best” models, to result in a small set of strong, yet dissimilar

component models. A linear-kappa regression is used to determine the model ensembling weights. The more accurate a given model is, the more weight it carries in the final score decision.

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

Model Validation

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

Initial Validation

Initial validation was conducted by applying each model to score a respective validation set of responses. The validation set is independent of the training set, in that none of the responses it contains have been used to build the model. Two or more professional raters will not always agree on what score to give a student’s response; therefore, modeling is considered successful when the engine produces scores that agree with professional raters to the same or greater extent than the raters agree with each other. The initial evaluation was made using the criteria shown in Table 55, based on criteria proposed by Williamson, Xi, and Breyer (2012). While Williamson et al. (2012) recommend an agreement between human and machine scores of 0.70 quadratic weighted kappa (QWK) for normally distributed data, a QWK threshold of 0.65 was adopted due to the prevalence of skewed distributions in response data. The degradation (QWK) criterion of .07 is slightly more stringent than proposed by Williamson et al. (2012). The evaluation process was used for both the item-specific scoring models and the condition code models.

Table 55. Initial Model Evaluation Criteria

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

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

Bias Considerations. Subgroup differences in responses to constructed response items can introduce construct-irrelevant variance in scores, in turn threatening valid score interpretations. MI investigated potential sources of bias annually, for newly modeled items, as part of the initial validation process using available data from previous summative administration. Table 56 shows the demographic variables and categories considered. MI received separate datafiles containing (1) handscored data and (2) student demographic data associated with responses.

Table 56. Demographic Variables and Categories

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

For each new item being modeled, analysis was performed on a subgroup if the number of observations (i.e., human-machine scores) was at least 10. A subgroup was flagged for bias if $|SMD| \geq 0.125$ and if the SMD was significant at an overall significance level of 95%. A Bonferroni correction was used to adjust the significance level for each subgroup comparison. An item was flagged for bias, excluded from automated scoring, and handscored if any subgroup comparison associated with the item was flagged.

Secondary Validation

All models associated with items that passed initial validation were subject to a secondary validation at the start of the spring 2024 administration using an early sample of operational responses from that administration. This sample was comprised of the first available 500 responses/item across states, at a minimum. Responses from this sample were scored by both the automated scoring engine and an expert rater. During this interval the human score was reported as the score of record. If the PEG scores were found to be consistent with the scores assigned by the expert raters, subsequent student responses for a given item were scored by PEG using a hybrid human-automated scoring approach. If not, the item was handscored. Table 57 presents the secondary validation criteria. Note that since expert raters are the only humans that score the secondary validation sample, a second human score is not collected and thus QWK degradation is not part of the criteria.

Table 57. Secondary Validation Criteria

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

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

Table 58 presents the secondary validation results. Of the 595 items with models subject to secondary validation, models associated with 454 of the items (76.3%) passed all secondary evaluation criteria.

Table 58. Summary of Secondary Validation Results, by Grade and Subject Area

Grade	Items with All Models Passing Initial Validation Criteria			Items with All Models Passing Secondary Validation Criteria		
	ELA/L		Mathematics	ELA/L		Mathematics
	Short-Answer	Essay		Short-Answer	Essay	
3	12	13	44	12	3	44
4	13	16	42	13	6	40
5	13	10	50	13	5	47
6	32	10	41	19	5	40
7	45	17	15	27	9	15
8	49	14	24	31	9	22
11	80	17	38	46	10	38
Total	244	97	254	161	47	246

Live Training and Validation

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

Table 59 summarizes the results of the live training and validation. Of the 356 items associated with models that underwent live training and validation, models associated with 211 of the items (59.3%) passed all evaluation criteria. While this pass rate is considerably lower than the pass rates during secondary (76.3%) validation efforts, it is most likely explained by the nature of the items modeled. Specifically, since all item models in this sample had failed a prior validation, by design the sample consisted of difficult-to-model items.

Table 59. Summary of Live Training and Validation Results, by Grade and Subject Area

Grade	Items Trained			Items with All Models Passing Initial Validation Criteria		
	ELA/L		Mathematics	ELA/L		Mathematics
	Short-Answer	Essay		Short-Answer	Essay	
3	1	25	9	1	16	4
4	3	24	9	3	19	1
5	1	25	33	1	14	19
6	24	16	10	15	10	4
7	28	20	7	18	12	4
8	26	25	9	17	6	7
11	36	21	4	24	12	4
Total	119	156	81	79	89	43

