

# Alternate Montana Science Assessment (AMSA)

## 2022–2023 Technical Report

Science Grades 5, 8, and 11



Submitted to  
Montana Office of Public Instruction  
by Cambium Assessment, Inc.

---

---

## TABLE OF CONTENTS

<b>1. OVERVIEW</b> .....	<b>6</b>
1.1 The Alternate Montana Science Assessment .....	6
1.2 Alternate Assessment Eligibility .....	7
1.3 Content Standards .....	7
1.4 Memorandum of Understanding on Item-Sharing Initiative .....	8
<b>2. TEST ADMINISTRATION</b> .....	<b>10</b>
2.1 Test Administrator Training .....	10
2.1.1 Online Training .....	10
2.2 Administration Manuals .....	11
2.3 Accommodations .....	12
2.3.1 Allowable Accommodations .....	12
2.3.2 Assistive Technology .....	14
2.4 Online Administration .....	15
2.5 Alternate Response Option Card Test Administration .....	15
2.6 Test Security .....	15
2.6.1 Student-Level Testing Confidentiality .....	15
2.6.2 System Security .....	16
2.7 Prevention and Recovery of Disruptions in the Test Delivery System .....	17
2.7.1 High-Level System Architecture .....	17
2.7.2 Automated Backup and Recovery .....	19
2.7.3 Other Disruption Prevention and Recovery .....	19
<b>3. SUMMARY OF SPRING 2023 OPERATIONAL TEST ADMINISTRATION</b> .....	<b>20</b>
3.1 Student Participation .....	20
3.2 Summary of Student Performance .....	22
3.3 Test-Taking Time .....	23
3.4 Distribution of Student Ability and Item Difficulty .....	24
<b>4. ITEM DEVELOPMENT</b> .....	<b>25</b>
4.1 Item Development for the MOU-Alt .....	25
4.1.1 Item Type and Scoring Rubrics .....	26
4.1.2 Item Development Procedure and Item Reviews .....	27
4.1.3 Development of Crosswalk and State Alternate Content Performance Standards .....	29
4.2 Field Testing .....	30
4.2.1 Item Statistics .....	30
4.2.2 Classical Statistics .....	31
4.2.3 Item Response Theory Statistics .....	32
4.2.4 Analysis of Differential Item Functioning .....	32
4.2.5 Summary of Item Statistics .....	33
4.2.6 Data Review Meeting .....	35
4.3 Scaling and Equating .....	35

---

4.3.1	Item Calibration .....	36
<b>5.</b>	<b>VALIDITY.....</b>	<b>37</b>
5.1	Intended Uses and Interpretations of AMSA Scores .....	37
5.2	Sources of Validity Evidence.....	37
5.2.1	Evidence Based on Test Content .....	38
5.2.2	Evidence Based on Response Process.....	40
5.2.3	Evidence Based on Internal Structure.....	41
5.2.4	Evidence Based on Relations to Other Variables.....	42
<b>6.</b>	<b>RELIABILITY .....</b>	<b>46</b>
6.1	Marginal Reliability .....	46
6.2	Standard Error Curves.....	47
6.3	Reliability of Performance Classification .....	48
6.4	Reliability for Content Strand Scores .....	51
<b>7.</b>	<b>SCORING.....</b>	<b>52</b>
7.1	Attemptedness Rules for Scoring.....	52
7.2	Estimating Student Ability Using Maximum Likelihood Estimation .....	52
7.3	Scoring All Correct and All Incorrect Cases.....	53
7.4	Rules for Transforming Theta to Scale Scores.....	53
7.5	Lowest/Highest Obtainable Scale Score (LOSS/HOSS) .....	54
<b>8.</b>	<b>PERFORMANCE STANDARDS.....</b>	<b>55</b>
8.1	Standard-Setting Procedures .....	55
8.2	Performance-Level Descriptors .....	55
8.3	Recommended Performance Standards.....	56
<b>9.</b>	<b>REPORTING AND INTERPRETING SCORES.....</b>	<b>57</b>
9.1	Centralized Reporting System for Students and Educators.....	57
9.1.1	Types of Online Score Reports .....	57
9.2	Interpretation of Reported Scores .....	61
9.2.1	Scale Score .....	61
9.2.2	Standard Error of Measurement .....	61
9.2.3	Performance Level .....	61
9.2.4	Aggregated Score .....	61
9.3	Appropriate Uses for Scores and Reports .....	62
<b>10.</b>	<b>QUALITY CONTROL PROCEDURES .....</b>	<b>63</b>
10.1	Operational Test Configuration.....	63
10.1.1	Platform Review .....	63
10.1.2	User Acceptance Testing and Final Review.....	64
10.2	Quality Assurance in Data Preparation.....	64
10.3	Quality Assurance in Test Scoring.....	64

*10.3.1 Score Report Quality Check*..... 65

**REFERENCES**.....**66**

## LIST OF TABLES

Table 1. Participation Criteria.....	7
Table 2. List of Available Accessibility Tools.....	13
Table 3. Total Number of Students with Allowed Accessibility Tools for AMSA.....	14
Table 4. AMSA Number of Participated Students.....	20
Table 5. AMSA Number of Attempted Students by Subgroup .....	20
Table 6. AMSA Number of Participated Students by Subgroup and Disability Category .....	21
Table 7. Student Performance Overall and by Subgroup, Grade 5 .....	22
Table 8. Student Performance Overall and by Subgroup, Grade 8.....	22
Table 9. Student Performance Overall and by Subgroup, Grade 11 .....	23
Table 10. Test-Taking Time .....	23
Table 11. Summary of the 2023 Field-Test Item Pool Across MOU-Alt States .....	30
Table 12. Thresholds for Flagging in Classical Item Analysis .....	32
Table 13. DIF Classification Rules Science.....	33
Table 14. 2023 MOU Item Sample Size Distribution.....	34
Table 15. Summary of Item Analyses Results for MOU-Alt Science .....	34
Table 16. Number of Items in Each DIF Classification Category .....	34
Table 17. Summary of the Item Data Review for MOU-Alt Shared Items .....	35
Table 18. Summary of AMSA Field-Test Items.....	35
Table 19. Percentage of Administered Tests Meeting Blueprint Requirements .....	39
Table 20. AMSA Correlations Among Strands .....	42
Table 21. Correlations Between LCI Descriptors and Total Scores in AMSA.....	45
Table 22. Marginal Reliability .....	47
Table 23. Average Conditional Standard Error of Measurement by Performance Level .....	48
Table 24. Classification Accuracy and Consistency for Performance Standards .....	50
Table 25. AMSA Marginal Reliability Coefficients for Content Strand Scores.....	51
Table 26. Scaling Constants on the Reporting Metric .....	53
Table 27. Range of Scale Scores by Performance level.....	54
Table 28. Recommended Performance Standards for AMSA .....	56
Table 29. Types of Online Score Reports by Level of Aggregation.....	58
Table 30. Types of Subgroups .....	58
Table 31. Overview of Quality Assurance Reports .....	65

**LIST OF FIGURES**

Figure 1. Distribution of Science Testing Time..... 24  
Figure 2. Student Ability–Item Difficulty Distribution for AMSA ..... 24  
Figure 3. Alternate Assessment Item Development Process ..... 26  
Figure 4. Conditional Standard Error of Measurement for Science..... 47

**LIST OF EXHIBITS**

Exhibit 1. Dashboard: State Level ..... 59  
Exhibit 2. Dashboard: District Level ..... 59  
Exhibit 3. Student Detail Page for Science ..... 60

## 1. OVERVIEW

This report provides a technical summary of the 2022–2023 Alternate Montana Science Assessment (AMSA) administered in grades 5, 8, and 11. The purpose of this technical report is to document the evidence supporting the claims made for how AMSA test scores can be interpreted. The report includes 10 chapters that discuss all the evidence accrued about the technical quality of the AMSA testing system. Analyses included in this report are based on Montana alternate assessment data and address all aspects of the technical requirements described in the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014) and in *A State’s Guide to the U.S. Department of Education’s Assessment Peer Review Process* (U.S. Department of Education, 2018).

Chapter 1 gives an overview of the AMSA. Chapter 2 documents the test administration procedures, including proctor training, the test administration manual, accommodations, and the prevention of disruptions in the Test Delivery System (TDS). Chapter 3 summarizes the results of the spring 2023 AMSA test administration. These sections provide summaries of the test-taking student population, their performance on the assessments, and the time spent taking the assessments. Chapter 4 describes the item-development process; specifically, the sequence of reviews that each item must pass through before being eligible for the AMSA test administration. This chapter also summarizes the field-test item analyses, data review, and procedures used to scale and calibrate the AMSA items for scoring and reporting. Chapter 5 provides validity evidence on the test contents, response processes, internal structure, and relations to other variables.

Chapter 6 provides evidence for the reliability of the AMSA, including marginal reliability, standard errors of measurement, and classification accuracy and consistency of performance standards. Chapter 7 describes the scoring procedures used in producing scale scores and performance levels. Chapter 8 describes the procedure to set performance standards based on data from the spring 2023 administration. Chapter 9 provides a description of the score reporting system and the interpretation of test scores. Chapter 10 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 ALTERNATE MONTANA SCIENCE ASSESSMENT

The AMSA is comprised of assessments that are based on the Montana Content Standards in science and is designed for students with the most significant cognitive disabilities. The purposes of the AMSA are: (1) to maximize access to the general education curriculum—the knowledge, skills, and abilities across the academic content standards for students with the most significant cognitive disabilities; (2) to ensure that all students with disabilities are included in Montana’s statewide assessments; and (3) to ensure that these students are included in the educational accountability system. Assessment results can inform instruction in the classroom by providing data that guide decision making. The AMSA is only for students with documented significant cognitive disabilities and adaptive behavior deficits who require extensive support across multiple settings (e.g., home, school, community). Typically, this student segment consists of about 1% of the total student population.

In 2020–2021, the Montana Office of Public Instruction (OPI) began the transition to a new, online computer-adaptive test (CAT) for the science alternate assessment for students with significant cognitive disabilities. The new assessment is designed to assess students in grades 5, 8, and 11. In the spring 2023

administration, each student was administered a 40-item operational test with 10 embedded field-test (EFT) items.

## 1.2 ALTERNATE ASSESSMENT ELIGIBILITY

Most students with disabilities can participate in the general state assessments when provided with the appropriate accommodations. However, for students with the most significant cognitive disabilities, it may be more appropriate to participate in the alternate assessment. Decisions concerning a student’s participation in statewide assessments are made by each student’s individualized education program (IEP) team. Guidance for IEP teams to inform decisions about which assessment is most appropriate for each student is provided in the OPI Alternate Assessment Eligibility Guidelines <https://opi.mt.gov/Portals/182/Page%20Files/Statewide%20Testing/Participation/Alternate%20Assessment%20Eligibility%20Guidelines.pdf>. The participation guidelines for Montana’s students to take the AMSA are summarized in Table 1.

Table 1. Participation Criteria

Participation Criteria	Participation Criteria Descriptors
1. The student has a significant cognitive disability.	Review of student records indicates a disability or multiple disabilities that significantly impact intellectual functioning and adaptive behavior.*  <i>*Adaptive behavior is defined as essential for someone to live independently and to function safely in daily life.</i>
2. The student is learning content that is linked to grade-level standards.	Goals and instructions listed in the student’s IEP are linked to the enrolled grade-level content standards and address knowledge and skills that are appropriate and challenging for this student.
3. The student requires extensive, direct, individualized instruction and substantial supports to achieve measurable gains in a grade- and age-appropriate curriculum.	The student (a) requires extensive, repeated, and individualized instruction and support that is not of a temporary or transient nature; and (b) uses 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 CONTENT STANDARDS

The publication of *A State’s Guide to the U.S. Department of Education’s Assessment Peer Review Process* (U.S. Department of Education, 2018) clearly indicates that content standards must specify what students are expected to know and be able to do. Standards should include coherent and rigorous content and encourage the use of advanced teaching pedagogy and research-based instructional practices.

The AMSA is aligned to the state content standards for science, which are linked to Essence Statements. The Essence Statements describe the core ideas within an achievement performance expectation (PE), distilled down to a level appropriate for the students participating in the alternate assessment. These Essence Statements serve as the foundation for the development of AMSA items.



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

Essence Statements for the AMSA are incorporated into the AMSA Performance-Level Descriptors (PLDs) for science. PLDs have been developed at four levels of complexity for each Essence Statement and were subject to approval during the spring 2022 standard-setting meeting. The levels are as follows:

- Level 4:** A student who is *Level 4* demonstrates a level of understanding that includes the ability to “bring together” the Disciplinary Core Ideas (DCI) and/or Science and Engineering Practices (SEP) and/or Cross-Cutting Concepts (CCC) associated with a PE.
- Level 3:** A student who is *Level 3* demonstrates an understanding of the DCI and/or SEP and/or CCC within PE at the level described in the Essence Statement.
- Level 2:** A student who is *Level 2* demonstrates some understanding of the content of the PE, but that understanding is incomplete and does not yet meet the expectations found in the Essence Statement. This student’s understanding is partial but emerging.
- Level 1:** A student who is *Level 1* demonstrates a level of understanding that is at a very preliminary level. This student’s understanding is nonexistent or incomplete, and he or she has difficulty meeting the expectations of a student who approaches expectations.

