Oregon Statewide Assessment System Science Assessment

2021–2022

Volume 2: Test Development

OREGON
DEPARTMENT OF EDUCATION
Oregon achieves . . . together!

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

1. Introduction

The Oregon State Board of Education adopted the Next Generation Science Standards (NGSS) in March of 2014. The Oregon Department of Education (ODE) and its assessment vendor, Cambium Assessment, Inc. (CAI; formerly the American Institutes for Research [AIR]), developed and administered a new online assessment to measure the new standards. The Oregon Statewide Assessment System (OSAS) Science Assessment was first administered operationally in 2018–2019. The OSAS Science Assessment uses an online, adaptively constructed test that makes use of technology-enhanced item types to measure the science knowledge and skills of Oregon students in grades 5, 8, 11. The content measures the three-dimensional science standards based on the National Research Council's *A Framework for K–12 Science Education* published in 2012.

Additional details on the implementation of the assessments can be found in Volume 1, of this technical report.

The interpretation, usage, and validity of test scores rely heavily on how the OSAS Science Assessment was developed. This volume of the technical report provides details on the test development process that contributes the validity of the test scores. Specifically, this volume provides evidence to support the following:

- The item Specifications offered detailed guidance for item writers and reviewers in order to ensure that science items were aligned to the performance expectations (PEs) they were intended to measure.
- The item development procedures employed for OSAS Science Assessment tests 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, 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 technical report focuses on the general test development activities.

For the OSAS Science Assessment, items are drawn from the Shared Science Assessment Item Bank, which consists of ICCR items, items owned by Oregon, and items owned by several other states that have signed a Memorandum of Understanding (MOU) to share content, leadership, and new ideas and methods. Specifically, all items developed under the MOU went through the same development process. In the remainder of this volume, the term *item bank* will refer to all items developed under the MOU 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 over August 22–23, 2016. The overarching goal of these 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 this end, CAI invited content and assessment leaders from 10 states as well as four nationally recognized experts that helped to author 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. The process began with detailed item specifications. The specifications, discussed in Section 2.2, Item Specifications, described the interaction types that could be used, gave guidelines for targeting the appropriate cognitive engagement, offered suggestions for controlling item difficulty, and provided sample items.

Items were written with the goal that virtually every item would be accessible to all students, either by itself or in conjunction with accessibility tools, such as text-to-speech (TTS), translations, or assistive technologies. This goal is supported by the delivery of the items on CAI's Test Delivery System (TDS), which has received Web Content Accessibility Guidelines (WCAG) 2.0 AA certification, offers a wide array of accessibility tools and is compatible with most assistive technologies.

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

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

These principles and the processes that support them have been incorporated into an item bank that measures the PEs with fidelity and does so in a way that minimizes construct-irrelevant variance and reduces barriers to access. The details of these processes follow.

1.3 Organization of This Volume

This volume addresses three subsequent sections:

1. An overview of the science item development process that supports the validity of the claims that science tests are designed to support.
2. An overview of the Shared Science Assessment Item Bank, the types of assessments the item bank is designed to support, and methods for refreshing the item bank.
3. A description of the test construction process for the OSAS Science Assessment, including the blueprint, the test design, an evaluation of 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 Memorandum of Understanding (MOU) participants by 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 each item change and comment. Reviewers, including internal CAI staff and external stakeholders in committee meetings, can review items in ITS as they will appear to the student, with all accessibility features and tools.

The process begins with the definition of item specifications, and continues with

- selection and training of item writers;
- writing and internal review of items;
- review by state personnel and stakeholder committees;
- markup for translation and accessibility features;
- field testing; and
- post-field-test reviews.

Each step of test development plays a role in ensuring that the items can support the claims on which they will be based. Table 1 shows how each step contributes to that goal and describes the steps in more detail.

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

Developmental Step	Supports Alignment to the Performance Expectations	Reduces Construct-Irrelevant Variance Through Universal Design	Expands Access Through Linguistic and Other Supports
Item specifications	Specifies item interactions, content limits, and guidelines for meeting task demands and levels of cognitive engagement requirements and adjusting difficulty.	Avoids the use of any item interactions with accessibility constraints and provides language guidelines. Allows for multiple response modes to accommodate different styles.	
Selection and training of item writers	Ensures that item writers have the background to understand the performance expectations and specifications. Teaches item writers about selection of item interactions for measurement and accessibility.	Training in language accessibility, bias, and sensitivity helps item writers avoid unnecessary barriers.	
Writing and internal review of items	Checks content alignment and evaluates and improves overall quality.	Eliminates editorial issues and flags and removes bias and accessibility issues.	
Markup for translation and accessibility features		Adds universal features, such as TTS for science, that reduce barriers.	Adds TTS, braille, translations, American Sign Language (ASL) and glossaries.
Review by state personnel and stakeholder committees	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 for issues.	May reveal usability or implementation issues with markup.
Post-field-test reviews	Provides final, more focused checks on flagged items. Rubric validation ensures that scoring reflects PEs.	Final, focused review on items flagged for differential item functioning (DIF).	

2.2 Item Specifications

CAI is working with a group of states and 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. Thirteen states (Connecticut, Hawaii, Idaho, Montana, New Hampshire, North Dakota, Oregon, Rhode Island, South Dakota, Utah, Vermont, West Virginia, and Wyoming) and one U.S. territory (the U.S. Virgin Islands) have participated in this initiative. This collaboration has yielded item specifications for PEs, sample item clusters for some specifications, and hundreds of science item clusters and stand-alone items. Under this collaboration, utilizing guidelines for item specifications proposed by WestEd in collaboration with the Council of Chief State School Officers (CCSSO), state and territory members, and content experts (CCSSO, 2015), states and one U.S. territory developed item specifications jointly.

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 comprising each science standard and which together form coherent item clusters and stand-alone items. Science item specifications include the following elements:

- **Performance Expectation.** This identifies the PE being assessed.
- **Dimensions.** This identifies the Science and Engineering Practices (SEPs), Crosscutting Concepts (CCCs), and Disciplinary Core Ideas (DCIs) that the PE assesses.
- **Clarifications and Content Limits.** This delineates the specific content that the PE measures and the parameters in which items must be developed to assess the PE accurately, including the lower and upper complexity limits of items. Specifically, content limits refine the intent of the PE 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.** This section identifies the relevant technical words that students are expected to know, and 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 imagination of the writers.
- **Content/Phenomena.** This section provides examples of the types of phenomena that would support the effective items related to the PE in question. In general, these are guideposts, and item writers seek comparable phenomena, rather than drawing on those within the documents.
- **Task Demands.** In this section, the PEs and associated evidence statements are broken down into specific task demands aligned to each PE. Task demands denote the specific ways in which students will 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. Each item should be clearly linked to one or more of the task demands, and the verbs guide the types of interactions writers might employ to elicit the student response.

Table 2 provides an example of the item specifications developed by content experts for a middle school Life Sciences PE.

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

<table>
<tr><td>Performance Expectation</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 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>
<tr><td>Science Vocabulary Students Are Not Expected to Know</td><td colspan="3">Differentiation, mitosis, meiosis, genetics, cellular respiration, energy transfer, RNA, protozoa, amoeba, histology, protista, archaea, nucleoid, plasmid, diatoms, cyanobacteria</td></tr>
<tr><td colspan="4" align="center">Phenomena</td></tr>
<tr><td>Context/ Phenomena</td><td colspan="3">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.</td></tr>
</table>

This Performance Expectation and Associated Evidence Statements Support the Following Task Demands
Task Demands
1. Identify from a list, including distractors, the materials/tools 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.

Note. [a]MS-LS1-1 is the performance expectation code for Middle School Life Sciences 1-1.
[b]Denotes those task demands which are 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 Selection and Training of Item Writers

All item writers developing science items at CAI have at least a bachelor's degree, and many bring 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, Appendix L and Appendix L. These include

- CAI's Language Accessibility, Bias, and Sensitivity Guidelines; and
- a training (presented using Microsoft PowerPoint) for the appropriate use of item interactions.

2.4 Internal Review

CAI's test development structure utilizes 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 items prior to the client review and provide training and feedback for all content-area team members.

ICCR and MOU science items go through a rigorous, multiple-level internal review process before they are sent to external review. Staff members are trained to review items for both content and accessibility throughout the entire process. A sample item review checklist that CAI test development use is included in Appendix B, Item Review Checklist. The ICCR and MOU science 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

Preliminary Review is performed by team leads or senior content staff. On occasion the Preliminary Review is conducted in a group setting, led by a senior test developer. During the process, team leads or senior content staff analyze items to ensure that the following criteria have been met:

- The item aligns with the performance expectation.
- 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 worthwhile in a reasonable way/it is a discrete observation that grounds a scenario, which allows for the meaningful assessment of something worthwhile).
- The item is properly aligned to 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 know 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 correct option); and
- free of obvious or subtle cuing.

2.4.2 Scoring Entry and Review

Before an item advances to the Content Review One level, the item writer inputs the machine scoring so that it can be reviewed by the team lead or senior staff that is reviewing the item prior to Content Review One. This step is kept separate from Preliminary Review so that the senior staff can suggest changes to the interaction at Preliminary Review without requiring the writer to overhaul the scoring that has already been created. It also allows the senior staff to ensure that the scoring proposed by the item writer is appropriate. This sequence ensures that the scoring is entered once, streamlining the process. At this level, the scoring is analyzed to ensure that the following criteria are met:

- The scoring functions as intended (i.e., the student gets 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 setup is unambiguous and matches the questions asked (i.e., if students are asked to round to a certain decimal place, their answer is scored accordingly).

