South Dakota Science Alternate Assessment (SDSAA)

2024–2025 Technical Report

Science Grades 5, 8, and 11

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

October 2025

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EXHIBITS

PREFACE

This report provides a technical summary of the 2024–2025 South Dakota Science Alternate Assessment (SDSAA) administered in grades 5, 8, and 11. The purpose of this technical report is to document evidence that supports the claims made for how SDSAA test scores can be interpreted. The report includes 12 chapters that discuss all the evidence accrued about the technical quality of South Dakota's testing system. This report is based on South Dakota operational test data for the alternate assessment, covering 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 [USDE], 2018).

Chapter 1, South Dakota Science Alternate Assessment, provides an overview of the SDSAA, the purposes and intended uses of SDSAA scores, the testing population, and the content standards. Chapter 2, Test Design and Development, describes the SDSAA tests, content specifications in test blueprints, and test assembly. Chapter 3, Item Development, describes the item-development process; specifically, the sequence of reviews that each item must pass through before being eligible for the SDSAA test administration. Chapter 4, Summary of Field-Test Item Analysis in Spring 2025, summarizes the field-test item analysis results and data review results from the spring 2025 test administration.

Chapter 5, Test Administration, documents the test administration procedures, including proctor training, the administration manual, accommodations, and the prevention of disruptions in the Test Delivery System (TDS). Chapter 6, Scoring, describes the scoring procedures used in producing scale scores and achievement levels. Chapter 7, Summary of Spring 2025 Operational Test Administration, summarizes the results of the spring 2025 SDSAA test administration, including the test-taking student population, their performance on the assessment, and the time spent taking the assessment.

Chapter 8, Validity, provides validity evidence on test contents, cognitive lab, and internal consistency. Chapter 9, Reliability, provides evidence on the reliability of the SDSAA scores, including marginal reliability, standard errors of measurement, and classification accuracy and consistency of achievement standards. Chapter 10, Achievement Standards, describes the procedure to set achievement standards. Chapter 11, Reporting and Interpreting Scores, provides a description of the score reporting system and the interpretation of test scores. Chapter 12, Quality Control Procedures, provides an overview of the quality assurance (QA) processes that are used to ensure that all test development, administration, scoring, and reporting activities are conducted with fidelity to the developed procedures.

1. SOUTH DAKOTA SCIENCE ALTERNATE ASSESSMENT

1.1 OVERVIEW

The South Dakota Science Alternate Assessment (SDSAA) is composed of tests that are based on the Core Content Connectors (CCCs) and is designed for students with the most significant cognitive disabilities. The purposes of the SDSAA are to (1) 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) ensure that South Dakota's statewide assessments are accessible to all students with disabilities; and (3) 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 SDSAA 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 South Dakota Department of Education (SDDOE) began the transition to a new computer-adaptive test (CAT) for the science alternate assessment for students with significant cognitive disabilities. The new science assessment was designed to assess students in grades 5, 8, and 11. Each student was administered a 40-item operational test with 10 embedded field-test (EFT) items. In spring 2021, interim achievement standards that adopted a statistical linking method were established to report SDSAA achievement levels. In spring 2022, a formal standard-setting workshop was conducted to establish the achievement standards used for score reporting. In fall 2022, an independent alignment study was conducted on the operational bank. The contractor for the alignment study compared the alignment of the items to the CCCs. The final analysis from that alignment study concluded that many items did not have a strong on-grade alignment to the CCCs. It became clear that it was necessary to remove the extra standard "layer" of Essence Statements that originally served as a gateway between the CCCs and the Achievement-Level Descriptors (ALDs). It was also deemed important to 1) make sure the ALDs had a strong alignment with the CCCs, and to 2) communicate clearly to all stakeholders that the ALDs would serve as the guide for item writing and alignment going forward. Thus, during academic year 2022–2023, two alignment workshops were conducted with South Dakota educators to re-examine and re-adjust, if necessary, the content alignment of all operational items after the educators revised and clarified the ALDs. These workshops were conducted in January and June 2023.

1.2 PURPOSES, INTERPRETATIONS, AND INTENDED USES OF THE SDSAA

The purposes, interpretations, and intended uses of the SDSAA serve as the foundation for test design and development. They play a crucial role in the validation process, as any statements about validity are tied to specific interpretations and uses.

Purposes and Intended Uses

The purposes and intended uses of the SDSAA are to measure students' academic performance and students' progress in meeting the state alternate academic achievement standards in science.

To fulfill its intended purposes, the SDSAA provides an overall scale score and an associated achievement level for each test. These achievement levels are determined based on the achievement standards established through a formal standard-setting process.

At the individual student level, the SDSAA test score can be used to estimate a student's academic performance; the associated achievement level, together with the ALDs, can indicate the knowledge and skills the student has attained in the assessed content area by the end of the academic year. Individual student scores and achievement levels can be compared across students who take the same test. Additionally, scores can also be aggregated to estimate the average performance of specific groups or to compare the average performance between different groups, such as by school, district, or gender.

Intended Users

Primary intended users of the SDSAA include the following:

- Students and families can use the results to stay informed about the student's learning progress in school.
- Teachers and educators can use the results to guide in-class instruction and identify students who need additional support.
- Educational agencies, organizations, and governments can use the test data and results to monitor educational improvement and make necessary changes to educational opportunities for all students.

1.3 ALTERNATE ASSESSMENT IDENTIFICATION

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 Student Participation Criteria from the *Summative Science Alternate Test Administration Manual* at https://sd.portal.cambiumast.com/resource-item/en/summative-science-alternate-assessment-test-administration-manual-tam. South Dakota's guidance documents for participation in the SDSAA can be found at https://doe.sd.gov/assessment/alternate.aspx. The participation guidelines are summarized in Table 1. All three criteria must be met for a student to be identified for participation inthe SDSAA. If one or more are not met, the student should take the regular assessments.

Table 1. Participation Criteria

Participation Criteria	Participation Criteria Descriptors
1. Student has a significant cognitive disability.	Review of student records indicates a disability or multiple disabilities that significantly impact intellectual functioning and adaptive behavior.* **Adaptive behavior is defined as essential for someone to live independently and to function safely in daily life.*
2. Student requires extensive instruction and support to acquire and maintain skills.	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.
3. Student learns through alternate academic achievement standards (AAAS).	Goals and instructions listed in the IEP for this student follow CCCs and address knowledge and skills that are appropriate and challenging for this student.

1.4 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) 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 SDSAA is aligned with the CCCs, which are linked to the 2015 South Dakota Science Standards.

The CCCs in science take the concepts from the South Dakota Science Standards and break them down to pinpoint the prevalent ideas that are accessible for students with significant cognitive disabilities. The CCCs address Science and Engineering Practices (SEPs), Disciplinary Core Ideas (DCIs), and Crosscutting Concepts from the standards.

To further break down the main ideas in the CCCs, SDDOE and Cambium Assessment, Inc. (CAI) staff prioritized the content and skills that were deemed most critical in the development of successful postsecondary outcomes for students with significant cognitive disabilities, creating Policy Achievement-Level Descriptors (ALDs) and Range ALDs.

Policy ALDs are used to provide a broad overview of the student's level of understanding of the science standards. These have been developed at four levels and were approved by SDDOE before the ALD review meeting. The levels are as follows:

- **Level 4 (Exceeded):** A student whose achievement level is Exceeded demonstrates a level of understanding that includes the ability to "bring together" the DCIs and/or SEPs and/or Crosscutting Concepts associated with a performance expectation.

- **Level 3 (Met):** A student whose achievement level is Met demonstrates an understanding of the DCIs and/or SEPs and/or Crosscutting Concepts within the performance expectations at the conceptual level described in the CCCs.

- **Level 2 (Nearly Met):** A student whose achievement level is Nearly Met demonstrates some understanding of the content of the performance expectations, but that understanding is incomplete and does not yet meet the expectations found in the CCCs. This student's understanding is partial but emerging.

- **Level 1 (Not Met):** A student whose achievement level is Not Met demonstrates a level of understanding that is at a very preliminary level. This student's understanding is nonexistent or incomplete, and the student has difficulty meeting the expectations.

In addition to these Policy ALDs, the Range ALDs have been developed for each CCC, reflecting 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 CCCs to craft the Just Barely and Reporting ALDs. See Section 10.2 for details.

The Range ALDs were created by the CAI content team starting with a small set of ALDs written to a subset of the CCCs that were posted on the SDDOE website. CAI took these and matched them to the appropriate CCCs and created ALDs for the remaining CCCs. CAI worked with SDDOE to finalize the wording for each of the remaining ALDs. All ALDs were brought to South Dakota educators before standard setting during an ALD review meeting for their discussion. Any edits suggested by the committee were reviewed by SDDOE, which made the final decision about which suggestions to incorporate before standard setting..

Furthermore, in January 2023, a second subset of South Dakota educators convened to review all the ALDs before the alignment workshops. The committee reviewed the ALDs to make sure they: (1) aligned with the CCCs, (2) described what students could do, (3) defined differences in content across the achievement levels, (4) described the contextual or scaffolding characteristics needed so a student could demonstrate the skill, (5) demonstrated a clear increase in cognitive demand across achievement levels, and (6) provided a mental picture of increases in skill across achievement levels. The committee suggested some edits, and SDDOE made the final determination on the implementation of those edits. These finalized ALDs serve as the foundation for the development of SDSAA items.

Items in the item bank align with the CCCs and ALDs, hitting a breadth of different levels of complexity 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.

2. TEST DESIGN AND DEVELOPMENT

2.1 TEST DESCRIPTIONS

The South Dakota Science Alternate Assessment (SDSAA) assesses science in grades 5, 8, and 11. The SDSAA has three tests in three grades. Each test comprises 40 operational items and 10 field-test items shared across states in the alternate assessment program (see Section 3.1 for details). The field-test items do not contribute to students' reported scores.

In the spring 2025 test administrations, the SDSAA was delivered as an online fixed-form test where each student took the same set of operational items in each grade. Students who were unable to fully access the online test had the option to receive an accommodation of paper response options.

2.2 TEST BLUEPRINTS

Content specifications for the SDSAA tests are aligned with the South Dakota Science content standards. Test blueprints outline the minimum and maximum number of operational items required from each domain (also known as strands) and from each substandard within those domains. Table 2 displays the blueprints at the domain level for each grade level of the SDSAA. Each full-length test consists of 40 operational items.

Table 2. SDSAA Test Blueprints

Grade	Domain	Minimum Required Items	Maximum Required Items
5	Earth and Space Science	12	15
	Life Science	12	15
	Physical Science	12	15
8	Earth and Space Science	12	15
	Life Science	13	15
	Physical Science	11	15
11	Earth and Space Science	12	15
	Life Science	12	15
	Physical Science	11	15

2.3 TEST ASSEMBLY

Each SDSAA test administered in the spring 2025 test administration was a fixed form and included three segments. The first segment included four fixed operational items, the second segment contained the remaining 36 operational items, the third segment contained 10 field-test items randomly selected from a larger field-test pool that was shared across multiple states (see Section 3.1, Memorandum of Understanding on Item-Sharing Initiative, for details).

The first segment is also known as the Early Stopping Rule (ESR) segment which was available for students who were non-responsive to the first four items on each test. Students and Proctors were required to follow the test administration guidelines put in place by the SDDOE Assessment Section. The ESR was used if the student had no consistent and observable mode of communication and was unable to respond to all of the

first four items in the test. If the student had a mode of communication and the ESR was used, the assessment was invalidated for misadministration.

If a student answers at least one item in the ESR segment, the test engine will proceed to the second segment**.** The four items in the ESR segment are selected to be relatively easy to encourage student engagement and participation, particularly for those who may struggle.

Since the fixed form was also used as an accommodation form for students who could not fully access the online tests, items with access limitations were not included.

3. ITEM DEVELOPMENT

3.1 MEMORANDUM OF UNDERSTANDING ON ITEM-SHARING INITIATIVE

The item development process for the alternate assessments is a collaborative effort among member states that have signed a Memorandum of Understanding (MOU) for item sharing in item development and field testing. Each MOU member state retains ownership of the items they develop, but these items are available for use by other MOU members. The number of items each state is responsible for developing is proportional to its alternate assessment population size. Given that the alternate assessment population in each state is small, the item-sharing initiative enables statistical and psychometric analysis based on combined data from all participating states. As a result, item parameter estimates are more stable compared to those derived from smaller sample sizes.

The MOU Alternate Assessment (MOU-Alt) was initiated in 2018 and originally signed by three states: Hawaii, South Carolina, and Wyoming, covering English language arts (ELA), mathematics, and science. In early 2019, Idaho and Vermont joined the MOU for ELA, mathematics, and science. Montana and South Dakota joined in 2020, but only for science. Vermont exited the MOU in 2022.

