North Dakota State Assessment for Science

2021–2022

Volume 2: Test Development

NORTH DAKOTA DEPARTMENT OF PUBLIC INSTRUCTION

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

1. Introduction

The North Dakota Department of Public Instruction (NDDPI) adopted the new North Dakota Science Content Standards in February, 2019. To measure the new standards, the NDDPI and its assessment vendor, Cambium Assessment, Inc. (CAI), developed a new online assessment referred to as the North Dakota State Assessment (NDSA) for Science. The assessment was administered operationally for the first time in 2020–2021. The NDSA for Science measures North Dakota students' science knowledge and skills in grades 4, 8, and 10 as an adaptive online assessment that features several technology-enhanced item types. The assessment's content measures the three-dimensional science standards based on the National Research Council's *A Framework for K–12 Science Education,* published in 2012.

Details about how the assessments were implemented is available in Volume 1 of this annual technical report.

The interpretation, usage, and validity of test scores rely heavily on the test development process. This volume provides details on the test development process for the NDSA for Science that contributes to the validity of the test scores. Specifically, this volume provides evidence to support the following:

- The item specifications contained detailed guidance for the item writers and reviewers to ensure that the science items were aligned to the performance standards (hereafter referred to as standards) they were intended to measure.
- The item development process employed for the NDSA for Science were consistent with industry standards.
- The development and maintenance of the Shared Science Assessment Item Bank, which contains item clusters and stand-alone items that cover the range of measured standards, grade-level difficulties, and levels of cognitive engagement.
- The test design summary/blueprint stipulated the range of operational items from each item type and content category required for each test administration. This document was implemented using the item-selection algorithm for science.

Note that for the science assessments, as outlined in Volume 1, *North Dakota State Assessment for Science 2021–2022 Annual Technical Report*, CAI collaborated with a group of states that share common item-development processes. In addition to developing items for each of those states, CAI developed and maintains the Independent College and Career Readiness (ICCR) item bank, which consists of items developed according to the same principles followed when the items owned by each of the collaborator states were created. This volume of the annual technical report focuses on general test development activities, even though the NDSA for Science draws exclusively from the ICCR item bank. This volume also outlines which ICCR item bank processes deviate from the overall process and how they do so.

For the remainder of this volume, the term *item bank* will refer to all items developed under the Memorandum of Understanding (MOU) among the groups of states unless explicitly stated otherwise.

1.1 Claim Structure

The goals, uses, and claims that the Shared Science Assessment Item Bank and subsequent tests would be designed to support were identified in a series of collaborative meetings held August 22–23, 2016. The overarching goal of those meetings was to support the development of statewide summative assessments using science content that measures the three-dimensional science standards based on *A Framework for K–12 Science Education* (National Research Council, 2012).

To that end, CAI invited four nationally recognized experts and content and assessment leaders from 10 states to participate in authoring the Next Generation Science Standards (NGSS). Two nationally recognized psychometricians also participated.

CAI staff and the participating states collaborated to develop items and test specifications designed to measure the three-dimensional science standards. In general, the item specifications were accompanied by sample item clusters that met those specifications. All specifications and sample item clusters were reviewed by state content experts and committees of educators in at least one of the states.

1.2 Underlying Principles Guiding Development

The Shared Science Assessment Item Bank was established using a highly structured, evidence-centered design (e.g., Mislevy, Behrens, Dicerbo, & Levy, 2012; Mislevy, & Riconscente, 2011). The process began with detailed item specifications. The specifications, discussed in Section 2.2, Item Specifications, described the interaction types that could be used, listed guidelines for targeting the appropriate cognitive engagement, offered suggestions for controlling item difficulty, and provided sample items.

1.2.1 Evidence-Centered Design to Item Development

The design of test items begins with performance expectations[1]. The PE is not a set of standards, but the learning goals for students and what they should be able to do at the end of instruction. The PE statement describes the explicit objectives of instruction. For examples, students may be able to make a claim about the merit of a solution to a problem when the environment changes.

The PE provides the knowledge, skills, and abilities (KSAs) that should be assessed. In the example, the student must be able to show the skills of making a valid claim, examining a solution, and understand certain aspects of environmental change.

Then we consider the assessment questions and tasks. We ask what tasks and/or questions will provide evidence of the student's KSAs. While not formalized, at this step, the task model is defined here. Strict adherence to the science PEs often requires the use of a specific task model

[1] For the definition and examples of performance expectations, see the link to the Next Generation Science Standards: https://www.nextgenscience.org/glossary/performance-expectation-pe.

(e.g., designing an experiment). Likewise, the work product is often specified. If a student is required to design an experiment, then a multiple-choice item will likely be insufficient.

We formalize these steps when creating the item in our authoring tool. Each interaction within an item is scored using scoring assertions, which describe the features(s) of a successful student response as well as the inference we can make about student knowledge and skills from each of those features.

This embodies evidence-centered design since, the assertions describe the expectations for the student response and how that evidence links to the claim about student understanding of the PE.

For example, the student drew a path of light from the candle to the toy, then to the eye. This provides evidence that the student can manipulate components of a model to show how light travels for an object to be seen.

Assertions are used to show how each point awarded to a student response reinforces the evidentiary argument.

1.2.2 Other Principles to Item Development

Item development supported the goal of creating high-quality item clusters and stand-alone items through rigorous development processes that were managed and tracked by a content development platform. This platform ensured that every item flowed through the correct sequence of reviews and captured every comment and change applied to each item.

CAI sought to ensure that the items measured the standards in a fair and meaningful way by engaging educators and other stakeholders at each step of the process. Educators evaluated the alignment of the items to the standards and offered guidance and suggestions for improvement. They participated in the review of items for fairness and sensitivity. Following item field testing, educators engaged in rubric validation, a process that refines rule-based rubrics upon review of student responses.

Combined, these principles and the processes that support them were incorporated into an item bank that measures the standards with fidelity and does so in a manner that minimizes construct-irrelevant variance and barriers to access. The details of these processes are described in this volume of the annual technical report.

1.3 Organization of This Volume

This volume is organized into three subsequent sections:

1. **Item Development Process.** This section outlines the science item development process that supports the validity of the claims that the science tests are designed to support.

2. **Shared Science Assessment Item Bank.** This section describes the Shared Science Assessment Item Bank, the types of assessments the item bank is designed to support, and the methods used to refresh the item bank.

3. **NDSA for Science Test Construction.** This section describes the test construction process for the NDSA for Science, including the blueprint, the test design, an evaluation of the simulated test sessions, the operational blueprint match results, and the item exposure rates.

2. ITEM DEVELOPMENT PROCESS THAT SUPPORTS VALIDITY OF CLAIMS

2.1 OVERVIEW

Cambium Assessment, Inc. (CAI) developed the Shared Science Assessment Item Bank in collaboration with the states that were part of the Memorandum of Understanding (MOU), using a rigorous, structured process that engaged stakeholders at critical junctures. This process was managed by CAI's Item Tracking System (ITS), which is an auditable content-development tool that enforces rigorous workflow and captures all changes made to and comments associated with each item. Reviewers, including internal CAI reviewers or stakeholders in committee meetings, can review items in ITS as they will appear to the student, along with all accessibility features and tools. Please note, the NDSA for Science draws exclusively from the ICCR item pool; the ICCR item pool is part of the larger Shared Science Assessment Item Bank.

The item development process begins with defining item specifications, and continues with

- selecting and training of item writers;
- writing and reviewing items internally;
- conducting item review by state personnel and stakeholder committees;
- marking up items for translation and accessibility features;
- field testing; and
- post-field-test reviews.

Each step plays a vital role in ensuring that the items support the claims on which they are based. Table 1 describes each step in the process and how each step contributes to this goal.

Table 1. Summary of How Each Development Step Supports the Validity of Claims

Developmental Steps	Supports Alignment to the Performance Standards	Reduces Construct-Irrelevant Variance Through Universal Design	Expands Access Through Linguistic and Other Supports
Defining Item specifications	Specifies item interactions and content limits and outlines the guidelines for meeting task demands, levels of cognitive engagement, and the parameters for adjusting difficulty.	Avoids using item interactions with accessibility constraints and provides language guidelines. Allows for multiple response modes to accommodate different styles.	
Selecting and training item writers	Ensures that item writers have the background to understand the performance standards and item specifications. Teaches item writers how to select item interactions for measurement and accessibility.	Training in language accessibility, and bias and sensitivity helps item writers avoid unnecessary barriers.	
Writing items and reviewing them internally	Checks content alignment and evaluates and improves overall quality.	Eliminates editorial issues and flags and removes bias and accessibility issues.	
Marking up items for translation and accessibility features		Adds universal features, such as test-to-speech (TTS) for science, which reduce barriers.	Adds TTS, braille, translations, American Sign Language (ASL), and glossaries.
State personnel and stakeholder committee reviews	Checks content and cognitive complexity alignment; evaluates and improves overall quality.	Flags sensitivity issues.	
Field testing	Provides statistical checks on quality and flags issues.	Flags items that appear to function differently for subsequent review to identify issues.	May reveal usability or implementation issues with markup.
Post-field-test review	Provides final, more focused checks on flagged items. Rubric validation ensures that scoring reflects performance standards.	Provides final, focused review on items flagged for differential item functioning (DIF).	

2.2 ITEM SPECIFICATIONS

CAI collaborates with a group of states and one U.S. territory, psychometricians, and science experts, including the authors of the Next Generation Science Standards (NGSS), to develop powerful innovative solutions to the challenges of measuring three-dimensional science standards based on the National Research Council's *A Framework for K–12 Science Education* published in 2012. Ten participating states, including Connecticut, Hawaii, Idaho, Montana, Oregon, Rhode Island, Utah, Vermont, West Virginia, and Wyoming, were involved in all activities. New Hampshire, North Dakota, South Dakota, and U.S. Virgin Islands participated in some activities. This collaboration has yielded item specifications for the performance standards, sample item clusters for some specifications, and hundreds of science item clusters and stand-alone items in various stages of development. Under this collaboration, these states and the U.S. Virgin Islands developed item specifications jointly using the guidelines proposed by WestEd in conjunction with the Council of Chief State School Officers (CCSSO), state and territory members, and content experts (CCSSO, 2015).

Item specifications are documents designed to guide item writers as they craft test questions and stakeholders as they review those items. These specifications are intended to serve as a roadmap for writers to facilitate the creation of items that are properly aligned to the three dimensions of each science standard and that together form coherent item clusters and stand-alone items.

The item specifications for science include the following elements:

- **Performance Standards.** The performance standards identify the standard being assessed.
- **Dimensions.** The dimensions identify the Science and Engineering Practices (SEPs), Crosscutting Concepts (CCCs), and Disciplinary Core Ideas (DCIs) that each standard assesses.
- **Clarifications and Content Limits.** The clarifications and content limits delineate the specific content that the standard measures and the parameters in which items must be developed to assess the standard accurately, including the lower and upper complexity limits of the items. Specifically, content limits refine the intent of the standard and provide limits of what may be asked of test takers. For example, content limits may identify the specific formulae that students are expected to know or not know.
- **Science Vocabulary.** The science vocabulary outlines the relevant technical words that students are expected to know, and the related words that they are explicitly not expected to know. These categories should not be considered exhaustive, as the boundaries of relevance are ambiguous, and the list is limited by the writers' imaginations.
- **Content/Phenomena.** Examples of the types of content/phenomena that support creating effective items related to a specific standard are provided. In general, these are guideposts, and item writers would seek comparable phenomena rather than drawing on those within the documents.