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 and Reporting PLDs.

Students participating in the AMSA also have communication skills ranging from symbolic or abstract, to concrete, to pre-symbolic. Accommodations may be provided to allow students to perceive and respond to test items in meaningful ways.

#### **1.4 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 for English language arts (ELA), mathematics, and science. Each state contributed a predetermined number of items proportional to their state’s student population for alternate assessments. In early 2019, Idaho and Vermont joined the collaborative item development and field-testing effort and participated in the spring 2019 field test for ELA, mathematics, and science. In spring 2020, Montana and South Dakota joined the MOU for the science assessment. In 2022, Vermont exited the MOU. Because the total number of students in alternate assessments is very small in each state, field testing common items in all MOU states allowed for the calibration of items based on the combined

data across all states. In addition to the MOU shared item pool, each state also developed some items that aligned to the state’s specific content standards or content specifications.

The item-sharing initiative is designed to implement an item development process that generates at least three times the number of items needed for each test administration for each grade and subject. With 40 operational items on the test, at least 120 calibrated items in the pool are needed for a CAT. The item-sharing initiative allowed for this item development effort. Each MOU member would own the items they developed, but their items would be available for use by the other MOU members. The number of items developed by each state would be proportional to the size of the alternate assessment population that would participate in the test.

## 2. TEST ADMINISTRATION

The spring 2023 testing window was open from March 13 to April 28, 2023, for the online adaptive operational tests and online fixed-form operational test. For all grades, the online adaptive operational tests were the default method of administration. In each grade, the online fixed-form test paired paper-pencil response cards and test visuals with the digital presentation of the stimuli and items. The online fixed-form test was provided as a special paper-pencil test form accommodation for students who were unable to fully access the online tests, even with the available accommodations. In the online adaptive test, the student took the assessment independently or with the test administrator’s (TA) assistance, as needed. In the fixed-form test, one TA administered the assessment to one student at a time. The online adaptive tests consisted of 40 operational items selected adaptively that match the student ability and meet the assessment blueprint, and 10 embedded field-test items selected from the Memorandum of Understanding (MOU) field-test item pool. The online fixed-form tests with paper-pencil accommodation comprised only 40 operational items.

### 2.1 TEST ADMINISTRATOR TRAINING

TA training 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 the standardization of test administration and test scoring rules. If TAs do not follow the same procedures, student performance cannot be meaningfully compared.

Authorized Representatives (ARs), System Test Coordinators (STCs), and Building Coordinator (BCs) oversee all aspects of testing at their schools and serve as the main points of contact, while TAs administer the online assessments. The online TA Certification Course, PowerPoint presentations, user guides, manuals, and regional trainings are used to train ARs and BCs in the online testing requirements and the mechanics of starting, pausing, and ending a test session. Training materials for the administration are available online at <https://mt.portal.cambiumast.com/resources>. STCs are responsible for training TAs.

#### 2.1.1 Online Training

Multiple online training opportunities are offered to key staff.

##### TA Certification Course

All school personnel who serve as TAs complete an online TA Certification Course before administering the secure and valid assessments. This web-based course is 30–45 minutes long and covers information on testing policies and the steps for administering a test session in the online system. The course is interactive, requiring participants to start test sessions under different scenarios. Throughout the training and at the end of the course, participants answer multiple-choice questions about the information provided.

##### System Tutorials

The following presentations are offered to explain how the assessment system works (each of these presentations lasts approximately 30 minutes; slides are available on the portal at <https://mt.portal.cambiumast.com/resources>):

*TA Interface for Online Testing.* This tutorial prepares ARs, BCs, and TAs for the assessments by providing an overview of the TA Interface and the Test Delivery System (TDS), including how to start and monitor a test session using the TA Interface.

*Student Interface For Online Testing.* This tutorial provides an overview of the online Student Interface in the TDS.

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

## **Manuals and User Guides**

The following user guides provide information on systems and preparation for testing:

*Alternate Accessibility Guidelines.* This manual provides guidance on appropriate supports and accommodations for Montana school district personnel who must make decisions about testing special student populations.

*Assistive Technology Manual.* This manual includes information about supported operating systems and required hardware and software for using assistive technology with the secure browser. It provides configuration requirements and recommendations for frequently used assistive technologies.

*Test Administration User Guide.* This user guide provides information about the TDS, including the TA Interface and Student Interface.

*TIDE User Guide.* This user guide provides information on how to use and navigate the TIDE system, including managing user and student account information and setting student test settings and accommodations.

## **Practice and Training Test Site**

In September 2021, separate training sites were opened for TAs and students. TAs can practice administering assessments and starting and ending test sessions on the TA training site, and students can practice taking online assessments on the student practice and training site. The Montana assessment provides a sample set of items corresponding to the summative assessments for the AMSA.

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

The practice test is available on the Montana portal at <https://mt.portal.cambiumast.com>.

## **2.2 ADMINISTRATION MANUALS**

The *2022–2023 Alternate Montana Science Assessment Test Administration Manual (TAM)* summarizes the AMSA and provides brief guidelines for test administration. It includes the following:

- Overview of the background, purpose, and content specifications for AMSA
- Assessment design
- Student inclusion and participation guidelines
- TA requirements

- Test delivery modes: online or online with fixed-form paper-pencil response cards and test visuals as a special accommodation
- Test administration procedures
- Test security guidelines

The 2022–2023 AMSA TAM can be found at [https://mt.portal.cambiumast.com/-/media/project/client-portals/montana/pdf/2022-2023\\_amsa\\_tam.pdf](https://mt.portal.cambiumast.com/-/media/project/client-portals/montana/pdf/2022-2023_amsa_tam.pdf)

Included in the 2022–2023 TAM are recommendations for the use of alternate response option cards for the paper-pencil test accommodation. This accommodation was provided for approved students and accompanied the administration of the fixed form.

There is no time limit besides the dates of the testing window during administration of the AMSA. If the student becomes tired, the TA can pause the assessment and restart it at the same point later.

## **2.3 ACCOMMODATIONS**

The AMSA was designed following universal design principles that incorporate supports that a student might need to access the assessment (e.g., picture arrays, oral reading of passages, the use of a student’s own receptive and expressive communication methods). The allowable accommodations listed in this section provide students the ability to access the items and provide a response.

### **2.3.1 Allowable Accommodations**

For the online assessment version, all items may be read and re-read using the read-aloud function in the online testing system. For the paper-pencil version, all items may be orally presented after the teacher uses the online digital interface to present the test item the first time. Testing for either test form is not timed, may be completed over multiple sessions, and can stop at any point within the test form, as needed.

A variety of universal tools are available for the AMSA. A list of available universal tools is provided in Table 2–Table 3. This list is by no means exhaustive, as students with significant cognitive disabilities vary widely in the type and amount of supports that may be required for access. The list of universal tools in Table 2 contains only examples of some of the supports that a student who takes the AMSA may need in order to demonstrate understanding. A general rule of thumb is to provide the same level of supports during the alternate assessment as are regularly provided during instruction.

Table 2. List of Available Accessibility Tools


<b>Topic</b>	<b>Description</b>
<b>Embedded Accessibility Features</b>	<ul style="list-style-type: none"> <li>▪ Color Choices</li> <li>▪ Highlighter</li> <li>▪ Human Voice Recording (HVR) (HVR can be used by selecting the  icon.)</li> <li>▪ Line Reader Tool</li> <li>▪ Mark for Review</li> <li>▪ Masking</li> <li>▪ Mouse Pointer</li> <li>▪ Notepad</li> <li>▪ Permissive Mode</li> <li>▪ Print-on-Request</li> <li>▪ Strikethrough</li> <li>▪ Streamlined Mode</li> <li>▪ Volume Control</li> <li>▪ Zoom</li> </ul>
<b>Non-Embedded Accessibility Features</b>	<ul style="list-style-type: none"> <li>▪ Amplification</li> <li>▪ Alternate Response Options</li> <li>▪ Bilingual Dictionary</li> <li>▪ Breaks</li> <li>▪ Color Contrast</li> <li>▪ Color Overlay</li> <li>▪ Magnification</li> <li>▪ Medical Supports</li> <li>▪ Noise Buffers</li> <li>▪ Physical Manipulatives</li> <li>▪ Read Aloud</li> <li>▪ Scribe</li> <li>▪ Separate Setting</li> <li>▪ Sign Language for Test Items</li> <li>▪ Specialized Calculator</li> <li>▪ Speech-to-Text</li> <li>▪ Simplified Test Directions</li> <li>▪ Translated Test Directions</li> <li>▪ Timing or Scheduling</li> <li>▪ Word Prediction</li> </ul>

Table 3 presents the number of students who were allowed accessibility tools in the science assessment.

Table 3. Total Number of Students with Allowed Accessibility Tools for AMSA

Accommodations	Grade		
	5	8	11
Alternate Response Options	6	3	
Color Choices		1	1
Line Reader			
Masking	8	5	
Mouse Pointer			1
Permissive Mode	1		
Print Size		1	1
Print on Request	3		
Sign Language for Test Items	3	1	
Streamlined Mode		4	
Strikethrough			
Specialized Calculator	3	3	
Speech-to-Text	9	3	
Timing or Scheduling	8	5	
Word Prediction	2	7	

### 2.3.2 Assistive Technology

Assistive technology (AT) that is documented in the student's Individualized Education Program (IEP) and used during regular instruction may be used to assist the student in accessing the AMSA via the TDS. Technology affords many ways to adapt student responses on the device. Any assistive technology that does not unfairly provide advantage or disadvantage to a student may be used, including, but not limited to, the following:

- Screen magnifier or screen magnification software
- Arm support
- Mouth stick, head pointer with standard or alternative keyboard
- Voice output device, both single and multiple message
- Tactile/voice output measuring devices (e.g., clock, ruler)
- Overhead projector

Eligible students take the AMSA and can access the assessment using the digital interface when provided the allowable supports. However, it is recognized that students with certain disabilities will still require access using the paper-pencil version of the assessment.

Some students with disabilities may be better able to access the assessment with the alternate response option card version of the AMSA. If a student's IEP care coordinator determines that the student requires the paper-pencil version of the AMSA due to the nature of his or her disability or disabilities, the student's

proctor will need to contact the STC, who will notify the OPI. The STC is responsible for printing the alternate response option cards or providing a PDF file to be printed by the TA.

## **2.4 ONLINE ADMINISTRATION**

During test administration, the student or TA touches the button bearing an ear icon for the stimulus, question, and response option portion of each item to be read aloud. The read-aloud script is a recorded human voice. The speed of narration is comparable to the average speed of narration when teachers read to students. Students respond to each item by clicking one of the response options presented, or the TA can click the student’s selected response option on their behalf. The online system automatically stores item responses when students touch their selected-response options.

For all test items, if no response is indicated or recorded by the student, the TA will access the context menu for the item and select the “No Response” option for that item. This marks the item as “No Response,” and the TA can advance to the next test item for administration.

In spring 2023, an Early Stopping Rule was available for students who were non-responsive to the first four items on each content-area test. Students and TAs were required to follow the administration guidelines put in place by the OPI. The Early Stopping Rule was instituted for a student’s test if the student did not respond to the first four items in the assessment and they were administered the student response check with no visible response.

## **2.5 ALTERNATE RESPONSE OPTION CARD TEST ADMINISTRATION**

In spring 2023, students who required an alternate response option card accommodation were administered a fixed-form test via the online testing system alongside printed response option cards which the TA placed in front of the student while listening to the test item read-aloud script via the online testing system. TAs completed and submitted the Learner Characteristics Inventory (LCI) as part of their Multistate Alternate Assessment (MSAA) outside of Cambium Assessment, Inc.’s (CAI) systems, which investigates the learning characteristics of students participating in alternate assessments based on alternate performance standards for each student. During test administration, the student’s item responses were entered into the online testing system directly by the TA after the student indicated their response option via the printed paper-pencil response option cards. No access-limited items were included on the fixed-form tests for paper-pencil administration. The number of students who received the fixed-form test in spring 2023 can be found in Table 4.

## **2.6 TEST SECURITY**

The Test Security Guidelines, included in the *2022–2023 Test Administration Manual*, indicate that photocopying any printed testing materials is strictly prohibited. Printed alternate response option cards and printed test visuals are secure materials. BCs are responsible for receiving, accounting for, and returning all test materials to CAI. If CAI does not receive the returned test materials within the scheduled time frame, CAI makes significant efforts to ensure that all secure materials are returned. Any known violations of test security are to be reported immediately.

