South Dakota

Science Alternate Assessment

2022-2023 Technical Report

Science Grades 5, 8, and 11

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

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EXHIBITS

1. OVERVIEW

This report provides a technical summary of the 2022-2023 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 10 chapters that discuss all the evidence accrued about the technical quality of a testing system. The analyses included in this report are based on South Dakota alternate assessment data and address all aspects of the technical requirements described in the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014) and in *A State's Guide to the U.S. Department of Education's Assessment Peer Review Process* (U.S. Department of Education, 2018).

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

Chapter 6 provides evidence for the reliability of the SDSAA, including marginal reliability, standard errors of measurement (SEMs), and classification accuracy and consistency of achievement standards. Chapter 7 describes the scoring procedures used in producing scale scores and achievement levels. Chapter 8 describes the procedure to set achievement standards. Chapter 9 provides a description of the score reporting system and the interpretation of test scores. Chapter 10 provides an overview of the quality assurance (QA) processes that are used to ensure that all test development, administration, scoring, and reporting activities are conducted with fidelity to the developed procedures.

1.1 THE SOUTH DAKOTA SCIENCE ALTERNATE ASSESSMENT

The SDSAA is composed of tests that are ultimately 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. See Table 5 in Section 3.1- Student Participation.

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 is 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

test scores; in spring 2022, a formal standard setting workshop was conducted to establish the Achievement Standards used for score reporting. In the fall of 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 Core Content Connectors. 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 ALDs. It was also deemed important to 1) make sure that the ALDs did have a strong alignment to 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 2022-23, 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 of 2023. Since the full alignment review was not completed until June 2023, and to ensure that only aligned items were administered to students in spring 2023, a fixed form was assembled and administered in each grade for the spring 2023 administration, using only items aligned during the first alignment workshop in January.

1.2 Alternate Assessment Eligibility

Most students with disabilities can participate in the general state assessments when provided with the appropriate accommodations. However, for students with the most significant cognitive disabilities, it may be more appropriate to participate in the alternate assessment. Decisions concerning a student's participation in statewide assessments are made by each student's individualized education program (IEP) team. Guidance for IEP teams to inform decisions about which assessment is most appropriate for each student is provided in the Student Participation Criteria from the *Spring 2023 SDSAA Test Administration Manual* at https://sd.portal.cambiumast.com/resources/educators/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 of these criteria must be met for a student to qualify to take the 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 are linked to the enrolled grade-level content standards and address knowledge and skills that are appropriate and challenging for this student.

1.3 CONTENT STANDARDS

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

The SDSAA is aligned to the content standards for science, which are linked to the CCCs.

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. The CCCs do address science and engineering practices (SEPs), disciplinary core ideas (DCIs), and crosscutting concepts (CCs) from the standards.

To further break down the big 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 and Range Achievement Level Descriptors (ALDs).

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. Cambium took these and matched them to the appropriate CCCs and created ALDs for the remaining CCCs. Cambium worked with SDDOE to finalize the wording for each of the remaining ALDs. All ALDs were brought to South Dakota educators prior to Standard Setting during an ALD Review Meeting for their review. Any edits suggested by the committee were reviewed by SDDOE and SDDOE made the decision about which suggested edits to incorporate before Standard Setting.

Furthermore, in January 2023, a second subset of South Dakota educators convened to review all of the ALDs prior to the alignment workshops. The committee reviewed the ALDs to make sure they: (1) align to the CCCs, (2) describe what students can do, (3) define difference in content across the achievement levels, (4) describe the contextual or scaffolding characteristics needed so a student can demonstrate the skill, (5) increase in cognitive demand across achievement levels in a cogent way, and (6) provide a mental picture of increases in skill across achievement levels. The committee suggested some edits and SDDOE made the final determination on the implimentation of those edits. These finalized ALDs serve as the foundation for the development of SDSAA items.

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

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 prior to the ALD review meeting in January 2023. The levels are as follows:

- **Level 4 (Exceeded):** A student who is Exceeded demonstrates a level of understanding that includes the ability to "bring together" the Disciplinary Core Ideas (DCI) and/or Science and Engineering Practices (SEP) and/or Crosscutting Concepts (CCC) associated with a PE.

- **Level 3 (Met):** A student who is Met demonstrates an understanding of the Disciplinary Core Ideas (DCI) and/or Science and Engineering Practices (SEP) and/or Crosscutting Concepts (CCC) within the PE at the conceptual level described in the Core Content Connectors.

- **Level 2 (Nearly Met):** A student who is Nearly Met demonstrates some understanding of the content of the PE, but that understanding is incomplete and does not yet meet the expectations found in the Core Content Connectors. This student's understanding is partial but emerging.

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

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.

1.4 MEMORANDUM OF UNDERSTANDING ON ITEM-SHARING INITIATIVE

In 2018, South Carolina, Hawaii, and Wyoming signed a Memorandum of Understanding (MOU) on item sharing in item development and field testing. Each state contributed a predetermined number of items proportional to their state's student population for alternate assessments. In early 2019, Idaho and Vermont joined the collaborative item development and field-testing effort and participated in the spring 2019 field

test. In spring 2020, Montana and South Dakota joined the MOU for science assessments. In 2022, Vermont exited the MOU. Because the total number of students in alternate assessments is very small in each state, field-testing common items in all MOU states allowed for the calibration of items based on the combined data across all states. In addition to the MOU shared item pool, each state also developed some items that aligned to the state's specific content standards or content specifications.

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

2. TEST ADMINISTRATION

In the spring 2023 administration, the South Dakota Science Alternate Assessment (SDSAA) was administered to students in grades 5, 8, and 11 from March 1 to May 12, 2023. 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 proctor and one for the student. The student's responses are captured in the student interface and the PR can respond on the students' behalf, if necessary. The default online fixed-form tests consisted of 40 fixed operational items and 10 field-test items randomly selected from the MOU field-test item pool. The paper-pencil tests with accommodation only comprised 40 operational items. The operational items in the paper-pencil test are identical to those in the online fixed-form test.

2.1 PROCTOR TRAINING

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

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

2.1.1 Online Training

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

PR 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 and valid assessments. This web-based course is 30–45 minutes long and covers information on testing policies and the steps for administering a test session in the online system. The course is interactive, requiring participants to start test sessions under different scenarios. Throughout the training and at the end of the course, participants answer multiple-choice questions about the information provided.

System Tutorials

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

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

Proctor (PR) Interface for Online Testing. This tutorial prepares DAs, SCs, and PRs for the assessments by providing an overview of the PR Interface and the TDS, including how to start and monitor a test session using the PR Interface.

Reporting Training. The module provides an overview of how to navigate the reporting system, to include generating, reading, and printing Individual Student Report (ISR), building longitudinal reports, and create rosters. In addition, slide notes and an additional presentation are provided as resources.

Student Interface Overview. This tutorial provides an overview of the online Student Interface in the TDS.

Test Information Distribution Engine (TIDE). This tutorial 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.

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 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.

2.2 ADMINISTRATION MANUALS

The *2022-2023 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 *2022-2023 TAM* can be found at https://sd.portal.cambiumast.com/resources/educators/summative-science-alternate-assessment-test-administration-manual-tam.

Included in the *2022-2023 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 *2022-2023 TAM* and as a separate quick guide.

The *2022-2023 TAM* also includes Appendix B: SDSAA Augmentative and Alternative Communication Guidelines, which provide protocols for administering the assessment and for capturing the students' responses. See below.

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.

2.3 ACCOMMODATIONS

2.3.1 Online Version of the SDSAA

2.3.1.1 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 the ability to access the items and make a response. For the online assessment version, all items may be read and re-read 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 are available for the SDSAA. The purpose of the universal tools is to provide the same level of supports during the alternate assessment as are regularly provided 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 separate from the testing platform. Tools are accessibility resources of the assuessment. Supports are features available for use by **any student** for whom the need has been indicated by an educator, or team of educators with parent or guardian and student. Accommodations are changes in procedures or materials that increase equitable access during the state assessments and not modification. A complete list of available universal tools is provided in Table 2.

