Idaho Alternate Assessment (IDAA)

2022–2023 Technical Report

English Language Arts: Grades 3–8, 10

Mathematics: Grades 3–8, 10

Science: Grades 5, 8, and 11

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

TABLE OF CONTENTS

1. OVERVIEW 7
1.1 The Idaho Alternate Assessment 7
1.2 Intended Uses and Interpretations of IDAA Scores 8
1.3 Alternate Assessment Eligibility 8
1.3.1 IDAA Participation Criteria 9
1.3.2 IDAA Non-Participation Criteria 9
1.4 Content Standards 10
1.5 Memorandum of Understanding on Item-Sharing Initiative 11
2. TEST ADMINISTRATION 12
2.1 Testing Windows 12
2.2 Test Options and Administrative Roles 12
2.2.1 Administrative Roles 12
2.2.2 Online Administration 14
2.2.3 Paper-Pencil Test Administration 14
2.3 Training and Information for Test Coordinators and Test Administrators 15
2.3.1 In-Person and Online Trainings 15
2.4 Test Security 16
2.4.1 Student-Level Testing Confidentiality 16
2.4.2 System Security 16
2.4.3 Security of the Testing Environment 17
2.4.4 Test Security Violations 18
2.5 Online Testing Features and Testing Accommodations 19
2.5.1 IDAA Accessibility Features and Accommodations 19
2.6 Prevention and Recovery of Disruptions in Test Delivery System 23
2.6.1 High-Level System Architecture 23
2.6.2 Automated Backup and Recovery 25
2.6.3 Other Disruption Prevention and Recovery 25
3. SUMMARY OF THE SPRING 2023 OPERATIONAL ADMINISTRATION 26
3.1 Student Participation 26
3.2 Summary of Student Performance 31

3.3 Test-Taking Time 37

3.4 Distribution of Student Ability and Item Difficulty in the IDAA Item Pool 40

4. ITEM DEVELOPMENT 44

4.1 Item Development for the MOU-ALT 44

4.1.1 Item Type and Scoring Rubrics 46

4.1.2 Item Development Procedure and Item Reviews 47

4.1.3 Development of Crosswalk and State Alternate Performance Standards 49

4.2 Field Testing 50

4.2.1 Item Statistics 51

4.2.2 Classical Statistics 51

4.2.3 Item Response Theory Statistics 52

4.2.4 Analysis of Differential Item Functioning 52

4.2.5 Summary of Item Statistics 56

4.2.6 Data Review Meeting 59

4.3 Scaling and Equating 60

5. VALIDITY 62

5.1 Evidence Based on Test Content 63

5.2 Evidence Based on Response Processes 67

5.3 Evidence Based on Internal Structure 68

5.4 Evidence Based on Relations to Other Variables 71

5.4.1 Learner Characteristics Inventory 71

5.4.2 Correlations with LCI Descriptors 74

6. RELIABILITY 76

6.1 Marginal Reliability 76

6.2 Standard Error Curves 77

6.3 Reliability of Performance Classification 81

6.4 Reliability of Content Strand Scores 83

7. SCORING 87

7.1 Attemptedness Rules for Scoring 87

7.2 Estimating Student Ability Using Maximum Likelihood Estimation 87

7.3 Rules for Transforming Theta to Scale Scores 88

7.4 Lowest/Highest Obtainable Scale Scores 89
7.5 Scoring All Correct and All Incorrect Cases 89
8. PERFORMANCE STANDARDS 90
8.1 Standard-Setting Procedures 90
8.2 Performance-Level Descriptors 91
8.3 Recommended Performance Standards 91
9. REPORTING AND INTERPRETING SCORES 93
9.1 Centralized Reporting System for Students and Educators 93
9.1.1 Types of Online Score Reports 93
9.1.2 Centralized Reporting System 94
9.2 Interpretation of Reported Scores 99
9.2.1 Scale Score 99
9.2.2 Standard Error of Measurement 99
9.2.3 Performance Level 99
9.2.4 Aggregated Score 99
9.3 Appropriate Uses for Scores and Reports 100
10. QUALITY CONTROL PROCEDURES 101
10.1 Operational Test Configuration 101
10.1.1 Platform Review 101
10.1.2 User Acceptance Testing and Final Review 102
10.2 Quality Assurance in Data Preparation 102
10.3 Quality Assurance in Test Scoring 102
10.3.1 Score Report Quality Check 103
REFERENCES 105

LIST OF TABLES

Table 1. Number of Students Using Accessibility Features 22

Table 2. Number of Participating Students and Attempted Tests in the IDAA 26

Table 3. Overall Alternate Assessment Participation Rate 27

Table 4. Number of Participating Students by Subgroup 28

Table 5. Number of Participating Students by Subgroup and Disability Category (Grades 3–6) 29

Table 6. Number of Participating Students by Subgroup and Disability Category (Grades 7–10) 30

Table 7. Number of Participating Students by Subgroup and Disability Category (Grade 11).............. 31

Table 8. Student Performance Overall and by Subgroup—ELA (Grades 3–6)...................................... 32

Table 9. Student Performance Overall and by Subgroup—ELA (Grades 7–10).................................... 33

Table 10. Student Performance Overall and by Subgroup—Mathematics (Grades 3–6)....................... 34

Table 11. Student Performance Overall and by Subgroup—Mathematics (Grades 7–10) 35

Table 12. Student Performance Overall and by Subgroup—Science (Grades 5, 8, and 11)................... 36

Table 13. Test-Taking Time 37

Table 14. Number of Progress Indicators - ELA 44

Table 15. Number of Core Content Connectors - Mathematics 45

Table 16. Number of Progress Indicators - Science 45

Table 17. Summary of Spring 2023 Field-Test Item Pool Across MOU-ALT States............................. 51

Table 18. Thresholds for Flagging in Classical Item Analysis 52

Table 19. DIF Classification Rules... 54

Table 20. Number of Items in Each DIF Classification Category.. 55

Table 21. Summary of Item Analyses Results for MOU-ALT ELA.. 56

Table 22. Summary of Item Analyses Results for MOU-ALT Mathematics.. 57

Table 23. Summary of Item Analyses Results for MOU-ALT Science.. 58

Table 24. Summary of Item Data Review for MOU-ALT Item Pool 59

Table 25. Summary of IDAA Field-Test Item Pool.. 60

Table 26. Percentage of Administered Tests Meeting Blueprint Requirements for ELA 65

Table 27. Percentage of Administered Tests Meeting Blueprint Requirements for Mathematics 66

Table 28. Percentage of Administered Tests Meeting Blueprint Requirements for Science 67

Table 29. Correlations Among Strand Scores for ELA 69

Table 30. Correlations Among Strand Scores for Mathematics 70

Table 31. Correlations Among Strand Scores for Science 71

Table 32. Correlations Between LCI Descriptors and Total IDAA Scores 75

Table 33. Marginal Reliability for ELA, Mathematics, and Science 77

Table 34. Average Conditional Standard Error of Measurement by Performance Level 80

Table 35. Classification Accuracy and Consistency for Performance Standards 83

Table 36. Marginal Reliability Coefficients for Content Strand Scores for ELA 84

Table 37. Marginal Reliability Coefficients for Content Strand Scores for Mathematics 85

Table 38. Marginal Reliability Coefficients for Content Strand Scores for Science 86

Table 39. Scaling Constants on the Reporting Metric 88

Table 40. Range of Scale Scores by Performance Level 89

Table 41. Final Recommended Performance Standards for IDAA 92

Table 42. Types of Online Score Reports by Level of Aggregation 94

Table 43. Types of Subgroups 94

Table 44. Overview of Quality Assurance Reports 103

LIST OF FIGURES

Figure 1. Distribution of Testing Time—ELA 38

Figure 2. Distribution of Testing Time—Mathematics.. 39

Figure 3. Distribution of Testing Time—Science 40

Figure 4. Student Ability—Item Difficulty Distribution for ELA 41

Figure 5. Student Ability—Item Difficulty Distribution for Mathematics 42

Figure 6. Student Ability—Item Difficulty Distribution for Science 43

Figure 7. Alternate Assessment Item Development Process 46

Figure 8. Conditional Standard Error of Measurement for ELA 78

Figure 9. Conditional Standard Error of Measurement for Mathematics 79

Figure 10. Conditional Standard Error of Measurement for Science 80

LIST OF EXHIBITS

Exhibit 1. Home Page: State Level.. 95

Exhibit 2. Home Page: District Level... 95

Exhibit 3. Subject Detail Page for English Language Arts by Gender: District Level........................... 96

Exhibit 4. Student Detail Page for English Language Arts 97

Exhibit 5. Student Detail Page for Mathematics 98

Exhibit 6. Student Detail Page for Science... 98

1. OVERVIEW

This report provides a technical summary of the 2022–2023 Idaho Alternate Assessment (IDAA) in English Language Arts (ELA), mathematics, and science. The IDAA was administered in ELA and mathematics in grades 3–8 and 10, and science in grades 5, 8, and 11. The purpose of this technical report is to document the evidence supporting the claims made for how IDAA test scores may be interpreted. The report includes 10 sections, including all the evidence accrued about the technical quality of a testing system. The findings are based on IDAA data, including all aspects of the technical qualities described in the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 2014) and the requirements in *A state's guide to the U.S. Department of Education's assessment peer review process* (U.S. Department of Education, 2018).

Section 2, Test Administration, documents the test administration procedures, testing conditions (including accessibility tools and accommodations), and test security procedures for all test administrations. Section 3, Summary of the Spring 2023 Operational Administration, summarizes the results of the spring 2023 IDAA in ELA, mathematics, and science. Sections 2 and 3 summarize the test-taking student population, student performance on the assessments, and the time spent taking the assessments. Section 4, Item Development, describes the item development process, specifically the sequence of reviews that each item must pass through before being eligible for the IDAA test administration. This section also summarizes the field-test item analyses, data review, and procedures used to scale and calibrate field-test items. Section 5, Validity, provides validity evidence based on test content, responses processes, internal structure, and relations to other variables.

Section 6, Reliability, provides evidence for the reliability of the IDAA, including marginal reliability, standard error of measurement (SEM), and the reliability of performance standards. Section 7, Scoring, describes the scoring procedures used in producing scale scores and performance levels. Section 8, Performance Standards, describes the Idaho State Department of Education's (SDE) procedures for identifying and adopting performance standards for the IDAA. Section 9, Reporting and Interpreting Scores, provides a description of the score reporting system and interpretation of test scores. Section 10, Quality Control Procedures, provides an overview of the quality assurance (QA) processes that are used to ensure that all test development, administration, scoring, and reporting activities are conducted with fidelity to the developed procedures.

1.1 THE IDAHO ALTERNATE ASSESSMENT

The Idaho Alternate Assessment (IDAA) for students with significant cognitive disabilities comprises tests based on alternate (extended) academic content and performance standards (i.e., the Idaho Extended Content Standards). The purposes of the IDAA are to maximize access to the general education curriculum for students with significant cognitive disabilities, ensure that all students with disabilities are included in Idaho's statewide assessments, and ensure that these students are included in the educational accountability system. The results of the assessments can inform instruction in the classroom by providing data that guide decision making. The IDAA is for students with documented significant cognitive disabilities only, who require extensive direct individualized instruction and substantial support to achieve measurable gains in grade- and age-appropriate curricula, and whose learning is linked to the content standards. Typically, this represents 1% of the total student population.

In spring 2023, Idaho administered the IDAA at grades 3–8 and 10 in ELA and mathematics, and grades 5, 8, and 11 in science. The IDAA included operational and field-test items from both the multistate shared items and Idaho-specific items. All IDAA tests were delivered primarily through the computer-adaptive tests (CATs) assembled using Cambium Assessment, Inc.'s (CAI) adaptive testing algorithm that selected operational items to meet the blueprint and match students' abilities. Fixed-form tests were also available as an accommodation for students who cannot access the IDAA in the test delivery system. The paper test forms included a student test booklet, a stimulus booklet, and paper response option cards. The test content was also accessed and submitted in the online test delivery system. Multistate field-test items were embedded among operational items in the CATs and randomly selected from the field-test item pool.

1.2 Intended Uses and Interpretations of IDAA Scores

Development and design of the IDAA are reflected in a theory of action that begins by answering fundamental questions about the purpose, uses, interpretations, and outcomes of the test and integrates evidence comprising theoretical, logical, and empirical components.

The intended uses of the IDAA score include

- measuring the academic achievements and progress of students with significant cognitive disabilities in core content areas taught in school,
- measuring achievement and progress toward meeting the state performance standards for this student population, and
- monitoring the education system and making necessary improvements to meet federal accountability requirements.

Intended test users include students and parents who would like to be informed of the students' learning progress in school; teachers and other educators in school who can use testing results to guide in-class instruction and identify students who need more help; and educational agencies, organizations, and governments that monitor the education system and make necessary changes in standards.

To assist these uses, the IDAA provides an overall scale score and an associated performance level for each test taken. The performance level is determined based on the performance standards that are set through a formal standard-setting process. Validity evidence on measuring achievement and progress toward meeting the state performance standards is documented separately in greater detail in the standard setting technical report. Section 8, Performance Standards, provides a high-level overview of the standard-setting procedure and results.

1.3 Alternate Assessment Eligibility

The IDAA is designed for students with significant cognitive disabilities who participate in a school curriculum consistent with the grade-level Idaho Extended Content Standards, which are directly linked to the Idaho state content standards. The Idaho SDE provides a summary of the IDAA on the Idaho portal, which states, "The IDAA has been developed to ensure that all students with significant cognitive impairment [SCI] are able to participate in an assessment that is a measure of what they know and can do in relation to the grade-level Idaho Extended Content Standards Core Content Connectors."

1.3.1 IDAA Participation Criteria

The IDAA participation criteria were developed by a committee of Idaho educational partners in winter 2018. The four participation criteria and 14 non-participation criteria outlined below were adopted by the Idaho Legislature in spring 2018 as incorporated by reference into the 2018 Idaho Special Education Manual (p. 83 – 85).

To participate in the IDAA, as stated in the Idaho Alternate Assessment Participation Criteria, a student must meet all four of the following participation criteria, as determined by a student's IEP team:

1. The student has a significant cognitive impairment.
    - Impairment has significant impact on adaptive skills and intellectual functioning
    - Adaptive skills well below average in two or more areas
    - Intellectual functioning well below average (i.e., IQ typically below 55)
2. The student is receiving academic instruction that is aligned with the Idaho Extended Content Standards.
    - The student's instruction and Individualized Education Program (IEP) goals, objectives, and benchmarks address knowledge and skills that are appropriate and challenging for the student.
3. The student's course of study is primarily adaptive-skills oriented, and typically not measured by state or district assessments.
    - Adaptive skills are essential to living independently and functioning safely in daily life, and include, but are not limited to, motor skills, socialization, communication, personal care, self-direction, functional academics, and personal health and safety.
4. The student requires extensive, direct, individualized instruction and substantial supports to achieve measurable gains in the grade- and age-appropriate curriculum.
    - The student consistently requires individualized instruction in core academic and adaptive skills at a substantially lower level relative to other peers with disabilities.
    - It is extremely difficult for the student to acquire, maintain, generalize, and apply academic and adaptive skills in multiple settings, across all content areas, even with high-quality extensive, intensive, pervasive, frequent, and individualized instruction.
    - The student requires pervasive supports, substantially adapted materials, and individualized methods of accessing information in alternative ways to acquire, maintain, generalize, demonstrate, and transfer skills across multiple settings.

1.3.2 IDAA Non-Participation Criteria

Students will not qualify to participate in the alternate assessments based on Alternate Achievement Standards solely based on any of the following 14 reasons:

1. Having a disability
2. Poor attendance or extended absences

3. Native language or social, cultural, or economic differences
4. Expected poor performance, or past basic or below basic performance on the regular education assessment
5. Academic and other services a student receives
6. Educational environment or instructional setting
7. Percentage of time receiving special education services
8. English learner (EL) status
9. Low reading level or achievement level
10. Anticipated disruptive behavior
11. Impact of student scores on the accountability system
12. Administration decision
13. Anticipated emotional distress
14. Need for accommodations (e.g., assistive technology, Augmentative and Alternate Communication [AAC]) to participate in the assessment

1.4 CONTENT STANDARDS

The Idaho Extended Content Standards that align with the Idaho state content standards were designed to make the IDAA more accessible to students with significant cognitive disabilities while maintaining the rigor and high expectations of Idaho State Content Standards. These standards ensure that all students with significant cognitive disabilities are provided with multiple ways to learn and demonstrate knowledge. The Idaho Extended Content Standards are posted on the Idaho Content Standards webpage (Idaho Content Standards / Content and Curriculum / SDE) under each content area dropdown menu.

The process of ensuring the Idaho Extended Content Standards is appropriately aligned to the Idaho Academic Content Standards began with Idaho's involvement in the National Center and State Collaborative (NCSC). Representatives from NCSC states developed the Core Content Connectors in direct alignment with the Common Core Standards used by all member states. Following this process, the former State Department of Education Director of Assessment and Accountability reviewed and cross-walked the Core Content Connectors with the Idaho Academic Content Standards and made minor adjustments to reflect Idaho-specific numbering conventions and verbiage. The resulting document became the Idaho Extended Content Standards, which went through the Idaho negotiated rule process, including an extensive public feedback period, and were adopted by the Idaho Legislature as incorporated by reference in 2018.

The Idaho Extended Content Standards are aligned with the Idaho Content Standards, but have been reduced in depth and complexity, which is suitable for students who are eligible for participation in the IDAA. The Idaho Extended Content Standards are referred to as Core Content Connectors at the standard level. This indicates that the Idaho Extended Content Standards are composed of the 'core content' of the Idaho Content Standards.

The IDAA item bank is comprised of items at different levels of complexity in order to test across the cognitive abilities in this population of students. This meets the requirements of both the Individuals with Disabilities Education Act (IDEA) and Every Student Succeeds Act (ESSA) to link assessments to grade-level standards, with the understanding that alternate assessments may include skills at lower levels of complexity.

The Idaho Extended Content Standards have been further parsed into Performance-Level Descriptors (PLDs) for ELA, mathematics, and science. The PLDs were created by CAI and reviewed by SDE before they were taken to the Standards Setting workshop in summer 2022. Representatives from each content area standard setting committee, all of whom were currenting practicing educators, made revisions to the PLDs with implementation in mind. The final PLDs were adopted by the Idaho State Board of Education in October 2022.

PLDs have been developed at four levels of cognitive complexity, which reflect the four achievement levels used for the Idaho Standards Achievement Tests, as outlined below:

- Advanced Performance Level — Highest level of performance expectation for the alternate test
- Proficient Performance Level — Proficient performance expectation for the alternate test
- Basic Performance Level — Basic performance expectation for the alternate test
- Below Basic Performance Level — Below Basic performance expectation for the alternate test

PLDs reflect different entry points into the grade-level state standards for students with significant cognitive disabilities and serve the following three purposes: 1) to assist teachers in providing access to the academic standards for students with significant cognitive disabilities, 2) to assist assessment personnel in developing test items that are accessible for students with a range of skill levels, and 3) to be used by standard setting committees in conjunction with Essence Statements to craft the Just Barely Statements, which describe what a student just barely scoring at the bottom of each performance level knows and can do, and the Reporting PLDs, which detail grade- and content-area-specific descriptions of exactly what students performing throughout the range of each performance level know and can do.

1.5 Memorandum of Understanding on Item-Sharing Initiative

In 2018, Hawaii, South Carolina, and Wyoming signed a Memorandum of Understanding (MOU) on item sharing in item development and field testing in ELA, mathematics, and science. Each state contributed a predetermined number of items proportional to their state's student population for alternate assessment in these three content areas. In early 2019, Idaho and Vermont joined the collaborative item development and field-testing agreement and participated in the spring 2019 field test in ELA, mathematics, and science. In spring 2020, Montana and South Dakota joined the MOU for science assessments. In 2022, Vermont exited the MOU. Because the total number of students in alternate assessments in each state was small, field-test items were calibrated based on the combined data across all MOU states. In addition to the shared MOU items, each state also developed items that align only with that state's content standards.

The item-sharing initiative was designed to implement an item development process generating enough items so that, for each grade and subject, the pool had at least three times the number of items required for each test administration in the beginning. With 40 operational items on the test, the goal was to have at least 120 calibrated items (40 X 3 forms) in the initial pool for the CAT. The items were developed using the item-sharing initiative, which stated that each MOU member would own the items they developed, but those items would be available for use by the other MOU members. The number of items to be developed by each state was proportional to their alternate assessment population size.

2. TEST ADMINISTRATION

2.1 Testing Windows

The spring 2023 Idaho Alternate Assessment (IDAA) testing window spanned two months, from March 13–May 12, 2023.

2.2 Test Options and Administrative Roles

The IDAA is primarily administered through online computer adaptive tests (CATs). If a student has a paper accommodation documented in their Individualized Education Program (IEP), they are eligible for paper-accommodated test materials. The paper version of the IDAA is a fixed form with an item booklet, a stimulus booklet, and printed response option cards. Qualifying students access IDAA test content in the secure browser, then the teacher presents the item booklet, stimulus booklet, and / or printed response option cards so the student can choose their answer. The IDAA paper test forms are intended for a very small number of students who cannot access the IDAA in the test delivery system; mainly students with seizure disorders or other conditions that prevent them from interacting with the computer. Schools can order paper-accommodated test materials that accompany an online fixed-form version of the assessments.

Test administrators (TAs) follow procedures outlined in the *Summative Test Administration Manual* (*TAM*) to ensure that standardized administration conditions are met. TAs must review the *TAM* before testing to ensure that the testing room is prepared (e.g., removing certain classroom posters, arranging desks). TAs must follow required administration procedures and directions.

2.2.1 Administrative Roles

The key personnel involved with test administration are district administrators (DAs), district coordinators (DCs), school coordinators (SCs), teachers (TEs), and TAs. The primary responsibilities of the key personnel are described in this section. More detailed descriptions can be found in the online *Summative TAM*.

Before the IDAA test administration, each DA, DC, SC, TE, and TA should review the *Summative TAM* to become familiar with the responsibilities of all parties.

District Administrator Responsibilities

DAs are assigned by the State. If assigned, a DA can upload, add, modify, and delete student records. The DA can also add DCs, SCs, TEs, TAs, District Instructional Supports (DISs), and Tools for Teachers—District and School roles (TFT_Ds and TFT_SCs) into the Test Information Distribution Engine (TIDE).

District Coordinator Responsibilities

DC responsibilities include the following:

- Adding SCs, TEs, TAs, DISs, TFT_Ds, and TFT_SCs into TIDE
- Ensuring that the SCs, TEs, and TAs in their districts are appropriately trained regarding the assessment administrations and security policies and procedures
- Reporting test security incidents to the State via the Test Improprieties module in TIDE and the Test Security Incidents Log

- Providing general oversight for all administration activities in their district/schools
- Entering and verifying test settings (i.e., Designated Supports and Accommodations) for students

School Coordinator Responsibilities

SC responsibilities include the following:

- Identifying TAs and ensuring that they are properly trained
- Adding TEs, TAs, and TFT_SCs into TIDE
- Coordinating with TAs so they administer all assessments
- Entering and verifying student test settings
- Creating or approving testing schedules and procedures for the school in a manner consistent with state and district policies
- Working with technology staff to ensure that necessary Idaho Secure Browsers are installed and that any other technical issues are resolved
- Monitoring testing progress during the testing window and ensuring that all students participate, as appropriate
- Addressing testing incidents, as needed
- Mitigating and reporting all test security incidents in a manner consistent with state and district policies

Teacher Responsibilities

TE responsibilities include the following:

- Completing assessment administration training and reviewing all Smarter Balanced, state, and district policy and administration documents prior to administering any assessments
- Viewing student information prior to testing to ensure that the correct student receives the proper test with the appropriate supports (TEs should report any potential data errors to SCs and DCs as appropriate)
- Administering the assessments under certain circumstances
- Reporting all potential test security incidents to their SC and DC in a manner consistent with state and district policies

Test Administrator Responsibilities

TA responsibilities include the following:

- Completing IDAA assessment administration training and reviewing all state, district policy, and administration documents prior to administering any assessments
- Viewing student information prior to testing to ensure that the correct student receives the proper test with the appropriate supports (TAs should report any potential data errors to SCs and DCs as appropriate)
- Administering the assessments under certain circumstances
- Reporting all potential test security incidents to their SC and DC in a manner consistent with state and district policies

2.2.2 Online Administration

Schools can set the testing schedule and customize the testing conditions within the state's testing window. For example, schools can allow students to test in intervals (i.e., multiple sessions). Schools are discouraged from testing in one long period, minimizing the interruption of classroom instruction and test fatigue. 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 points of contact. TAs administer the online assessments one-on-one with a student. TAs are trained in the online testing requirements and the mechanics of starting, pausing, and ending a test session. Training materials for the test administration are provided online. All school personnel who serve as TAs must complete an online TA Certification Course to receive a certificate of completion and administer assessments.

To start a test session, the TA must first enter the TA Interface of the online Test Delivery System (TDS) using his or her computer. A session ID is generated when the test session is created. Students taking the assessment with the TA must enter their EDUID, first name, and session ID into the Student Interface using computers provided by the school. The TA then verifies that the students are taking the appropriate assessments with the appropriate accessibility features (refer to Section 2.5, Prevention and Recovery of Disruptions in Test Delivery System, for a list of accommodations). Students can begin testing after the TA confirms that the settings are appropriate. The TA may aid the students through the login process.

Once an assessment has started, the student must answer the test question presented on the page before proceeding to the next page. Skipping questions is not permitted. For the online test, students may review and edit the responses they have previously completed before submitting that segment of the assessment.

The operational items are administered in two segments with the first segment comprised of four items and the second segment comprised of the remaining items. Segment 1, also referred to as the early stopping rule segment, is to identify students who do not have a consistent and observable mode of communication. If a student moves through the first segment and does not respond to four consecutive items, which are then marked "No Response" by the TA, the Early Stopping Rule is implemented. The TA must confirm the early-stopped records (ESR) should be implemented, and the test ends.

During a test session, TAs may pause the test for the student to take a break. It is up to the TA to determine an appropriate stopping point. When continuing testing on a different day, the TA must start a new test session, and the student will resume the test from the point where he or she paused.

The TA must always remain seated next to the student during a test session to monitor student testing. The test is administered one-on-one to the student by the TA; therefore, no observers are allowed other than translators, interpreters, and student aides. The TA must ensure that each student has successfully logged out of the system when the test session ends.

2.2.3 Paper-Pencil Test Administration

DCs or DAs need to flag students who need paper accommodations (i.e., students with limited vision, students who are blind, students who benefit from manipulative response options) as an "Alternate Assessment Paper Tester" in TIDE, and then provide them with paper-accommodated test materials. This process gives students access to a fixed form. The student will still use the Student Interface to complete the test, but rather than viewing the response options on the screen, the student will use paper response options that accompany the fixed form to select an answer. The TA will then assist the student in selecting his or her response on the computer. TAs might also serve as scribes, as is often the case with the printed response option cards and paper test forms. DCs are responsible for ordering paper accommodated test materials based on the number of students who need this accommodation.

2.3 Training and Information for Test Coordinators and Test Administrators

All SCs oversee all aspects of testing at their schools and serve as the main points of contact. TAs administer the online assessments. The in-person trainings, online TA Certification Course, user guides, and manuals are used to teach TAs about the online testing requirements and the mechanics of starting, pausing, and ending a test session. Training materials for the administration are provided online and in person.

2.3.1 In-Person and Online Trainings

Districts assume all responsibility for a non-standard test administration or a testing irregularity resulting in a test invalidation due to administration error (e.g., unexpected interruptions that impact students while testing). As with all statewide testing, districts must provide annual training on test security and standards for the ethical use of tests to all employees who have access to state tests and access to the students who are administered the state tests.

TA Certification Course

Each year, all TAs must complete an online TA Certification Course to be approved to administer assessments. The TA Certification Course must be taken (and passed) by all users that will administer an in-person summative assessment. Users learn how to log in to the TA Interface, start a test session, approve students to test, pause and stop a session, and access the mobile interface. They will also learn how a student logs in. The course is complete with audio and visual instructions; interactive slides that allow for guided practice; and multiple-choice questions. A user who completes the course will obtain a printable certificate of completion. The course can be taken as many times as needed.

Manuals and User Guides

The following manuals and user guides are available on the Idaho Assessment Program Portal:

The *Summative Test Administration Manual* outlines how DCs and TAs should prepare for and administer the IDAA. This manual includes participation requirements, an overview of the assessment, and detailed instructions for administration.

The *TIDE User Guide* is designed to help users navigate TIDE. This guide provides information on managing user account information, student account information, student test settings and tools, appeals, rosters, and voice packs.

The *Reporting System User Guide* provides information on using the Centralized Reporting System to view student performance information for the IDAA.

The *IDAA Participation Criteria* outlines criteria for student participation in the IDAA (decisions regarding participation should be made with the student's IEP team).

The *Data Entry Interface User Guide* covers general and training information specific to the IDAA.

2.4 TEST SECURITY

The security of assessment instruments and the confidentiality of student information are vital to maintaining the results' validity, reliability, and fairness. 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 trainings and in the user guides and manuals. Features in the testing system also protect test security. This section describes student confidentiality, system security, testing environment security, and policies on testing incidents.

2.4.1 Student-Level Testing Confidentiality

All secured websites and software systems enforce role-based security models that protect individual privacy and confidentiality in a manner consistent with the Family Educational Rights and Privacy Act (FERPA) and other federal 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. The Idaho systems use role-based security models that ensure that users may access only the data to which they are entitled and may edit data only in accordance with their user rights.