- **Task Demands.** The task demands denote the specific ways students will be expected to provide evidence of their understanding of the concept or skill. Specifically, the task demands identify the types of interactions and activities that item writers should employ. The standards and associated evidence statements are broken down into specific task demands aligned to each standard. Item writers are required to clearly link each item to one or more task demand, and the verbs guide the types of interactions writers might employ to elicit the student response.

Table 2 provides a sample of the item specifications developed by content experts for a middle school Life Sciences performance standard.

Table 2. Sample Science Item Cluster Specifications for Middle School Life Sciences Performance Standard

<table>
<tr><td>Performance Standard</td><td colspan="3">MS-LS1-1[a]
Conduct an investigation to provide evidence that living things are made of cells; either one cell or many different numbers and types of cells.</td></tr>
<tr><td>Dimensions</td><td>Planning and Carrying Out Investigations
• Conduct an investigation to produce data to serve as the basis for evidence that meets the goals of an investigation.</td><td>LS1.A: Structure and Function
• All living things are made up of cells, which is the smallest unit that can be said to be alive. An organism may consist of one single cell (unicellular) or many different numbers and types of cells (multicellular).</td><td>Scale, Proportion, and Quantity
• Phenomena that can be observed at one scale may not be observable at another scale.</td></tr>
<tr><td>Clarifications and Content Limits</td><td colspan="3">Clarification Statements
• Emphasis is placed on developing evidence that living things are made of cells, distinguishing between living and non-living things, and understanding that living things may be made of one cell or many varying cells.
Content Limits
• <u>Students do not need to know the following</u>:
o The structures or functions of specific organelles or different proteins
o Systems of specialized cells
o The mechanisms by which cells are alive
o Specifics of DNA and proteins or of cell growth and division
o Endosymbiotic theory
o Histological procedures</td></tr>
<tr><td>Science Vocabulary Students Are Expected to Know</td><td colspan="3">Multicellular, unicellular, cell, tissue, organ, system, organism hierarchy, bacteria, colony, yeast, prokaryote, eukaryote, magnify, microscope, DNA, nucleus, cell wall, cell membrane, algae, chloroplast(s), chromosome, cork</td></tr>
</table>

Science Vocabulary Students Are Not Expected to Know	Differentiation, mitosis, meiosis, genetics, cellular respiration, energy transfer, RNA, protozoa, amoeba, histology, protists, archaea, nucleoid, plasmid, diatoms, cyanobacteria
Phenomena	
Context/ Phenomena	Some example phenomena for MS-LS1-1 include the following: • Plant leaves and roots have tiny, box-like structures that can be seen under a microscope. • Small creatures can be seen swimming in samples of pond water viewed through a microscope. • Different parts of a frog's body (e.g., muscles, skin, tongue) are observed under a microscope and are seen to be composed of cells. • One-celled organisms (e.g., bacteria, protists) perform the eight necessary functions of life, but nothing smaller has been seen to do this. • Swabs from the human cheek are observed under a microscope. Small cells can be seen.
This Performance Standard and Associated Evidence Statements Support the Following Task Demands	
Task Demands	
1. Identify from a list the materials/tools, including distractors, needed for an investigation to find the smallest unit of life (cell).	
2. Identify the outcome data that should be collected in an investigation of the smallest unit of living things.	
3. Evaluate the sufficiency and limitations of data collected to explain that the smallest unit of living things is the cell.	
4. Make and/or record observations about whether the sample contains cells.[b]	
5. Interpret and/or communicate data from the investigation to determine if a specimen is alive.	
6. Construct a statement to describe the overall trend suggested by the observed data.	

[a]MS-LS1-1 is the performance standard code for Middle School Life Sciences 1-1.
[b]Denotes task demands deemed appropriate for use in stand-alone item development.

The specifications help test developers create item clusters and stand-alone items that will support a range of difficulty, furthering the goal of measuring the full range of performance found in the population, but remaining at grade level.

2.3 Selecting and Training Item Writers

All CAI item writers developing science items have at least a bachelor's degree, and many have teaching experience. All item writers are trained in

- the principles of universal design;
- the appropriate use of item interactions; and

- the science item specifications.

Key materials are shown in Appendix A, Item Writer Training Materials, Appendix B, Item Specifications Grade 3 through High School, and Appendix C, Style Guide for Science Items. These include

- CAI's Language Accessibility, Bias, and Sensitivity Guidelines
- a training (presented using Microsoft PowerPoint) for the appropriate use of item interactions
- Item Specifications for Science for Grades 3 through High School
- Style Guide for Science Items

2.4 Internal Review

CAI's test development structure employs highly effective units organized around each content area. Unit directors oversee team leaders who work with team members to ensure item quality and adherence to best practices. All team members, including item writers, are content-area experts. Teams include senior content specialists who review the items before the client review phase and provide training and feedback for all content-area team members.

ICCR and MOU science items undergo a rigorous, multiple-level internal review process before they are sent for external review. Staff members are trained to review items for both content and accessibility throughout the process. A sample of the item review checklist used by CAI's test development specialists is available in Appendix D, Item Review Checklist.

The ICCR item bank and MOU science items' internal review cycle includes the following phases:

- Preliminary Review
- Scoring Entry and Review
- Content Review One
- Edit Review
- Senior Review

2.4.1 Preliminary Review

Team leads or senior content staff conduct the Preliminary Review. Sometimes, Preliminary Review is conducted in a group setting, led by a senior test developer. During the Preliminary Review process, team leads or senior content staff analyze items to ensure the following requirements have been met:

- The item aligns with the performance standard.
- The item matches the item specification for the skills being assessed.

- The item is based on a quality scientific phenomenon (i.e., it assesses something in a reasonable way, and it is a discrete observation that grounds a scenario that allows for the assessment of something worthwhile in a meaningful way).
- The item aligns appropriately with the task demands.
- The vocabulary used in the item is appropriate for the grade and subject matter.
- The item considers language accessibility, bias, and sensitivity.
- The content is accurate and straightforward.
- The graphic and stimulus materials are necessary to answer the question.
- The item follows the approved style guide.
- The stimulus is clear, concise, and succinct (i.e., it contains enough information to convey what is being asked, it is stated positively, and it does not rely on negatives—such as *no, not, none,* or *never*—unless necessary).

For selected-response item interactions, test developers also check to ensure that the set of response options are

- as succinct and short as possible (without repeating text);
- parallel in structure, grammar, length, and content;
- sufficiently distinct from one another;
- all plausible (but with only one correct option); and
- free of obvious or subtle cueing.

2.4.2 Scoring Entry and Review

During Scoring Entry, the item writer submits the machine scoring to the team lead or senior staff for review before the Content Review One level. This step is separate from Preliminary Review and allows senior staff to suggest changes to the interaction at Preliminary Review without requiring the writer to overhaul the scoring they have already created. This step also allows senior staff to ensure that the scoring suggested by the writer at the Preliminary Review phase is appropriate. This step streamlines the process by ensuring that the scoring is entered once. At this level, scoring is analyzed to ensure the following criteria are met:

- The scoring works as intended (i.e., the student receives a point for ALL correct responses and no points for ALL incorrect responses).
- The student receives a point for every unique piece of information they reveal about their understanding through their responses.
- Dependent scoring between and within interactions is captured.

- The scoring system is unambiguous and matches the questions asked (i.e., if students are asked to round a number to a particular decimal place, they are scored accordingly).

The senior staff approves the intent of the scoring from the Preliminary Review. At the Scoring Entry level, the writer inputs this approved scoring, after which senior staff checks the functionality of the scoring. Once the scoring is determined to be working correctly, the senior staff signs off on the item and moves it to Content Review One.

2.4.3 Content Review One

Content Review One is conducted by a senior content specialist who was not part of the Preliminary Review. This reviewer carefully examines each item based on the same criteria identified for Preliminary Review. They also ensure that the revisions made during the Preliminary Review did not introduce errors or content inaccuracies. This reviewer approaches the item from the perspective of potential clients and their expertise in test development.

2.4.4 Edit Review

During Edit Review, editors have four primary tasks:

1. Editors perform basic line editing for correct spelling, punctuation, grammar, and mathematical and scientific notation, ensuring consistency of style across the items.
2. Editors ensure that all items are accurate in content. Editors compare reading passages against the original publications to ensure that all information is internally consistent across stimulus materials and items, including names, facts, or cited lines of text that appear in the item. They ensure that the answer key(s) and all information in the item are correct. For items with mathematical tasks, editors perform all calculations to ensure accuracy.
3. Editors review all material for fairness and language accessibility issues.
4. Editors confirm that the items reflect the accepted guidelines for good item construction. They examine all items for language that is simple, direct, and free of ambiguity with minimal verbal difficulty. Editors confirm that a problem or task and its stem are clearly defined and concisely worded with no unnecessary information. For multiple-choice interactions, editors check that options are parallel in structure and fit logically and grammatically with the stem. They also ensure that the key answers the question posed accurately and correctly, is not inappropriately obvious, and is the only correct answer to an item among the distractors. For constructed-response interactions, editors review the rubrics for appropriate style and grammar.

2.4.5 Senior Review

By the time a science item arrives at Senior Review, both content reviewers and editors have thoroughly vetted it. Senior reviewers (in particular, senior content specialists) look at the item's entire review history, ensuring that all the issues identified in that item have been adequately addressed. Senior reviewers verify the overall content of each item, confirming its accuracy, alignment to the performance standard, and consistency with expectations for the highest quality.

They check whether the scoring is working as intended and ensure that the scoring assertions adequately address the evidence the student provides with each type of response.

2.5 Review by State Personnel and Stakeholder Committees

All science items undergo an exhaustive external review process. Items in the Shared Science Assessment Item Bank were reviewed by content experts in one or several states and reviewed and approved by multiple stakeholder committees that evaluated them for both content and bias and sensitivity. Please note, the NDSA for Science draws exclusively from the ICCR item pool; the ICCR item pool is part of the larger Shared Science Assessment Item Bank.

2.5.1 State Review

After items have been developed for a state participating in the MOU, content experts from the state that owns the item review any eligible items before they are sent for committee review. The review of ICCR items is distributed over participating state committee reviews, where some items are reviewed by content experts from more than one state. Clients can request edits, such as wording edits, scoring edits, alignment changes, or task demand updates at this stage in the review process. A CAI science content expert reviews all client-requested edits considering the science item specifications, other clients' requests, and existing items in the bank to determine whether the requested edits will be made. At this stage, clients can either present these items to the committee (based on the edits made) or withhold them from committee review.

At least one or two states review the ICCR items. The state(s) then provide feedback on the ICCR items, and CAI's science leadership gathers suggestions and makes edits that improve the ICCR items. Not all suggestions are implemented, as CAI owns these items. Further, most MOU states accept or reject ICCR and MOU items (as they appear at the time) to be presented to their committees. Some clients skip this step and allow CAI to review all items with their committees before reviewing them. These items can be either set for field testing in a future administration or become a part of the locked operational pool.

2.5.2 Content Advisory Committee Reviews

During the Content Advisory Committee (CAC) reviews, items are reviewed for content validity, grade-level appropriateness, and alignment to the performance standard. CAC members are typically grade-level and subject-matter experts. During this review, educators also ensure that the scoring assertions clearly identify what is being scored as correct and give credit where they should (refer to Section 2.7.1, Rubric Validation).