In the 2024–25 academic year, there are six MOU member states: Hawai‘i, Idaho, Montana, South Carolina, South Dakota, and Wyoming. South Dakota and Montana administered the assessment in science only, while the remaining four states tested in all three subjects. All member states participated in field-testing items. Psychometric analyses were conducted using the combined sample across all states. In addition to the items jointly developed by the MOU member states, each state may also develop items that are specifically aligned with its own content standards.

Following the spring 2025 test administration, South Carolina and Montana withdrew from the MOU. As a result, field-test items owned by these states are no longer eligible for operational use in the remaining member states.

Each state in the MOU follows a similar process for developing and reviewing their items in collaboration with CAI. Items are developed by each state to fulfill their agreed-upon contribution to the MOU each school year. CAI requires Department of Education (DOE) staff in each participating state to review the items contributed by their partner MOU states for field testing each school year and provide a state-specific alignment to their own state’s content standards at the shared grade level for each item. Following yearly field testing and data review, DOE staff in each participating state make a final determination on whether shared items are accepted for operational use by confirming the state-specific content alignment for each item.

3.2 ITEM TYPES

There are multiple-choice (MC) items and multi-select (MS) items in the MOU item banks. The MC items have two to four 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 1 point; otherwise, the student receives 0 points. For MS items, if a student selects two keys, he or she earns 2 points; if the student selects one key, he or she earns 1 point; otherwise, the student earns 0 points. Each item measures a specific content standard. The final item difficulties are determined through field testing.

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 (TEIs) should be developed for the MOU shared field-test

items. The item types included MS, equation editor, table match, and animation. Neither equation editor nor table match proved to be a successful item type for this population of students, and therefore, states will not develop anymore of these item formats. MS items were successful for middle school and high school students with significant cognitive disabilities who perform strongly on alternate assessments. These items will continue to be developed for this segment of the alternate assessment population.. Animations were successful across all states and grade levels. This item format is present in the current item pool.

3.3 DEVELOPMENT OF CROSSWALK OF STATE ALTERNATE CONTENT STANDARDS

A crosswalk across individual state alternate academic achievement standards was completed for the first year of the MOU-Alt shared field-test item development. This crosswalk has been updated as more states joined the MOU since 2018. Content of the standards from each of the MOU states was reviewed and compared by special education and content experts at Cambium Assessment, Inc. (CAI) to determine which standards are on-grade and overlapped across states. For example, CAI looked at all grade 5 science standards for each MOU state and determined which standards contained common content. If standard A in the first state contained the same content as standard B in the next state, and standard C in the third state, then the three standards in the three states were common. When aligning items to standards in each state, with this crosswalk available, CAI knew instantly which standards items should be aligned to. The opposite is true, as well. Some standards did not have similar content to other states' on-grade standards, so items aligned to those standards were not aligned to other states.

The crosswalk was created by senior CAI test development specialists and reviewed by the state departments of education. The crosswalk was based on each state's blueprint and included the common core or Next Generation Science Standards (NGSS) and the general education and alternate academic achievement standards for each state. Each state has a unique set of alternate academic achievement standards as follows:

- Hawaii Essence Statements and Performance-Level Descriptors (PLDs)
- Idaho Extended Content Standards Core Content Connectors
- Montana Alternate Academic Achievement Standards and PLDs
- South Carolina Alternate Academic Achievement Standards and PLDs
- South Dakota Content Standards and Core Content Connectors
- Wyoming Extended Standards and Instructional Achievement-Level Descriptors (ALDs)

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

Once all individual state standards were aligned across all participating states, item development plans (IDPs) were created for each state. These IDPs were based on identified areas where additional items were needed to ensure that all the MOU-Alt standards aligned on the crosswalk were addressed in the item-sharing pool. Items for each state-specific standard that were not aligned to the MOU-Alt crosswalk standards were created to meet the state's test blueprint, if the state decided to create additional items for their state. These IDPs guided the development of 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 IDP is created. As new states joined the MOU-Alt agreement, or when states changed their standards, the individual state standards were added to the crosswalk so that items from the state could be aligned across all states.

3.4 Development of Item Specifications

The development of item specifications was informed by the crosswalk of state alternate academic achievement standards. The item specifications are for the MOU, instead of for individual states. For each common standard in the crosswalk, CAI examined the states' content extensions and PLD or ALD documents to identify which extensions were aligned with that common standard. Each item specification included the General Education standard, followed by the state-specific alternate content standards that aligned with the General Education standard. Item specifications also included complexity statements and task demands. The language of the complexity statements and task demands were derived from each state's content standards, where applicable, and synthesized to drive items aligned to multiple states. Once completed, the item specifications were sent to each state for review to confirm alignment and overall approach.

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

The analyses were then used to create sample items at each of the proficiency levels. Each sample item was annotated with information regarding its proficiency level, as well as which sample items address the Science and Engineering Practice (SEPs) and Crosscutting Concepts for the associated content standard.

3.5 Item Development Process

Items are developed by each of the states that joined the shared item development agreement. In each state, item development for each year begins in the spring. Before item development, item writers are trained on aspects of items that are unique to students with significant cognitive disabilities. Items are written by professional item writers with a background in education and expertise in the assigned content area and alternate assessments. A group of senior test-development specialists monitor and support the item development activities.

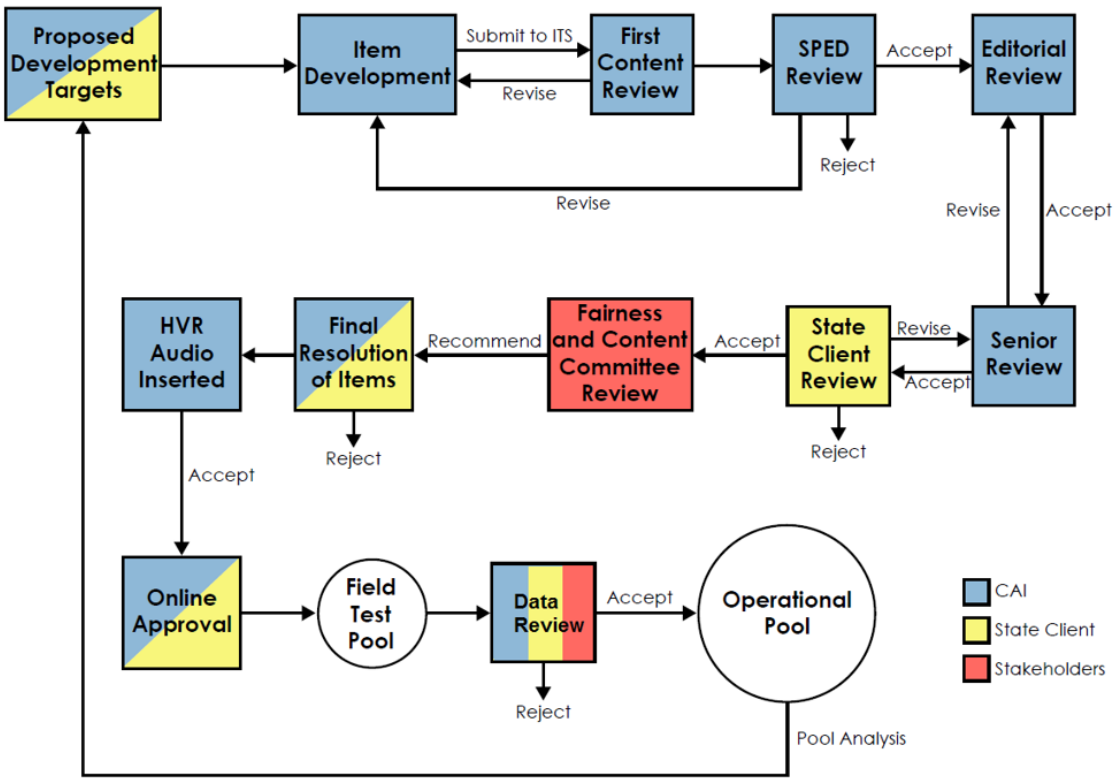
Items are written by CAI content staff, in compliance with the item specifications and style guide documents to ensure that items meet the expected alignment, complexity, and style criteria. The item specifications and style guide documents are created by CAI and reviewed and approved by the state departments of education. The item specifications are for the MOU, instead of for individual states. If a particular standard is under one state only, that standard is not included in the MOU item specifications. Rather, the state creates separate field-test slots for items associated with state-specific standards.

Item development begins with establishing CAI's 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 items are initially developed, they undergo a group review that includes content and senior reviewers, followed by an individual content review, where edits are made based on group reviews, and then a special education review. After the items are reviewed by the special education reviewer, they go through an editorial review. After editorial review, the items go back through a senior review, which is the

last review step at CAI before the items are sent to each state for client review. At this step, the client may accept the items, recommend edits, or reject the items.

After client comments are resolved, all accepted items are then submitted to a stakeholder Content and Fairness Committee for review. At the same time the Content and Fairness Committee reviews the items, the other members of the MOU-Alt also review the items and provide feedback. After the Content and Fairness Committee makes its recommendations, the state and CAI convene a resolution meeting at which all comments from the Content and Fairness Committee and the other MOU-Alt states are reviewed. The state approves final edits to the items based on the Content and Fairness Committee and other state comments. Items then go through a final edit resolution. Lastly, CAI verifies that the items will appear on the test as expected through the platform review process. Figure 1 shows the full item development process.

Figure 1. Alternate Assessment Item Development Process



CAI Review

Items are reviewed at CAI at various levels.

- CAI Internal Group Review: Before making any changes to draft items, content and senior reviewers meet to discuss items and determine revisions to content, alignment, and style. Reviewers use the item specifications and a style guide to make sure the items fit all guidelines.

- CAI Internal Preliminary Review: Following group review, a preliminary review is conducted by a member of CAI's content team assigned to the alternate assessment. Items are revised to

eliminate initial errors, meet content standards, and satisfy internal style and clarity expectations, as agreed on in the group review.

- CAI Internal Content Review: A second content review occurs after the preliminary review to further ensure that changes based on the group review are implemented, and to revise items to address any errors and issues on content, alignment, clarity, and accessibility.

- 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 broad spectrum of cognitive and physical disabilities. When appropriate, the special education expert designates items as "Access Limited," meaning that a task is inappropriate to administer to students with a specific physical disability (e.g., blindness). If revisions are required, the special education reviewer will send items back to the content reviewer to implement changes.

- Editorial Review: After the special education reviewer approves 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 previously requested edits are in place.

- CAI Batch Review: This is the last step in the CAI internal review process and is designed as a final quality control check to ensure that the items are ready for state review.

State Review

At this level, items are compared to the CCCs, reviewed against the ALDs at all difficulty levels, and compared to the blueprint. Items are further reviewed to ensure that they align with the support guides and item specifications 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 South Dakota Department of Education (SDDOE) 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

In each state, items owned and accepted by the state are prepared for review by a statewide Content and Fairness Committee. The Content and Fairness Committee is composed of stakeholders from around the state with teaching experience in grades K–12 and experience working with students with disabilities. Additional statewide stakeholders with expertise in specific disability categories and multicultural or

foreign language expertise are invited to participate in the committee meetings. The review committee includes special educators, general educators, vision and hearing specialists, school principals, special education directors, and university professors with expertise in special education. The review committee represents a diversity of gender, ethnicity, disability, race, and cultural subgroups across the state.

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

Committee stakeholders review the items and provide feedback to ensure that all accepted items are correct, meet bias and sensitivity guidelines, align with content standards, and abide by universal design principles. Most importantly, these educators ensure that this population of students can understand the language used in the items and that the included visuals and audio directions will aid and not distract students.

The common criteria used for item review are as follows:

- Content accuracy and clarity
- Alignment to the content specifications
- 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

3.6 FIELD TESTING

After going through various stages of reviews, items are moved into the field-test item pool to be field-tested the following spring during the operational testing window. For example, the items developed in 2023–24 were field-tested in spring 2025; the items developed during the academic year of 2024–25 will be field-tested in spring 2026. In spring 2025, the computer-adaptive test (CAT) was the primary test delivery mode in all MOU states except South Dakota where fixed-form tests were used. Field-test items were embedded among operational items in CATs and appended at the end of the test in fixed-form tests. This approach yields item parameter estimates that capture all the contextual effects contributing to item difficulty in operational test administrations. Field testing in an operational setting is beneficial in the context of a pre-equating model and CATs 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 standalone field testing.

After the operational test administration, CAI psychometricians perform both classical item analysis and item response theory (IRT) analysis for all field-test items. Items are flagged for review based on predetermined statistical criteria. Details of the psychometric analysis and flagging rules on field-test items are presented in Chapter 4, Summary of Field-Test Item Analysis in Spring 2025.