The following three elements are involved in ensuring that students are accessing appropriate test content:

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

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

- Providing login information (usernames and passwords) to other authorized TIDE users or unauthorized individuals
- Sending a student's name and EDUID number together in an email message
- Having students log in and test under another student's EDUID number

Test materials and score reports should not be exposed to identify student names with test scores except by authorized individuals with an appropriate need to know. If information about a test must be sent via email or fax, administrators must include only the EDUID number, not the student's name.

All students must be enrolled or registered to take the IDAA at their testing schools. Student enrollment information, including demographic data, is generated using an SDE file and is uploaded nightly via a Secure File Transfer Protocol (SFTP) site to the online testing system during the testing period.

2.4.2 System Security

The objective of system security is to ensure that all data are kept protected and are accessed only by the appropriate user groups. The end goals for system security are to protect and maintain data and system integrity, safeguarding personal information, and ensuring that data transfer is accurate and that users are assigned the appropriate level of user access.

Hierarchy of Control

As described in Section 2.2.1, Administrative Roles, DAs, DCs, SCs, TEs, and TAs have well-defined roles and levels of access to the testing system. DCs are responsible for selecting and entering the SCs' information into TIDE; SCs are responsible for entering the TAs' information in TIDE. Throughout the year, the DC and SC are also expected to delete data in TIDE for any staff members who have transferred to other schools, have resigned, or no longer serve as TAs.

Password Protection

All access points by different roles (i.e., state, district, school-principal, and school-staff levels) require a password to log in to the system. Newly added SCs and TAs receive separate passwords through their personal, school-assigned email addresses.

CAI Secure Browser

Developed by CAI, the Secure Browser prevents students from accessing other computers or Internet applications and copying test information while testing. It suppresses access to commonly used browsers such as Internet Explorer and Firefox and prevents students from searching for answers on the Internet or communicating with other students. The assessments can be accessed only through the CAI Secure Browser and not by other Internet browsers. Technology coordinators (TCs) are tasked with ensuring that the CAI Secure Browser is installed properly on the computers used for administering the online assessments.

2.4.3 Security of the Testing Environment

Maintaining test security is an important responsibility of personnel involved in alternate assessment administration. SCs and TAs are required to follow their district's written procedures for protecting the security of test materials at all times. Paper response options must be kept secure at all times.

Unlike the general assessment, the alternate assessment allows for TA support during testing and, in the case of using paper accommodations, requires that the TAs review the test materials before administering the assessment. However, it is illegal and unethical to reproduce or disclose any secure materials. Each test contains materials that will be used on future tests. Therefore, security is vital for current and future test administrations, and TAs are responsible for ensuring the security of the test materials, which is required even when materials are returned.

SCs are responsible for maintaining the security of all paper accommodations while they are in the SCs' possession. SCs are also responsible for ensuring that the TAs act in accordance with all security requirements while TAs are in possession of paper accommodations. Paper test materials should be kept in a locked, secure location with limited access when they are not in use. Only individuals authorized by school policy should have access to these materials. The Test Security Agreement must be reviewed and signed by each TA following their security training. It is the responsibility of the building test coordinator (BC) or principal to retain the signed agreements for at least two years; the agreements may be stored electronically.

Some examples of test security violations, as outlined in the Assessment Integrity Guide, may include, but are not limited to, the following:

- Giving any student access to secure test materials except in the regular course of an authorized administration of the state assessment system
- Giving unauthorized individuals or other persons access to secure test materials
- Copying, reproducing, using, or otherwise disclosing in any manner inconsistent with test security regulations and procedures for any portion of secure test materials
- Providing answers orally, in writing, or by any other means to any student during the administration of the test
- Coaching any student during testing by giving the student answers to secure test questions, otherwise directing or guiding a response, or altering or interfering with the student's response in any way
- Failing to follow security regulations and procedures for the storage, distribution, collection, and return of secure test materials, or failing to account for all secure test materials before, during, and after testing
- Failing to monitor the test administration properly or failing to return materials used by the students during testing
- Emailing, faxing, or inappropriately reproducing any student identification numbers associated with student names or other personally identifiable information (PII)
- Producing unauthorized printed copies of test materials, failing to properly destroy printed copies as authorized, or allowing printed copies to leave the test site
- Allowing tests to be administered by unauthorized personnel
- Participating in, directing, aiding, counseling, assisting, encouraging, or failing to report any of the prohibited acts
- Refusing to disclose information regarding test security violations
- Refusing to cooperate in investigating a suspected breach of test security, whether this investigation is conducted by a school district, the SDE, or others (the investigation will include a review of mitigating circumstances, if applicable)
- Changing students' incorrect answers to correct answers
- Discussing test questions with other people
- Taking home test materials
- Emailing information to anyone regarding the content of a test

If, at any time, a TA believes that a test security violation has occurred, they should contact their SC and follow the procedures established by their school district to handle the alleged test security violation.

2.4.4 Test Security Violations

Everyone who administers or proctors the assessments is responsible for understanding the security procedures for administering the assessments. Prohibited practices, as detailed in the TAM, are categorized into the following three categories:

1. **Impropriety.** This is a test security incident that has a low impact on the individual student or group of students who are testing and has a low risk of potentially affecting student performance on the test, test security, or test validity (e.g., students leaving the testing room without authorization).
2. **Irregularity.** This is a test security incident that affects an individual student or group of students who are testing and may potentially affect student performance, test security, or test

validity (e.g., disruption during the test session such as a fire drill). These circumstances can be contained at the local level.

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

District and school personnel are required to document all test security incidents in the test security incident log. This log is the document of record for all test security incidents and should be maintained at the district level and submitted to the SDE at the end of testing.

2.5 ONLINE TESTING FEATURES AND TESTING ACCOMMODATIONS

The IDAA contains universal tools 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 of that system.

State-level users, DCs, SCs, and TAs can set embedded and non-embedded accommodations based on their user role in TIDE. Accommodations must be set in TIDE before starting a test session.

All embedded and non-embedded universal tools will be activated for use by all students during a test session. One or more of the pre-selected universal tools can be deactivated by a DC or SC in TIDE before a student is tested or by a TA in the TA Interface of the testing system for a student who may be distracted by the ability to access a specific tool during a test session.

2.5.1 IDAA Accessibility Features and Accommodations

Accessibility features and accommodations are access features of an assessment that are embedded or non-embedded components of the test administration system. Accessibility features and accommodations were available to all students based on their preference and selection and were preset in TIDE. In the 2022–2023 test administration, the features of accessibility features and accommodations described in this section were available for all students to access.

2.5.1.1 Embedded Accessibility Features

Color Contrast. This feature allows for different background or font color, based on student needs or preferences. Available options are: black on white (default), black on magenta, yellow on blue, medium gray on light gray, and reverse contrast (white on black).

Highlighter. A digital tool for marking all or parts of desired text, item questions, and item answers in yellow.

Human Voice Recording (HVR). Text is read aloud to the student via embedded HVR technology. The TE or student activates the HVR by clicking the ear icon.

IDAA Fixed Form. The IDAA Fixed Form should be administered to students who are deaf or hard of hearing. Although the student will access test content in the online test delivery system, the “IDAA Paper Tester” flag must be checked in TIDE before the IDAA Fixed Form will appear in the test delivery system.

Language Presentation. Language selection (English) for IDAA tests.

Line Reader Tool. Allows students to highlight an individual line of text in a passage or question**.**

Mark for Review. Allows students to flag items for future review during the assessment.

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

Permissive Mode. An accommodation option that allows students to use accessibility software in addition to the secure browser.

Print on Demand. Items can be printed for students from the Student Interface in the TDS. This feature it not intended to be used to administer a paper-pencil test to students.

Print Size/ Zoom. A tool for making text or other graphics in a window or frame appear larger on the screen. The test page can zoom in up to five levels.

Streamlined Mode. This tool will be required when setting additional print size/zoom levels (5x–20x).

Strikethrough. Allows users to cross out answer options.

Volume Control. Audio can be controlled for embedded HVRs.

Non-Embedded Accessibility Features

Assistive Technology. Hardware and software tools used to increase, maintain, or improve the functional capabilities of children with disabilities.

Augmentative and Alternative Communication (AAC). Forms of communication used to supplement or replace oral speech that are used to express thoughts, needs, wants, and ideas. These systems of communication may be aided or unaided.

Hand-held Calculator. Familiar hand-held calculator with the same functions as those available on the online calculator.

Mathematics Manipulatives. Mathematics materials, such as counters or other concrete materials, that a student might use to solve mathematic equations and/or problems. If a student regularly uses manipulatives to solve math problems, those manipulatives should be made available for students during testing.

Scribe. A scribe enters the student-selected response on behalf of the student. Trained TE may enter student responses in the Idaho Secure Browser for the student as such: Student is unable to control the mouse to click an answer. Student may use an alternate mode of communication to indicate their answer choice, and student does not use the mouse with intention. Student may use an alternate mode of communication to indicate their answer choice. When administering the fixed-form test with printed response option cards, the trained TE records student responses in the Idaho Secure Browser.

2.5.1.2 Accommodations and Fixed Forms

Accommodations are changes in procedures or materials that increase equitable access during testing. 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 only. Students using these accommodations must also take the IDAA Fixed Form.

Non-Embedded Accommodations

Printed Response Option Cards. Printed Response Option Cards are printed cards that correspond with the answer options for each test item. They are intended for students who need to manipulate and/or interact with printed cards to indicate their answer choice. They would be appropriate for students who use an augmentative or alternate communication system; such as a picture exchange communication system, a communication device, an eye-gaze board to communicate, etc. The Printed Response Option Cards are used with the IDAA Fixed Form.

Read Aloud by Familiar Adult. Some students may need IDAA test items read-aloud by a familiar adult, as opposed to relying on the HVRs to access test content. The TE will first play all HVRs, then read text and describe images, tables, etc. as modeled by the HVR. The TE will use the test administration script based on the test form administered to the student.

Paper Test Form. The IDAA paper test forms are intended for a very small number of students who cannot access the IDAA in the test delivery system; mainly students with seizure disorders or other conditions that prevent them from interacting with the computer. The paper test forms include a student test booklet and a stimulus booklet.

Table 1 presents the number of students using accessibility features in the 2022–2023 administration.

Table 1. Number of Students Using Accessibility Features

Accessibility Features	**Grade**						
	3	**4**	**5**	**6**	**7**	**8**	**10/11**
English Language Arts							
Color Choices (Black on Rose)		1		1			
Color Choices (Yellow on Black)					1		
Highlight (On with multiple colors)	9	10	15	7	6	6	10
Permissive Mode (On)	84	104	87	100	99	83	92
Print Size (1.5X)				1			
Print Size (1.75X)				1			
Print Size (3X)	1	1			1		1
Print Size (5X (Streamlined Mode required))				1			
Streamlined Mode (On)	83	104	84	94	99	83	93
Strikethrough (Enhanced Strikethrough)	9	10	15	7	6	6	10
Non-Embedded Accommodations (Read Aloud)	25	27	20	34	29	28	6
Non-Embedded Designated Supports (Amplification)		2				3	
Non-Embedded Designated Supports (Medical Device)		1		1			
Non-Embedded Designated Supports (Noise Buffers)	6	2	9	8	4	6	1
Mathematics							
Color Choices (Black on Rose)				1			
Highlight (On with multiple colors)	9	11	15	8	6	7	8
Permissive Mode (On)	83	101	83	98	98	83	87
Print Size (1.5X)				1			
Print Size (1.75X)				1			
Print Size (3X)	1	1					1
Print Size (5X (Streamlined Mode required))				1			
Streamlined Mode (On)	82	101	83	93	97	83	88
Strikethrough (Enhanced Strikethrough)	9	11	15	8	6	7	8
Non-Embedded Accommodations (Read Aloud)	25	26	19	34	27	28	6
Non-Embedded Designated Supports (Amplification)		3				3	
Non-Embedded Designated Supports (Calculator)	5	5	12	11	4	16	3
Non-Embedded Designated Supports (Medical Device)		1	1	1			
Non-Embedded Designated Supports (Noise Buffers)	6	2	9	8	4	4	2
Science							
Highlight (On with multiple colors)			15			7	17
Masking (On)			94			88	93
Permissive Mode (On)			84			82	84
Print on Request (On)			95			87	91
Streamlined Mode (On)			84			82	84
Strikethrough (Enhanced Strikethrough)			15			7	17
Non-Embedded Accommodations (Read Aloud)			17			28	8
Non-Embedded Designated Supports (Amplification)			1			3	
Non-Embedded Designated Supports (Calculator)			8			14	3
Non-Embedded Designated Supports (Noise Buffers)			9			6	17

2.6 PREVENTION AND RECOVERY OF DISRUPTIONS IN TEST DELIVERY SYSTEM

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

CAI has developed a unique monitoring system that is sensitive to changes in server performance. Most monitoring systems provide warnings when something is going wrong; our monitoring system also provides warnings when any server performs differently from its performance over the few hours prior or from the other servers performing the same jobs. Subtle changes in performance often precede actual failure by hours or days, allowing us to detect potential problems, investigate them, and mitigate them before a failure. On multiple occasions, this has enabled us to adjust and replace equipment before any problems occurred.

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

Section 2.6.1, High-Level System Architecture, describes CAI system architecture and how it recovers from device failures, Internet interruptions, and other problems.

2.6.1 High-Level System Architecture

CAI system architecture provides the redundancy, robustness, and reliability required by a large-scale, high-stake testing program. The general approach is pragmatic and well supported by the architecture.

Any system built around an expectation of flawless performance of computers or networks within schools and districts is bound to fail. The CAI 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. Key elements of the testing system, including the data integrity processes at work at each point in the system, are described in this section. Fault tolerance and automated recovery are built into every component of the system.

Student Machine

Student responses are conveyed to our servers in real time as students respond. 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 proceeding 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, the student may be unaware of the momentary interruption.
- If connectivity cannot be silently restored, the student is prevented from testing and given the option of logging out or retrying to save.
- If the system fails, upon logging back into the system, the student returns to the item at which the failure occurred.

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

Test Delivery Satellites

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

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

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

Hub

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

Demographic and History Servers

The demographic and history servers store student data for the duration of the testing window. They are clustered database servers, also equipped with RAID subsystems, providing redundant capability to prevent data loss in the event of server or disk failure. At the normal conclusion of a test, these servers receive completed tests from the test delivery satellites. Upon successful completion of information storage, 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, monitors real-time item function, and evaluates test integrity. Every completed test runs through the QA system, and any anomalies (e.g., unscored or missing items, unexpected test lengths, other unlikely issues) are flagged, and a notification immediately goes out to our psychometricians and project team.

Database of Record

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

2.6.2 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 the loss of student data. Redundant systems at every point; real-time data integrity protection and checks; and well-considered real-time backup processes prevent the loss of student data, even in the unlikely event of system failure.

2.6.3 Other Disruption Prevention and Recovery

These testing systems are designed to be extremely fault-tolerant and can withstand the failure of any component with little to no interruption. The robustness is achieved through redundancy. Key redundant systems include the following attributes:

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

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

3. SUMMARY OF THE SPRING 2023 OPERATIONAL ADMINISTRATION

3.1 STUDENT PARTICIPATION

The spring 2023 Idaho Alternate Assessment (IDAA) was administered by subject and grade level. All students in grades 3–8 and 10 were assessed in English language arts (ELA) and mathematics. Students in grades 5, 8, and 11 were also assessed in science. For a test to have been considered "attempted" for scoring, a student must have responded to at least one item, or a test administrator (TA) must have marked "No Response" on at least one item. Table 2 presents the number of students who participated in the online adaptive tests and online fixed-form tests with accommodations by subject and grade. A test was marked "incomplete" if a student did not reach the end of the test. As shown, most eligible students completed the tests in spring 2023. Table 3 presents the alternate assessment participation rate, computed as the number of students taking the IDAA divided by the total number of students in the state taking the general education summative test and the IDAA. Table 4 presents the total number of students who participated by subgroups. Table 5 to Table 7 provide the total number of students who participated by the Individuals with Disabilities Education Act's (IDEA) disability categories.

Table 2. Number of Participating Students and Attempted Tests in the IDAA

Subject	Grade	Not Attempted	Attempted Tests							Total
			Online Adaptive			Online Fixed Form			Total Attempted	
			Completed	Incomplete	ESR*	Completed	Incomplete	ESR*		
ELA	3		119	1	21	9			150	150
	4		144	3	30	9	1		187	187
	5		125	3	18	13		2	161	161
	6	1	158	1	18	6		1	184	185
	7	2	144	4	11	4		1	164	166
	8		142		17	6			165	165
	10		115	2	12	9	1		139	139
Mathematics	3		117	1	22	9			149	149
	4		139	2	31	10		1	183	183
	5		125		19	13		2	159	159
	6		156		19	7		1	183	183
	7		143		14	5		1	163	163
	8		141	2	15	7			165	165
	10		112	1	12	8			133	133
Science	5		127		18	13		2	160	160
	8		138		13	7			158	158
	11		112	2	9	17			140	140

*Early Stopping Rule

Table 3. Overall Alternate Assessment Participation Rate

Subject	Grade	Number of IDAA Test Participants	Number of Idaho State Summative Test Participants	Overall Idaho State Alternate Assessment Participation Rate (%)
ELA	3	150	23,266	0.6%
	4	187	23,457	0.8%
	5	161	23,398	0.7%
	6	185	23,619	0.8%
	7	166	23,920	0.7%
	8	165	24,284	0.7%
	10	139	16,602	0.8%
	Overall	**1,153**	**158,546**	**0.7%**
Mathematics	3	149	23,356	0.6%
	4	183	23,548	0.8%
	5	159	23,437	0.7%
	6	183	23,702	0.8%
	7	163	23,974	0.7%
	8	165	24,351	0.7%
	10	133	18,990	0.7%
	Overall	**1,135**	**161,358**	**0.7%**
Science	5	160	23,508	0.7%
	8	158	24,438	0.6%
	11	140	21,276	0.7%
	Overall	**458**	**69,222**	**0.7%**

Table 4. Number of Participating Students by Subgroup

Group	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 10/11
ELA							
All	150	187	161	185	166	165	139
Female	60	64	59	73	59	56	56
Male	90	123	102	112	107	109	83
African American	1	2	5	1	2	5	1
American Indian or Alaskan Native	2	4	4	3	3	2	3
Asian	4	7	3		9	3	2
Hispanic or Latino	36	39	33	47	37	31	24
White	103	122	108	128	111	117	104
Pacific Islander	1	2	1				1
Multi-Racial	3	11	7	6	4	7	4
Mathematics							
All	149	183	159	183	163	165	133
Female	59	64	57	72	58	56	56
Male	90	119	102	111	105	109	77
African American	1	2	5	1	2	5	1
American Indian or Alaskan Native	2	4	4	3	3	2	3
Asian	4	7	3		8	3	2
Hispanic or Latino	35	38	33	47	37	31	22
White	103	120	106	126	109	117	100
Pacific Islander	1	2	1				1
Multi-Racial	3	10	7	6	4	7	4
Science							
All			160			158	140
Female			58			54	59
Male			102			104	81
African American			5			4	4
American Indian or Alaskan Native			4			2	2
Asian			3			3	
Hispanic or Latino			34			30	22
White			106			112	110
Pacific Islander			1				
Multi-Racial			7			7	2

Table 5. Number of Participating Students by Subgroup and Disability Category (Grades 3–6)

Group	ASD	ID	DD	EMD	LI	MD	OHI	SLD	TBI	VI	N/A
Grade 3											
All Students	26	12	6		1	17	5				83
Female	8	5	2			10	4				31
Male	18	7	4		1	7	1				52
African American											1
American Indian or Alaskan Native	1	1									
Asian	3										1
Hispanic or Latino	5	2	3		1	7	1				17
White	17	8	3			10	4				61
Pacific Islander											1
Multi-Racial		1									2
Grade 4											
All Students	28	21	2		2	21	6		2	1	104
Female	8	8	1		2	9	4				32
Male	20	13	1			12	2		2	1	72
African American						1					1
American Indian or Alaskan Native	1										3
Asian						1					6
Hispanic or Latino	5	9				4	1				20
White	19	11	2		2	15	5		1	1	66
Pacific Islander	1								1		
Multi-Racial	2	1									8
Grade 5											
All Students	24	29			1	17	1				89
Female	4	14				10	1				30
Male	20	15			1	7					59
African American						2					3
American Indian or Alaskan Native	1					1					2
Asian		1									2
Hispanic or Latino	3	11				4					16
White	18	16			1	9	1				62
Pacific Islander											1
Multi-Racial	2	1				1					3
Grade 6											
All Students	20	37	2	1		14	8	1	2	1	99
Female	5	17		1		5	7		1	1	36
Male	15	20	2			9	1	1	1		63
African American						1					
American Indian or Alaskan Native		1						1			1
Asian											
Hispanic or Latino	4	15				4	2		1	1	20
White	15	20	2			9	6		1		75
Pacific Islander											
Multi-Racial	1	1		1							3

Note. ASD=Autism Spectrum Disorder; ID=Intellectual Disability; DB=Deaf/Blindness; DE=Deaf and Hard of Hearing; DD=Developmental Delay; EMD=Emotional Behavioral Disturbance; LI=Language Impairment; MD=Multiple Disabilities; OHI=Other Health Impairment; OI=Orthopedic Impairment; SLD=Specific Learning Disability; SI=Speech Impairment; TBI=Traumatic Brain Injury; VI=Visual Impairment Including Blindness; N/A=Not Applicable.

Table 6. Number of Participating Students by Subgroup and Disability Category (Grades 7–10)

Group	ASD	ID	DD	EMD	LI	MD	OHI	SLD	TBI	VI	N/A
Grade 7											
All Students	21	29				23	4		1		88
Female	4	13				12	2				28
Male	17	16				11	2		1		60
African American											2
American Indian or Alaskan Native	1	2									
Asian											9
Hispanic or Latino	8	9				6			1		13
White	12	18				17	4				60
Pacific Islander											
Multi-Racial											4
Grade 8											
All Students	24	22			2	9	3		1	1	103
Female	3	10			1	5	2				35
Male	21	12			1	4	1		1	1	68
African American		1								1	3
American Indian or Alaskan Native		1			1						
Asian											3
Hispanic or Latino	1	3			1	3					23
White	21	16				6	3		1		70
Pacific Islander											
Multi-Racial	2	1									4
Grade 10											
All Students	17	21		2	1	12	1		2		81
Female	3	10			1	7			1		34
Male	14	11		2		5	1		1		47
African American											1
American Indian or Alaskan Native	1			1	1						
Asian											2
Hispanic or Latino	3	8							1		12
White	12	12		1		12	1		1		63
Pacific Islander	1										
Multi-Racial		1									3

Note. ASD=Autism Spectrum Disorder; ID=Intellectual Disability; DB=Deaf/Blindness; DE=Deaf and Hard of Hearing; DD=Developmental Delay; EMD=Emotional Behavioral Disturbance; LI=Language Impairment; MD=Multiple Disabilities; OHI=Other Health Impairment; OI=Orthopedic Impairment; SLD=Specific Learning Disability; SI=Speech Impairment; TBI=Traumatic Brain Injury; VI=Visual Impairment Including Blindness; N/A=Not Applicable.

Table 7. Number of Participating Students by Subgroup and Disability Category (Grade 11)

Group	ASD	ID	DD	EMD	LI	MD	OHI	SLD	TBI	VI	N/A
Grade 11											
All Students	15	29		1		12	3				80
Female	3	16				4	2				34
Male	12	13		1		8	1				46
African American											4
American Indian or Alaskan Native		1		1							
Asian											
Hispanic or Latino	2	9				1	1				9
White	13	19				11	2				65
Pacific Islander											
Multi-Racial											2

Note. ASD=Autism Spectrum Disorder; CI=Intellectual Disability; DB=Deaf/Blindness; DE=Deaf and Hard of Hearing; DD=Developmental Delay; EMD=Emotional Behavioral Disturbance; LI=Language Impairment; MD=Multiple Disabilities; OHI=Other Health Impairment; OI=Orthopedic Impairment; SLD=Specific Learning Disability; SI=Speech Impairment; TBI=Traumatic Brain Injury; VI=Visual Impairment Including Blindness; N/A=Not Applicable.

3.2 Summary of Student Performance

Table 8–Table 12 present a summary of the spring 2023 IDAA test results for all students and by subgroup, including the average and the standard deviation of scale scores, the percentage of students in each performance level, and the percentage of Proficient (Proficient + Advanced) students. The results are based on the students who meet attempt requirements for scoring and reporting of the IDAA.

Table 8. Student Performance Overall and by Subgroup—ELA (Grades 3–6)

Group	Number Tested	Scale Score Mean	Scale Score SD	% Below Basic	% Basic	% Prof.	% Adv.	% Prof. & Adv.
Grade 3								
All Students	150	270.11	79.49	37	22	21	19	41
Female	60	274.18	80.25	40	15	23	22	45
Male	90	267.40	79.30	36	27	20	18	38
African American	1*							
American Indian or Alaskan Native	2*							
Asian	4*							
Hispanic or Latino	36	267.36	75.52	36	28	22	14	36
White	103	272.25	82.90	34	22	21	22	44
Pacific Islander	1*							
Multi-Racial	3*							
Grade 4								
All Students	187	256.29	78.10	41	28	21	10	31
Female	64	249.11	79.91	45	25	25	5	30
Male	123	260.03	77.20	38	30	20	12	32
African American	2*							
American Indian or Alaskan Native	4*							
Asian	7	254.29	80.58	43	29	14	14	29
Hispanic or Latino	39	255.26	68.98	41	36	18	5	23
White	122	263.36	76.33	38	26	25	11	36
Pacific Islander	2*							
Multi-Racial	11	224.18	91.61	64	18	9	9	18
Grade 5								
All Students	161	272.53	76.62	43	19	24	14	38
Female	59	265.97	83.61	47	15	20	17	37
Male	102	276.32	72.43	40	22	26	12	38
African American	5*							
American Indian or Alaskan Native	4*							
Asian	3*							
Hispanic or Latino	33	249.67	76.99	64	12	21	3	24
White	108	279.69	71.84	38	24	21	17	38
Pacific Islander	1*							
Multi-Racial	7	282.14	88.08	29	0	57	14	71
Grade 6								
All Students	184	280.38	77.04	37	22	20	21	41
Female	73	287.81	82.29	30	19	23	27	51
Male	111	275.49	73.35	41	24	18	16	34
African American	1*							
American Indian or Alaskan Native	3*							
Asian								
Hispanic or Latino	46	284.09	62.76	37	28	13	22	35
White	128	279.28	79.59	38	20	23	20	43
Pacific Islander								
Multi-Racial	6	295.17	105.39	17	33	17	33	50

*To protect individual student confidentiality, results are not reported for 5 or fewer students.

Table 9. Student Performance Overall and by Subgroup—ELA (Grades 7–10)

Group	Number Tested	Scale Score Mean	Scale Score SD	% Below Basic	% Basic	% Prof.	% Adv.	% Prof. & Adv.
Grade 7								
All Students	164	283.78	62.54	33	24	26	17	43
Female	58	275.00	69.02	38	24	24	14	38
Male	106	288.58	58.48	30	25	26	19	45
African American	2*							
American Indian or Alaskan Native	3*							
Asian	8	260.88	69.54	38	38	25	0	25
Hispanic or Latino	37	282.59	65.28	32	27	24	16	41
White	110	286.71	60.91	33	22	26	19	45
Pacific Islander								
Multi-Racial	4*							
Grade 8								
All Students	165	290.87	78.05	33	16	24	27	51
Female	56	309.16	66.07	23	16	25	36	61
Male	109	281.47	82.24	38	17	23	23	46
African American	5*							
American Indian or Alaskan Native	2*							
Asian	3*							
Hispanic or Latino	31	282.77	86.08	32	23	26	19	45
White	117	293.33	78.62	32	15	23	30	53
Pacific Islander								
Multi-Racial	7	280.86	24.16	57	14	29	0	29
Grade 10								
All Students	139	279.19	66.50	44	15	29	12	41
Female	56	271.36	68.58	48	16	23	13	36
Male	83	284.47	64.95	41	14	34	11	45
African American	1*							
American Indian or Alaskan Native	3*							
Asian	2*							
Hispanic or Latino	24	282.46	50.75	54	17	21	8	29
White	104	275.87	70.94	42	14	33	11	43
Pacific Islander	1*							
Multi-Racial	4*							

*To protect individual student confidentiality, results are not reported for 5 or fewer students.