Items developed for each state under the MOU are reviewed by the state that owns the items. In contrast, ICCR items are reviewed by the CAC of one or more states. In most cases, items are reviewed by multiple state committees before they are field tested or locked for operational use.

In 2022, all MOU states participated in a single CAC process wherein educators and administrators from multiple states reviewed items. The items were edited and returned to the respective state of ownership for final approval.

A summary of the committee meetings and number of participants and items reviewed is presented in Table 3, with additional details about the participants available in Appendix E, Content Advisory Committee Participant Details. Twelve North Dakota educators participated in the July 2021 ICCR meeting listed below. Appendix E also contains detailed information about the participants of Content Advisory Committee meetings of previous years.

Table 3. Summary of Content Advisory Committee Meetings

State/Item Bank	Meeting	Number of Committee Members	Number of Items Reviewed
Connecticut	July 2021	26	26[c]
	September 2021	27	25
ICCR	July 2021	[a]	141[c]
Idaho	July 2021	12	0[b, c]
	November 2021	11	317
Montana	July 2021	1	36[c]
	October 2021	6	41
Multi-State Science Assessment (Rhode Island and Vermont)	July 2021	7	32[c]
	August 2021	11	93
Oregon	August 2021	14	375
Utah	July 2021	0	55[c]
	August 2021	14	62
West Virginia	July 2021	10	16[c]
Wyoming	June/July 2021	14	39
	July 2021	14	39[c]

[a]Number of Content Advisory Committee Members is not available at the time of writing this report.
[b]Number of science items reviewed by Content Advisory Committees is unavailable at the time of writing this report.
[c]Items were reviewed in a combined multi-state Content Advisory Committee meeting.

2.5.3 Language Accessibility, Bias, and Sensitivity Committee Reviews

During the bias and sensitivity reviews, stakeholders review items to check for issues that might unfairly impact students based on their background. For example, some states include representatives from student populations such as special education, low vision, and the hearing impaired. Further, diverse members of this committee represent students of various ethnic and economic backgrounds to ensure that all items are free of bias and sensitivity concerns.

Due to the COVID-19 pandemic in 2020, 2021, and 2022, CAI reviewed items that contained references to virus, vaccine, bacteria, disease, infection, and related words and phrases. CAI content experts reviewed 65 items and rejected one item for sensitivity concerns.

In 2022, the MOU states were all involved in a single review process where participants from multiple states reviewed items. The items were edited and returned to the respective state of ownership for final approval.

A summary of the committee meetings and number of participants and items reviewed is presented in Table 4, with additional details about the participants available in Appendix F, Fairness Committee Participant Details. One of the North Dakota educators participated in the July 2021 ICCR meeting listed below. Appendix F also contains detailed information about the participants of Fairness Committee meetings of previous years.

Table 4. Summary of Fairness Committee Meetings

State/Item Bank	Meeting	Number of Committee Members	Number of Items Reviewed	Number of Items Rejected
Connecticut	July 2021	6	20[a]	0
	September 2021	7	111	23
ICCR	July 2021	15	157[a]	1
Idaho	December 2021	21	179	0
Montana	July 2021	3	41[a]	0
Multi-State Science Assessment (Rhode Island and Vermont)	July 2021	3	30[a]	1
	August 2021	3	93	3
Oregon	August 2021	7	353	13
U.S. Virgin Islands	October 2021	6	299	28
Utah	July 2021	11	64[a]	0
	August 2021	6	62	62
West Virginia	July 2021	2	12[a]	1
Wyoming	June/July 2021	6	39	39
	July 2021	4	28[a]	0

[a]Items were reviewed in a combined multi-state Fairness Committee Meeting.

2.5.4 Markup for Translation and Accessibility Features

After all approved state- and committee-recommended edits have been applied, the items are considered “locked” and ready for a portion of the accessibility tagging. TTS tagging is applied prior to field testing, while Spanish translations and braille are applied post-field testing. Accessibility markup is embedded into each item as part of the item development process rather than as a *post-hoc* process applied to completed tests.

Accessibility markup, whether translations or TTS, follows similar processes. One trained expert enters the markup, and then a second expert reviews the work and recommends changes if necessary. If there is disagreement, a third expert is engaged to resolve the conflict.

Currently, science items are tagged with TTS. Spanish translations, including Spanish TTS and braille, are available for a subset of items. The common Shared Science Assessment Item Bank is reviewed to identify items that are appropriate for braille embossing and/or Spanish translation/Spanish TTS. The braille and translated pool include a subset of items for each grade band.

2.6 FIELD TESTING

A large pool of science field-test items was administered in the following nine states in spring 2018: Connecticut, Hawaii, New Hampshire, Oregon, Rhode Island, Utah, Vermont, West Virginia, and Wyoming. Items were embedded as field-test items in the legacy science test for Hawaii, Oregon, and Wyoming. Connecticut and Rhode Island conducted an independent field test in which all students participated, but no scores were reported. Finally, in New Hampshire, Utah, Vermont, and West Virginia, an operational field test was administered.

In 2019, a second pool of field-test items was administered in the following nine states: Connecticut, Hawaii, Idaho, New Hampshire, Oregon, Rhode Island, Vermont, West Virginia, and Wyoming. Unscored field-test items were added as a separate segment to the operational (scored) legacy science test in Hawaii, Idaho (elementary school), and Wyoming. An independent field test in which students were administered a full set of items was conducted for a sample of Idaho middle schools. Finally, in Connecticut, New Hampshire, Oregon, Rhode Island, Vermont, and West Virginia, field-test items were administered as unscored items embedded within the operational items.

In 2021, a third pool of field-test items was administered in 12 states. An independent field test, in which students were administered a full set of items, was conducted for Idaho and Montana. Unscored field-test items were added as a separate segment to the operational (scored) legacy science test in Wyoming. Finally, in Connecticut, Hawaii, New Hampshire, North Dakota, Rhode Island, South Dakota, Vermont, Utah, and West Virginia, field-test items were administered as unscored items embedded within the operational items.

In 2022, a fourth pool of field-test items was administered in 13 states and one U.S. territory. Field-test items were administered as unscored items embedded within the operational items in those 13 states and one U.S. territory (Connecticut, Hawaii, Idaho, Montana, New Hampshire, North Dakota, Oregon, Rhode Island, South Dakota, Utah, Vermont, West Virginia, Wyoming, and the U.S. Virgin Islands). CAI's field-test process is detailed in Section 3.2, Field Testing, in Volume 1 of this technical report.

2.7 POST-FIELD-TEST REVIEW

Following the field test, items were subjected to a substantial validation process, including rubric validation and data review. That validation process is described in the following sections (Section 2.7.1, Rubric Validation, and Section 2.7.2, Data Review).

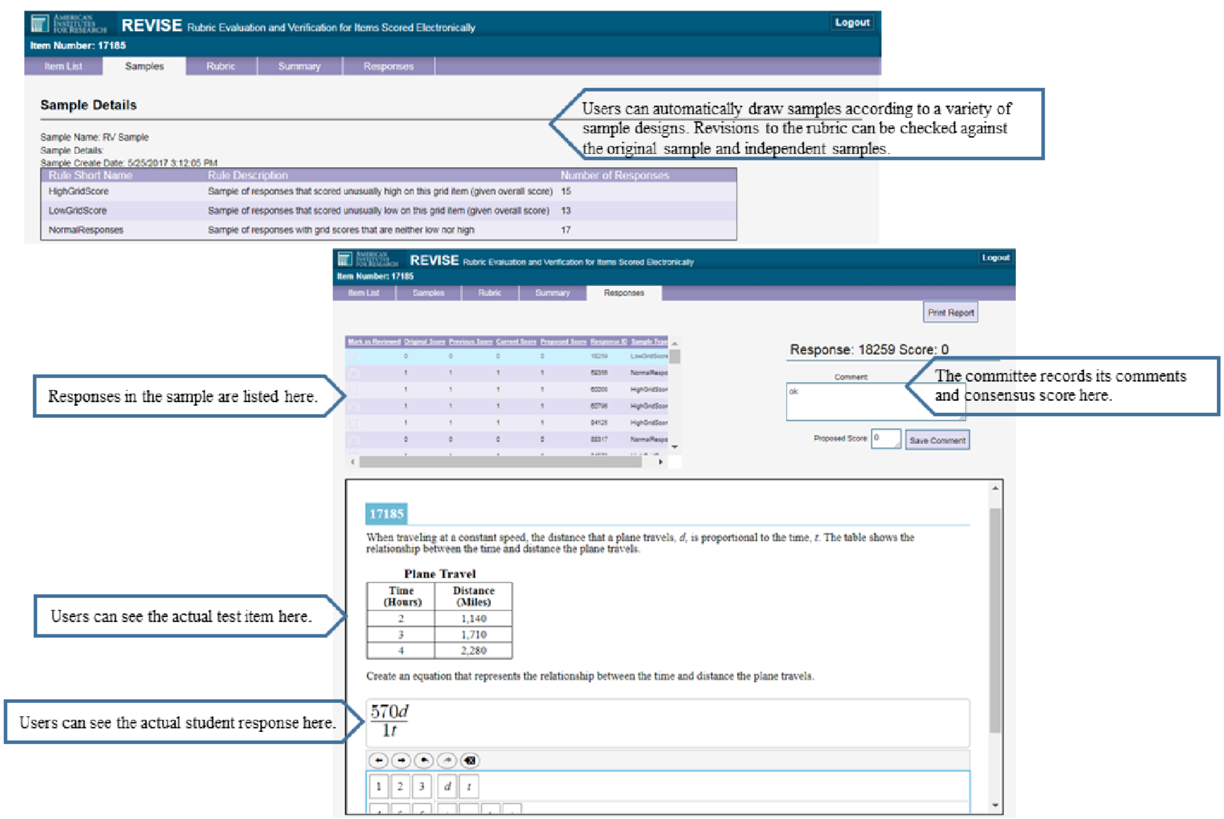
2.7.1 Rubric Validation

The validation process for the field-test items begins with rubric validation to verify and make any necessary revisions to the scoring rubrics. The rubric validation process occurs in two phases. During the first phase, CAI content experts work with the analysis team to prepare for the rubric validation meetings. CAI content experts use the Rubric Evaluation and Verification for Items Scored Electronically (REVISE) system to generate student responses that are scientifically sampled to overrepresent responses most likely to have been mis-scored. Specifically, the sample overrepresents: (1) low-scored responses from otherwise high-scoring students, and (2) high-scored responses from otherwise low-scoring students. This process allows CAI to identify any potential scoring concerns before the rubric validation meeting, such as unanticipated (but accurate) responses, equivalent responses that were not originally considered, and responses receiving credit but should not (based on the content and the item rubric). At this point, the rubrics may be adjusted, and responses rescored.

The second phase of rubric validation involves committees of educators in each state. The committees review the response samples generated by CAI to make recommendations to change or confirm each item's rubric. The committee recommendations are then discussed with the state of ownership to resolve any inconsistencies. Finally, rubric is edited or confirmed based on this resolution.

Figure 1 illustrates the features provided by the REVISE system.

Figure 1. Features of the REVISE Software



After the rubric validation meetings, CAI staff apply the approved revisions to the rubrics, and any items rejected as part of the process are rejected in the Item Tracking System (ITS). ITS archives critical information regarding the scoring certification completed during the rubric validation process. This includes any rubric changes made during the scoring decision meetings and the sign-off completed by the senior content expert once the rubric has been changed, rescoring the entire sample, and the verification that the final rubric functioned as intended.