3.7 POST-FIELD-TESTING ITEM DATA REVIEW

Following the psychometric analysis, items are categorized into four groups for further action:

- Items with a sample size of fewer than 50 are archived for future re-field-testing.
- Items with negative biserial/polyserial correlations are rejected after an additional key verification from CAI.
- Items not flagged by the statistics and owned by other MOU member states are reviewed through a Roman Voting process by the educator committees.
- Items flagged by the statistics undergo an item data/content review (IDCR) process.

Roman Voting

The purpose of the Roman Voting process is to provide states and their educators with an additional opportunity to review items before they are used in future operational administrations. This process is carried out independently within each state. In South Dakota, the Content and Fairness Committee first votes on whether to move items into the operational item pool. If the committee votes Yes, the items are added to the operational item pool without future review. If the committee votes No, SDDOE makes the final decision on whether to include them in the operational item pool.

Item Data/Content Review

Items flagged by the statistics are reviewed in IDCR meetings involving all MOU states. The MOU-Alt data review committee consists of staff across MOU states, CAI content specialists, special education specialists, and psychometricians. Before IDCR, CAI psychometricians train reviewers on how flagged statistics can be used to identify potential content flaws in items. During IDCR, the committee evaluates whether flagged items contain features that might result in undesirable statistics. They then decide whether to reject the item completely, accept it with modifications for further field testing, or accept it as is. Additionally, content experts from each state ensure that items from other states are included only if they align with the state's standards.

The IDCR process has two phases.

1. **Individual State Review:** In this initial phase, state staff or educators from each state independently review the items and decide whether to accept or reject them. After all states complete their reviews, the decisions are summarized into four categories:

   - Items that are accepted by all states
   - Items that are rejected by all states
   - Items that are rejected by the source state but accepted by at least one destination state
   - Items that are accepted by the source state but rejected by at least one destination state

   Items in the first category are added to the item pools of all states, while those in the second category are rejected from all state item pools. Items in the third or fourth categories, where there is disagreement among states, proceed to the second phase: group review.

2. **Group Review:** In this phase, all states participate in group IDCR meetings where they share rationales for their decisions. After discussing and considering the perspectives of other states, states can revise their initial decisions from the individual state review.

Upon completion of the Roman Voting and IDCR process, all field-test items accepted by each state will be added to their operational item pool, ready for administration in the following year. Item data review

results in the spring 2025 administration are presented in Chapter 4, Summary of Field-Test Item Analysis in Spring 2025.

4. SUMMARY OF FIELD-TEST ITEM ANALYSIS IN SPRING 2025

The SDSAA spring 2025 field-test item pool included Memorandum of Understanding (MOU) items. Table 3 provides the number of MOU items administered in South Dakota by source state. South Dakota Department of Education (SDDOE) approved a total of 108 items for field testing in spring 2025. With South Carolina and Montana's withdrawal from the MOU, 44 items are eligible for the potential SDSAA operational pool.

Table 3. Number of Field-Test Items administered in Spring 2025

Subject	Grade	MOU Items by Source State						MOU Items (All States)	MOU Items (Exclude SC and MT)
		HI	ID	MT	SC	SD	WY		
Science	5	3	4	3	26	2	3	41	12
	8	2	10	2	22	2	1	39	15
	11	4	2	2	9	5	6	28	17
Total		**9**	**16**	**7**	**57**	**9**	**10**	**108**	**44**

4.1 FIELD-TEST ITEM ANALYSIS

After the close of the spring testing window, Cambium Assessment, Inc. (CAI) psychometrics staff analyzed field-test data based on combined data from all MOU states, to prepare for item data review meetings and to promote high-quality test items to operational item pools. Analysis of field-test items included the following:

- **Classical item analysis**, used to evaluate the relationship of each item to the overall scale and assess the quality of the distractors
- **Item response theory (IRT) analysis**, used to assess how well items fit the measurement model and provide the statistical foundation for constructing operational forms and test scoring and reporting
- **Differential item functioning (DIF) analysis**, used to identify items that may exhibit bias across subgroups

4.1.1 Classical Item Analyses

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

Items that are extremely difficult or easy are flagged for review but not necessarily rejected if they align with the test and content specifications. For dichotomous items, the proportion of test takers in the sample selecting the correct answer (*p*-value) is computed as well as those selecting the incorrect responses. For polytomous items with 0–2 score points, item difficulty is calculated both as the item's mean score and the average proportion correct (analogous to *p*-value and indicating the ratio of an item's mean score divided by the maximum score point possible). Items are flagged for review if the *p*-value or average proportion correct is less than 1 divided by the number of response options or greater than 0.95.

The item discrimination index indicates the extent to which each item differentiates between those test takers who possess the skills being measured and those who do not. In general, the higher the value, the better the item could differentiate between high- and low-achieving students. The discrimination index is calculated as the correlation between the item score and the student's IRT-based ability estimate. Items are flagged for subsequent reviews if the correlation for the keyed (correct) response is less than 0.20. For polytomous items, the mean total number correct score is computed for students scored within each possible score category; 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 dichotomously scored multiple-choice items is used to identify items with marginal distractors or ambiguous correct responses. The discrimination value of the correct response should be substantial and positive, and the discrimination values for distractors should be lower and, generally, negative. The biserial correlation for distractors is the correlation between the item score, treating the target distractor as the correct response, and the student's IRT-based ability estimate, restricting the analysis to those students selecting either the target distractor or the keyed response. Items are flagged for subsequent reviews if the biserial correlation for the distractor response is greater than 0.05.

The flagging criteria based on classical item analysis statistics are summarized in Table 4.

Table 4. Thresholds for Flagging in Classical Item Analysis

Analysis Type	Flagging Criteria
Item Difficulty	*p*-value (for dichotomous items) or average proportion correct (for polytomous items) is < 1/number of response options or > 0.95.
Item Discrimination	Biserial or polyserial correlation for the correct response is < 0.20.
Mean Score for Two-Point Items	Mean total score for a lower score point > Mean total score for a higher score point
Distractor Analysis	Biserial correlation for any distractor response is > 0.05.

4.1.2 Item Response Theory Analysis

The Item Response Model

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 θ 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 Masters' (1982) partial credit model (PCM). Under Masters' PCM, the probability of getting a score of x_i on item i given ability θ can be written as

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

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

Item Calibration

Calibration is the process by which we estimate the statistical relationship between item responses and the underlying trait being measured. Winsteps is used to estimate the Rasch and Masters' (PCM) item parameters for the MOU-Alt. Winsteps, provided by Mesa Press, is publicly available software that utilizes a joint maximum likelihood estimation (JMLE) approach. This method simultaneously estimates both person and item parameters.

The Winsteps output, which includes item statistics, is reviewed. Item fit is evaluated via the mean square Infit and mean square Outfit statistics, which are based on weighted and unweighted standardized residuals for each item response. These residual statistics reflect the discrepancy between the observed item scores and predicted item scores according to the IRT model. The expected value for both fit statistics is 1. Values substantially greater than 1 indicate model underfit, while values substantially less than 1 indicate model overfit (Linacre, 2004). Items are flagged if Infit or Outfit values are less than 0.5 or greater than 2.0.

Embedding randomly selected field-test items among operational items in computer-adaptive tests (CATs) results in a sparse data matrix. In this matrix, both operational and field-test items are calibrated concurrently for each grade and subject, with the parameter estimates of the operational items fixed. The operational items were previously calibrated and scaled to the existing MOU-Alt scale during the years they were used as field-test items. Consequently, the field-test item parameter estimates are also on the MOU-Alt scale. Completed records from all MOU states are included in the IRT analysis, with items not presented being treated as not administered.

4.1.3 Differential Item Functioning Analysis

Differential item functioning (DIF) refers to items that appear to function differently across identifiable groups, typically across different demographic groups. Identifying DIF is important because it can indicate that an item contains a cultural or other bias. Not all items that exhibit DIF are biased; some characteristics of the educational system may also lead to DIF. For example, if schools in low-income areas are less likely to offer geometry classes, students at those schools might perform more poorly on geometry items than

would be expected, given their proficiency on other types of items. In this example, it is not the item that exhibits bias but the curriculum. However, because DIF can indicate bias, all field-tested items were evaluated for DIF. Items exhibiting DIF were flagged for further examination by CAI and the MOU states.

CAI conducts DIF analysis on all field-tested items to detect potential item bias among the following group comparisons:

- Female vs. Male
- African American vs. White
- Hispanic or Latino vs. White

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 (EFT) items, design effects in student samples are rarely reduced to the level of a simple random sample.

The ability distribution is divided into 10 intervals to compute the generalized Mantel–Haenszel chi-square ($GMH_{\chi 2}$) DIF statistics. For dichotomous items, the analysis program computes the $GMH\chi^2$ DIF statistic, the log-odds ratio, and the *MH*-delta (Δ_{MH}); for the polytomous items, the program computes the $GMH\chi 2$ DIF statistic, the item score standard deviation (σ), and the standardized mean difference (*SMD*).

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 5. 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 fell 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. DIF classification rules are presented in Table 5. 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).

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

Table 5. DIF Classification Rules

Dichotomous Items	
Category	*Rule*
C	$GMH\chi^2$ is significant at .05 and $\lvert\Delta_{MH}\rvert > 1.5$.
B	$GMH\chi^2$ is significant at .05 and $1 < \lvert\Delta_{MH}\rvert \leq 1.5$.
A	$GMH\chi^2$ is not significant at .05 or $\lvert\Delta_{MH}\rvert \leq 1$.
Polytomous Items	
Category	*Rule*
C	$GMH\chi^2$ is significant at .05 and $\frac{\lvert SMD\rvert}{\sigma} > .25$.
B	$GMH\chi^2$ is significant at .05 and $.17 < \frac{\lvert SMD\rvert}{\sigma} \leq .25$.
A	$GMH\chi^2$is not significant at .05 or $\frac{\lvert SMD\rvert}{\sigma} \leq .17$.

4.2 Results of the Spring 2025 Field-Test Item Analysis

This section presents a summary of results from the classical item analysis, IRT analysis, and DIF analysis of items field-tested in South Dakota in spring 2025. Table 6 presents the average sample size and the sample size at various percentiles for the analysis. Table 7 provides summaries of item statistics. For each item statistic (e.g., *p*-values), the percentiles are computed across items administered in South Dakota. Table 8 provides the DIF analysis summary.

Table 6. Sample Size Distribution

Subject	**Grade**	**Total # of Items**	**Average Sample Size**	**Sample Size in Percentiles**								
				Min	**5th**	**10th**	**25th**	**50th**	**75th**	**90th**	**95th**	**Max**
Science	5	12	189	125	125	127	157	202	217	226	228	228
	8	15	176	151	151	158	171	182	185	188	189	189
	11	17	242	85	85	88	227	265	295	303	316	316
	Overall	44	205	85	95	127	173	193	252	295	299	316

Table 7. Summary of Item Analyses

Grade	Total # of Items	Statistics	Min	P10	P25	P50	P75	P90	Max
5	12	*p*-value	0.35	0.37	0.40	0.45	0.53	0.58	0.65
		Biserial/Polyserial	0.04	0.05	0.08	0.28	0.49	0.53	0.61
		Step Difficulty	-1.19	-0.80	-0.44	-0.20	0.12	0.25	0.32
		Infit	0.87	0.88	0.94	1.02	1.13	1.18	1.18
		Outfit	0.82	0.85	0.90	1.08	1.17	1.22	1.23
8	15	*p*-value	0.34	0.34	0.36	0.43	0.57	0.64	0.77
		Biserial/Polyserial	-0.02	0.02	0.15	0.26	0.50	0.52	0.64
		Step Difficulty	-1.75	-1.04	-0.72	-0.14	0.15	0.27	0.30
		Infit	0.83	0.89	0.97	1.01	1.08	1.15	1.16
		Outfit	0.76	0.89	0.94	1.00	1.10	1.21	1.23
11	17	*p*-value	0.28	0.34	0.36	0.56	0.61	0.72	0.78
		Biserial/Polyserial	-0.04	0.05	0.27	0.33	0.46	0.51	0.65
		Step Difficulty	-1.76	-1.45	-0.91	-0.66	0.28	0.46	0.63
		Infit	0.87	0.93	0.94	1.01	1.03	1.16	1.20
		Outfit	0.75	0.81	0.90	0.99	1.05	1.22	1.23

Table 8. Number of Items in Each DIF Classification Category

Subject	Female vs. Male							African American vs. White							Hispanic vs. White						
Grade	Total	+A	-A	+B	-B	+C	-C	Total	+A	-A	+B	-B	+C	-C	Total	+A	-A	+B	-B	+C	-C
Science																					
5	9	3	6					2	2												
8	14	7	6			1															
11	14	6	5		1		2	13	3	9				1							

4.3 ITEM DATA REVIEW RESULTS

Table 9 presents the item data review results in spring 2025. Out of the 44 science items field-tested in South Dakota, no item had negative biserials/polyserials that were rejected without further review. SDDOE and their Content and Fairness Committee reviewed the remaining items, rejecting those that did not align with state content standards, were deemed inappropriate for South Dakota, or had content flaws as indicated by statistical analysis. Ultimately, 42 field-test items passed the review and were added to the SDSAA operational item pool.