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

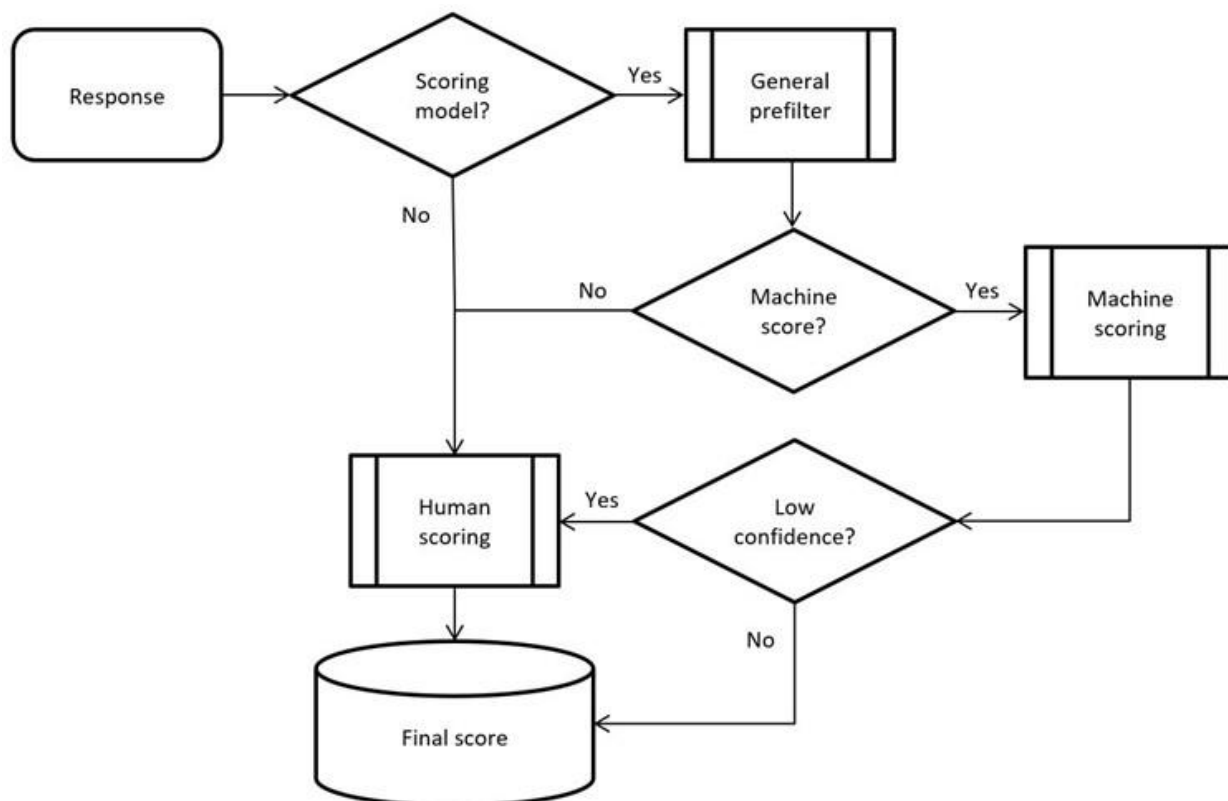
6.8.3. Automated Scoring Processes

Hybrid Scoring Process

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

Figure 14 shows the response routing rules under the hybrid scoring process. In the hybrid model, responses with associated scoring models were first pre-processed for automated scoring and to filter alert responses and certain non-scorable cases (e.g., insufficient text to score or high proportion of copied prompt text). Flags were used to indicate condition codes as defined in the handscoring criteria (see Table 60 and Table 61). For example, PEG flags responses that lack proper development, lack enough content to be scored, are written in an unsupported language, or contain vulgar language or other alert words or phrases that indicate that the response should be reviewed by the client. Responses were then sent to the automated scoring engine, where text features were extracted, the scoring model(s) applied, and responses assigned a score and measure of score confidence. Low-confidence responses straddle the lines between score point values on a rubric and are difficult to score accurately because they exhibit characteristics of multiple score points. Higher-confidence responses received the engine score as the score of record, while lower-confidence responses were routed directly to expert raters, who assigned the score of record. Note that the expert rater pool was dynamic, and raters were added or removed several times each day based on their current performance. Overall, approximately 15% of responses to engine-scored items were flagged as low confidence and scored by expert raters.

Figure 14. Response Routing Rules



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

Table 60. Flags Currently Established

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

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

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

Table 61. Model Setting

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

Scoring Infrastructure

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

Quality Assurance

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

“Alert” Procedures

MI implemented a formal process for informing clients when student responses reflect a possibly dangerous situation for the test-taker. Specifically, MI employed a set of alert procedures to notify the client of responses indicating endangerment, abuse, or psychological and/or emotional difficulties. PEG employed a rule-based detection system to flag responses that are indicative of potentially dangerous situations. Responses flagged by PEG as possible alerts were reviewed by scoring leadership, who decided whether each response should be forwarded to the client. Once vetted, all alerts were provided to CAI, who associated the pertinent student information with the response(s) and contacts the state. In addition, CAI separately evaluates all responses and student-generated text for possible alerts.

Score Delivery

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

6.8.4. PEG-Human Agreement

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

Tables 62 through 64 present the human-machine agreement on the initial and secondary validation samples for ELA/L SA items, ELA/L essay items, and mathematics SA items, respectively. For the PEG-scored items, the human-machine agreement was computed based on the combined data across all states with hybrid scoring in the 2023–2024 summative assessment.

Table 62. Human-Machine Agreement for ELA/L Short-Answer Items on Initial and Secondary Validation Samples, by Grade

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

Table 63. Human-Machine Agreement for ELA/L Essay Items on Initial and Secondary Validation Samples, by Grade

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

Table 64. Human-Machine Agreement for Mathematics Items on Initial and Secondary Validation Samples, by Grade

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

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

Tables 65 through 67 present the human-machine agreement on the live validation samples for ELA/L SA items, ELA/L essay items, and mathematics SA items, respectively. Recall live training did not involve a secondary validation since 2023–24 operational data were used to build the models.

Table 65. Human-Machine Agreement for ELA/L Short-Answer Items on Live Validation Sample, by Grade

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

Table 66. Human-Machine Agreement for ELA/L Essay Items on Live Validation Sample, by Grade

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

Table 67. Human-Machine Agreement for Mathematics Items on Live Validation Samples, by Grade

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

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

6.8.5. Recommendations

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

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

Furthermore, a review of the scoring outcomes revealed that while the mean QWK values for inter-rater agreement generally met expectations, there were concerns regarding the relatively low minimum QWKs observed for some ELA/L short-answer items, as indicated by the minimum QWK values in Table 51. These low QWK values suggest variability in rater agreement for certain items, which could undermine the overall reliability of the scoring process. To address this issue, it is recommended that additional targeted training and calibration sessions be conducted for raters assigned to items with historically low QWK values. This could include additional focused trainings on interpreting and applying scoring rubrics for those items, the development of supplemental materials, as well as implementing more frequent monitoring and feedback loops during the scoring process.

7. REPORTING AND INTERPRETING SCORES

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

This section contains a description of the types of scores reported in the Reporting System and a description of the ways to interpret and use these scores in detail.

7.1. REPORTING SYSTEM

The Reporting System is designed to help educators and students answer questions about how well students have performed on English language arts/literacy (ELA/L) and mathematics assessments. The Reporting System is the online tool that provides all stakeholders with timely, relevant score reports. The Reporting System for the South Dakota assessments was designed such that score reports are easy to read and understand for all stakeholders. This is achieved by using plain, non-technical language to facilitate review by parents and the general public. The Reporting System is also designed to present student performance in a uniform format. For example, similar colors are used for groups of similar elements, such as achievement levels, throughout the design. This design strategy allows readers to compare similar elements and avoid comparing dissimilar elements.

Generally, the Reporting System provides two categories of online score reports: (1) aggregate score reports, and (2) student score reports. Table 68 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 navigating the Reporting System can be found in the *Reporting System User Guide*, located via a help button on the Reporting System.