Table 10. Student Performance Overall and by Subgroup—Mathematics (Grades 3–6)

Group	Number Tested	Scale Score Mean	Scale Score SD	% Below Basic	% Basic	% Prof.	% Adv.	% Prof. & Adv.
Grade 3								
All Students	149	266.44	83.06	48	12	17	22	40
Female	59	268.61	74.40	51	10	20	19	39
Male	90	265.02	88.66	47	13	16	24	40
African American	1*							
American Indian or Alaskan Native	2*							
Asian	4*							
Hispanic or Latino	35	261.74	74.29	60	9	14	17	31
White	103	267.78	88.05	45	12	18	25	44
Pacific Islander	1*							
Multi-Racial	3*							
Grade 4								
All Students	183	260.14	82.32	52	14	20	14	34
Female	64	253.91	82.77	56	16	20	8	28
Male	119	263.50	82.24	50	13	20	17	37
African American	2*							
American Indian or Alaskan Native	4*							
Asian	7	282.43	94.53	57	0	14	29	43
Hispanic or Latino	38	262.32	75.49	47	16	32	5	37
White	120	263.63	79.37	53	13	18	16	34
Pacific Islander	2*							
Multi-Racial	10	244.70	106.51	60	10	10	20	30
Grade 5								
All Students	159	255.09	74.58	43	25	17	15	32
Female	57	248.14	74.98	46	23	16	16	32
Male	102	258.98	74.45	41	26	18	15	32
African American	5*							
American Indian or Alaskan Native	4*							
Asian	3*							
Hispanic or Latino	33	250.79	71.09	39	30	27	3	30
White	106	257.68	71.63	43	27	13	16	29
Pacific Islander	1*							
Multi-Racial	7	277.00	89.53	29	0	29	43	71
Grade 6								
All Students	183	276.45	73.67	38	14	30	19	48
Female	72	283.69	72.11	29	7	46	18	64
Male	111	271.76	74.60	44	18	19	19	38
African American	1*							
American Indian or Alaskan Native	3*							
Asian								
Hispanic or Latino	47	279.32	60.08	38	23	23	15	38
White	126	275.75	77.11	40	10	30	21	51
Pacific Islander								
Multi-Racial	6	274.33	86.71	17	17	67	0	67

*To protect individual student confidentiality, results are not reported for 5 or fewer students.

Table 11. Student Performance Overall and by Subgroup—Mathematics (Grades 7–10)

Group	Number Tested	Scale Score Mean	Scale Score SD	% Below Basic	% Basic	% Prof.	% Adv.	% Prof. & Adv.
Grade 7								
All Students	163	272.50	69.22	42	24	17	18	34
Female	58	268.97	72.40	40	24	19	17	36
Male	105	274.45	67.68	43	24	15	18	33
African American	2*							
American Indian or Alaskan Native	3*							
Asian	8	269.13	77.07	38	25	25	13	38
Hispanic or Latino	37	263.05	71.32	43	27	16	14	30
White	109	275.48	68.17	43	21	17	19	36
Pacific Islander								
Multi-Racial	4*							
Grade 8								
All Students	165	276.13	73.22	42	18	23	17	40
Female	56	285.61	65.07	38	18	27	18	45
Male	109	271.26	76.91	44	18	21	17	38
African American	5*							
American Indian or Alaskan Native	2*							
Asian	3*							
Hispanic or Latino	31	264.77	65.61	55	16	19	10	29
White	117	280.59	77.62	37	20	22	21	44
Pacific Islander								
Multi-Racial	7	270.71	27.46	71	0	29	0	29
Grade 10								
All Students	133	267.44	66.25	41	23	20	16	36
Female	56	257.70	68.49	46	25	18	11	29
Male	77	274.52	64.1	36	22	22	19	42
African American	1*							
American Indian or Alaskan Native	3*							
Asian	2*							
Hispanic or Latino	22	280.05	61.63	36	23	14	27	41
White	100	263.10	68.82	42	24	21	13	34
Pacific Islander	1*							
Multi-Racial	4*							

*To protect individual student confidentiality, results are not reported for 5 or fewer students.

Table 12. Student Performance Overall and by Subgroup—Science (Grades 5, 8, and 11)

Group	Number Tested	Scale Score Mean	Scale Score SD	% Below Basic	% Basic	% Prof.	% Adv.	% Prof. & Adv.
Grade 5								
All Students	160	264.79	73.41	44	29	21	6	27
Female	58	258.05	80.52	43	31	21	5	26
Male	102	268.62	69.17	45	27	21	7	27
African American	5*							
American Indian or Alaskan Native	4*							
Asian	3*							
Hispanic or Latino	34	252.12	71.12	50	24	26	0	26
White	106	270.55	67.45	44	31	17	8	25
Pacific Islander	1*							
Multi-Racial	7	271.86	87.36	29	0	57	14	71
Grade 8								
All Students	158	284.38	66.64	38	20	29	13	42
Female	54	298.33	62.05	28	22	33	17	50
Male	104	277.13	68.06	43	19	27	11	38
African American	4*							
American Indian or Alaskan Native	2*							
Asian	3*							
Hispanic or Latino	30	285.33	61.65	40	20	30	10	40
White	112	282.53	71.77	38	20	27	15	42
Pacific Islander								
Multi-Racial	7	291.43	24.99	29	29	43	0	43
Grade 11								
All Students	140	277.43	60.63	31	36	17	16	33
Female	59	275.10	58.81	31	32	24	14	37
Male	81	279.12	62.23	32	38	12	17	30
African American	4*							
American Indian or Alaskan Native	2*							
Asian								
Hispanic or Latino	22	260.45	46.54	55	36	5	5	9
White	110	280.20	61.27	28	35	20	17	37
Pacific Islander								
Multi-Racial	2*							

* To protect individual student confidentiality, results are not reported for 5 or fewer students.

3.3 TEST-TAKING TIME

The test is administered one-on-one between the student and the person administering the test. The IDAA is not timed. The time spent on each item may vary among students, which may provide useful information about student testing behaviors and motivation. Since the length of a test session could be monitored by TAs who are knowledgeable about their students, additional time for students who need it could be arranged.

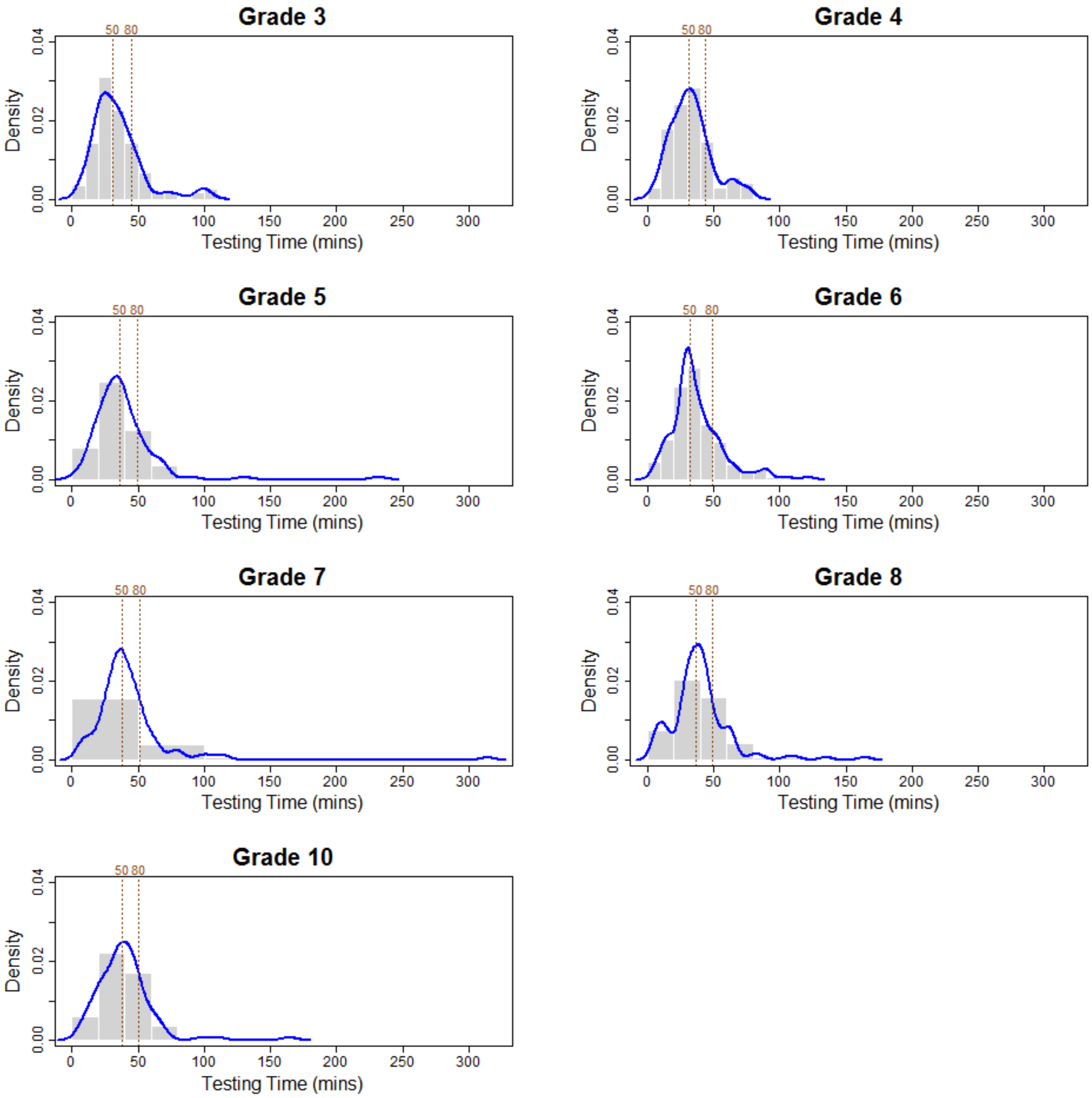
Item response time is captured as the item page time (i.e., the time that a student spends on each item page) in milliseconds in the Test Delivery System (TDS). Discrete items appear on the screen one item at a time, and items associated with a stimulus appear on the screen together with the page time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

Table 13 presents the average testing time and the testing time at various percentiles for the overall test. The distribution of testing time is also provided in Figure 1–Figure 3.

Table 13. Test-Taking Time

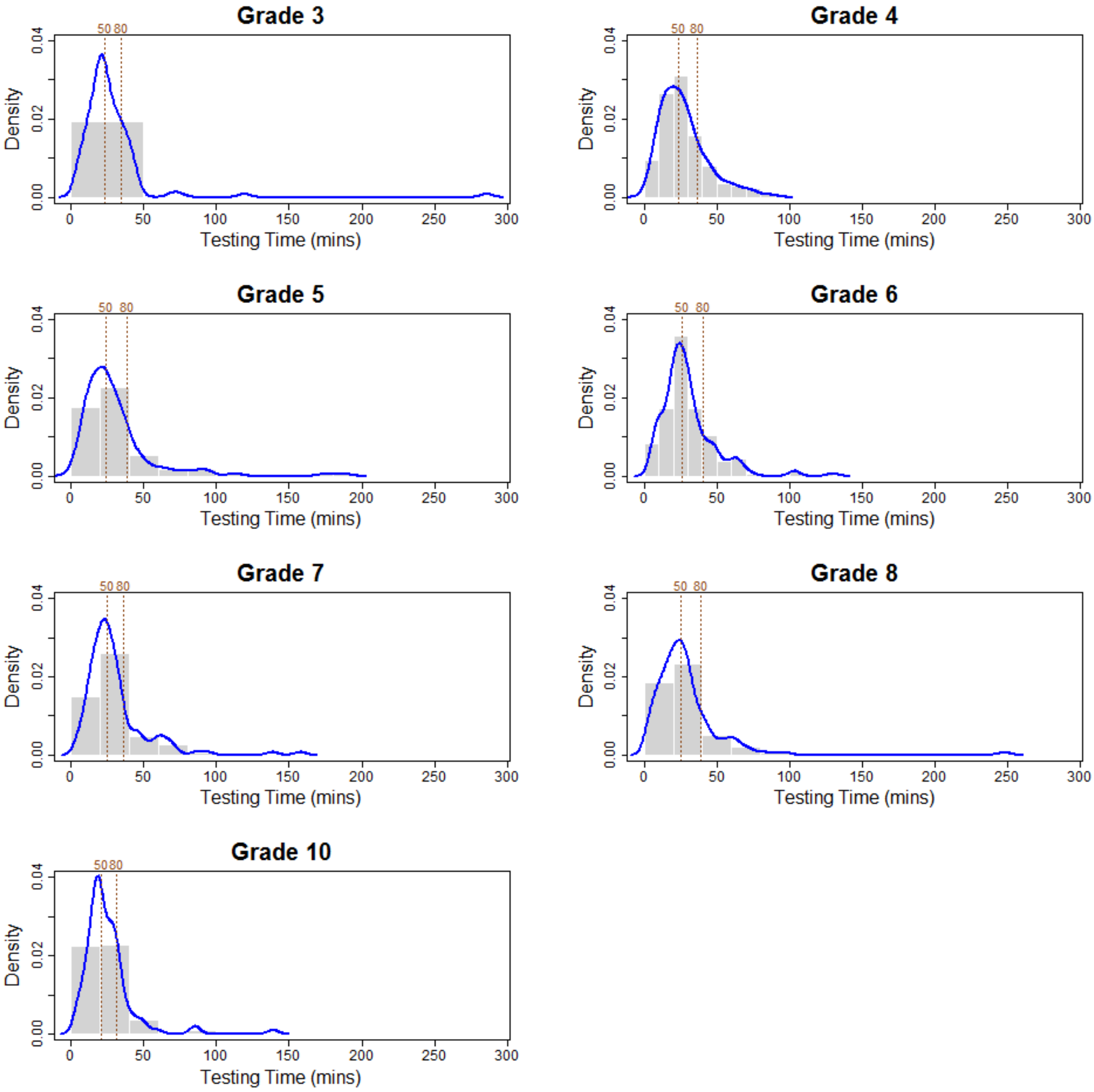
Grade	Average Testing Time (hh:mm)	Median Testing Time (hh:mm)	Testing Time in Percentiles (hh:mm)				
			75th	80th	85th	90th	Max
ELA							
3	00:34	00:30	00:41	00:44	00:46	00:53	01:43
4	00:33	00:31	00:41	00:44	00:48	00:54	01:19
5	00:38	00:35	00:45	00:48	00:52	00:58	03:52
6	00:38	00:33	00:45	00:50	00:54	01:00	02:01
7	00:42	00:37	00:48	00:52	00:54	01:05	05:14
8	00:39	00:37	00:46	00:49	00:56	01:02	02:44
10	00:39	00:38	00:47	00:50	00:55	01:00	02:44
Mathematics							
3	00:27	00:23	00:32	00:34	00:37	00:40	04:46
4	00:27	00:24	00:33	00:36	00:42	00:47	01:27
5	00:31	00:24	00:35	00:40	00:43	00:54	03:07
6	00:30	00:26	00:36	00:40	00:47	00:51	02:09
7	00:29	00:24	00:32	00:35	00:44	00:57	02:38
8	00:28	00:24	00:32	00:38	00:42	00:52	04:08
10	00:25	00:21	00:30	00:31	00:33	00:39	02:19
Science							
5	00:26	00:23	00:32	00:34	00:35	00:43	01:57
8	00:27	00:24	00:30	00:33	00:37	00:42	02:31
11	00:26	00:24	00:30	00:33	00:38	00:42	01:43

Figure 1. Distribution of Testing Time—ELA



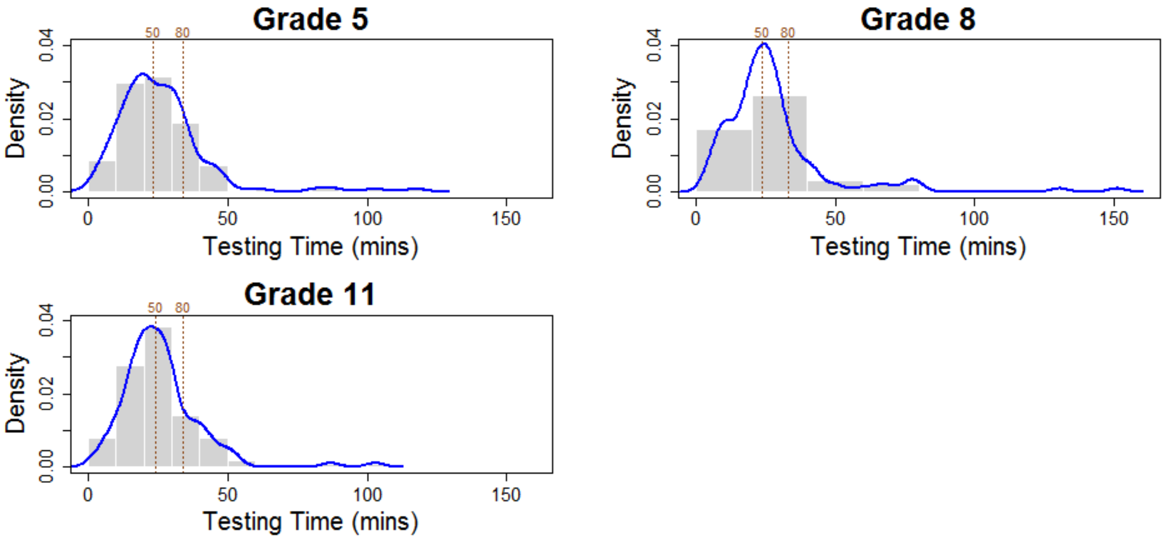
Note: 50 is median; 80 is 80th percentile

Figure 2. Distribution of Testing Time—Mathematics



Note: 50 is median; 80 is 80th percentile

Figure 3. Distribution of Testing Time—Science



Note: 50 is median; 80 is 80th percentile

3.4 Distribution of Student Ability and Item Difficulty in the IDAA Item Pool

Figure 4–Figure 6 display the empirical distribution of students' overall theta scores in the spring 2023 administration and the distribution of the item difficulty parameter estimates in the 2023 IDAA operational item pool. The student ability distributions in ELA, mathematics, and science tests are based on the completed test results from both the adaptive and fixed form tests. These charts provide a visual presentation on whether the difficulty levels of the items in the pool match the ability distribution of the population being assessed. They can also inform a direction for future item development. For example, in some mathematics tests, more easier items are needed in the item pool that target for students with lower academic achievement.

Figure 4. Student Ability—Item Difficulty Distribution for ELA

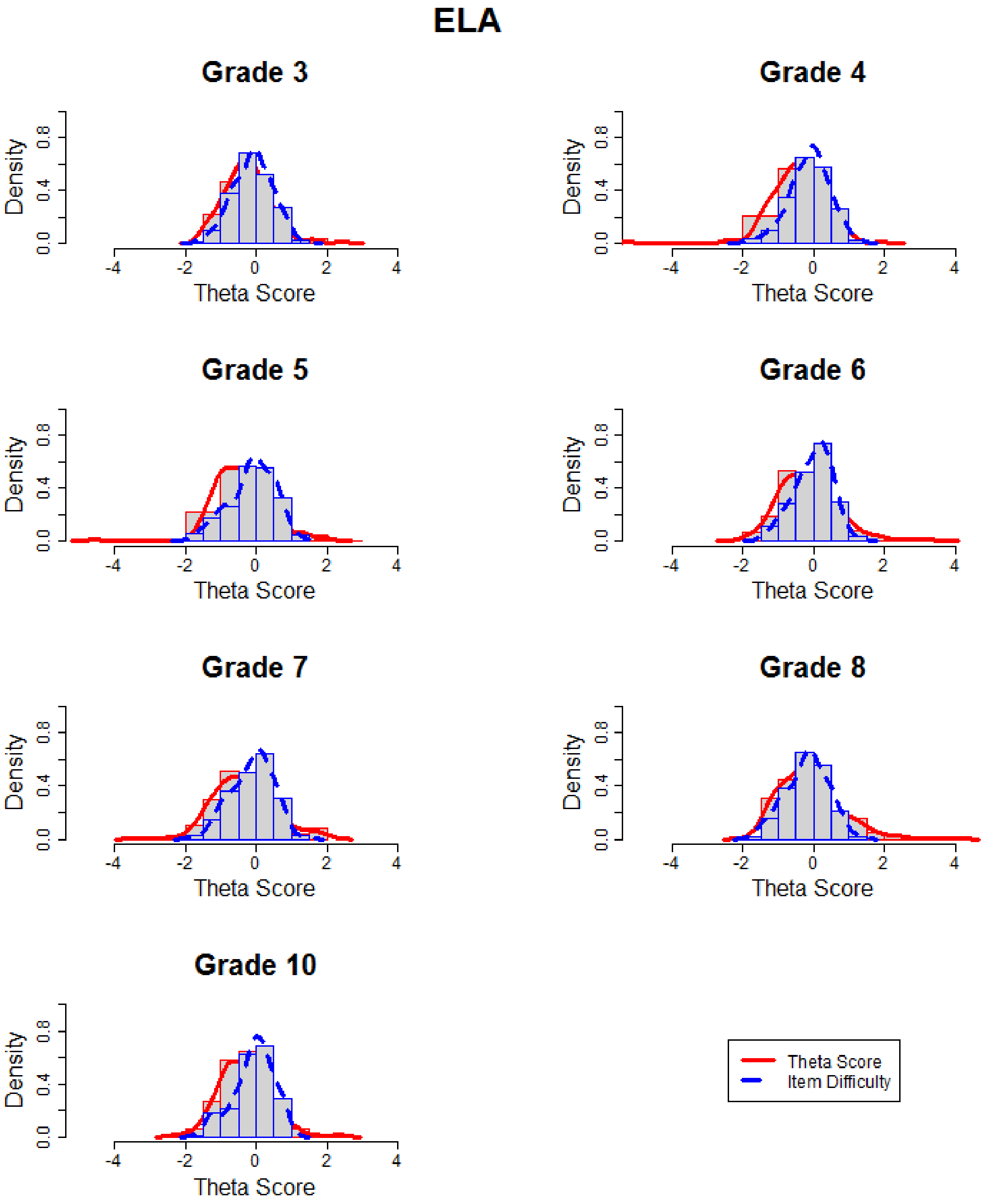


Figure 5. Student Ability—Item Difficulty Distribution for Mathematics

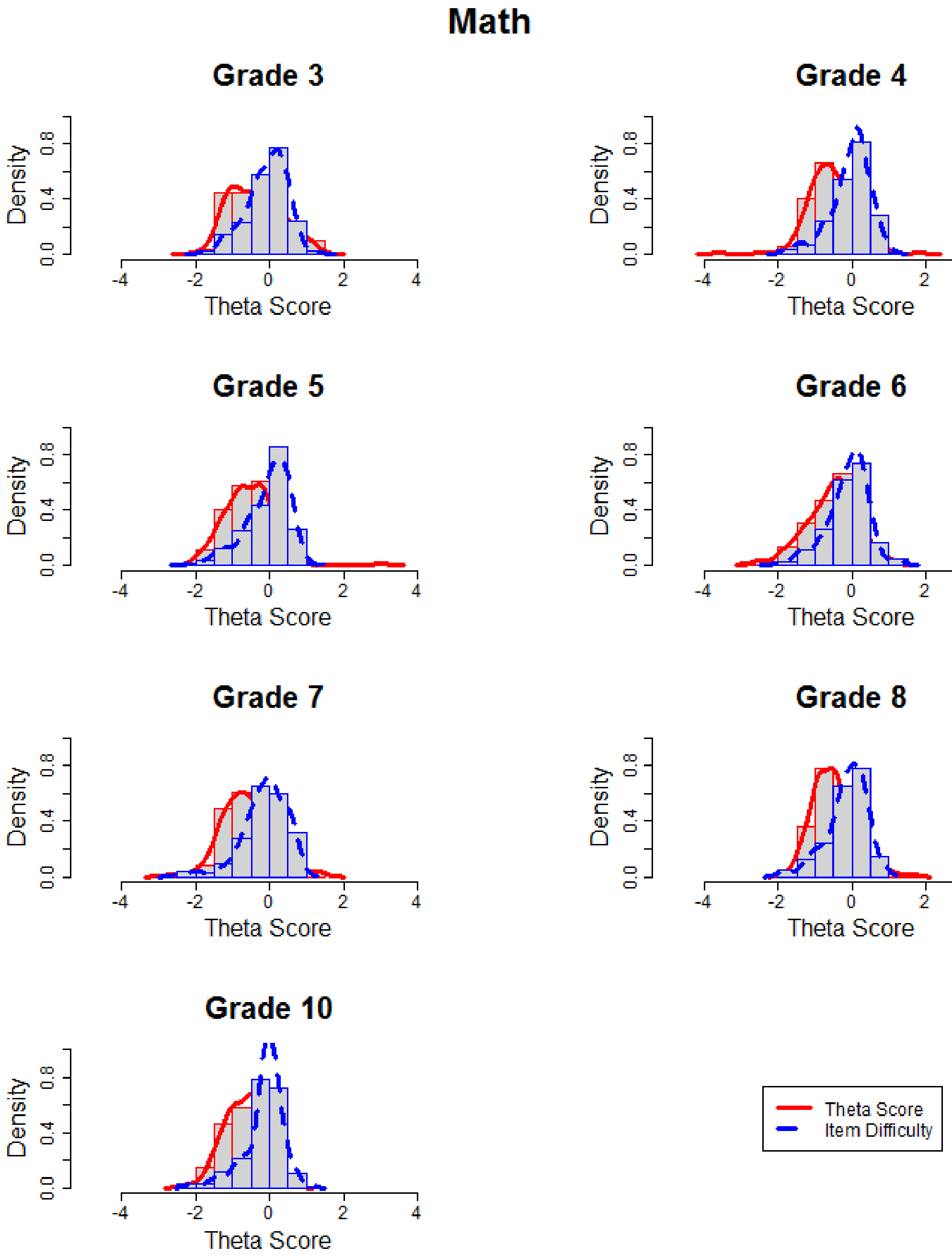
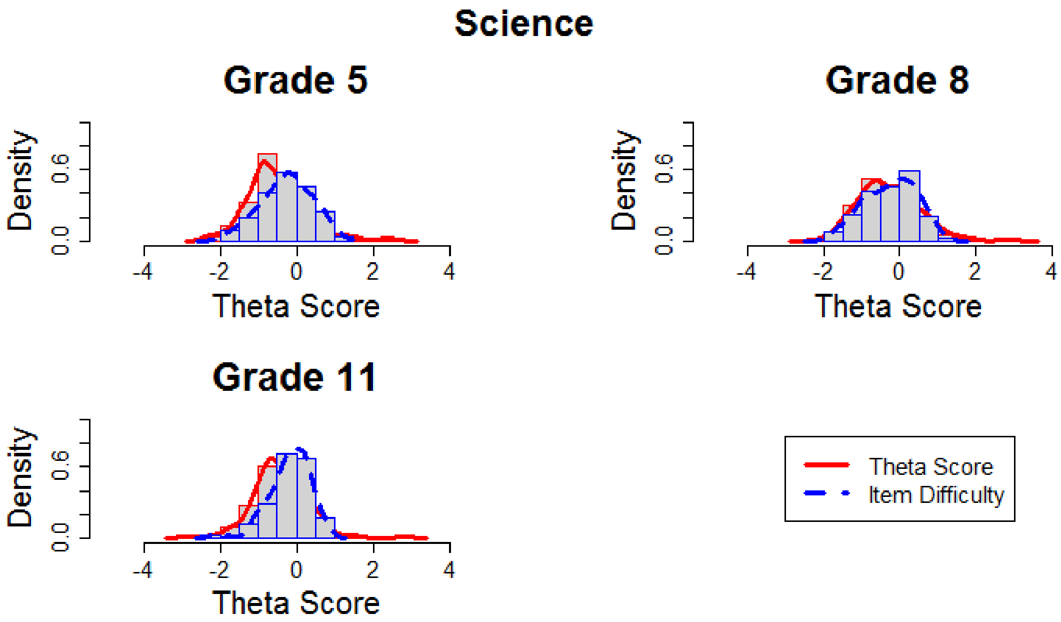


Figure 6. Student Ability—Item Difficulty Distribution for Science



4. ITEM DEVELOPMENT

4.1 ITEM DEVELOPMENT FOR THE MOU-ALT

A Memorandum of Understanding (MOU) on creating an alternate assessment program (MOU-Alt) was initiated in 2018 and signed among the three original states of South Carolina, Hawaii, and Wyoming. The purpose of the MOU was to create a shared field test item pool by having states participate in item development and field testing. Each state contributed a predetermined number of items proportional to their state's student population for alternate assessment for three content areas (English language arts [ELA], mathematics, and science). In early 2019, Idaho and Vermont joined in collaborative item development and field testing and participated in the spring 2019 field test in ELA, mathematics, and science. In spring 2020, Montana and South Dakota joined the MOU for science and participated in the spring 2021 field test. In 2022, Vermont exited the MOU.

A crosswalk across all the individual state alternate assessment standards was completed for the first year of the MOU Alternate Assessment shared field test development. Specifically, the content of the standards from each of the MOU states were reviewed, contrasted and compared by special education content experts at Cambium Assessment, Inc. (CAI) to determine which standards are on-grade and overlapped across states. For example, CAI looked at all of the grade 3 math standard for each MOU state and determined which standards contained common content. If standard A in the first state contained the same content as standard B in the next state, and standard C in the third state, then the three standards in the three states are common. When aligning items to standards in each state, with this crosswalk available, CAI knew instantly which standards items should be aligned to. The opposite is true as well. There were standards that did not have similar content to other states' on-grade standards, so items aligned to those standards were not aligned to other states.

Table 14 to Table 16 below indicate the number of Core Content Connectors (CCC) or Progress Indicators (PI) on the IDAA Blueprint that align to other MOU states' standards and the number of Idaho's CCC and PI that are considered Idaho only standards because they do not crosswalk to other MOU states' standards.