Following rubric validation, all items are subject to statistical checks, and flagged items are presented in data review committees.

2.7.2 Data Review

Following rubric validation, all items are rescored, and classical item statistics are computed for the scoring assertions, including item difficulty and item discrimination statistics, testing time, and differential item functioning (DIF) statistics. The states' established standards for the statistics, and any items violating these standards, are flagged for a second educator review. Even though the scoring assertions were the basic units of analysis used to compute classical item statistics, the business rules to flag items for additional educator review were established at the item level because assertions cannot be reviewed in isolation. A common set of business rules was defined for all the states participating in the field test. The classical item statistics were computed on the data of the students testing in the state that owned the item. For Rhode Island and Vermont, which share their item development, statistics were computed on the combined data of students testing in both states. For ICCR items, the data from students testing in Connecticut, Idaho, New Hampshire, North Dakota, Oregon, South Dakota, Rhode Island, Utah, Vermont, and West Virginia were combined (states that administered ICCR items and used either an independent field test or operational test).

Section 4, Field-Test Classical Analysis, in Volume 1 of this technical report, describes the statistical flags used to designate items for data review. The flags are designed to highlight potential content weaknesses, miskeys, or possible bias issues. Committee members are taught to interpret these flags and are given guidelines for examining the items for content or fairness issues.

A data review committee reviewed each flagged item from each state participating in the MOU. In general, the data review committees consisted of content experts from the state's department of education or state educators (in this case, the state educators were science teachers) and were supported by CAI content experts. ICCR field-test items were sent to committee members from several states participating in the MOU. The outcomes were decided by CAI's science content leadership, who took the committees' recommendations into consideration.

Each state-owned item-data review meeting began with CAI staff leading participants through a training session to familiarize them with the purpose of the data review committee, the item development process, the data review process, and the flagging system. Committee members were taught to interpret each of the flags and provided with guidelines for examining the items for content or fairness issues. The training sessions included a group review of item cards, which detailed specific item attributes (including grade level and alignment to the science standards, the content and rubric of the item, and various item statistics). A sample of the training materials used for these data review meetings is presented in Appendix G, Sample Data Review Training Materials. Participants used an online environment via laptop computers to review the items and interact with them in a manner similar to that of students, and to view the statistics associated with each item.

The items were then reviewed by the participants who were most familiar with the particular grade (band) level and the items' content domain. CAI's content specialists, who were well versed in item statistics, facilitated the discussion in each room with CAI's psychometricians available to answer questions. At the end of each meeting day, CAI's content specialists met with state content specialists to review the committee recommendations and decide whether to accept or reject the item for inclusion in the operational pool. Items that were rejected become eligible for potential changes and additional field test items.

Table 5 summarizes the data review committee meetings. Details, including the composition of each committee, are available in Appendix H, Data Review Committee Participant Details.

Table 5. Summary of Data Review Committee Meetings

Owner	Meeting	Number of Committee Members	Item Type	Number of Items Reviewed	Number of Items Rejected
Connecticut	August 2018	29	**Total**	**18**	**11**
			Cluster	7	5
			Stand-Alone	11	6
	August 2019	29	**Total**	**53**	**17**
			Cluster	14	6
			Stand-Alone	39	11
	August 2021	19	**Total**	**51**	**12**
			Cluster	8	2
			Stand-Alone	43	10
	August 2022	15	**Total**	**19**	**6**
			Cluster	5	4
			Stand-Alone	14	2
Hawaii	August 2018	18	**Total**	**32**	**3**
			Cluster	7	1
			Stand-Alone	25	2
	August 2019	18	**Total**	**37**	**13**
			Cluster	17	5
			Stand-Alone	20	8
	August 2021	25[d]	**Total**	**26**	**8**
			Cluster	6	0
			Stand-Alone	20	8
	August 2022	12[d]	**Total**	**49**	**8**
			Cluster	11	2
			Stand-Alone	38	6
ICCR	July 2018	18	**Total**	**84**	**8**
			Cluster	33	2
			Stand-Alone	51	6
	August 2019	N/A[c]	**Total**	**43**	**3**
			Cluster	0	1
			Stand-Alone	43	2
	August 2021	25[d]	**Total**	**75**	**6**

Owner	Meeting	Number of Committee Members	Item Type	Number of Items Reviewed	Number of Items Rejected
			Cluster	11	2
			Stand-Alone	64	4
	August 2022	20[d]	**Total**	**68**	**14**
			Cluster	12	1
			Stand-Alone	56	13
Idaho	August 2019	10	**Total**	**12**	**6**
			Cluster	4	3
			Stand-Alone	8	3
	August 2021	25[d]	**Total**	**60**	**5**
			Cluster	26	1
			Stand-Alone	34	4
	August 2022	8[d]	**Total**	**4**	**0**
			Cluster	3	0
			Stand-Alone	1	0
Montana	September 2021	4	**Total**	**17**	**4**
			Cluster	3	2
			Stand-Alone	14	2
	September 2022	5	**Total**	**17**	**3**
			Cluster	5	2
			Stand-Alone	12	1
Multi-State Science Assessment (Rhode Island and Vermont)	August 2018	N/A[a]	**Total**	**9**	**6**
			Cluster	2	0
			Stand-Alone	7	6
	August 2019	N/A[a]	**Total**	**14**	**4**
			Cluster	2	1
			Stand-Alone	12	3
	August 2021	N/A[a]	**Total**	**18**	**9**
			Cluster	4	4
			Stand-Alone	14	5
	September 2022	N/A[a]	**Total**	**11**	**7**
			Cluster	1	1
			Stand-Alone	10	6
Oregon	September 2018	11	**Total**	**44**	**6**
			Cluster	28	5
			Stand-Alone	16	1
	August 2019	4	**Total**	**8**	**7**
			Cluster	1	1
			Stand-Alone	7	6
	August 2022	8[d]	**Total**	**31**	**8**
			Cluster	11	2
			Stand-Alone	20	6

Owner	Meeting	Number of Committee Members	Item Type	Number of Items Reviewed	Number of Items Rejected
South Dakota	September 2021	N/A[b]	**Total**	**15**	**0**
			Cluster	0	0
			Stand-Alone	15	0
Utah	August 2018	16	**Total**	**40**	**6**
			Cluster	40	6
			Stand-Alone	0	0
	September 2021	6	**Total**	**11**	**3**
			Cluster	11	3
			Stand-Alone	0	0
	September 2022	13	**Total**	**11**	**6**
			Cluster	11	6
			Stand-Alone	0	0
West Virginia	July 2018	4	**Total**	**3**	**1**
			Cluster	3	1
			Stand-Alone	0	0
	September 2019	4	**Total**	**7**	**6**
			Cluster	1	1
			Stand-Alone	6	5
	August 2021	25[d]	**Total**	**7**	**3**
			Cluster	1	1
			Stand-Alone	6	2
	August 2022	9[d]	**Total**	**10**	**4**
			Cluster	4	2
			Stand-Alone	6	2
Wyoming	October 2018	19	**Total**	**16**	**6**
			Cluster	6	1
			Stand-Alone	10	5
	August 2019	10	**Total**	**16**	**5**
			Cluster	4	3
			Stand-Alone	12	2
	August 2021	25[d]	**Total**	**16**	**4**
			Cluster	3	1
			Stand-Alone	13	3
	August 2022	12[d]	**Total**	**19**	**3**
			Cluster	2	0
			Stand-Alone	17	3

[a]Conducted by the Rhode Island Department of Education and the Vermont Agency of Education science content experts.
[b]Reviewed by South Dakota Department of Education.
[c]In summer 2019, ICCR field-test items were taken to Connecticut, Hawaii, and Idaho for committee review.

[d]Combined Data Review for multiple states (184 Hawaii, Idaho, West Virginia, Wyoming, and ICCR items in 2021 and 181 Hawaii, Idaho, Oregon, West Virginia, Wyoming, and ICCR items in 2022). There were 25 total participants in 2021 and 38 total participants in 2022. Items are broken out by owning state.

3. SHARED SCIENCE ASSESSMENT ITEM BANK SUMMARY

Tests based on *A Framework for K–12 Science Education* (National Research Council, 2012), such as the North Dakota State Assessment (NDSA) for Science, adopt a three-dimensional conceptualization of science understanding, including Science and Engineering Practices (SEPs), Crosscutting Concepts (CCCs), and Disciplinary Core Ideas (DCIs). Accordingly, the new science assessments are comprised mostly of item clusters representing a series of interrelated student interactions directed towards describing, explaining, and predicting scientific phenomena. Some stand-alone items are added to increase the coverage of the test without increasing the testing time or testing burden.

CAI has built the Shared Science Assessment Item Bank in partnership with multiple states and one U.S. territory. The science item bank is robust and has been constructed to support multiple statewide science assessments. As described earlier, the science items are written to the three-dimensional science standards. The Shared Science Assessment Item Bank comprises Independent College and Career Readiness (ICCR) item bank items and items developed for specific states, which are all shared with Memorandum of Understanding (MOU) partner states. Please note, the NDSA for Science draws exclusively from the ICCR item pool; the ICCR item pool is part of the larger Shared Science Assessment Item Bank. These items follow the same specifications, test development processes, and review processes. In 2018, CAI field tested more than 540 item clusters and stand-alone items, of which 451 (including items from all sources) were accepted and made available as operational items in 2019. In 2019, 347 item clusters and stand-alone items were field tested, of which 268 were accepted and made available as operational items in 2020. In 2021, CAI field tested 545 item clusters and stand-alone items, of which 458 have passed rubric validation and item data review. In 2022, CAI field tested 471 item clusters and stand-alone items, of which 403 have passed rubric validation and item data review.

Each partnered user of the Shared Science Assessment Item Bank selects items that are appropriately aligned and have passed required reviews (as described in Section 2, Item Development Process That Supports Validity of Claims) for use on its statewide assessment. The Shared Science Assessment Item Bank continues to grow as participating states and territory continue to field test new items. These participating partners collectively share the items and agree to field test new items yearly. The NDSA for Science draws exclusively from the ICCR item pool; because the ICCR item pool is part of the larger Shared Science Assessment Item Bank, the full item bank is described in Section 3.1, Current Composition of the Shared Science Assessment Item Bank.

3.1 CURRENT COMPOSITION OF THE SHARED SCIENCE ASSESSMENT ITEM BANK

The Shared Science Assessment Item Bank contains item clusters and stand-alone items. Item clusters represent a series of interrelated student interactions directed toward describing, explaining, and predicting scientific phenomena. Item clusters can consist of several item parts, requiring the student to interact with the item in various ways. In addition, shorter items (stand-alone items) are included to increase the coverage of the assessments without also increasing the testing time or testing burden.

Within each item (item cluster and stand-alone item), a series of explicit assertions is made about the knowledge and skills that a student has demonstrated based on specific features of the student's responses across multiple interactions. For example, a student may correctly graph data points indicating that they can construct a graph showing the relationship between two variables, but they may make an incorrect inference about the relationship between the two variables. In this case, the student's performance would not support the assertion that they can interpret relationships expressed graphically. Table 6 lists and describes the science items by interaction type. Some examples of various interaction types are presented in Appendix I, Example Item Interactions.