Table 68. Types of Online Score Reports by Level of Aggregation

Level of Aggregation	Types of Online Score Reports
State District School Teacher Roster	<ul style="list-style-type: none"> • Number of students tested and percentage of students proficient (for overall students and by subgroup) • Average scale score and standard error of average scale score on the overall test and claim (for overall students and by subgroup) • Percentage of students at each achievement level on the overall test and claim (for overall students and by subgroup) • Performance category in each target (for overall students and by subgroup) • On-demand student roster report
Student	<ul style="list-style-type: none"> • Total scale score and standard error of measurement • Achievement level on overall score and claim score with achievement-level descriptors • Average scale scores and standard errors of average scale scores for student’s school, district, and state • Student growth in scale score and achievement level over time • Writing performance descriptors and scores by dimensions

Aggregate score reports at a selected aggregate level are provided for students overall and by subgroup. Users can view student assessment results in any of the subgroups. Table 69 presents the types of subgroups and subgroup categories provided in the Reporting System.

Table 69. Types of Subgroups

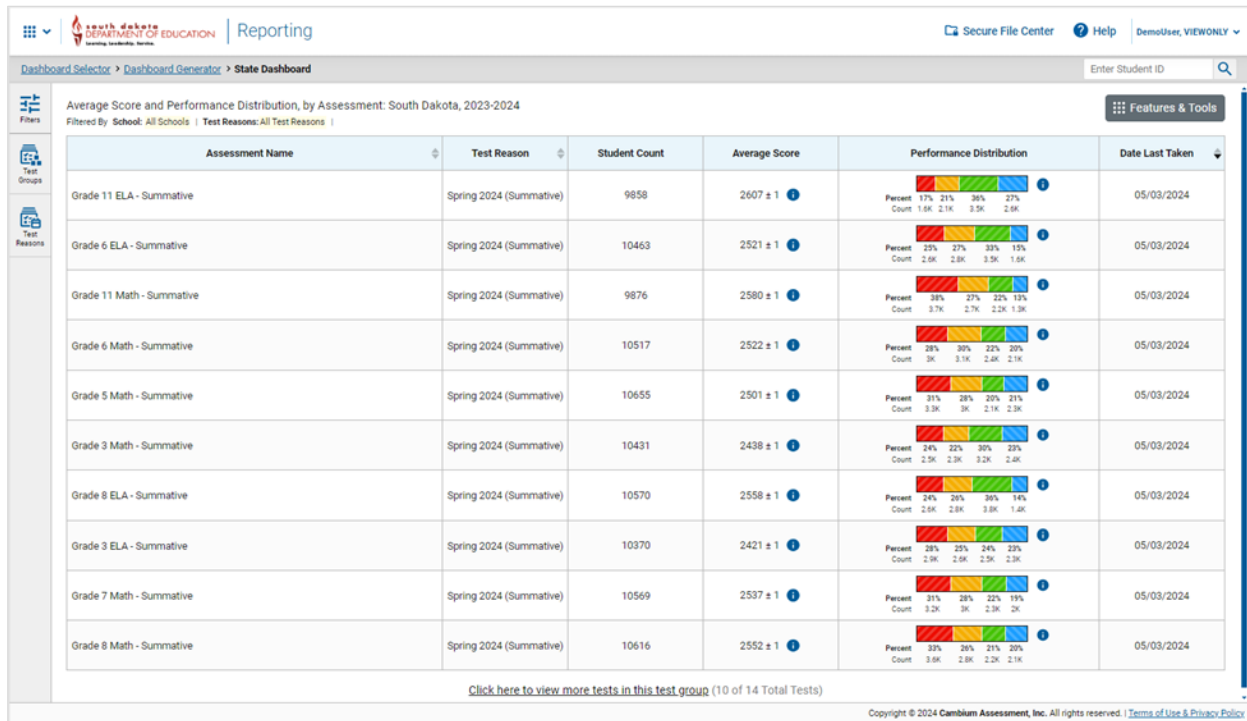
Subgroup	Subgroup Category
Gender	Male Female
IDEA Indicator	Yes No
Limited English Proficiency (LEP) Status	Yes No
Section 504 Status	Yes No Unknown/Cannot Provide
Race/Ethnicity	American Indian or Alaskan Native Asian Black or African American Hispanic or Latino Multi-Racial Native Hawaiian or Other Pacific Islander White Declined to Report

7.1.1. Dashboard

The Reporting System provides a state dashboard for authorized state-level users to track student performance for a test across the entire state. The dashboard summarizes students’ performance for both ELA/L and mathematics in each grade, including (1) student count, (2) average scale score and standard error of the average scale score, (3) percentage and counts of students at each achievement level, and (4) test date last taken. Users can specify the test and administration year to display in the report.

Exhibit 1 presents an example dashboard page at the state level.

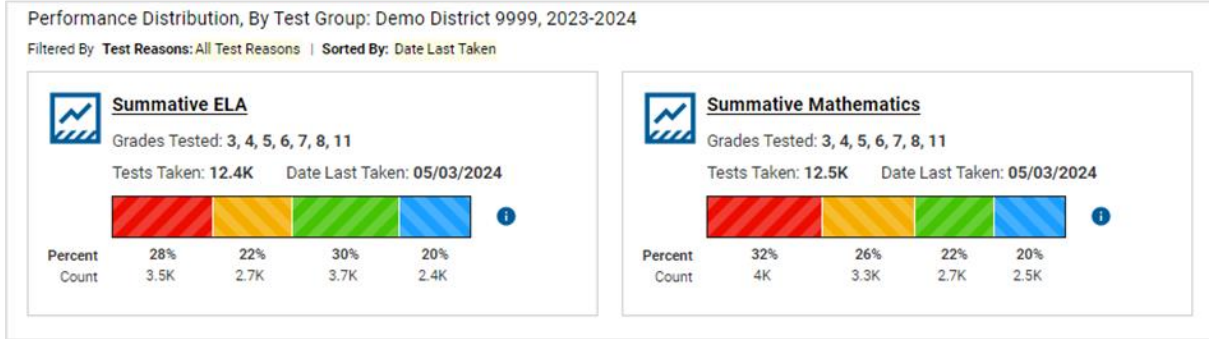
Exhibit 1. Dashboard: State Level



Once authorized users in the district, school, and teacher level log in to the Reporting System, the dashboard page shows overall test results for all tests that the students have taken grouped by test family (e.g., South Dakota Summative ELA/L). The dashboard summarizes students’ performance by test family for both ELA/L and mathematics across all grades, including (1) the grades of the students who have tested, (2) the number of tests taken, (3) the test date last taken, and (4) the percentage and counts of students at each achievement level. District personnel see district summaries, school personnel see school summaries, and teachers see summaries of their students.

Exhibit 2 presents an example dashboard page at the district level.