Table 2. List of Available Accessibility Tools

	Tools	Supports	Accommodations
Embedded	Breaks Calculator[1] Digital Notepad English Dictionary[2] English Glossary Expandable Passages and/or Items Global Notes[3] Highlighter Keyboard Navigation Line Reader Mark for Review Math Tools[4] Reference Guide Spell Check Strikethrough Tutorials Thesaurus[5] Writing Tools[6] Zoom	Color Contrast Illustration Glossaries[7] Masking Mouse Pointer Streamline Text-to-Speech[8] Text-to-Speech in Spanish Translated Test Directions[9] Translations (Glossaries)[10] Translations (Dual Language)[11] Turn off Any Tools Zoom (1.5X – 20X)	American Sign Language[12] Braille Braille Transcript Closed Captioning[13] Speech-to-Text Text-to-Speech[14] Word Prediction
Non-embedded	Breaks English Dictionary[15] Reference Guide Scratch Paper Thesaurus[16]	Amplification Bilingual Dictionary[17] Color Contrast Color Overlay Illustration Glossaries[18] Magnification Medical Supports Noise Buffers Printed test directions in English Read Aloud[19] Read Aloud in Spanish[20] Separate Setting Simplified Test Directions Translated Test Directions (also ASL) Translations (Glossaries) [21]	100s Number Table Abacus Alternate Response Options[22] Braille[23] Calculator[24] Large Print Multiplication Table Print on Demand Read Aloud[25] Scribe[26] Speech-to-Text Word Prediction

[1] For calculator-allowed items only in grades 6 – 8 and 11; Science – available for all; [2] For ELA performance task full writes; [3] For ELA performance tasks; [4] Includes embedded ruler, embedded protractor; [5] For ELA performance task full writes; [6] Includes bold, italic, underline, indent, cut, paste, spell check, bullets, undo/redo; [7] For math items; [8] For math & science stimuli and items and ELA items (not for reading passages); [9] For math items; [10] For math items; [11] For math items; [12] For ELA listening Items and math items; [13] For ELA listening items; [14] Available for ELA reading passages, all grades; [15] For ELA performance task full writes; [16] For ELA performance task full writes; [17] For ELA performance task full writes; [18] For paper pencil math items; [19] For math & science stimuli and items and ELA items (not for reading passages); [20] For math, all grades; [21] For math items on the paper-pencil test; [22] Includes adapted keyboards, large keyboards, StickyKeys, Mouse Keys, FilterKeys, adapted mouse, touch screen, head wand, and switches; [23] Paper-pencil assessment; [24] For calculator-allowed items only in grades 6–8 and 11 (Braille/talking calculators); [25] For ELA reading passages; [26] For ELA performance task full write.

Table 3 presents the number of students who used the accessability tools in the SDSAA.

Table 3. Total Number of Students Who Used Accessibility Tools

Accessibility Tools	Grade		
	5	8	11
Permissive Mode (On)	1		
Print on Request (Items)	1		
Print on Request (Stimuli and Items)	1	1	1
Non-Embedded Accommodations (Scribe)	2	1	4
Non-Embedded Designated Supports (Simplified Test Directions)	1		2

2.3.1.2 Allowable Accommodations – Assistive Technology

Assistive technology (AT) that is documented in the student's 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

2.3.2 Paper-Pencil Response Card Version of the SDSAA

Eligible students take 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 the first time. If a student's IEP care coordinator 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 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.

2.4 Online Administration

During test administration, the student or PR touches the button bearing an ear icon for the stimulus, question, and response option portion of each item to be read aloud. The read-aloud script is a recorded human voice. The speed of narration is comparable to the average speed of narration when teachers read to students. Students respond to each item by clicking one of the response options presented, or the 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, 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.

In spring 2023, an Early Stopping Rule was available for students who were non-responsive to the first four items on each test. Students and PRs were required to follow the administration guidelines put in place by the SDDOE Assessment Section. The Early Stopping Rule 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 early stopping rule was used, the test was invalidated for misadministration.

2.5 PAPER-PENCIL RESPONSE CARD TEST ADMINISTRATION

In spring 2023, 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 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 on the paper-pencil tests.

2.6 TEST SECURITY

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

2.6.1 Student-Level Testing Confidentiality

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

FERPA prohibits the public disclosure of student information or test results. To comply with the secure standards, student names and IDs are communicated via a Secure File Transfer Protocol (SFTP). Student login information is associated with the particular tests to which they are assigned. 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 in order 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 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 within their classrooms. The school will provide a printed copy of each child's score reports to their parent/guardian.

2.6.2 System Security

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

Password Protection. This security measure ensures that all access points by different roles—at the state, districts, 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.

2.7 Prevention and Recovery of Disruptions in the Test Delivery System

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

CAI has developed a unique monitoring system that is sensitive to changes in server performance. Most monitoring systems provide warnings when something is going wrong; in addition to general warnings of malfunction, our monitoring system also provides warnings when any given server is performing differently from its performance over the few hours prior, or differently than the other servers performing the same

jobs. Subtle changes in performance often precede actual failure by hours or days, allowing us to detect potential problems, investigate them, and mitigate them before a failure. On multiple occasions, this has enabled us to make adjustments and replace equipment before any problems occurred.

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

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

2.7.1 High-Level System Architecture

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

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

The following sections describe key elements of the testing system, 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 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., unscored 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.

2.7.2 Automated Backup and Recovery

Every system is backed up nightly. Industry-standard backup and recovery procedures are in place to ensure the safety, security, and integrity of all data. This set of systems and processes is designed to provide complete data integrity and prevent the loss of student data. Redundant systems at every point, real-time

data integrity protection and checks, and well-considered, real-time backup processes prevent the loss of student data, even in the unlikely event of system failure.

2.7.3 Other Disruption Prevention and Recovery

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

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

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

3. SUMMARY OF SPRING 2023 OPERATIONAL TEST ADMINISTRATION

3.1 STUDENT PARTICIPATION

The South Dakota Science Alternate Assessment (SDSAA) was administered by grade level. All students meeting alternate assessment eligibility criteria in grades 5, 8, and 11 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* to at least one item.

Table 4 presents the total number of students who participated in the assessment by grade. No students took the paper-pencil tests with accommodation. Table 5 presents the alternate assessment participation rate, computed as the number of students meeting alternate assessment elibigility criteria and taking the SDSAA divided by the total number of students in the state taking the general education summative tests and the SDSAA. Table 6 presents the total number of students who participated by demographic subgroup. Table 7 presents the total number of students who participated by demographic subgroup and the Individuals with Disabilities Education Act (IDEA) disability category for each grade.

Table 4. Number of Attempted Students in SDSAA

Grade	Online Fixed Form				Total
	Completed	ESR	Incomplete	Not Attempted	
5	88	2	1		91
8	80	6			86
11	73	4			77

Note. ESR=Early Stopping Rule

Table 5. Overall Alternate Assessment Participation Rate

Subject	Grade	Number of Participants in SDSAA	Number of Participants in General Education	Overall State Alternate Assessment Participation Rate (%)
Science	5	91	10,605	0.9%
	8	86	10,744	0.8%
	11	73	9,474	0.8%
	Overall	250	30,823	0.8%

Table 6. Number of Participated Students by Subgroup

Group	Grade 5	Grade 8	Grade 11
All	91	86	77
Female	26	36	22
Male	65	50	55
American Indian or Alaskan Native	11	8	13
Asian	1	1	1
Black or African American	5	6	4
Hispanic or Latino	10	9	5
White	59	56	51
Native Hawaiian or Other Pacific Islander	0	0	0
Multi-Racial	5	6	3
LEP	6	4	2
plan504	1	1	1

Table 7. Number of Participated Students by Subgroup and Disability Category

Group	AUT	ID	MD	TBI	OHI	DB
Grade 5						
All Students	17	42	26	2		
Female	4	16	5	1		
Male	13	26	21	1		
American Indian or Alaskan Native	1	5	4	1		
Asian			1			
Black or African American	2	2	1			
Hispanic or Latino	2	6	2			
White	9	28	18	1		
Native Hawaiian or Other Pacific Islander						
Multi-Racial	3	1				
LEP	1	4	1			
Plan504		1				
Grade 8						
All Students	10	46	27		1	
Female	2	23	10		1	
Male	8	23	17			
American Indian or Alaskan Native	1	5	2			
Asian			1			
Black or African American	3	1	1			
Hispanic or Latino		5	4			
White	5	32	18		1	
Native Hawaiian or Other Pacific Islander						
Multi-Racial	1	3	1			
LEP		3	1			
Plan504		1				
Grade 11						
All Students	7	38	28			1
Female	2	12	7			1
Male	5	26	21			
American Indian or Alaskan Native	2	7	4			
Asian						
Black or African American		4				
Hispanic or Latino	1	2	2			
White	4	23	21			1
Native Hawaiian or Other Pacific Islander						
Multi-Racial		2	1			
LEP		2				
Plan504		1				

Note. AUT = Autism; ID = Intellectual Disability; MD = Multiple Disabilities; TBI = Traumatic Brain Injury; OHI = Other Health Impairment; DB = Deaf-blindness.