Senior staffs approve the intent of the scoring from the Preliminary Review. At the Scoring Entry level, item writers input the approved scoring. After that, senior staffs check the functionality of the scoring. Once the scoring is determined to be working correctly, senior staffs sign-off on the item and move 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 all the criteria identified for Preliminary Review. He or she also ensures 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 applies his or her own experience 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 make sure 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 keys are correct and that all information in items is 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 items reflect the accepted guidelines for good item construction. In all items, they ensure that language 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 and that the key accurately and correctly answers the question as posed, that the answer is not inappropriately obvious, and that it 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 PE, and consistency with expectations for the highest quality. They check whether the scoring is working as intended and 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 have been through an exhaustive external review process. Items in the Shared Science Assessment Item Bank were reviewed by content experts in one or several states or territory and reviewed and approved by multiple stakeholder committees to evaluate both content and bias/sensitivity.

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 prior to committee review. At this stage in the review process, clients can request edits, such as wording edits, scoring edits, alignment changes, or task demand updates. An 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 have the option to present these items to the committee (based on the edits made) or withhold them from committee review.

ICCR items are reviewed by at least one or two states or territory. The states or territory provide feedback on the ICCR items, and the CAI science leadership gathers suggestions and makes edits that improve the ICCR item. Not all suggestions are implemented, as these items are owned by CAI. 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 already at 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 expectation. CAC members are typically grade-level and subject-matter experts. During this review, educators also ensure that the scoring assertions make clear what is being scored as correct and give credit where they should (see Section 2.7.1, Rubric Validation, for more information).

Items developed for each state under the MOU are reviewed by the state that owns the items. ICCR items are reviewed by the CAC of one or more states or territory. In most cases, items are seen by multiple state or territory committees prior to their field-test or operational use.

In 2022, all MOU states engaged in a single CAC process in which participants from multiple states reviewed items. The items were edited and returned to their respective owner states for final approval.

A summary of the 2021-2022 committee meetings appears in Table 3, with further details about the participants in Appendix C, Content Advisory Committee Participant Details. Appendix C 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]

Note. [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 include representatives from student populations such as Special Education, low vision, and the hearing impaired. Furthermore, diverse members of this committee represent the interests of students from various ethnic and economic backgrounds in order to ensure that test items are free of bias and sensitivity concerns.

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

In 2022, the MOU states were all involved in a single review process whereby participants from multiple states would review items. The items were edited and then returned to the owning state for final approval.

A summary of the committee meetings appears in Table 4, with additional details about the participants in Appendix D, Fairness Committee Participant Details. Appendix D 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

Note. [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 test. 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 for 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.

2.6 Field-Testing

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

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

In 2022, a fourth wave 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 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. This included rubric validation and data review. These processes are described in the following sections.

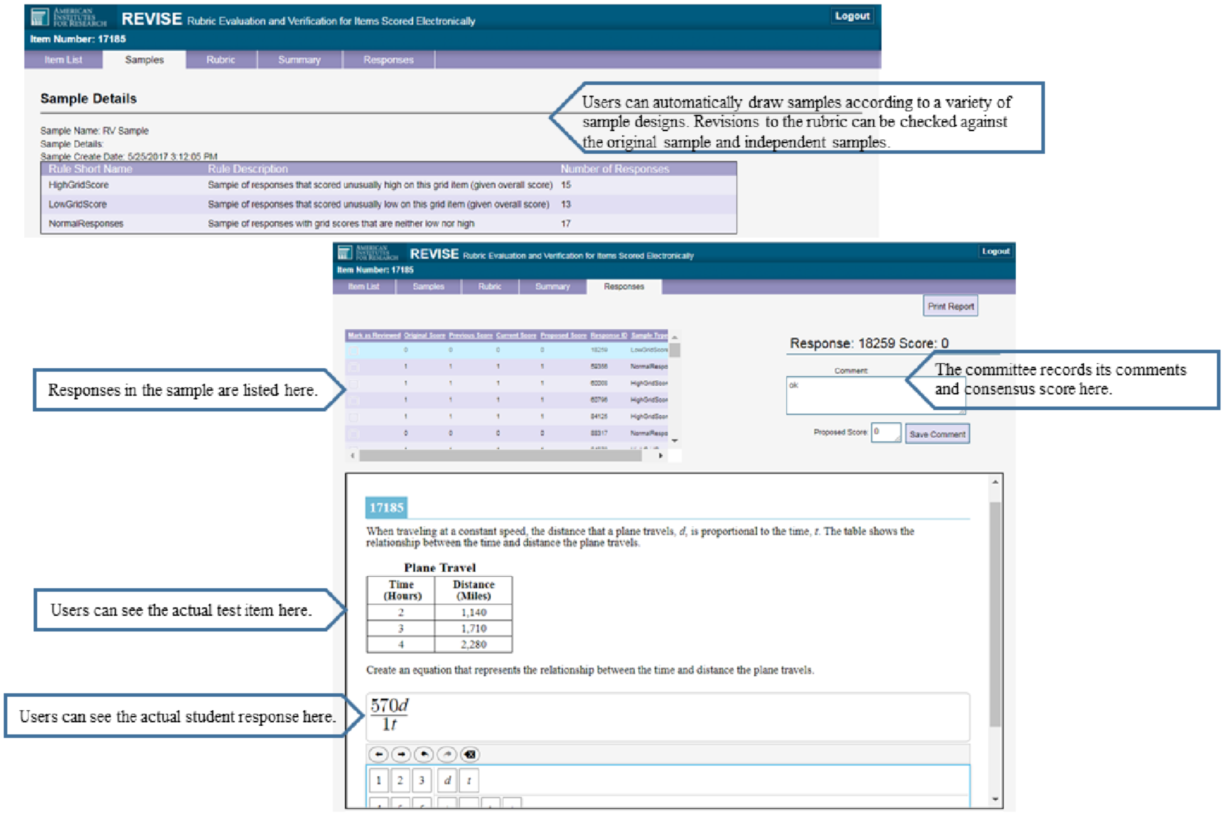
2.7.1 Rubric Validation

The validation process of 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. The 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: (a) low-scored responses from otherwise high-scoring students, and (b) 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 that received credit but should not have (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 to confirm the rubrics of each item. The committee recommendations are then discussed with the state of ownership to resolve any inconsistencies. The rubrics are then 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; any items rejected as part of the process are rejected in the Item Tracking System (ITS), which archives critical information regarding the scoring certification completed during the rubric validation process. This includes recording 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 verifying 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 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 utilized either an independent field test or operational test).

Volume 1, Section 4, Field-Test Classical Analysis Overview, describes in detail the statistical flags that send items to 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.

For each of the states participating in the MOU, flagged items owned by the state were reviewed by a data review committee. The composition of the data review committees generally 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 taken to committee members from several states participating in the MOU. Outcomes were decided by CAI science content leadership, taking the committees' recommendations into consideration.

At the start of each state-owned item data review meeting, CAI staff leads participants in a training session to familiarize them with the item development process, the purpose of data review, the meaning of the various flags, and the purpose of the data review committee. Committee members are taught to interpret the various flags and are given guidelines for examining the items for content or fairness issues. The training includes a group review of item cards, which detail specific item attributes (including grade level and alignment to the science performance expectations, the content and rubric of the item, and the various item statistics). A sample of the training materials used for these data review meetings appears in Appendix E, Sample Data Review Training Materials. Participants use 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 are then reviewed by the participants who are most familiar with the particular grade (or grade band) level and the items' content domain. CAI content specialists, who are also well versed in item statistics, facilitate the discussion in each room with CAI psychometricians available to answer questions as they arise. At the end of each meeting day, CAI content specialists meet with the 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 2022 data review committee meetings. Details, including the composition of each committee, appear in Appendix F, 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**
			Cluster	11	2

Owner	Meeting	Number of Committee Members	Item Type	Number of Items Reviewed	Number of Items Rejected
			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

Owner	Meeting	Number of Committee Members	Item Type	Number of Items Reviewed	Number of Items Rejected
			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
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

Owner	Meeting	Number of Committee Members	Item Type	Number of Items Reviewed	Number of Items Rejected
			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

Note. [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) 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 composed 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 also 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, science items were written to the three-dimensional science standards. The Shared Science Assessment Item Bank is comprised of ICCR items and items developed for specific states, which are all shared with MOU partner states. 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 state or territory using 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. Participating states and territory collectively share the items and agree to field-test new items each year.

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 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, therefore not supporting the assertion that the student can interpret relationships expressed graphically. Table 6 lists the science interaction types. Examples of various interaction types can be found in Appendix G, Sample Item Interactions.

Table 6. Science Interaction Types and Descriptions

Interaction Type	Associated Sub-Types	Description
Choice	Multiple-Choice	Traditional multiple-choice interaction allows the student 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 response 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	Interaction allows students to check a box to indicate if the information from a column header matches information from a row header.

Interaction Type	Associated Sub-Types	Description
	Table Input	Interaction solicits a student to complete tabular data.
Edit Task	Edit Task	A student clicks a word and replaces it with another word that they type to revise a sentence.
	Edit Task with Choice	A student clicks a word or phrase and chooses the replacement from a number of options.
	Edit Task Inline Choice	Drop-down menus are placed through the text, and a student chooses the right 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.
Equation	N/A	Equation interactions require students to enter a response into input boxes. These boxes may stand alone, or they may be in line with text or embedded in a table. The equation interaction may have an on-screen keypad which may consist of special mathematic characters. Students may also enter their response 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 allow you to create 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 allow you to create 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 the student to investigate a phenomenon by selecting variables to get output data. Some simulations are accompanied by animations.

Table 7 through Table 11 on the following pages provide the number of items in the Shared Science Assessment Item Bank available for use in the spring 2022 statewide assessments. Appendix H, Shared Science Assessment Item Bank, provides the items within the bank available by grade band, PE, and origin.