### **2.6.1 Student-Level Testing Confidentiality**

The online adaptive and fixed-forms tests are administered through secure websites. All the secure websites 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 the basic features of the current system and ensure authorized data access. All aspects of the system, including item development and review, test delivery, and reporting, are secured by password-protected logins. The systems use role-based security models to ensure that users may access only the data to which they are entitled and may edit data only in accordance with their user rights.

FERPA prohibits the public disclosure of student information or test results. To comply with the secure standards, student names and IDs are communicated via a Secure File Transfer Protocol (SFTP). Student login information is associated with the particular tests to which they are assigned.

Student login information is entered only at the beginning of a test after an authorized TA creates and manages the test session and after the TA reviews and approves a test (and its settings) for the student. Accommodation settings can only be changed by users with the AR, STC, or BC TIDE user role(s). Designated support settings can be set by users with the TA user role. Test materials and reports are carefully protected so that student names and test results cannot be identified and accessed by unauthorized individuals.

All students must be enrolled or registered at their testing schools in order to take the online tests. Student enrollment information, including demographic data, is generated by the OPI and uploaded nightly to the online testing system via an SFTP site during the testing period.

Only staff with the administrative roles of AR, STC, or BC can view students' scores. STCs and ARs have access to all scores within their district. BCs have access to all scores within their school. Teachers have access to all scores within their classrooms. Parents receive only a printed copy of their children's online score reports if the school or teacher provides one.

## **2.6.2 System Security**

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

**Password Protection.** This security measure ensures that all access points by different roles—at the state, district, school principal, and school staff levels—require a password to log in to the system. Newly added BCs and TAs receive separate passwords (assigned by the school or district) through their personal email addresses.

**CAI Secure Browser.** With this security measure, the technology coordinator must ensure that the CAI Secure Browser is properly installed on the computers used for the administration of the online assessments. Developed by the testing contractor, the CAI Secure Browser prevents students from accessing other computers or Internet applications and from copying test information. The Secure Browser suppresses access to commonly used browsers such as Chrome and Firefox and prevents students from searching for answers on the Internet or communicating with other students. Assessments can be accessed only through the CAI Secure Browser and not by other Internet browsers.

Testing personnel are reminded in the online training and user manuals that assessments should be administered in an appropriate testing environment.

## **2.7 PREVENTION AND RECOVERY OF DISRUPTIONS IN THE 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. Our 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; in addition to general warnings of malfunction, our monitoring system also provides warnings when any given server is performing differently from its performance over the few hours prior, or differently than the other servers performing the same jobs. Subtle changes in performance often precede actual failure by hours or days, allowing us to detect potential problems, investigate them, and mitigate them before a failure. On multiple occasions, this has enabled us to make adjustments and replace equipment before any problems occurred.

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.

The next section describes CAI system architecture and how it recovers from device failures, Internet interruptions, and other performance issues.

### **2.7.1 High-Level System Architecture**

CAI system architecture provides the redundancy, robustness, and reliability required by a large-scale, high-stakes testing program. Our general approach is pragmatic and well supported by its 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 can 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. Fault tolerance and automated recovery are built into every component of the system.

Key elements of the testing system, including the data integrity processes at work at each step, are described in this section.

#### **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 (e.g., computer, iPad) waiting for confirmation of successfully stored data on the server. If confirmation is not received within the designated time (usually 30–90 seconds), the system will prevent the student from doing any more work until connectivity is restored. The student is offered the choice of asking the system to try again or pausing the test and returning later. For example:

- If connectivity is lost and restored within the designated time period, the student may be unaware of the momentary interruption.
- If connectivity cannot be silently restored, the student is prevented from testing and given the option of logging out or retrying to save.

- If the system fails completely, upon logging back in to 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 a Redundant Array of Independent Disks (RAID) system to mitigate the risk of disk failure. Each response is stored on multiple independent disks.

One server 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 system failure, data are completely protected. Satellites are automatically monitored and, upon failure, are removed from service. Real-time student data are immediately recoverable from the satellite, hub, or backup hub, with backup copies remaining on the drive arrays of the disabled satellite.

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

### **Hub**

Hub servers are redundant clusters of database servers with RAID 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 with RAID subsystems, providing redundant capability to prevent data loss in the event of server or disk failure. At the normal conclusion of a test, these servers receive completed tests from the test delivery satellites. Upon successful completion of the storage of the information, these servers notify the hub and satellites that it is safe to delete student data.

### **Quality Assurance System**

The quality assurance (QA) system gathers data, 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) are flagged, and a notification goes out to our psychometricians and project team immediately.

### **Database of Record**

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

### **2.7.2 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.

### **2.7.3 Other Disruption Prevention and Recovery**

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

- The system’s hosting provider has redundant power generators that can continue to operate for up to 60 hours without refueling. With multiple refueling contracts in place, these generators can operate 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 a complete service failure caused by an unlikely network cable cut.
- On the network level, we have redundant firewalls and load balancers throughout the environment.
- The system uses redundant power and switching within all of our server cabinets.
- Data are protected by nightly backups. We complete a full weekly backup and incremental backups nightly. Should a catastrophic event occur, CAI is able to reconstruct real-time data using the data retained on the TDS satellites and hubs.
- The server backup agents send alerts to notify system administration staff in the event of a backup error, at which time they will inspect the error to determine whether the backup was successful or needs to be re-run.

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 data are always stored in at least two locations in the event of failure. The engineering that led to this system protects the loss of student-response data.

### 3. SUMMARY OF SPRING 2023 OPERATIONAL TEST ADMINISTRATION

#### 3.1 STUDENT PARTICIPATION

The Alternate Montana Science Assessment (AMSA) was administered by grade level. All students in grades 5, 8, and 11 were assessed in science. In the AMSA, a student needs to respond to at least one item for the test to be considered as attempted and the student to be considered participated.

Table 4 presents the total number of students who participated in the assessment by grade. Table 5 presents the total number of students who participated by demographic subgroup. Table 6 presents the total number of students who participated by demographic subgroup and the Individuals with Disabilities Education Act (IDEA) disability category for each grade.

Table 4. AMSA Number of Participated Students

Grade	Online Adaptive				Fixed-Form with Paper-pencil Accommodation				Total Attempted
	Completed	ESR*	Incomplete	Not Attempted	Completed	ESR*	Incomplete	Not Attempted	
5	95	7	1		6				109
8	90	2			2	1			95
11	85	4							89

\* Early Stopping Rule

Table 5. AMSA Number of Attempted Students by Subgroup

Group	Grade 5	Grade 8	Grade 11
All	109	95	89
Female	45	39	37
Male	64	56	52
American Indian or Alaskan Native	15	12	9
Asian	1	4	
Black or African American		1	2
Hispanic or Latino	9	5	3
White	77	66	71
Native Hawaiian or Other Pacific Islander			
Multi-Racial	7	7	4
LEP			
Section 504 Plan			

Table 6. AMSA Number of Participated Students by Subgroup and Disability Category

Group	AU	CD	DB	ED	LD	MD	OI	OHI	SL	TB	VI
Grade 5											
All Students	20	20			2	50	2	11	1	2	1
Female	4	13				16	2	7	1	2	
Male	16	7			2	34		4			1
American Indian or Alaskan Native	1	3			1	7		2		1	
Asian							1				
Black or African American											
Hispanic or Latino	5					4					
White	11	15			1	37	1	9	1	1	1
Native Hawaiian or Other Pacific Islander											
Multi-Racial	3	2				2					
Grade 8											
All Students	16	27	1		1	42	1	6	1		
Female	3	14	1			15	1	4	1		
Male	13	13			1	27		2			
American Indian or Alaskan Native		5	1		1	4		1			
Asian	1					2			1		
Black or African American		1									
Hispanic or Latino	1	2				2					
White	13	17				30	1	5			
Native Hawaiian or Other Pacific Islander											
Multi-Racial	1	2				4					
Grade 11											
All Students	11	24		1	2	44	1	5			
Female	1	15			2	17	1				
Male	10	9		1		27		5			
American Indian or Alaskan Native		4				4					
Asian											
Black or African American	1							1			
Hispanic or Latino		1				1		1			
White	9	18		1	1	38	1	3			
Native Hawaiian or Other Pacific Islander											
Multi-Racial	1	1			1	1					

Note. AU=Autism; CD=Cognitive Delay; DB=Deaf-Blindness; ED=Emotional Disturbance; HI=Hearing Impairment; LD=Learning Disability; MD=Multiple Disabilities; OI=Orthopedic Impairment; OHI=Other Health Impairment; SL=Speech-Language Impairment; TB=Traumatic Brain Injury; VI=Visual Impairment

### 3.2 SUMMARY OF STUDENT PERFORMANCE

Table 7–Table 9 present a summary of the spring 2023 AMSA test results for all students and by subgroup, including the average and the standard deviation (SD) of scale scores, the percentage of students in each performance level, and the percentage of proficient (Level 3 and Level 4) students. The results were based on the students who met attemptedness requirements for scoring and reporting of the AMSA.

Table 7. Student Performance Overall and by Subgroup, Grade 5

<b>Group</b>	<b>Number Tested</b>	<b>Scale Score Mean</b>	<b>Scale Score SD</b>	<b>% Level 1</b>	<b>% Level 2</b>	<b>% Level 3</b>	<b>% Level 4</b>	<b>% Proficient</b>
All Students	109	277.43	65.14	36	28	28	8	37
Female	45	275.33	65.04	33	31	27	9	36
Male	64	278.91	65.68	38	25	30	8	38
American Indian or Alaskan Native	15	290.2	63.37	27	20	40	13	53
Asian	1*							
Black or African American								
Hispanic or Latino	9*							
White	77	279.49	67.84	34	27	30	9	39
Native Hawaiian or Other Pacific Islander								
Multi-Racial	7*							

\* Results for n<10 are suppressed.

Table 8. Student Performance Overall and by Subgroup, Grade 8

<b>Group</b>	<b>Number Tested</b>	<b>Scale Score Mean</b>	<b>Scale Score SD</b>	<b>% Level 1</b>	<b>% Level 2</b>	<b>% Level 3</b>	<b>% Level 4</b>	<b>% Proficient</b>
All Students	95	283.92	47.38	24	36	28	12	40
Female	39	276.03	38.44	26	44	31	0	31
Male	56	289.41	52.36	23	30	27	20	46
American Indian or Alaskan Native	12	289.75	33.75	17	58	8	17	25
Asian	4*							
Black or African American	1*							
Hispanic or Latino	5*							
White	66	281.29	52.43	27	30	32	11	42
Native Hawaiian or Other Pacific Islander								
Multi-Racial	7*							

\* Results for n<10 are suppressed.

Table 9. Student Performance Overall and by Subgroup, Grade 11

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
All Students	89	293.19	62.61	28	29	16	27	43
Female	37	290.92	65.68	30	30	16	24	41
Male	52	294.81	60.93	27	29	15	29	44
American Indian or Alaskan Native	9*							
Asian								
Black or African American	2*							
Hispanic or Latino	3*							
White	71	290.51	65.14	28	32	11	28	39
Native Hawaiian or Other Pacific Islander								
Multi-Racial	4*							

\* Results for n<10 are suppressed.

### 3.3 TEST-TAKING TIME

The AMSA tests are not timed. The time spent on each item may vary among individual students, which may provide useful information about student testing behaviors and motivation. Since the length of a test session was monitored by Test Administrators (TAs) who are familiar with their students, additional time for students who needed it was arranged.

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

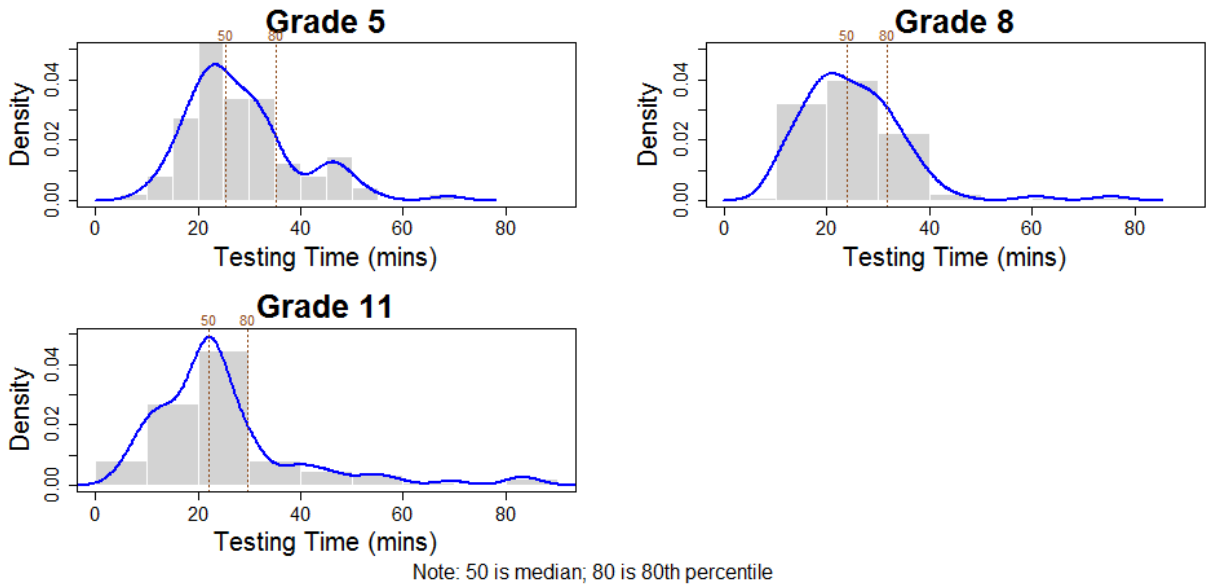
Table 10 presents an average testing time and the testing time at various percentiles for the overall test. The distribution of testing time is provided in Figure 1.