Exhibit 2. Dashboard: District Level



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

Exhibit 3 presents an example detailed dashboard page for summative ELA/L at the district level.

Exhibit 3. Detailed Dashboard: District Level

Average Score and Performance Distribution, by Assessment: Demo District 9999, 2023-2024
Filtered By: School: All Schools | Test Reasons: All Test Reasons | Features & Tools

Assessment Name	Test Group	Test Grade	Test Reason	Student Count	Average Score	Performance Distribution	Date Last Taken
Grade 11 ELA - Summative	Summative	11	Spring 2024 (Summative)	1704	2619 ± 3	Percent: 15% 17% 37% 31% Count: 261 288 627 528	05/03/2024
Grade 8 ELA - Summative	Summative	8	Spring 2024 (Summative)	1821	2555 ± 2	Percent: 26% 25% 35% 14% Count: 476 453 636 254	05/03/2024
Grade 7 ELA - Summative	Summative	7	Spring 2024 (Summative)	1769	2542 ± 2	Percent: 27% 24% 35% 13% Count: 473 416 619 261	05/02/2024
Grade 6 ELA - Summative	Summative	6	Spring 2024 (Summative)	1748	2517 ± 2	Percent: 27% 26% 31% 16% Count: 473 455 536 282	05/03/2024
Grade 5 ELA - Summative	Summative	5	Spring 2024 (Summative)	1768	2489 ± 2	Percent: 33% 20% 29% 18% Count: 582 347 511 328	04/24/2024
Grade 4 ELA - Summative	Summative	4	Spring 2024 (Summative)	1828	2452 ± 2	Percent: 37% 21% 21% 21% Count: 672 380 393 383	04/26/2024
Grade 3 ELA - Summative	Summative	3	Spring 2024 (Summative)	1746	2414 ± 2	Percent: 33% 23% 22% 22% Count: 573 405 385 383	05/01/2024

Rows per page: 70 | 7 Items: 1 of 1

7.1.2. Aggregate Score Reports: Overall Performance

Student performance for each grade in a subject area for a selected aggregate level is presented when users select a specific assessment name. On each aggregate report, the summary report presents the summary results for the selected aggregate unit and the summary results for the state and the aggregate unit both above and below the selected aggregate. For example, if a district is selected, the summary results of the state and individual schools within the district and the district summary results are provided to facilitate a comparison between the district's performance and the other aggregate levels. The aggregated subject summary report provides the summaries on a specific grade in a subject, including (1) student count, (2) the average scale score and standard error of the average scale score, (3) the percentage and counts of

students in each achievement level, and (4) the percentage of proficient students. The summaries are also presented for students overall and by subgroup.

Exhibit 4 presents an example overall performance summary result for grade 8 ELA/L at the district level, and Exhibit 5 presents an example summary by gender.

Exhibit 4. Overall Performance Summary Results for Grade 8 ELA/L: District Level

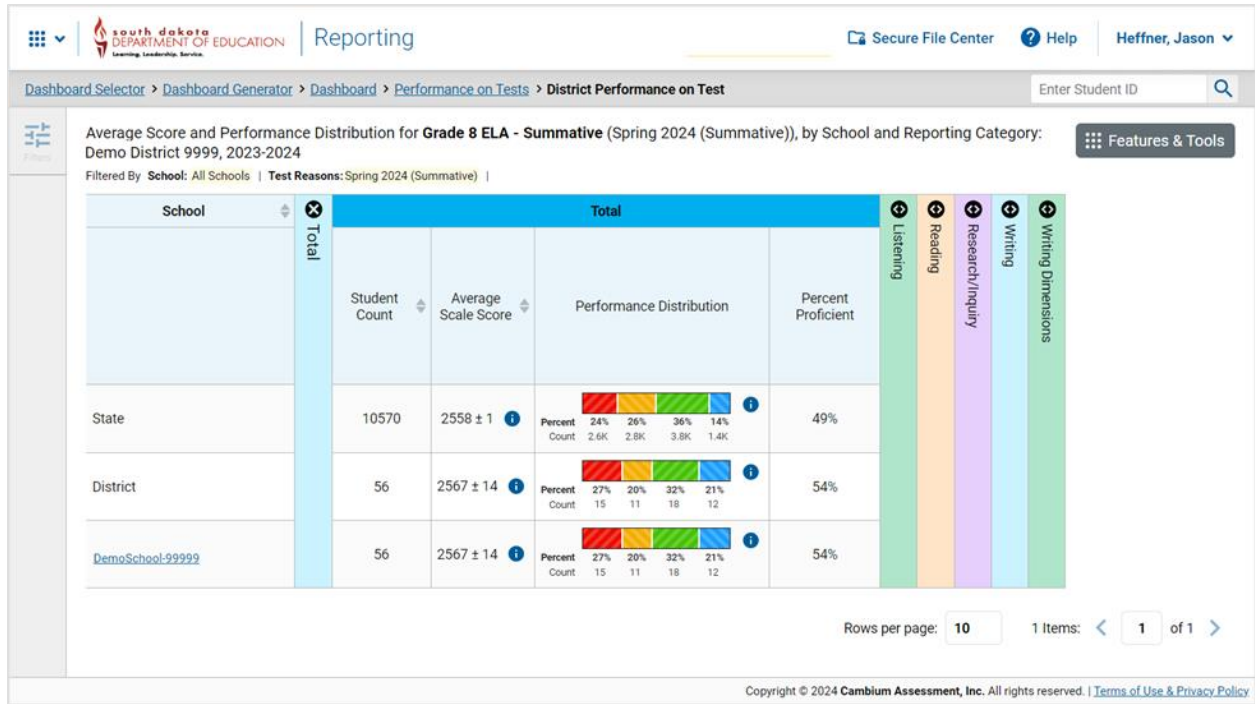
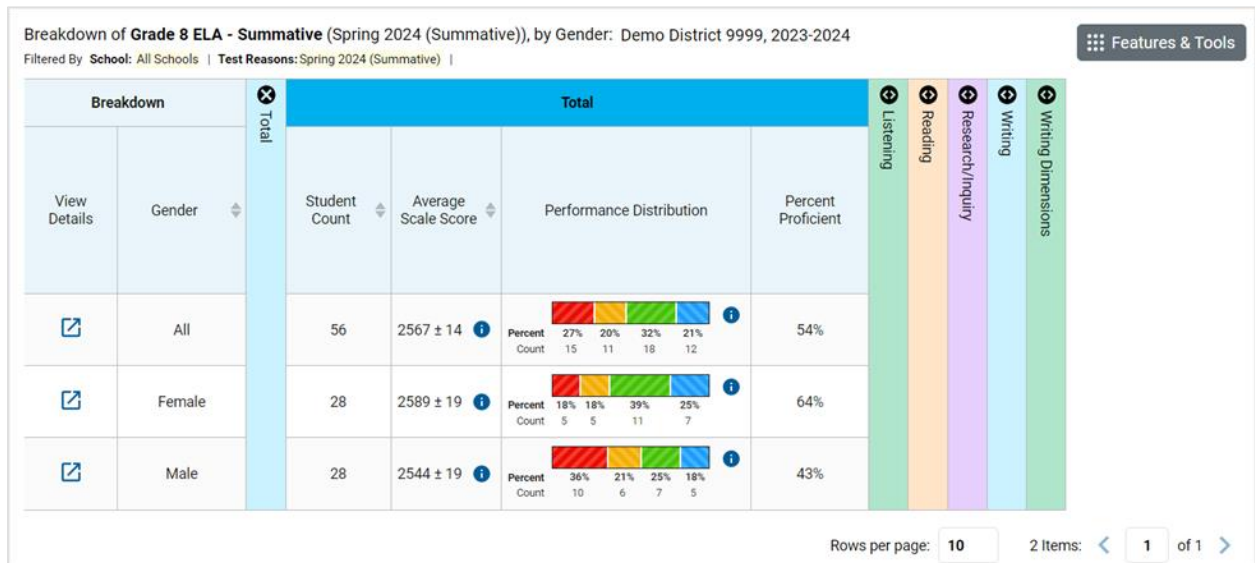