3.2 SUMMARY OF OVERALL STUDENT PERFORMANCE

Table 8–Table 10 present a summary of the spring 2023 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 (level3 + level4) students. The results were based on the students who met attemptedness requirements for scoring and reporting of the SDSAA.

Table 8. 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	91	295.32	41.65	25	29	35	11	46
Female	26	307.96	32.02	15	19	58	8	65
Male	65	290.26	44.14	29	32	26	12	38
American Indian or Alaskan Native	11	284.36	65.59	18	45	27	9	36
Asian	1*							
Black or African American	5*							
Hispanic or Latino	10	295.8	19.27	30	30	40	0	40
White	59	296.41	42.04	27	24	36	14	49
Native Hawaiian or Other Pacific Islander								
Multi-Racial	5*							
LEP	6*							
Plan504	1*							

* Results for $n < 10$ are suppressed.

Table 9. 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	86	284.02	62.16	16	42	26	16	42
Female	36	285.06	65.67	11	42	31	17	47
Male	50	283.28	60.17	20	42	22	16	38
American Indian or Alaskan Native	8*							
Asian	1*							
Black or African American	6*							
Hispanic or Latino	9*							
White	56	292.93	50.64	9	46	29	16	45
Native Hawaiian or Other Pacific Islander								
Multi-Racial	6*							
LEP	4*							
Plan504	86	284.02	62.16	16	42	26	16	42

* Results for $n < 10$ are suppressed.

Table 10. 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	77	292.38	55.67	17	48	22	13	35
Female	22	287.36	71.84	32	32	18	18	36
Male	55	294.38	48.37	11	55	24	11	35
American Indian or Alaskan Native	13	289.23	13.77	15	62	23	0	23
Asian	1*							
Black or African American	4*							
Hispanic or Latino	5*							
White	51	292.2	57.04	14	49	24	14	37
Native Hawaiian or Other Pacific Islander								
Multi-Racial	3*							
LEP	2*							
Plan504	1*							

* Results for $n < 10$ are suppressed.

3.3 TEST-TAKING TIME

The SDSAA tests are not timed. The time spent on each item may vary among individual students, which may provide useful information about student testing behaviors and motivation. Since the length of a test session could be monitored by proctors who are knowledgeable about their students, additional time for students who need it would be arranged.

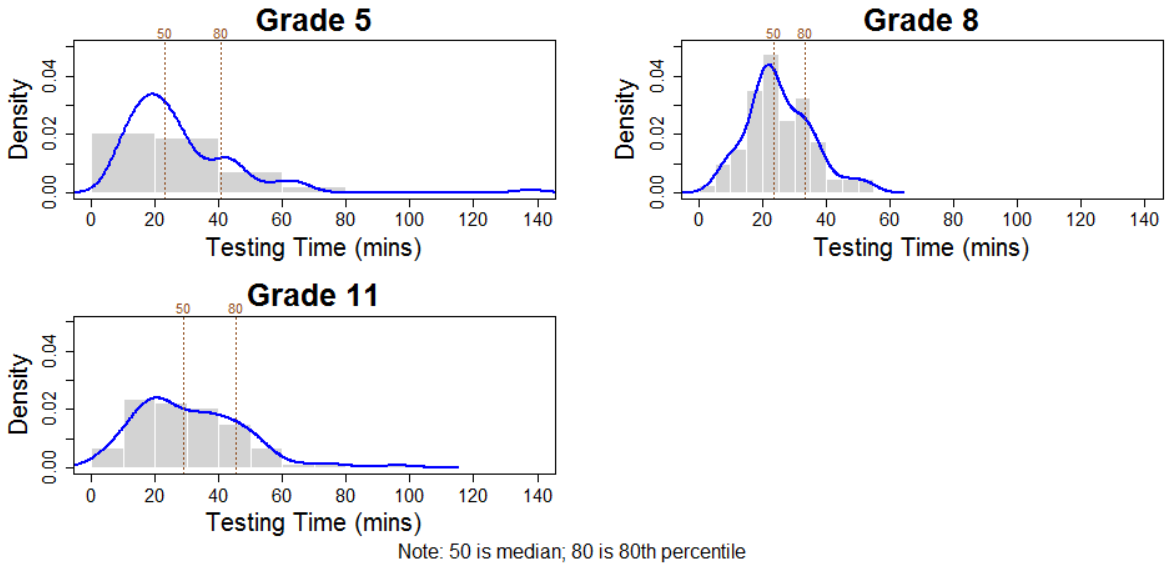
In the Test Delivery System (TDS), item response time is captured as the item page time (the time that a student spends on each item page) in milliseconds. Discrete items appear on the screen one item at a time, and items associated with a stimulus appear on the screen together with the page time measured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

Table 11 presents an average testing time and the testing time at various percentiles for the overall test. The distribution of testing time is provided in Figure 1.

Table 11. 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	00:28	00:23	00:07	00:35	00:41	00:43	00:46	02:18
8	00:25	00:23	00:05	00:32	00:34	00:36	00:38	00:53
11	00:31	00:29	00:04	00:41	00:46	00:50	00:51	01:36