Table 10. Test-Taking Time

Grade	Average Testing Time (hh:mm)	Median Testing Time (hh:mm)	Testing Time in Percentiles (hh:mm)					
			Min	75 <sup>th</sup>	80 <sup>th</sup>	85 <sup>th</sup>	90 <sup>th</sup>	Max
5	00:28	00:25	00:09	00:32	00:35	00:40	00:45	01:08
8	00:24	00:24	00:09	00:30	00:31	00:32	00:35	00:44
11	00:24	00:22	00:04	00:27	00:29	00:32	00:41	01:24



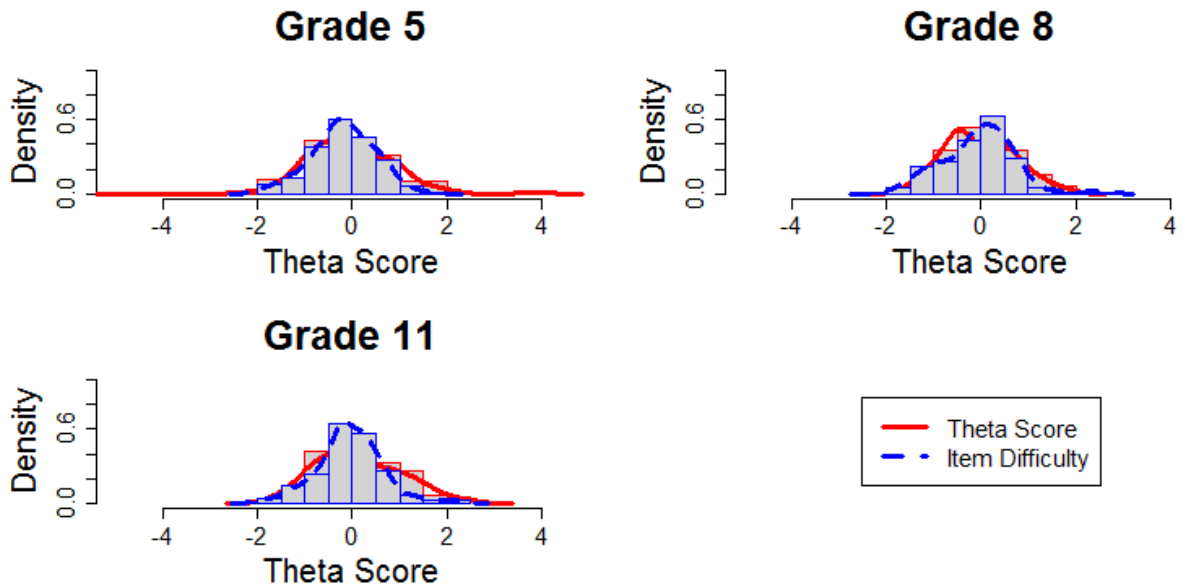
Figure 1. Distribution of Science Testing Time



### 3.4 DISTRIBUTION OF STUDENT ABILITY AND ITEM DIFFICULTY

Figure 2 displays the empirical distribution of students’ overall theta scores and the distribution of the operational item difficulty parameter estimates. The distributions were based on the completed test results from both the adaptive and fixed-form tests.

Figure 2. Student Ability–Item Difficulty Distribution for AMSA



## 4. ITEM DEVELOPMENT

### 4.1 ITEM DEVELOPMENT FOR THE MOU-ALT

Hawaii, South Carolina, and Wyoming signed a Memorandum of Understanding (MOU) on item development item sharing and field testing in 2018 for English language arts (ELA), mathematics, and science. Each state contributed a predetermined number of items proportional to their state’s student population for the alternate assessment. In early 2019, Idaho and Vermont joined the collaborative item development and field testing efforts and participated in the spring 2019 field test for ELA, mathematics, and science. In 2020, Montana and South Dakota joined the MOU for science. In 2022, Vermont exited the MOU.

For the first year of the alternate assessment MOU shared field test item development, a crosswalk among all the individual state alternate assessment standards was completed. Test items from each of the original three states could then be aligned across states. Once all individual state items were aligned, item development plans were created for each state. These plans were based on identified areas where additional items were needed to ensure that all MOU standards aligned on the crosswalk were addressed in the shared field test items, and items for each state-specific standard or content specification that was not aligned to the MOU 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 is conducted, and a new item development plan is created. As new states joined the MOU Alternate Assessment agreement, or in cases where states changed their standards, the individual state standards were added to the crosswalk so that items from the state could be aligned across all participating states.

Starting in 2017, items were developed for the state-shared MOU Alternate Assessment field test pool. All items were developed by a group of professional item writers that included both 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 item aspects that would be unique to students with significant cognitive disabilities. A group of senior test development specialists monitored and supported item development activities.

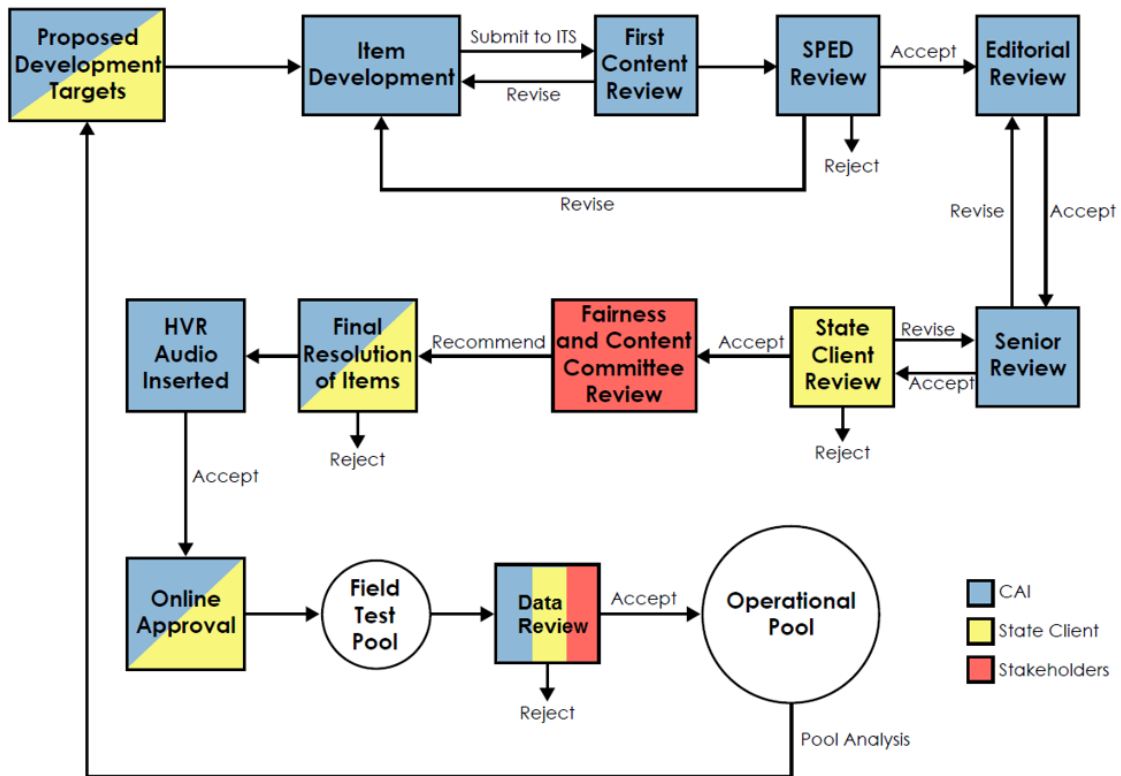
The development process begins with establishing Cambium Assessment, Inc.’s (CAI) proposed development targets and working with individual states to edit them and accept a final plan. The CAI Content Team then starts item development. After the initial round of development, the items go through a group review that includes content and senior reviewers, followed by an individual content review phase, where edits are made based on group reviews, and then a special education review. After items are reviewed by the special education reviewer, the items go through an editorial review then go back through a senior review as the last step of review at CAI before the items are sent to each state for client review.

During client review, state experts either accept the items as they are provided, recommend edits, or reject the items. After client comments and issues are resolved, all accepted items are then taken to a stakeholder Content and Fairness Committee review. After the committee makes its recommendations, the states and CAI go through a final edit resolution process. The items then go through an approval phase 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 closes, all field-tested items are analyzed. Items with sample sizes smaller than 50 are archived and field tested in future years. Items with a negative biserial/polyserial correlation are first verified by CAI content specialists to ensure they were not mis-keyed 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. Finally, accepted items are moved into the operational item pool for each individual state.

Figure 3 presents the item development process described above.

Figure 3. Alternate Assessment Item Development Process



Notes. SPED = Special Education; HVR = Human Voice Recording.

#### 4.1.1 Item Type and Scoring Rubrics

The MOU shared field test pool has multiple-choice (MC) items and multi-select (MS) items. The MC items have 2–4 options with one key. The MS items have up to five options with two keys. For MC items, if a student selects the key, he or she receives one point; otherwise, the student receives zero points. For MS items, if a student selects two keys, he or she earns two points; if the student selects one key, he or she earns one point; otherwise, the student earns zero points. Each item measures a specific content standard.

Starting in late spring 2018, cognitive 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 field test items. The item types included MS, equation editor, table match, and animation items. Neither equation editor nor table match proved to be a successful item type for this population of students, and therefore, states will not

develop these item formats in future. MS items were successful for high-functioning middle school and high school students and will continue to be developed for this segment of the alternate assessment population.

## **4.1.2 Item Development Procedure and Item Reviews**

### **4.1.2.1 Item Development Procedure**

Items were 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. After items passed the required stages of CAI internal reviews, described at length in the following section, items were then presented to the state for department review and acceptance. Following a state’s item approval, the other sharing state partners were notified that the items could be reviewed and commented on. During this review step, states could also verify whether the items aligned to their own state standards. Any comments regarding item content and suggested revisions were sent to the state that owned the items, and it was that state’s determination whether these comments should be acted upon.

In each state, items owned by the state that were accepted were prepared for review by a state-wide Content and Fairness Committee convened for each content area in each state. The Content and Fairness Committee was comprised of stakeholders from around the state with teaching experience in the content area in grades K–12 and/or experience working with students with disabilities. Additional stakeholders with expertise in specific disability categories, stakeholders with multi-cultural/foreign language expertise, and stakeholder parents were invited to participate in the committee meetings. These specific stakeholders reviewed the items and provided feedback to ensure that all accepted items are correct and free from fairness and bias issues. Most importantly, these educators made sure that this population of students will be able to 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 were then shared across the state item banks for field testing. See Figure 3 for a flowchart that documents the item development process.

### **4.1.2.2 Item Reviews**

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

*CAI Review: Items are reviewed at CAI at various levels.*

- **CAI Internal Group 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:** Following group review, the preliminary review is conducted by a member of CAI’s content team assigned to the alternate assessments. As agreed upon in the group review, items are revised to eliminate initial errors, meet content standards, and meet internal style and clarity expectations.
- **CAI Internal Content Review:** A second content review occurs after the preliminary review to further ensure changes based on the group review, and to revise items further, as necessary, to address any content, alignment, clarity, accessibility, and errors.

- **Special Education Review:** At this stage, items are reviewed by a CAI special education expert. The 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 wide 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 sends items back to the content reviewer to implement changes.
- **Edit Review:** After the special education reviewer approves the items, they send them through an editorial review. At this stage, 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, spelling error check), and the 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 items are ready for state review.

#### *State Review*

At this level, items are compared to the state standards and state content specifications. The items are also reviewed against the Performance-Level Descriptors at all difficulty levels and compared to the blueprint. Items are further reviewed to ensure that they align to 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 resubmit for a second state 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 brought to the Content and Fairness Committee for further review. The review committee includes special educators, general educators, vision and hearing specialists, school principals, and special education directors. The review committee members represent a diversity of gender, ethnicity, disability, race, and cultural subgroups across the state. During the review meeting, each item is reviewed and assured to meet bias and sensitivity guidelines, is aligned to content standards, and is determined to abide by the principles of universal design (UD).