Table 6. Science Interaction Types and Descriptions

Interaction Type	Associated Sub-Types	Description
Choice	Multiple-Choice	Traditional multiple-choice interaction allows students to select a single option from a list of possible answer options.
	Multi-Select	Traditional multi-select interaction (checkboxes) allows students to select one or more options from a list of possible answer choices.
Text Entry	Simple Text Entry	Students type a response in a text box.
	Embedded Text Entry	Students type their responses in one or more text boxes that are embedded in a section of read-only text.
	Natural Language	Students are directed to provide a short, written response.
	Extended-Response	Students are directed to provide a longer, written response in the form of an essay.
Table	Table Match	This interaction type allows students to check a box to indicate if the information from a column header matches information from a row header.
	Table Input	This interaction type asks students to complete tabular data.
Edit Task	Edit Task	Students click a word and replace it with another word that they type to revise a sentence.
	Edit Task with Choice	Students click a word or phrase and select a replacement from several options.
	Edit Task Inline Choice	Drop-down menus are placed in the text, and students select an option to complete the text.
Hot-Text	Selectable	Selectable hot-text interactions require students to select one or more text elements in the response area.
	Re-orderable	Re-orderable hot-text interactions require students to click and drag hot-text elements into a different order.
	Drag-from-Palette	Drag-from-Palette hot-text interactions require students to drag elements from a palette into the available blank table cells or "gaps" (text boxes) in the response area.
	Custom	Custom hot-text interactions combine the functionality of the other hot-text interaction sub-types. Students responding to a custom hot-text interaction may need to select text elements, rearrange text elements, and/or drag text elements from a palette to blank table cells or drop targets in the response area.

Interaction Type	Associated Sub-Types	Description
Equation	N/A	Equation interactions require students to enter a response into input boxes. These boxes may stand alone or be in line with text or embedded in a table. The equation interaction may have an on-screen keypad that might consist of special mathematic characters. Students may also enter their responses via a physical keyboard.
Grid	Grid	Grid interactions require students to enter a response by interacting with a grid area in the answer space. Students may be required to draw a line or shape, plot a point, or create a graph. Students may also drag and drop or click on selectable hot spots.
	Hot-Spot	Hot-spot interaction sub-types facilitate grid interactions with specific hot-spot functionality. These interactions require students to select hot-spot regions in the grid area.
	Graphic Gap Match	Graphic gap match interactions facilitate grid interactions with specific drag-and-drop functionality. These interactions require students to drag image objects from a palette to specified regions (gaps) in the grid area.
Simulation	N/A	Simulation interactions allow students to investigate a phenomenon by selecting variables to get output data. Some simulations are accompanied by animations.

Table 7 through Table 11 present the number of items in the Shared Science Assessment Item Bank available for use in the spring 2022 statewide assessments. Appendix J, Shared Science Assessment Item Bank, presents the items available in the shared item bank by grade band and standards.

Table 7. Spring 2022 Shared Science Assessment Operational and Field-Test Item Bank

Grade Band and Item Type	ICCR Items	MOU Items[a]	Total Bank Items
Elementary School	**148**	**425**	**573**
Cluster	49	245	294
Stand-Alone	99	180	279
Middle School	**163**	**432**	**595**
Cluster	55	238	293
Stand-Alone	108	194	302
High School	**145**	**291**	**436**
Cluster	50	122	172
Stand-Alone	95	169	264
Total	**456**	**1148**	**1604**

[a]MOU state item sources include Connecticut, Hawaii, Idaho, Montana, MSSA (Rhode Island and Vermont), Oregon, Utah, West Virginia, and Wyoming.

Table 8. Spring 2022 Shared Science Assessment Operational Item Bank

Grade Band and Item Type	ICCR Operational Items	MOU Operational Items[a]	Total Bank Operational Items
Elementary School	**116**	**287**	**403**
Cluster	40	166	206
Stand-Alone	76	121	197
Middle School	**101**	**304**	**405**
Cluster	29	176	205
Stand-Alone	72	128	200
High School	**103**	**222**	**325**
Cluster	37	94	131
Stand-Alone	66	128	194
Total	**320**	**813**	**1133**

[a]MOU state operational item sources include Connecticut, Hawaii, Idaho, MSSA (Rhode Island and Vermont), Oregon, Utah, West Virginia, and Wyoming.

Table 9. Spring 2022 Shared Science Assessment Field-Test Item Bank

Grade Band and Item Type	ICCR Field-Test Items	MOU Field-Test Items[a]	Total Bank Field-Test Items
Elementary School	**32**	**138**	**170**
Cluster	9	79	88
Stand-Alone	23	59	82
Middle School	**62**	**128**	**190**
Cluster	26	62	88
Stand-Alone	36	66	102
High School	**42**	**69**	**111**
Cluster	13	28	41
Stand-Alone	29	41	70
Total	**136**	**335**	**471**

[a]MOU state field-test item sources include Connecticut, Hawaii, Idaho, Montana, MSSA (Rhode Island and Vermont), Utah, West Virginia, and Wyoming.

Table 10. Spring 2022 Shared Science Assessment Operational and Field-Test Item Bank by Science Discipline

Grade Band	Science Discipline	Item Type	ICCR Items	MOU Items[a]	Total Bank Items[b]
Elementary School	Earth and Space Sciences	Cluster	18	81	99
Elementary School	Earth and Space Sciences	Stand-Alone	29	59	88
Elementary School	Life Sciences	Cluster	14	68	82
Elementary School	Life Sciences	Stand-Alone	32	51	83
Elementary School	Physical Sciences	Cluster	17	96	113
Elementary School	Physical Sciences	Stand-Alone	38	70	108
Middle School	Earth and Space Sciences	Cluster	16	62	78
Middle School	Earth and Space Sciences	Stand-Alone	29	55	84
Middle School	Life Sciences	Cluster	22	91	113
Middle School	Life Sciences	Stand-Alone	47	67	114
Middle School	Physical Sciences	Cluster	17	78	95
Middle School	Physical Sciences	Stand-Alone	32	71	103
High School	Earth and Space Sciences	Cluster	13	24	37
High School	Earth and Space Sciences	Stand-Alone	22	38	60
High School	Life Sciences	Cluster	20	60	80
High School	Life Sciences	Stand-Alone	50	74	124
High School	Physical Sciences	Cluster	17	37	54
High School	Physical Sciences	Stand-Alone	23	57	80
Total			**456**	**1139**	**1595**

[a]MOU state item sources include Connecticut, Hawaii, Idaho, Montana, MSSA (Rhode Island and Vermont), Oregon, Utah, West Virginia, and Wyoming.
[b]Count excludes nine MOU items that do not align to the Next Generation Science Standards (NGSS).

Table 11. Spring 2022 Shared Science Assessment Operational and Field-Test Item Bank by Disciplinary Core Idea

Grade Band	Science Discipline	Disciplinary Core Idea	ICCR Items	MOU Items[a]	Total Bank Items[b]
Elementary School	Earth and Space Sciences	ESS1	12	38	50
		ESS2	15	64	79
		ESS3	20	38	58
	Life Sciences	LS1	16	48	64
		LS2	6	22	28
		LS3	5	16	21
		LS4	19	33	52
	Physical Sciences	PS1	14	41	55
		PS2	15	38	53
		PS3	20	57	77
		PS4	6	30	36
Middle School	Earth and Space Sciences	ESS1	15	31	46
		ESS2	16	42	58
		ESS3	14	44	58
	Life Sciences	LS1	22	54	76
		LS2	24	43	67
		LS3	5	19	24
		LS4	18	42	60
	Physical Sciences	PS1	13	49	62
		PS2	6	43	49
		PS3	19	36	55
		PS4	11	21	32
High School	Earth and Space Sciences	ESS1	12	17	29
		ESS2	11	25	36
		ESS3	12	20	32
	Life Sciences	LS1	20	38	58
		LS2	21	41	62
		LS3	11	18	29
		LS4	18	37	55
	Physical Sciences	PS1	19	33	52

Grade Band	Science Discipline	Disciplinary Core Idea	ICCR Items	MOU Items[a]	Total Bank Items[b]
		PS2	9	23	32
		PS3	5	22	27
		PS4	7	16	23
Total			**456**	**1139**	**1595**

[a]MOU state item sources include Connecticut, Hawaii, Idaho, Montana, MSSA (Rhode Island and Vermont), Oregon, Utah, West Virginia, and Wyoming. [b]Count excludes nine MOU items that do not align to the NGSS.

3.2 Strategy for Item Bank Evaluation and Replenishment

CAI and the participating MOU states and one U.S. territory continue to develop items to replenish and grow the Shared Science Assessment Item Bank. The general strategy for targeting item development involves gathering information from three sources:

1. Characteristics of released items to be replaced
2. Characteristics of items that are overused
3. Tabulations of content coverage and ranges of difficulty that help to identify gaps in the item bank

Before a test goes live, simulations are used to fine-tune the parameters of the algorithm that govern the item selection in an adaptive test design. Among the many reports from the simulator are items seen by more than 20% of students. The characteristics of these items are the primary targets for development. Overused items become candidates for release in two years, once replacements have been introduced into the operational bank.

4. North Dakota State Assessment for Science Test Construction

4.1 Test Design

The North Dakota State Assessment (NDSA) for Science was administered online to students in grades 4, 8, and 10 using an adaptive test design in spring 2022. Appendix K, NDSA for Science Item Pool, presents the 2022 item pool by grade band, performance expectation, and origin. In an adaptive test, operational items are selected on the fly based on a student's performance on past items while ensuring the test blueprint is followed for each student. An advantage of adaptive testing is that it can provide more precise scores for students with lower and higher proficiencies, in contrast to fixed forms and linear-on-the fly tests (LOFTs) that are typically targeted to provide the best precision for students with medium proficiencies. Another advantage of adaptive testing that fixed forms and LOFTs do not offer is that every student is likely to see a different set of items that adapts to the student's ability, thus offering a better testing experience.

Items are selected by an item-selection algorithm that is based on content and information value. At any given point during the test, the content value of an item is determined by its contribution to meeting the blueprint, given the content characteristics of the items that have already been administered. During the test, the content value increases for items that exhibit features that have not met their designated minimum as the end of the test approaches. Conversely, the content value decreases for items with content features that met the minimum. The information value of an item is based on the information function evaluated at the estimated proficiency. The proficiency estimate is updated throughout the test.

The adaptive item-selection algorithm is the same algorithm CAI uses to deliver English language arts (ELA) and mathematics tests, but with some modifications to make it suitable for using item clusters. Specifically, the proficiencies estimated during the test are computed under an item response theory (IRT) model that incorporates cluster effects. To avoid over-selecting items with many scoring assertions, the information of an item at an estimated proficiency level is normalized by the number of assertions in the item (similar to how information is computed for item sets in ELA and mathematics assessments). Additional details about CAI's adaptive testing algorithm are available in Appendix L, Adaptive Algorithm Design.

A non-segmented test design was used for the NDSA for Science. Students received items from different disciplines in random order. Compared to a segmented design, in which items are administered by science discipline, a non-segmented test design provides more freedom when selecting items that target a current best estimate of proficiency in an adaptive test. Embedded field-test items were randomly positioned in the test and randomly distributed across students. Every student received either one item cluster or four stand-alone items as embedded field-test items in their test.

4.2 TEST BLUEPRINTS

Test blueprints provide the parameters for the following elements:

- Test length
- Science disciplines to be covered and the acceptable number of items across performance standards (standards) within each science discipline and disciplinary core idea (DCI)

The blueprints for the NDSA for Science are presented in Table 12 through Table 14. CAI and NDDPI developed the blueprints collaboratively, along with review and input from North Dakota educators. The details of the blueprints' development and educator participation are available in Appendix M, Science Blueprint Development.