Table 14. Number of Progress Indicators - ELA

	Grade	# of Idaho Progress Indicators on the Blueprint that Crosswalk to Other MOU States' Standards	# of Idaho only Progress Indicators
ELA	3	42	0
	4	35	0
	5	40	0
	6	37	1
	7	34	1
	8	36	0
	HS	29	3

Table 15. Number of Core Content Connectors - Mathematics

Math	Grade	# of Idaho Core Content Connectors on the Blueprint that Crosswalk to Other MOU States' Standards	# of Idaho only Core Content Connectors
	3	32	13
	4	41	13
	5	25	17
	6	43	15
	7	35	13
	8	25	18
	HS	30	42

Table 16. Number of Progress Indicators - Science

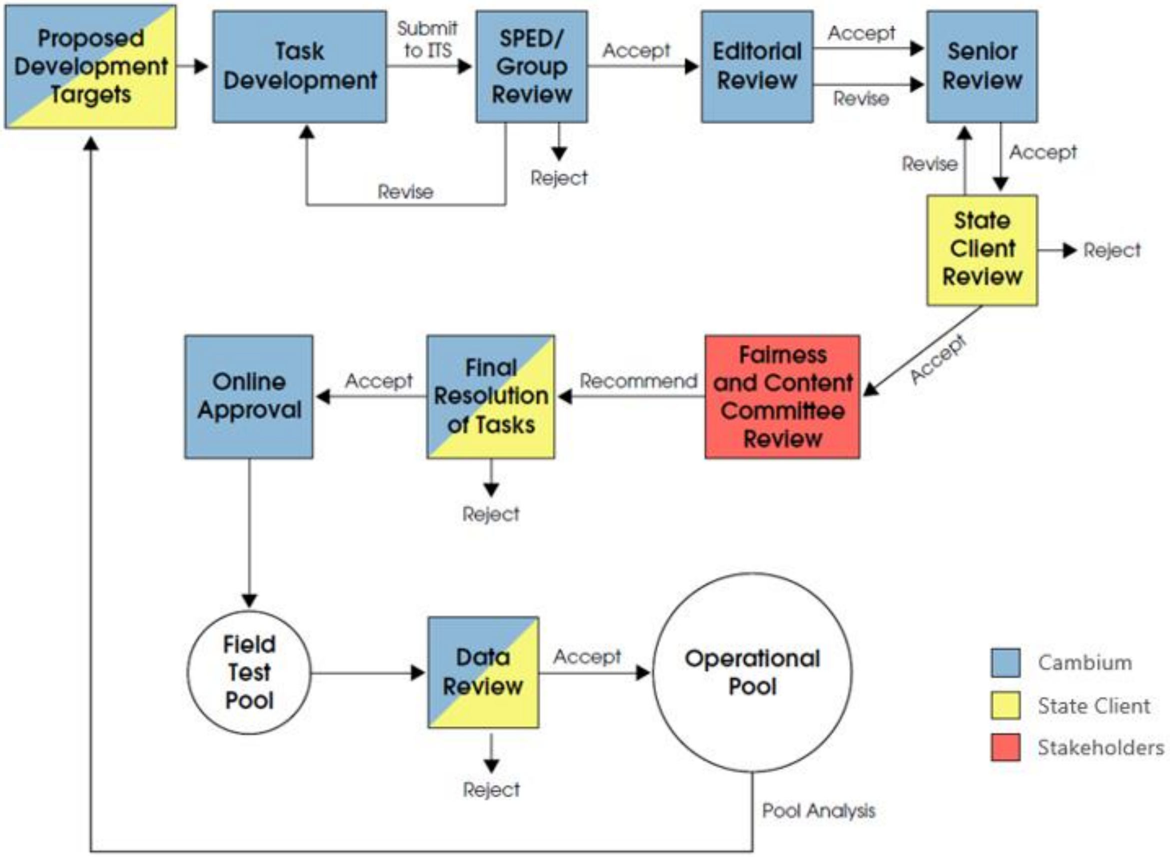
Science	Grade	# of Idaho Progress Indicators on the Blueprint that Crosswalk to Other MOU States' Standards	# of Idaho only Progress Indicators
	ES	40	0
	MS	49	1
	HS	41	0

Once all individual state items were aligned across all the states, item development plans were created for each state. These item development plans were based on identified areas where additional items were needed to ensure that all the MOU standards aligned on the crosswalk were addressed in the item-sharing pool, and items for each state-specific standard that was not aligned to the MOU crosswalk standards were created to meet the state's test blueprint. These item development plans guided the development of the new items to be field tested across states. Each year, following data review of the field-test items, an item-pool analysis was conducted and a new item development plan was created. As new states joined the MOU Alternate Assessment agreement, or when states changed their standards, the individual state standards were added to the crosswalk so that items from the state could be aligned across all the states.

Starting in 2017, items were developed each year for the state-shared MOU Alternate Assessment field test item pool. All the items were developed by a group of professional item writers that included experienced item writers with a background in education and expertise in the assigned content area and specialists in alternate assessments with experience in teaching students with significant cognitive disabilities. Prior to item development, item writers were trained on aspects of items that were unique to students with significant cognitive disabilities. A group of senior test-development specialists monitored and supported the item development activities.

Item development begins with establishing CAI's proposed development targets and working with the individual states to edit the development targets and accept a final plan. The CAI Content team then starts item development. After the items are initially development, they undergo a group review that includes content and senior reviewers, followed by an individual content review, where edits are made based on group reviews, and then a special education review. After the items are reviewed by the special education reviewer, they go through an editorial review. After editorial review, the items go back through a senior review, which is the last review step at CAI before the items are sent to each state for client review. At this step, the client may accept, recommend edits, or reject the items. After the client comments are resolved, all accepted items are then submitted to a Content and Fairness Committee (i.e., Content Advisory Committee in Idaho) for review. At the same time the Content and Fairness Committee reviews the items, the other members of the Alt MOU also review the items and provide feedback. After the Content and Fairness Committee makes its recommendations, the state and CAI convene a resolution meeting at which all of the comments from the Content and Fairness Committee and the other Alt MOU states are reviewed. The items then go through a final edit resolution. The items then go through an approval step in which CAI verifies that the items will appear on the test as expected. Items are then moved into the field-test item pool and are field tested. After the testing window is closed, all field-tested items are analyzed. Items with a sample size smaller than 50 are archived and will be field tested in future years; items with a negative biserial/polyserial correlation are first verified by CAI content specialists to ensure that the items are not miskeyed before they are rejected from the item bank; items with borderline statistics are reviewed in an item data review meeting with CAI and the states. Items are then accepted and/or rejected. The accepted items are then moved into the state's operational item pool. Figure 7 presents a flowchart documenting the item development process.

Figure 7. Alternate Assessment Item Development Process



4.1.1 Item Type and Scoring Rubrics

The MOU shared field test items have multiple-choice (MC) items and multi-select (MS) items. Note that the IDAA item pool does not contain any MS items. The MC items have 2–4 options with one key. For MC items, if a student selects the key, the student receives one point; otherwise, the student receives zero points. Each item measures a specific content standard.

The items can be stand-alone, grouped in short passages with two to three items, or grouped in long passages with four or more items. The test administration algorithm ensures that the items within a passage are always administered together.

Starting in late spring 2018, cognitive labs (cog labs) were conducted in each of the original three states to determine if certain types of technology-enhanced items should be developed for the shared MOU field test items. The item types included MS, equation editor, table match, and animation items. Neither equation-editor nor table-match items proved to be successful item types for this population of students; therefore, states will not develop any more of these item formats in the future.

4.1.2 Item Development Procedure and Item Reviews

4.1.2.1 Item Development Procedure

Items are developed by each of the states that joined the shared item development agreement. In each state, item development for each year begins in the spring. Items are written by CAI content and senior content staff or by third party item development vendors, in compliance with the item specifications and style guide documents to ensure items meet the expected alignment, complexity, and style criteria. The item specifications and style guide documents are created by CAI, reviewed, and approved by the department of education in each individual state. The item specifications are for the MOU, instead of for individual states. If a particular standard is only under one state, that standard is not included in the MOU item specifications. Rather, the state creates separate field-test slots for items associated with state-specific standards.

After items pass CAI's four required stages of internal reviews, described at length in the following sections, items are then presented to the state for department review and acceptance. Following a state's approval of their items, the other state partners are notified that the items are available for review and comments. During this review step, states can also verify whether the items align with their state standards. Any comments regarding item content and suggested revisions are sent to the state that owns the item(s), and that state determines if those comments should be acted upon.

In each state, items owned and accepted by the state are prepared for review by a statewide Content and Fairness Committee convened for each content area. The Content and Fairness Committee comprises stakeholders from around the state with teaching experience in grades K–12 and experience working with students with disabilities. In Idaho, general educators and administrations are also invited to participate in the committee meetings. These stakeholders review the items and provide feedback to ensure that all accepted items were correct and free from fairness and bias issues. Most importantly, these educators ensure that this population of students can understand the language used in the items and that the included visuals and audio directions will aid and not distract students.

Following the committee reviews in these states, the accepted items are shared across the other state item banks for field testing. Figure 7 illustrates a flowchart that documents the development process.

4.1.2.2 Item Reviews

Draft items are reviewed with CAI at various stages, followed by a review from the state staff and the state special education and general education teachers.

CAI Review

Items are reviewed by CAI at the following levels:

- CAI Internal Group Review: In this review, prior to making any changes to draft items, content and senior reviewers meet to discuss items and determine revisions to content, alignment, and style.
- CAI Internal Preliminary Review: This is a preliminary review conducted by a member of CAI's content team assigned to the IDAA. Items are revised to eliminate initial errors, meet content standards, and satisfy internal style and clarity expectations, as agreed on in the group review.
- CAI Internal Content Review: This second content review occurs after the preliminary review to further ensure that changes based on the group review are implemented, and to revise items to address any errors and content, alignment, clarity, and accessibility issues.
- Special Education Review: At this stage, a CAI special education expert reviews and revises the items to ensure that they not only meet the content standards but are also as accessible as possible to students across a broad spectrum of cognitive and physical disabilities. When appropriate, the special education expert designates items as *Access Limited*, meaning that a task is inappropriate to administer to students with a specific physical disability (e.g., blindness). If revisions are required, the special education reviewer will send items back to the content reviewer to implement changes.
- Editorial Review: This review takes place after the special education reviewer approves items. Reviewers then send the items through an editorial review, where a CAI content editor reviews each item to verify that the language used conforms to the standard editorial and style conventions outlined in the item development style guide.
- Senior Review: At this stage, a CAI senior content specialist reviews all items to ensure that they meet the content standards, are free of typographical and technical errors (e.g., key check, spell check), and previously requested edits are in place.
- CAI Batch Review: This is the last step in the CAI internal review process and is designed as a final quality control check to ensure that the items are ready for state review.

State Review

At this level, items are compared to the state standards, reviewed against the Performance -Level Descriptors (PLDs) at all difficulty levels, and compared to the blueprint. Items are further reviewed to ensure that they align with the support guides for each subject area. At this stage, state staff review each item and make the following decisions:

- Accept without modification ("Accept as Appears")
- Request minor revisions ("Accept as Revised")
- Request substantial changes and re-submit for a second State Department of Education (SDE) review ("Revise and Resubmit")
- Reject entirely (e.g., failure to meet content standards, inappropriateness for the targeted grade, general lack of clarity)

Content and Fairness Committee Review

Following revisions and state approval, items are submitted to the Content and Fairness Committee for further review. The state recruited the following to be on the review committee: special educators, general educators, vision and hearing specialists, school principals, and special education directors. The review committee's members represent the diverse gender, ethnicity, disability, race, and cultural subgroups across the state. During the review meeting, each item is assessed to ensure that it meets bias and sensitivity guidelines, aligns with content standards, and abides by universal design (UD) principles.

The following are the common criteria used for item review:

- Content accuracy and clarity
- Alignment to the content specifications
- Appropriate scoring rubrics
- Correct answer key and appropriate distractors for each MC item
- Appropriate item format for item content
- Precision and clarity of wording in directions and items
- Appropriate graphics for color-blindness issues and standardized font size
- Accessibility for students with vision impairment
- Appropriate, fair, and nonbiased content

At the beginning of each meeting, a CAI item development specialist provides a training session to ensure that the committee members understand the expectations and are familiar with the training materials that encompass the pertinent content and bias guidelines. Because the MOU shared items are used in each state for its online assessment, the committee members conduct the review online to view the items in the same way that the student will view them.

4.1.3 Development of Crosswalk and State Alternate Performance Standards

Before item development began, the alternate performance standards for each state were compared in a crosswalk created by senior test development specialists in CAI and reviewed by the state Department of Education. The crosswalk was based on each state's blueprint and included the common core standards and the general education and alternate performance standards for each state. Each state had a unique set of alternate performance standards as follows:

- Hawaii Essence Statements and PLDs
- Idaho Extended Content Standards Core Content Connectors
- Montana Content Standards in Science
- South Carolina Prioritized Standards and PLDs
- South Dakota Science Standards and Core Content Connectors
- Vermont Common Core State Standards (CCSS) (for ELA and mathematics) and Next Generation Science Standards (NGSS), and Achievement-Level Descriptors (ALDs)
- Wyoming Extended Standards and Instructional ALDs

These performance standards were examined to determine how they aligned with the general education standards and to each other. This examination revealed the standards to which items could be developed to meet the needs of each of the states.

The crosswalk then informed the development of item specifications. Each item specification included the Common Core or General Education standard (for Idaho, only General Education standards were included), followed by the state-specific alternate performance standards that aligned with the General Education standard. The item specifications also included complexity statements and task demands. The language of the complexity statements and task demands were derived from each state's performance standards, where applicable, and were synthesized to drive items aligned to multiple states. Once completed, the item specifications were sent to each state for review to confirm alignment and overall approach.

The content extensions of the MOU states were internally examined to create a content extension crosswalk between the states. For each Common Core standard, CAI examined the states' content extensions and PLD documents to identify which extensions were aligned to that Common Core standard. This crosswalk was used as the basis for the structure of the MOU Item Specifications, informing the "Common Core Standard" and "Content Extensions by State" sections of the *MOU Item Specifications*.

The states' content extensions and PLDs or ALDs were further analyzed to cull relevant concepts, skills, and vocabulary. Based on MOU state feedback, these were compiled and displayed in the form of a Complexity matrix and a Vocabulary matrix, revealing which concepts, skills, and vocabulary were relevant to each state. The intent was to provide an "at-a-glance" perspective on content extension overlap across the states. The Complexity and Vocabulary matrices were subdivided into three categories of cognitive complexity: (1) Low, (2) Moderate, and (3) High. The states' content extensions and PLDs were also analyzed to reveal state-specific and cross-state content limits in the content extensions.

The analyses outlined were then used to create a numbered list of task demands describing the essential tasks students were expected to perform based on the language of the content extensions and PLDs or ALDs. Additionally, these task demands were annotated with information regarding complexity and any special exceptions for individual states. A sample items section was added to the list of task demands. Each sample item was annotated with information regarding complexity and special state exceptions. Each sample item also referred to the numbered list of task demands as a reference.

4.2 FIELD TESTING

Items that passed the Content and Fairness Committee review were field tested in the spring 2023 test administration and embedded among operational items. The IDAA was administered online as computer-adaptive tests (CATs) in ELA, mathematics, and science at all grade levels. CATs were assembled using CAI's adaptive testing algorithm. The adaptive item selection algorithm selected items based on their content value that met blueprint and information value that matched students' ability.

Embedding field-test items among operational items yields item parameter estimates that capture all the contextual effects contributing to item difficulty in operational test administrations. Field testing in an operational setting is beneficial in the context of a pre-equating model for scoring and reporting test results. Because the test administration context remains the same as subsequent operational test administration, item parameter estimates are more stable over time than they may be when obtained through stand-alone field testing.

Following the spring 2023 test administrations, all field-test items were calibrated anchoring on the operational items for each grade and subject, creating a pre-calibrated operational item pool that could be used in future administrations. Items field tested in spring 2023 were not used for scoring in the same year.

The spring 2023 field-test item pool consisted of the items shared across MOU states and the unique items within each state. The items shared across MOU states were administered in the MOU states that reviewed and agreed to field test the items, while state-specific items were administered in that state only. The spring 2023 MOU item pool is summarized in Table 17.

Table 17. Summary of Spring 2023 Field-Test Item Pool Across MOU-ALT States

Subject	Grade	State Only	MOU						
		ID	HI	ID	MT	SC	SD	WY	MOU Total
ELA	3		4	8		18		2	32
	4		4	6		18		4	32
	5		4	7		18		2	31
	6		4	6		17		2	29
	7		5	6		17		2	30
	8		4	7		17		2	30
	HS		4	6		17		2	29
Mathematics	3		4	7		13		2	26
	4		4	8		12		3	27
	5		5	7		13		2	27
	6		6	8		12		2	28
	7		5	7		13		2	27
	8		4	7		13		3	27
	HS		4	6		11		2	23
Science	ES		22	10	5		7	30	74
	MS		21	6	2		6	4	39
	HS	1	8	7	3	1	7	20	46

Note. ES=Elementary School; MS=Middle School; HS=High School.

4.2.1 Item Statistics

After the close of the spring 2023 testing window, CAI psychometrics staff analyzed field-test data in preparation for item data review meetings and promotion of high-quality test items to operational item pools. Analysis of field-test items included classical item statistics and item response theory (IRT) item calibrations. Item analyses were conducted on the combined data across MOU-ALT states.

Classical item statistics are designed to evaluate the relationship of each item to the overall scale, assess the quality of the distractors, and identify items that may exhibit bias across subgroups (differential item functioning [DIF] analyses). The IRT item analyses allow examination of the fit of items to the measurement model and provide the statistical foundation for operational form construction and test scoring and reporting. Items are flagged if analyses indicate resulting values are out of range. Flagged items are reviewed by CAI and MOU-ALT states staff. Items that pass CAI and MOU states' statistical review process are accepted for future operational use.

4.2.2 Classical Statistics

Classical item analyses ensure that the field-test items function as intended, according to the MOU-ALT's underlying scales. CAI's analysis program computes the required item and test statistics for each dichotomous and polytomous item to check the integrity of the item and verify the appropriateness of the item's difficulty level. Key statistics that are computed and examined include item difficulty, item discrimination, and distractor analysis.

Items that are extremely difficult or easy are flagged for review, but not necessarily rejected, if they align with the test and content specifications. For dichotomous items, the proportion of test takers in the sample selecting the correct answer (*p*-value) is computed, as well as those selecting the incorrect responses. For items with 0–2 score points, item difficulty is calculated both as the item's mean score and the average proportion correct (analogous to *p*-value and indicating the ratio of an item's mean score divided by the maximum score point possible). Items are flagged for review if the *p*-value was less than 0.25 or greater than 0.95.

The item discrimination index indicates the extent to which each item differentiates between those test takers who possess the skills being measured and those who do not. In general, the higher the value, the better the item can differentiate between high- and low-achieving students. The discrimination index is calculated as the correlation between the item score and the student's IRT-based ability estimate. Items are flagged for subsequent reviews if the biserial/polyserial correlation for the keyed (correct) response is less than 0.20. For polytomous items, we also compute the mean total points earned on the entire test within each possible score category of the item. Items are flagged for review if the mean total score for a lower score point is greater than the mean total score for a higher score point.

Distractor analysis for the dichotomous items is used to identify items with marginal distractors or ambiguous correct responses. The discrimination value of the correct response should be substantial and positive, and the discrimination values for distractors should be lower and, generally, negative. The biserial correlation for distractors is the correlation between the item score (treating the target distractor as the correct response) and the student's IRT ability estimate, restricting the analysis to those students selecting either the target distractor or the keyed response. Items were flagged for subsequent reviews if the biserial correlation for the distractor response was greater than 0.05.

Flagging criteria for field test items are shown in Table 18.

Table 18. Thresholds for Flagging in Classical Item Analysis

Analysis Type	Flagging Criteria
Item Difficulty	The proportion of students (*p*-value) is < 0.25 or > 0.95.
Item Discrimination	Biserial or polyserial correlation for the correct response is < 0.20.
Mean Score for Two-Point Items	Mean total score for a lower score point > Mean total score for a higher score point
Distractor Analysis	Point biserial correlation for any distractor response is > 0.05.

4.2.3 Item Response Theory Statistics

Rasch and Masters' partial credit model (PCM) is used to estimate the IRT model parameters for dichotomously and polytomously scored items, respectively. We reviewed the Winsteps output showing the item statistics resulting from the anchored estimation of parameters for items in the operational tests. Item fit is evaluated via the mean square Infit and mean square Outfit statistics reported by Winsteps, which are based on weighted and unweighted standardized residuals for each item response. These residual statistics indicate the discrepancy between observed item responses and the predicted item responses based on the IRT model. Both fit statistics have an expected value of 1. Values substantially greater than 1 indicate model underfit, while values substantially less than 1 indicate model overfit (Linacre, 2004). Items are flagged if Infit or Outfit values are less than 0.5 or greater than 2.0.

4.2.4 Analysis of Differential Item Functioning

DIF refers to items that appear to function differently across identifiable groups, typically across different demographic groups. Identifying DIF is important because it can indicate that an item contains a cultural or other bias. Not all items that exhibit DIF are biased; some characteristics of the educational system may also lead to DIF. For example, if schools in low-income areas are less likely to offer geometry classes, students at those schools might perform more poorly on geometry items than would be expected, given their proficiency on other types of items. In this example, it is not the item that exhibits bias, but the curriculum. However, because DIF can indicate bias, all field-tested items are evaluated for DIF. Items exhibiting DIF are flagged for further examination by CAI and the MOU-ALT states.

CAI conducts DIF analysis on all field-tested items to detect potential item bias across major ethnic and gender groups. For MOU-ALT, DIF is investigated among the following group comparisons:

- Female vs. Male
- African American vs. White
- Hispanic or Latino vs. White
- Severe and Moderate Intellectual Disability vs. Other. Severe and moderate Intellectual disability is defined by each state based on their primary disability code.

CAI uses a generalized Mantel-Haenszel (MH) procedure to evaluate DIF. The two generalizations include (1) adaptation to polytomous items, and (2) improved variance estimators to render the test statistics valid under complex sample designs. Because students within a district, school, and classroom are more similar than would be expected in a simple random sample of students statewide, the information provided by students within a school is not independent. Therefore, standard errors assume that simple random samples are underestimated. We compute design-consistent standard errors that reflect the clustered nature of educational systems. While clustering is mitigated through random administration of large numbers of embedded field-test (EFT) items, design effects in student samples are rarely reduced to the level of a simple random sample.

The ability distribution is divided into a configurable number of intervals to compute the MH chi-square ($MH\ \chi^2$) DIF statistics. The analysis program computes the MH chi-square value, the log-odds ratio, the standard error of the log-odds ratio, and the MH-delta ($\Delta_{\text{hat } MH}$) for the dichotomous items and the MH chi-square, the standardized mean difference (SMD), and the standard error of the SMD for the polytomous items.

Items are classified into three categories (A, B, or C), ranging from no evidence of DIF to severe DIF according to the DIF classification convention listed in Table 19. Items are also categorized as positive DIF (i.e., +A, +B, or +C), signifying that the item favors the focal group (e.g., African American/Black, Hispanic, female), or negative DIF (i.e., –A, –B, or –C), signifying that the item favors the reference group (e.g., white, male).

Table 19. DIF Classification Rules

Dichotomous Items	
Category	Rule
C	MH_{X^2} is significant and $\left\|\hat{\Delta}_{MH}\right\| \geq 1.5$.
B	MH_{X^2} is significant and $1 \leq \left\|\hat{\Delta}_{MH}\right\| < 1.5$.
A	MH_{X^2} is not significant or $\left\|\hat{\Delta}_{MH}\right\| < 1$.
Polytomous Items	
Category	Rule
C	MH_{X^2} is significant and $\|SMD\|/\ \|SD\| > .25$.
B	MH_{X^2} is significant and $.17 < \|SMD\|/\ \|SD\| \leq .25$.
A	MH_{X^2} is not significant or $\|SMD\|/\ \|SD\| \leq .17$.

Table 20 presents the number of items in each DIF classification category for items with sample size larger than or equal to 50 in both the focal and reference groups. Items are flagged if their DIF statistics fall into the "C" category for any group and the sample size for both focal and reference groups are larger than or equal to 50. A DIF classification of "C" indicates that the item shows significant DIF and should be reviewed for potential content bias, differential validity, or other issues that may reduce item fairness. Because of the unreliability of the DIF statistics when calculated on small samples, caution must be used when evaluating DIF classifications for items where focal or reference groups are fewer than 200 students (Mazor, Clauser, & Hambleton, 1992; Camilli & Shepard, 1994; Muniz, Hambleton, & Xing, 2001; Sireci & Rios, 2013).

All items flagged due to DIF were reviewed during the item data/content review process by content specialists in Idaho SDE. Reviewers were instructed to examine whether there were any content reasons that may have led to the item being flagged. Items that were determined to be biased were rejected and not included in the state's operational item pool.

Table 20. Number of Items in Each DIF Classification Category

Female vs Male								**African American vs. White**							
Subject/Grade	Total	+A	-A	+B	-B	+C	-C	Subject/Grade	Total	+A	-A	+B	-B	+C	-C
ELA	32	18	14					**ELA**	14	6	5			2	1
3	32	12	18			2		3	18	5	13				
4	31	17	14					4	31	17	14				
5	29	14	13		1		1	5	21	9	11				1
6	30	14	14			2		6	13	6	7				
7	29	18	10				1	7	27	10	17				
8	29	15	13			1		8	29	9	16				4
HS	26	13	12		1			HS	26	11	14			1	
Mathematics	27	12	15					**Mathematics**	25	14	11				
3	27	12	15					3	27	18	9				
4	28	14	14					4	24	8	14			1	1
5	27	13	14					5	22	8	12				2
6	27	13	14					6	27	9	18				
7	23	13	8		2			7	23	10	13				
8	15	9	6					8	2	1	1				
HS	18	7	10				1	HS	18	5	12		1		
Science	22	11	11					**Science**	19	10	9				
ES	32	18	14					ES	14	6	5			2	1
MS	32	12	18			2		MS	18	5	13				
HS	31	17	14					HS	31	17	14				
Hispanic vs. White								**Severe/Moderate Disability vs.Other**							
Subject/Grade	Total	+A	-A	+B	-B	+C	-C	Subject/Grade	Total	+A	-A	+B	-B	+C	-C
ELA								**ELA**	1						1
3								3	26	16	10				
4								4	29	9	19				1
5								5	29	12	15				2
6	6	2	3				1	6	30	14	15				1
7								7	29	11	17				1
8								8	29	9	18				2
HS								HS	14	9	5				
Mathematics								**Mathematics**	24	12	12				
3	6	5	1					3	26	9	16				1
4	1						1	4	28	12	13			1	2
5	4	2	2					5	26	9	16				1
6	6	4	2					6	27	15	12				
7								7	23	14	5			1	3
8								8	7	4	3				
HS	1			1				HS	18	5	11			1	1
Science								**Science**	19	7	12				
ES								ES	1						1
MS	4		4					MS	26	16	10				
HS	2	1	1					HS	29	9	19				1

Note. This table includes only items with sample size > = 50 in both the focal and reference groups. ES = Elementary School (grades 3–5); MS = Middle School (grades 6–8); HS = High School (grades 9–12).

4.2.5 Summary of Item Statistics

This section summarizes results from the classical item analysis and item calibration analysis of the 2023 MOU-ALT field-test items conducted after the testing window closed. Table 21—Table 23 summarize item statistics for *p*-values, biserials/polyserials, item difficulties, Infit and Outfit by percentile, and the range, by grade and subject, for all MOU items administered in ELA, mathematics, and science. For each item statistic (e.g., *p*-values, the percentiles are computed across items), the column "Total MOU Items" shows the number of items in the MOU-ALT field-test pool used to compute the percentiles.

Table 21. Summary of Item Analyses Results for MOU-ALT ELA

Grade	Total MOU Items	Statistics	Min	P10	P25	P50	P75	P90	Max
3	32	*p*-value	0.31	0.43	0.47	0.54	0.59	0.62	0.65
		Biserial/Polyserial	-0.07	0.03	0.16	0.31	0.43	0.52	0.7
		Step Difficulty	-1.03	-0.84	-0.74	-0.52	-0.16	0.07	0.51
		Infit	0.82	0.91	0.94	1.01	1.1	1.17	1.23
		Outfit	0.78	0.86	0.93	1.02	1.11	1.18	1.29
4	32	*p*-value	0.28	0.35	0.42	0.48	0.56	0.61	0.75
		Biserial/Polyserial	0.05	0.13	0.2	0.31	0.46	0.51	0.63
		Step Difficulty	-1.55	-0.92	-0.65	-0.25	0.01	0.29	0.62
		Infit	0.85	0.91	0.96	1.02	1.07	1.11	1.15
		Outfit	0.81	0.88	0.95	1.01	1.09	1.16	1.23
5	31	*p*-value	0.34	0.4	0.49	0.54	0.61	0.72	0.75
		Biserial/Polyserial	-0.17	0.15	0.21	0.36	0.47	0.53	0.69
		Step Difficulty	-1.5	-1.38	-0.8	-0.58	-0.31	-0.01	0.36
		Infit	0.86	0.91	0.94	1	1.07	1.09	1.25
		Outfit	0.81	0.85	0.9	0.99	1.09	1.13	1.41
6	29	*p*-value	0.3	0.38	0.48	0.54	0.63	0.69	0.71
		Biserial/Polyserial	0.1	0.13	0.24	0.3	0.46	0.54	0.68
		Step Difficulty	-1.24	-1.12	-0.85	-0.33	-0.09	0.36	0.77
		Infit	0.86	0.91	0.94	1.03	1.09	1.15	1.17
		Outfit	0.77	0.85	0.93	1.06	1.12	1.28	1.39
7	30	*p*-value	0.25	0.4	0.47	0.54	0.64	0.68	0.7
		Biserial/Polyserial	0.02	0.06	0.14	0.25	0.41	0.48	0.6
		Step Difficulty	-1.31	-1.26	-0.96	-0.54	-0.23	0.11	0.92
		Infit	0.88	0.95	0.96	1.07	1.12	1.16	1.21
		Outfit	0.84	0.87	0.96	1.08	1.16	1.23	1.24
8	30	*p*-value	0.19	0.39	0.47	0.6	0.64	0.71	0.79
		Biserial/Polyserial	-0.24	0.07	0.24	0.31	0.5	0.61	0.64
		Step Difficulty	-1.76	-1.39	-1.03	-0.81	-0.2	0.26	1.34
		Infit	0.84	0.87	0.94	1	1.07	1.19	1.27
		Outfit	0.75	0.81	0.88	0.97	1.13	1.28	1.66
HS	29	*p*-value	0.37	0.42	0.52	0.61	0.67	0.78	0.83
		Biserial/Polyserial	0.12	0.17	0.35	0.46	0.58	0.65	0.68
		Step Difficulty	-1.95	-1.57	-1	-0.63	-0.3	0.08	0.4
		Infit	0.84	0.88	0.89	0.95	1.03	1.12	1.18
		Outfit	0.69	0.77	0.84	0.95	1.02	1.13	1.21

Note. HS = High School (grades 9–12).