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

Grade Band and Item Type	ICCR Items	Oregon Items	MOU Items[a]	Total Bank Items[b]
Elementary School	**148**	**41**	**384**	**573**
Cluster	49	25	220	294
Stand-Alone	99	16	164	279
Middle School	**163**	**41**	**391**	**595**
Cluster	55	23	215	293
Stand-Alone	108	18	176	302
High School	**145**	**40**	**251**	**436**
Cluster	50	25	97	172
Stand-Alone	95	15	154	264
Total	**456**	**122**	**1026**	**1604**

Note. [a]Other MOU state item sources include Connecticut, Hawaii, Idaho, Montana, MSSA (Rhode Island and Vermont), Utah, West Virginia, and Wyoming.
[b]Count excludes 85 legacy items (60 Oregon-owned and 25 South Dakota-owned).

Table 8. Spring 2022 Shared Science Assessment Operational Item Bank

Grade Band and Item Type	ICCR Operational Items	Oregon Operational Items	MOU Operational Items[a]	Total Bank Operational Items[b]
Elementary School	**116**	**22**	**265**	**403**
Cluster	40	16	150	206
Stand-Alone	76	6	115	197
Middle School	**101**	**19**	**185**	**405**
Cluster	29	12	164	205
Stand-Alone	72	7	121	200
High School	**103**	**23**	**199**	**325**
Cluster	37	15	79	131
Stand-Alone	66	8	120	194
Total	**320**	**64**	**749**	**1133**

Note. [a]Other MOU state operational item sources include Connecticut, Hawaii, Idaho, MSSA (Rhode Island and Vermont), Utah, West Virginia, and Wyoming.
[b]Count excludes 15 South Dakota-owned legacy operational items.

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

Grade Band and Item Type	ICCR Field-Test Items	Oregon Field-Test Items	MOU Field-Test Items[a]	Total Bank Field-Test Items[b]
Elementary School	**32**	**19**	**119**	**170**

Grade Band and Item Type	ICCR Field-Test Items	Oregon Field-Test Items	MOU Field-Test Items[a]	Total Bank Field-Test Items[b]
Cluster	9	9	70	88
Stand-Alone	23	10	49	82
Middle School	**62**	**22**	**106**	**190**
Cluster	26	11	51	88
Stand-Alone	36	11	55	102
High School	**42**	**17**	**52**	**111**
Cluster	13	10	18	41
Stand-Alone	29	7	33	70
Total	**136**	**58**	**277**	**471**

Note. [a]Other MOU state field-test item sources include Connecticut, Hawaii, Idaho, MSSA (Rhode Island and Vermont), Utah, West Virginia, and Wyoming.
[b]Count excludes 70 legacy field-test items (60 Oregon-owned and 10 South Dakota-owned).

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

Grade Band	Science Discipline	Item Type	ICCR Items	Oregon Items	MOU Items[a]	Total Bank Items[b]
Elementary School	Earth and Space Sciences	Cluster	18	7	73	99
		Stand-Alone	29	5	54	88
	Life Sciences	Cluster	14	11	57	82
		Stand-Alone	32	5	46	83
	Physical Sciences	Cluster	17	7	89	113
		Stand-Alone	38	6	64	108
Middle School	Earth and Space Sciences	Cluster	16	5	57	78
		Stand-Alone	29	5	50	84
	Life Sciences	Cluster	22	9	82	113
		Stand-Alone	47	7	60	114
	Physical Sciences	Cluster	17	9	69	95
		Stand-Alone	32	6	65	103
High School	Earth and Space Sciences	Cluster	13	8	16	37
		Stand-Alone	22	5	33	60
	Life Sciences	Cluster	20	8	52	80
		Stand-Alone	50	3	71	124
	Physical Sciences	Cluster	17	9	28	54
		Stand-Alone	23	7	50	80
Total			**456**	**122**	**1017**	**1595**

Note. [a]Other MOU state item sources include Connecticut, Hawaii, Idaho, Montana, MSSA (Rhode Island and Vermont), Utah, West Virginia, and Wyoming.
[b]Count excludes 85 legacy items (60 Oregon-owned and 25 South Dakota-owned) and 9 MOU items that do not align to the 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	Oregon Items	MOU Items[a]	Total Bank Items[b]
Elementary School	Earth and Space Sciences	ESS1	12	3	35	50
		ESS2	15	8	56	79
		ESS3	20	1	37	58
	Life Sciences	LS1	16	6	42	64
		LS2	6	3	19	28
		LS3	5	3	13	21
		LS4	19	4	29	52
	Physical Sciences	PS1	14	4	37	55
		PS2	15	6	32	53
		PS3	20	2	55	77
		PS4	6	1	29	36
Middle School	Earth and Space Sciences	ESS1	15	2	29	46
		ESS2	16	4	38	58
		ESS3	14	4	40	58
	Life Sciences	LS1	22	5	49	76
		LS2	24	4	39	67
		LS3	5	0	19	24
		LS4	18	7	35	60
	Physical Sciences	PS1	13	6	43	62
		PS2	6	4	39	49
		PS3	19	2	34	55
		PS4	11	3	17	32
High School	Earth and Space Sciences	ESS1	12	4	13	29
		ESS2	11	5	20	36
		ESS3	12	4	16	32
	Life Sciences	LS1	20	5	33	58
		LS2	21	3	38	62
		LS3	11	2	16	29
		LS4	18	1	36	55
	Physical Sciences	PS1	19	6	27	52
		PS2	9	3	20	32

Grade Band	Science Discipline	Disciplinary Core Idea	ICCR Items	Oregon Items	MOU Items[a]	Total Bank Items[b]
		PS3	5	4	18	27
		PS4	7	3	13	23
Total			**456**	**122**	**1017**	**1595**

Note. [a]Other MOU state item sources include Connecticut, Hawaii, Idaho, Montana, MSSA (Rhode Island and Vermont), Utah, West Virginia, and Wyoming. [b]Count excludes 85 legacy items (60 Oregon-owned and 25 South Dakota-owned) and 9 MOU items that do not align to the NGSS.

3.2 STRATEGY FOR ITEM BANK EVALUATION AND REPLENISHMENT

Both CAI and the participating MOU states continue to develop items to replenish and grow the Shared Science Assessment Item Bank. The general strategy for targeting item development gathers 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 to identify gaps in the 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 that are seen by more than 20% of students. The characteristics of these items are the primary targets for development. Overused items become candidates for release two years, once replacements have been introduced into the operational bank.

4. CONSTRUCTION OF THE OSAS SCIENCE ASSESSMENT

4.1 TEST DESIGN

The Oregon Statewide Assessment System (OSAS) Science Assessment was administered online to students in grades 5, 8, and 11 using an adaptive test design in spring 2022. Appendix I, OSAS Science Assessment Item Pool, provides the 2022 item pool by grade band, performance expectation (PE), and origin. In an adaptive test, operational items are selected on the fly based on the performance of a student on past items while ensuring the test blueprint is followed for each individual 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. Also, as opposed to a fixed form and a LOFT, every student has the potential to see a different set of items that adapt to the student's ability, thus offering a better testing experience.

Items are selected by an item-selection algorithm based on the 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 item 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 ELA and mathematics tests but with some modifications to make it suitable for using item clusters. Specifically, the proficiencies that are estimated during the test are computed under an item response theory (IRT) model that incorporates cluster effects. In order to avoid over-selection of 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). Details for CAI's adaptive testing algorithm are described in Appendix J, Adaptive Algorithm Design.

A non-segmented test design was used. Students received items from different disciplines in a random order. In an adaptive test, the use of a non-segmented test design provides more freedom when selecting items targeting a current best estimate of proficiency. 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 field-test items throughout the test.

4.2 Test Blueprints

Test blueprints provide the following guidelines:

- Length of the test; and
- Science disciplines to be covered and the acceptable number of items across performance expectations (PEs) within each science discipline and disciplinary core idea (DCI).

The blueprint for science is given in Table 12 through Table 14.

Table 12. Science Test Blueprint, Grade 5

Grade 5	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 Sciences, PE Total = 17	**1**	**2**	**4**	**5**	**6**	**6**
DCI—Motion and Stability: Forces and Interactions	**0**	**1**	**0**	**2**	**0**	**3**
3-PS2-1: Forces-balanced and unbalanced forces	0	1	0	1	0	1
3-PS2-2: Forces-pattern predicts future motion	0	1	0	1	0	1
3-PS2-3: Forces-between objects not in contact	0	1	0	1	0	1
3-PS2-4: Forces-magnets*	0	1	0	1	0	1
5-PS2-1: Space Systems	0	1	0	1	0	1
DCI—Energy	**0**	**1**	**0**	**2**	**0**	**3**
4-PS3-1: Energy-relationship between speed and energy of object	0	1	0	1	0	1
4-PS3-2: Energy-transfer of energy	0	1	0	1	0	1
4-PS3-3: Energy-changes in energy when objects collide	0	1	0	1	0	1
4-PS3-4: Energy-converting energy from one form to another*	0	1	0	1	0	1
5-PS3-1: Matter and Energy	0	1	0	1	0	1
DCI—Waves and Their Applications in Technologies for Information Transfer	**0**	**1**	**0**	**2**	**0**	**3**
4-PS4-1: Waves-waves can cause objects to move	0	1	0	1	0	1
4-PS4-2: Structure, Function, Information Processing	0	1	0	1	0	1
4-PS4-3: Waves-using patterns to transfer information*	0	1	0	1	0	1
DCI—Matter and Its Interactions	**0**	**1**	**0**	**2**	**0**	**3**
5-PS1-1: Structure and Properties of Matter	0	1	0	1	0	1
5-PS1-2: Structure and Properties of Matter	0	1	0	1	0	1