Figure 1. Distribution of Testing Time

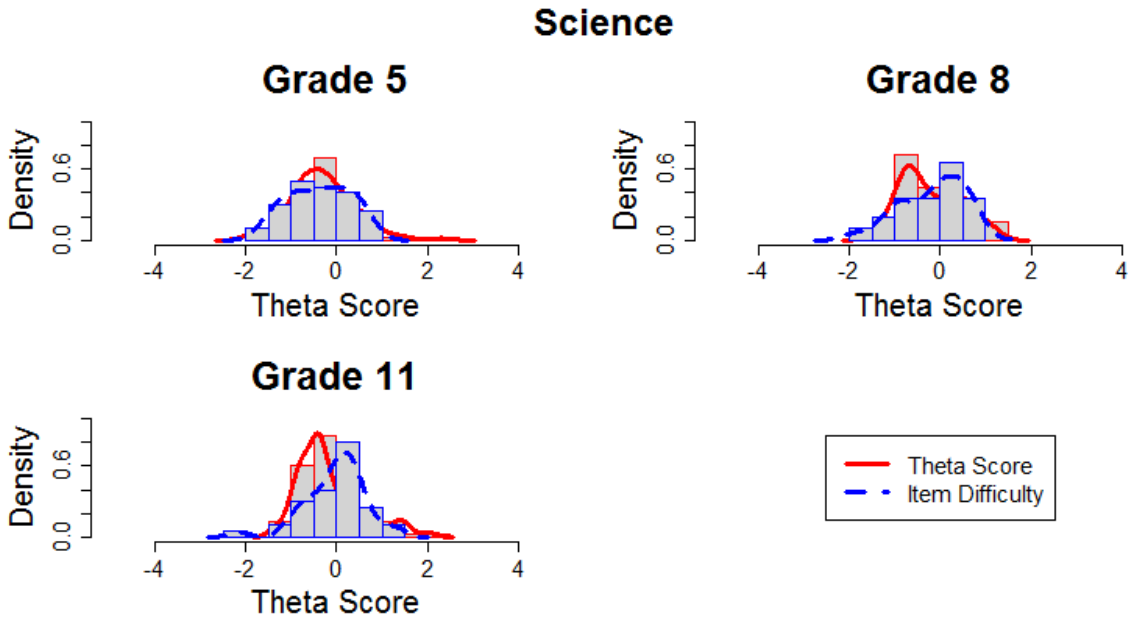


Note: 50 is median; 80 is 80th percentile

3.4 Distribution of Student Ability and Item Difficulty for SDSAA

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

Figure 2. Student Ability–Item Difficulty Distribution for SDSAA



4. ITEM DEVELOPMENT

4.1 ITEM DEVELOPMENT FOR THE MOU-ALT

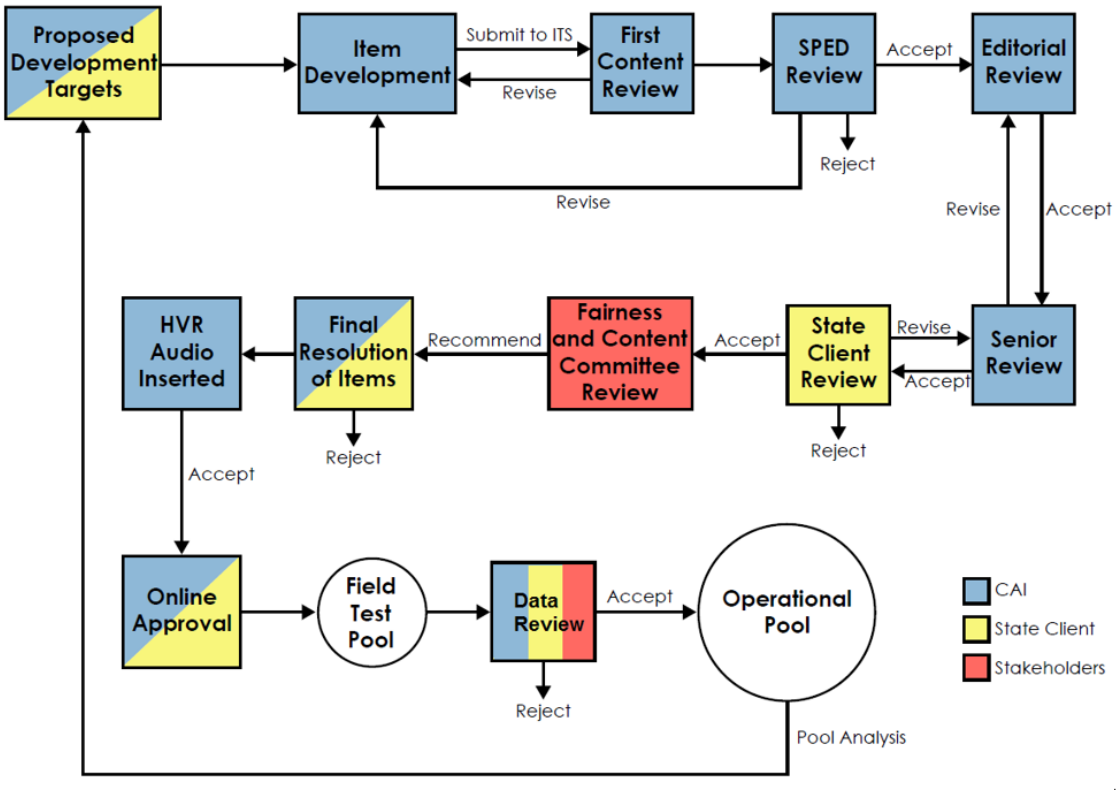
South Carolina, Hawaii, and Wyoming signed a Memorandum of Understanding (MOU) on item sharing in item development and field testing in 2018. Each state contributed a predetermined number of items proportional to their state's student population for the alternate assessment. In early 2019, Idaho and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

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

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

The development process begins with establishing CAI's proposed development targets and working with the individual states to edit the development targets and accept a final plan. The CAI content team then starts item development. After the initial round of development, the items go through 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 items are accepted by the special education reviewer, the items go through an editorial review. After editorial review, the items go back through a senior review as the last step of review at CAI before the items are sent to each state for the client review. At this step, the client accepts the items, recommends edits, or rejects the items. After client comments are resolved, all accepted items are then taken to a stakeholder Content and Fairness Committee review. After the Content and Fairness Committee makes its recommendations, the states and CAI go through a final edit resolution. The items then go through an approval step in which CAI verifies that the items will appear on the test as expected. Items are then moved into the field-test item pool and are field tested. After the testing window, all the field-test items are reviewed in an item data review meeting within each MOU state. This is followed with a stakeholder item data review meeting across all MOU states. At either item data review meeting, items may be accepted and/or rejected by each individual state. The accepted items are then moved into the operational item pools for each individual state.

Figure 3. Alternate Assessment Item Development Process



4.1.1 Item Type and Scoring Rubrics

The MOU shared field test item pool has multiple-choice (MC) items and multi-select (MS) items. The MC items have 2–4 options with one key. The MS items have up to five options with two keys. For MC items, if a student selects the key, he or she receives one point; otherwise, the student receives zero points. For MS items, if a student selects two keys, he or she earns two points; if the student selects one key, he or she earns one point; otherwise, the student earns zero points. Each item measures a specific content standard. Items were written to five difficulty levels: low, low-medium, medium, medium-high, and high. The final item difficulties are determined through field testing.

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

Starting in late Spring 2018, cog labs were conducted in each of the original three states to determine if certain types of technology-enhanced items should be developed for the shared field test items. The item types included multi-select, equation editor, table match, and animations. Neither equation editor or table match proved to be a successful item type for this population of students and therefore states will not develop any more of these item formats. MS items were successful for high functioning middle school and high school students and will continue to be developed for this segment of the Alternate Assessment population.

4.1.2 Item Development Procedure and Item Reviews

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

In each state, items owned by the state that were accepted by the state were prepared for review by a state-wide Content and Fairness Committee convened for each content area in each state. The Content and Fairness Committee was composed of stakeholders from around the state with teaching experience in grades K–12 and experience working with students with disabilities. Additional stakeholders with expertise in specific disability categories, stakeholders with multi-cultural/foreign language expertise, and stakeholder parents were invited to participate in the committee meetings. These specific stakeholders reviewed the items and provided feedback to ensure that all accepted items are correct and are free from fairness and bias issues. Most importantly, these educators made sure that this population of students will be able to understand the language used in the items and that the included visuals and audio directions will aid and not distract students.

Following the committee reviews in these states, the accepted items were then shared across the state item banks for field testing. See Figure 3 for a flowchart that documents the development process.

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

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

- CAI Internal Group Review: Prior to making any changes to draft items, content and senior reviewers meet to discuss items and determine revisions to content, alignment, and style.
- CAI Internal Preliminary Review: Following the group review, the preliminary review is conducted by a member of CAI's content team assigned to the Alternate Assessments. Items are revised, as agreed upon in the group review, to eliminate initial errors, meet content standards, and meet internal style and clarity expectations.
- CAI Internal Content Review: A second content review occurs after the preliminary review to further ensure changes based on the group review, and to revise items further, as necessary, to address any content, alignment, clarity, accessibility, and errors.
- Special Education Review: At this stage, items are reviewed by an CAI special education expert. The expert reviews and revises the items to ensure that they not only meet the content standards but are also as accessible as possible to students across a wide spectrum of cognitive and physical disabilities. When appropriate, the special education expert designates items as "Access Limited," meaning that a task is inappropriate to administer to students with a specific physical disability (e.g., blindness). If revisions are required, the special education reviewer will send items back to the content reviewer to implement changes.

- Edit Review: After the special education reviewer approves items, they send them through an editorial review. At this stage, an 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, an CAI senior content specialist reviews all items to ensure that they meet the content standards, they are free of typographical and technical errors (e.g., key check, spelling error check), and the previously requested edits are in place.
- CAI Batch Review: This is the last step in the CAI internal review process and is designed as a final quality control check to ensure the items are ready for state review.

State Review: At this level, items are compared to the extended and prioritized standards, state standards, and state content specifications. The items are also reviewed against the achievement level descriptors at all difficulty levels and compared to the blueprint. Items are further reviewed to ensure that they align to the support guides for each subject area. At this stage, state staff review each item and make the following decisions:

- accept without modification ("Accept as Appears")
- request minor revisions ("Accept as Revised")
- request substantial changes and resubmit for a second 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: Following revisions and state approval, items are brought to the Content and Fairness Committee for further review. The review committee includes special educators, general educators, vision and hearing specialists, school principals, special education directors, parents of special education students, and university professors with expertise in special education. The review committees represent a diversity of gender, ethnicity, disability, race, and cultural subgroups across the state. During the review meeting, each item is reviewed and assured to that it meets bias and sensitivity guidelines, is aligned to content standards, and is determined to abide by the principles of universal design (UD).

The common criteria used for item review are:

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

At the beginning of each meeting, a CA special education item development specialist provides a training session to ensure that the committee members understand the expectations and are familiar with the training materials that encompass the pertinent content and bias guidelines. Because the MOU shared items are used in each state for its online assessment, the committee members conduct the review online to see the item just as the student will see it.

4.1.3 Development of Crosswalk and State Alternate Content Standards

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

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

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

The crosswalk then informed the development of item specifications. Each item specification included the General Education standard, followed by the state-specific alternate standards that align to the general education standard. The item specifications also included content extensions at four different levels, from Level 1 to Level 4.. The language of the content extensions was derived from each state's standards and ALDs/PLDs, where applicable, and synthesized in an effort to drive items that aligned to multiple states. Once completed, the item specifications were sent to each state for review to confirm alignment and overall approach.

The states' content standards were further analyzed to cull relevant concepts, skills and vocabulary. Based on MOU state feedback, these were compiled and displayed in the form of a and Vocabulary matrix, revealing which concepts, skills, and vocabulary were relevant to each state. The intent was to provide an "at-a-glance" perspective on content standard overlap across the states. The states' content standards were also analyzed to reveal state specific and cross-state content limits in the content extensions. These were listed in the Content Limits section.