Table 9. Summary of SDSAA Field-Test Item Review

Grade	Total # of Items Administered	# of Items with n < 50	# of Items with biserial < 0	# of Items Rejected	# of Items Eligible for Operational Use
5	12	0	0	1	11
8	15	0	1	0	14
11	17	0	1	0	16
Total	44	0	2	1	41

5. TEST ADMINISTRATION

In the spring 2025 administration, the South Dakota Science Alternate Assessment (SDSAA) was administered to students in grades 5, 8, and 11 from March 24 to May 2, 2025. There was an online fixed form in each grade, which was the default method of administration, and a paper-pencil test as a special accommodation for students who were unable to fully access the online tests, even with the available accommodations. Each test was administered one-on-one, with one proctor (PR) administering the assessment to one student at a time, for both online and paper-pencil tests. The administration requires two machines in order to test; one for the PR and one for the student. The student's responses are captured in the student interface, and the PR can respond on the student's behalf, if necessary. The default online fixed-form tests consisted of 40 fixed operational items and 10 field-test items randomly selected from the Memorandum of Understanding (MOU) field-test item pool. The-paper-pencil tests with accommodation comprised only 40 operational items. The operational items in the paper-pencil test are identical to those in the online fixed-form test.

5.1 PROCTOR TRAINING

PR training is critical in producing reliable and valid test scores. Comparability of test scores between students and schools is based on the standardization of test administration and test scoring rules. If PRs do not follow the same procedures, student performance cannot be compared meaningfully.

Assessment coordinators (ACs), district administrators (DAs), and school coordinators (SCs) oversee all aspects of testing at their schools and serve as the main points of contact, while teachers (TE) and PRs administer the online assessments. The online Proctor Certification Course, PowerPoint, user guides, manuals, and regional trainings are used to train ACs and SCs in the online testing requirements and the mechanics of starting, pausing, and ending a test session. Training materials for the administration are found online at https://sd.portal.cambiumast.com/resources. ACs and SCs are responsible for training TEs and PRs.

Multiple online training opportunities are available and strongly recommended to key staff.

5.1.1 Proctor Certification Course

All school personnel who serve as TEs and PRs are highly recommended to complete an online PR Certification Course before administering the secure 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.

5.1.2 System Tutorials

The following presentations are offered to explain how the assessment system works (each of these presentations lasts approximately 60 minutes; slides are available on the Gateway at https://sd.portal.cambiumast.com/resources#refine=type:Training):

Reporting Training Webinar. This webinar provides an overview of how to navigate the reporting system, including generating, reading, and printing individual student reports (ISRs), building longitudinal reports, and creating rosters. In addition, slide notes and an additional presentation are provided as resources.

South Dakota Science Alternate Assessment (SDSAA) Training Webinar. This webinar provides an overview of the components of the SDSAA and explains how to access and administer online tests through the Proctor and Student Interfaces.

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

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

5.1.3 Practice and Training Test Site

In August 2020, separate training sites were opened for TEs, PRs, and students. TEs and PRs can practice administering assessments and starting and ending test sessions on the PR training site, and students can practice taking online assessments on the student practice and training site. The South Dakota State Assessment provides a sample set of items corresponding to the summative assessments for SDSAA.

A student can log in directly to the practice and training test site as a "Guest" without a PR-generated test session ID, or the student can log in through a training test session created by the TE or PR in the PR training site. Items in the student training test include all item types that are in the operational item pool.

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

5.2 Administration Manuals

The *Summative Science Alternate Assessment Test Administration Manual* (TAM) summarizes the SDSAA and provides brief guidelines for test administration. It includes the following:

- Overview of the background, purpose, and content specifications for SDSAA
- Assessment design
- Student inclusion and participation guidelines
- PR 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 *2024–2025 TAM* can be found at the following location:
https://sd.portal.cambiumast.com/content/contentresources/en/SDSAA-2025-Summative-TAM_FINAL.031225.pdf.

Included in the *2024–2025 TAM* is a short guide for the use of paper-pencil response option cards and printed test visuals for students approved for the paper-pencil test accommodation. This was provided to PRs who administered the paper-pencil tests to approved students. This guide can be found in the *2024–2025 TAM* and as a separate quick guide.

The *2024–2025 TAM* also includes Appendix B: SDSAA Augmentative and Alternative Communication Guidelines, which provide protocols for administering the assessment and for capturing the students' responses.

AAC Protocol for the SDSAA

The PR must adhere to the AAC Protocol to ensure that the student's response is generated in a manner that allows for accurate measurement of the student's ability.

Words/symbols/pictures/phrases that the student typically uses to communicate during instruction can be provided and should be words/pictures/symbols/phrases that are familiar to the student (e.g., events, descriptive words).
Introduce vocabulary related to the test item, but do not practice or teach the vocabulary in the context of the assessment. • For example, if the test item refers to "solar energy," it is appropriate to define and describe "solar energy" and its uses to familiarize the student with the related symbol(s) using the AAC device.
Any content represented in the grade-specific stimulus materials can be added to the student's AAC device (e.g., list of temporal words, problem/solution cards, words from mentor text or sample essay) to support student responding. • Ensure the words/pictures/symbols/phrases used from the stimulus materials are familiar or can readily be understood.
A response **cannot** be the result of a series of dichotomous choices of words, phrases, or sentences selected by the PR. The following is an example of a series of dichotomous choices that would **not** be allowed: The teacher asks, "Do you want to say that the amount in the table should be 5 or 4?" The student chooses 5. The teacher then asks, "Do you want to make it balls or pens?" The student chooses pens.
A response can be the result of the student completing a process directed by the PR using a series of two categories to communicate his or her word/picture/symbol/phrase preference. The following is an example of a series of dichotomous choices that is allowable: The Teacher asks, "Do you want People- Thing words or Action words?" The student selects People-Thing words and the Teacher then gives the choice of People or Thing words. The student chooses People words. The teacher then presents a series of choices of People words to allow the student to select the preferred person from those provided on the board. (As stated above, this should not result in a series of dichotomous choices of words, phrases, or sentences selected by the PR.)
Words/symbols/pictures/phrases **cannot** be arranged by the PR on a student's communication board so that any selection would be correct. *An exception to this would be if the student requests or selects a specific category level or board that has all words that could be used in a response (e.g., the student selects or requests the board filled with nouns or numbers and all would apply to the response).*

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

5.3 ACCOMMODATIONS

5.3.1 Online Version of the SDSAA

5.3.1.1 Allowable Accommodations – Accessibility Tools

The SDSAA 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 with the ability to access items and make a response. For the online assessment version, all items may be read and reread using the read-aloud function in the online testing system. Testing

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 is available for the SDSAA. The purpose of the universal tools is to provide the same level of supports during the alternate assessment as is provided regularly during instruction. Tools, supports, and accommodations are delivered to the student either as digitally delivered, embedded, components of the test administration system, or as non-embedded, delivered separately from the testing platform. Tools are accessibility resources of the assessment. Supports are features available for use by **any student** for whom the need has been indicated by an educator or team of educators with the parent or guardian and student. Accommodations are not modifications but rather changes in procedures or materials that increase equitable access during the state assessments. A complete list of available universal tools is provided in Table 10.

Table 10. List of Available Accessibility Tools

	Tools	Supports
Embedded	Breaks Calculator Digital Notepad Expandable Passages and/or Items Highlighter Keyboard Navigation Line Reader Mark for Review Strikethrough Tutorials Zoom	Color Contrast Masking Mouse Pointer Streamline Text-to-Speech Turn Off Any Tools Zoom (1.5X – 20X)
Non-embedded	Breaks Scratch Paper	Amplification Color Contrast Color Overlay Magnification Medical Supports Noise Buffers Periodic Table Printed test directions in English Read Aloud Separate Setting Simplified Test Directions

Table 11 presents the number of students who used the accessibility tools in the SDSAA.

Table 11. Total Number of Students Who Used Accessibility Tools

Accessibility Tools	Grade		
	5	8	11
Non-Embedded Designated Supports (Magnification)	0	1	0
Non-Embedded Designated Supports (Periodic Table)	0	0	2
Non-Embedded Designated Supports (Read Aloud Items)	6	8	1
Non-Embedded Designated Supports (Read Aloud Stimuli)	5	8	0
Non-Embedded Designated Supports (Separate Setting)	7	8	10
Non-Embedded Designated Supports (Simplified Test Directions)	4	8	3
Masking	93	69	57
Streamlined Mode	93	69	57
Text-to-Speech	16	9	13

5.3.1.2 Allowable Accommodations – 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 SDSAA via the TDS. Technology affords many ways to adapt student responses on the device. Any assistive technology that helps the student either access the assessment or provide their answers 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

5.3.2 Paper-Pencil Response Card Version of the SDSAA

Students participating in the SDSAA can access the assessment using the digital interface when provided the allowable supports. However, it is recognized that some students with disabilities may be better able to access the assessment with the paper-pencil response card version of the SDSAA. For the paper-pencil version, all items may be orally presented after the teacher uses the online digital interface to present the test item for the first time. If a student's IEP case manager determines that the student requires the paper-pencil version of the SDSAA due to the nature of his or her disability or disabilities, the student's PR will need to contact the SC or AC, who will notify the South Dakota Department of Education (SDDOE). The SC or AC is responsible for printing the paper-pencil response cards or providing a PDF file to be printed by the PR.

5.4 Online Administration

During the test administration, the student or PR selects 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 the TEs read to students. Students respond to each item by clicking one of the response options presented, or the PR 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 in the Early Stopping Rule segment, if no response is indicated or recorded by the student despite multiple attempts on different days, the PR 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 PR can advance to the next test item for administration.

An Early Stopping Rule (ESR) was available for students who were non-responsive to the first four items on each assessment. Students and PRs were required to follow the test administration guidelines put in place by the SDDOE Assessment Section. The ESR was used if the student had no consistent and observable mode of communication and was unable to respond to all of the first four items in the assessment. If the

student had a mode of communication and the ESR was used, the assessment was invalidated for misadministration.

5.5 Paper-Pencil Response Card Test Administration

In spring 2025, students who required a paper-pencil response card accommodation were administered a fixed-form test via the online testing system alongside printed response option cards which the PR placed in front of the student while listening to the test item read-aloud script via the online testing system. During the test administration, the student's item responses were entered into the online testing system directly by the PR after the student indicated their response option via the printed paper-pencil response option cards. No access-limited items were included in the paper-pencil tests.

5.6 Test Security

The Test Security Guidelines, included in the *Summative Science Alternate Test Administration Manual*, indicate that photocopying any printed testing materials is strictly prohibited. Printed paper-pencil response cards and test visuals are secure materials. SCs are responsible for receiving, accounting for, and returning all test materials to Cambium Assessment, Inc. (CAI). If CAI does not receive the returned test materials within the scheduled time frame, CAI makes significant effort to ensure that all secure materials are returned. Any known violations of test security are to be reported immediately.

5.6.1 Student-Level Testing Confidentiality

The online fixed-form 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 according to 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 tests to which they are assigned. If information must be sent via email or fax, only the Statewide Student Identifier (SSID) number, not the student's name, is included. A student cannot take a test under another student's SSID.

Student login information is entered only at the beginning of a test after an authorized PR creates and manages the test session and after the PR reviews and approves a test (and its settings) for the student. Only authorized users can make changes to the test registration system. 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 to take the online tests. Student enrollment information, including demographic data, is generated by the SDDOE and uploaded nightly to the online testing system via a secured file transfer protocol site during the testing period.

Only staff with the administrative roles of AC, DA, SC, or TE can view students' scores. ACs and DAs have access to all scores within their district. SCs have access to all scores within their school. TEs have

access to all scores of students rostered to them. The school will provide a printed copy of each child's score reports to their parent or guardian.

5.6.2 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 SCs and PRs receive separate passwords (assigned by the school) 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.

5.7 Prevention and Recovery of Disruptions in the Test Delivery System

CAI is continuously improving its ability to protect its 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 the 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 adjust and replace equipment before any problems occurred.

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

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

5.7.1 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.

The following sections describe key elements of the TDS, including the data integrity processes applied at each step.

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 to either retry to save or to log out; or
- if the system fails completely, the student is returned to the item at which place the failure occurred when he or she logs back in to the system.

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

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; any anomalies (e.g., nonscored or missing items, unexpected test lengths) are flagged, and a notification is immediately sent to our psychometricians and project team.

Database of Record

The Database of Record (DOR) is a cluster of database servers that, along with RAID systems, hold the finalized student data.