Exhibit 5. Overall Performance Results for Grade 8 ELA/L by Gender: District Level



7.1.3. Aggregate Score Reports: Claim and Target Performance

Detailed summaries on aggregated claim and target results are also available on the same report page when a claim on the right side of the page is selected. For the claim result, both the average scale score and standard error of the average scale score are presented. For the target result, the strength or weakness indicators on each target within a claim are presented. These strength or weakness indicators are presented in two ways. The "Proficient?" measure indicates whether the group's performance on each target is better than (checkmark), less than (x mark), or not different from (half-filled circle) the proficiency standard for the selected test. The "Weak or Strong?" measure presents whether the group's performance on each target is lower than (minus sign), higher than (plus sign), or not different from (equal sign) the group's overall performance. If there is insufficient information in the "Proficient?" measure or "Weak or Strong?" measure, this is indicated with a star sign (*).

Like the overall performance summary results, the summary report presents results for the selected aggregate unit, the state, and the aggregate units above and below the selected aggregate unit. Also, the summaries on claim- and target-level performance can be presented for overall students and by subgroup.

Exhibit 6 presents an example of claim- and target-level results for grade 5 mathematics at the district level.

Exhibit 6. Claim- and Target-Level Results for Grade 5 Mathematics: District Level

School	Claim Average Scale Score	Performance Distribution	Concepts and Procedures									
			Target A		Target B		Target C		Target D		Target E	
			Proficient?	Weak or Strong?	Proficient?	Weak or Strong?	Proficient?	Weak or Strong?	Proficient?	Weak or Strong?	Proficient?	Weak or Strong?
State	2553 ± 1	Percent: 29%, 33%, 38% Count: 4.1K, 3.9K, 2.8K	✗	-	✗	=	✗	=	✗	+	✗	+
District	2593 ± 16	Percent: 25%, 33%, 42% Count: 14, 18, 23	✗	-	✓	+	⊖	-	✓	+	✓	+
DemoSchool-9999	2593 ± 16	Percent: 25%, 33%, 42% Count: 14, 18, 23	✗	-	✓	+	⊖	-	✓	+	✓	+

7.1.4. Roster Performance Report

Class, teacher, and school performance rosters provide users with performance data for a group of students belonging to a system-defined or user-defined class. The report includes (1) the student's overall subject scale scores with standard error of measurement, (2) the achievement level, and (3) the performance category for each claim.

Exhibit 7 shows a sample roster performance report for grade 5 mathematics.

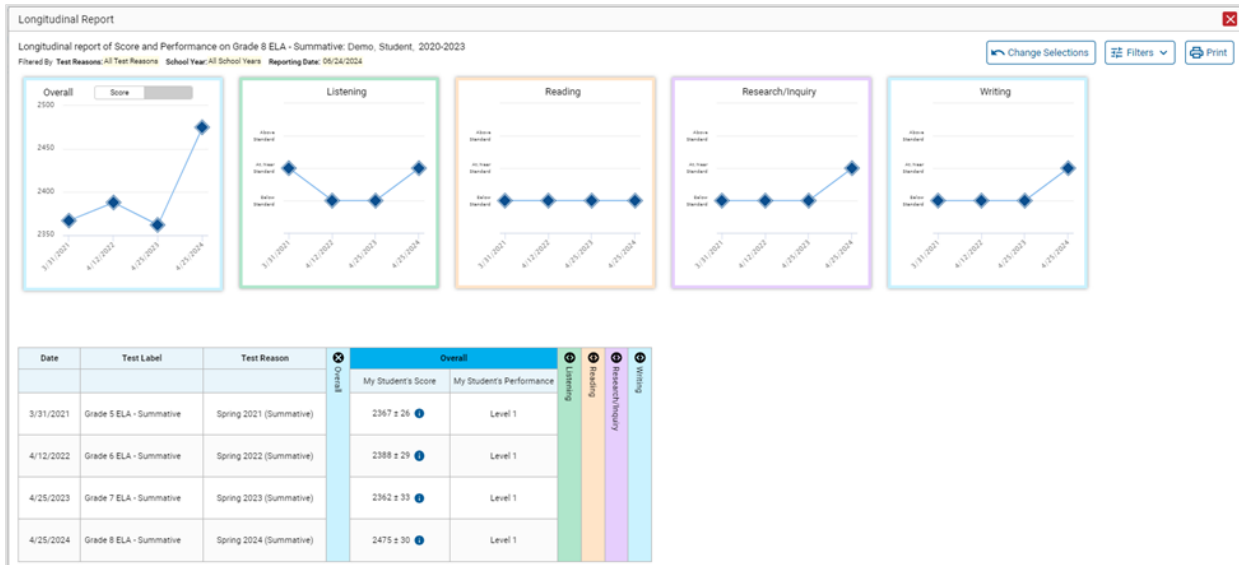
Exhibit 7. Roster Performance Report for Grade 5 Mathematics



7.1.5. Trend Report

The trend (i.e., longitudinal) page provides the trend of student performance for individual level and aggregate level over time. The trend report can be set to plot either average scale scores or percentage of students in each achievement level on the graph for the selected aggregate unit. The trend report is also available at the individual student level. Exhibit 8 presents an example trend report page for ELA/L at the individual student level.