All the above analysis was then used to create 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 SEP and CC for the associated content standard.

4.2 FIELD-TESTING

Items that survived Content and Fairness Committee review were field-tested in the spring 2023 test administration as embedded field-test items. The 2023 alternate assessment science tests were administered online using a CAT design in all MOU states except the SDSAA which had a fixed-form design. The CATs

were assembled using CAI's adaptive testing algorithm. The adaptive item selection algorithm selected items based on their content value and information value.

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

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

The spring 2023 field-test item pool was consisted of items that were shared across MOU-Alt states and the items that were unique within each state. The field-test items shared across MOU states were administered in all MOU states while the state only field-test items were administered in the state only. The spring 2023 item pool is summarized in Table 12.

Table 12. Summary of the 2023 Field-Test Item Pool

Grade	State-Only	MOU						Total
	SD	HI	ID	MT	SC	SD	WY	
ES		22	10	5		7	30	74
MS		21	6	2		6	4	39
HS	3	8	7	3	1	7	20	46

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

4.2.1 Item Statistics

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

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

4.2.2 Classical Statistics

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

Items that were either extremely difficult or extremely easy were flagged for review but not necessarily rejected if they aligned with the test and content specifications. For dichotomous items, the proportion of test takers in the sample selecting the correct answer (*p*-value) was computed, as well as those selecting the incorrect responses. For items with 0-2 score points, item difficulty was calculated both as the item's mean score and as the average proportion correct (analogous to *p*-value and indicating the ratio of an item's mean score divided by the number of points possible). Items were flagged for review if the *p*-value was less than .25 or greater than .95.

The item discrimination index indicated 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 was able to differentiate between high- and low- achieving students. The discrimination index for dichotomous items was calculated as the correlation between the item score and the student's IRT-based ability estimate. For polytomous items, we computed the mean total number correct for student scoring within each of the possible score categories. Items were flagged for subsequent reviews if the biserial correlation for the keyed (correct) response was less than .20.

Distractor analysis for the dichotomous items was used to identify items that had marginal distractors or ambiguous correct responses. The discrimination value of the correct response should be substantial and positive, and the discrimination values for distractors should be lower and, generally, negative. The biserial correlation for distractors was the correlation between the item score, treating the target distractor as the correct response, and the student's IRT ability estimate, restricting the analysis to those students selecting either the target distractor or the keyed response. Items were flagged for subsequent reviews if the biserial correlation for the distractor response was greater than .05.

The flagging criteria based on classical item analysis are summarized in Table 13.

Table 13. Thresholds for Flagging in Classical Item Analysis

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

4.2.3 Item Response Theory Statistics

Rasch and Masters' Partial Credit Model were used to estimate the item response theory (IRT) model parameters for dichotomously and polytomously scored items, respectively. The Winsteps output showed the item statistics resulting from anchoring the field-test items on the operational items. Item fit was evaluated via the mean square Infit and mean square Outfit statistics reported by Winsteps, which were based on weighted and unweighted standardized residuals for each item response, respectively. These residual statistics indicated the discrepancy between observed item responses and the predicted item responses based on the IRT model. Both fit statistics had an expected value of 1. Values substantially greater than 1 indicated model underfit, while values substantially less than 1 indicated model overfit (Linacre, 2004). Items were flagged if Infit or Outfit values were less than 0.5 or greater than 2.0.

4.2.4 Analysis of Differential Item Functioning

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

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

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

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

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

Items were classified into three categories (A, B, or C), ranging from no evidence of DIF to severe DIF according to the DIF classification convention listed in Table 14. 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, or female), or negative DIF (i.e., –A, –B, or –C), signifying that the item favors the reference group (e.g., white or male). Items were flagged if their DIF statistics fall into the "C" category for any group. A DIF classification of "C" indicates that the item shows significant DIF and should be reviewed for potential content bias, differential validity, or other issues that may reduce item fairness. DIF classification rules are presented in Table 14. Because of the unreliability of the DIF statistics when calculated on small samples, caution must be used when evaluating DIF classifications for items where focal or reference groups are less than 200 students (Mazor, Clauser, & Hambleton, 1992, Camilli & Shepard, 1994, Muniz, Hambleton, & Xing, 2001, Sireci & Rios, 2013).

Table 14. DIF Classification Rules

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

4.2.5 Summary of Item Statistics

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

Table 15. 2023 MOU Item Sample Size Distribution

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

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

Table 16. Summary of Item Analyses Results for MOU-Alt Science

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

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

4.2.6 Data Review Meeting

4.2.6.1 MOU-Alt Item Pool

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

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

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

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

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

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

4.3 SCALING AND EQUATING

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

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

where *Z* represents the pattern of item responses, and 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 Rasch family Masters' (1982) partial credit model. Under Masters' partial credit model, the probability of getting a score of x_i on item i given ability 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.

4.3.1 Item Calibration

The field-test items were calibrated anchoring on the operational item parameters. All completed records were included in IRT analysis. Through this anchoring process, field-test item parameter estimates were placed on the same MOU scale as the operational items. These operational item parameters will be used to calibrate new field-test items in the following years.

Winsteps was used to estimate Rasch and Masters' partial credit model item parameters for the MOU-Alt. Winsteps is a publicly available software program from Mesa Press. Winsteps employs a joint maximum likelihood approach towards estimation (JMLE), which jointly estimates the person and item parameters. The Rasch model estimates the parameters for student responses to dichotomous (0/1 point) items. Masters'

(1982) partial credit model, an extension of the one parameter Rasch model which allows for partial credit to be given on items, estimates the responses for polytomous items.

5. VALIDITY

According to the *Standards for Educational and Psychological Testing* (AERA, APA, & 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 interpretation for specified uses, and thus, the validation process starts logically with well-articulated statements on intended uses of test scores. Arguments of logic, theoretical, and empirical evidence are then provided to support the intended uses.

This chapter will first present the statements on intended uses of the SDSAA test scores, followed by various sources of evidence validating the interpretation of test scores for the intended uses.

5.1 INTENDED USES AND INTERPRETATIONS OF THE SDSAA SCORES

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

The intended uses of the SDSAA score include

- measuring students' academic achievements and progress in core content areas taught in school,
- measuring achievement and progress toward meeting the state performance standards, and
- monitoring the education system and make necessary improvement to meet federal accountability requirements.

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

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

5.2 SOURCES OF VALIDITY EVIDENCE

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

Providing sufficient and solid validity evidence is also required of the State to meet federal peer review requirements. In the guidance provided by the United States Department of Education for assessing peer

review process (U.S. Department of Education, 2018), the requirements related to validity are represented by three critical elements.

Validity evidence for the SDSAA are gathered from the following four sources, as outlined in the Standards. The particular critical element in the peer review guidance corresponding to each source is included in the parathesis.

- Evidence based on test content
  (Critical Element 3.1- Overall Validity, Including Validity Based on Content)
- Evidence based on response processes
  (Critical Element 3.2- Validity Based on Cognitive Process/Linguistic Processes)
- Evidence based on internal structure
  (Critical Element 3.3- Validity Based on Internal Structure)

Evidence on test content validity is provided with both theorical and empirical evidence related to content specifications, 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 response to items. Evidence on internal structure is examined in the results of intercorrelations among content strand scores.

5.2.1 Evidence Based on Test Content

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

- Content specifications
- Test blueprint
- Item development
- Test administration conditions, and
- 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.

5.2.1.1 Content Specifications

Content standards and content specification is 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 the content standards for science, which are linked to the Core Content Connectors (CCCs). 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. The

CCCs do address science and engineering practices (SEPs), disciplinary core ideas (DCIs), and crosscutting concepts (CCs) from the standards. To further break down the big 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 and Range Achievement Level Descriptors (ALDs). For more details, see Section 1.3 - Content Standards in this technical report.

5.2.1.2 Test Blueprint

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

For the SDSAA, each student receives 40 operational items and 10 field-test items from the MOU pool. Only operational items contribute to student scores. In spring 2023, a fixed form for operational items was created for each grade that met all requirements of the SDSAA blueprints, as shown in Table 18.

Table 18. 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%

5.2.1.3 Item Development

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

5.2.1.4 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 Test Administration Manual (TAM). See Chapter 2 - Test administration in this technical report for details.

5.2.1.5 Item and Test Scoring

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

5.2.2 Evidence Based on Response 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 Content Standards. Cognitive lab studies conducted in each state explored student performance on items aligned to the state standards in knowledge and skill level. The results of these studies demonstrated students' application of their knowledge and skills.