Table 22. Summary of Item Analyses Results for MOU-ALT Mathematics

Grade	Total MOU Items	Statistics	Min	P10	P25	P50	P75	P90	Max
3	26	*p*-value	0.28	0.32	0.37	0.45	0.57	0.62	0.64
		Biserial/Polyserial	0.08	0.1	0.18	0.24	0.32	0.38	0.51
		Step Difficulty	-1.09	-1.01	-0.82	-0.22	0.13	0.35	0.58
		Infit	0.92	0.98	1	1.04	1.08	1.11	1.18
		Outfit	0.9	0.96	0.99	1.05	1.09	1.17	1.26
4	27	*p*-value	0.27	0.3	0.33	0.47	0.54	0.61	0.74
		Biserial/Polyserial	-0.06	0.07	0.11	0.25	0.4	0.51	0.62
		Step Difficulty	-1.73	-1	-0.79	-0.4	0.17	0.42	0.48
		Infit	0.86	0.93	0.96	1.03	1.09	1.13	1.19
		Outfit	0.82	0.9	0.94	1.03	1.15	1.2	1.24
5	27	*p*-value	0.25	0.29	0.33	0.46	0.6	0.64	0.7
		Biserial/Polyserial	-0.17	-0.14	0.07	0.2	0.35	0.4	0.64
		Step Difficulty	-1.28	-1.13	-0.93	-0.29	0.29	0.5	0.77
		Infit	0.86	0.93	0.97	1.03	1.08	1.14	1.19
		Outfit	0.81	0.93	0.96	1.02	1.09	1.15	1.22
6	28	*p*-value	0.27	0.29	0.34	0.44	0.58	0.65	0.67
		Biserial/Polyserial	-0.26	-0.12	0.01	0.13	0.34	0.41	0.41
		Step Difficulty	-1.39	-1.26	-1.02	-0.3	0.1	0.31	0.48
		Infit	0.94	0.95	0.98	1.06	1.11	1.16	1.26
		Outfit	0.93	0.94	0.96	1.07	1.12	1.28	1.39
7	27	*p*-value	0.23	0.28	0.32	0.37	0.5	0.67	0.69
		Biserial/Polyserial	-0.16	-0.11	0.07	0.2	0.3	0.37	0.49
		Step Difficulty	-1.56	-1.38	-0.64	-0.05	0.11	0.36	0.64
		Infit	0.93	0.95	1	1.03	1.09	1.14	1.17
		Outfit	0.91	0.96	0.99	1.03	1.11	1.22	1.26
8	27	*p*-value	0.23	0.27	0.31	0.45	0.56	0.63	0.76
		Biserial/Polyserial	-0.23	-0.12	-0.04	0.07	0.16	0.21	0.25
		Step Difficulty	-1.85	-1.16	-0.92	-0.4	0.2	0.44	0.53
		Infit	0.98	1	1.02	1.04	1.09	1.11	1.15
		Outfit	0.95	0.98	1.02	1.05	1.12	1.14	1.18
HS	23	*p*-value	0.29	0.36	0.46	0.55	0.6	0.64	0.78
		Biserial/Polyserial	-0.17	-0.03	0.14	0.19	0.29	0.41	0.49
		Step Difficulty	-1.97	-1.15	-1.02	-0.79	-0.41	0.05	0.41
		Infit	0.89	0.92	0.99	1.03	1.07	1.14	1.2
		Outfit	0.81	0.9	0.96	1.02	1.08	1.15	1.33

Note. HS = High School (grades 9–12).

Table 23. Summary of Item Analyses Results for MOU-ALT Science

Grade	Total MOU Items	Statistics	Min	P10	P25	P50	P75	P90	Max
ES	74	*p*-value	0.13	0.26	0.35	0.45	0.56	0.69	0.81
		Biserial/Polyserial	-0.5	-0.03	0.14	0.3	0.53	0.63	0.94
		Step Difficulty	-1.98	-1.14	-0.55	-0.07	0.41	0.73	1.52
		Infit	0.7	0.86	0.92	1.03	1.13	1.23	1.51
		Outfit	0.52	0.79	0.89	1.04	1.16	1.26	2.86
MS	39	*p*-value	0.27	0.34	0.4	0.52	0.61	0.73	0.8
		Biserial/Polyserial	-0.01	0.11	0.18	0.29	0.49	0.57	0.83
		Step Difficulty	-1.89	-1.45	-0.86	-0.34	0.17	0.47	0.94
		Infit	0.85	0.87	0.93	1	1.09	1.14	1.2
		Outfit	0.64	0.79	0.87	0.99	1.1	1.16	1.18
HS	46	*p*-value	0.21	0.3	0.39	0.49	0.6	0.67	0.79
		Biserial/Polyserial	-0.2	-0.09	0.11	0.28	0.4	0.6	0.82
		Step Difficulty	-1.55	-1.19	-0.73	-0.25	0.27	0.79	1.54
		Infit	0.79	0.85	0.96	1.02	1.12	1.25	1.34
		Outfit	0.63	0.8	0.92	1.04	1.14	1.32	1.59

Note. ES = Elementary School (grades 3–5); MS = Middle School (grades 6–8); HS = High School (grades 9–12).

4.2.6 Data Review Meeting

4.2.6.1 MOU-ALT Shared Items

The MOU-ALT item data review committee reviewed items flagged for undesired statistics. In addition to the statistical flag, CAI flagged and removed the items with a sample size less than 50 or negative biserial/polyserial correlations for the key before the data review. The data review committee did not review these items. Items with sample sizes less than 50 will be re-field tested in future administrations.

The MOU-ALT data review committee comprised staff across MOU states, CAI content specialists, special education specialists, and psychometricians. During the meeting, the committee identified defects that led to the undesired statistics of the items and then decided to reject the item completely, accept the item with modifications for further field testing, or accept the item as it is. Items accepted without modifications were included in the shared MOU-Alt operational items. In addition, Idaho content experts ensured that items from other states were included only if the standards were aligned to Idaho standards.

Table 24 presents a summary of the MOU-ALT data review results.

Table 24. Summary of Item Data Review for MOU-ALT Item Pool

Subject	Grade	Total Number of MOU Items	Items with N < 50	Items with biserial < 0	Total Number of Reviewed Items for IDR	Items Rejected by IDR Committee
ELA	3	32	0	3	10	0
	4	32	0	0	13	0
	5	31	0	1	7	0
	6	29	0	0	9	0
	7	30	0	0	13	1
	8	30	0	2	6	0
	HS	29	0	0	9	0
Mathematics	3	26	0	0	11	0
	4	27	0	2	10	0
	5	27	0	3	12	0
	6	28	0	7	13	0
	7	27	0	4	12	0
	8	27	0	11	10	0
	HS	23	0	3	11	0
Science	ES	74	14	5	20	1
	MS	39	3	0	14	2
	HS	46	8	6	8	0

4.2.6.2 IDAA Item Pool

All IDAA field-test items in spring 2023 were from the MOU-ALT shared items and the Idaho-only items. Idaho state staff confirmed the content alignments for all items in the Idaho item pool and rejected items that did not align to the Idaho Extended Content Standards specified in the test blueprints.

Table 25 presents a summary of the IDAA field-test item pool.

Table 25. Summary of IDAA Field-Test Item Pool

Grade	Total # of Items Administered	Items with n < 50	Items with bis < 0	Rejected Items	Eligible Items
			ELA		
3	32	0	3	5	24
4	32	0	0	2	30
5	31	0	1	4	26
6	29	0	0	3	26
7	30	0	0	4	26
8	30	0	2	0	28
10	29	0	0	3	26
			Mathematics		
3	25	0	0	5	20
4	26	0	2	2	22
5	27	0	3	4	20
6	28	0	7	1	20
7	27	0	4	2	21
8	26	0	10	2	14
10	23	0	3	3	17
			Science		
5	54	3	5	4	42
8	31	0	0	3	28
11	26	0	1	5	20

4.3 Scaling and Equating

Calibration is the process by which we estimate the statistical relationship between item responses and the underlying trait being measured. Traditional item response models assume a single underlying trait and that items are independent given that underlying trait. In other words, the models assume that given the value of the underlying trait, knowing the response to one item provides no information about responses to other items. This basic simplifying assumption allows the likelihood function for these models to take the relatively simple form of a product over items for a single student:

$$L(Z) = \prod_{j=1}^{n} P(z|\theta),$$

where *Z* represents the pattern of item responses, and *θ* represents a student's true proficiency.

Traditional item response models differ only in the form of the function *P*(*Z*). The one-parameter model (1PL; also known as the Rasch model) is used to calibrate MOU-ALT items that are scored either right or wrong, and takes the form of

$$P(X_i = 1|\theta) = \frac{\exp(\theta - b_i)}{1 + \exp(\theta - b_i)},$$

where b_i is the difficulty parameter for item i.

The b parameter is often called the *location* or *difficulty* parameter; the greater the value of b, the greater the item's difficulty. The one-parameter model assumes that the probability of a correct response approaches zero as proficiency decreases toward negative infinity. In other words, the one-parameter model assumes that no guessing occurs. In addition, the one-parameter model assumes that all items are equally discriminating.

For items that have multiple, ordered-response categories (i.e., partial-credit items), MOU-ALT items were calibrated using the Rasch family Masters' (1982) PCM. Under Masters' PCM, the probability of getting a score of x_i on item i given ability θ can be written as

$$P(X_i = x_i|\theta) = \frac{\exp \sum_{k=0}^{x_i}(\theta - b_{ki})}{\sum_{l=0}^{m_i} \exp \sum_{k=0}^{l}(\theta - b_{ki})},$$

with the constraint that $\sum_{k=0}^{0}(\theta - b_{ki}) \equiv 0$. b_{ki} is the item location parameter for category k of item i.

4.3.1 Item Calibration

The design of embedding randomly selected field-test items among operational items produces the item response data in a sparse data matrix. The items in the sparse data matrix are concurrently calibrated for each grade and subject, fixing parameter estimates of the operational items which are calibrated and scaled in prior administrations. The item parameter estimates of field-test items are then placed on the same MOU-ALT scale.

Winsteps is used to estimate the Rasch and Masters' PCM item parameters for MOU-ALT. Winsteps from Mesa Press is publicly available software. Winsteps employs a joint maximum likelihood approach towards estimation (JMLE), which jointly estimates the person and item parameters. The Rasch model estimates the parameters for student responses to dichotomous (0/1 point) items. Masters' (1982) PCM, an extension of the one-parameter Rasch model, which allows for partial credit to be given on items, estimates the responses for polytomous items.

5. VALIDITY

According to the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014, hereafter referred to as the *Standards*), “Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests” (p.11). Statements about validity should refer to interpretations for specified uses, and thus, the validation process logically starts with well-articulated statements on intended uses of test scores (see Section 1.2). Arguments of logical, theoretical, and empirical evidence are then provided to support the intended uses.

A sound validity argument integrates various strands of evidence into a coherent account of the degree to which exiting evidence and theory support the intended interpretation of test scores for specific uses (AERA, APA, & NCME, 2014; p. 21). Validity of an intended interpretation of test scores relies on all the evidence accrued about the technical quality of a testing system, including item and test development and construction procedures, test score reliability, accurate scaling and equating, procedures for setting meaningful performance standards, standardized test administration and scoring procedures, and attention to fairness for all test takers. The appropriateness and usefulness of the IDAA depends on the assessments meeting the relevant standards of validity.

The state is also required to provide sufficient and solid validity evidence to meet federal peer review requirements. In the guidance provided by the U.S. Department of Education for assessing peer review process (U.S. Department of Education, 2018), the requirements related to validity are represented in Critical Element 3.

Validity evidence for the IDAA is gathered from the following four sources, as outlined in the *Standards*; the particular Critical Element in the peer review guidance corresponding to each source is included in the paratheses:

- Evidence based on test content
  (Critical Element 3.1—Overall Validity, Including Validity Based on Content)

- Evidence based on response processes
  (Critical Element 3.2—Validity Based on Cognitive Process/Linguistic Processes)

- Evidence based on internal structure
  (Critical Element 3.3—Validity Based on Internal Structure)

- Evidence based on relations to other variables
  (Critical Element 3.4—Validity Based on Relations to Other Variables)

Evidence on test content validity is provided with both theorical and empirical evidence related to content specifications, test specifications, blueprints, the item and test development process, the administration process, and scoring. Evidence on response processes is gathered by conducting cognitive lab studies of student responses to items. Evidence on internal structure is examined in the results of intercorrelations among content strand scores. Evidence on relations to other variables is provided with the correlations between test scores and the Learner Characteristics Inventory (LCI) questions.

5.1 Evidence Based on Test Content

Content evidence for validity is based on the appropriateness of test content and the procedures used to create test content, which should be well aligned with the required statewide standards implemented by teachers in daily instruction at schools. This evidence is based on the justification for and connections among several factors, including the following:

- Content specifications
- Test blueprints
- Item development
- Test administration conditions
- Item and test scoring

These resources are developed by content and measurement experts and are consistent with state standards. Collectively, they help connect the assessment results to learning and instruction. The descriptions of the evidence, most of which are documented in this technical report, are summarized in this section.

Content Specifications

Content standards and specification are the starting points for test development. The IDAA is aligned with the Idaho Extended State Standards and is designed for students with the most significant cognitive disabilities. The purpose of the IDAA is to maximize access of this student population to the general education curriculum, ensure that all students with disabilities are included in the statewide assessments, and make certain that they are included in the educational accountability system. Refer to Section 1.4, Content Standards, in this technical report for details.

Test Blueprints

Test blueprints specify the content standards to be covered in the test, and the minimum and maximum number of items in each content domain. The goal is to ensure that the test has a balanced representation of items from each content standard.

For the IDAA, each student received 40 operational items that were selected adaptively in each test. In the adaptive item-selection algorithm, item selection occurred in two discrete stages: (1) blueprint satisfaction and (2) match-to-ability. Table 26–Table 28 present the percentages of administered tests (i.e., tests delivered in the Test Delivery System [TDS]) aligned with the test blueprint constraints for ELA, mathematics, and science. The blueprint match rates were based on the completed online tests only.

As shown, the adaptive algorithm selected items for all tests according to the blueprint requirements: 100% match at the overall strand level, except ELA grade 7 and 8 domain language with a matching rate of 97% and 98% respectively.

Item Development

Section 4, Item Development, provides a detailed description on how items are developed. The number and type of items to be developed are based on an evaluation of content needs and available sample size for field testing that can result in reliable statistics. Item writers are carefully chosen and well trained to follow standardized procedures and templates when creating items. All items undergo rigorous multiple rounds of internal and external reviews from the content and fairness perspective before they are field tested in an operational context. After field testing, item analysis is conducted to examine whether items perform as expected. All items are reviewed by special education teachers and content experts in the state before they are moved to the final operational item pool.

Test Administration Conditions

Standardized test administration is critical in producing reliable and valid test scores. Comparability of test scores, whether between students and schools or across time for the same students, is based on standardization of test administration and test scoring rules. If test administrators (TAs) do not follow the same procedures, student performance cannot be compared meaningfully. TAs are required to complete and pass an online TA Certification Course before they can administer the IDAA to their students. The guidelines for test administration are summarized in the Test Administration Manual (TAM). See Section 2, Test Administration, for details.

Item and Test Scoring

Item and test scores are the most critical element. All interpretations are established around students' test results; therefore, every effort is made to ensure absolute accuracy on item and test scores. Section 10.3, Quality Assurance in Test Scoring, provides a detailed description of quality control and monitoring procedures implemented within Cambium Assessment, Inc. (CAI) to assure that accurate scores are generated and reported.

Table 26. Percentage of Administered Tests Meeting Blueprint Requirements for ELA

Grade	Content Strands	Minimum Required Items	Maximum Required Items	% BP Match
3	Language	8	10	100
	Reading—Informational Text	11	13	100
	Reading—Literature	11	13	100
	Writing	8	10	100
4	Language	8	10	100
	Reading—Informational Text	11	13	100
	Reading—Literature	11	13	100
	Writing	8	10	100
5	Language	8	10	100
	Reading—Informational Text	11	13	100
	Reading—Literature	11	13	100
	Writing	8	10	100
6	Language	8	10	100
	Reading—Informational Text	11	13	100
	Reading—Literature	11	13	100
	Writing	8	10	100
7	Language	8	10	97
	Reading—Informational Text	11	13	100
	Reading—Literature	11	13	100
	Writing	8	10	100
8	Language	8	10	98
	Reading—Informational Text	11	13	100
	Reading—Literature	11	13	100
	Writing	8	10	100
10	Language	8	10	100
	Reading—Informational Text	11	13	100
	Reading—Literature	11	13	100
	Writing	8	10	100

Table 27. Percentage of Administered Tests Meeting Blueprint Requirements for Mathematics

Grade	**Content Strands**	**Minimum Required Items**	**Maximum Required Items**	**% BP Match**
3	Data Analysis, Probability, and Statistics (DPS)	5	8	100
	Geometry (GM)	3	5	100
	Measurement (ME)	7	10	100
	Number and Operations (NO)	12	14	100
	Patterns, Relations, and Functions (PRF)	4	6	100
	Symbolic Expression (SE)	2	3	100
4	Data Analysis, Probability, and Statistics (DPS)	5	7	100
	Geometry (GM)	5	7	100
	Measurement (ME)	5	7	100
	Number and Operations (NO)	14	17	100
	Patterns, Relations, and Functions (PRF)	3	5	100
	Symbolic Expression (SE)	2	4	100
5	Data Analysis, Probability, and Statistics (DPS)	5	7	--*
	Geometry (GM)	5	7	100
	Measurement (ME)	6	8	100
	Number and Operations (NO)	11	13	100
	Patterns, Relations, and Functions (PRF)	5	7	100
	Symbolic Expression (SE)	1	2	100
6	Data Analysis, Probability, and Statistics (DPS)	8	10	100
	Geometry (GM)	4	6	100
	Measurement (ME)	6	8	100
	Number and Operations (NO)	10	12	100
	Patterns, Relations, and Functions (PRF)	7	9	100
	Symbolic Expression (SE)	2	4	100
7	Data Analysis, Probability, and Statistics (DPS)	11	13	100
	Geometry (GM)	4	6	100
	Measurement (ME)	5	7	100
	Number and Operations (NO)	9	11	100
	Patterns, Relations, and Functions (PRF)	4	6	100
	Symbolic Expression (SE)	2	3	100
8	Data Analysis, Probability, and Statistics (DPS)	11	13	100
	Geometry (GM)	6	8	100
	Measurement (ME)	3	5	100
	Number and Operations (NO)	6	8	100
	Patterns, Relations, and Functions (PRF)	8	10	100
	Symbolic Expression (SE)	1	2	100
10	Data Analysis, Probability, and Statistics (DPS)	8	10	100
	Geometry (GM)	4	6	100
	Measurement (ME)	6	8	100
	Number and Operations (NO)	8	10	100
	Patterns, Relations, and Functions (PRF)	10	12	100

* There are only three items in the pool.

Table 28. Percentage of Administered Tests Meeting Blueprint Requirements for Science

Grade	Content Strands	Minimum Required Items	Maximum Required Items	% BP Match
5	Earth and Space Sciences (ESS)	11	14	100
	Life Sciences (LS)	10	13	100
	Physical Sciences (PS)	14	17	100
8	Earth and Space Sciences (ESS)	12	15	100
	Life Sciences (LS)	12	15	100
	Physical Sciences (PS)	12	15	100
11	Physical Sciences (PS)	7	10	100
	Earth and Space Sciences (ESS)	10	13	100
	Life Sciences (LS)	19	22	100

5.2 Evidence Based on Response Processes

Cognitive lab studies document validity evidence to show that the assessments measure the intended cognitive processes appropriate for each grade level as represented in the state's alternate academic content standards. Cognitive lab studies explored student performance on items aligned to the state standards in knowledge and skill level. The results of these studies demonstrated students' application of their knowledge and skills.

Students with significant cognitive disabilities represent about 1% of a state's total assessed population. The students who participate in the alternate assessments for students with significant cognitive disabilities represent various disability categories and demonstrate many concomitant learning difficulties. Students in this population can exhibit difficulties in responding to stimuli; challenges committing information to working, short-term, or long-term memory; difficulties generalizing learning to familiar and novel environments; meta-cognition; or self-regulating behaviors. Furthermore, students with significant cognitive disabilities may also demonstrate significant communication or sensory deficits; limited fine or gross motor skill abilities; specialized health care needs; or inability to synthesize learned skills. Students with significant cognitive disabilities require multiple opportunities to engage with academic content and daily activities and multiple ways to express and represent their knowledge.

Although Idaho has not yet had an opportunity to implement a cognitive lab study, results from the cognitive labs in other Memorandum of Understanding (MOU) states who share testing items can also provide insights.

In spring 2019, Hawaii and Wyoming conducted the cognitive lab studies. Students with significant cognitive disabilities at all grade levels from each of the three cognitive levels (i.e., low ability, moderate ability, and high ability) were included in these studies, including 4–5 students per grade. The estimation of low, moderate, or high ability level was determined either by the student's score on the previous year's alternate assessment administration or teacher recommendation. In addition to grade-level and ability-level considerations, students selected for this study represented the Individuals with Disabilities Education Act (IDEA) disability categories with the greatest number of students in each state's significantly cognitively disabled student population, intellectual disability, autism spectrum, and multiple disabilities.

Items from the state's item bank were selected for this study based on their closeness of fit to the cognitive demands of the standard the item was intended to assess. For each ELA, mathematics, and science item for each grade level, CAI content experts and state content experts agreed on the item's alignment to the state standards and the thought processes that the student would have to engage in to answer the question. Five items for each content area and grade level were selected for these studies. Each student within each grade level answered the same five items for ELA, mathematics, and science. All items were based on standards with higher cognitive demands (cognitive demand does not equal Depth of Knowledge [DOK]) so the experts could examine the students who could respond successfully to items at a cognitive level that matched the standards.

The data for these studies were obtained from three sources: student behaviors while responding to each item; student oral responses to questions that asked them to reflect on how they answered each item; and teacher observations about the student's behaviors and their cognitive processing implications. Not all the students in the alternate population were verbal, and not all students had full mobility, and some may have used eye gaze to indicate their responses. Therefore, several different methods had to be used to document their responses and thought processes. The students were video recorded as they interacted with the computer-delivered items so that researchers could return to the video to verify the student's responses. The student's teacher and two observers entered each student's behaviors and oral responses to prompts on a data collection protocol as the student took each item. Following the delivery of each item, the teacher recorded the observed student's behaviors and their interpretation of these behaviors. The student responses to items that matched the cognitive demands and skills included in the aligned standard were collected from all states.

5.3 Evidence Based on Internal Structure

The measurement and reporting model used in the IDAA assumes a single underlying latent trait, with performance reported as a total score. The evidence on the internal structure is examined based on the correlations among content strand scores. The correlations among content strand scores are presented in Table 29–Table 31. The correction for attenuation indicates what the correlation would be if strand scores could be measured with perfect reliability and corrected (adjusted) for measurement error estimates.

The observed correlation between two claim scores with measurement errors can be corrected for attenuation as $r_{x\prime y\prime} = \frac{r_{xy}}{\sqrt{r_{xx}} * \sqrt{r_{yy}}}$, where $r_{x\prime y\prime}$ is the correlation between x and y corrected for attenuation, r_{xy} is the observed correlation between x and y, r_{xx} is the reliability coefficient for x, and r_{yy} is the reliability coefficient for y. Note that when the reliability estimate is negative, the disattenuated correlation cannot be computed. There are a few strands in both ELA and mathematics that have negative reliability estimates due to a small number of items (<=5) in each strand.

When corrected for attenuation (above diagonal), the correlations among strand scores are higher than observed correlations. Note that disattenuated correlation equals 1 if the disattenuated correlation is greater than 1.

The correlations among strand scores are based on operational items. The number of items in each strand varies across students in computer-adaptive tests (CATs). In some strands, the number of items is very small, and the strand scores are less reliable.

Table 29. Correlations Among Strand Scores for ELA

Grade	**Content Strand**	**Observed & Disattenuated Correlation**				
		Strand 1	**Strand 2**	**Strand 3**	**Strand 4**	**Strand 5**
3	Strand 1: Reading—Informational Text	**0.48**	0.75		0.80	0.25
	Strand 2: Reading—Literature	0.36	**0.48**		0.89	0.77
	Strand 3: Reading—Writing Literature	0.15	0.21	**-0.61**		
	Strand 4: Writing	0.39	0.44	0.13	**0.50**	0.80
	Strand 5: Writing—Across All Types	0.11	0.33	0.24	0.34	**0.37**
4	Strand 1: Reading—Informational Text	**0.37**	0.86		1.00	1.00
	Strand 2: Reading—Literature	0.38	**0.52**		0.89	1.00
	Strand 3: Reading—Writing Literature	0.27	0.45	**-0.05**		
	Strand 4: Writing	0.41	0.43	0.18	**0.45**	1.00
	Strand 5: Writing—Across All Types	0.26	0.22	0.30	0.32	**0.05**
5	Strand 1: Reading—Informational Text	**0.46**	1.00		0.85	0.97
	Strand 2: Reading—Literature	0.57	**0.51**		0.89	0.88
	Strand 3: Reading—Writing Literature	0.35	0.27	**-0.27**		
	Strand 4: Writing	0.43	0.48	0.10	**0.56**	0.87
	Strand 5: Writing—Across All Types	0.41	0.39	0.14	0.41	**0.39**
6	Strand 1: Reading—Informational Text	**0.57**	0.89	1.00	0.96	1.00
	Strand 2: Reading—Literature	0.54	**0.64**	1.00	0.94	1.00
	Strand 3: Reading—Writing Literature	0.52	0.49	**0.30**	1.00	1.00
	Strand 4: Writing	0.53	0.55	0.51	**0.54**	1.00
	Strand 5: Writing—Across All Types	0.39	0.31	0.26	0.33	**0.04**
7	Strand 1: Reading—Informational Text	**0.59**	1.00	1.00	0.98	1.00
	Strand 2: Reading—Literature	0.59	**0.57**	1.00	0.93	1.00
	Strand 3: Reading—Writing Literature	0.56	0.59	**0.40**	1.00	1.00
	Strand 4: Writing	0.51	0.47	0.44	**0.46**	1.00
	Strand 5: Writing—Across All Types	0.34	0.40	0.44	0.27	**0.15**
8	Strand 1: Reading—Informational Text	**0.56**	0.93	1.00	0.90	0.60
	Strand 2: Reading—Literature	0.58	**0.69**	1.00	0.92	0.82
	Strand 3: Reading—Writing Literature	0.31	0.43	**0.09**	1.00	1.00
	Strand 4: Writing	0.47	0.54	0.39	**0.48**	1.00
	Strand 5: Writing—Across All Types	0.28	0.43	0.40	0.46	**0.40**
10	Strand 1: Reading—Informational Text	**0.39**	1.00	1.00	1.00	1.00
	Strand 2: Reading—Literature	0.53	**0.51**	1.00	1.00	0.91
	Strand 3: Reading—Writing Literature	0.26	0.34	**0.13**	1.00	1.00
	Strand 4: Writing	0.42	0.46	0.37	**0.42**	0.91
	Strand 5: Writing—Across All Types	0.32	0.30	0.22	0.27	**0.21**