The common criteria used for item review are:

- Content accuracy and clarity
- Alignment to the content specifications
- Appropriate scoring rubrics

- Correct answer key and appropriate distractor(s) 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 in order to see the item just as the student will see it.

### **4.1.3 Development of Crosswalk and State Alternate Content Performance Standards**

Before item development began, the alternate performance content standards for each state were compared in a crosswalk created by senior test development specialists. The crosswalk was based on each state’s blueprint and includes the general education and alternate performance standards for each state. Each state has a unique set of alternate performance content standards as follows:

- Hawaii Essence Statements and Performance Level Descriptors
- Idaho Extended Content Standards Core Content Connectors
- Montana Content Standards in Science
- South Carolina Prioritized Standards and Performance Level Descriptors
- South Dakota Science Standards and Core Content Connectors, and ALDs
- Wyoming Extended Standards and Instructional Performance Level Descriptors

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

The crosswalk then informed the development of item specifications. Each item specification included the general education standard, followed by the state-specific alternate performance standards that align to the general education standard. The item specifications also included complexity statements and task demands. The language of the complexity statements and task demands was derived from each state’s performance standards, where applicable, and synthesized in an effort to drive items that aligned to multiple states. Once completed, the item specifications were sent to each state for review to confirm alignment and overall approach.

The states’ content performance standards 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 standard overlap across the states. The complexity and vocabulary matrices were subdivided into three categories of cognitive complexity: Low, Moderate, and High. The states’ content performance standards were also analyzed to reveal state-

specific and cross-state content limits in the content extensions. These were listed in the Content Limits section.

All the above analysis was 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 performance standards. 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 refers to the numbered list of task demands as a reference.

## 4.2 FIELD TESTING

Items that survived Content and Fairness Committee review were field tested in the spring test administration of the following year as embedded field-test items. The 2023 alternate assessment operational tests were administered online using a computer-adaptive testing (CAT) design for grades 5, 8, and 11. The CATs were assembled using CAI’s adaptive testing algorithm. The adaptive item selection algorithm selected items based on their content value and information value.

Embedding field-test items among operational items yields item parameter estimates that capture all the contextual effects that contribute to item difficulty in operational test administrations and is especially useful 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 administrations, item parameter estimates are more stable over time than they may be when obtained through stand-alone field testing.

Following the spring 2023 operational test administration, all field-test items were calibrated anchoring on the operational item parameter estimates and placed on the same scale as the existing operational items in the pool.

The spring 2023 field-test item pool consisted of items that were shared across MOU-Alt states and the items that were unique within each state. The field-test items shared across MOU states were administered in the MOU states after obtaining the state’s approval, while state-only field-test items were administered in the state only. The spring 2023 item pool is summarized in Table 11. There are no Montana-specific field-test items.

Table 11. Summary of the 2023 Field-Test Item Pool Across MOU-Alt States

Grade	MOU						Total
	HI	ID	MT	SC	SD	WY	
5	22	10	5		7	30	74
8	21	6	2		6	4	39
11	8	7	3	1	7	20	46

### 4.2.1 Item Statistics

Following the close of spring testing windows, 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. Analyses included classical item statistics and item response theory (IRT) calibrations. Item analyses were conducted based on the combined data across all MOU-Alt states.

Classical item statistics were designed to evaluate the relationship of each item to the overall scale, evaluate the quality of the distractors, and identify items that may exhibit bias across subgroups (differential item functioning [DIF] analyses). The IRT item analyses allowed examination of the fit of items to the measurement model and provided the statistical foundation for operational form construction and test scoring and reporting. Items were flagged if analyses indicated resulting values out of range. Flagged items were reviewed by Montana stakeholders, CAI, and MOU-Alt state staff. Items that passed CAI and MOU states statistical review were accepted for future operational use.

#### 4.2.2 Classical Statistics

Classical item analyses ensured that the field-test items function as intended with respect to the MOU-Alt's underlying scales. CAI's analysis program computed the required item and test statistics for each dichotomous and polytomous item to check the integrity of the item and to verify the appropriateness of its difficulty level. Key statistics included item difficulty, item discrimination, and distractor analysis.

Items that were either extremely difficult or extremely easy were flagged for review but not necessarily rejected if they aligned with the test and content specifications. For dichotomous items, the proportion of test takers in the sample selecting the correct answer ( $p$ -value) was computed, as well as those selecting the incorrect responses. For items with 0–2 score points, item difficulty was calculated both as the item's mean score and as the average proportion correct (analogous to  $p$ -value and indicating the ratio of an item's mean score divided by the number of points possible). Items were flagged for review if the  $p$ -value was less than .25 or greater than .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 could differentiate between high- and low-achieving students. The discrimination index was calculated as the correlation between the item score and the student's IRT-based ability estimate. Items were flagged for subsequent reviews if the biserial/polyserial correlation for the keyed (correct) response was less than 0.20. For polytomous items, we also computed 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 was used to identify items that had 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 was 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 .05.

The flagging criteria based on classical item analysis are summarized in Table 12.



Table 12. 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-points 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 were used to estimate the IRT model parameters for dichotomously and polytomously scored items, respectively. The Winsteps output showed the item statistics resulting from anchoring the field-test items on the operational items. Item fit was evaluated via the mean square Infit and mean square Outfit statistics reported by Winsteps, which were based on weighted and unweighted standardized residuals for each item response, respectively. These residual statistics indicated the discrepancy between observed item responses and the predicted item responses based on the IRT model. Both fit statistics had an expected value of 1. Values substantially greater than 1 indicated model underfit, while values substantially less than 1 indicated model overfit (Linacre, 2004). Items were flagged if Infit or Outfit values were 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 sometimes it is a clue that an item contains a cultural or other bias. Not all items that exhibit DIF are biased; 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, DIF can indicate bias, so all field-tested items were evaluated for DIF, and all items exhibiting DIF were flagged for further examination by CAI and the MOU-Alt states.

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

- Female vs. Male
- African-American vs. White
- Hispanic or Latino vs. White
- Severe and Moderate Mental Disability vs. Other

CAI uses a generalized Mantel-Haenszel (MH) procedure to evaluate DIF. The 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, so that standard errors based on the assumption of 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 items, design effects in student samples are rarely reduced to the level of a simple random sample.

The ability distribution was divided into a configurable number of intervals to compute the Mantel-Haenszel chi-square ( $MH \chi^2$ ) DIF statistics. The analysis program computed 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; the MH chi-square, the standardized mean difference (SMD), and the standard error of the SMD for the polytomous items.

Items were 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 13. Items were 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). Items were flagged if their DIF statistics fall into the “C” category for any group. 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 less than 200 students (Mazor, Clauser, & Hambleton, 1992; Camilli & Shepard, 1994; Muniz, Hambleton, & Xing, 2001; Sireci & Rios, 2013).

Table 13. DIF Classification Rules Science

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

#### 4.2.5 Summary of Item Statistics

This section presents a summary of results from the classical item analysis and item calibration analysis of the spring 2023 MOU-Alt embedded field-test items. Table 14 presents the average sample size and the sample size at various percentiles for the MOU field-test items. Table 15 summarizes item statistics for  $p$ -values, biserials/polyserials, item difficulties, infit and outfit by percentile, and the range for all MOU science items. For each item statistics, e.g.,  $p$ -values, the percentiles were computed across items. The column “Total MOU Items” shows the number of items in the MOU field-test pool that were used in the computation of the percentiles. Table 16 presents the number of items in each DIF classification category.

Table 14. 2023 MOU Item Sample Size Distribution

Subject	Grade	Total MOU Items	Average Sample Size	Sample Size in Percentiles								
				Min	5 <sup>th</sup>	10 <sup>th</sup>	25 <sup>th</sup>	50 <sup>th</sup>	75 <sup>th</sup>	90 <sup>th</sup>	95 <sup>th</sup>	Max
Science	ES	74	88	13	23	33	53	71	94	188	201	218
	MS	39	218	21	41	66	92	115	374	394	395	399
	HS	46	180	27	33	35	63	119	314	356	369	386
	<b>Overall</b>	<b>159</b>	<b>146</b>	<b>13</b>	<b>27</b>	<b>35</b>	<b>58</b>	<b>91</b>	<b>203</b>	<b>367</b>	<b>383</b>	<b>399</b>

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

Table 15. Summary of Item Analyses Results for MOU-Alt Science

Grade	Total MOU Items	Statistics	Min	P10	P25	P50	P75	P90	Max
ES	74	<i>p</i> -value	0.13	0.26	0.35	0.45	0.56	0.69	0.81
		Biserial/Polyserial	-0.50	-0.03	0.14	0.30	0.53	0.63	0.94
		Step Difficulty	-1.98	-1.14	-0.55	-0.07	0.41	0.73	1.52
		Infit	0.70	0.86	0.92	1.03	1.13	1.23	1.51
MS	39	Outfit	0.52	0.79	0.89	1.04	1.16	1.26	2.86
		<i>p</i> -value	0.27	0.34	0.40	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
HS	46	Infit	0.85	0.87	0.93	1.00	1.09	1.14	1.20
		Outfit	0.64	0.79	0.87	0.99	1.10	1.16	1.18
		<i>p</i> -value	0.21	0.30	0.39	0.49	0.60	0.67	0.79
		Biserial/Polyserial	-0.20	-0.09	0.11	0.28	0.40	0.60	0.82
HS	46	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	0.80	0.92	1.04	1.14	1.32	1.59

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

Table 16. 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
<b>Science</b>								<b>Science</b>							
ES	15	9	6					ES	2	1	1				
MS	18	7	10				1	MS	18	5	12		1		
HS	22	11	11					HS	19	10	9				
		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
<b>Science</b>								<b>Science</b>							
ES								ES	7	4	3				
MS	4		4					MS	18	5	11			1	1
HS	2	1	1					HS	19	7	12				

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

## 4.2.6 Data Review Meeting

### 4.2.6.1 MOU-Alt Shared Items

Items flagged for undesired statistics were reviewed in the MOU and reviewed by content experts in OPI. In addition to the statistical flag, CAI flagged and removed the items with the sample size less than 50 or negative biserial/polyserial correlations for the key. These items were removed from the item pool before data review and were not seen by the data review committees.

The MOU-Alt data review committee consisted of staff across MOU states, CAI content specialists, special education specialists, and psychometricians. During the meetings, the committees were charged with identifying any defects that might have led to the undesired statistics of the items and then asked to render a decision on the items. Committees could choose to reject the item completely, accept the item with modifications for further field testing, or accept the item without any changes. Items accepted without modification are included in the Montana state operational item pool.

Table 17 presents a summary of the MOU-Alt data review results.

Table 17. Summary of the Item Data Review for MOU-Alt Shared Items

Subject	Grade	Total Number of MOU Items	Items with N < 50	Items with Biserial < 0	Total Reviewed Items for IDR	Items Rejected by IDR Committee
Science	5	74	14	5	20	1
	8	39	3	0	14	2
	11	46	8	6	8	0

### 4.2.6.2 AMSA Item Pool

All AMSA field-test items in spring 2023 were from the MOU-ALT shared items. Montana Office of Public Instruction (OPI) confirmed the content alignments for all items in the AMSA item pool and rejected items that did not align to the Montana Content Standards in science specified in the test blueprints.

Table 18 presents a summary of the AMSA field-test items in spring 2023.

Table 18. Summary of AMSA Field-Test Items

Grade	Total # of Items Administered	Items with n < 50	Items with bis < 0	Rejected Items	Eligible Items for Operational Pool
5	45	5	2	4	34
8	21	0	0	4	17
11	29	1	5	5	18

## 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

assume 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  $\theta$  represents a student's true proficiency.

Traditional item response models differ only in the form of the function  $P(Z)$ . The one-parameter logistic 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

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

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 difficulty of the item. The 1PL 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) partial credit model. Under Masters' partial credit model, the probability of getting a score of  $x_i$  on item  $i$  given ability  $\theta$  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 field-test items were calibrated by anchoring on the operational item parameters under the CAT test design. All completed records were included in the IRT analysis. Through this anchoring process, field-test item parameter estimates were placed on the same MOU scale as the operational items. These operational item parameters will be used as a reference scale to calibrate new items in the following years.

Winsteps was used to estimate Rasch and Masters' partial credit model item parameters for the MOU-Alt. Winsteps is a publicly available software program from Mesa Press. Winsteps employs a joint maximum likelihood estimation (JMLE) approach towards estimation, 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) partial credit model, 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 particular interpretations for specified uses, and thus, the validation process starts logically with well-articulated statements on intended uses of test scores. Arguments of logical, theoretical, and empirical evidence are then provided to support the intended uses.

This chapter will first present the statements on intended uses of the Alternate Montana Science Assessment (AMSA) test scores, followed by various sources of evidence validating the interpretation of test scores for the intended uses.