5.7.2 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.

5.7.3 Other Disruption Prevention and Recovery

These testing systems are designed to be extremely fault-tolerant. The system can withstand the 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 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 can reconstruct real-time data using the data retained on the TDS satellites and hubs.
- The server backup agents send alerts to notify system administration staff in the event of a backup error, at which time they will inspect the error to determine whether the backup was successful or needs to be rerun.

CAI's TDS is hosted in an industry-leading facility, with redundant power, cooling, state-of-the-art security, and other features that protect the system from failure. The system itself is redundant at every component, and the unique design ensures that 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.

6. SCORING

For the South Dakota Science Alternate Assessment (SDSAA), each student receives an overall scale score and an overall achievement level. No subscores are reported. This section describes the rules used in generating overall scores.

6.1 ITEM SCORING RULES

For multiple-choice items scored dichotomously, students receive 1 point for selecting the correct response option and 0 points for any incorrect response options. For multi-select items with two correct response options, students earn 2 points for selecting both options, 1 point for selecting only one, and 0 points for selecting none. If the proctor marks an item as *No Response* (NR), the student receives 0 points.

6.2 ATTEMPTEDNESS RULES FOR SCORING

When a student logs in to the test administration system and is presented with one item, they are considered to have participated if they provide a valid response for that item. A valid response includes either marking one or more response options or an NR marked by the proctor on the item. Participated students are counted as attempted. For accountability purposes, only students who receive a score will be counted as participants.

Scores are generated only for attempted tests. Detailed scoring rules are as follows (see Section 2.3, Test Assembly, for the description of test segments):

- If a student answers all items in Segments 1 and 2, the test is completed.
- If a student does not complete Segments 1 and 2, the test is considered incomplete with skipped items scored zero. All operational items in Segments 1 and 2 contribute to students' final scores.
- If a student has four consecutive NRs for items in Segment 1 (i.e., Early Stopping Rule [ESR] segment), the student is given the lowest obtainable scale score (LOSS) of the test. The SEM and theta score will be set to BLANK.

Table 14, Number of Attempted Students in SDSAA, in Section 7.1, lists the number of "Completed" tests, the number of "Incomplete" tests, and the number of ESR tests receiving the LOSS.

6.3 ESTIMATING STUDENT ABILITY USING MAXIMUM LIKELIHOOD ESTIMATION

The item response theory model (IRT) used to generate student scores employs the Rasch model for dichotomous items and the Partial Credit Model (PCM) for polytomous items. SDSAA tests 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

$$L_j\left(\theta_j \middle| \boldsymbol{z}_j, b_1, \ldots b_k\right) = \prod_{i=1}^{I} p_{ij}\left(z_{ij} \middle| \theta_j, b_{i,1}, \ldots b_{i,m_i}\right),$$

where $b_i' = \left(b_{i,1}, \ldots, b_{i,m_i}\right)$ 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 the person *j*, and *k* indexes the step of the item *i*.

Depending on the item score points, the probability $p_{ij}(z_{ij}|\theta_j, b_i, \ldots, b_{i,m_i})$ takes either the form of the Rasch model for items with 1 point or the form based on the PCM for items with 2 or more points.

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

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

in the case of items 2 two or more points,

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

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

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

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

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

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

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

6.4 Scoring All Correct and All Incorrect Cases

Using the MLE method, a test where no items are answered correctly (i.e., all incorrect) would receive a theta estimate of negative infinity, and a test where all items are answered correctly (i.e., all correct) would receive a theta estimate of positive infinity. To obtain real-valued theta score estimates for these extreme cases, 0.3 is added to an item score among the administered operational items for all incorrect cases, and 0.3 is subtracted from an item score for all correct cases.

6.5 Rules for Transforming Theta Scores to Scale Scores

The student's performance in each test is summarized in an overall test score referred to as a *scale score*. Student theta scores, which are based on the number of items answered correctly and the difficulty of those items, are converted into scale scores. This conversion involves a linear transformation using the formula

$SS = a * \theta + b$, where a is the transformation slope and b is the transformation intercept. Table 12 presents the scaling slope and intercept for each test. The final scale scores are rounded to the nearest integer.

Table 12. Scaling Constants

Subject	Grade	Slope (a)	Intercept (b)
Science	5	41.8737	311.2994
	8	56.5832	309.8030
	11	46.5680	314.6903

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

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

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

6.6 Lowest/Highest Obtainable Scale Scores

Extremely unreliable student ability estimates are truncated to the lowest obtainable scale score (LOSS) or the highest obtainable scale score (HOSS). For the SDSAA, the minimum and maximum scale scores are set at 100 and 500, respectively. Overall scale scores below 100 are truncated to 100, and those above 500 are truncated to 500. The standard error for LOSS and HOSS is calculated using the estimated theta scores derived from the responded items.

6.7 Achievement Levels

The scale scores are mapped into four achievement levels. Table 13 provides the range of scale scores at each achievement level by grade. These cut scores were established through a standard-setting process, as described in Chapter 10: Achievement Standards.

Table 13. Range of Scale Scores at Each Achievement Level

Grade	Level 1 (Not Met)	Level 2 (Nearly Met)	Level 3 (Met)	Level 4 (Exceeded)
5	100–277	278–299	300–338	339–500
8	100–260	261–299	300–336	337–500
11	100–273	274–299	300–329	330–500

7. SUMMARY OF SPRING 2025 OPERATIONAL TEST ADMINISTRATION

7.1 STUDENT PARTICIPATION

The South Dakota Science Alternate Assessment (SDSAA) was administered by grade level. All students in grades 5, 8, and 11 identified for participation in the SDSAA were assessed with the SDSAA. For a test to be considered attempted for scoring, a student needs to respond to at least one item, or the proctor marks *No Response* (NR) to at least one item.

Table 14 displays the total number of students who participated in the assessment by grade. Table 15 presents the total number of students who participated by demographic subgroup. Table 16 presents a detailed breakdown of the total number of participating students, categorized by both demographic subgroup and the Individuals with Disabilities Education Act (IDEA) disability category for each grade.

Table 14. Number of Attempted Students in SDSAA

Grade	Online Fixed Form				Total
	Completed	*ESR	Incomplete	Not Attempted	
5	105	5	1	1	112
8	81	7	1		89
11	70	3			73

*ESR=Early Stopping Rule

Table 15. Number of Participated Students by Subgroup

Group	Grade 5	Grade 8	Grade 11
All	112	89	73
Female	43	36	30
Male	69	53	43
American Indian or Alaskan Native	16	17	13
Asian	2	5	1
Black or African American	11	4	5
Hispanic or Latino	7	8	5
White	70	53	45
Native Hawaiian or Other Pacific Islander			
Multi-Racial	6	2	4
LEP	8	3	5
Section 504 Plan	3	1	2

Table 16. Number of Participated Students by Subgroup and Disability Category

Group	ASD	HI	ID	MD	TBI
Grade 5					
All Students	18	1	36	53	1
Female	4	1	16	19	1
Male	14		20	34	
American Indian or Alaskan Native	3		6	6	1
Asian				2	
Black or African American	3		3	5	
Hispanic or Latino	2		1	4	
White	10	1	25	31	
Native Hawaiian or Other Pacific Islander					
Multi-Racial			1	5	
LEP			4	4	
Section 504 Plan	1		1	1	
Grade 8					
All Students	18		37	31	1
Female	6		17	12	
Male	12		20	19	1
American Indian or Alaskan Native	6		6	5	
Asian	1		2	2	
Black or African American	1			3	
Hispanic or Latino	2		5	1	
White	6		24	20	1
Native Hawaiian or Other Pacific Islander					
Multi-Racial	2				
LEP			2	1	
Section 504 Plan			1		
Grade 11					
All Students	11		29	29	1
Female			17	12	1
Male	11		12	17	
American Indian or Alaskan Native	3		6	4	
Asian				1	
Black or African American			2	3	
Hispanic or Latino	1		3	1	
White	6		16	19	1
Native Hawaiian or Other Pacific Islander					
Multi-Racial	1		2	1	
LEP			3	2	
Section 504 Plan				2	

Note. ASD = Autism; HI = Hearing Impairment; ID = Intellectual Disability; MD = Multiple Disabilities; TBI = Traumatic Brain Injury.

7.2 SUMMARY OF STUDENT PERFORMANCE

Table 17–Table 19 present a summary of the spring 2025 SDSAA test results for all students and by subgroup, including the average and the standard deviation of scale scores, the percentage of students in each achievement level, and the percentage of proficient (level 3 + level 4) students. The results were based on the students who met attemptedness requirements for scoring and reporting of the SDSAA.

Table 17. Grade 5 Student Performance Overall and by Subgroup

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
All Students	111	286.95	53.27	32	32	25	11	36
Female	42	278.02	65.11	40	26	21	12	33
Male	69	292.38	44.22	28	35	28	10	38
American Indian or Alaskan Native	16	263.56	66.87	50	31	19	0	19
Asian	2*							
Black or African American	11	313	33.53	0	45	27	27	55
Hispanic or Latino	7*							
White	69	291.41	50.93	29	30	28	13	41
Native Hawaiian or Other Pacific Islander								
Multi-Racial	6*							
LEP	8*							
Section 504 Plan								

* Results for $n < 10$ are suppressed.

Table 18. Grade 8 Student Performance Overall and by Subgroup

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
All Students	89	272.92	61.92	38	30	18	13	31
Female	36	264.97	60.97	42	33	17	8	25
Male	53	278.32	62.56	36	28	19	17	36
American Indian or Alaskan Native	17	247.35	79.2	53	24	12	12	24
Asian	5*							
Black or African American	4*							
Hispanic or Latino	8*							
White	53	283.4	52.36	32	32	19	17	36
Native Hawaiian or Other Pacific Islander								
Multi-Racial	2*							
LEP	3*							
Section 504 Plan								

* Results for $n < 10$ are suppressed.

Table 19. Grade 11 Student Performance Overall and by Subgroup

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
All Students	73	288.14	46.57	27	38	23	11	34
Female	30	284.17	55.14	20	50	23	7	30
Male	43	290.91	40	33	30	23	14	37
American Indian or Alaskan Native	13	303.69	37.36	31	23	23	23	46
Asian	1*							
Black or African American	5*							
Hispanic or Latino	5*							
White	45	284.13	54.68	27	38	24	11	36
Native Hawaiian or Other Pacific Islander								
Multi-Racial	4*							
LEP	5*							
Section 504 Plan								

* Results for $n < 10$ are suppressed.

7.3 TEST-TAKING TIME

SDSAA tests are not timed. The time spent on each item may vary among individual students, which can provide valuable information about student testing behaviors and motivation. Proctors, who are familiar with their students, monitor the duration of test sessions. If needed, they can arrange additional time for students requiring more time to complete the tests.

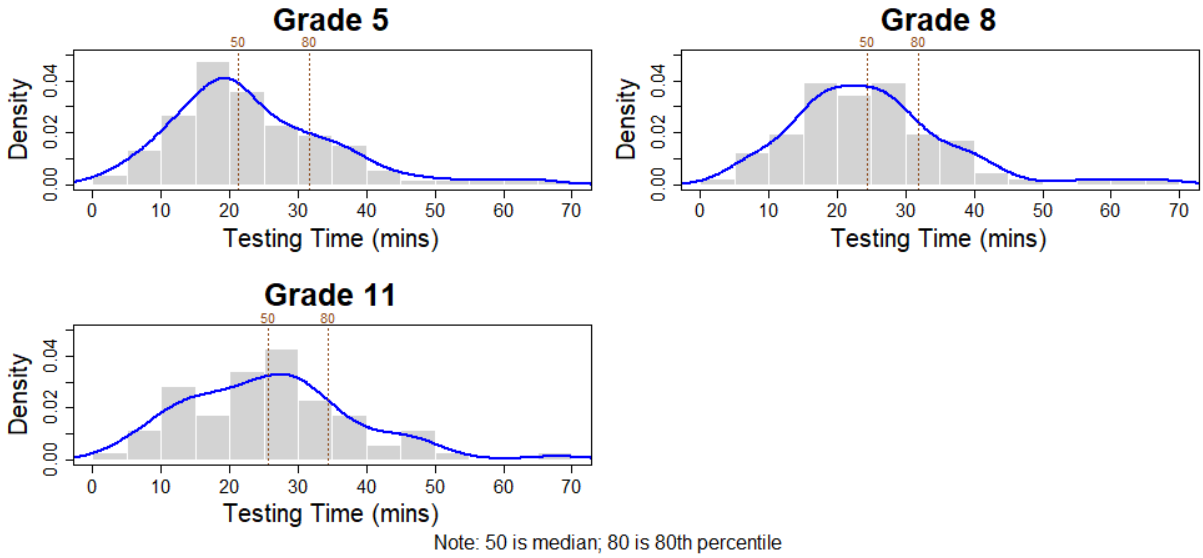
In the Test Delivery System (TDS), item response time is captured as the item page time (the time that a student spends on each item page) in milliseconds. Discrete items appear on the screen one item at a time, while items associated with a stimulus appear on the screen together, with the page time measured as the total time spent on all associated items. In this case, the time spent on each item is the average time of all 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 20 presents an average testing time and the testing time at various percentiles for the overall test. The analysis included all completed test records. The distribution of testing time is provided in Figure 2.