Grade 5	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
5-PS1-3: Structure and Properties of Matter	0	1	0	1	0	1
5-PS1-4: Structure and Properties of Matter	0	1	0	1	0	1
Discipline—Life Sciences, PE Total = 12	**1**	**2**	**4**	**5**	**6**	**6**
DCI—From Molecules to Organisms: Structure and Function	**0**	**1**	**0**	**3[a]**	**0**	**3**
3-LS1-1: Inheritance	0	1	0	1	0	1
4-LS1-1: Structure, Function, Information Processing	0	1	0	1	0	1
4-LS1-2: Structure, Function, Information Processing	0	1	0	1	0	1
5-LS1-1: Matter and Energy	0	1	0	1	0	1
DCI—Ecosystems: Interactions, Energy, and Dynamics	**0**	**1**	**0**	**3[a]**	**0**	**3**
3-LS2-1: Ecosystems	0	1	0	1	0	1
5-LS2-1: Matter and Energy	0	1	0	1	0	1
DCI—Inheritance and Variation of Traits	**0**	**1**	**0**	**3[a]**	**0**	**3**
3-LS3-1: Inheritance	0	1	0	1	0	1
3-LS3-2: Inheritance	0	1	0	1	0	1
DCI—Biological Evolution: Unity and Diversity	**0**	**1**	**0**	**3[a]**	**0**	**3**
3-LS4-1: Ecosystems	0	1	0	1	0	1
3-LS4-2: Inheritance	0	1	0	1	0	1
3-LS4-3: Ecosystems	0	1	0	1	0	1
3-LS4-4: Ecosystems*	0	1	0	1	0	1
Discipline—Earth and Space Sciences, PE Total = 13	**1**	**2**	**4**	**5**	**6**	**6**
DCI—Earth's Systems	**0**	**1**	**0**	**3[a]**	**0**	**3**
3-ESS2-1: Weather and Climate	0	1	0	1	0	1

Grade 5	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-ESS2-2: Weather and Climate	0	1	0	1	0	1
4-ESS2-1: Earth's Systems and Processes	0	1	0	1	0	1
4-ESS2-2: Earth's Systems and Processes	0	1	0	1	0	1
5-ESS2-1: Earth's Systems	0	1	0	1	0	1
5-ESS2-2: Earth's Systems	0	1	0	1	0	1
DCI—Earth and Human Activity	**0**	**1**	**0**	**3[a]**	**0**	**3**
3-ESS3-1: Weather and Climate*	0	1	0	1	0	1
4-ESS3-2: Earth's Systems and Processes*	0	1	0	1	0	1
4-ESS3-1: Energy	0	1	0	1	0	1
5-ESS3-1: Earth's Systems	0	1	0	1	0	1
DCI—Earth's Place in the Universe	**0**	**1**	**0**	**3[a]**	**0**	**3**
4-ESS1-1: Earth's Systems and Processes	0	1	0	1	0	1
5-ESS1-1: Space Systems	0	1	0	1	0	1
5-ESS1-2: Space Systems	0	1	0	1	0	1
PE Total = 42	**4**	**4**	**14**	**14**	**18**	**18**

Note. *These PEs have an engineering component.

[a]Because of the limitation of the item pool in the Life Sciences (LS) and Earth and Space Sciences (ESS) disciplines, the maximum number of stand-alone items allowed for these DCIs was changed from 2 to 3 while keeping the maximum number of items (item clusters + stand-alone items) allowed at 3 in the Earth's Systems DCI.

Table 13. Science Test Blueprint, Grade 8

Grade 8	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 Sciences, PE Total = 19	**1**	**2**	**4**	**5**	**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: Chemical Reactions	0	1	0	1	0	1
MS-PS1-3: Structure and Properties of Matter	0	1	0	1	0	1
MS-PS1-4: Structure and Properties of Matter	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 Interactions*	0	1	0	1	0	1
MS-PS2-2: Forces and Interactions	0	1	0	1	0	1
MS-PS2-3: Forces and Interactions	0	1	0	1	0	1
MS-PS2-4: Forces and Interactions	0	1	0	1	0	1
MS-PS2-5: Forces and 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	0	1	0	1	0	1
MS-PS3-3: Energy*	0	1	0	1	0	1
MS-PS3-4: Energy	0	1	0	1	0	1
MS-PS3-5: Energy	0	1	0	1	0	1
DCI—Waves and Their Applications in Technologies for Information Transfer	**0**	**1**	**0**	**2**	**0**	**3**
MS-PS4-1: Waves and Electromagnetic Radiation	0	1	0	1	0	1

Grade 8	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
MS-PS4-2: Waves and Electromagnetic Radiation	0	1	0	1	0	1
MS-PS4-3: Waves and Electromagnetic Radiation	0	1	0	1	0	1
Discipline—Life Sciences, PE Total = 21	**1**	**2**	**4**	**5**	**6**	**6**
DCI—From Molecules to Organisms: Structures and Processes	**0**	**1**	**0**	**2**	**0**	**3**
MS-LS1-1: Structure, Function, Information Processing	0	1	0	1	0	1
MS-LS1-2: Structure, Function, Information Processing	0	1	0	1	0	1
MS-LS1-3: Structure, Function, Information Processing	0	1	0	1	0	1
MS-LS1-4: Growth, Development, Reproduction	0	1	0	1	0	1
MS-LS1-5: Growth, Development, Reproduction	0	1	0	1	0	1
MS-LS1-6: Matter and Energy	0	1	0	1	0	1
MS-LS1-7: Matter and Energy	0	1	0	1	0	1
MS-LS1-8: Structure, Function, Information Processing	0	1	0	1	0	1
DCI—Ecosystems: Interactions, Energy, and Dynamics	**0**	**1**	**0**	**2**	**0**	**3**
MS-LS2-1: Matter and Energy	0	1	0	1	0	1
MS-LS2-2: Interdependent Relationships in Ecosystems	0	1	0	1	0	1
MS-LS2-3: Matter and Energy	0	1	0	1	0	1
MS-LS2-4: Matter and Energy	0	1	0	1	0	1
MS-LS2-5: Interdependent Relationships in Ecosystems*	0	1	0	1	0	1
DCI—Heredity: Inheritance and Variation of Traits	**0**	**1**	**0**	**2**	**0**	**3**
MS-LS3-1: Growth, Development, Reproduction	0	1	0	1	0	1

Grade 8	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
MS-LS3-2: Growth, Development, Reproduction	0	1	0	1	0	1
DCI—Biological Evolution: Unity and Diversity	**0**	**1**	**0**	**2**	**0**	**3**
MS-LS4-1: Natural Selection and Adaptation	0	1	0	1	0	1
MS-LS4-2: Natural Selection and Adaptation	0	1	0	1	0	1
MS-LS4-3: Natural Selection and Adaptation	0	1	0	1	0	1
MS-LS4-4: Natural Selection and Adaptation	0	1	0	1	0	1
MS-LS4-5: Growth, Development, Reproduction	0	1	0	1	0	1
MS-LS4-6: Natural Selection and Adaptation	0	1	0	1	0	1
Discipline—Earth and Space Sciences, PE Total = 15	**1**	**2**	**4**	**5**	**6**	**6**
DCI—Earth's Place in the Universe	**0**	**1**	**0**	**2**	**0**	**3**
MS-ESS1-1: Space Systems	0	1	0	1	0	1
MS-ESS1-2: Space Systems	0	1	0	1	0	1
MS-ESS1-3: Space Systems	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 Systems	0	1	0	1	0	1
MS-ESS2-2: History of Earth	0	1	0	1	0	1
MS-ESS2-3: History of Earth	0	1	0	1	0	1
MS-ESS2-4: Earth's Systems	0	1	0	1	0	1
MS-ESS2-5: Weather and Climate	0	1	0	1	0	1
MS-ESS2-6: Weather and Climate	0	1	0	1	0	1
DCI—Earth and Human Activity	**0**	**1**	**0**	**2**	**0**	**3**
MS-ESS3-1: Earth's Systems	0	1	0	1	0	1

Grade 8	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
MS-ESS3-2: Human Impacts	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: Weather and Climate	0	1	0	1	0	1
PE Total = 55	**5**	**5**	**13**	**13**	**18**	**18**

Note. *These PEs have an engineering component.