### **5.1 INTENDED USES AND INTERPRETATIONS OF AMSA SCORES**

Development and design of the AMSA 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 comprised of theoretical, logical, and empirical components.

The intended uses of the AMSA score include

- measuring students’ academic achievements and progress in core content areas taught in school;
- measuring achievement and progress toward meeting the state performance standards; 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.

In realizing these uses, the AMSA 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 performance and progress toward meeting the state performance standards is documented separately in greater detail in the standard setting technical report. Chapter 8 in this technical report provides a high-level overview of the standard-setting procedure and results.

### **5.2 SOURCES OF VALIDITY EVIDENCE**

According to the Standards (AERA, APA, & NCME, 2014), “A sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses” (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 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 AMSA depends on the assessments meeting the relevant standards of validity.

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

Validity evidence for the AMSA 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 parentheses.

- 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)

For the AMSA, evidence on test content validity is provided with both theoretical 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.

### **5.2.1 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 state-wide standards implemented by teachers in daily instruction at schools. This evidence is based on the justification for and connections among several factors, including:

- 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 early chapters, are summarized as follows.

#### **Content Specifications**

Content standards and content specification is the starting point for test development. The AMSA is aligned with the Montana Content Standards in Science. It is designed for students with the most significant cognitive disabilities. The purpose of the AMSA is to maximize access for 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 1.3, 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 the test has a balanced representation of items from each content standard.

For the AMSA, each student received 40 operational items. Item selection took place in two discrete stages: blueprint satisfaction and match-to-ability. Table presents the matching rates based on the completed online adaptive tests in grades 5, 8, and 11. The adaptive algorithm selected items for all tests according to the blueprint requirements - 100% match at the overall strand level.

Table 19. Percentage of Administered Tests Meeting Blueprint Requirements

Grade	Standard	Minimum Required Items	Maximum Required Items	% BP Match
5	Earth and Space Science	12	15	100
	Life Science	12	15	100
	Physical Science	12	15	100
8	Earth and Space Science	12	15	100
	Life Science	12	15	100
	Physical Science	12	15	100
11	Earth and Space Science	12	15	100
	Life Science	18	21	100
	Physical Science	9	12	100

**Item Development**

Chapter 4 – Item Development, provides 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 template 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 Administrations (TAs) do not follow the same procedures, student performance cannot be compared meaningfully. For the AMSA, TAs are required to complete an online TA Certification Course before they can administer the AMSA tests to their students. The guidelines for test administration are summarized in the Test Administration Manual (TAM). See Chapter 2 – Test Administration for details.

### **Item and Test Scoring**

Item and test scores are probably the most critical element. All interpretations are established around students' test results. Every effort is made to ensure absolute accuracy on item and test scores. Section 10.3 Assurance in Test Scoring, provides detailed description on quality control and monitoring procedures implemented within CAI to assure accurate scores are generated and reported.

### **5.2.2 Evidence Based on Response Process**

Cognitive lab studies document validity evidence to show that the assessments tap the intended cognitive processes appropriate for each grade level as represented in the State's alternate academic contents standards. Cognitive lab studies conducted in each state 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 a variety of disability categories and demonstrate many concomitant learning difficulties. Students in this population can exhibit difficulties responding to stimuli; challenges committing information to working, short-term, or long-term memory; difficulties generalizing learning to familiar and novel environments; difficulties with meta-cognition; or self-regulating behaviors. Furthermore, students with significant cognitive disabilities may also demonstrate significant communication and/or sensory deficits; limited fine or gross motor skill abilities; specialized health care needs; or an inability to synthesize learned skills. Students with significant cognitive disabilities require multiple opportunities to engage with academic content and daily activities, as well as multiple ways to express and represent their knowledge.

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

In spring 2019, Hawaii and Wyoming conducted cognitive lab studies. Students with significant cognitive disabilities at all grade levels from each of the three cognitive levels (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, including intellectual disability, autism spectrum, and multiple handicaps.

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 we 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 are verbal, and not all students have full mobility, and some may use eye gaze to indicate their responses. Therefore, several different methods must 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.

### 5.2.3 Evidence Based on Internal Structure

The measurement and reporting model used in the AMSA 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 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  $r_{x|y'} = \frac{r_{xy}}{\sqrt{r_{xx}} \cdot \sqrt{r_{yy}}}$ , where  $r_{x|y'}$  is the correlation between  $x$  and  $y$  corrected for attenuation,  $r_{xy}$  is the observed correlation between  $x$  and  $y$ ,  $r_{xx}$  is the reliability coefficient for  $x$ , and  $r_{yy}$  is the reliability coefficient for  $y$ .

Correlations among content strand scores are presented in Table 20. Values above the diagonal are disattenuated correlations, values below the diagonal are observed correlations, and values on the diagonal are marginal reliabilities of the strands. When corrected for attenuation (above diagonal), the correlations among strand scores are higher than observed correlations.

Table 20. AMSA Correlations Among Strands

Grade	Strand	Observed & Disattenuated Correlation		
		Strand 1	Strand 2	Strand 3
5	Strand 1: Earth and Space Science	<b>0.56</b>	1	1
	Strand 2: Life Science	0.64	<b>0.61</b>	1
	Strand 3: Physical Science	0.63	0.71	<b>0.71</b>
8	Strand 1: Earth and Space Science	<b>0.49</b>	0.86	0.70
	Strand 2: Life Science	0.46	<b>0.58</b>	0.91
	Strand 3: Physical Science	0.40	0.57	<b>0.66</b>
11	Strand 1: Earth and Space Science	<b>0.59</b>	1	0.78
	Strand 2: Life Science	0.66	<b>0.75</b>	0.95
	Strand 3: Physical Science	0.44	0.61	<b>0.55</b>

## 5.2.4 Evidence Based on Relations to Other Variables

The publication of *A State’s Guide to the U.S. Department of Education’s Assessment Peer Review Process* (U.S. Department of Education, 2018) indicates that adequate validity evidence show the state’s assessment scores are related as expected with other variables. The Montana Office of Public Instruction (OPI) required all teachers of students with severe cognitive disabilities who took the AMSA to complete the Learner Characteristics Inventory (LCI) as part of their Multistate Alternate Assessment (MSAA), which was outside of CAI’s system. CAI then analyzed the results and ran a correlational study. Several of the LCI questions related to student behaviors that might directly impact student performance on the alternate assessment, and all of the grade-specific teacher rating questions of student skills and knowledge in a content area were used. The results of this study are discussed below, following an initial discussion of the purpose and questions extracted from the LCI.

### 5.2.4.1 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 is funded through a four-year General Supervision Enhancement Grant 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 Montana consists of 16 questions covering the following topics which a teacher provides about each student:

1. Student’s grade

2. Student’s age
3. Student’s primary IDEA disability label
4. Whether student’s primary language is English or not
5. Student’s primary language
6. Student’s primary classroom setting
7. Student’s expressive communication skills
8. Student’s use of an 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

The information gleaned that address LCI criteria help the state understand the characteristics of the state’s alternate assessment student population. The LCI is designed to be a descriptive instrument for the state to define this population and to develop participation guidelines for their states’ alternate assessments.

While reviewing the LCI questions administered in Montana, it was observed that several of these questions did yield evidence relevant to the academic performance of these students. These questions include 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** inquiry asks teachers to describe the student’s oral/written or augmentative communication ability. Three levels of descriptors are 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 would do less well on an assessment that relied on symbolic communication. The LCI “expressive communication skills” question would therefore predict, at a broad level, the student’s final score on an assessment.

The **student’s receptive language skills** indicator includes four levels of descriptors, including:

1. The first, or highest, descriptor states that the student can independently follow without additional cues 1–2 step directions presented through words.
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 alerts 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, and students who are able to do so will receive a higher score on an assessment than those who cannot. Therefore, 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 **student’s engagement descriptor**) also has four descriptive statements, including:

1. The first, or highest, 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 into social interactions with others—even if they do not initiate the interaction, as this is not necessary on an assessment—would have more of a chance of success on an assessment than students who do not enter into social interactions with others.

The **student’s reading skills** descriptor relates directly to the student’s reading ability, as well as 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 do better on an assessment than students who read only with literal understanding, and students who read with literal understanding will do 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, as well as 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 with 1:1 correspondence to at least 10
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 only do computation 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.

#### **5.2.4.2 Correlations with LCI Descriptors**

The LCI descriptors on Expressive Language, Receptive Language, Engagement, Reading, and Mathematics, and a composite score by adding five LCI descriptors were correlated with the AMSA scores.

As shown in Table 21, each of these descriptors is moderately correlated with the AMSA scores. The correlations are relatively low between AMSA scores and the LCI Expressive descriptor in grade 11 (0.27) and in grade 8 (0.28). Among the five LCI descriptors, the highest correlation is between the AMSA scores and the “Mathematics Skills” in grade 11 with a value of 0.57.

A teacher’s description of a student’s ability level, as required when completing the LCI, correlates moderately with students’ overall score on the AMSA. 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 the student’s teacher’s independent judgment of that student possessing a higher level of skill.

Table 21. Correlations Between LCI Descriptors and Total Scores in AMSA

Grade	N	Composite	Expressive Communication Skills	Receptive Language Skills	Engage with Others	Reading Skills	Mathematics Skills
<b>Science</b>							
5	102	0.61	0.53	0.40	0.52	0.44	0.48
8	92	0.50	0.28	0.36	0.30	0.31	0.45
11	85	0.58	0.27	0.52	0.39	0.52	0.57

## 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 the test information, which describes the amount of information provided by the test at each score point along the ability continuum. Test information is inversely related to the measurement error of the test; the larger the measurement error, the less test information is being provided.

The reliability evidence of the Alternate Montana Science Assessment (AMSA) is provided with marginal reliability, SEM, conditional SEM, and classification accuracy and consistency for each performance standard.

### 6.1 MARGINAL RELIABILITY

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

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

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

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

Another way to examine test reliability is with SEM. Under item response theory (IRT), SEM is estimated as a function of test information provided by a given set of items that makes up the test. In computer-adaptive testing (CAT), since items administered vary among all students, the SEM can vary across students, which yields conditional SEM. The average conditional SEM across all students can be computed as

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

The smaller the value of average conditional SEM, the greater the accuracy of test scores. Table 22 presents the marginal reliability coefficients and the average conditional SEM for the overall scale scores, based on all completed tests, excluding the early stopped tests.

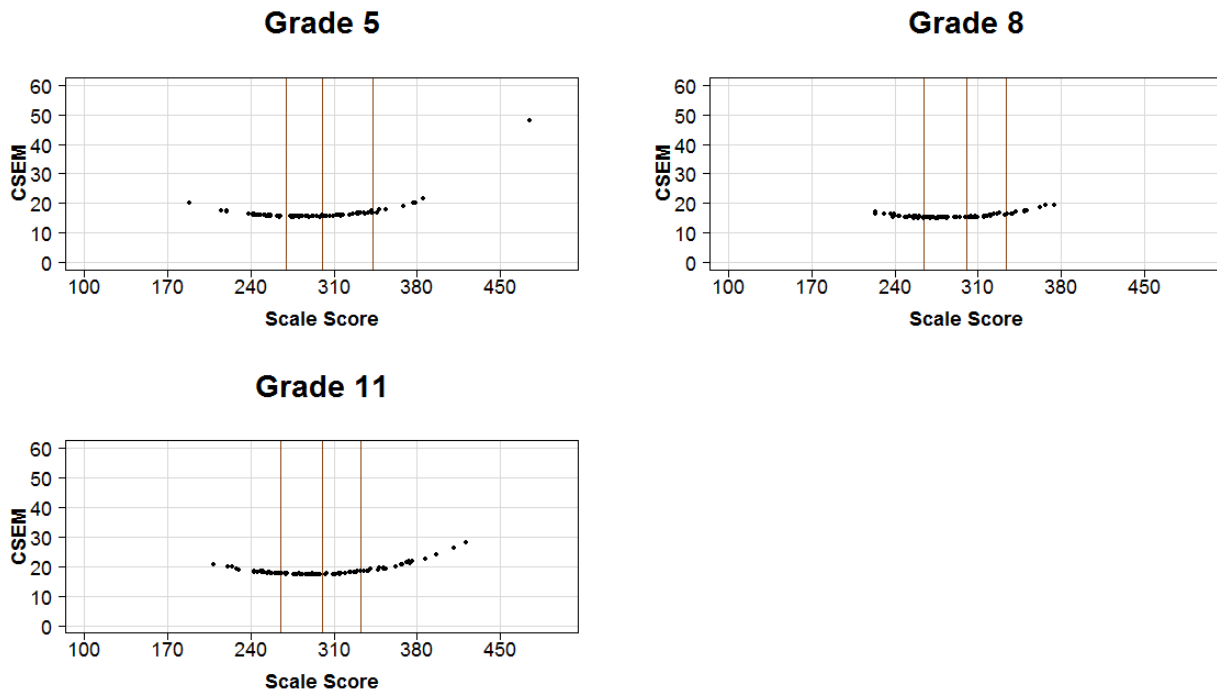
Table 22. Marginal Reliability

Grade	Number of Items	Marginal Reliability	Scale Score Mean	Scale Score Standard Deviation (SD)	Average CSEM
5	95	0.85	293.08	43.74	17.01
8	90	0.79	289.24	34.16	15.80
11	85	0.84	302.28	47.39	18.68

## 6.2 STANDARD ERROR CURVES

Figure 4 presents a plot of the conditional SEM of scale scores across the range of ability for each grade. The vertical lines indicate the cut scores for Level 2, Level 3, and Level 4. Overall, the standard error curves suggest that students are measured with a similar precision along the range of score distribution.