Students with significant cognitive disabilities represent about 1% of a state's total assessed population. The students who participate in the alternate assessments for students with significant cognitive disabilities represent a variety of disability categories and demonstrate many concomitant learning difficulties. Students in this population can exhibit difficulties responding to stimuli; committing information to working, short-term, or long-term memory; generalizing learning to familiar and novel environments; 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 had an opportunity to implement a cognitive lab study yet, results from the cognitive labs in other MOU states who share testing items can also provide insights.

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

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

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

5.2.3 Evidence Based on Internal Structure

The measurement and reporting model used in the 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 correlations among content strand scores are presented in Table 19. The correction for attenuation indicates what the correlation would be if strand scores could be measured with perfect reliability and corrected (adjusted) for measurement error estimates.

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

When corrected for attenuation (above diagonal), the correlations among strand scores are higher than observed correlations. Note that disattenuated correlation is set to one if the disattenuated correlation is greater than one. The values on the diagonal are reliabilities for strand scores. If the reliability is negative, correlations cannot be computed.

Table 19. SDSAA Correlations Among Strands

Grade	Strand	Observed & Disattenuated Correlation		
		Strand 1	**Strand 2**	**Strand 3**
5	Strand 1: Earth and Space Science	0.40	0.92	0.91
	Strand 2: Life Science	0.46	0.63	0.84
	Strand 3: Physical Science	0.42	0.48	0.52
8	Strand 1: Earth and Space Science	0.34	0.90	0.98
	Strand 2: Life Science	0.36	0.49	0.90
	Strand 3: Physical Science	0.44	0.49	0.60
11	Strand 1: Earth and Space Science	0.62	1	1
	Strand 2: Life Science	0.57	0.45	1
	Strand 3: Physical Science	0.50	0.60	0.37

6. RELIABILITY

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

The reliability evidence of the SDSAA is provided with marginal reliability, SEM, conditional SEM (CSEM), and classification accuracy and consistency for each achievement standard.

6.1 MARGINAL RELIABILITY

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

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

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

where N is the number of students; $CSEM_i$ is the 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 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

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

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

Table 20 presents the marginal reliability coefficients and the average CSEMs for the overall science test scale scores, based on all completed tests, excluding the early stopped tests.

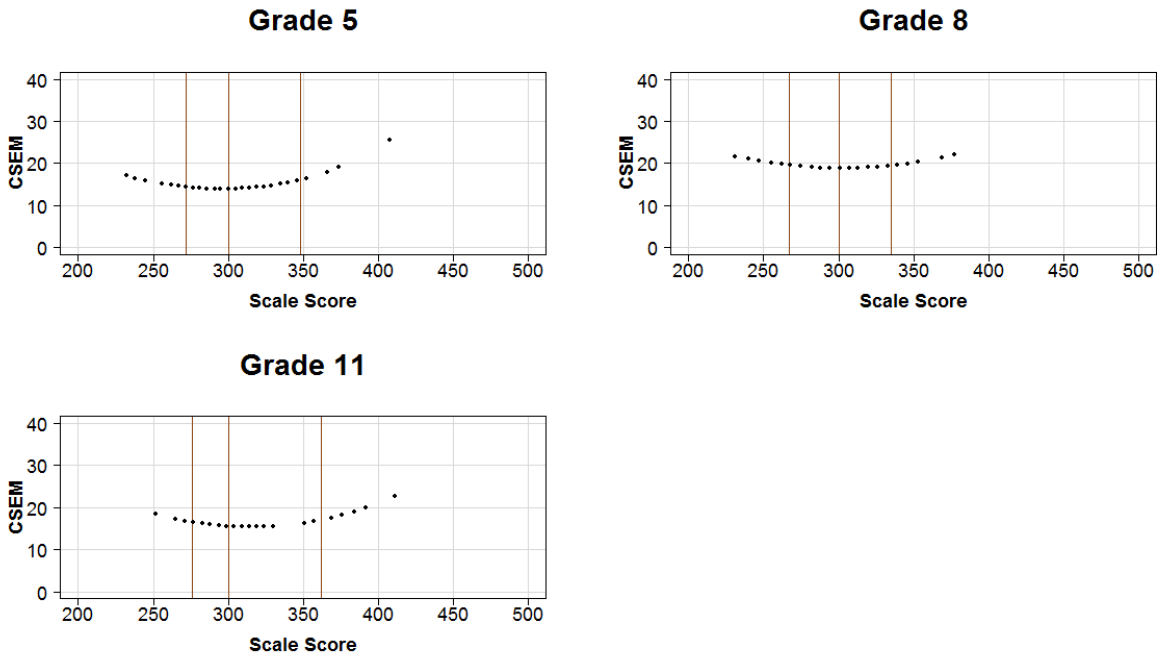
Table 20. Marginal Reliability for SDSAA

Grade	Number of Items	Marginal Reliability	Scale Score Mean	Scale Score SD	Average CSEM
5	40	0.76	300.02	29.82	14.74
8	40	0.72	297.83	37.31	19.67
11	40	0.76	302.92	33.20	16.33

6.2 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.

Figure 4. Conditional Standard Error of Measurement for SDSAA



The SEMs presented in Figure 4 are summarized in Table 21 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 21. Average Conditional Standard of Error Measurement by Achievement level

Grade	Level 1	Level 2	Level 3	Level 4	Average CSEM
5	14.87	14.01	14.28	17.50	14.74
8	20.83	19.39	19.11	20.59	19.67
11	17.28	15.92	15.41	18.37	16.33

6.3 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 on the basis of all sets of items administered across students, using an IRT-based method (Guo, 2006).

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

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

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

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

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

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

If we are interested in only the classification at each cut, *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

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

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

Classification Accuracy

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

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

where $n_{a11} = \sum_{pl_i=\mathrm{below}}(1-p_i)$, 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 $n_{a12} = \sum_{pl_i=\mathrm{below}} p_i$, $n_{a21} = \sum_{pl_i=\mathrm{at\ or\ above}}(1-p_i)$, and $n_{a22} = \sum_{pl_i=\mathrm{at\ or\ above}} p_i$. In the above table, the row represents the observed level and the column represents the expected level.

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

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

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

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

and the overall classification accuracy for the cut is estimated by

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

Classification Consistency

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

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

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

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

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

the classification consistency (*CC*) for the below the cut is estimated by

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

and the overall classification consistency is

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

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

Table 22 shows classification accuracy and consistency indexes for the spring 2023 SDSAA. Accuracy classifications are slightly higher than the consistency classifications in all achievement standards. The consistency classification rate can be somewhat lower than the accuracy rate because consistency assumes two test scores, both of which include measurement error, but the accuracy index assumes only a single test score and a true score, which does not include measurement error.

Table 22. Classification Accuracy and Consistency for Achievement standards

Grade	Accuracy			Consistency		
	Level 2	Level 3	Level 4	Level 2	Level 3	Level 4
5	0.87	0.84	0.94	0.85	0.78	0.94
8	0.81	0.89	0.89	0.79	0.84	0.86
11	0.84	0.80	0.95	0.79	0.74	0.96

6.4 Reliability for 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 23. The reliability coefficients were computed on the basis of completed tests only.

Table 23. Marginal Reliability Coefficients for 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	0.40	302.10	34.32	26.48
	Life Science	13	13	0.63	300.13	47.03	27.64
	Physical Science	14	14	0.52	297.31	37.27	25.45
8	Earth & Space Science	13	13	0.34	296.58	42.61	34.60
	Life Science	13	13	0.49	292.60	49.91	35.45
	Physical Science	14	14	0.60	301.02	58.46	36.00
11	Earth & Space Science	13	13	0.62	293.21	51.33	30.50
	Life Science	12	12	0.45	307.99	41.85	30.70
	Physical Science	15	15	0.38	305.88	34.48	27.03

7. SCORING

For the 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.

7.1 ATTEMPTEDNESS RULES FOR SCORING

If a student logged into the test administration system, was presented one item, and a valid response was entered for that first item, the student is counted as participated. A valid response is recorded when the student marks on one or more response options, or the proctor marks *No Response* on the item. Participated students are counted as attempted.

Scores are generated only for tests with Test Attempted = Y.

- If a student answered all items in Segment 1 and 2, the test was scored without penalty.
- If a student did not complete Segment 1 and 2, the student was scored with penalty. The penalty was the theta estimate minus one CSEM for the estimated theta value.
- If a student had four consecutive NR responses for items within Segment 1, the student was given the lowest obtainable score of the test. The SEM and theta score was set to BLANK.