Table 14. Science Test Blueprint, Grade 11

Grade 11	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 Sciences, PE Total = 24	**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-PS1-3: Structure and Properties of Matter	0	1	0	1	0	1
HS-PS1-4: Chemical Reactions	0	1	0	1	0	1
HS-PS1-5: Chemical Reactions	0	1	0	1	0	1
HS-PS1-6: 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
HS-PS2-4: Types of Interactions	0	1	0	1	0	1
HS-PS2-5: Types of Interactions	0	1	0	1	0	1
HS-PS2-6: Chemical Reactions*	0	1	0	1	0	1
DCI—Energy	**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*	0	1	0	1	0	1
HS-PS3-4: Energy	0	1	0	1	0	1

Grade 11	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
HS-PS3-5: Energy	0	1	0	1	0	1
DCI—Waves and Their Applications in Technologies for Information Transfer	**0**	**1**	**0**	**2**	**0**	**3**
HS-PS4-1: Wave Properties	0	1	0	1	0	1
HS-PS4-2: Wave Properties	0	1	0	1	0	1
HS-PS4-3: Wave Properties/Electromagnetic Radiation	0	1	0	1	0	1
HS-PS4-4: Electromagnetic Radiation	0	1	0	1	0	1
HS-PS4-5: Electromagnetic Radiation*	0	1	0	1	0	1
Discipline—Life Sciences, PE Total = 24	**2**	**2**	**4**	**4**	**6**	**6**
DCI—From Molecules to Organisms: Structures and Processes	**0**	**1**	**0**	**2**	**0**	**3**
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
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**	**2**	**0**	**3**
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

Grade 11	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
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
HS-LS2-8: Social Interactions and Group Behavior	0	1	0	1	0	1
DCI—Heredity: Inheritance and Variation of Traits	**0**	**1**	**0**	**2**	**0**	**3**
HS-LS3-1: Structure and Function	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**	**2**	**0**	**3**
HS-LS4-1: Evidence of Common Ancestry and Diversity	0	1	0	1	0	1
HS-LS4-2: Natural Selection	0	1	0	1	0	1
HS-LS4-3: Natural Selection	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*	0	1	0	1	0	1
Discipline—Earth and Space Sciences, PE Total = 19	**2**	**2**	**4**	**4**	**6**	**6**
DCI—Earth's Place in the Universe	**0**	**1**	**0**	**2**	**0**	**3**
HS-ESS1-1: The Universe and Its Stars	0	1	0	1	0	1

Grade 11	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
HS-ESS1-2: The Universe and Its Stars	0	1	0	1	0	1
HS-ESS1-3: The Universe and Its Stars	0	1	0	1	0	1
HS-ESS1-4: Earth and the Solar System	0	1	0	1	0	1
HS-ESS1-5: The History of Planet Earth	0	1	0	1	0	1
HS-ESS1-6: The History of Planet Earth	0	1	0	1	0	1
DCI—Earth's Systems	**0**	**1**	**0**	**2**	**0**	**3**
HS-ESS2-1: Earth Materials and Systems	0	1	0	1	0	1
HS-ESS2-2: Earth Materials and Systems	0	1	0	1	0	1
HS-ESS2-3: Earth Materials and Systems	0	1	0	1	0	1
HS-ESS2-4: Weather and Climate	0	1	0	1	0	1
HS-ESS2-5: The Roles of Water in Earth's Surface Processes	0	1	0	1	0	1
HS-ESS2-6: Weather and Climate	0	1	0	1	0	1
HS-ESS2-7: Weather and Climate	0	1	0	1	0	1
DCI—Earth and Human Activity	**0**	**1**	**0**	**2**	**0**	**3**
HS-ESS3-1: Natural Resources	0	1	0	1	0	1
HS-ESS3-2: Natural Resources*	0	1	0	1	0	1
HS-ESS3-3: Human Impacts on Earth Systems	0	1	0	1	0	1
HS-ESS3-4: Human Impacts on Earth Systems*	0	1	0	1	0	1
HS-ESS3-5: Global Climate Change	0	1	0	1	0	1
HS-ESS3-6: Global Climate Change*	0	1	0	1	0	1
PE Total = 67	**6**	**6**	**12**	**12**	**18**	**18**

Note. *These PEs have an engineering component.

Main characteristics of the blueprint were that any performance expectation (PE) could be tested only once (indicated by the values of 0 and 1 for the minimum and maximum values of the individual PEs in Table 12 through Table 14); in general, no more than one item cluster or two stand-alone items could be sampled from the same disciplinary core idea (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 Oregon Department of Education (ODE) published estimated testing times for the OSAS Science Assessment. The 85th percentile of the testing times (in minutes) is presented in Table 15.

Table 15. OSAS Science Assessment 85th Percentile Testing Times by Grade

Subject	Grade	85th Percentile Testing
Science	5	83.12
	8	80.52
	11	79.57

4.3 TEST CONSTRUCTION

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

The 2022 OSAS Science Assessment item pools were built by CAI test developers to match items exactly to the detailed test blueprints. Operational items were selected from ten item banks (ICCR, Connecticut, Hawaii, Idaho, Montana, MSSA [Rhode Island and Vermont], Oregon, Utah, West Virginia, and Wyoming) to fulfill the blueprints. Table 16 through Table 20 summarize the 2022 OSAS Science Assessment item pool. Appendix I, OSAS Science Assessment Item Pool, provides the 2022 item pool by grade, PE, and origin.

Table 16. Spring 2022 OSAS Science Assessment Item Pool

Grade and Item Type	ICCR Items	Oregon Items	MOU Items[a]	Total Item Pool[b]
Grade 5	**116**	**41**	**242**	**399**
Cluster	37	25	136	198
Stand-Alone	79	16	106	201
Grade 8	**86**	**41**	**197**	**324**
Cluster	22	23	109	154
Stand-Alone	64	18	88	170
Grade 11	**109**	**40**	**180**	**329**

Grade and Item Type	ICCR Items	Oregon Items	MOU Items[a]	Total Item Pool[b]
Cluster	39	25	68	132
Stand-Alone	70	15	112	197
Total	**311**	**122**	**619**	**1052**

Note. [a]Other MOU states include Connecticut, Hawaii, Idaho, Montana, MSSA (Rhode Island and Vermont), Utah, West Virginia, and Wyoming.
[b]Count excludes 60 Oregon-owned legacy items.

Table 17. Spring 2022 OSAS Science Assessment Operational Item Pool

Grade and Item Type	ICCR Operational Items	Oregon Operational Items	MOU Operational Items[a]	Total Operational Pool Items
Grade 5	**108**	**22**	**228**	**358**
Cluster	37	16	130	183
Stand-Alone	71	6	98	175
Grade 8	**77**	**19**	**189**	**285**
Cluster	19	12	107	138
Stand-Alone	58	7	82	147
Grade 11	**91**	**23**	**172**	**286**
Cluster	35	15	66	116
Stand-Alone	56	8	106	170
Total	**276**	**64**	**589**	**929**

Note. [a]Other MOU states include Connecticut, Hawaii, Idaho, Montana, MSSA (Rhode Island and Vermont), Utah, West Virginia, and Wyoming.

Table 18. Spring 2022 OSAS Science Assessment Field-Test Item Pool

Grade and Item Type	ICCR Field-Test Items	Oregon Field-Test Items	MOU Field-Test Items[a]	Total Field-Test Pool Items[b]
Grade 5	**8**	**19**	**14**	**41**
Cluster	0	9	6	15
Stand-Alone	8	10	8	26
Grade 8	**9**	**22**	**8**	**39**
Cluster	3	11	2	16
Stand-Alone	6	11	6	23
Grade 11	**18**	**17**	**8**	**43**
Cluster	4	10	2	16
Stand-Alone	14	7	6	27
Total	**35**	**58**	**30**	**123**

Note. [a]Other MOU state includes Hawaii.
[b]Count excludes 60 Oregon-owned legacy field-test items.

Table 19. Spring 2022 OSAS Science Assessment Item Pool by Science Discipline

Grade	Science Discipline	Item Type	ICCR Items	Oregon Items	MOU Items[a]	Total Item Pool[b]
Grade 5	Earth and Space Sciences	Cluster	12	7	45	64
		Stand-Alone	24	5	38	67
	Life Sciences	Cluster	12	11	38	61
		Stand-Alone	26	5	30	61
	Physical Sciences	Cluster	13	7	53	73
		Stand-Alone	29	6	38	73
Grade 8	Earth and Space Sciences	Cluster	7	5	28	40
		Stand-Alone	18	5	24	47
	Life Sciences	Cluster	7	9	45	61
		Stand-Alone	29	7	29	65
	Physical Sciences	Cluster	8	9	36	53
		Stand-Alone	17	6	35	58
Grade 11	Earth and Space Sciences	Cluster	11	8	11	30
		Stand-Alone	17	5	22	44
	Life Sciences	Cluster	19	8	34	61
		Stand-Alone	33	3	51	87
	Physical Sciences	Cluster	9	9	23	41
		Stand-Alone	20	7	39	66
Total			**311**	**122**	**619**	**1052**

Note. [a]Other MOU states include Connecticut, Hawaii, Idaho, Montana, MSSA (Rhode Island and Vermont), Utah, West Virginia, and Wyoming.
[b]Count excludes 60 Oregon-owned legacy items.

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

Grade	Science Discipline	Disciplinary Core Idea	ICCR Items	Oregon Items	MOU Items[a]	Total Item Pool[b]
Grade 5	Earth and Space Sciences	ESS1	10	3	24	37
		ESS2	11	8	41	60
		ESS3	15	1	18	34
	Life Sciences	LS1	15	6	27	48
		LS2	6	3	12	21
		LS3	3	3	12	18
		LS4	14	4	17	35
	Physical Sciences	PS1	14	4	23	41
		PS2	9	6	20	35
		PS3	15	2	33	50
		PS4	4	1	15	20
Grade 8	Earth and Space Sciences	ESS1	10	2	17	29
		ESS2	6	4	19	29
		ESS3	9	4	16	29
	Life Sciences	LS1	13	5	30	48
		LS2	12	4	20	36
		LS3	4	0	5	9
		LS4	7	7	19	33
	Physical Sciences	PS1	5	6	24	35
		PS2	3	4	17	24
		PS3	10	2	18	30
		PS4	7	3	12	22
Grade 11	Earth and Space Sciences	ESS1	10	4	9	23
		ESS2	8	5	12	25
		ESS3	10	4	12	26
	Life Sciences	LS1	17	5	21	43
		LS2	12	3	27	42
		LS3	8	2	12	22
		LS4	15	1	25	41

Grade	Science Discipline	Disciplinary Core Idea	ICCR Items	Oregon Items	MOU Items[a]	Total Item Pool[b]
	Physical Sciences	PS1	15	6	27	48
		PS2	5	3	15	23
		PS3	4	4	12	20
		PS4	5	3	8	16
Total			**311**	**122**	**619**	**1052**

Note. [a]Other MOU states include Connecticut, Hawaii, Idaho, Montana, MSSA (Rhode Island and Vermont), Utah, West Virginia, and Wyoming.
[b]Count excludes 60 Oregon-owned legacy items.