Table 12. NDSA for Science Test Blueprint, Grade 4

Grade 4, Arranged by DCI	Min Item Clusters	Max Item Clusters	Min Stand-Alone Items	Max Stand-Alone Items	Min Item Clusters + Min Stand-Alone Items	Max Item Clusters + Max Stand-Alone Items
Discipline—Physical Science, Performance Standard Total = 10	**2**	**2**	**4**	**4**	**6**	**6**
DCI —Motion and Stability: Forces and Interactions and Waves and Their Applications	**0**	**1**	**0**	**2**	**0**	**3**
3-PS2-1: Forces and Motion, Types of Interactions	0	1	0	1	0	1
3-PS2-2: Forces and Motion	0	1	0	1	0	1
3-PS2-3: Types of Interactions	0	1	0	1	0	1
3-PS2-4: Types of Interactions*	0	1	0	1	0	1
DCI—Energy	**0**	**1**	**0**	**2**	**0**	**3**
4-PS3-1: Energy	0	1	0	1	0	1
4-PS3-2: Conservation and Transfer of Energy	0	1	0	1	0	1
4-PS3-3: Conservation and Transfer of Energy, Energy and Forces	0	1	0	1	0	1
4-PS3-4: Conservation and Transfer of Energy*	0	1	0	1	0	1
DCI—Waves and Their Applications in Technologies for Information Transfer	**0**	**1**	**0**	**2**	**0**	**2**
4-PS4-1: Wave Properties	0	1	0	1	0	1
4-PS4-3: Information Technologies*	0	1	0	1	0	1
Discipline—Life Science, Performance Standard Total = 9	**2**	**2**	**4**	**4**	**6**	**6**
DCI—From Molecules to Organisms: Structure and Function	**0**	**1**	**0**	**2**	**0**	**3**
3-LS1-1: Growth and Development of Organisms	0	1	0	1	0	1
4-LS1-1: Structure, Function	0	1	0	1	0	1
4-LS1-2: Information Processing	0	1	0	1	0	1
DCI—Ecosystems: Interactions, Energy, and Dynamics	**0**	**1**	**0**	**1**	**0**	**1**

Grade 4, Arranged by DCI	Min Item Clusters	Max Item Clusters	Min Stand-Alone Items	Max Stand-Alone Items	Min Item Clusters + Min Stand-Alone Items	Max Item Clusters + Max Stand-Alone Items
3-LS2-1: Social Interactions and Group Behavior	0	1	0	1	0	1
DCI—Inheritance and Variation of Traits	**0**	**1**	**0**	**2**	**0**	**2**
3-LS3-1: Inheritance and Variation of Traits	0	1	0	1	0	1
3-LS3-2: Inheritance and Variation of Traits	0	1	0	1	0	1
DCI—Biological Evolution: Unity and Diversity	**0**	**1**	**0**	**2**	**0**	**3**
3-LS4-1: Evidence of Common Ancestry and Diversity	0	1	0	1	0	1
3-LS4-2: Natural Selection	0	1	0	1	0	1
3-LS4-3: Adaptation	0	1	0	1	0	1
Discipline—Earth and Space Science, Performance Standard Total = 8	**2**	**2**	**4**	**4**	**6**	**6**
DCI—Earth's Systems	**0**	**1**	**0**	**2**	**0**	**3**
3-ESS2-1: Weather and Climate	0	1	0	1	0	1
3-ESS2-2: Weather and Climate	0	1	0	1	0	1
4-ESS2-1: Earth Materials and Systems	0	1	0	1	0	1
4-ESS2-2: Plate Tectonics and System Interactions	0	1	0	1	0	1
DCI—Earth and Human Activity	**0**	**1**	**0**	**2**	**0**	**3**
3-ESS3-1: Natural Hazards*	0	1	0	1	0	1
4-ESS3-2: Natural Hazards*	0	1	0	1	0	1
4-ESS3-1: Natural Resources	0	1	0	1	0	1
DCI—Earth's Place in the Universe	**0**	**1**	**0**	**1**	**0**	**1**
4-ESS1-1: Earth's History	0	1	0	1	0	1
Performance Standard Total = 27	**6**	**6**	**12**	**12**	**18**	**18**

*These performance standards have an engineering component.

Table 13. NDSA for Science Test Blueprint, Grade 8

Grade 8, Arranged by DCI	Min Item Clusters	Max Item Clusters	Min Stand-Alone Items	Max Stand-Alone Items	Min Clusters + Stand Alone Items	Max Item Clusters + Stand-Alone Items
Discipline—Physical Science, Performance Standard Total = 19	**2**	**2**	**4**	**4**	**6**	**6**
DCI—Matter and Its Interactions	**0**	**1**	**0**	**2**	**0**	**3**
MS-PS1-1: Structure and Properties of Matter	0	1	0	1	0	1
MS-PS1-2: Structure and Properties of Matter, Chemical Reactions	0	1	0	1	0	1
MS-PS1-3: Structure and Properties of Matter, Chemical Reactions	0	1	0	1	0	1
MS-PS1-4: Structure and Properties of Matter, Energy	0	1	0	1	0	1
MS-PS1-5: Chemical Reactions	0	1	0	1	0	1
MS-PS1-6: Chemical Reactions*	0	1	0	1	0	1
DCI—Motion and Stability: Forces and Interactions	**0**	**1**	**0**	**2**	**0**	**3**
MS-PS2-1: Forces and Motion*	0	1	0	1	0	1
MS-PS2-2: Forces and Motion	0	1	0	1	0	1
MS-PS2-3: Types of Interactions	0	1	0	1	0	1
MS-PS2-4: Types of Interactions	0	1	0	1	0	1
MS-PS2-5: Types of Interactions	0	1	0	1	0	1
DCI—Energy	**0**	**1**	**0**	**2**	**0**	**3**
MS-PS3-1: Energy	0	1	0	1	0	1
MS-PS3-2: Energy, Relationship Between Energy and Forces	0	1	0	1	0	1
MS-PS3-3: Energy, Conservation of Energy*	0	1	0	1	0	1
MS-PS3-4: Energy, Conservation and Transfer of Energy	0	1	0	1	0	1
MS-PS3-5: Conservation and Transfer of Energy	0	1	0	1	0	1

Grade 8, Arranged by DCI	Min Item Clusters	Max Item Clusters	Min Stand-Alone Items	Max Stand-Alone Items	Min Clusters + Stand Alone Items	Max Item Clusters + Stand-Alone Items
DCI—Waves and Their Applications in Technologies for Information Transfer	**0**	**1**	**0**	**2**	**0**	**3**
MS-PS4-1: Wave Properties	0	1	0	1	0	1
MS-PS4-2: Wave Properties and Electromagnetic Radiation	0	1	0	1	0	1
MS-PS4-3: Information Technologies	0	1	0	1	0	1
Discipline—Life Science, Performance Standard Total = 20	**2**	**2**	**4**	**4**	**6**	**6**
DCI—From Molecules to Organisms: Structures and Processes	**0**	**1**	**0**	**2**	**0**	**3**
MS-LS1-1: Structure and Function	0	1	0	1	0	1
MS-LS1-2: Structure and Function	0	1	0	1	0	1
MS-LS1-3: Structure and Function	0	1	0	1	0	1
MS-LS1-4: Growth and Development of Organisms	0	1	0	1	0	1
MS-LS1-5: Growth and Development of Organisms	0	1	0	1	0	1
MS-LS1-6: Organization of Matter and Energy Flow in Organisms	0	1	0	1	0	1
MS-LS1-7: Organization of Matter and Energy Flow in Organisms	0	1	0	1	0	1
DCI—Ecosystems: Interactions, Energy, and Dynamics	**0**	**1**	**0**	**2**	**0**	**3**
MS-LS2-1: Interdependent Relationships in Ecosystems	0	1	0	1	0	1
MS-LS2-2: Interdependent Relationships in Ecosystems	0	1	0	1	0	1
MS-LS2-3: Cycle of Matter and Energy in Ecosystems	0	1	0	1	0	1
MS-LS2-4: Ecosystem Dynamics, Functioning, and Resilience	0	1	0	1	0	1
MS-LS2-5: Ecosystem Dynamics, Biodiversity, and Humans*	0	1	0	1	0	1
DCI—Heredity: Inheritance and Variation of Traits	**0**	**1**	**0**	**2**	**0**	**2**
MS-LS3-1: Inheritance and Variation of Traits	0	1	0	1	0	1
MS-LS3-2: Inheritance and Variation of Traits	0	1	0	1	0	1

Grade 8, Arranged by DCI	Min Item Clusters	Max Item Clusters	Min Stand-Alone Items	Max Stand-Alone Items	Min Clusters + Stand Alone Items	Max Item Clusters + Stand-Alone Items
DCI—Biological Evolution: Unity and Diversity	**0**	**1**	**0**	**2**	**0**	**3**
MS-LS4-1: Evidence of Common Ancestry and Diversity	0	1	0	1	0	1
MS-LS4-2: Evidence of Common Ancestry and Diversity	0	1	0	1	0	1
MS-LS4-3: Evidence of Common Ancestry and Diversity	0	1	0	1	0	1
MS-LS4-4: Natural Selection	0	1	0	1	0	1
MS-LS4-5: Natural Selection	0	1	0	1	0	1
MS-LS4-6: Adaptation	0	1	0	1	0	1
Discipline—Earth and Space Science, Performance Standard Total = 15	**2**	**2**	**4**	**4**	**6**	**6**
DCI—Earth's Place in the Universe	**0**	**1**	**0**	**2**	**0**	**3**
MS-ESS1-1: The Universe and Its Stars, Earth, and the Solar System	0	1	0	1	0	1
MS-ESS1-2: The Universe and Its Stars, Earth, and the Solar System	0	1	0	1	0	1
MS-ESS1-3: Earth and the Solar System	0	1	0	1	0	1
MS-ESS1-4: History of Earth	0	1	0	1	0	1
DCI—Earth's Systems	**0**	**1**	**0**	**2**	**0**	**3**
MS-ESS2-1: Earth's Materials and Systems	0	1	0	1	0	1
MS-ESS2-2: Earth's Materials and Systems, Roles of Water	0	1	0	1	0	1
MS-ESS2-3: Plate Tectonics	0	1	0	1	0	1
MS-ESS2-4: Roles of Water in Earth's Surface Processes	0	1	0	1	0	1
MS-ESS2-5: Roles of Water, Weather, and Climate	0	1	0	1	0	1
MS-ESS2-6: Roles of Water, Weather, and Climate	0	1	0	1	0	1
DCI—Earth and Human Activity	**0**	**1**	**0**	**2**	**0**	**3**
MS-ESS3-1: Natural Resources	0	1	0	1	0	1

Grade 8, Arranged by DCI	Min Item Clusters	Max Item Clusters	Min Stand-Alone Items	Max Stand-Alone Items	Min Clusters + Stand Alone Items	Max Item Clusters + Stand-Alone Items
MS-ESS3-2: Natural Hazards	0	1	0	1	0	1
MS-ESS3-3: Human Impacts*	0	1	0	1	0	1
MS-ESS3-4: Human Impacts	0	1	0	1	0	1
MS-ESS3-5: Global Climate Change	0	1	0	1	0	1
Performance Standard Total = 54	**6**	**6**	**12**	**12**	**18**	**18**

*These performance standards have an engineering component.