Figure 4. Conditional Standard Error of Measurement for Science



The SEMs presented in Figure 4 are summarized in Table 23, which provides the average conditional SEM for all scores and for scores in each performance level. As shown in Figure 4, the average CSEMs are similar in Level 2 and Level 3 but slightly larger in Level 1 and Level 4, which are expected for tests with extreme scores.



Table 23. Average Conditional Standard Error of Measurement by Performance Level

Grade	Level 1	Level 2	Level 3	Level 4	Average CSEM
5	16.37	15.62	16.31	23.96	17.01
8	15.77	15.25	15.75	17.69	15.80
11	18.38	17.33	17.77	20.76	18.68

### 6.3 RELIABILITY OF PERFORMANCE CLASSIFICATION

When student performance is reported in terms of performance levels, a 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 the *Standards for Educational and Psychological Testing* (AERA, APA, and NCME, 2014). The indexes consider the accuracy and consistency of classifications.

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

The classification index can be examined in terms of the classification accuracy and the classification consistency. Classification accuracy refers to the agreement between the classifications based on the form actually taken and the classifications that would be made based on the test takers’ true scores, if their true scores could somehow be known. Classification consistency refers to the agreement between the classifications based on the form (adaptively administered items) actually taken and the classifications that would be made based on an alternate form (another set of adaptively administered items given the same ability), that is, the percentages of students who are consistently classified in the same performance 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  $\hat{\theta}_i \sim N(\theta_i, se^2(\hat{\theta}_i))$ , assuming a normal distribution where  $\theta_i$  is the unknown true ability of the  $i$ th student. The probability of the true score at performance level  $l$  based on the cut scores  $c_{l-1}$  and  $c_l$  is estimated as

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

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

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

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

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

where the likelihood function based on Rasch IRT models is

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

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  $2 \times 2$  table as

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

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

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

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

the classification accuracy (CA) for the below the cut is estimated by

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

and the overall classification accuracy for the cut 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  $2 \times 2$  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 either below, or at or above the cut, respectively, based on observed scores and hypothetical scores from an equivalent test form.

The classification consistency ( $CC$ ) for the at or above the cut is estimated by

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

the classification consistency ( $CC$ ) for the below the cut 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 24 shows classification accuracy and consistency indexes for the spring 2023 AMSA. 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, but the accuracy index assumes only a single test score and a true score, which does not include measurement error.

Table 24. Classification Accuracy and Consistency for Performance Standards

Grade	Accuracy			Consistency		
	Level 2	Level 3	Level 4	Level 2	Level 3	Level 4
5	0.89	0.91	0.93	0.84	0.86	0.90
8	0.87	0.90	0.94	0.82	0.85	0.91
11	0.90	0.90	0.92	0.85	0.86	0.88

#### 6.4 RELIABILITY FOR CONTENT STRAND SCORES

For the AMSA, although only the overall score was reported, the marginal reliability coefficients and the measurement errors were also computed for strand scores, as shown in Table 25. The reliability coefficients were computed based on the complete tests only.

Table 25. AMSA Marginal Reliability Coefficients for Content Strand Scores

Grade	Strand	Number of Items		N	Marginal Reliability	Scale Score Mean	Scale Score SD	Average CSEM
		Min	Max					
5	Earth & Space Science	13	14	95	0.56	283.95	44.35	28.69
	Life Science	12	14	95	0.61	298.51	49.90	30.36
	Physical Science	13	14	95	0.71	295.27	57.44	30.28
8	Earth & Space Science	12	14	90	0.49	282.86	38.85	27.45
	Life Science	13	14	90	0.58	299.09	47.40	29.63
	Physical Science	13	14	90	0.66	287.51	50.67	29.20
11	Earth & Space Science	12	13	85	0.59	300.56	56.81	35.46
	Life Science	18	19	85	0.75	301.27	56.11	27.86
	Physical Science	9	10	85	0.55	308.05	65.82	42.64

## 7. SCORING

For the Alternate Montana Science Assessment (AMSA), 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

If a student logged in to the test administration system, was presented with one item, and a valid response was entered for that first item, the student was counted as *attempted*. Scores were only generated for attempted tests.

- If a student answered all items in Segment 1 and 2, the test was scored without penalty.
- If a student did not complete Segment 1 and 2, the student was scored with penalty. The penalty was the theta estimate minus one conditional standard error of measurement (SEM) associated with the estimated theta value.
- If a student had four consecutive No Response (NR) recorded by the test administrator (TA) for items within Segment 1, the student was given the lowest obtainable scale score of the test. The SEM and theta score were set to BLANK.

### 7.2 ESTIMATING STUDENT ABILITY USING MAXIMUM LIKELIHOOD ESTIMATION

The AMSA are scored using maximum likelihood estimation (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  $m_i = 1$ ,

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

in the case of items with two or more points,

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

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

The MLE theta is then estimated by finding the value of theta that maximizes the loglikelihood, i.e.,

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

### Standard Error of Measurement

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

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

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

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

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

### 7.3 SCORING ALL CORRECT AND ALL INCORRECT CASES

With item response theory (IRT) maximum likelihood (ML) ability estimation methods, 0 and perfect 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.

### 7.4 RULES FOR TRANSFORMING THETA TO SCALE SCORES

The scale score is the linear transformation of the IRT ability estimate using the scaling constants of  $a$  and  $b$ ,  $SS = a * \theta + b$ , where  $a$  is the slope and  $b$  is the intercept.

Table 26 provides the linear transformation slope and intercept with four decimals. The final reported score for each student is the scale score rounded to the nearest integer.

Table 26. Scaling Constants on the Reporting Metric

Subject	Grade	Slope (a)	Intercept (b)
Science	5	47.2401	292.7852
	8	45.4985	291.2859
	11	52.4858	295.5625

Standard errors of the MLEs are transformed onto the reporting scale using the following formula:

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

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

The scale scores are mapped into four performance levels. Table 27 provides the range of scale scores in each performance level by grade.

Table 27. Range of Scale Scores by Performance level

Grade	Level 1	Level 2	Level 3	Level 4
5	100–269	270–299	300–342	343–500
8	100–263	264–299	300–333	334–500
11	100–264	265–299	300–332	333–500

### 7.5 LOWEST/HIGHEST OBTAINABLE SCALE SCORE (LOSS/HOSS)

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

## 8. PERFORMANCE STANDARDS

After the spring 2022 operational administration, formal standard-setting workshops were conducted in all three grades to recommend performance standards for the Alternate Montana Science Assessment (AMSA). The spring 2022 standard-setting results were used to report student scores in the spring 2023 administration.

In August 2022, following the close of the testing window, Cambium Assessment, Inc. (CAI) under contract to Montana’s Office of Public Instruction (OPI), invited a panel of 36 teachers and administrators to recommend performance standards (new cut scores) for the assessments. OPI recruited a broadly representative panel, ensuring that a diverse range of perspectives informed the standard-setting process. Panelists included special education teachers, curriculum specialists, education administrators, and other stakeholders. The panel was also broadly representative of Montana’s special education teacher population in terms of gender, race/ethnicity, and regional composition. OPI designated the most knowledgeable and experienced panelists at the workshop as table leaders.

For each test, the panelists recommended three cut scores, or performance standards: Level 2 (Nearing Proficiency), Level 3 (Proficient), and Level 4 (Advanced).

### 8.1 STANDARD-SETTING PROCEDURES

Montana used the Bookmark method (Mitzel, Lewis, Patz, & Green, 2001), which is the most common procedure used throughout the country. Using this procedure, the panelists reviewed items ordered by difficulty in an ordered-item booklet (OIB) for each test. Each OIB contains a set of items that meets the test blueprint. The panelists also reviewed the corresponding Montana content standards and Performance-Level Descriptors (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 OPI. The plan was reviewed and approved by OPI before the workshop occurred.

The standard-setting workshop was conducted over two days. The first day was devoted to training and review of materials, and the second day was devoted to two rounds of standard setting. At the end of the activity, the panelists completed a survey that evaluated the workshop.

### 8.2 PERFORMANCE-LEVEL DESCRIPTORS

A prerequisite to standard setting is to determine the nature of the categories into which students are classified. These categories, or performance levels, are associated with Performance-Level Descriptors (PLDs). PLDs link the content standards (range PLDs) to the performance standards. There are four types of PLDs used within the AMSA program (Egan, 2012):

1. **Policy PLDs.** Policy PLDs provide a brief description of the policy goals of each performance level that do not vary across grade or content.
2. **Range PLDs.** These PLDs describe what students should know and be able to do at different proficiency levels and what students know and are able to do throughout the range of each



performance level. For example, the range PLD for Level 3 (Proficient) describes what students know and can do at that level up to just below the Level 4 (Advanced) cut score.

3. **“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.
4. **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.

Montana uses four performance levels to describe student performance:

- 1) Level 1: Novice
- 2) Level 2: Nearing Proficiency
- 3) Level 3: Proficient
- 4) Level 4: Advanced.

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

### **8.3 RECOMMENDED PERFORMANCE STANDARDS**

Panelists were tasked with recommending three performance standards that resulted in four performance levels. Table 28 presents the performance standard in scale score metrics associated with the percentage of students reaching each standard based on the 2022 AMSA results.

Table 28. Recommended Performance Standards for AMSA

Grade	Performance Standards			Impact Data		
	Level 2	Level 3	Level 4	Level 2	Level 3	Level 4
5	270	300	343	79%	42%	13%
8	264	300	334	76%	45%	20%
11	265	300	333	77%	37%	21%

## 9. REPORTING AND INTERPRETING SCORES

The Centralized Reporting System (CRS) generates a set of online score reports that includes information describing student performance for parents, educators, students, and other stakeholders. The online score reports are generally produced immediately after students complete the tests. Since 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 it to improve student learning.

In addition to individual students' score reports, the CRS also produces aggregate score reports by class, school, and state. The timely accessibility of aggregate score reports can help users to monitor students' performance in each grade by subject area and evaluate the effectiveness of instructional strategies. It can also inform the adoption of strategies to improve student learning and teaching and inform professional development for educators and help educators make curriculum decisions for the state over time.

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 students have performed on the assessments. It is an online tool that provides educators and other stakeholders with timely, relevant score reports. The CRS for the AMSA has been designed with stakeholders who are not technical measurement experts in mind in order to make score reports easy to read and understand. This is achieved by using simple language so that users can quickly understand assessment results and make inferences about student 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 score reporting design. This design strategy allows readers to compare similar elements and to avoid comparing dissimilar elements.

Once authorized users log in to the CRS, the dashboard shows overall test results for all tests that the students have taken grouped by test family (e.g., grade 5 science). Once the user clicks on the test family that he or she wants to explore, it will take him or her to a detailed dashboard. Additionally, when authorized state-level users log in to the CRS and select "State View," the system 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 29 summarizes the types of online score reports available at the aggregate level and the individual student level. Detailed information about the online score reports and instructions on how to navigate the online score reporting system can be found in the *Reporting System User Guide*, located via a help button with the CRS.

Table 29. Types of Online Score Reports by Level of Aggregation

Level of Aggregation	Types of Online Score Reports
State District School Teacher Roster	<ul style="list-style-type: none"> <li>• Number of students tested and percentage of proficient students (for overall students and by subgroup)</li> <li>• Average scale score (for overall students and by subgroup)</li> <li>• Percentage of students at each performance level</li> <li>• Participation rate (for overall students)<sup>1</sup></li> <li>• On-demand student roster report</li> </ul>
Student	<ul style="list-style-type: none"> <li>• Total scale score and standard error of measurement</li> <li>• Performance level for overall score with Performance Level Descriptors</li> <li>• Average scale scores for individual schools, district, and states</li> </ul>

<sup>1</sup> Participation rate reports are provided at the state, district, and school level.

Aggregate score reports at a selected aggregate level are provided for overall students and by subgroup. Users can see student assessment results by any of the subgroups. Table 30 presents the types of subgroups and subgroup categories provided in the CRS.