More information about *p*-values, biserial correlations, and IRT parameters can be found in Volume 1, Annual Technical Report. The details on calibration, equating, and scoring of the OSAS Science Assessment can also be found in Volume 1.

4.4 Remote Testing Forms

In the 2021–2022 school year, remote testing forms were constructed to assess science among students taking the test remotely. They were built as fixed forms to reduce the risk of the item content being compromised. To minimize the number of items on remote forms across states, the remote forms for the OSAS Science Assessment used the ICCR remote forms as a starting point. Therefore, the construction of the ICCR remote forms is first described in this section. Adaptations to the specific blueprint and psychometric characteristics of online adaptive forms for the OSAS Science Assessment are discussed subsequently.

The items from the ICCR pool were considered eligible for the remote testing forms using the criteria outlined in Table 21.

Table 21. Criteria for Selecting Items to Build Possible Remote Fixed Forms

Criteria	Reason for Criteria
Items that are present in all three ICCR states	To have a common form across ICCR states.
Items that have a Spanish translation	The remote fixed form would be administered in English and Spanish. Therefore, this criterion was established to avoid having to translate items.
Items that do not have a braille translation	Not all items in the bank are eligible for a braille translation.
Items that are aligned to a PE for which there are other items	To avoid having a shallow operational pool.

Items fulfilling all four criteria constituted the original item pool. The item pool was used to build all possible forms adhering to the blueprint. When a form could not be built with the initial item pool, items were iteratively added relaxing one criterion at a time. Subsequently, forms were selected based on how closely they matched the aggregated psychometric characteristics of simulated online test forms for each of the ICCR states (i.e., New Hampshire, North Dakota, and South Dakota), while maximizing the number of items taken from the original pool. The following psychometric characteristics were considered:

- Total number of assertions
- Average *b* (difficulty) value in the test
- Expected time (the 80th percentile was used as the expected testing time for each item)

Table 22 illustrates these results per grade and state using the simulation results.

Table 22. Mean Values Across Forms for Total Number of Assertions, Average b Value, and Expected Time

Grade	Statistic	Mean per State		
		South Dakota	*North Dakota*	*New Hampshire*
Elementary School	N. Assertions	62.5	67.1	64.2
	Average *b*	−0.204	−0.247	−0.211
	Expected Time	92.9	101.3	90.4
Middle School	N. Assertions	76.8	76.2	79.8
	Average *b*	0.082	0.089	0.134
	Expected Time	81.7	81.3	86.2
High School	N. Assertions	84.3	90.7	86.1
	Average *b*	0.861	0.681	0.995
	Expected Time	71.0	73.5	77.8

In addition to these form statistics, for each grade and state, the average test information function (TIF) was computed using the simulated test forms. The average TIF was computed as the average of the information functions of the simulated test forms evaluated over a grid of theta values. For each grade, this procedure yielded one average TIF per state.

To establish boundaries for the TIF for the selected form, the minimum and maximum (across states) of the average TIF at each theta value was considered. Figure 2 through Figure 5 illustrate the minimum and maximum boundaries of the TIF per grade.

Figure 2. Test Information Function, Selected Elementary School Form (All ICCR States) and Boundaries

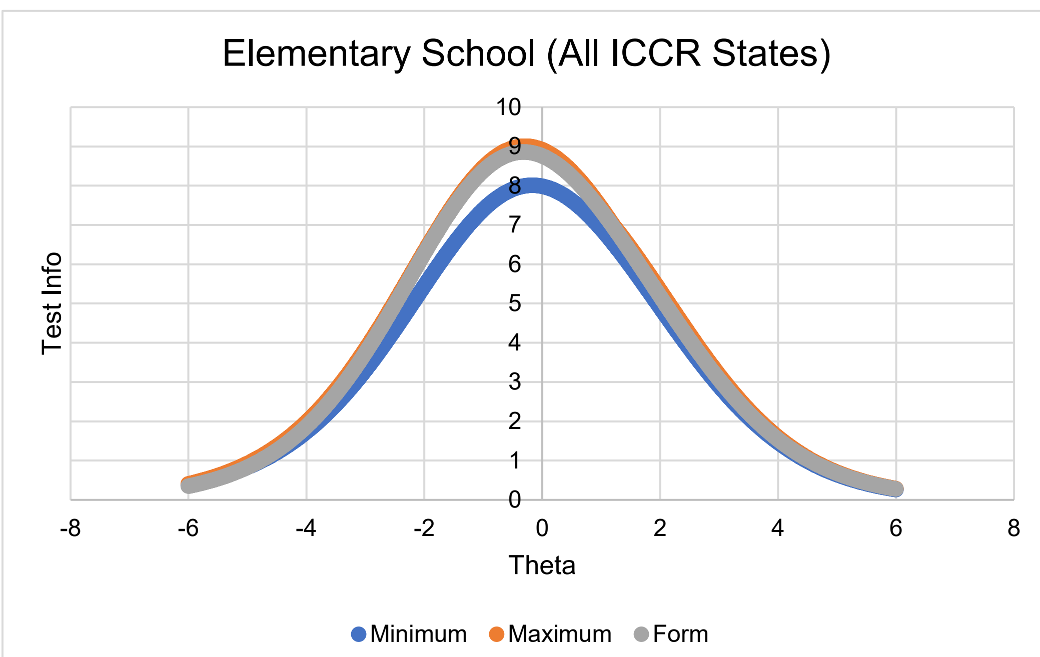


Figure 3. Test Information Function, Selected Middle School Form (All ICCR States) and Boundaries

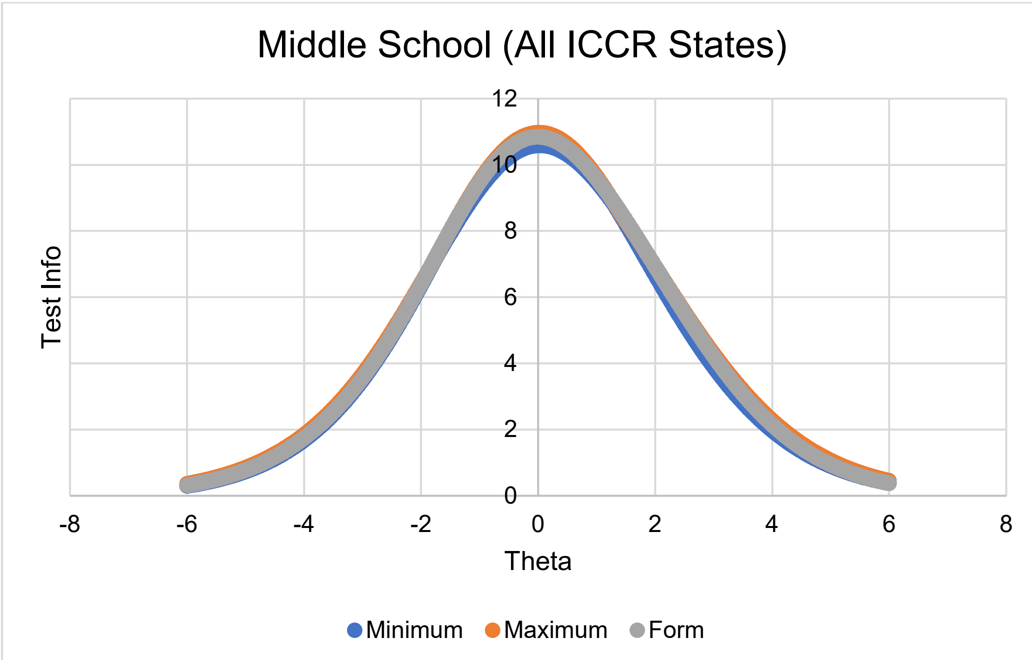


Figure 4. Test Information Function, Selected High School Form (New Hampshire and South Dakota) and Boundaries