Table 14. NDSA for Science Test Blueprint, Grade 10

Grade 10, Arranged by DCI	Min Item Clusters	Max Item Clusters	Min Stand-Alone Items	Max Stand-Alone Items	Min Item Clusters + Stand-Alone Items	Max Item Clusters + Stand-Alone Items
Discipline—Physical Science, Performance Standard Total = 13	**2**	**2**	**4**	**4**	**6**	**6**
DCI—Matter and Its Interactions	**0**	**1**	**0**	**2**	**0**	**3**
HS-PS1-1: Structure and Properties of Matter	0	1	0	1	0	1
HS-PS1-2: Structure and Properties of Matter	0	1	0	1	0	1
HS-PS1-5: Chemical Reactions	0	1	0	1	0	1
HS-PS1-7: Chemical Reactions	0	1	0	1	0	1
HS-PS1-8: Nuclear Processes	0	1	0	1	0	1
DCI—Motion and Stability: Forces and Interactions	**0**	**1**	**0**	**2**	**0**	**3**
HS-PS2-1: Forces and Motion	0	1	0	1	0	1
HS-PS2-2: Forces and Motion	0	1	0	1	0	1
HS-PS2-3: Forces and Motion*	0	1	0	1	0	1
DCI—Energy and Waves and Their Applications	**0**	**1**	**0**	**2**	**0**	**3**
HS-PS3-1: Energy	0	1	0	1	0	1
HS-PS3-2: Energy	0	1	0	1	0	1
HS-PS3-3: Energy in Chemical Processes*	0	1	0	1	0	1
HS-PS3-4: Energy Conservation and Transfer	0	1	0	1	0	1
HS-PS4-1: Wave Properties	0	0	0	1	0	1
Discipline—Life Science, Performance Standard Total = 23	**4**	**4**	**8**	**8**	**12**	**12**
DCI—From Molecules to Organisms: Structures and Processes	**0**	**1**	**0**	**3**	**0**	**4**
HS-LS1-1: Structure and Function	0	1	0	1	0	1
HS-LS1-2: Structure and Function	0	1	0	1	0	1
HS-LS1-3: Structure and Function	0	1	0	1	0	1

Grade 10, Arranged by DCI	Min Item Clusters	Max Item Clusters	Min Stand-Alone Items	Max Stand-Alone Items	Min Item Clusters + Stand-Alone Items	Max Item Clusters + Stand-Alone Items
HS-LS1-4: Growth and Development of Organisms	0	1	0	1	0	1
HS-LS1-5: Organization for Matter and Energy Flow in Organisms	0	1	0	1	0	1
HS-LS1-6: Organization for Matter and Energy Flow in Organisms	0	1	0	1	0	1
HS-LS1-7: Organization for Matter and Energy Flow in Organisms	0	1	0	1	0	1
DCI—Ecosystems: Interactions, Energy, and Dynamics	**0**	**1**	**0**	**3**	**0**	**4**
HS-LS2-1: Interdependent Relationships in Ecosystems	0	1	0	1	0	1
HS-LS2-2: Interdependent Relationships in Ecosystems	0	1	0	1	0	1
HS-LS2-3: Cycles of Matter and Energy Transfer in Ecosystems	0	1	0	1	0	1
HS-LS2-4: Cycles of Matter and Energy Transfer in Ecosystems	0	1	0	1	0	1
HS-LS2-5: Cycles of Matter and Energy Transfer in Ecosystems	0	1	0	1	0	1
HS-LS2-6: Ecosystem Dynamics, Functioning, and Resilience	0	1	0	1	0	1
HS-LS2-7: Ecosystem Dynamics, Functioning, and Resilience*	0	1	0	1	0	1
DCI—Heredity: Inheritance and Variation of Traits	**0**	**1**	**0**	**3**	**0**	**4**
HS-LS3-1: Structure and Function, Inheritance of Traits	0	1	0	1	0	1
HS-LS3-2: Variation of Traits	0	1	0	1	0	1
HS-LS3-3: Variation of Traits	0	1	0	1	0	1
DCI—Biological Evolution: Unity and Diversity	**0**	**1**	**0**	**3**	**0**	**4**
HS-LS4-1: Evidence of Common Ancestry and Diversity	0	1	0	1	0	1
HS-LS4-2: Natural Selection, Adaptation	0	1	0	1	0	1
HS-LS4-3: Natural Selection, Adaptation	0	1	0	1	0	1
HS-LS4-4: Adaptation	0	1	0	1	0	1
HS-LS4-5: Adaptation	0	1	0	1	0	1
HS-LS4-6: Adaptation, Biodiversity, and Humans*	0	1	0	1	0	1

Grade 10, Arranged by DCI	Min Item Clusters	Max Item Clusters	Min Stand-Alone Items	Max Stand-Alone Items	Min Item Clusters + Stand-Alone Items	Max Item Clusters + Stand-Alone Items
Performance Standard Total = 36	**6**	**6**	**12**	**12**	**18**	**18**

*These performance standards have an engineering component.

The main characteristics of the blueprints were that a standard could be tested only once (indicated by the values of 0 and 1 for the minimum and maximum values of the individual standards in Table 12 through Table 14). In most cases, no more than one item cluster or two stand-alone items could be sampled from the same DCI, and no more than three total items could be sampled from the same DCI (as indicated by the minimum and maximum values in the rows representing DCIs).

While tests are not timed, the North Dakota Department of Public Instruction (NDDPI) published estimated testing times for the NDSA for Science. The 85th percentile of the real testing times is presented by grade in Table 15.

Table 15. NDSA for Science 85th Percentile Testing Times by Grade

Subject	Grade	85th Percentile Testing
Science	4	103.68
	8	78.69
	10	78.73

4.3 ONLINE TEST CONSTRUCTION

During fall 2021, CAI's psychometricians and content experts worked with NDDPI's content specialists and leadership to build item pools for the spring 2022 administration. The NDSA for Science test construction used a structured test construction plan, explicit blueprints, and active collaborative participation from all parties.

CAI test developers built the 2022 NDSA for Science item pool to match items exactly to the detailed test blueprints. Operational items were selected from the ICCR item pool to fulfill the blueprint for each grade. Table 16 through Table 20 summarize the 2022 NDSA for Science item pool. Appendix K, NDSA for Science Item Pool, outlines the 2022 item pool by grade and standards.

Table 16. Spring 2022 NDSA for Science Operational and Field-Test Item Pool

Grade and Item Type	NDSA Item Pool[a]
Grade 4	**92**
Cluster	30
Stand-Alone	62
Grade 8	**111**
Cluster	32
Stand-Alone	79
Grade 10	**85**
Cluster	28
Stand-Alone	57
Total	**288**

Table 17. Spring 2022 NDSA for Science Operational Item Pool

Grade Band and Item Type	NDSA Operational Item Pool[a]
Grade 4	**80**
Cluster	26
Stand-Alone	54
Grade 8	**99**
Cluster	28
Stand-Alone	71
Grade 10	**73**
Cluster	24
Stand-Alone	49
Total	**252**

Table 18. Spring 2022 NDSA for Science Field-Test Item Pool

Grade Band and Item Type	NDSA Field-Test Item Pool
Grade 4	**12**
Cluster	4
Stand-Alone	8
Grade 8	**12**
Cluster	4
Stand-Alone	8
Grade 10	**12**
Cluster	4
Stand-Alone	8
Total	**36**

Table 19. Spring 2022 NDSA for Science Operational and Field-Test Item Pool by Reporting Category

Grade	Science Discipline	Item Type	NDSA Item Pool
Grade 4	Earth and Space Sciences	Cluster	9
		Stand-Alone	17
	Life Sciences	Cluster	12
		Stand-Alone	24
	Physical Sciences	Cluster	9
		Stand-Alone	21
Grade 8	Earth and Space Sciences	Cluster	10
		Stand-Alone	21
	Life Sciences	Cluster	10
		Stand-Alone	36
	Physical Sciences	Cluster	12
		Stand-Alone	22
Grade 10	Life Sciences	Cluster	18
		Stand-Alone	47
	Physical Sciences	Cluster	10
		Stand-Alone	10
Total			**288**

Table 20. Spring 2022 NDSA for Science Operational and Field-Test Item Pool by Disciplinary Core Idea

Grade	Science Discipline	Disciplinary Core Idea	NDSA Item Pool
Grade 4	Earth and Space Sciences	ESS1	2
		ESS2	9
		ESS3	15
	Life Sciences	LS1	13
		LS2	4
		LS3	5
		LS4	14
	Physical Sciences	PS2	12
		PS3	15
		PS4	3
Grade 8	Earth and Space Sciences	ESS1	12
		ESS2	8
		ESS3	11
	Life Sciences	LS1	10
		LS2	19
		LS3	5
		LS4	12
	Physical Sciences	PS1	9
		PS2	3
		PS3	14
		PS4	8
Grade 10	Life Sciences	LS1	18
		LS2	18
		LS3	11
		LS4	18
	Physical Sciences	PS1	10
		PS2	4
		PS3	4
		PS4	2
Total			**288**

Additional information on *p*-values, biserial correlations, and item response theory (IRT) parameters is available in Volume 1 of this technical report. Information about calibrating, equating, and scoring the NDSA for Science is also available in Volume 1.

4.4 Paper-Based Braille Accommodation Form Construction

Student scores should not depend upon the mode of administration or type of test form. In spring 2022, the NDSA for Science was primarily administered via CAI's online Test Delivery System (TDS), with only one student in grade 10 taking the paper-based braille test. The braille tests are the only paper-based, fixed forms for NDSA. Of note, scores obtained via alternate modes of administration must be established as comparable to scores obtained via online testing. This section outlines the overall test development plans that ensured comparability between the online and paper-based forms.

To build paper-based braille forms, content specialists began with the online pool and removed any items that could not render on paper and would therefore be inaccessible to visually impaired students taking the braille tests. Next, content specialists constructed fixed forms adhering to the test blueprint. All overall, discipline (reporting category), DCI, and performance standard-level blueprint requirements were met.

4.5 Remote Testing Forms

Testing forms were constructed to assess science among students taking the test remotely. As the NDDPI did not see a need, none of the remote testing forms were administered during the 2021–2022 school year.

5. Simulation Summary Report

This section describes the results of the simulated test administrations used to configure and evaluate the adequacy of the adaptive algorithm used to administer the 2021–2022 North Dakota State Assessment (NDSA) for Science for grades 4, 8, and 10. Simulations were conducted to configure the algorithm's settings and evaluate whether individual tests adhered to the test blueprint.

Some important settings included "Select Candidate Set 1" (cset1) and "Select Candidate Set 2" (cset2), which represent subsets of the item pool that were eligible for item selection. Refer to Appendix L, Adaptive Algorithm Design, for additional details about the current item selection algorithm. In spring 2022, cset1 and cset2 values were set to 10 and 1. Psychometricians reviewed the simulation results and configured settings based on some key diagnostics, including:

- **Match-to-Test Blueprint.** This diagnostic determines whether the tests contain the correct number of overall test items and whether those items are in the appropriate proportion by content categories at each level of the content hierarchy, as specified in the test blueprints for each science grade.
- **Item Exposure Rate.** This diagnostic evaluates the utility of the item pools and identifies overexposed and underexposed items.