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



7.1.6. Individual Student Report

When a student completes a test and any handscored items have been scored, an individual student report (ISR) can be generated and exported as a PDF file. The ISR shows the student's overall performance on the test with detailed information on multiple pages. The ISR provides (1) the scale score and standard error of measurement (SEM); (2) the achievement level for the overall test; (3) the performance category in each claim; (4) the average scale scores for student's state, district, and school in each subject area; and (5) the writing scores and performance descriptors in each dimension (ELA/L only).


The student's name, scale score with the SEM, and achievement level are shown at the top of the first page of the ISR. In the middle section, the student's performance is described in detail using a barrel chart. The student's scale score is presented with the SEM using a "±" sign in the barrel chart. The SEM represents the precision of the scale score, or the range in which the student would likely score if a similar test were administered multiple times. The achievement-level descriptors (ALDs) with cut scores at each achievement level are also provided in the barrel chart. This defines the content-area knowledge, skills, and processes that test takers at the achievement level are expected to possess.

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

On the following page, the student's performance on each claim is displayed alongside a description of his or her performance on each claim. At the bottom of the page, the student's performance in writing dimension scores (ELA/L only) is displayed alongside a description of his or her performance on each writing dimension. The third page shows the trend of student performance over time.

Exhibit 9 presents an example of an ISR for grade 8 ELA/L.

Exhibit 9. Individual Student Report for ELA/L



Reporting

Individual Student Report

Demo, Student

Student ID: 999999999 | Student DOB: 12/21/2009 | Enrolled Grade: 8
Date Taken: 4/25/2024

Grade 8 ELA - Summative 2023-2024

Demo District
Demo School

Scale Score: 2475±30 Performance: Level 1

How Did Your Child Do on the Test?

Score
2475 ±30

Meets State Standard

Does Not Meet State Standard

2989

2668

2567

2487

2097

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

Level 3 The student has met the achievement standard and demonstrates progress toward mastery of the knowledge and skills in English language arts/literacy needed for likely success in entry-level credit-bearing college coursework after high school.

Level 2 The student has nearly met the achievement standard and may require further development to demonstrate the knowledge and skills in English language arts/literacy needed for likely success in entry-level credit-bearing college coursework after high school.

Level 1 The student has not met the achievement standard and needs substantial improvement to demonstrate the knowledge and skills in English language arts/literacy needed for likely success in entry-level credit-bearing college coursework after high school.


How Does Your Child's Score Compare?

Name	Average Scale Score
South Dakota	2558±1
Demo District	2555±2
Demo School	2561±6

Information on Standard Error of Measurement

A student's score is best interpreted when recognizing that the student's knowledge and skills fall within a score range and not just a precise number. For example, 2300 (±10) indicates a score range between 2290 and 2310.

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



Reporting

Individual Student Report

Demo, Student

Student ID: 999999999 | Student DOB: 12/21/2009 | Enrolled Grade: 8

Date Taken: 4/25/2024

Grade 8 ELA - Summative 2023-2024





Demo District
Demo School

Scale Score: 2475±30 Performance: Level 1

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

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

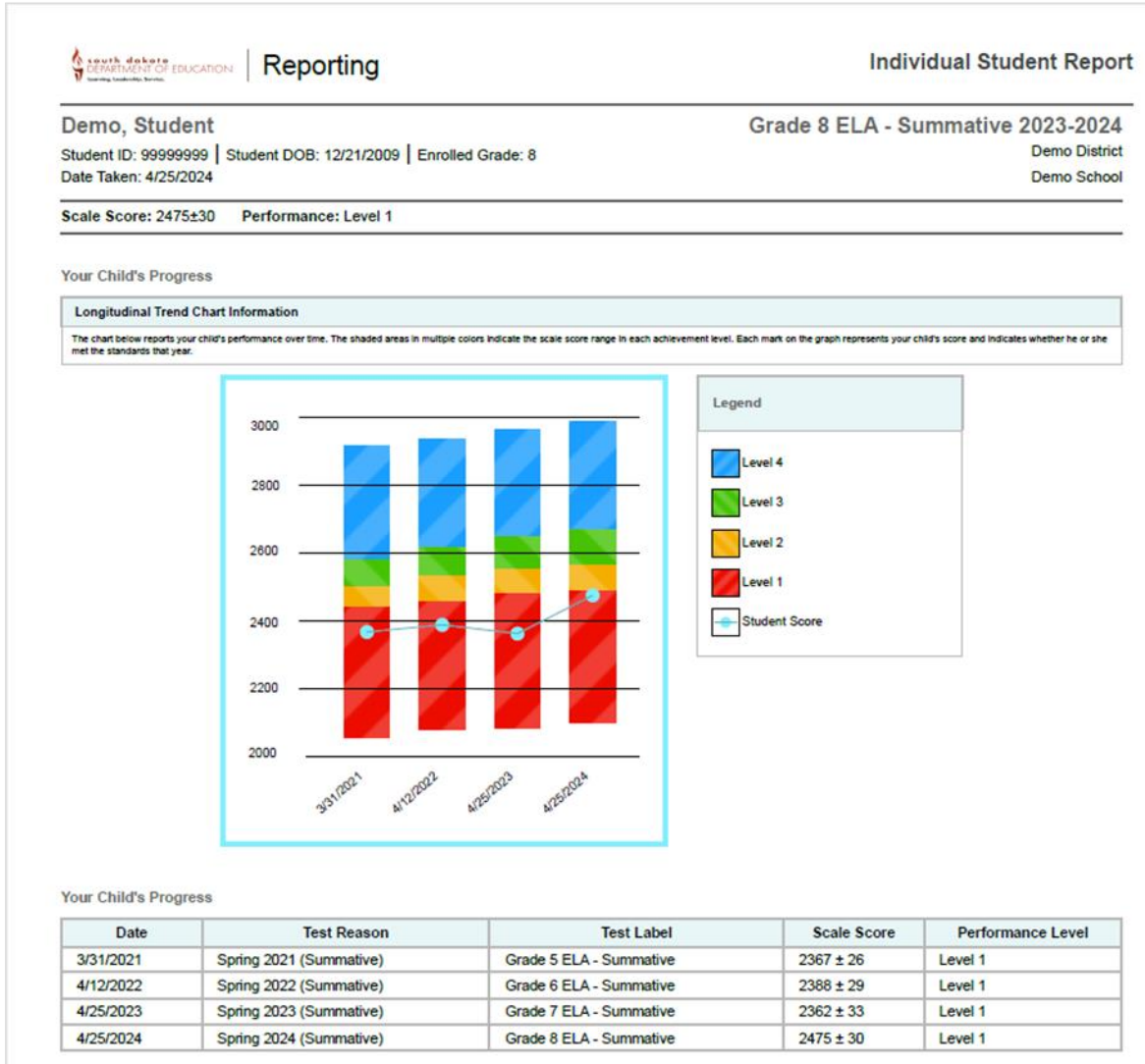
▲ Below Standard ▢ At/Near Standard ● Above Standard