Table 4 in Section 3.1 lists the number of "completed" tests without scoring penalty, the number of "Incomplete" tests with scoring penalty, and the number of early stopping rule (ESR) tests receiving the lowest obtainable scale score.Estimating Student Ability Using Maximum Likelihood Estimation

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

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

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

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

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

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

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

in the case of items with two or more points,

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

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

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

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

7.1.1 Standard Error of Measurement

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

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

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

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

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

7.2 Rules for Transforming Theta to Scale Scores

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

Table 24 provides the linear transformation constants, intercept, and slope values with four decimals. For the score reports, the scale scores computed applying the slope and intercept for individual students will be rounded to the nearest integer.

Table 24. Scaling Constants on the Reporting Metric

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. The scale scores are mapped onto four achievement levels. Table 25 provides the range of scale scores at each achievement level by grade.

Table 25. Range of Scale Scores at Each Achievement Level by Grade

Grade	Does Not Meet	Approaching	Meets	Exceeds
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

7.3 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. For the overall scale scores, scale scores lower than 100 or higher than 500 are truncated to 100 or 500. The standard error for LOSS and HOSS is computed using the estimated theta scores based on the responded items.

7.4 Scoring All Correct and All Incorrect Cases

With item response theory (IRT) maximum likelihood (ML) ability estimation methods, 0 and perfect scores are assigned the ability of minus and plus infinity. All incorrect tests are scored by adding 0.3 to an item score among the administered operational items for a test. All correct tests are scored by subtracting 0.3 from an item score among the administered operational items for a student.

8. ACHIEVEMENT STANDARDS

After the spring 2022 operational administration, formal standard setting workshops were conducted in all three grades to recommend achievement standards for the SDSAA. The standard setting results replaced the interim achievement standards derived using a statistical linking method in the spring 2021 administration.

In July 2022, following the close of the testing window, Cambium Assessment, Inc. (CAI) under contract to South Dakota Department of Education (SDDOE), invited a panel of 18 teachers and administrators to recommend achievement standards (new cut scores) for the assessments. 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).

8.1 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.

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

2. **Range** ALD**s.** These ALDs, also called Instructional ALDs, describe what students know and are able to 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. Cambium took these and matched them to the appropriate CCCs and created the remaining ALDs for the remaining CCCs. In July of 2020, Cambium 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 prior to 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 prior to standard setting, creating the final document that was then reposted to the SDDOE website.

3. **"Just Barely"** ALD**s.** 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** ALD**s.** 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.

8.3 RECOMMENDED ACHIEVEMENT STANDARDS

Panelists were tasked with recommending three achievement standards that resulted in four achievement levels. Table 26 presents the achievement standard in scale score metric associated with the percentage of students reaching each standard based on the 2022 SDSAA results.

Table 26. 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%

9. REPORTING AND INTERPRETING SCORES

The Centralized Reporting System (CRS) 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. Score reports from the Spring 2023 assessment were provided to districts and students through the CRS, beginning on March 27, 2023. The CRS provided 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 have access to students' performance scores and use them to improve student learning. In addition to individual student reports (ISRs), the CRS also produces aggregate score reports by class, school, and state. The timely accessibility of aggregate score reports can help users to monitor students' performance in each grade by subject area and evaluate the effectiveness of instructional strategies; it can also inform 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 CRS and how to interpret and use these scores.

9.1 CENTRALIZED REPORTING SYSTEM FOR STUDENTS AND EDUCATORS

9.1.1 Types of Online Score Reports

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

Once authorized users log in to the CRS, the dashboard page shows overall test results grouped by test family (e.g., grade 5 science, etc.) for all tests that a student has taken. Once the user clicks on a test family, they are taken to a detailed dashboard where the results are shown by test (e.g., grade 8 science, etc.). In addition, when authorized state-level users log in to the CRS and select "State View," the CRS generates a summary of student performance data for a specified test across the entire state.

Generally, the CRS provides two categories of online score reports: (1) aggregate score reports and (2) student score reports. Table 27 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 *Centralized Reporting System User Guide*, accessed via a HELP button on the CRS.

Table 27. 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 schools, district, and states

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

Aggregate score reports at a selected aggregate level are provided for overall students and by subgroup. Users can see student assessment results by any of the subgroups. Table 28 presents the types of subgroups and subgroup categories provided in the CRS.

Table 28. Types of Subgroups

Subgroup	Category
Gender	Male Female
ELL	ELL Not ELL
Disability	With Disability No Disability
Migrant Status	Migrant Not Migrant
Disadvantaged	Disadvantaged Not Disadvantaged
Ethnicity	American Indian or Alaska Native Asian Black or African American Hispanic or Latino Native Hawaiian or Pacific Islander White Two or More Races

9.1.2 Centralized Reporting System

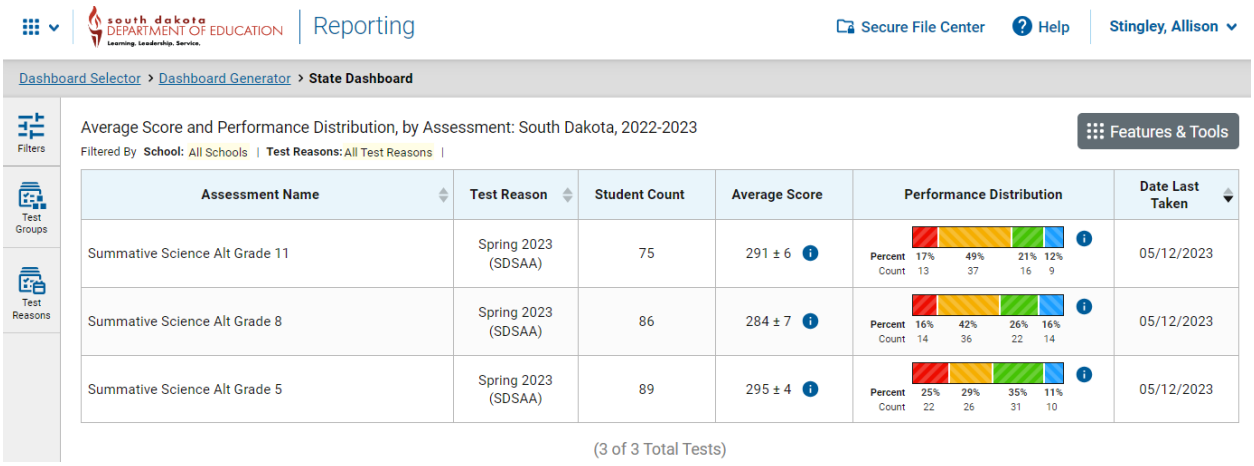
9.1.2.1 Dashboard

The first page users see when they log onto the CRS contains 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 their students. State personnel and district area personnel would need to select the district in order to view the aggregate results.

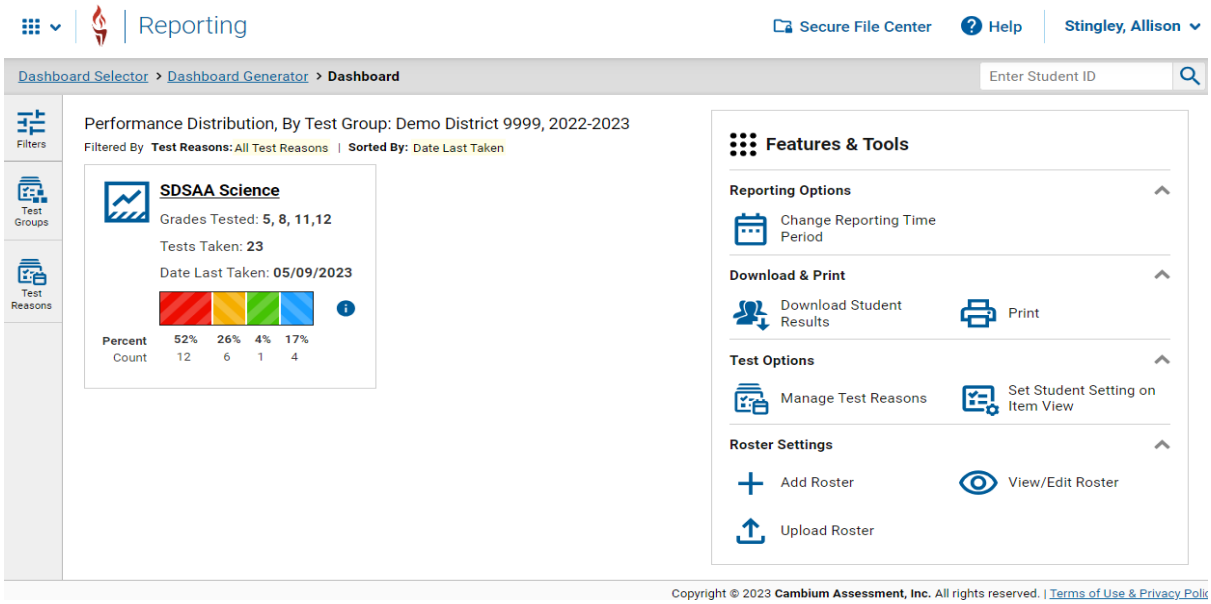
The dashboard summarizes students' performance by test family, including (1) the number of students tested, (2) the grades of the students who have tested, and (3) the percentage and counts of students at each achievement level. Exhibit 1 presents a sampled dashboard pages at the state level.