Table 20. Test-Taking Time

Grade	Average Testing Time (hh:mm)	Median Testing Time (hh:mm)	Testing Time in Percentiles (hh:mm)					
			Min	75th	80th	85th	90th	Max
5	00:23	00:21	00:02	00:29	00:32	00:36	00:38	01:05
8	00:25	00:24	00:05	00:30	00:32	00:35	00:39	01:08
11	00:26	00:26	00:05	00:32	00:35	00:38	00:44	01:08

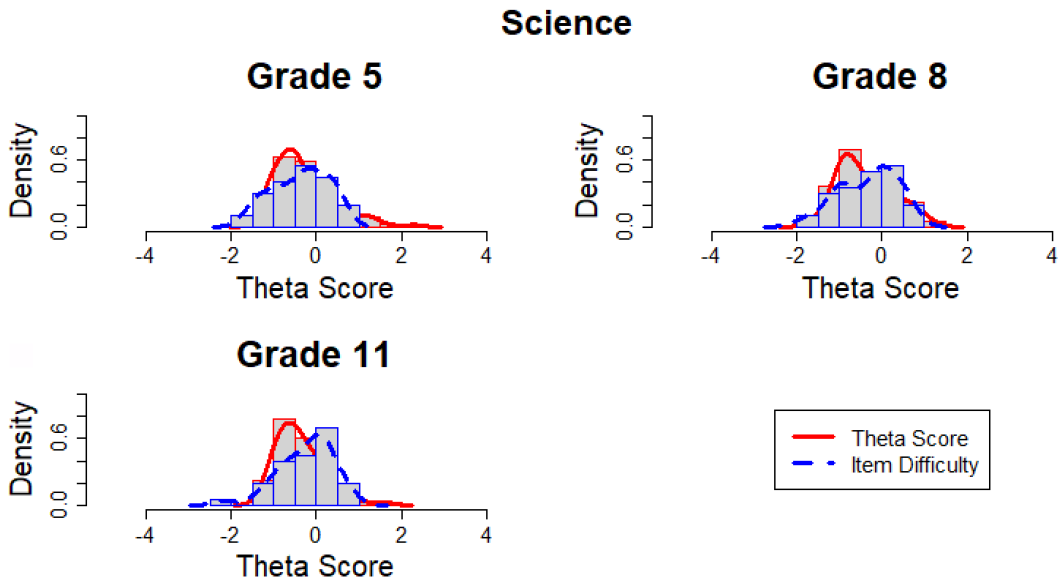
Figure 2. Distribution of Testing Time



7.4 Distribution of Student Ability and Item Difficulty for the SDSAA

Figure 3 displays the empirical distribution of students' overall theta scores and the distribution of the operational item difficulty parameter estimates. The student theta score distributions were based on completed test records, and the item difficulty parameter distributions were based on the online fixed-form tests.

Figure 3. Student Ability–Item Difficulty Distribution for SDSAA



8. 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 logic, and theoretical and empirical evidence are then provided to support the intended uses.

The South Dakota Science Alternate Assessment (SDSAA) was created to answer fundamental questions such as: What are the purposes of the assessment? Who are the intended users and what are the intended uses? Section 1.2, Purposes, Interpretations, and Intended Uses of the SDSAA, illustrates that the purposes and intended uses of the SDSAA are to measure students' academic performance and students' progress in meeting the state alternate academic achievement standards in science. The validation progress and validity argument for the SDSAA, documented in this chapter, are established around these uses.

Purposes and Intended Uses

The purposes, interpretations, and intended uses of the SDSAA serve as the foundation for test design and development. They play a crucial role in the validation process, as any statements about validity are tied to specific interpretations and uses.

The purposes and intended uses of the SDSAA are to measure students' academic performance and students' progress in meeting the state alternate academic achievement standards in science.

To fulfill its intended purposes, the SDSAA provides an overall scale score and an associated achievement level for each test. These achievement levels are determined based on the achievement standards established through a formal standard-setting process.

At the individual student level, the SDSAA test score can be used to estimate a student's academic performance; the associated achievement level, together with the ALDs, can indicate the knowledge and skills the student has attained in the assessed content area by the end of the academic year. Individual student scores and achievement levels can be compared across students who take the same test. Additionally, scores can also be aggregated to estimate the average performance of specific groups or to compare the average performance between different groups, such as by school, district, or gender.

Primary intended users of the SDSAA include the following:

- Students and families can use the results to stay informed about the student's learning progress in school.
- Teachers and educators can use the results to guide in-class instruction and identify students who need additional support.
- Educational agencies, organizations, and governments can use the test data and results to monitor educational improvement and make necessary changes to educational opportunities.

Validity Evidence

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; AERA, APA, & NCME, 2014). 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 SDSAA depend on the assessment meeting the relevant standards of validity.

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

Validity evidence for the SDSAA 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 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 Processes/Linguistic Processes)
- Evidence based on internal structure
  (Critical Element 3.3—Validity Based on Internal Structure)
- Evidence based on relations to other variables
  (Critical Element 3.4—Validity Based on Relations to Other Variables)

Evidence on test content validity is provided with both theoretical and empirical evidence related to content standards, test specifications, blueprints, item and test development process, administration process, and scoring. Evidence on response processes is gathered by conducting cognitive laboratory studies of student responses to items. Evidence on internal structure is examined in the results of intercorrelations among content strand scores. Due to lack of data, evidence on relations to other variables is not available for the SDSAA.

8.1 Evidence Based on Test Content

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

- Content standards
- 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.

8.1.1 Content Standards

Content standards are the starting point for test development. The SDSAA is developed based on the South Dakota Science Standards and designed for students with the most significant cognitive disabilities. The purpose of the SDSAA is to maximize access of this student population to the general education curriculum, ensure that all students with disabilities are included in the statewide assessments, and make certain that they are included in the educational accountability system.

The SDSAA is aligned to South Dakota's AAASs, the CCCs, which are linked to the South Dakota Science Standards.. The CCCs in science take the concepts from the South Dakota Science Standards and break them down to pinpoint the big ideas that are accessible for students with significant cognitive disabilities, at a reduced achievement level. The CCCs do address Science and Engineering Practices (SEPs), Disciplinary Core Ideas (DCIs), and crosscutting concepts (CCCs) from the standards. To further break down the big ideas in the CCCs, the South Dakota Department of Education (SDDOE) and Cambium Assessment, Inc. (CAI) staff prioritized the content and skills that were deemed most critical in the development of successful postsecondary outcomes for students with significant cognitive disabilities, creating Policy Achievement-Level Descriptors and Range Achievement-Level Descriptors (ALDs). For more details, see Section 1.4, Content Standards, in this technical report.

8.1.2 Test Blueprints

Content specifications in test blueprints specify the content standards to be covered in the test, and the minimum and maximum number of items from each content domain and sub-standards under those domains. The goal is to ensure the test has a balanced representation of items from each content standard.

For the SDSAA, each student receives 40 operational items. In spring 2025, a fixed form for operational items was created for each grade that met all requirements of the SDSAA blueprints, as shown in Table 21.

Table 21. Percentage of Administered Tests Meeting Blueprint Requirements

Grade	Standard	Minimum Required Items	Maximum Required Items	Percentage of 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	13	15	100%
	Physical Science	11	15	100%
11	Earth and Space Science	12	15	100%
	Life Science	12	15	100%
	Physical Science	11	15	100%

8.1.3 Item Development

Chapter 3, Item Development, provides a detailed description of how items are developed. The number and type of items to be developed are based on an evaluation of content needs and available sample size for field testing that can result in reliable statistics. Item writers are carefully chosen and well-trained to follow

standardized procedures and templates when creating items. All items undergo rigorous multiple rounds of internal and external reviews from the content and fairness perspective before they are field-tested in an operational context. After field testing, item analysis is conducted to examine whether items perform as expected. All items are reviewed by special education teachers and content experts in South Dakota before they are moved to the final operational item pool.

8.1.4 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 proctors do not follow the same procedures, student performance cannot be compared meaningfully. For the SDSAA, proctors are strongly encouraged to complete an online certification course before they can administer the test to their students. The guidelines for test administration are summarized in the *Summative Science Test Administration Manual* (TAM). See Chapter 5, in this technical report for details.

8.1.5 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 12.3, Quality Assurance in Test Scoring, provides a detailed description of quality control and monitoring procedures implemented within CAI to ensure that accurate scores are generated and reported.

8.2 Evidence Based on Response Processes

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 achievement 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. Students who participate in the alternate assessments represent a variety of disability categories and demonstrate many concomitant learning difficulties. Students in this population can exhibit difficulties responding to stimuli; committing information to working, short-term, or long-term memory; generalizing learning to familiar and novel environments; adaptive skills; 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 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 SDSAA has not yet had an opportunity to implement a cognitive lab study, results from the cognitive labs in other MOU states that share testing items can also provide insights. In spring 2019, Hawaii and Wyoming conducted cognitive lab studies. A summary of the cognitive lab studies and their findings is provided below.

Study Sample

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 levels was determined either by the student's score on the previous year's alternate assessment administration or teacher recommendation. In addition to the grade- 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 students with intellectual disabilities, autism spectrum disorders, and multiple disabilities.

Items Selected

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 English language arts (ELA), mathematics, and science item for each grade level, CAI content experts and state content experts agreed on the item's alignment with 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.

Data Collection

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 or fully mobile, and some use eye gaze to indicate their responses. Therefore, several different methods were 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.

Findings

The evidence and insights gained from the cognitive lab studies supported the validity argument that the alternate assessment elicited the intended cognitive process inherent in the grade level state content standards mediated by the Range Performance Level Descriptors (PLD). Students were challenged by many of the items but were able to apply some of the skills that they had learned in the classroom to answer test items successfully. Insights gained through the critical analysis of off-target student responses resulted in several completed and planned initiatives. An updated style guide and test specifications that included the consideration of language complexity, vocabulary, and audio and visual supports were created by the multi-state collaborative.

8.3 EVIDENCE BASED ON INTERNAL STRUCTURE

The measurement and reporting model used in the SDSAA assumes a single underlying latent trait, with achievement reported as a total score and scores for each content strand measured. The evidence on the internal structure is examined based on the correlations among content strand scores. The observed correlation between two claim scores with measurement errors can be corrected for attenuation (i.e., disattenuated correlation) as $r_{x'y'} = \frac{r_{xy}}{\sqrt{r_{xx}}*\sqrt{r_{yy}}}$, 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*. 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. Disattenuated correlations are higher than observed correlations. When the reliability estimate of either test is negative, disattenuated correlations cannot be computed.

The correlations among content strand scores are presented in Table 22. Values below the diagonal represent observed correlations, values above the diagonal represent disattenuated correlations, and values on the diagonal (bolded) represent strand score reliability estimates. Disattenuated correlations are capped at 1 for values exceeding 1.

Table 22. SDSAA Correlations Among Strands

Grade	Strand	Observed & Disattenuated Correlation		
		Strand 1	**Strand 2**	**Strand 3**
5	Strand 1: Earth and Space Science	**0.45**	1.00	1.00
	Strand 2: Life Science	0.59	**0.54**	1.00
	Strand 3: Physical Science	0.57	0.55	**0.41**
8	Strand 1: Earth and Space Science	**0.55**	0.42	0.91
	Strand 2: Life Science	0.22	**0.49**	1.00
	Strand 3: Physical Science	0.46	0.62	**0.46**
11	Strand 1: Earth and Space Science	**0.30**	1.00	1.00
	Strand 2: Life Science	0.36	**0.29**	0.90
	Strand 3: Physical Science	0.45	0.31	**0.42**

9. RELIABILITY

Reliability refers to the consistency in test scores. Reliability is evaluated in terms of the standard errors of measurement (SEM). In classical test theory, reliability is defined as the ratio of the true score variance to the observed score variance, assuming the error variance is the same for all scores. Within the item response theory (IRT) framework, measurement error varies conditioning on ability. The amount of precision in estimating achievement can be determined by the test information, which describes the amount of information provided by the test at each score point along the ability continuum. Test information is a value that is the inverse of the squared measurement error of the test; the larger the measurement error, the less test information is being provided.

The reliability evidence of the South Dakota Science Alternate Assessment (SDSAA) is provided with marginal reliability, SEM, conditional standard error of measurement (CSEM), and classification accuracy and consistency for each achievement standard.