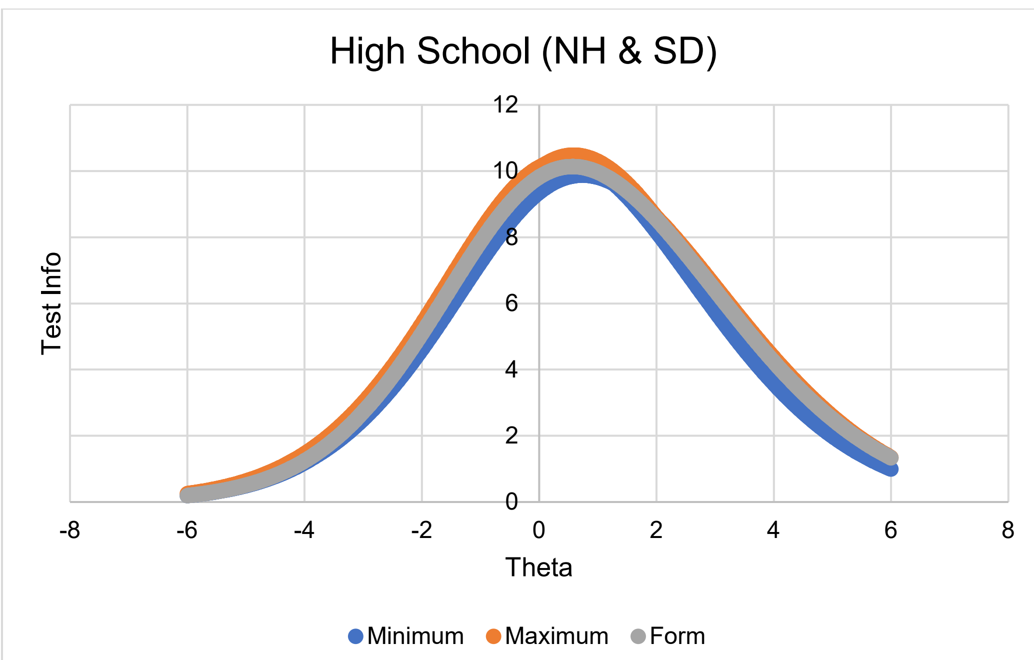
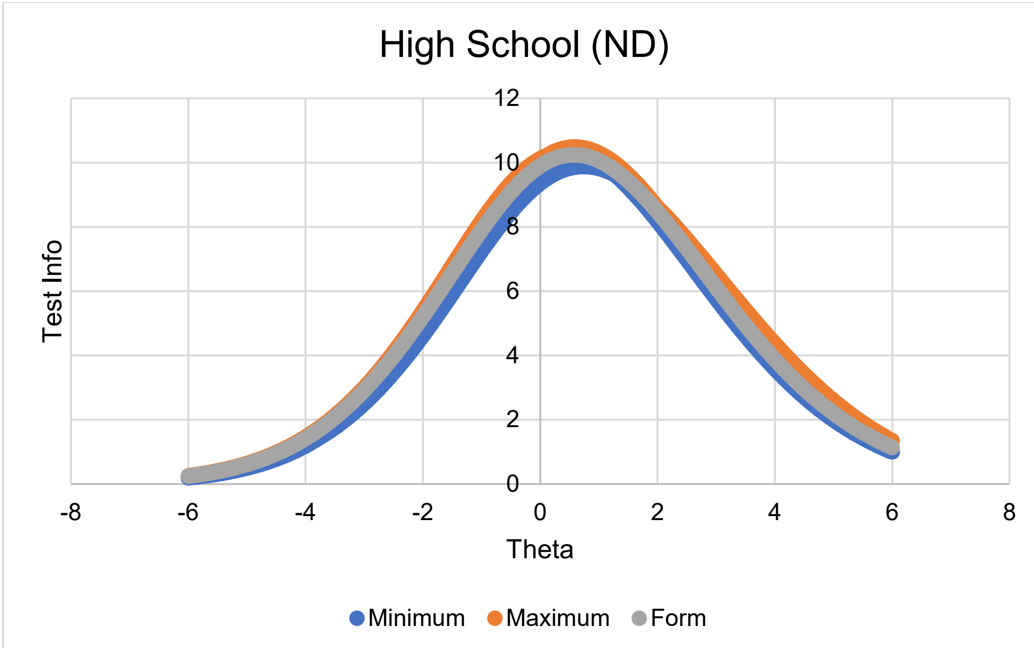


Figure 5. Test Information Function, Selected High School Form (North Dakota) and Boundaries



All possible forms adhering to the blueprint for the pool of eligible items were ranked according to how well their TIF mimicked the target information function (fell within the boundaries). Other psychometric criteria were later considered to select a form that most closely resembled the psychometric characteristics of the average simulated forms across states. When a state called for a different form due to differences in blueprint (e.g., North Dakota high school), a form was selected that most closely resembled the psychometric characteristics of that state.

The psychometric characteristics of the selected remote forms for the OSAS Science Assessment are summarized in Table 23, and the TIF is displayed in Figure 6 for grade 5, Figure 7 for grade 8, and Figure 8 for high school.

Table 23. Total Number of Assertions, Average b Value, and Expected Time for the Selected Remote Forms

Test	Number of Assertions	Average b	Expected Time
Grade 5	57	−0.167	89
Grade 8	65	0.120	76.7
Grade 11	97	0.826	86.9

Figure 6. Test Information Function, Grade 5 Remote Form

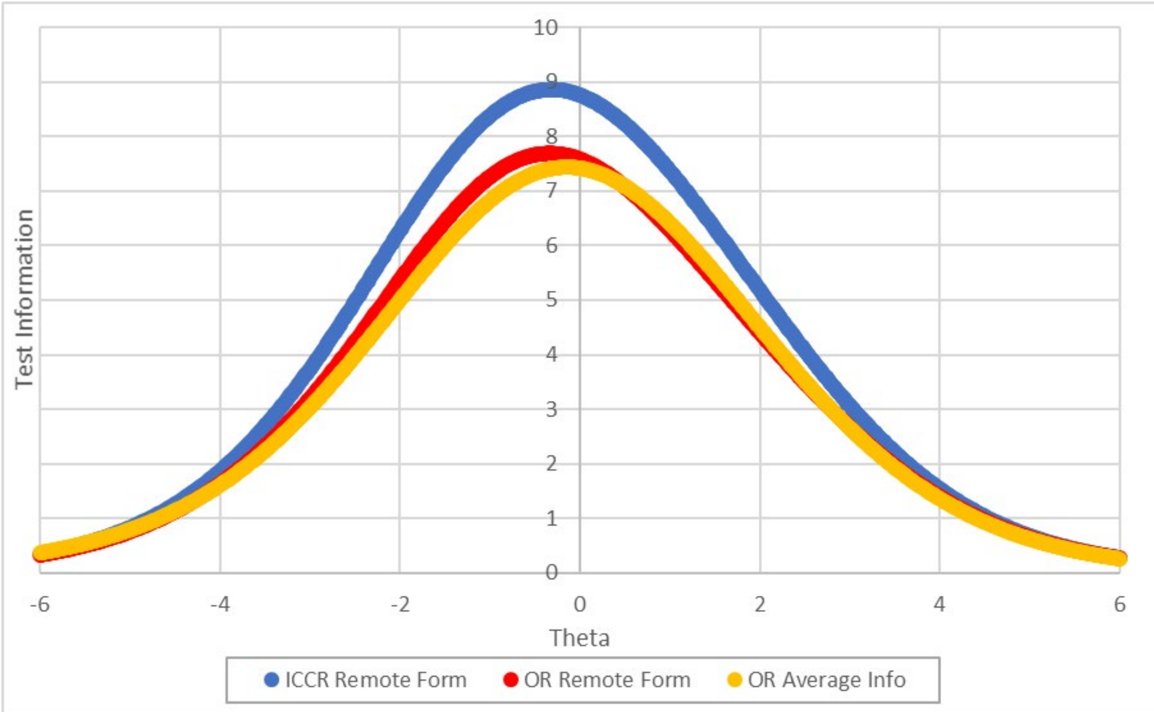


Figure 7. Test Information Function, Grade 8 Remote Form

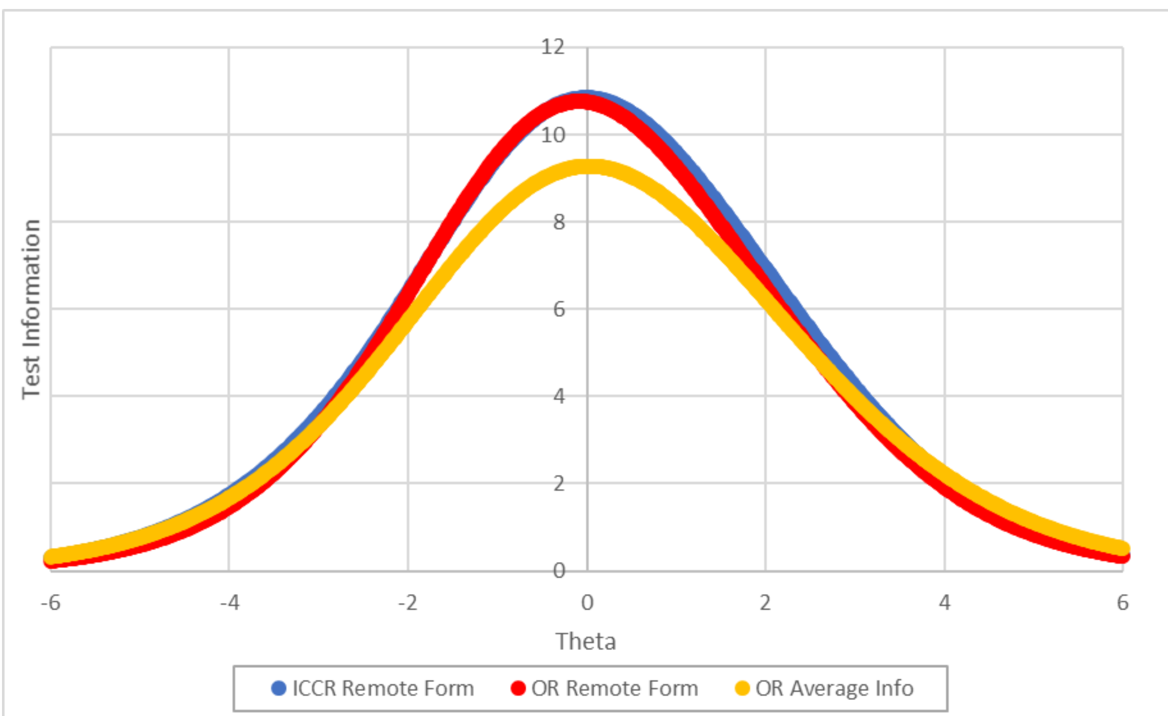
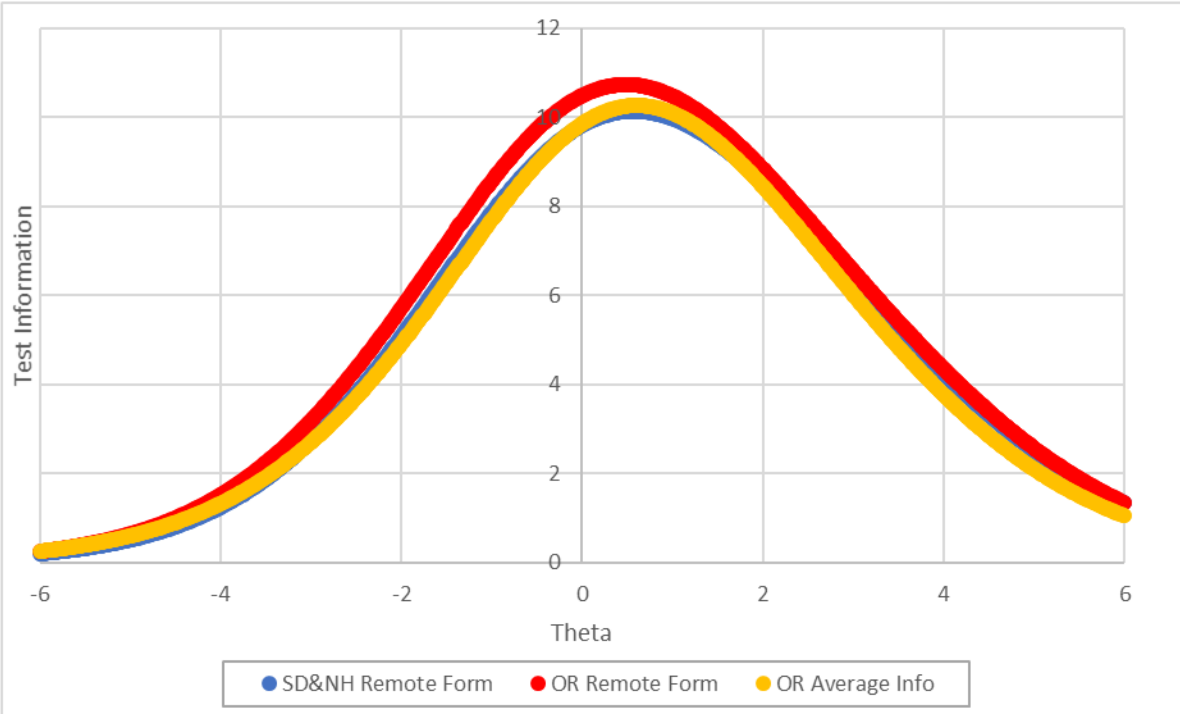