Table 30. Correlations Among Strand Scores for Mathematics

Grade	Strand	Observed & Disattenuated Correlation					
		Strand 1	Strand 2	Strand 3	Strand 4	Strand 5	Strand 6
3	Strand 1: Data Analysis, Probability, and Statistics	**0.28**	0.91	0.80	1.00	1.00	
	Strand 2: Geometry	0.21	**0.19**	1.00	1.00	0.40	
	Strand 3: Measurement	0.33	0.40	**0.61**	0.72	0.43	
	Strand 4: Number and Operations	0.41	0.31	0.37	**0.44**	1.00	
	Strand 5: Patterns, Relations, and Functions	0.36	0.08	0.15	0.37	**0.19**	
	Strand 6: Symbolic Expression	0.16	0.20	0.33	0.38	0.28	**-0.21**
4	Strand 1: Data Analysis, Probability, and Statistics	**-0.08**					
	Strand 2: Geometry	0.32	**0.36**	0.90	1.00		
	Strand 3: Measurement	0.20	0.22	**0.17**	0.48		
	Strand 4: Number and Operations	0.28	0.45	0.15	**0.56**		
	Strand 5: Patterns, Relations, and Functions	0.35	0.40	0.07	0.24	**-0.01**	
	Strand 6: Symbolic Expression	0.20	0.13	0.17	0.13	0.34	**-0.33**
5	Strand 1: Data Analysis, Probability, and Statistics	**-0.28**					
	Strand 2: Geometry	0.24	**0.32**	1.00	0.86	0.93	
	Strand 3: Measurement	0.46	0.36	**0.28**	1.00	1.00	
	Strand 4: Number and Operations	0.30	0.25	0.41	**0.25**	0.89	
	Strand 5: Patterns, Relations, and Functions	0.40	0.28	0.29	0.24	**0.28**	
	Strand 6: Symbolic Expression	-0.02	0.18	-0.06	0.13	0.10	**-0.88**
6	Strand 1: Data Analysis, Probability, and Statistics	**0.22**	1.00	0.97	0.44	0.81	
	Strand 2: Geometry	0.19	**0.08**	1.00	1.00	0.65	
	Strand 3: Measurement	0.33	0.58	**0.51**	1.00	0.41	
	Strand 4: Number and Operations	0.14	0.28	0.50	**0.48**	0.77	
	Strand 5: Patterns, Relations, and Functions	0.22	0.11	0.17	0.30	**0.33**	
	Strand 6: Symbolic Expression	0.09	0.36	0.57	0.45	0.20	**-0.20**
7	Strand 1: Data Analysis, Probability, and Statistics	**0.55**		0.82	0.84	1.00	
	Strand 2: Geometry	0.19	**-0.01**				
	Strand 3: Measurement	0.32	0.16	**0.27**	1.00	1.00	
	Strand 4: Number and Operations	0.43	0.22	0.38	**0.48**	1.00	
	Strand 5: Patterns, Relations, and Functions	0.40	0.10	0.29	0.4	**0.10**	
	Strand 6: Symbolic Expression	0.22	0.17	0.23	0.29	0.27	**-0.25**
8	Strand 1: Data Analysis, Probability, and Statistics	**0.25**	1.00			0.83	
	Strand 2: Geometry	0.28	**0.14**			1.00	
	Strand 3: Measurement	0.28	0.17	**-0.01**			
	Strand 4: Number and Operations	0.25	0.16	0.12	**-0.24**		
	Strand 5: Patterns, Relations, and Functions	0.24	0.33	0.08	0.17	**0.33**	
	Strand 6: Symbolic Expression	-0.10	-0.20	0.06	0.07	0.02	**-1.20**
10	Strand 1: Data Analysis, Probability, and Statistics	**0.39**	0.98	0.70	0.93	0.30	
	Strand 2: Geometry	0.17	**0.08**	0.91	1.00	1.00	
	Strand 3: Measurement	0.26	0.15	**0.36**	1.00	0.29	
	Strand 4: Number and Operations	0.20	0.23	0.24	**0.12**	1.00	
	Strand 5: Patterns, Relations, and Functions	0.12	0.19	0.11	0.27	**0.38**	

Table 31. Correlations Among Strand Scores for Science

Grade	Strand	Observed & Disattenuated Correlation		
		Strand 1	Strand 2	Strand 3
5	Strand 1: Earth and Space Science	**0.57**	0.62	0.81
	Strand 2: Life Science	0.35	**0.54**	0.99
	Strand 3: Physical Science	0.51	0.61	**0.69**
8	Strand 1: Earth and Space Science	**0.59**	1.00	0.91
	Strand 2: Life Science	0.59	**0.52**	0.95
	Strand 3: Physical Science	0.54	0.53	**0.6**
11	Strand 1: Earth and Space Science	**0.46**	0.74	0.52
	Strand 2: Life Science	0.41	**0.67**	1.00
	Strand 3: Physical Science	0.22	0.57	**0.39**

5.4 EVIDENCE BASED ON RELATIONS TO OTHER VARIABLES

In *A State's Guide to the U.S. Department of Education's Assessment Peer Review Process* (U.S. Department of Education, 2018), the U.S. Department of Education listed the results of a correlational study between assessments results (i.e., student test scores) and variables related to test takers as "adequate validity evidence that the state's assessment scores are related as expected with other variables." The Idaho State Department of Education (SDE) and CAI implemented a study that required all teachers of students with severe cognitive disabilities who took the IDAA to complete the Learner Characteristics Inventory (LCI) for each of those students. CAI then analyzed the results and ran a correlational study with several of the LCI questions related to variables of student behaviors that directly impact student performance on the spring 2023 IDAA. The results of this study are discussed following an initial discussion of the purpose and questions included on the LCI.

5.4.1 Learner Characteristics Inventory

The LCI was developed by a committee of experts brought together by the National Center and State Collaborative (NCSC) project across all of the 18 core partner states. NCSC was funded through a four-year General Supervision Enhancement Grant (GSEG) from the Office of Special Education Programs at the U.S. Department of Education. "Its purpose is to create a system of high-quality supports and resources for educators who work with students with the most significant cognitive disabilities" (Towles-Reeves, E., Kearns, J., Flowers, C., Hart, L., Kerbel, A., Kieinert, H., Quenemoen, R., & Thurlow, M., 2012, p. 1). According to these experts, the LCI was based on the work of Pellegrino, Chudowsky, & Glaser, 2001, who defined three pillars on which every assessment must rest: "A model of how students represent knowledge and develop competence in the subject domain, tasks, or situations that allow one to observe students' performance, and an interpretation method for drawing inferences from the performance evidence thus obtained" (p. 2).

The final version of the LCI administered in Idaho comprises 48 questions that a teacher answers about each student administered the IDAA. The LCI results do not affect students' IDAA test scores. Rather, these characteristics, taken together across all students participating in an alternate assessment, help the state understand the characteristics of their population of alternate assessment test-takers.

The questions inquire about the following:

1–4. Participation criteria
5. Student's primary language
6. Student's dominant language
7. Student's expressive communication skills
8. Student's augmentative communication system
9. Student's receptive language skills
10. Student's vision
11. Student's hearing
12. Student's motor skills
13. Student's ability to engage with others
14. Student's health/attendance issues
15. Student's reading skills
16. Student's mathematics skills
17. Student's writing skills
18. Students' accommodations included in the IEP and used during instruction
19–24. Student's skills related to reading text
25–30. Student's skills related to solving mathematics problems
31–36. Student's skills related to science
37. Instructional minutes
38–39. Student's inclusion in general education instruction
40. Parents' educational expectation of the student
41–42. Student's ability to interact with others
43. Student's career aspirations
44. Career skills instruction received
45–48. Student's work experience

The LCI results provide a description of the student. The results also provide a description of the state's students who are classified as having significant cognitive disabilities across all students in that state. The LCI is designed to be a descriptive instrument for the states to use to define this population of students and to then develop participation guidelines for their states' alternate assessments.

While reviewing the questions included in the Idaho LCI, it was observed that several of these questions yielded evidence relevant to the academic performance of these students. These questions inquired about the following:

- Student's expressive communication skills
- Student's receptive language skills
- Student's ability to engage with others
- Student's reading skills
- Student's mathematics skills

The **student's expressive communication skills** question asked teachers to describe the student's oral/written or augmentative communication ability. The following three levels of descriptors were defined:

1. The first, or highest-level, descriptor states that the student uses symbolic language to communicate.
2. The second, or middle-level, descriptor states that the student uses intentional communication but not at a symbolic level.
3. The third, or lowest-level, descriptor states that the student communicates predominately through cries, facial expressions, change in muscle tone, or other indicators.

Students who communicate symbolically are able to respond to items on the assessment and be more successful on an assessment that requires the use of symbolic communication. Students with limited or no symbolic communication skills perform less well on an assessment that relies on symbolic communication. The LCI "expressive communication skills" question will therefore predict, at a broad level, the student's final score on an assessment.

The **student's receptive language skills** question includes the following four levels of descriptors:

1. The first, or highest, descriptor states that the student can independently follow 1–2 step directions presented through words without additional cues.
2. The second descriptor states that the student can follow 1–2 step directions with additional cues.
3. The third descriptor states that the student is receptive and alert to sensory input from another person, but the student requires actual physical assistance to follow simple directions.
4. The fourth, or lowest, descriptor states that the student demonstrates an uncertain response to sensory stimuli.

On an academic assessment, a student must be able to respond independently to directions. Students who can respond independently tend to receive a higher score on an assessment than those who cannot. Therefore, the receptive language descriptors relate to a student's performance on a symbolic-language-based assessment.

The **student's ability to engage with others** (i.e., the students' engagement descriptor) question also has the following four descriptive statements:

1. The first, or highest, descriptor states that the student can initiate and sustain social interactions.
2. The second descriptor describes the student as responding but not initiating social interactions.
3. The third descriptor defines a student who alerts to others.
4. The fourth, or lowest, descriptor defines a student who does not alert to others.

An academic assessment situation is a social interaction, and the computer audio voice reads the questions and options to the student; students who enter social interactions with others—even if they do not initiate the interaction, as this is not necessary on an assessment—would have more chance of success on an assessment than students who do not enter social interactions with others.

The **student's reading skills** descriptor relates directly to the student's reading ability and the student's ability to understand all instruction in the content areas, as much of the instruction requires the student to read; even if the instruction does not require reading letters and words, it may include numbers and operation signs. The reading descriptors progress as follows:

1. Reads fluently with critical understanding
2. Reads fluently with literal understanding
3. Reads basic sight words
4. Is aware of text
5. Demonstrates no observable awareness of print

Students who can read critically will perform better on an assessment than students who read with literal understanding only, and students who read with literal understanding will perform better on an assessment than students who read only sight words. These descriptors seem to have the potential of being predictive of high and low scores on an academic assessment.

The **student's mathematics skills** descriptor relates to mathematics instruction and assessment and any other content areas, such as science or the reading of graphs and charts, that require the use of mathematics or an understanding of numerical values. The mathematics descriptors progress as follows:

1. Applies computation procedures to solve real-life or routine word problems
2. Does computational procedures with or without a calculator
3. Counts to at least 10 with 1:1 correspondence
4. Counts by rote to 5
5. Demonstrates no observable awareness or use of numbers

A student who can apply computational procedures to real-life problems will perform better on an assessment than a student who can do only computational procedures, and a student who can do computational procedures will do better than a student who counts with 1:1 correspondence to 10. Just as with the reading descriptors, the mathematics descriptors also have the potential of being predictive of high and low scores on an academic assessment.

5.4.2 Correlations with LCI Descriptors

The LCI descriptors on Expressive Language, Receptive Language, Engagement, Reading, and Mathematics, and a composite score calculated by adding five LCI descriptors, were correlated with the IDAA scores in ELA, mathematics, and science (estimated by Pearson correlation coefficient). As shown in Table 32, the reading skills and mathematics skills descriptors tend to have higher correlations with reading and mathematics scores than the other three descriptors for most grades. In mathematics grade 8, the correlations between the mathematics scale score and Expressive, Receptive and Ability to Engagement descriptors are among the lowest with values lower than 0.15.

A teacher's description of a student's ability level, as required when completing the LCI, correlates moderately with students' overall score on the IDAA. It provides supporting validity evidence for the assessments, in that the assessment itself reflects the range of student skills in an academic content area with the higher scores correlating with an independent judgment of a higher student skill level.

Table 32. Correlations Between LCI Descriptors and Total IDAA Scores

Grade	**Composite**	**Expressive Communication Skills**	**Receptive Language Skills**	**Ability to Engage with Others**	**Reading Skills**	**Mathematics Skills**
			ELA			
3	0.30	0.37	0.27	0.29	0.15	0.14
4	0.38	0.27	0.21	0.26	0.38	0.48
5	0.33	0.40	0.18	0.15	0.27	0.31
6	0.47	0.29	0.31	0.39	0.37	0.36
7	0.53	0.33	0.29	0.34	0.49	0.51
8	0.34	0.23	0.29	0.30	0.25	0.30
10	0.47	0.41	0.27	0.24	0.49	0.39
			Mathematics			
3	0.25	0.27	0.11	0.05	0.26	0.27
4	0.34	0.26	0.19	0.23	0.34	0.38
5	0.18	0.21	0.10	0.07	0.11	0.15
6	0.24	0.25	0.16	0.05	0.31	0.19
7	0.38	0.18	0.24	0.24	0.31	0.37
8	0.16	0.07	0.15	0.02	0.19	0.13
10	0.31	0.17	0.18	0.11	0.31	0.31
			Science			
5	0.35	0.36	0.22	0.28	0.29	0.27
8	0.34	0.29	0.22	0.31	0.18	0.33
11	0.56	0.34	0.45	0.44	0.46	0.51

* Significant at the .05 level.

6. RELIABILITY

Reliability refers to consistency in test scores. Reliability is evaluated in terms of the 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. Within the item response theory (IRT) framework, measurement error varies conditioning on ability. The amount of precision in estimating performance can be determined by test information, which describes the amount of information provided by the test at each score point along the ability continuum. Test information is a value that is the inverse of the squared measurement error of the test; the larger the measurement error, the less test information is being provided.

Each item in a computer-adaptive test (CAT) is selected based on content values that meet the blueprint and information values that match students' ability. The reliability evidence of the Idaho Alternate Assessment (IDAA) is provided with marginal reliability, SEM, and classification accuracy and consistency in each performance level.

6.1 MARGINAL RELIABILITY

For reliability, the marginal reliability is 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 (CSEM), estimated at different points on the ability scale, for all students.

The marginal reliability ($\bar{\rho}$) is defined as

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

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

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

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

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

Table 33 presents the marginal reliability coefficients and the average CSEM for the total scale scores. Since the analysis is based on completed online test cases, the minimum and maximum items are always equal to the number of operational items. Among the three subjects, reliability of mathematics tests is lower than the other two subjects.

Table 33. Marginal Reliability for ELA, Mathematics, and Science

Grade	Range of Items Used for Score Reports Across Tests		Marginal Reliability	Scale Score Mean	Scale Score SD	Average CSEM
	Min	Max				
English Language Arts (ELA)						
3	40	40	0.74	299.42	38.63	19.82
4	40	40	0.75	288.46	37.08	18.71
5	40	40	0.79	297.82	39.93	18.10
6	40	40	0.85	301.82	46.66	18.34
7	40	40	0.83	298.56	34.92	14.29
8	40	40	0.85	311.76	45.22	17.67
10	40	40	0.76	296.10	38.42	18.63
Mathematics						
3	40	40	0.77	296.18	46.50	22.28
4	40	40	0.70	296.74	36.08	19.65
5	40	40	0.67	278.34	41.65	23.94
6	40	40	0.77	297.79	42.58	20.61
7	40	40	0.75	289.45	44.38	22.22
8	40	40	0.52	290.90	41.97	29.00
10	40	40	0.61	284.91	41.07	25.65
Science						
5	40	40	0.81	287.58	40.70	17.75
8	40	40	0.80	299.01	35.93	16.11
11	40	40	0.76	289.00	36.19	17.64

6.2 STANDARD ERROR CURVES

Figure 8–Figure 10 present plots of the CSEM of scale scores across the range of ability. The vertical lines indicate the cut scores for Basic, Proficient, and Advanced. Overall, the standard error curves suggest that students are measured with a similar precision across the range of score distribution, except for a few outliers with extreme scores.

Figure 8. Conditional Standard Error of Measurement for ELA

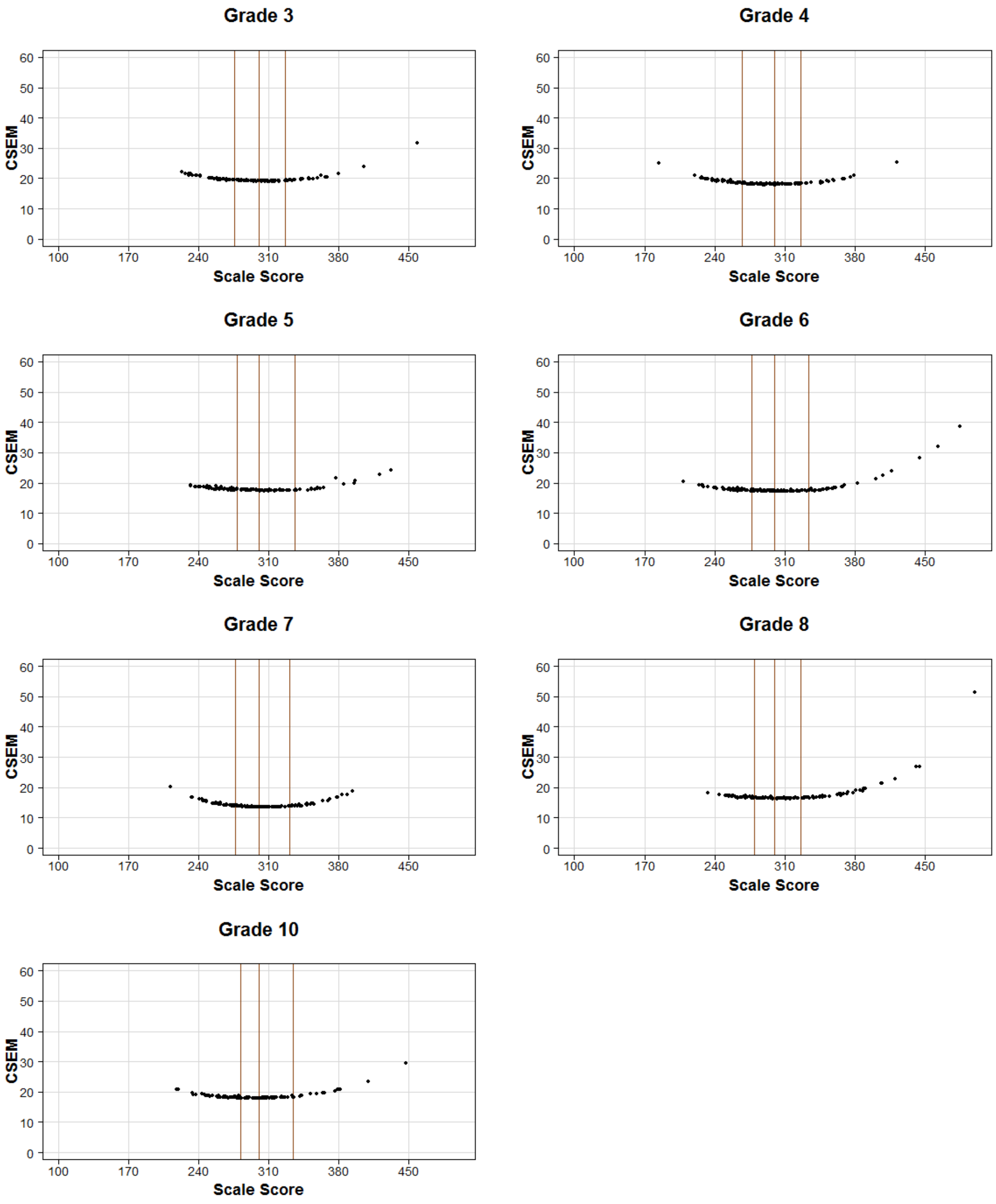


Figure 9. Conditional Standard Error of Measurement for Mathematics

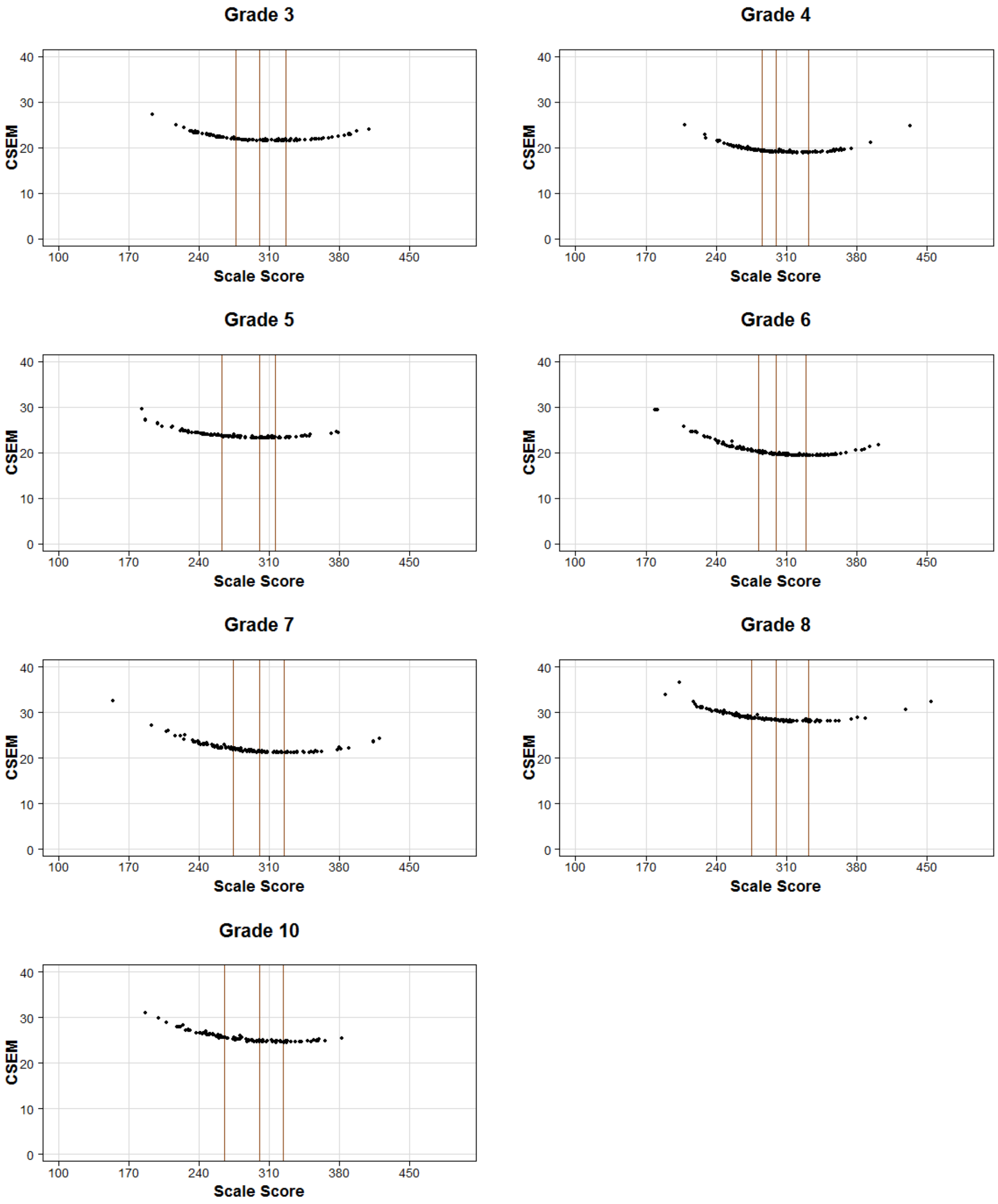
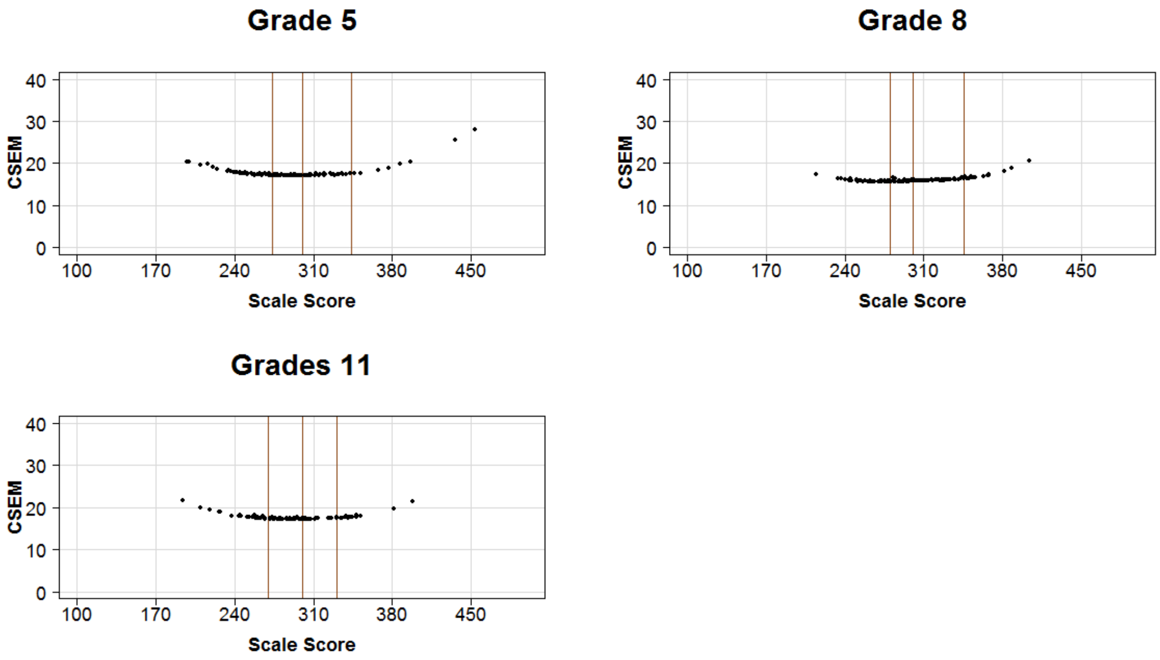


Figure 10. Conditional Standard Error of Measurement for Science



Grade 5

Grade 8

Grades 11

The SEMs presented in the figures are summarized in Table 34, which provides the average CSEM by performance level. As shown in Figure 8–Figure 10, the average CSEMs in Basic and Proficient are similar, but slightly larger in Below Basic and Advanced, which can be expected for tests with extreme scores.

Table 34. Average Conditional Standard Error of Measurement by Performance Level

Grade	Below Basic	Basic	Proficient	Advanced	Average CSEM
ELA					
3	20.28	19.30	19.18	20.53	19.82
4	19.41	18.20	18.20	19.58	18.71
5	18.25	17.75	17.54	19.11	18.10
6	18.12	17.50	17.38	20.31	18.34
7	14.90	13.73	13.60	15.10	14.29
8	16.96	16.49	16.42	19.93	17.67
10	18.60	18.02	18.14	20.59	18.63
Mathematics					
3	22.92	21.70	21.61	22.16	22.28
4	20.23	19.19	19.04	19.57	19.65
5	24.82	23.54	23.33	23.66	23.94
6	22.51	19.97	19.56	19.75	20.61
7	23.32	21.67	21.27	21.67	22.22
8	30.02	28.57	28.16	28.43	29.00
10	26.96	25.17	24.80	24.83	25.65
Science					
5	17.86	17.21	17.37	21.08	17.75
8	15.90	15.88	16.03	17.23	16.11
11	18.07	17.28	17.33	17.99	17.64

6.3 Reliability of Performance Classification

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

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

The classification index can be examined in terms of the classification accuracy and consistency. *Classification accuracy* refers to the agreement between the classifications based on the form 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) taken and the classifications that would be made based on an alternate form (another set of adaptively administered items given the same ability). That is, the percentages of students consistently classified in the same achievement levels on two equivalent test forms.

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

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

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

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

The likelihood function of theta, given a student's item scores, represents the likelihood of the student's ability at that theta value. Integrating the likelihood values over the range of theta at and above the cut score (with proper normalization) represents the probability of the student's latent ability or the true score being at or above that cut score. If a student with estimated theta is below the cut score, a probability of at or above the cut score is an estimate of the chance that this student is misclassified as below the cut score, and 1 minus that probability is the estimate of the chance that the student is correctly classified as below the cut score. Using this logic, we can define various classification probabilities.

If we are interested in only the classification at each cut score (*cut*), the probability of the *i*th student being classified as at or above the cut score given the item scores $\mathbf{z}_i = (z_{i1}, \cdots, z_{iJ})$ and item parameters $\mathbf{b} = (\mathbf{b}_1, \cdots, \mathbf{b}_J)$ with *J* administered items, can be estimated as

$$p_i = P(\theta_i \geq cut|\mathbf{z}, \mathbf{b}) = \frac{\int_{cut}^{+\infty} L(\theta|\mathbf{z},\mathbf{b})d\theta}{\int_{-\infty}^{+\infty} L(\theta|\mathbf{z},\mathbf{b})d\theta},$$

where the likelihood function based on Rasch IRT models is

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

where d stands for dichotomous and p stands for polytomous items; $\mathbf{b}_j = (b_j)$ if the *j*th item is a dichotomous item, and $\mathbf{b}_j = (b_{j1}, \dots, b_{jK_j})$ if the *j*th item is a polytomous item.

Classification Accuracy

Using p_i, we can construct a 2 × 2 table as

$$\begin{pmatrix} n_{a11} & n_{a12} \\ n_{a21} & n_{a22} \end{pmatrix}$$

where $n_{a11} = \sum_{pl_i=\text{below}}(1 - p_i)$, which is the expected number of students below the cut score when the *i*th student's performance level, pl_i, is below the cut score. Similarly, we can define $n_{a12} = \sum_{pl_i=\text{below}} p_i$, $n_{a21} = \sum_{pl_i=\text{at or above}}(1 - p_i)$, and $n_{a22} = \sum_{pl_i=\text{at or above}} p_i$. In the aforementioned table, the row represents the observed level, and the column represents the expected level.