9.1 MARGINAL RELIABILITY

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

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

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

where N is the number of students; $CSEM_i$ is the conditional standard error of measurement of the scale score for student i; and σ^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 IRT, the SEM is estimated as a function of test information provided by the set of items that make up the test. Because items administered in a computer-adaptive test (CAT) can vary among all students, the SEM can also vary across students, which yields a CSEM. The average CSEM across all students can be computed as

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

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

Table 23 presents the marginal reliability coefficients and the average CSEMs for the overall SDSAA scores, based on all completed tests, excluding the Early Stopping Rule test records. The reliability estimate for grade 11 is the lowest at 0.62, largely due to its comparatively small standard deviation of 25.66. By contrast, the same form produced a reliability estimate of 0.82 when it was administered last year.

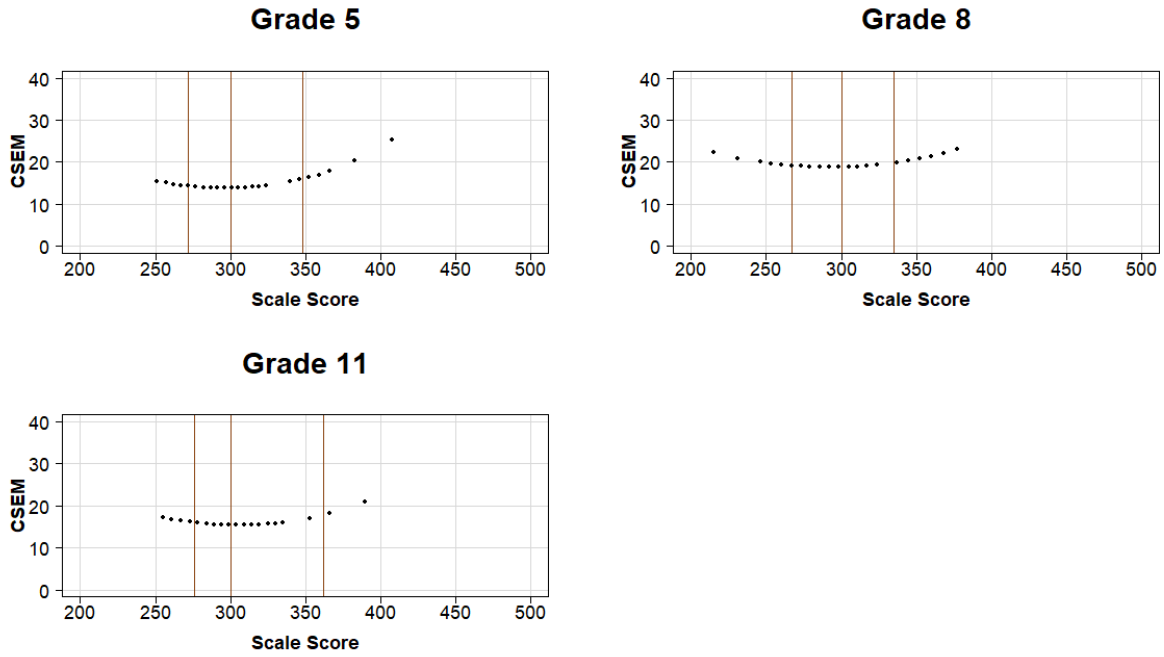
Table 23. Marginal Reliability of SDSAA Scores

Grade	Number of Students	Marginal Reliability	Scale Score Mean	Scale Score SD	Average CSEM
5	105	0.75	297.63	29.50	14.65
8	81	0.72	287.86	37.09	19.63
11	70	0.62	296.20	25.66	15.91

9.2 Standard Error Curves

Figure 4 presents a plot of the CSEM 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 except for a few extreme scores.

Figure 4. Conditional Standard Error of Measurement of SDSAA Scores



The CSEMs presented in Figure 4 are summarized in Table 24 by achievement 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 results for tests with extreme scores.

Table 24. Average Conditional Standard of Error Measurement by Achievement Level

Grade	Level 1	Level 2	Level 3	Level 4	Average CSEM
5	14.50	13.95	14.15	17.85	14.65
8	19.92	18.99	19.13	20.94	19.63
11	16.39	15.58	15.43	16.99	15.91

9.3 RELIABILITY OF PERFORMANCE CLASSIFICATION

When student performance is reported in terms of achievement 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, & 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 (CA) and the classification consistency (CC). CA 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. CC 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 achievement levels on two equivalent test forms.

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

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

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

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

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

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

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

where the likelihood function based on Rasch IRT models is

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

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

Classification Accuracy

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

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

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

The *CA* for the at or above the cut is estimated by

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

the *CA* for the below the cut is estimated by

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

and the overall CA 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 2×2 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 *CC* for the at or above the cut is estimated by

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

the *CC* for the below the cut is estimated by

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

and the overall CC 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 25 shows CA and CC indexes for the spring 2025 SDSAA. CAs are slightly higher than the CCs all achievement standards. The CC rate can be somewhat lower than the CA rate because CC assumes two test scores, both of which include measurement error, but the CA index assumes only a single test score and a true score, which does not include measurement error.

Table 25. Classification Accuracy and Consistency for Achievement Standards

Grade	Accuracy			Consistency		
	Level 2	Level 3	Level 4	Level 2	Level 3	Level 4
5	0.84	0.86	0.96	0.80	0.80	0.95
8	0.83	0.90	0.93	0.77	0.85	0.90
11	0.83	0.83	0.91	0.77	0.76	0.98

9.4 Reliability of Content Strand Scores

For the SDSAA, 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 26. The reliability coefficients were computed based on completed tests only.

Table 26. Marginal Reliability Coefficients of Content Strand Scores

Grade	Strand	Number of Items		Marginal Reliability	Scale Score Mean	Scale Score SD	Average CSEM
		Min	Max				
5	Earth & Space Science	13	13	0.45	298.45	35.48	26.15
	Life Science	13	13	0.53	300.43	40.21	26.71
	Physical Science	14	14	0.40	293.43	32.43	24.95
8	Earth & Space Science	13	13	0.55	282.98	54.05	35.97
	Life Science	13	13	0.49	288.01	50.47	35.59
	Physical Science	14	14	0.46	288.98	46.85	34.37
11	Earth & Space Science	13	13	0.30	286.59	33.64	28.05
	Life Science	12	12	0.29	296.54	35.86	30.03
	Physical Science	15	15	0.42	301.94	35.49	26.94

10. ACHIEVEMENT STANDARDS

After the spring 2022 operational administration, formal standard-setting workshops were conducted in all three grades to recommend achievement standards for the South Dakota Science Alternate Assessment (SDSAA). The standard-setting results replaced the interim achievement standards derived using a statistical linking method in the spring 2021 operational test administration.

In July 2022, following the close of the testing window, Cambium Assessment, Inc. (CAI) under contract to the South Dakota Department of Education (SDDOE), invited a panel of 18 teachers and administrators to recommend achievement standards (new cut scores) for the assessment. SDDOE 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 South Dakota's special education teacher population in terms of gender, race/ethnicity, and regional composition. SDDOE designated the most knowledgeable and experienced panelists at the workshop as table leaders.

For each test, the panelists recommended three cut scores, or achievement standards: Level 2 (Nearly Met), Level 3 (Met), and Level 4 (Exceeded).

10.1 STANDARD-SETTING PROCEDURES

South Dakota 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 meet the test blueprint. The panelists also reviewed the corresponding South Dakota Content Standards and Achievement-Level Descriptors (ALDs) 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 SDDOE. The plan was reviewed by the South Dakota Technical Advisory Committee and approved by SDDOE before the workshop.

The standard-setting workshop was held 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.

10.2 ACHIEVEMENT-LEVEL DESCRIPTORS

A prerequisite to standard setting is determining the nature of the categories into which students are classified. These categories, or achievement levels, are associated with ALDs that link the content standards to the achievement standards. There are four types of ALDs:

1. **Policy ALDs.** These ALDs describe the policy goals of each achievement level, which do not vary across grades or content.

2. **Range ALDs.** These ALDs, also called Instructional ALDs, describe what students know and can do throughout the range of each achievement level. For example, the Range ALD for Level 2 (Nearly Met) describes what students know and can do at that level up to just below the Level 3 (Met) cut score. The Range ALDs were created by the CAI content team starting with a small set of ALDs written to a subset of the Core Content Connectors (CCCs) that were posted on the

SDDOE website. CAI took these and matched them to the appropriate CCCs and created the remaining ALDs for the remaining CCCs. In July 2020, CAI sent SDDOE a draft of all Essence Statements and ALDs. SDDOE provided feedback, and posted the updated document to their website once suggested edits were incorporated. All ALDs were brought to South Dakota educators before standards setting during an ALD review meeting. At this full-day meeting, teachers reviewed and discussed the existing ALDs. They provided suggestions for edits to the wording of some ALDs to best fit the needs of South Dakota students. SDDOE reviewed the suggested edits from the committee and decided which edits to incorporate into the ALDs before standard setting, creating the final document that was then reposted to the SDDOE website.

3. **"Just Barely" ALDs.** These ALDs are sometimes called "threshold" or "target" ALDs. "Just Barely" ALDs are created by South Dakota educators during the standard-setting workshop and are used for standard setting only. The "Just Barely" ALDs describe what a student just barely scoring at the bottom of each achievement level knows and can do.

4. **Reporting ALDs.** These are abbreviated ALDs (typically 350 or fewer characters in length) created after standard setting has been completed, and they are used on the score reports to describe what students know and can do.

South Dakota uses four achievement levels to describe student performance: Level 1: Not Met, Level 2: Nearly Met, Level 3: Met, and Level 4: Exceeded. The standard-setting panelists used Range ALDs and Just Barely ALDs in the workshop.

10.3 Recommended Achievement Standards

Panelists were tasked with recommending three achievement standards that resulted in four achievement levels. Table 27 presents the achievement standard in scale score metric associated with the percentage of students reaching each standard based on the 2022 SDSAA results.

Table 27. Recommended Achievement Standards for SDSAA

Grade	Cut Scores			Impact Data		
	Level 2	Level 3	Level 4	Level 2	Level 3	Level 4
5	278	300	339	84%	48%	20%
8	261	300	337	81%	50%	20%
11	274	300	330	73%	46%	19%

11. REPORTING AND INTERPRETING SCORES

The Reporting System generates multiple online score reports that include information on student performance for presentation to students, parents, educators, and other stakeholders. The online score reports are generally produced immediately after testing has been completed. The Reporting System provides information on student performance and aggregated summaries at several levels—district, school, and roster. Since the performance score report is updated each time a student completes a test, authorized users (e.g., school principals, teachers) can quickly access students' performance scores and use them to improve student learning. In addition to individual student reports (ISRs), the Reporting System also produces aggregate score reports by class, school, and state. The timely accessibility of aggregate score reports can help users monitor students' performance in each grade by subject area and evaluate the effectiveness of instructional strategies; it can also inform not only the adoption of strategies to improve student learning and teaching but also professional development for educators and curriculum decisions for the state over time.

This section describes in detail both the types of scores that are reported in the Reporting System and how to interpret and use these scores.

11.1 REPORTING SYSTEM FOR STUDENTS AND EDUCATORS

11.1.1 Types of Online Score Reports

The Reporting System is designed to help educators and students answer questions about how students have performed on the assessments. The Reporting System is an online tool that provides educators and other stakeholders with timely, relevant score reports. In order to make score reports easy to read and understand, the Reporting System for the South Dakota Science Alternate Assessment (SDSAA) has been designed with stakeholders who are not technical measurement experts in mind. Simple language is used so that stakeholders can quickly understand assessment results and make inferences about student achievement. The Reporting System is also designed to present student performance in a uniform format. For example, similar colors are used for groups of similar elements, such as achievement levels, throughout the design. This design strategy allows readers to compare similar elements and avoid comparing dissimilar elements.

Once authorized users log in to the Reporting System, the dashboard page shows overall test results grouped by test family (e.g., grade 5 science) for all tests that a student has taken. Once the user clicks a test family, they are taken to a detailed dashboard. In addition, when authorized state-level users log in to the Reporting System and select "State View," the Reporting System generates a summary of student performance data for a specified test across the entire state.

Generally, the Reporting System provides two categories of online score reports: (1) aggregate score reports and (2) student score reports. Table 28 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*, accessed via a HELP button in the Reporting System.

Table 28. Types of Online Score Reports by Level of Aggregation

Level of Aggregation	Types of Online Score Reports
State District School Teacher Roster	• Number of students tested and percentage of proficient students (for overall students and by subgroup) • Average scale score (for overall students and by subgroup) • Percentage of students at each achievement level • Participation rate (for overall students)[1] • On-demand student roster report
Student	• Total scale score and standard error of measurement • Achievement level for overall score with Achievement-Level Descriptors • Average scale scores for individual school, district, and the state

[1] Participation rate reports are provided at the state, district, and school levels.