Category	Performance	Performance	Performance Description
Listening		▢	<p>What These Results Mean Student may be able to employ effective listening skills for a range of purposes and audiences.</p> <p>Next Steps Have your child listen to a documentary and explain how presentation (graphics, tone of voice, music) relates to purpose. Point out when there is not enough evidence or when evidence is unrelated to the topic and why.</p>
Reading		▲	<p>What These Results Mean Student has difficulty reading closely and analytically to comprehend a range of increasingly complex literary and informational texts.</p> <p>Next Steps Have your child explain how authors build arguments or stories: How do they organize texts (the order of the ideas, the details used)? How do they connect ideas or plot elements (characters, settings, story line)?</p>
Research/Inquiry		▢	<p>What These Results Mean Student may be able to engage in research and inquiry to investigate topics, and to analyze, integrate, and present information.</p> <p>Next Steps Have your child conduct a short research project based on his or her questions on a topic. He or she should include several sides of the topic, combine data, use quotations from sources, and his or her own ideas.</p>
Writing		▢	<p>What These Results Mean Student may be able to produce effective and well-grounded writing for a range of purposes and audiences.</p> <p>Next Steps Help your child write argumentative essays, which address opposing views and include a counterclaim, logical reasoning, and support. All essays need direct quotations and formal, subject-specific language.</p>

How Did Your Child Perform on the Essay?

Essay	Raw Score	Conventions	Elaboration	Purpose
Argumentative	5 out of 10 points	The argumentative response shows a partial understanding of correct sentence formation, punctuation, capitalization, grammar usage, and spelling. (1 out of 2 points)	The argumentative response provides uneven elaboration to support the claim including few facts and details cited from sources, weak elaborative techniques and ineffective language for the audience and purpose. (2 out of 4 points)	The argumentative response has an inconsistent structure including an unclear claim, uneven development, few transitions, and loosely connected ideas. If present, the introduction or conclusion may be weak. The response may address the opposing argument. (2 out of 4 points)

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



7.2. INTERPRETATION OF REPORTED SCORES

Students' test performance is reported in a scale score and as an achievement level for the overall test and each claim. Students' scores and achievement levels are also summarized at the aggregate levels. The next section describes how to interpret these scores.

7.2.1. Scale Score

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

meaningful when the scale scores are used along with achievement levels and achievement-level descriptors (ALDs).

7.2.2. Conditional Standard Error of Measurement

A scale score (the observed score on any test) is an estimate of the true score. If a student takes a similar test multiple times, the resulting scale score will vary across administrations, sometimes being a little higher, a little lower, or the same. The standard error of measurement (SEM) represents the precision of the scale score, or the range in which the student would likely score if a similar test were 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 likely to contain the true score. For example, $2,680 \pm 10$ indicates that if a student were tested again, it is likely that he or she would receive a score between 2,670 and 2,690. The SEM can be different for the same scale score, depending on how closely the administered items match the student’s ability.

7.2.3. Achievement Level

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

7.2.4. 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 claim is evaluated with respect to the “Meets Standard” achievement standard. For students performing at “Below Standard” or “Above Standard,” this can be interpreted to mean that their performance is clearly below or above the “Meets Standard” cut score for a specific claim. For students performing at “At/Near Standard,” this can be interpreted to mean that their performance does not provide enough information to tell whether they reached the “Meets Standard” mark for the specific claim.

7.2.5. Performance Category for Targets

In addition to the claim-level reports, teachers and educators can ask for additional reports on student performance for instructional needs. Target-level reports are produced for the aggregate units only, not for individual students, because each student is administered too few items in a target to produce a reliable score for each target.

Target reports are produced for each target within a claim. Two types of relative strength and weakness scores for each target within a claim are reported. The strengths and weaknesses reports are generated for aggregate units of classrooms, schools, and districts and provide information about how a group of students in a class, school, or district performed on each target, either relative to the proficiency standard (i.e., "Proficient?" target measure) or their overall performance on the test (i.e., "Weak or Strong?" target measure).

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

For the "Proficient?" target measure, students' observed performance on items within the reporting element is compared to the expected performance on those items of someone who has an ability equal to the proficiency cut (i.e., the Achievement Level 3 cut). At the aggregate level, when observed performance within a target is greater than the proficiency cut, the reporting unit shows relative strength in that target compared to the proficiency standard. Conversely, when observed performance within a target is below the proficiency cut, the reporting unit shows a relative weakness in that target.

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

7.2.6. Aggregated Scale Score

Students' scale scores are aggregated at the 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 average scale scores can be interpreted as an estimate of the knowledge and skills that a group of students possesses. Given that student scale scores are estimates, the average scale scores are also average estimates and subject to measures of uncertainty. In addition to the average scale scores, the percentage of students in each achievement level overall and by claim are reported at the aggregate level to represent how well a group of students performs.

7.3. APPROPRIATE USES OF TEST RESULT

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

Assessment results on student achievement on the test can help teachers or schools make decisions on how best to support students' learning. Aggregate score reports at the teacher and school levels provide information regarding the strengths and weaknesses of their students and can be utilized to improve teaching and student learning. For example, a group of students may perform very well overall on the test but

potentially not perform as well in several targets compared to their overall performance. In this case, teachers and schools would be able to identify their students' strengths and weaknesses through the group performance by claim and target. They could then promote instruction in the specific claim or target areas in which their students performed relatively lower. Further, by narrowing down student performance results by subgroup, teachers and schools can determine which strategies may best improve student learning, particularly for students from disadvantaged subgroups. For example, teachers can examine student assessment results by LEP status and observe that LEP students need help in a particular area, such as reading literary responses and analysis. Teachers can then provide additional focused instruction for these students to enhance their achievement in any specific target or claim with which they are struggling.