Exhibit 1. Dashboard: State Level



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

Exhibit 2. Dashboard: District Level



9.1.2.2 Subject Detail Page

Detailed summaries of student performance for each grade in a subject area for a selected aggregate level are presented when users select a specific assessment name. On each aggregate report, the summary report presents the summary results for the selected aggregate unit and the summary results for the state and the aggregate unit above the selected aggregate. For example, if a school is selected, the summary results of the 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.

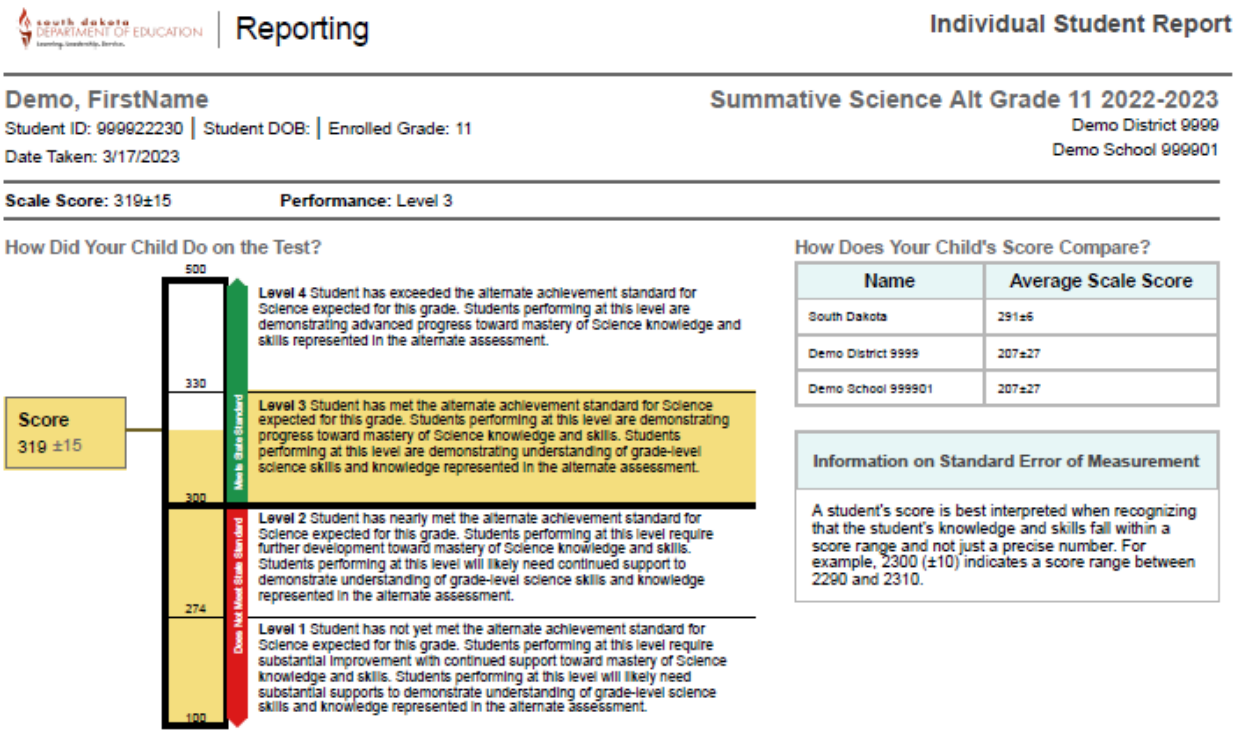
The aggregated subject summary report provides the summaries on a specific subject area, including (1) the number of students tested, (2) the average scale score and standard error (SE) associated with the average scale score, (3) the percentage of proficient students, and (4) the percentage and counts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

9.1.2.3 Student Detail Page

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

Underneath, average scale scores and SEs 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 SE of the average scale scores.

Exhibit 3. Student Detail Page for Science



9.1.2.4 Participation Rate

The Test Information Distribution Engine (TIDE) provides participation rate reports for states, districts, and schools to help monitor the student participation rate. Participation data are updated each time a student completes a test. Included in the participation table are (1) the number and percentage of students who are tested and not tested and (2) the percentage of proficient students.

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

Exhibit 4. Participation Rate Report at the District Level

Test	Total Student	Total Student Started	Total Student Completed	Percent Started	Percent Completed
Science Alt Grade 5	109	92	92	84.40%	84.40%
Science Alt Grade 8	102	84	84	82.35%	82.35%
Science Alt Grade 11	114	97	96	85.09%	84.21%

9.2 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.

9.2.1 Scale Score

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

9.2.2 Standard Error of Measurement

A scale score (observed score on any test) is an estimate of the true score. If a student takes a similar test multiple times, the resulting scale score would vary across administrations, being sometimes a little higher, a little lower, or the same. The 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.

9.2.3 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). Achievement-level descriptors are a description of the content area knowledge and skills that test-takers at each achievement level are expected to possess. Thus, achievement levels can be interpreted based on achievement-level descriptors.

9.2.4 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.

9.3 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 are able to 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 among different groups. Teachers can evaluate how their students performed compared with students in other schools, districts, and the state overall.

Although assessment results provide valuable information to understand student performance, these scores and reports should be used with caution. It is important to note that scale scores reported are estimates of true scores and, therefore, do not represent a precise measure of student performance. A student's scale score is associated with measurement error, and, thus, users need to consider measurement error when using student scores to make decisions about student 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.

10. QUALITY CONTROL PROCEDURES

Quality assurance (QA) procedures are enforced through all stages of the alternate assessment development, administration, and scoring and reporting of results. CAI uses a series of quality control steps to ensure the error-free production of score reports. The quality of the information produced in the test delivery system (TDS) is tested thoroughly before, during, and after the testing window opens.

10.1 OPERATIONAL TEST CONFIGURATION

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

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

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

10.1.1 Platform Review

CAI's TDS supports a variety of item layouts. Each item goes through an extensive platform review on different operating systems such as Windows, Linux, and iOS to ensure that the item looks consistent in all of them. For the 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.

10.1.2 User Acceptance Testing and Final Review

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

10.2 Quality Assurance in Data Preparation

CAI's TDS has a real-time quality-monitoring component built in. After a test is administered to a student, the TDS passes the resulting data to our QA system. QA conducts a series of data integrity checks, ensuring, for example, that the record for each test contains information for each item, keys for multiple-choice items, score points in each item, total number of field-test items and operation items, and that the test record contains no data from items that have been invalidated.

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

10.3 Quality Assurance in Test Scoring

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

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

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 29 presents an overview of the QA reports.

Table 29. 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

10.3.1 Score Report Quality Check

Online Report Quality Assurance

Scores for online assessments are assigned by automated systems in real time. During operational testing, actual item responses are compared to expected item responses (given the item response theory [IRT] parameters), which can detect miskeyed items, item score distribution, or other scoring problems. Potential issues are 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 Centralized Reporting System (CRS), which is responsible for presenting individual-level results and calculating and presenting aggregate results. Absolutely no score is reported in the ORS until it passes all of the QA system's validation checks.

REFERENCES

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

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

Guo, F. (2006). Expected Classification accuracy using the latent distribution. *Practical, Assessment, Research & Evaluation, 11*(6).

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing, *Journal of Educational Measurement, 13*(4), 253–264.

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

Livingston, S. A., & Lewis, C. (1995).Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, *32*, 179–197.

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

Masters, G. N. (1982). *Psychometrika*, *47*, 149.

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

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

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

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

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

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

U.S. Department of Education. (2018). *A State's Guide to the U.S. Department of Education's Assessment Peer Review Process.* Retrieved from https://www2.ed.gov/admins/lead/account/saa/assessmentpeerreview.pdf