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 29 presents the types of subgroups and subgroup categories provided in the Reporting System.

Table 29. Types of Subgroups

Subgroup	Category
Enrolled Grade	05 08 11
IDEA Indicator	Yes No
Ethnicity	American Indian or Alaska Native Asian Black or African American Hispanic or Latino Native Hawaiian or Pacific Islander White Multi-Racial
Gender	Male Female
Limited English Proficiency Status	Yes No
Section 504 Status	Yes No

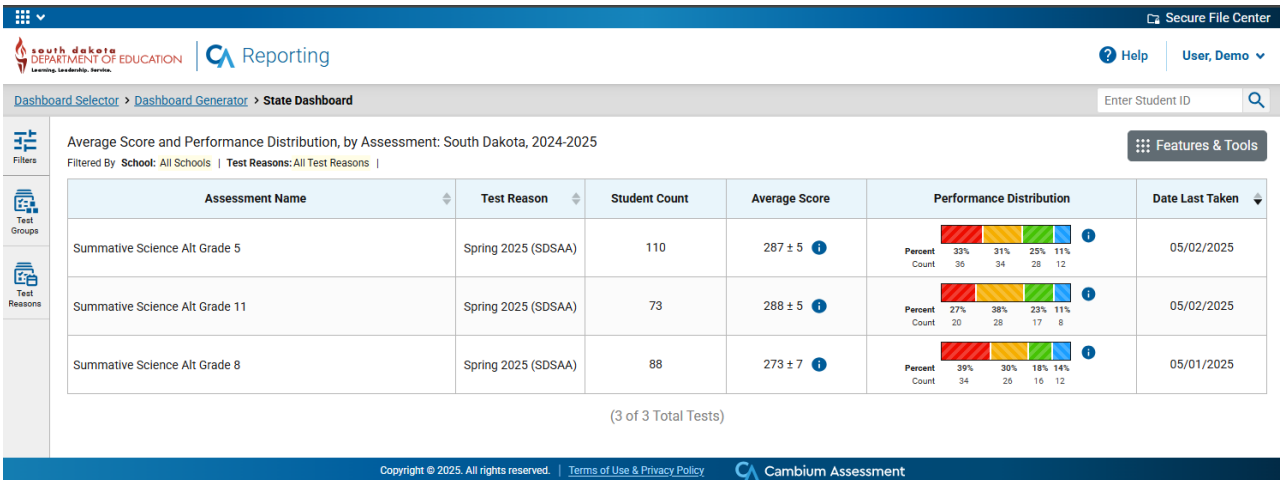
11.1.2 Reporting System

11.1.2.1 Dashboard

The first page users see when they log in to the Reporting System contain summaries of student performance by test family (i.e., Summative Science Alt Grade 5). District personnel see district summaries, school personnel see school summaries, and teachers see summaries of the students rostered to them. State personnel and district area personnel would need to select the district 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 achievement level. Exhibit 1 presents a sample dashboard page at the state level.

Exhibit 1. Dashboard: State Level



Educators can click the subject group to view individual test results for the selected test group. Once the user clicks 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) the average score and standard error of the means, and (3) the percentage and counts of students at each achievement level. Exhibit 2 presents a sample detailed dashboard page for SDSAA at the district level.

Exhibit 2. Dashboard: District Level

11.1.2.2 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 state and district of the school are provided above the school summary results as well so that school performance can be compared with the aggregate levels.

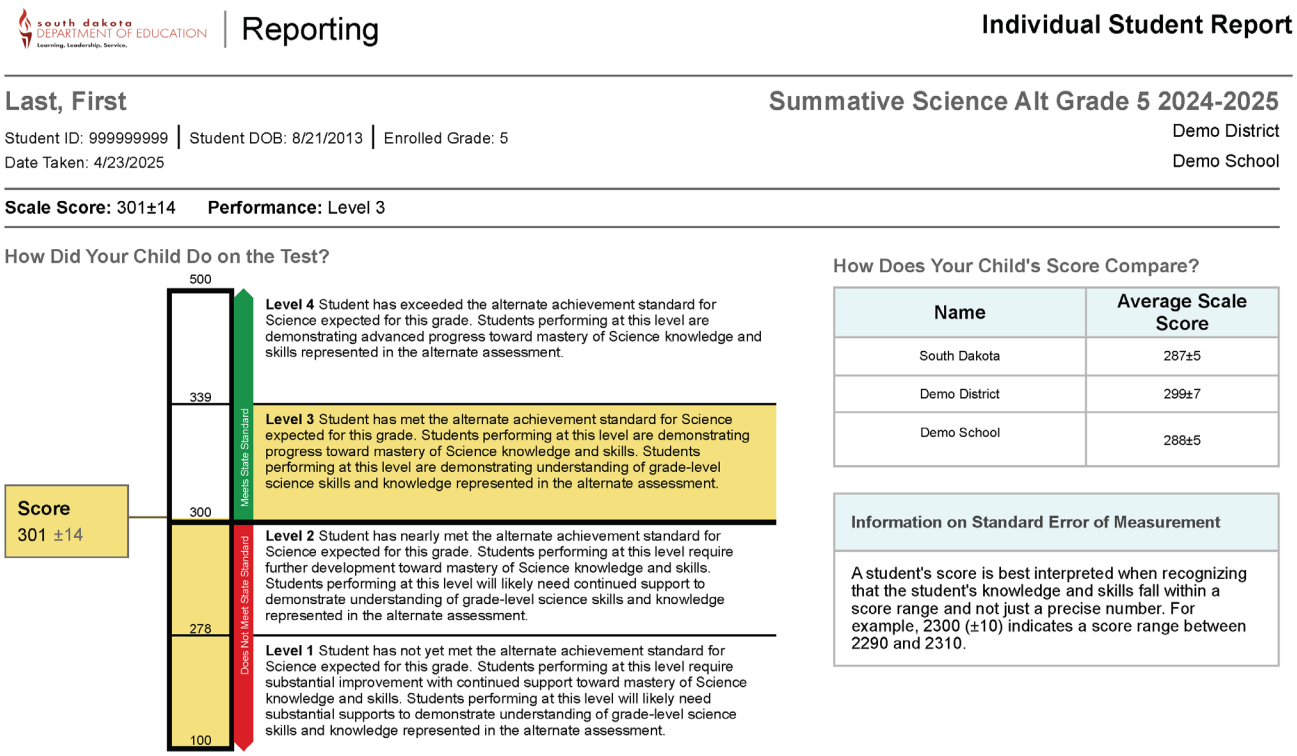
The aggregated subject summary report provides summaries on a specific subject area, including (1) the number of students tested, (2) the average scale score and standard error associated with the average scale score, (3) the percentage of proficient students, and (4) the percentage and counts of students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

11.1.2.3 Student Detail Page

When a student completes a test, an online score report appears in the individual student report (ISR) in the Reporting System. Exhibit 3 presents a sample student detail page. 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); and (2) the achievement level for the overall test.

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

Exhibit 3. Student Detail Page for Science



11.1.2.4 Participation Rate

The Test Information Distribution Engine (TIDE) provides participation rate reports for states, districts, and schools to help monitor the student participation rate. During and after the testing window, participation data are updated each time a student completes a test, allowing authorized users to track participation in real time and evaluate final participation outcomes.

Included in the participation table are the total number of students who have registered to take the SDSAA, the total number and percentage of students who have started the tests, and the total number and percentage of students who have completed the tests.

Exhibit 4 presents a sample participation rate report at the state level.

Exhibit 4. Participation Rate Report at the State Level

Grade	Total Student	Total Student Started	Total Student Completed	Percent Started	Percent Completed
5	91	73	73	80.22	80.22
8	127	112	110	88.19	86.61
11	104	89	88	85.58	84.62

11.2 Interpretation of Reported Scores

A student's performance on a test is reported with a scale score and an associated achievement level for the overall test. Students' scores and achievement levels are summarized at the aggregate levels. The next section describes how to interpret these scores.

11.2.1 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. Interpretation of scale scores is more meaningful when the scale scores are used along with achievement levels and Achievement-Level Descriptors (ALDs).

11.2.2 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 test administrations, being sometimes a little higher, a little lower, or the same. The SEM represents the precision of the scale score, or the range in which the student would likely score if a similar test 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. SEM can be different for the same scale score, depending on how closely the administered items match the student's ability.

11.2.3 Achievement Level

Achievement levels are proficiency categories on a test that students fall into based on their scale scores. For the SDSAA, scale scores are mapped into four achievement levels (i.e., Level 1, Level 2, Level 3, and Level 4) using three achievement standards (i.e., cut scores). ALDs are a description of the content area knowledge and skills that test takers at each achievement level are expected to possess. Thus, achievement levels can be interpreted based on ALDs.

11.2.4 Aggregated Score

Student scale scores are aggregated at 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 achievement level for the overall test is reported at the aggregate level to represent how well a group of students performed overall.

11.3 Appropriate Uses for Scores and Reports

Assessment results can provide information about individual students' achievement on the test. Overall, assessment results tell what students know and can do in certain subject areas.

Assessment results for student achievement on the test can be used to help teachers or schools make decisions on how to support student learning. Aggregate score reports at the teacher and school level provide information regarding the strengths and weaknesses of their students 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 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 achievement. Moreover, although student scores may be used to help make important decisions about students' placement and retention, or teachers' instructional planning and implementation, the assessment results should not be used as the only source of information. Given that assessment results measured by a test provide limited information, other sources on student achievement, such as classroom assessment and teacher evaluation, should be considered when making decisions about student learning.

12. 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.

12.1 OPERATIONAL TEST CONFIGURATION

For the operational test, a test configuration file is a 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 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 South Dakota Science Alternate Assessment (SDSAA) tests. 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.

12.1.1 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 SDSAA, 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 expected.

12.1.2 User Acceptance Testing and Final Review

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

12.2 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 in each item, total number of field-test items and operation items, and that the test record contains no data from items that have been invalidated.

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

12.3 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 exceed 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 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 testing window for early detection of any unexpected issues. Any deviations from the expected outcome are flagged, investigated, and resolved.

Blueprint match and item exposure reports allow psychometricians to verify that test administrations conform to the simulation results. The QA reports can be generated on any desired schedule. Item analysis and blueprint match reports are 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 30 presents an overview of the QA reports.

Table 30. Overview of Quality Assurance Reports

QA Reports	Purpose	Rationale
Item Statistics	To confirm whether items work as expected	Early detection of errors (key errors for selected-response items)
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

12.4 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 mis keyed items, item score distribution, or other scoring problems. Potential issues are automatically flagged 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 Reporting System, which is responsible for presenting individual-level results and calculating and presenting aggregate results. No score is reported in the Reporting System until it passes all the QA system's validation checks.

REFERENCES

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education. (NCME). (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association. https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage, 1994.

Guo, F. (2006). Expected classification accuracy using the latent distribution. *Practical Assessment, Research & Evaluation, 11*(6), 1–9.
https://openpublishing.library.umass.edu/pare/article/1302/galley/1253/view/

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement, 13*(4), 253–264. https://doi.org/10.1111/j.1745-3984.1976.tb00016.x

Linacre, J. M. (2004). Rasch model estimation: Further topics. *Journal of Applied Measurement*, *5*(1), 95–110. https://www.winsteps.com/a/Linacre-estimation-further-topics.pdf

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, *32*(2), 179–197. https://doi.org/10.1002/j.2333-8504.1993.tb01559.x

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

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. https://doi.org/10.1007/BF02296272

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. https://doi.org/10.1177/001316449205200202Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The Bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249–281). Mahwah, NJ: Lawrence Erlbaum Associates.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The Bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249–281). Lawrence Erlbaum Associates.

Muñiz, J., Hambleton, R. K., & Xing, D. (2001). Small sample studies to detect flaws in item translations, *International Journal of Testing*, *1*(2), 115–135. https ://doi.org/10.1207/S15327574IJT0102_2

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). Knowing what students know: The science and design of educational assessment**.** National Academy Press.

Sireci, S. G., & Rios, J. A. (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation: An International Journal on Theory and Practice*, *19*(2–3), 170–187. https://doi.org/10.1080/13803611.2013.767621

Subkoviak, M. J. (1976). Estimating reliability from a single administration of a criterion-referenced test*. *Journal of Educational Measurement*, *13*(4), 265–276. https://doi.org/10.1111/j.1745-3984.1976.tb00017.x

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)*. University of Minnesota, National Center and State Collaborative. http://www.ncscpartners.org/media/default/pdfs/lci-project-report-08-21-12.pdf

U.S. Department of Education. (2018, September 24). *A state's guide to the U.S. Department of Education's assessment peer review process.* https://oese.ed.gov/files/2020/07/assessmentpeerreview.pdf