In addition, assessment results can be used to compare performance among different students and groups. Teachers can evaluate how their students perform compared with other students in their school and district by overall scores and by claims. Although all students are administered different sets of items in each CAT, scale scores are comparable across students. Furthermore, scale scores can be used to measure the growth of individual students over time when data are available. In the South Dakota 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 to measure the growth. However, caution is advised, particularly when comparing scores across non-adjacent grade levels.

While assessment results provide valuable information to understand students' performance, these scores and reports should be used with caution. It is important to note that scale scores reported are estimates of true scores and hence do not represent the precise measure for student performance. A student's scale score is associated with measurement error, and thus users need to consider measurement error when using student scores to make decisions about student achievement. Moreover, although student scores may be used to 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 of student achievement such as classroom assessment and teacher evaluation should be considered when making decisions on student learning. Finally, when student performance is compared across groups, users need to consider the group size. The smaller the group size, the larger the measurement error related to these aggregate data, thus requiring interpretation with more caution.

8. QUALITY CONTROL PROCEDURE

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

8.1. ADAPTIVE TEST CONFIGURATION

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

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

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

8.1.1. Platform Review

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

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

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

8.1.2. User Acceptance Testing and Final Review

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

8.2. QUALITY ASSURANCE IN DOCUMENT PROCESSING

South Dakota summative assessments are administered primarily online; however, a few students took 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) is created to verify all possible responses and demographic grids, including typical errors that require 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 whether the output received from the scanners was correct. MI staff carefully compared the documents and the data file created from them to further ensure that results from the scanner, editing process (validation and data correction), and transfer to the CAI database are correct.

8.3. QUALITY ASSURANCE IN DATA PREPARATION

CAI's TDS has a real-time quality-monitoring component built-in. After a test is administered to a student, the TDS passes the resulting data to CAI's 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, and the total number of field-test items and operational items. The QA system also ensures that the test record does not contain 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 is the tool that pulls data from the DOR for delivery to the Department. CAI staff ensure that data in the extract files match the DOR before it is delivered.

8.4. QUALITY ASSURANCE IN ONLINE TEST DELIVERY SYSTEM

To monitor the performance of the TDS during the test administration window, CAI statisticians examine the delivery demands, including the number of tests to be delivered, the length of the test administration window, and the historic state-specific behaviors to model the likely peak loads. Using the data from the load tests, CAI can calculate the number of each type of server necessary to provide continuous, responsive service, and contracts for service exceeding this amount. Once deployed, the servers are monitored at the hardware, operating system, and software platform levels with software that alerts CAI engineers at the first signs that trouble may be ahead. The applications log errors, exceptions, and item response time information for crucial database calls. This information enables CAI to know instantly whether the system is performing as designed, starting to slow down, or experiencing a problem. In addition, item response time data are captured for each assessed student, such as data about how long it takes to load, view, or respond to an item. All this information is logged, enabling CAI to automatically identify schools or districts experiencing unusual slowdowns, often before they even notice.

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

For example, an item statistics analysis report allows psychometricians to ensure that items are performing as intended and serves as an empirical key check throughout the operational testing window. The item statistics analysis report monitors 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. These types of problems include incorrect designation of a keyed response or other scoring errors, and potential breaches of test security that may be indicated by changes in the difficulty of test items. This report generates classical item analysis indicators including the item p-value and item discrimination index, and IRT item fit statistics. The report is configurable and can be produced so that only items with statistics falling outside a specified range are flagged for reporting or generating reports based on all items in the pool.

For the CAT component, other reports such as blueprint match and item exposure reports allow psychometricians to verify that test administrations conform to the simulation results. The QA reports can be generated on any desired schedule. Item analysis and blueprint match reports are evaluated frequently at the opening of the testing window to ensure that test administrations conform to the blueprint and items are performing as anticipated. Table 70 presents an overview of the QA reports.

Table 70. Overview of Quality Assurance Reports

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

8.4.1. Score Report Quality Check

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

8.4.1.1 Online Report Quality Assurance

The systems automatically assign scores on the online assessments in real time. Every test undergoes a series of validation checks. Once the QA system signs off, data are passed to the DOR, which serves as the centralized location for all student scores and responses, ensuring that there is only one place where the official record is stored. Only after scores have passed the QA checks and are uploaded to the DOR are they passed to the Reporting System, which is responsible for presenting individual-level results and calculating and presenting aggregate results. Absolutely no score is reported in the Reporting System until it passes all

the QA system's validation checks. All of the previously mentioned processes take milliseconds to complete so that within less than one second after CAI receives handcores and they pass QA validation checks, the composite score will be available in the Reporting System.

8.4.1.2 Paper Report Quality Assurance

Statistical Programming

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

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

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

Display Programming

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

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

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

REFERENCES

- American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC.
- Billingsley, P. (1995). *Probability and Measure* (3rd ed.). New York, NY: John Wiley & Sons.
- Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38*(1), 67–86.
- Guo, F. (2006). Expected classification accuracy using the latent distribution. *Practical Assessment, Research & Evaluation*, *11*(6).
- Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement*, *13*(4), 253–264.
- Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, *32*(2), 179–197.
- Livingston, S. A., & Wingersky, M. S. (1979). Assessing the reliability of tests used to make pass/fail decisions. *Journal of Educational Measurement*, *16*(4), 247–260.
- Raczynski, K. R., Cohen, A. S., Engelhard, G., Jr., & Lu, Z. (2015). Comparing the effectiveness of self-paced and collaborative frame-of-reference training on rater accuracy in a large-scale writing assessment. *Journal of Educational Measurement*, *52*(3), 301–318.
- Snijders, T. A. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika*, *66*(3), 331–342.
- Sotaridona, L. S., Pornel, J. B., & Vallejo, A. (2003). Some applications of item response theory to testing. *The Philippine Statistician*, *52*(1–4), 81–92.
- Subkoviak, M. J. (1976). Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement*, *13*(4), 265–276.
- U.S. Department of Education. (2015). *Peer Review of State Assessment Systems: Non-Regulatory Guidance for States*. Washington, D.C. Retrieved from <https://www2.ed.gov/policy/elsec/guid/assessguid15.pdf>
- Williamson, D., Xi, X., & Breyer, J. (2012). A framework for the evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, *31*(1), 2–13.