These diagnostics are interrelated. For example, if the test pool for a particular content category is limited (i.e., there are only a few test items available), achieving a 100% match to the blueprint for this content level will lead to a high item exposure rate, which means that a large number of students will see the same items. The software system that performs the simulation allows adjustments to the setting parameters to attain the best possible balance among these diagnostics. The simulation involves an iterative process that reviews the initial results, adjusts the system parameters, runs new simulations, reviews the new results, and repeats the exercise until an optimal balance is achieved. The final setting would then be applied to the operational tests.

5.1 Factors Affecting Simulation Results

Several factors may influence the simulation results in an adaptive test administration. These factors include

- *The proportional relationship between the pool and the constraints to be met.* Proportionally distributed pools tend to make better use of the pool (i.e., more uniform item exposure) and make it easier to meet blueprint and other constraints. For example, if the specifications call for at least one item cluster per disciplinary core idea (DCI), but the pool has no item cluster for some DCIs, it may be impossible to meet this constraint.
- *The correlational structure between constraints.* It is easier to satisfy a constraint if there are instances of the constraint at all levels of another constraint. For example, if stand-alone items within a discipline are associated only with a specific DCI, it may be difficult to meet both the desired distribution of content and the desired distribution of item type.
- *Whether or not there is a strict maximum on a given constraint.* This means that the requirement must be met exactly in each test administration.

5.2 Results of Simulated Test Administrations: English

This section presents the simulation results for the English online test, which is the test taken by almost all students (99.95%). Simulations were evaluated for all content areas using 5,000 simulated cases per grade.

5.2.1 Summary of Blueprint Match

The simulation results showed no blueprint violations at all content levels for all three grades.

5.2.2 Item Exposure

The simulator output also reports the degree to which the constraints outlined in the blueprints may yield greater exposure of items to students. This finding is reported by examining the percentage of test administrations in which an item appears. For instance, in a fixed paper-based form, 100% of the items appear on 100% of the test administrations because every test taker takes the same form. In an adaptive or linear-on-the-fly (LOFT) test with a sufficiently large item pool, it is expected that most of the items would appear on a relatively small percentage of the test administrations only.

When this condition holds, it suggests that test administrations between students are more or less unique. Therefore, the item exposure rate was calculated for each item across by dividing the total number of test administrations in which an item appeared by the total number of tests administered. Then the distribution of the item exposure rate (r) was reported in eight bins. The bins were $r = 0\%$ (unused), $0\% < r \leq 1\%$, $1\% < r \leq 5\%$, $5\% < r \leq 20\%$, $20\% < r \leq 40\%$, $40\% < r \leq 60\%$, $60\% < r \leq 80\%$, and $80\% < r \leq 100\%$. If global item exposure is minimal, it is expected that the largest proportion of items would appear in the bins of $0\% < r \leq 20\%$, indicating that most of the items appeared on a very small percentage of the test forms.

Table 21 presents the percentage of items that fell into each exposure bin for all grades. Most test items (89.05% or more) were administered in 1–60% of the test administrations. No item had an exposure rate less than 1% in grade 4 and 10, and only 6.06% of items had an exposure rate less than 1% in grade 8. A few items had an exposure rate higher than 80% because of the limitation of the current pool for some content categories.

Table 21. Item Exposure Rates by Grade: Percentage of Items by Exposure Rate, Across All English Online Simulation Sessions

Grade	Total Items	0%	(0,1]%	(1,5]%	(5,20]%	(20,40]%	(40,60]%	(60,80]%	(80,100]%
4	80	0	0	10	47.5	26.25	12.5	2.5	1.25
8	99	0	6.06	14.14	46.46	22.22	10.1	1.01	0
10	73	0	0	24.66	24.66	27.4	12.33	8.22	2.74

5.3 Results of Simulated Test Administrations: Spanish

This section presents the simulation results for the Spanish tests. The Spanish item pool consisted of a subset of ICCR items that had Spanish translations available. Table 22 presents the number of items available for the Spanish tests in spring 2022.

Table 22. Spring 2022 Spanish Operational Item Pool

Grade	Item Type	Total Number of Items
4	Cluster	15
	Stand-Alone	26
8	Cluster	18
	Stand-Alone	39
10	Cluster	15
	Stand-Alone	28
Total		**141**

Simulations were evaluated for all content areas using 1,000 simulated cases per grade.

5.3.1 Summary of Blueprint Match

The simulation results showed no blueprint violations at all content levels for all three grades.

5.3.2 Item Exposure

Table 23 presents the percentage of items that fell into each exposure bin for all grades. Most items were administered in more than 20% of the test administrations. Some items had an exposure rate of 100% because of the limited Spanish item pool. Only those items were available to satisfy the blueprint constraints.

Table 23. Item Exposure Rates by Grade: Percentage of Items by Exposure Rate Across All Spanish Simulation Sessions

Grade	Total Items	0%	(0,1]%	(1,5]%	(5,20]%	(20,40]%	(40,60]%	(60,80]%	(80,100]%
A	41	0	0	7.32	14.63	24.39	26.83	14.63	12.2
8	57	0	1.75	5.26	29.82	29.82	21.05	10.53	1.75
10	43	0	0	6.98	23.26	23.26	18.6	20.93	6.98

6. Operational Test Administration Summary Report

This section presents the blueprint match reports and item exposure rates for the spring 2022 operational test administrations.

6.1 Blueprint Match

Table 24 presents the percentages of the spring 2022 tests that aligned with the blueprint requirements. Every English and Spanish test administered met the blueprint specifications with a 100% match at all content levels across all grades.

Table 24. Spring 2022 Blueprint Match for Tests Delivered, Science

Grade	Content Level	MinItems	MaxItems	% of Cases Meeting BP	% of Cases Violating BP			
					1	2	-1	-2
			English					
5	Discipline	6	6	100	-	-	-	-
5	Discipline–Cluster	2	2	100	-	-	-	-
5	Discipline–Stand-alone	4	4	100	-	-	-	-
5	DCI	0	3	100	-	-	-	-
5	DCI–Cluster	0	1	100	-	-	-	-
5	DCI–Stand-alone	0	2	100	-	-	-	-
5	Performance Expectation (PE)	0	1	100	-	-	-	-
8	Discipline	6	6	100	-	-	-	-
8	Discipline–Cluster	2	2	100	-	-	-	-
8	Discipline–Stand-alone	4	4	100	-	-	-	-
8	DCI	0	3	100	-	-	-	-
8	DCI–Cluster	0	1	100	-	-	-	-
8	DCI–Stand-alone	0	2	100	-	-	-	-
8	PE	0	1	100	-	-	-	-
11	Discipline (Physical Science)	6	6	100	-	-	-	-
11	Discipline (Physical Science)–Cluster	2	2	100	-	-	-	-
11	Discipline (Physical Science)–Stand-alone	4	4	100	-	-	-	-
11	DCI (Physical Science)	0	3	100	-	-	-	-
11	DCI (Physical Science)–Cluster	0	1	100	-	-	-	-
11	DCI (Physical Science)–Stand-alone	0	2	100	-	-	-	-
11	Discipline (Life Science)	12	12	100	-	-	-	-
11	Discipline (Life Science)–Cluster	4	4	100	-	-	-	-
11	Discipline (Life Science)–Stand-alone	8	8	100	-	-	-	-
11	DCI (Life Science)	0	4	100	-	-	-	-
11	DCI (Life Science)–Cluster	0	1	100	-	-	-	-
11	DCI (Life Science)–Stand-alone	0	3	100	-	-	-	-
11	PE	0	1	100	-	-	-	-

Grade	Content Level	MinItems	MaxItems	% of Cases Meeting BP	% of Cases Violating BP			
					1	2	-1	-2
			Spanish					
5	Discipline	6	6	100	-	-	-	-
	Discipline–Cluster	2	2	100	-	-	-	-
	Discipline–Stand-alone	4	4	100	-	-	-	-
	DCI	0	3	100	-	-	-	-
	DCI–Cluster	0	1	100	-	-	-	-
	DCI–Stand-alone	0	2	100	-	-	-	-
	PE	0	1	100	-	-	-	-
8	Discipline	6	6	100	-	-	-	-
	Discipline–Cluster	2	2	100	-	-	-	-
	Discipline–Stand-alone	4	4	100	-	-	-	-
	DCI	0	3	100	-	-	-	-
	DCI–Cluster	0	1	100	-	-	-	-
	DCI–Stand-alone	0	2	100	-	-	-	-
	PE	0	1	100	-	-	-	-
11	Discipline (Physical Science)	6	6	100	-	-	-	-
	Discipline (Physical Science)–Cluster	2	2	100	-	-	-	-
	Discipline (Physical Science)–Stand-alone	4	4	100	-	-	-	-
	DCI (Physical Science)	0	3	100	-	-	-	-
	DCI (Physical Science)–Cluster	0	1	100	-	-	-	-
	DCI (Physical Science)–Stand-alone	0	2	100	-	-	-	-
	Discipline (Life Science)	12	12	100	-	-	-	-
	Discipline (Life Science)–Cluster	4	4	100	-	-	-	-
	Discipline (Life Science)–Stand-alone	8	8	100	-	-	-	-
	DCI (Life Science)	0	4	100	-	-	-	-
	DCI (Life Science)–Cluster	0	1	100	-	-	-	-
	DCI (Life Science)–Stand-alone	0	3	100	-	-	-	-
	PE	0	1	100	-	-	-	-

6.2 ITEM EXPOSURE

Table 25 presents the item exposure rates for the spring 2022 test administration. In general, the exposure rates were very similar to the simulation results described in Section 5.2.2, Item Exposure, for the English test administrations. The item exposure rate for field-test items ranged from 15.96% to 17.14% across three grades. For the Spanish tests, more items had high exposure rates compared to the English tests due to a smaller item pool. In addition, the operational exposure rates were slightly different from the simulation results because of the small population sizes in all three grades. In spring 2022, five students or fewer took the Spanish test in each grade.

Table 25. Item Exposure Rates by Grade: Percentage of Items by Exposure Rate Across All Spring 2022 Test Administrations

Grade	Total Items	0%	(0,1]%	(1,5]%	(5,20]%	(20,40]%	(40,60]%	(60,80]%	(80,100]%
English									
4	80	0	0	11.25	45	25	15	2.5	1.25
8	99	0	3.03	20.2	43.43	18.18	12.12	3.03	0
10	73	0	0	28.77	28.77	16.44	13.7	10.96	1.37
Spanish									
4	41	36.59	0	0	0	21.95	0	14.63	26.83
8	57	24.56	0	0	28.07	22.81	15.79	7.02	1.75
10	43	25.58	0	0	23.26	9.3	11.63	18.6	11.63

7. References

Council of Chief State School Officers. (2015). *Science Assessment Item Collaborative (SAIC) Assessment Framework for the Next Generation Science Standards*. Washington, DC: Council of Chief State School Officers. Retrieved from: https://ccsso.org/sites/default/files/2017-12/SAICAssessmentFramework_FINAL.pdf.

Mislevy, R. J., Behrens, J. T., Dicerbo, K. E., & Levy, R. (2012). Design and discovery in educational assessment: Evidence-centered design, psychometrics, and educational data mining. *Journal of Educational Data Mining, 4*(1), 11–48.

Mislevy, R. J., & Riconscente, M. M. (2011). Evidence-centered assessment design. In T. M. Haladyna & S. M. Downing (Eds.), Handbook of test development (pp. 75–104). New York: Routledge.

National Research Council. (2012). *A framework for K–12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: The National Academies Press.