Table 30. Types of Subgroups

Subgroup	Subgroup Category
Gender	Male Female
Section 504 Plan Status	Yes No
IDEA Indicator	Yes No
Limited English Proficiency Status	Yes No
Economic Disadvantaged Status	Disadvantaged Not Disadvantaged
Race/Ethnicity	American Indian or Alaska Native Asian Black or African American Hispanic or Latino Native Hawaiian or Pacific Islander White Two or More Races
Enrolled Grade	KG, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12

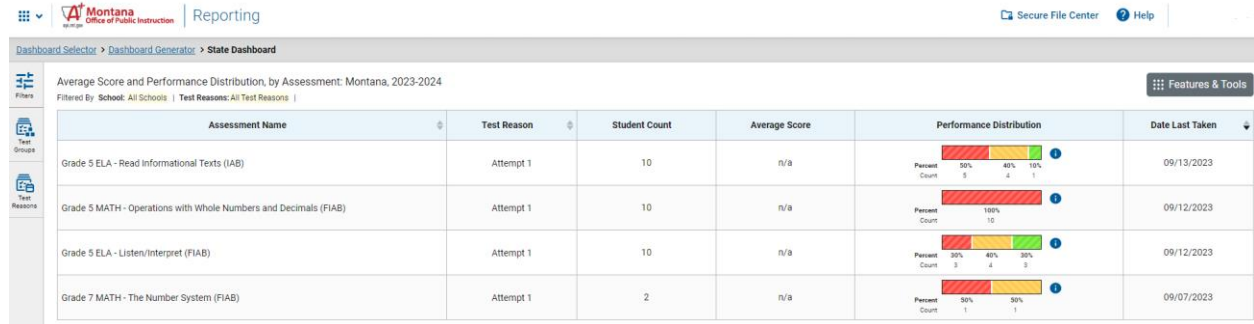
## Centralized Reporting System

### 9.1.2.1 Dashboard

The first page users see when they log in to the CRS contains summaries of student performance by test family (e.g., Summative Science Alt Grade 5). District personnel see district summaries, school personnel see school summaries, and teachers see summaries of their students. State personnel and district area personnel need to select a district in order to view the aggregate results.

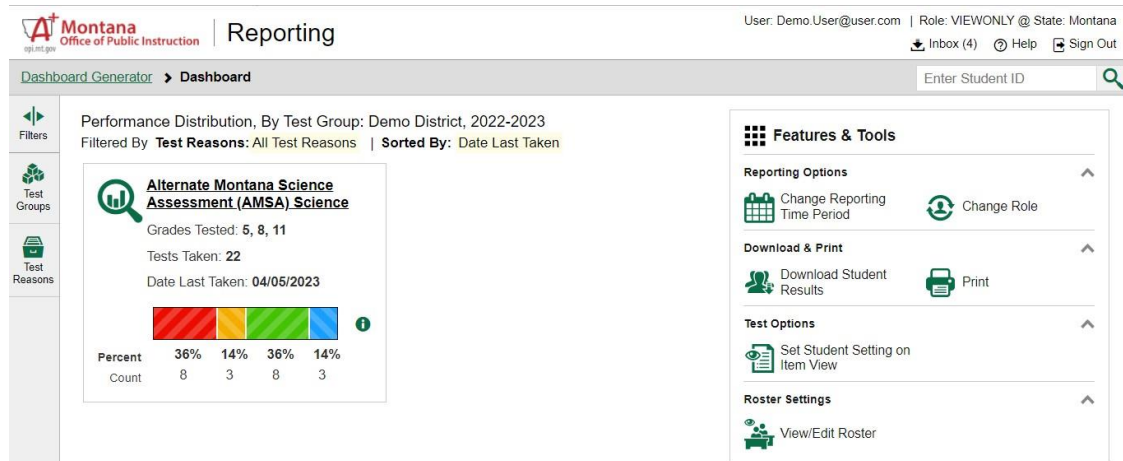
The dashboard summarizes students’ performance by test family, including (1) the number of students tested, (2) the grades of the students who have tested, and (3) the percentage and counts of students at each performance level. Exhibit 1 presents a sampled dashboard pages at the state level.

Exhibit 1. Dashboard: State Level



Educators can click on the subject group to view individual test results for the selected test group. Once the user clicks on the test family that he or she wants to explore further, the detailed dashboard page will appear. The detailed dashboard summarizes students’ performance by test, including (1) the number of students tested, (2) average score and standard error of the means, and (3) the percentage and counts of students at each performance level. Exhibit 2 presents a sample detailed dashboard page for AMSA at the district level.

Exhibit 2. Dashboard: District Level



9.1.2.2 Subject Detail Page

Detailed summaries of student performance for each grade in a subject area for a selected aggregate level are presented when users select a specific assessment name. On each aggregate report, the summary

report presents the summary results for the selected aggregate unit and the summary results for the state and the aggregate unit above the selected aggregate. For example, if a school is selected, the summary results of the school’s state and district are provided above the school summary results, as well, so that school performance can be compared with the aggregate levels.

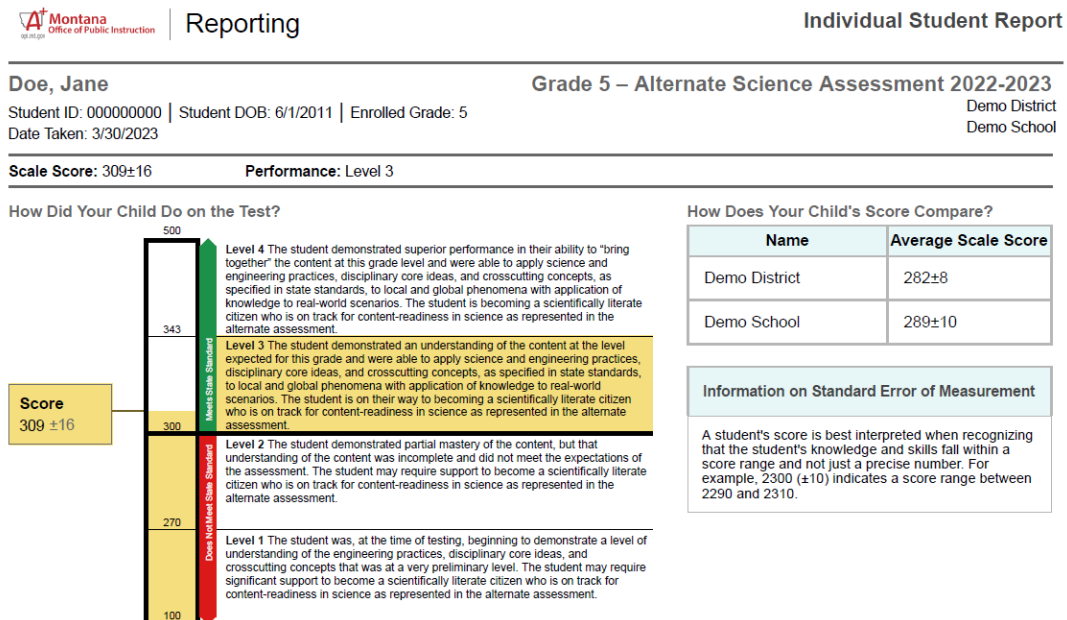
The aggregated subject summary report provides the summaries on a specific subject area, including (1) the number of students tested, (2) the average scale score and standard error associated with the average scale score, (3) the percentage of proficient students, and (4) the percentage and counts of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

### 9.1.2.3 Student Detail Page

When a student completes a test, an online score report appears in the individual student report (ISR) in the CRS. The ISR shows individual student performance on the test. In each subject area, the ISR provides (1) the scale score and standard error of measurement (SEM); (2) performance level for overall test; and (3) average scale scores for the student’s state, district, and school.

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

Exhibit 3. Student Detail Page for Science



## 9.2 INTERPRETATION OF REPORTED SCORES

A student’s performance on a test is reported with a scale score and a performance level for the overall test. Students’ scores and performance levels are summarized at the aggregate levels. The next section describes 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 students’ knowledge and skills. The scale score is the transformed score from a theta score estimated based on mathematical models. Low scale scores can be interpreted to mean that the student does not possess sufficient knowledge and skills measured by the test. Conversely, high scale scores can be interpreted to mean that the student has proficient knowledge and skills measured by the test. Scale scores can be used to measure student growth across school years. Interpretation of scale scores is more meaningful when the scale scores are used along with performance levels and Performance-Level Descriptors.

### 9.2.2 Standard Error of Measurement

A scale score (observed score on any test) is an estimate of the true score. If a student takes a similar test multiple times, the resulting scale score would vary across administrations, being sometimes a little higher, a little lower, or the same. The standard error of measurement (SEM) represents the precision of the scale score, or the range in which the student would likely score if a similar test were administered multiple times. When interpreting scale scores, it is recommended to consider the range of scale scores incorporating the SEM of the scale score.

The “±” next to the student’s scale score provides information about the certainty, or confidence, of the score’s interpretation. The boundaries of the score band are one SEM above and below the student’s observed scale score, representing a range of score values that is likely to contain the true score. For example, “312 ± 18” indicates that if a student were tested again, he or she would likely receive a score between 294 and 330. 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 AMSA, scale scores are mapped into four performance levels (i.e., Level 1, Level 2, Level 3, Level 4) using three performance standards (i.e., cut scores). PLDs are a description of the 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.

### 9.2.4 Aggregated Score

Student 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. Given that student scale scores are estimates, the aggregated scale scores are also estimates and are subject to measures of uncertainty. In addition to the aggregated scale scores, the percentage of students in each

performance level for the overall test is 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 provide information about individual students' performance on the test. Overall, assessment results tell what students know and are able to do in certain subject areas.

Assessment results for student performance on the test can be used to help teachers or schools make decisions on how to support student learning. Aggregate score reports at the teacher and school level provide information regarding the strengths and weaknesses of their students and can be used to improve teaching and student learning. Furthermore, by narrowing down the student performance result by subgroup, teachers and schools can determine what strategies may need to be implemented to improve teaching and student learning, particularly for students from a disadvantaged subgroup.

In addition, assessment results can be used to compare student performance among different students and among different groups. Teachers can evaluate how their students performed compared with students in other schools, districts, and the state overall.

Although assessment results provide valuable information to understand student performance, these scores and reports should be used with caution. It is important to note that scale scores reported are estimates of true scores and, therefore, do not represent a precise measure of student performance. A student's scale score is associated with measurement error, and, thus, users need to consider measurement error when using student scores to make decisions about student performance. Moreover, although student scores may be used to help make important decisions about students' placement and retention, or teachers' instructional planning and implementation, the assessment results should not be used as the only source of information. Given that assessment results measured by a test provide limited information, other sources on student performance, such as classroom assessment and teacher evaluation, should be considered when making decisions about student learning.

## 10. QUALITY CONTROL PROCEDURES

Quality assurance (QA) procedures are enforced through 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, such as the test blueprint specification, slopes and intercepts for theta-to-scale score transformation, cut scores, and the item information (i.e., answer keys, item attributes, item parameters, and passage information). The accuracy of the information in the configuration file is independently checked and confirmed numerous times by multiple staff members before the testing window opens.

To verify the accuracy of the scoring engine, we use simulated test administrations. The simulator generates a sample of students with an ability distribution that matches that of the population. 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 Alternate Montana Science Assessment (AMSA). The purpose of the simulations is to configure the algorithm to optimize item selection to meet blueprint specifications as well as to 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 in all of them. For the AMSA, there are two commonly used layouts: one has the stimulus and item response options/response area displayed side by side, where stimulus and response options have independent scroll bars; the other has the item stem and responses on the full screen.

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

Platform review is conducted by a team. The team leader projects the item as it was web approved in the Item Tracking System (ITS), and team members, each using a different platform, look at the same item to confirm that it 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 where they are subject to user acceptance testing (UAT). UAT of the TDS serves as both a software evaluation and a content approval role. The UAT period provides Montana’s Office of Public Instruction (OPI) with an opportunity 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. QA conducts a series of data integrity checks, ensuring, for example, that the record for each test contains information for each item, keys for multiple-choice items, score points for each item, total number of field-test items and operational items, and that the test record contains no data from items that have been invalidated.

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

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

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

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

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

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

Table 31 presents an overview of the QA reports.

Table 31. Overview of Quality Assurance Reports

<b>QA Reports</b>	<b>Purpose</b>	<b>Rationale</b>
Item Statistics	To confirm whether items work as expected	Early detection of errors (key errors for selected-response 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/passages)	Early detection of any oversight in the blueprint specification

### 10.3.1 Score Report Quality Check

#### Online Report Quality Assurance

Scores for online assessments are assigned by automated systems in real time. 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 of the QA system’s validation checks.

## 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. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage, 1994.
- Egan, K. L., Schneider, C., Ferrara, S. (2012). Performance level descriptors: History, practice, and a proposed framework. In Cizek, G. J. (Ed.), *Setting performance standards* (2nd edition). New York: Routledge.
- 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., and Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32*(2), 179–197.
- Livingston, S. A., and 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* (pp. 249–281). Mahwah, NJ: Lawrence Erlbaum Associates.
- 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>.