The classification accuracy (*CA*) at or above the cut score is estimated by

$$CA_{\text{at or above}} = \frac{n_{a22}}{n_{a21}+n_{a22}},$$

the classification accuracy below the cut score is estimated by

$$CA_{\text{below}} = \frac{n_{a11}}{n_{a11}+n_{a12}},$$

and the overall classification accuracy for the cut score is estimated by

$$CA = \frac{n_{a22}+n_{a11}}{n_{a21}+n_{a22}+n_{a11}+n_{a12}}.$$

Classification Consistency

Using p_i, which is similar to accuracy, we can construct another 2 × 2 table by assuming the test is administered twice independently to the same student group, hence we have

$$\begin{pmatrix} n_{c11} & n_{c12} \\ n_{c21} & n_{c22} \end{pmatrix}$$

where $n_{c11} = \sum_{i=1}^{N}(1 - p_i)(1 - p_i)$, $n_{c12} = \sum_{i=1}^{N}(1 - p_i)p_i$, $n_{c21} = \sum_{i=1}^{N} p_i(1 - p_i)$, and $n_{c22} = \sum_{i=1}^{N} p_i p_i$. In each of the above four equations, the first and the second probabilities are the probabilities of the *i*th student being classified at one of two things: (1) below or (2) at or above the cut score, respectively, based on observed scores and hypothetical scores from an equivalent test form.

The classification consistency (*CC*) at or above the cut score is estimated by

$$CC_{\text{at or above}} = \frac{n_{c22}}{n_{c21}+n_{c22}},$$

the classification consistency below the cut score is estimated by

$$CC_{\text{below}} = \frac{n_{c11}}{n_{c11}+n_{c12}},$$

and the overall classification consistency is

$$CC = \frac{n_{c22}+n_{c11}}{n_{c21}+n_{c22}+n_{c11}+n_{c12}}.$$

The analysis of the classification index is performed based on overall scale scores.

Table 35 shows classification accuracy and consistency indexes for the spring 2023 IDAA tests. Accuracy classifications are slightly higher than the consistency classifications in all performance standards. The consistency classification rate can be somewhat lower than the accuracy rate because consistency assumes two test scores, both of which include measurement error; however, the accuracy index assumes only a single test score and a true score, which does not include measurement error.

Table 35. Classification Accuracy and Consistency for Performance Standards

Grade	Accuracy			Consistency		
	Basic	Proficient	Advanced	Basic	Proficient	Advanced
			ELA			
3	0.87	0.84	0.88	0.82	0.78	0.83
4	0.87	0.86	0.91	0.81	0.80	0.88
5	0.86	0.87	0.93	0.80	0.82	0.90
6	0.87	0.88	0.93	0.82	0.83	0.89
7	0.88	0.88	0.94	0.84	0.83	0.91
8	0.89	0.86	0.91	0.84	0.81	0.87
10	0.85	0.85	0.93	0.79	0.79	0.90
			Mathematics			
3	0.88	0.88	0.89	0.82	0.83	0.85
4	0.83	0.85	0.91	0.77	0.78	0.88
5	0.84	0.85	0.88	0.79	0.79	0.83
6	0.88	0.85	0.88	0.83	0.79	0.83
7	0.84	0.86	0.91	0.78	0.80	0.87
8	0.81	0.81	0.87	0.74	0.74	0.82
10	0.85	0.83	0.88	0.79	0.77	0.83
			Science			
5	0.83	0.88	0.96	0.78	0.82	0.95
8	0.86	0.88	0.94	0.82	0.83	0.91
11	0.86	0.86	0.92	0.81	0.81	0.89

6.4 Reliability of Content Strand Scores

Although only the overall score is reported for IDAA tests, the marginal reliability coefficients and the measurement errors are also computed for strand scores. The reliabilities for strand scores are similar in

ELA and science but lower in mathematics. When the number of items in a strand is very low (e.g., <= 5), the marginal reliability will be low and, in certain cases, negative.

Table 36. Marginal Reliability Coefficients for Content Strand Scores for ELA

Grade	**Strand**	**Blueprints Requirement**		**Marginal Reliability**	**Scale Score Mean**	**Scale Score SD**	**Average CSEM**
		Min	**Max**				
3	Reading—Informational Text	11	11	0.48	295.47	58.66	42.24
	Reading—Literature	11	11	0.48	301.63	57.19	41.40
	Reading—Writing Literature	1	5	-0.61	302.42	76.43	97.07
	Writing	8	9	0.50	296.92	68.91	48.80
	Writing—Across All Types	5	8	0.37	298.92	62.92	49.93
4	Reading—Informational Text	11	12	0.37	281.75	47.54	37.69
	Reading—Literature	11	12	0.52	287.31	55.32	38.27
	Reading—Writing Literature	3	5	-0.05	285.81	78.93	81.03
	Writing	9	10	0.45	287.56	53.63	39.93
	Writing—Across All Types	4	5	0.05	298.26	61.87	60.37
5	Reading—Informational Text	11	12	0.46	297.18	48.22	35.54
	Reading—Literature	11	13	0.51	298.41	51.22	35.71
	Reading—Writing Literature	2	5	-0.27	300.13	80.36	90.62
	Writing	8	10	0.56	293.10	61.98	41.12
	Writing—Across All Types	4	7	0.39	297.98	72.87	56.95
6	Reading—Informational Text	11	12	0.57	297.78	55.70	36.58
	Reading—Literature	11	13	0.64	294.85	62.74	37.82
	Reading—Writing Literature	4	6	0.30	306.63	74.10	62.11
	Writing	8	10	0.54	304.84	61.26	41.47
	Writing—Across All Types	3	5	0.04	311.60	70.47	68.90
7	Reading—Informational Text	11	12	0.59	295.56	47.42	30.19
	Reading—Literature	11	12	0.57	303.38	43.77	28.74
	Reading—Writing Literature	5	6	0.40	290.97	64.53	50.15
	Writing	8	9	0.46	296.47	43.16	31.79
	Writing—Across All Types	4	5	0.15	302.33	57.31	52.95
8	Reading—Informational Text	11	11	0.56	313.58	51.90	34.37
	Reading—Literature	11	13	0.69	310.48	68.69	37.97
	Reading—Writing Literature	3	4	0.09	302.65	74.18	70.80
	Writing	8	9	0.48	310.70	59.59	42.75
	Writing—Across All Types	5	7	0.40	318.51	66.34	51.16
10	Reading—Informational Text	11	11	0.39	299.70	47.52	37.31
	Reading—Literature	11	12	0.51	292.33	54.49	38.21
	Reading—Writing Literature	3	4	0.13	307.70	76.36	71.31
	Writing	8	10	0.42	294.83	53.82	40.95
	Writing—Across All Types	5	5	0.21	284.34	69.78	62.03

Table 37. Marginal Reliability Coefficients for Content Strand Scores for Mathematics

Grade	Strand	Number of Items Used in Score Reports		Marginal Reliability	Scale Score Mean	Scale Score SD	Average CSEM
		Min	Max				
3	Data Analysis, Probability, and Statistics	5	6	0.28	299.23	84.37	71.64
	Geometry	4	4	0.19	285.30	93.05	83.96
	Measurement	9	9	0.61	294.33	85.82	53.42
	Number and Operations	13	14	0.44	295.20	51.91	38.80
	Patterns, Relations, and Functions	5	6	0.19	286.95	81.68	73.53
	Symbolic Expression	3	3	-0.21	309.30	85.09	93.58
4	Data Analysis, Probability, and Statistics	5	5	-0.08	294.78	63.02	65.36
	Geometry	6	6	0.36	304.83	71.48	57.39
	Measurement	6	7	0.17	296.75	61.18	55.72
	Number and Operations	15	16	0.56	291.68	48.45	32.30
	Patterns, Relations, and Functions	4	5	-0.01	289.60	73.93	74.39
	Symbolic Expression	3	3	-0.33	293.91	73.71	85.11
5	Data Analysis, Probability, and Statistics	3	3	-0.28	269.30	95.99	108.76
	Geometry	7	7	0.32	279.89	75.84	62.35
	Measurement	8	8	0.28	276.46	68.33	57.85
	Number and Operations	13	13	0.25	279.31	49.52	42.73
	Patterns, Relations, and Functions	7	7	0.28	267.45	74.65	63.47
	Symbolic Expression	2	2	-0.88	303.84	90.92	124.65
6	Data Analysis, Probability, and Statistics	8	8	0.22	284.79	58.01	51.10
	Geometry	5	5	0.08	281.33	76.09	72.87
	Measurement	6	7	0.51	294.40	86.16	60.15
	Number and Operations	11	11	0.48	300.63	56.73	40.93
	Patterns, Relations, and Functions	7	7	0.33	298.82	70.51	57.75
	Symbolic Expression	2	3	-0.20	323.54	89.79	98.43
7	Data Analysis, Probability, and Statistics	11	12	0.55	291.97	65.16	43.92
	Geometry	5	5	-0.01	282.88	69.66	69.89
	Measurement	5	6	0.27	288.51	85.56	72.86
	Number and Operations	10	11	0.48	285.34	66.42	47.90
	Patterns, Relations, and Functions	5	6	0.10	283.92	79.32	75.36
	Symbolic Expression	3	3	-0.25	279.73	84.66	94.84
8	Data Analysis, Probability, and Statistics	11	12	0.25	290.91	64.36	55.92
	Geometry	6	6	0.14	281.56	87.65	81.35
	Measurement	3	4	-0.01	300.86	111.94	112.41
	Number and Operations	7	8	-0.24	283.71	65.94	73.48
	Patterns, Relations, and Functions	9	10	0.33	293.89	80.74	65.88
	Symbolic Expression	2	2	-1.20	279.44	106.10	157.19
10	Data Analysis, Probability, and Statistics	8	9	0.39	286.13	75.09	58.86
	Geometry	4	4	0.08	281.18	97.73	93.89
	Measurement	6	7	0.36	260.23	92.81	74.47
	Number and Operations	9	9	0.12	290.84	61.19	57.25
	Patterns, Relations, and Functions	11	12	0.38	280.69	64.92	51.00

Table 38. Marginal Reliability Coefficients for Content Strand Scores for Science

Grade	Strand	Number of Items Used in Score Reports		Marginal Reliability	Scale Score Mean	Scale Score SD	Average CSEM
		Min	Max				
5	Life Science	13	14	0.57	284.79	49.12	32.02
	Earth and Space Science	10	11	0.54	295.91	54.01	36.73
	Physical Science	15	17	0.69	281.97	54.05	29.83
8	Life Science	12	14	0.59	299.65	44.96	28.95
	Earth and Space Science	12	14	0.52	298.50	41.48	28.64
	Physical Science	13	14	0.60	298.06	46.59	29.54
11	Life Science	10	12	0.46	281.07	46.79	34.30
	Earth and Space Science	20	22	0.67	287.29	43.31	24.93
	Physical Science	7	8	0.39	301.71	57.79	45.04

7. SCORING

For the Idaho Alternate Assessment (IDAA), each student receives an overall scale score and an overall performance level; no subscores are reported. This section describes the rules used in generating overall scores.

7.1 Attemptedness Rules for Scoring

Scores are generated for attempted tests only. For an operational test to be considered attempted for scoring, a student must respond to at least one operational item, or the test administrator (TA) marks *No Response* (NR) for at least one item. Score reports are generated for all attempted tests only. In scoring item responses, NR is a valid response and scored as 0. If a student has four consecutive NRs for the first four items indicating no mode of communication, the test stops, and the student receives the lowest obtainable scale score (LOSS).

7.2 Estimating Student Ability Using Maximum Likelihood Estimation

The IDAA is scored using the maximum likelihood estimate (MLE). The likelihood function for generating the MLEs is based on a mixture of item score points.

Indexing items by *i*, the likelihood function based on the *j*th person's score pattern for *I* items is

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

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

Depending on the item score points, the probability $p_{ij}(z_{ij}|\theta_j, b_i, \dots, b_{i,m_i})$ takes either the form of the Rasch model for items with one point or the form based on the partial credit model (PCM) for items with two or more points.

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

$$p_{ij}(z_{ij}|\theta_j, b_{i,1}) = \begin{cases} \dfrac{exp\left((\theta_j - b_{i,1})\right)}{1 + exp\left((\theta_j - b_{i,1})\right)}, & if\ z_{ij} = 1 \\ \dfrac{1}{1 + exp\left((\theta_j - b_{i,1})\right)}, & if\ z_{ij} = 0 \end{cases}$$

and in the case of items with two or more points,

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

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

The MLE theta is then estimated by finding the value of theta that maximizes the log likelihood, or,

$$\hat{\theta}_j = \text{argmax} \log\left(L_j\left(\theta_j \middle| \mathbf{z}_j, \mathbf{b}_1, \dots, \mathbf{b}_I\right)\right).$$

Standard Error of Measurement

With MLE, the standard error measurement (SEM) 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:

$$I(\theta_j) = \sum_{i=1}^{I}\left(\frac{\sum_{l=1}^{m_i} l^2 Exp\left(\sum_{k=1}^{l}\left(\theta_j - b_{i,k}\right)\right)}{s_{ij}\left(\theta_j, b_{i,1}, \dots, b_{i,m_i}\right)} - \left(\frac{\sum_{l=1}^{m_i} l Exp\left(\sum_{k=1}^{l}\left(\theta_j - b_{i,k}\right)\right)}{s_{ij}\left(\theta_j, b_{i,1}, \dots, b_{i,m_i}\right)}\right)^2\right),$$

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

7.3 Rules for Transforming Theta to Scale Scores

The student's performance in each subject is summarized with an overall test score referred to as a *scale score*. The number of items a student answers correctly and the difficulty of the items presented are used to statistically transform theta scores to scale scores so that scores from different sets of items can be compared meaningfully. The scale scores represent a linear transformation of the ability estimates (theta scores) using the formula $SS = a * \theta + b$. Table 39 presents the scaling constants (i.e., slope and intercept) for each test for the theta-to-scale score linear transformation. Scale scores are rounded to the nearest integer.

Table 39. Scaling Constants on the Reporting Metric

Subject	Grade	Slope (a)	Intercept (b)
ELA	3	59.0955	314.5867
	4	55.5391	314.5631
	5	53.7878	321.4206
	6	52.8890	310.3006
	7	41.7220	313.8509
	8	50.6408	320.2352
	10	55.2314	318.0673
Mathematics	3	66.4632	321.8955
	4	58.5650	325.6774
	5	71.4949	319.9075
	6	60.0227	331.6446
	7	65.4896	331.3325
	8	87.0955	346.0162
	10	76.1369	337.0709
Science	5	52.2393	319.1104
	8	48.2369	315.0874
	11	52.9720	313.5356

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 performance levels (Below Basic, Basic, Proficient, Advanced). Table 40 provides the range of scale scores in each achievement level by grade and subject.

Table 40. Range of Scale Scores by Performance Level

Subject	Grade	Below Basic	Basic	Proficient	Advanced
ELA	3	100–275	276–299	300–325	326–500
	4	100–266	267–299	300–325	326–500
	5	100–277	278–299	300–335	336–500
	6	100–276	277–299	300–333	334–500
	7	100–276	277–299	300–330	331–500
	8	100–279	280–299	300–325	326–500
	10	100–281	282–299	300–333	334–500
Mathematics	3	100–276	277–299	300–325	326–500
	4	100–285	286–299	300–331	332–500
	5	100–262	263–299	300–315	316–500
	6	100–281	282–299	300–328	329–500
	7	100–273	274–299	300–324	325–500
	8	100–274	275–299	300–331	332–500
	10	100–264	265–299	300–323	324–500
Science	5	100–272	273–299	300–343	344–500
	8	100–279	280–299	300–345	346–500
	11	100–269	270–299	300–330	331–500

7.4 Lowest/Highest Obtainable Scale Scores

Extremely unreliable estimates of student ability are truncated to the lowest obtainable scale score (LOSS) or the highest obtainable scale score (HOSS). For the IDAA, the minimum and maximum scale scores are set at 100 and 500. For the overall scale scores, scale scores lower than 100 or higher than 500 are truncated to 100 or 500. The standard error for LOSS and HOSS is computed based on the estimated theta scores derived from the answered items.

7.5 Scoring All Correct and All Incorrect Cases

With item response theory (IRT) and MLE methods, all incorrect and correct scores are assigned the ability of minus and plus infinity. All incorrect tests are scored by adding 0.3 to an item score among the administered operational items for a test. All correct tests are scored by subtracting 0.3 from an item score among the administered operational items for a student.

8. PERFORMANCE STANDARDS

In summer 2022, following the close of the operational testing window, Cambium Assessment, Inc. (CAI) convened panels of Idaho educators to recommend performance standards on each of the Idaho Alternate Assessment (IDAA) tests. The details of the panels, procedures, and outcomes were documented in the *Idaho Alternate Assessment Standard Setting Technical Report*.

This section briefly describes the procedures used by educators to recommend standards and resulting performance standards. Performance standards used in spring 2023 score reporting are based on the spring 2022 standard setting results.

8.1 STANDARD-SETTING PROCEDURES

Student performance on the IDAA is classified into four performance levels: Below Basic, Basic, Proficient, and Advanced. Interpretation of the IDAA test scores rests fundamentally on how test scores relate to performance standards that define the extent to which students have achieved the expectations defined according to the Idaho Extended Content Standards. The cut score establishing the Proficient performance level is the most critical score because it indicates that students are meeting grade-level expectations for performance of the Idaho Extended Content Standards, that they are prepared to benefit from instruction at the next grade level, and that they are on track to enter the workforce. Procedures used to adopt performance standards for the IDAA are therefore central to the validity of test score interpretations.

Following the operational administration of the IDAA in spring 2022, a standard-setting workshop was conducted to recommend a set of performance standards for reporting student performance of the Idaho Extended Content Standards. The workshop comprised a series of standardized and rigorous procedures that Idaho educators serving as standard-setting panelists followed to recommend performance standards. The workshops employed the Bookmark procedure (Mitzel, Lewis, Patz, & Green, 2001), a widely used method where standard-setting panelists used their expert knowledge of the Extended Standards and student performance to map the Performance-Level Descriptors (PLDs) to an ordered-item booklet (OIB) based on the operational test administered in spring 2022.

The panelists also reviewed the corresponding content standards and PLDs for each test. With this information in mind, the panelists selected pages in the OIB that best represented the cut scores on the test. The Bookmark standard-setting process was described in a standard-setting plan submitted to the Idaho State Department of Education (SDE). The plan was reviewed by the Idaho Technical Advisory Committee (TAC) and approved by the SDE prior to the workshop.

Panelists were also provided with contextual information to help inform their primarily content-driven, cut-score recommendations. Benchmarking provided panelists with an external reference that allowed them to gauge how their recommendations compared with standards in other similar assessments or populations. For Idaho, the 2022 general education assessment scores provided benchmark data that panelists could use to evaluate and adjust their bookmarks. By comparing each round's results against the percentage proficient on the general education tests, panelists could judge the reasonableness and rigor of the proposed performance standards for the alternate tests.

Panelists were also provided with feedback about the vertical articulation of their recommended performance standards to evaluate how the locations of their recommended cut scores for each grade-level assessment related to the cut-score recommendations at the other grade levels. This approach allowed panelists to view their cut-score recommendations as a coherent system of performance standards and further reinforced the interpretation of test scores as indicating the performance of current grade-level standards and preparedness to benefit from instruction in the subsequent grade level.

8.2 Performance-Level Descriptors

A prerequisite to standard setting is determining the nature of the categories into which students are classified. These categories, or performance levels, are associated with PLDs. PLDs link the extended standards to the performance standards. There are the following four types of PLDs:

- **Policy PLDs.** These PLDs describe the policy goals of each performance level, which do not vary across grades or content.
- **Range PLDs.** These PLDs, also called Instructional PLDs, describe what students know and are able to do throughout the range of each performance level. For example, the Instructional PLD for "Basic" describes what students know and can do at that level up to just below the "Proficient" cut score. Additional information about the Instructional (Range) PLDs for the IDAA can be found on the SDE website.
- **"Just Barely" PLDs.** These PLDs are sometimes called "threshold" or "target" PLDs. "Just Barely" PLDs are created during the standard-setting workshop and are used for standard setting only. The "Just Barely" PLDs describe what a student just barely scoring at the bottom of each performance level knows and can do.
- **Reporting PLDs.** These are abbreviated PLDs (typically 350 or fewer characters in length) created following standard setting and are used to describe what students know and can do on the score reports.

The standard-setting panelists used Range PLDs and Just Barely PLDs in the workshop.

8.3 Recommended Performance Standards

Panelists were tasked with recommending three performance standards (Basic, Proficient, and Advanced) that resulted in four performance levels (Below Basic, Basic, Proficient, and Advanced). Table 41 presents the performance standard associated with panelist-recommended OIB page numbers in scale scores and the percentage of students classified as meeting or exceeding each standard.

Table 41. Final Recommended Performance Standards for IDAA

Subject/ Grade	Cut Scores			Impact Data			Benchmark Data
	Basic	Proficient	Advanced	Basic	Proficient	Advanced	Proficient
ELA							
3	276	300	326	75%	49%	25%	49%
4	267	300	326	76%	49%	20%	52%
5	278	300	336	77%	54%	25%	57%
6	277	300	334	71%	48%	16%	53%
7	277	300	331	77%	51%	26%	58%
8	280	300	326	73%	51%	30%	54%
10	282	300	334	73%	53%	23%	61%
Mathematics							
3	277	300	326	71%	49%	22%	51%
4	286	300	332	62%	47%	19%	49%
5	263	300	316	74%	49%	24%	43%
6	282	300	329	64%	49%	21%	41%
7	274	300	325	64%	41%	22%	42%
8	275	300	332	65%	41%	15%	36%
10	265	300	324	74%	41%	21%	33%
Science							
5	273	300	344	76%	46%	14%	43%
8	280	300	346	72%	50%	13%	41%
11	270	300	331	70%	40%	11%	38%

9. REPORTING AND INTERPRETING SCORES

The Centralized Reporting System (CRS) generates a set of online score reports that includes the information describing student performance for students, parents, educators, and other stakeholders. The online score reports are generally produced immediately after students complete the tests. Because the performance score report is updated each time a student completes a test, authorized users (e.g., school principals, teachers) can have quickly available information on students' performance scores and use them to improve student learning. In addition to individual student reports (ISRs), the CRS also produces aggregate score reports by classes, schools, districts, and states. The timely accessibility of aggregate score reports can help users monitor students' performance in each subject by grade area, evaluate the effectiveness of instructional strategies, and inform the adoption of strategies to improve student learning and teaching during the school year.

This section describes the types of scores reported in the CRS and a description of the ways to interpret and use these scores in detail.

9.1 CENTRALIZED REPORTING SYSTEM FOR STUDENTS AND EDUCATORS

9.1.1 Types of Online Score Reports

The CRS is designed to help educators and students answer questions about how well students have performed on the Idaho Alternate Assessment (IDAA) English language arts (ELA), mathematics, and science tests. The CRS is the online tool that provides educators and other stakeholders with timely, relevant score reports. The CRS for the alternate assessment has been designed with stakeholders, who are not technical measurement experts, in mind to make score reports easier to read. This is achieved by using simple language that allows users to quickly understand assessment results and make inferences about student performance. The CRS is also designed to present student performance in a uniform format. For example, similar colors are used for groups of similar elements, such as performance levels, throughout the design. This design strategy allows readers to compare similar elements and to avoid comparing dissimilar elements.

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

Generally, the CRS provides two categories of online score reports: (1) aggregate score reports and (2) student score reports. Table 42 summarizes the types of online score reports available at the aggregate level and the individual student level. Detailed information about the online score reports and instructions on how to navigate the CRS can be found in the *Centralized Reporting System User Guide*, located via the help button on the CRS website.

Table 42. Types of Online Score Reports by Level of Aggregation

Level of Aggregation	Types of Online Score Reports
State District School Teacher Roster	• Number of students tested and percentage of proficient students (for overall students and by subgroup) • Average scale score (for overall students and by subgroup) • Percentage of students at each performance level on the overall test (for overall students and by subgroup) • Counts of students at each performance level on the overall test (for overall students and by subgroup) • On-demand student roster report
Student	• Total scale score • Performance level on overall score with Performance-Level Descriptors (PLDs) • Average scale scores for student's school, district, and state

Aggregate score reports at a selected aggregate level are provided for overall students and by subgroup. Users can view student assessment results by any of the subgroups. Table 43 presents the types of subgroup and subgroup categories provided in the CRS.

Table 43. Types of Subgroups

Subgroup	Subgroup Category	
Gender	Female Male	
Ethnicity	American Indian or Alaskan Native Asian Black or African American Hispanic or Latino Native Hawaiian or Other Pacific Islander White Multi-Racial	
Migrant Student	Migrant Migrant—N	

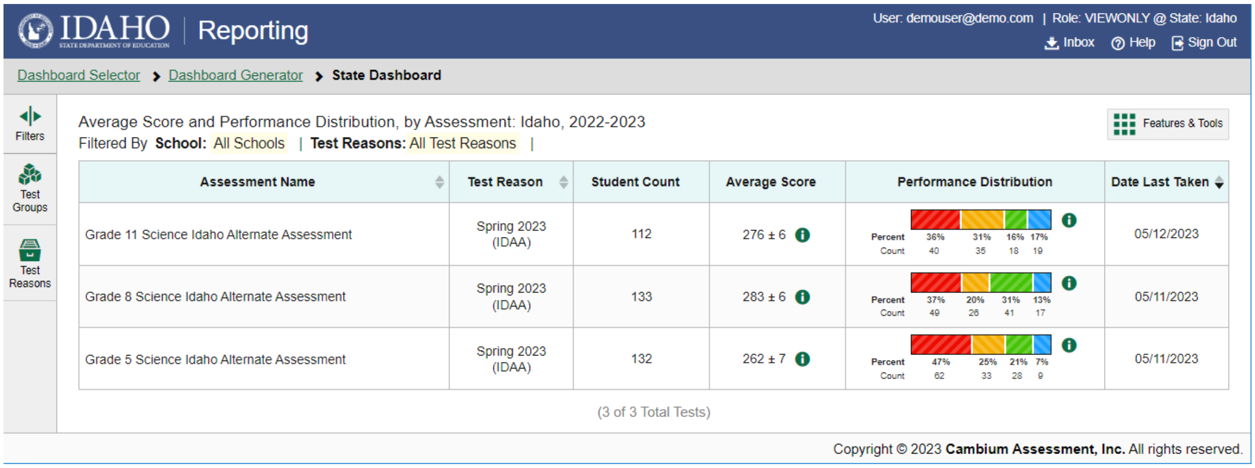
9.1.2 Centralized Reporting System

9.1.2.1 Home Page

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

The home page summarizes students' performance, including the number of students tested and the percentage of proficient students. Exhibits 1 and 2 present a sample of home pages at the state and district levels, respectively.

Exhibit 1. Home Page: State Level



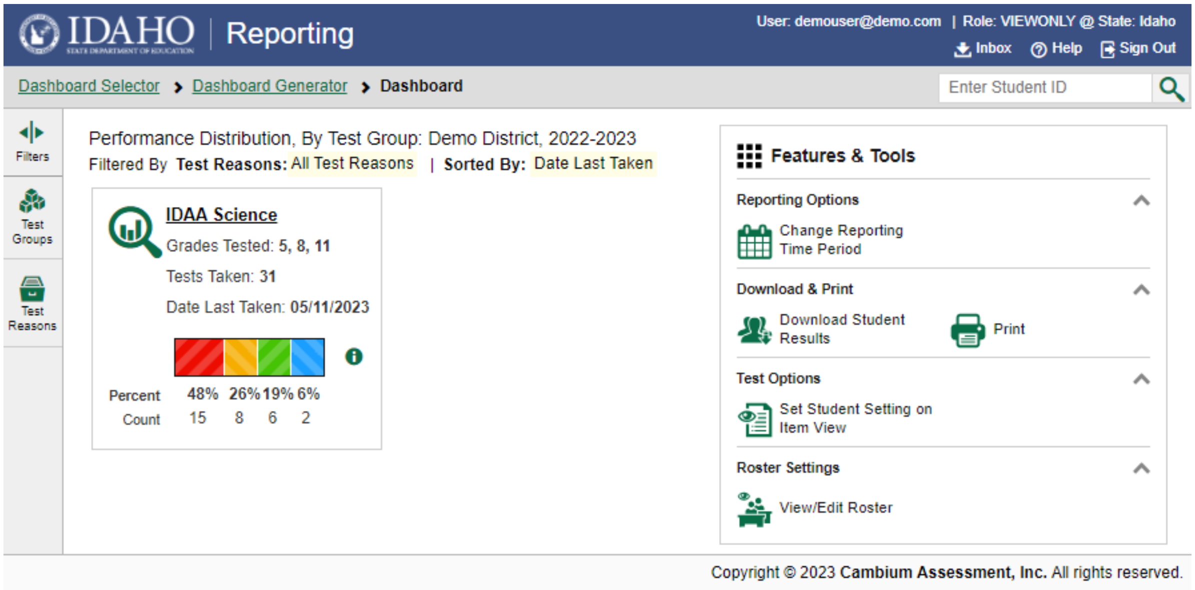
Average Score and Performance Distribution, by Assessment: Idaho, 2022-2023
Filtered By **School:** All Schools | **Test Reasons:** All Test Reasons |

Assessment Name	Test Reason	Student Count	Average Score	Performance Distribution	Date Last Taken
Grade 11 Science Idaho Alternate Assessment	Spring 2023 (IDAA)	112	276 ± 6	Percent 36% 31% 16% 17% Count 40 35 18 19	05/12/2023
Grade 8 Science Idaho Alternate Assessment	Spring 2023 (IDAA)	133	283 ± 6	Percent 37% 20% 31% 13% Count 49 26 41 17	05/11/2023
Grade 5 Science Idaho Alternate Assessment	Spring 2023 (IDAA)	132	262 ± 7	Percent 47% 25% 21% 7% Count 62 33 28 9	05/11/2023

(3 of 3 Total Tests)

Copyright © 2023 **Cambium Assessment, Inc.** All rights reserved.

Exhibit 2. Home Page: District Level



Performance Distribution, By Test Group: Demo District, 2022-2023
Filtered By **Test Reasons:** All Test Reasons | **Sorted By:** Date Last Taken

IDAA Science
Grades Tested: **5, 8, 11**
Tests Taken: **31**
Date Last Taken: **05/11/2023**

Percent	48%	26%	19%	6%
Count	15	8	6	2

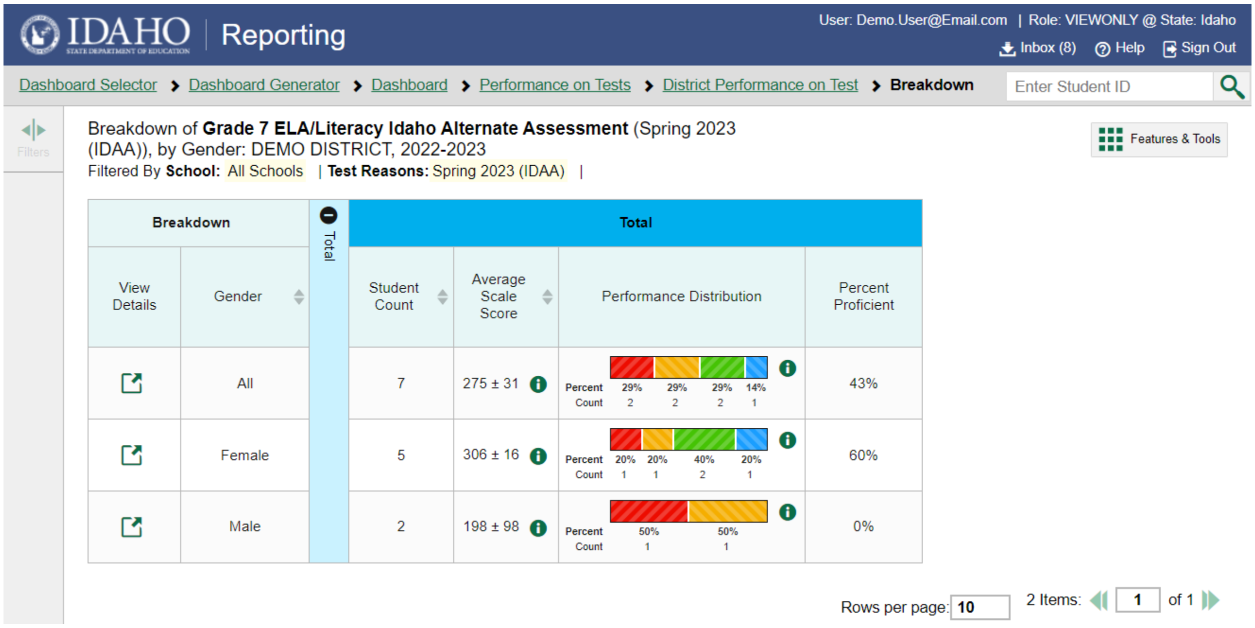
Copyright © 2023 **Cambium Assessment, Inc.** All rights reserved.

9.1.2.2 Subject Detail Page

More detailed summaries of student performance in each grade on a subject area for a selected aggregate level are presented when users select a grade within a subject on the home page. On each aggregate report, the summary report presents the summary results for the selected aggregate unit and the summary results for the state and all aggregate units above the selected aggregate. The CRS reports summaries for all aggregate levels above the selected aggregate level. For example, summaries appear for the teacher, school, district, and state aggregates at the roster level. Roster performance can be compared with the above-aggregate levels.

The subject detail page provides the aggregate summaries on a specific subject area, including student count, average scale score, percentage of proficient students, percentage of students at each performance level, and counts of students at each performance level. The summaries are also presented for overall students and by subgroup. Exhibit 3 shows an example of a subject detail page for ELA at the district level when a user selects a gender subgroup.

Exhibit 3. Subject Detail Page for English Language Arts by Gender: District Level


IDAHO
STATE DEPARTMENT OF EDUCATION
Reporting
User: Demo.User@Email.com | Role: VIEWONLY @ State: Idaho
Inbox (8)
Help
Sign Out
Dashboard Selector > Dashboard Generator > Dashboard > Performance on Tests > District Performance on Test > Breakdown
Enter Student ID
Filters
Features & Tools

Breakdown of **Grade 7 ELA/Literacy Idaho Alternate Assessment** (Spring 2023 (IDAA)), by Gender: DEMO DISTRICT, 2022-2023
Filtered By **School:** All Schools | **Test Reasons:** Spring 2023 (IDAA) |

Breakdown		Total	Total			
View Details	Gender		Student Count	Average Scale Score	Performance Distribution	Percent Proficient
	All		7	275 ± 31	Percent 29% 29% 29% 14% Count 2 2 2 1	43%
	Female		5	306 ± 16	Percent 20% 20% 40% 20% Count 1 1 2 1	60%
	Male		2	198 ± 98	Percent 50% 50% Count 1 1	0%

Rows per page: 10 2 Items: 1 of 1

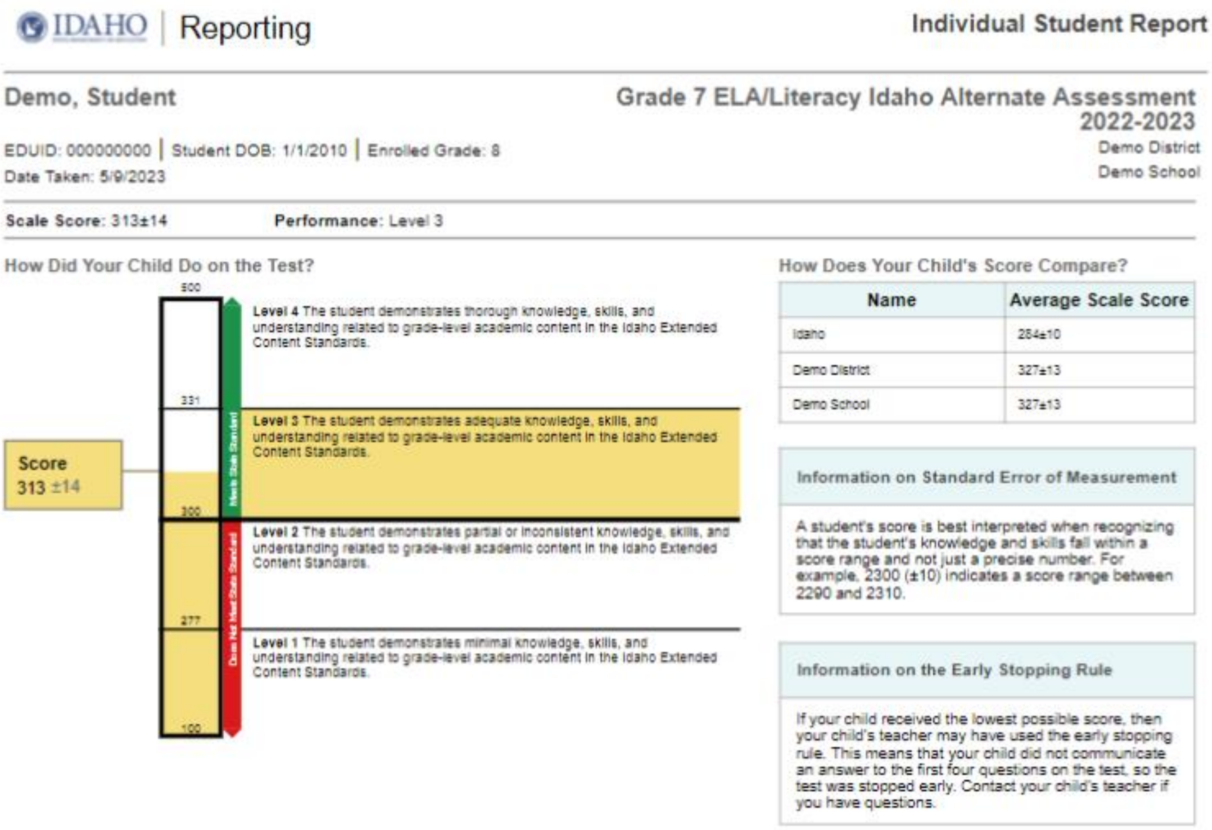
9.1.2.6 Student Detail Page

When a student completes a test, an online score report appears on the student detail page in the CRS. The student detail page provides individual student performance on the test. In each subject area, the student detail page provides scale score, performance level for the overall test, and average scale scores for the student's state, district, and school.

Specifically, the student's name, scale score, and performance level are presented at the top of the page. On the left middle section, the student's performance is described in detail using a barrel chart. Further, in the barrel chart, PLDs with cut scores at each performance level are provided, which define the content-area knowledge, skills, and processes that test takers at the performance level are expected to possess. On the top right section, average scale scores for the student's state, district, and school are displayed to compare student performance with the aggregate levels above. Below that are two text boxes with descriptive sections. The first section, "Information on Standard Error of Measurement", describes the use of standard error in interpreting the student level score range. The second section, "Information on the Early Stopping Rule", describes the Early Stopping Rule.

Exhibits 4, 5, and 6 show examples of a student detail pages for ELA, mathematics, and science.

Exhibit 4. Student Detail Page for English Language Arts



IDAHO | Reporting

Individual Student Report

Demo, Student

Grade 7 ELA/Literacy Idaho Alternate Assessment
2022-2023

EDUID: 000000000 | Student DOB: 1/1/2010 | Enrolled Grade: 8
Date Taken: 5/9/2023

Demo District
Demo School

Scale Score: 313±14 **Performance:** Level 3

How Did Your Child Do on the Test?

Level 4 The student demonstrates thorough knowledge, skills, and understanding related to grade-level academic content in the Idaho Extended Content Standards.

Level 3 The student demonstrates adequate knowledge, skills, and understanding related to grade-level academic content in the Idaho Extended Content Standards.

Level 2 The student demonstrates partial or inconsistent knowledge, skills, and understanding related to grade-level academic content in the Idaho Extended Content Standards.

Level 1 The student demonstrates minimal knowledge, skills, and understanding related to grade-level academic content in the Idaho Extended Content Standards.

How Does Your Child's Score Compare?

Name	Average Scale Score
Idaho	284±10
Demo District	327±13
Demo School	327±13

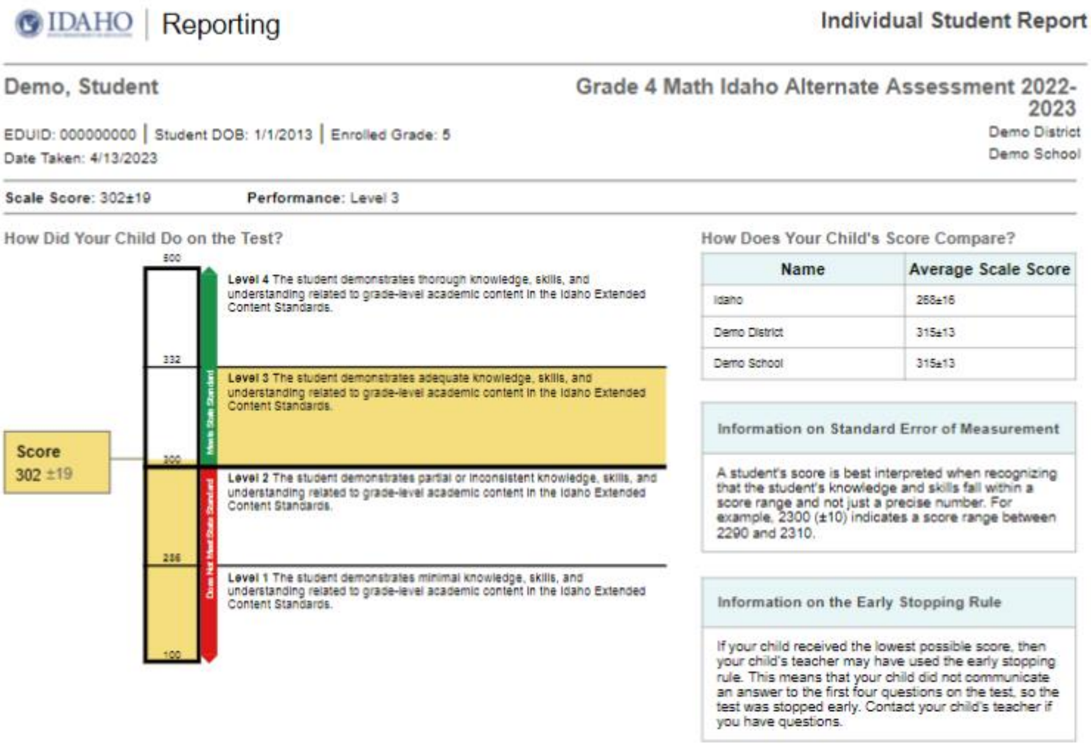
Information on Standard Error of Measurement

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.

Information on the Early Stopping Rule

If your child received the lowest possible score, then your child's teacher may have used the early stopping rule. This means that your child did not communicate an answer to the first four questions on the test, so the test was stopped early. Contact your child's teacher if you have questions.

Exhibit 5. Student Detail Page for Mathematics



IDAHO | Reporting

Individual Student Report

Demo, Student

Grade 4 Math Idaho Alternate Assessment 2022-2023

EDUID: 000000000 | Student DOB: 1/1/2013 | Enrolled Grade: 5
Date Taken: 4/13/2023

Demo District
Demo School

Scale Score: 302±19 **Performance:** Level 3

How Did Your Child Do on the Test?

Score
302 ±19

Level 4 The student demonstrates thorough knowledge, skills, and understanding related to grade-level academic content in the Idaho Extended Content Standards.

Level 3 The student demonstrates adequate knowledge, skills, and understanding related to grade-level academic content in the Idaho Extended Content Standards.

Level 2 The student demonstrates partial or inconsistent knowledge, skills, and understanding related to grade-level academic content in the Idaho Extended Content Standards.

Level 1 The student demonstrates minimal knowledge, skills, and understanding related to grade-level academic content in the Idaho Extended Content Standards.

How Does Your Child's Score Compare?

Name	Average Scale Score
Idaho	268±16
Demo District	315±13
Demo School	315±13

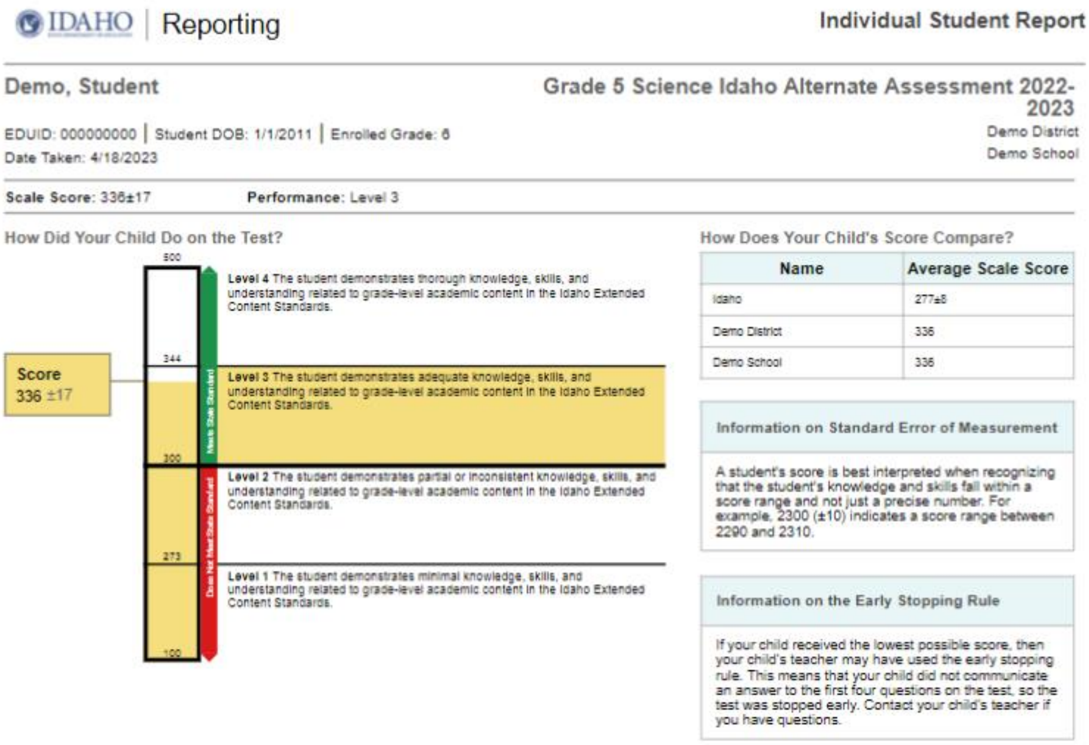
Information on Standard Error of Measurement

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.

Information on the Early Stopping Rule

If your child received the lowest possible score, then your child's teacher may have used the early stopping rule. This means that your child did not communicate an answer to the first four questions on the test, so the test was stopped early. Contact your child's teacher if you have questions.

Exhibit 6. Student Detail Page for Science



IDAHO | Reporting

Individual Student Report

Demo, Student

Grade 5 Science Idaho Alternate Assessment 2022-2023

EDUID: 000000000 | Student DOB: 1/1/2011 | Enrolled Grade: 6
Date Taken: 4/18/2023

Demo District
Demo School

Scale Score: 336±17 **Performance:** Level 3

How Did Your Child Do on the Test?

Score
336 ±17

Level 4 The student demonstrates thorough knowledge, skills, and understanding related to grade-level academic content in the Idaho Extended Content Standards.

Level 3 The student demonstrates adequate knowledge, skills, and understanding related to grade-level academic content in the Idaho Extended Content Standards.

Level 2 The student demonstrates partial or inconsistent knowledge, skills, and understanding related to grade-level academic content in the Idaho Extended Content Standards.

Level 1 The student demonstrates minimal knowledge, skills, and understanding related to grade-level academic content in the Idaho Extended Content Standards.

How Does Your Child's Score Compare?

Name	Average Scale Score
Idaho	277±8
Demo District	336
Demo School	336

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.

Information on the Early Stopping Rule

If your child received the lowest possible score, then your child's teacher may have used the early stopping rule. This means that your child did not communicate an answer to the first four questions on the test, so the test was stopped early. Contact your child's teacher if you have questions.

9.2 INTERPRETATION OF REPORTED SCORES

A student's performance on a test is reported using a scale score and a performance level for the overall test. Students' scores and performance levels are also summarized at the aggregate levels. The following section provides a description of how to interpret these scores.

9.2.1 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 by the test. The scale score is the transformed score from a theta score, estimated based on mathematical models. Low scale scores can be interpreted to mean the student does not possess sufficient knowledge and skills measured by the test. Conversely, high scale scores can be interpreted to mean the student has proficient knowledge and skills measured by the test. Interpretation of scale scores is more meaningful when the scale scores are used along with performance levels and PLDs.

9.2.2 Standard Error of Measurement

A scale score (i.e., the observed score on any test) is an estimate of the true score. If a student takes a similar test multiple times at around the same time without additional instruction, the resulting scale score will vary across repeated test attempts, sometimes being a little higher, a little lower, or the same. The standard error of measurement (SEM) represents the precision of the scale score, or the range in which the student would likely score if a similar test was administered multiple times. When interpreting scale scores, it is recommended to consider the scale scores' range, incorporating the scale scores' SEM.

The "±" next to the student's scale score provides information about the certainty, or confidence, of the score's interpretation. The boundaries of the score band are one SEM above and below the student's observed scale score, representing a range of score values that is likely to contain the true score. For example, 315 ± 13 indicates that if a student tested again, the student would likely receive a score between 302 and 328. The SEM can be different for the same scale score, depending on how closely the administered items match the student's ability.

9.2.3 Performance Level

Performance levels are proficiency categories on a test that students fall into based on their scale scores. For the IDAA, scale scores are mapped into four performance levels (i.e., Below Basic, Basic, Proficient, and Advanced) using three performance standards (i.e., cut scores). PLDs describe content-area knowledge and skills that test takers at each performance level are expected to possess. Thus, performance levels can be interpreted based on PLDs and the PLD interpretation constitutes a content standards-referenced understanding of the student performance.

9.2.4 Aggregated Score

Students' scale scores are aggregated at the roster, teacher, school, district, and state levels to represent how a group of students performed on a test. When students' scale scores are aggregated, the aggregated scale scores can be interpreted as an estimate of the knowledge and skills that a group of students possesses. In addition to the aggregated scale scores, the percentage of students in each performance level for the overall subject are reported at the aggregate level to represent how well a group of students performed overall.

9.3 Appropriate Uses for Scores and Reports

Assessment results can be used to provide information on individual students' performance on the test. Overall, assessment results show what students know and can do in certain subject areas.

Assessment results on student performance on the test can help teachers or schools make decisions on how to support students' learning. Aggregate score reports at the teacher and school levels provide information regarding students' strengths and weaknesses that can be used to improve teaching and student learning. By narrowing down the student performance results by subgroup, teachers and schools can determine what strategies may need to be implemented to improve teaching and student learning, particularly for disadvantaged subgroups. For example, teachers can view student assessment results by the Migrant Student subgroup and observe that students in the subgroup category "Migrant" struggle with ELA. Teachers can then provide additional instruction for students in the Migrant Student subgroup to enhance their performance in ELA.

Assessment results can also be used to compare students' performance among different students and among different groups. Teachers can evaluate how their students perform compared with other students in schools and districts overall.

While assessment results provide valuable information to understand students' performance, these scores and reports should be used with caution. Most importantly, assessment results should always be used in combination with information about student from other sources and should not be used in isolation. It is important to note that reported scale scores are estimates of true scores and hence do not represent the precise measure for student performance. Moreover, although student scores may help educators make important decisions about students' placement and retention, or teachers' instructional planning and implementation, the assessment results should not be the only source of information. Given that assessment results measured by a test provide limited information, other sources on student performance, such as classroom assessment and teacher evaluation, should be considered when making decisions on student learning.

10. QUALITY CONTROL PROCEDURES

Quality assurance (QA) procedures are enforced throughout all stages of the alternate assessment development, administration, and scoring and reporting of results. Cambium Assessment, Inc. (CAI) uses 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 opens.

10.1 OPERATIONAL TEST CONFIGURATION

For the operational test, a *test configuration file* is the key file that contains all specifications for the item selection algorithm and the scoring algorithm. The test configuration file includes the test blueprint specification; slopes and intercepts for theta-to-scale score transformation, cut scores; and the item information (e.g., answer keys, item attributes, item parameters, passage information). The accuracy of the information in the configuration file is independently checked and confirmed numerous times by multiple staff members before the testing window opens.

To verify the accuracy of the scoring engine, a simulated test administrations is used. The simulator generates a sample of students with an ability distribution that matches that of the population. The ability of each simulated student is used to generate a sequence of item response scores consistent with the underlying ability distribution.

Simulations are generated using the production, item selection, and scoring engine to ensure that verification of the scoring engine is based on a wide range of student response patterns. The results of simulated test administrations are used to configure and evaluate the adequacy of the item selection algorithm used to administer the Idaho Alternate Assessment (IDAA). The purpose of the simulations is to configure the algorithm to optimize item selection to meet blueprint specifications and check score accuracy. The scores in the simulated data file are checked independently, following the scoring rules specified in the scoring specifications.

10.1.1 Platform Review

CAI’s TDS supports a variety of item layouts. Each item goes through an extensive platform review on different operating systems, such as Windows, Linux, and iOS, to ensure that the item looks consistent across the board. When the stimulus and item response options and response areas are displayed side by side, each side has an independent scroll bar.

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

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

10.1.2 User Acceptance Testing and Final Review

Prior to deployment, the testing system and content are deployed to a staging server to undergo user acceptance testing (UAT). UAT of the TDS serves as both a software evaluation and a content approval role. The UAT period allows the Department to interact with the exact test that the students will use.

10.2 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 our QA system. The QA system conducts a series of data integrity checks, ensuring, for example, that the record for each test contains information for each item, keys for multiple-choice items, score points for each item, and the total number of field-test items and operational items. It also ensures that the test record contains no data from items that have been invalidated.

Data pass directly from the Quality Monitoring System (QMS) to the Database of Record (DOR), which serves as the repository for all test information from which all test information for reporting is pulled. The Data Extract Generator (DEG) pulls data from the DOR for delivery to the Idaho State Department of Education (SDE). CAI staff ensures that data in the extract files match the DOR before delivering the findings to the SDE.

10.3 QUALITY ASSURANCE IN TEST SCORING

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

A series of QA reports, such as blueprint match rate, item exposure rate, and item statistics, can also be generated at any time during the online testing window for early detection of any unexpected issues. Any deviations from the expected outcome are flagged, investigated, and resolved.

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 frequently evaluated at the opening of the testing window to ensure that test administrations conform to the blueprint and that items are performing as anticipated.

The item statistics analysis report is used to monitor the performance of test items throughout the testing window. It serves as a key check for the early detection of potential problems with item scoring, including the incorrect designation of a keyed response or other scoring errors and potential breaches of test security that may be indicated by changes in the difficulty of test items. This report generates classical item analysis indicators of difficulty and discrimination, including proportion correct and biserial/polyserial correlation.

The report is configurable and can be produced so that only items with statistics falling outside of a specified range are flagged for reporting or to generate reports based on all items in the pool.

Table 44 presents an overview of the QA reports.

Table 44. Overview of Quality Assurance Reports

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

10.3.1 Score Report Quality Check

Two types of score reports were produced for the IDAA: (1) online score reports and (2) family score reports.

10.3.1.1 Online Report Quality Assurance

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

Every test undergoes a series of validation checks. Once the QA system signs off, data are passed to the DOR, which serves as the centralized location for all student scores and responses, ensuring that there is only one place where the "official" record is stored. Only after scores have passed the QA checks and are uploaded to the DOR are they passed to the Centralized Reporting System (CRS), which is responsible for presenting individual-level results and calculating and presenting aggregate results. Absolutely no score is reported in the CRS until it passes all the QA system's validation checks. The aforementioned processes take milliseconds to complete. The composite score becomes available in the CRS within less than a second of handscores being received and passing QA validation checks.

10.3.1.2 Paper Report Quality Assurance

Statistical Programming

The family score reports contain custom programming and require rigorous QA processes to ensure their accuracy. All custom programming is guided by detailed and precise specifications in our 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. Two senior statisticians and one senior programmer review the final programs to ensure that the teams have implemented the agreed-on procedures. Custom programming is implemented independently by two statistical programming teams working from the specifications. Only when the output from both teams matches exactly are the scripts released for production.

Because much of the statistical processing is repeated, CAI has implemented a structured software-development process to ensure that the repeated tasks are implemented correctly and identically each time. We write small programs (called macros) that take specified data as input and produce data sets containing derived variables as output. Approximately 30 such macros reside in our library for the family score reports. Each macro is extensively tested and stored in a central development server. Once a macro is tested and stored, the director of score reporting, the director of psychometrics, and the project directors for affected projects must approve changes to the macro.

Each change is followed by a complete retesting of the entire collection of scenarios on which the macro was originally tested. The main statistical program is mostly made up of calls to various macros, including macros that verify the data, conversion tables, and macros that complete 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 CAI designers create backgrounds, our VIPP programmers write code that indicates where to place all variable information (i.e., data, graphics, 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 process 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; the output is formatted as VIPP input. This process allows us to test the entire system. Programmed output goes through multiple stages of review and revision by graphics editors and the CAI Score Reporting team to ensure that design elements are accurately reproduced and that data are correctly displayed. Once we receive final data and VIPP programs, the CAI Score Reporting team reviews proofs that contain actual data based on our standard QA documentation.

Additionally, data that are independently calculated by CAI psychometricians are compared with data on the reports. Several CAI staff members review a large sample of reports to ensure that all data are correctly placed on reports. This rigorous review is typically 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. Prior to printing the reports, CAI provides a live data file and individual student reports (ISRs) with sample districts for SDE staff review. CAI works closely with the SDE to resolve questions and correct any problems. The reports are not delivered unless the SDE approves the sample reports and data file.

REFERENCES

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

Camilli, G., & Shepard, L. A. (1994). Methods for identifying biased test items. Thousand Oaks [Calif.: Sage Publications.

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.

Linacre, J. M. (2004). Rasch model estimation: Further topics. *Journal of Applied Measurement, 5*(1), 95–110.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement,* 32: 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: 247–260.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement, 52*(2), 443–451.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The Bookmark Procedure: Psychological perspectives. In G. Cizek (Ed.), *Setting performance standards*. Mahwah, NJ: Lawrence Erlbaum.

Muniz, J., Hambleton, R., & Xing, D. (2001). Small sample studies to detect flaws in item translations. *International Journal of Testing, 1*(2), 115–135.

Sireci, S. G., & Rios, J. (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation, 19*(2–3), 170–187.

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

Towles-Reeves, E., Kearns, J., Flowers, C., Hart, L., Kerbel, A., Kleinert, H., Quenemoen, R., & Thurlow, M. (2012). Learner characteristics inventory project report (A product of the NCSC validity evaluation). Minneapolis, MN: University of Minnesota, National Center and State Collaborative.

U.S. Department of Education (2018). *A state's guide to the U.S. Department of Education's assessment peer review process*. Retrieved from https://www2.ed.gov/admins/lead/account/saa/assessmentpeerreview.pdf