Figure 8. Test Information Function, Grade 11 Remote Form



Adaptations were made for the selected remote forms for the OSAS Science Assessment based on the ICCR remote forms. All adaptations were made to fulfill the blueprints for the OSAS Science Assessment and to closely resemble the psychometric characteristics of the average simulated forms for the OSAS Science Assessment. For grade 5, the blueprint that the ICCR form adhered to required six item clusters and 12 stand-alone items, whereas the Oregon blueprint required four item clusters and 14 stand-alone items. Two item clusters from the ESS discipline and LS discipline were therefore removed from the ICCR form, and two stand-alone items from the same disciplines were added. In addition, one item cluster in the ESS discipline was rejected in the Oregon bank and was replaced by another cluster in the same discipline. Similarly for grade 8, the blueprint that the ICCR form adhered to required six item clusters and 12 stand-alone items, whereas the Oregon blueprint required five item clusters and 13 stand-alone items. Therefore, one item cluster from the Physical Science discipline was swapped for a stand-alone item in the same discipline. For grade 11, the number of item clusters and stand-alone items in the ICCR form met the Oregon blueprint. However, one item cluster from the Life Science discipline, one item cluster from the Physical Science discipline, and one stand-alone item from the Earth and Space Science discipline in the ICCR form were rejected in the Oregon bank. Therefore, they were replaced by items from the same disciplines in the Oregon bank.

5. Simulation Summary Report

This section describes the results of simulated test administrations used to configure and evaluate the adequacy of the item selection algorithm used to administer the 2021–2022 Oregon Statewide

Assessment System (OSAS) Science Assessments for grades 5, 8, and 11. Simulations were carried out to configure the settings of the algorithm and to 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 J, Adaptive Algorithm Design, for more details of 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:** Determines that the tests have the correct number of test items overall and the appropriate proportion by content categories at each level of the content hierarchy, as specified in the test blueprints for every science grade.
- **Item Exposure Rate**: Evaluates the utility of 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 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 for the operational tests.

5.1 Factors Affecting Simulation Results

Several factors may influence simulation results for an adaptive test administration. These include the following:

- *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 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 online tests administered in English, which is the test taken by most students (98.5%). 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 set forth in the blueprints may yield greater exposure of items to students. This is reported by examining the percentage of test administrations in which an item appears. For instance, in a fixed form, 100% of the items appear on 100% of the test administrations because every test taker takes the same form. In an adaptive test or a linear-on-the-fly test (LOFT) 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 by dividing the total number of test administrations in which an item appears by the total number of tests administered. Then the distribution of the item exposure rate (r) in eight bins. The bins are r = 0% (unused), 0% < r ≤ 1%, 1% < r ≤ 5%, 5% < r ≤ 20%, 20% < r ≤ 40%, 40% < r ≤ 60%, 60% < r ≤ 80%, and 80% < r ≤ 100%. If an item bank is relatively large, most of the items is expected to appear in the bins of 0% < r ≤ 20%. This is an indication that most of the items appear on a very small percentage of the test forms.

Table 24 presents the percentage of items that falls into each exposure bin for all grades. Most of the items had item exposure rates less than 20%.

Table 24. Item Exposure Rates by Grade: Percentage of Items by Exposure Rate, Across All English Online Simulation Sessions

Grade	Total Items	(0,0]%	(0,1]%	(1,5]%	(5,20]%	(20,40]%	(40,60]%	(60,80]%	(80,100]%
5	358	0	21.79	50.00	23.74	3.63	0.84	0	0
8	285	0	31.58	27.72	33.68	6.32	0.70	0	0
11	286	0	28.32	34.97	29.72	5.59	1.40	0	0

5.3 Results of Simulated Test Administrations: Spanish

This section presents the simulation results for the Spanish tests. The Spanish item pool consists of a subset of ICCR items and some MOU items for which Spanish translations were available. Table 25 presents the number of items available for the Spanish tests in spring 2022.

Table 25. Spring 2022 Spanish Operational Item Pool

Grade	Item Type	Total Number of Items
5	Cluster	26
	Stand-Alone	35
8	Cluster	21
	Stand-Alone	32
11	Cluster	23
	Stand-Alone	28
Total		165

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 any content level in all three grades.

5.3.2 Item Exposure

Table 26 presents the percentage of items that falls into each exposure bin for all grades. Due to the limited size of the Spanish pool, most items had exposure rates less than 40% rather than less than 20%.

Table 26. Item Exposure Rates by Grade: Percentage of Items by Exposure Rate, Across All Spanish Simulation Sessions

Grade	Total Items	(0,0]%	(0,1]%	(1,5]%	(5,20]%	(20,40]%	(40,60]%	(60,80]%	(80,100]%
5	61	0	0	8.20	36.07	22.95	22.95	6.56	3.28
8	53	0	0	7.55	28.30	28.30	18.87	13.21	3.77
11	51	0	0	11.76	25.49	25.49	17.65	9.80	9.80

5.4 Results of Simulated Test Administrations: Braille

This section presents the simulation results for the Braille tests. The braille item pool consists of a subset of ICCR items and some MOU items for which braille versions were available. Table 27 presents the number of items available for the braille tests in spring 2022.

Table 27. Spring 2022 Braille Operational Item Pool

Grade	Item Type	Total Number of Items
5	Cluster	14

Grade	Item Type	Total Number of Items
	Stand-Alone	20
8	Cluster	22
	Stand-Alone	20
11	Cluster	13
	Stand-Alone	22
Total		111

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

5.4.1 Summary of Blueprint Match

The simulation results showed no blueprint violations at any content level in all three grades.

5.4.2 Item Exposure

Table 28 presents the percentage of items that falls into each exposure bin for all grades. Most items were administered in more than 20% of the test administrations. A few items had an exposure rate of 100% because of the limited braille item pool. Only those items were available to satisfy the blueprint constraints.

Table 28. Item Exposure Rates by Grade: Percentage of Items by Exposure Rate, Across All Spanish Simulation Sessions

Grade	Total Items	(0,0]%	(0,1]%	(1,5]%	(5,20]%	(20,40]%	(40,60]%	(60,80]%	(80,100]%
5	34	0	0	0	17.65	26.47	14.71	17.65	23.53
8	42	0	0	4.76	26.19	19.05	23.81	9.52	16.67
11	35	0	0	0	11.43	31.43	25.71	11.43	20.00

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

No blueprint violation was found at any content level in all three grades.

6.2 ITEM EXPOSURE

Table 29 presents the item exposure rates of the spring 2022 test administration. The exposure rates closely resembled the simulation results described in Section 5.2.2, Item Exposure, for the English test administrations. More items on the Spanish tests had high exposure rates as compared with items on the English tests because of a smaller item pool. Also, the operational exposure rates differed slightly from the simulation results because of small population sizes in all three grades. In spring 2022, 795 students took the Spanish test in grade 5, 645 students in grade 8, and 103 in high school. The item exposure rate for the Braille tests is not presented due to the small number of students taking the test. In spring 2022, only one student in grade 5 and one students in grade 8 took the Braille test. The item exposure rate for the Remote tests is not presented because they are fixed forms. Within a grade, every student received the same set of items. In spring 2022, 777 students took the Remote test in grade 5, 852 students in grade 8, and 720 in high school.

Table 29. Item Exposure Rates by Grade: Percentage of Items by Exposure Rate, Across All Spring 2022 Test Administrations

Grade	Total Items	(0,0]%	(0,1]%	(1,5]%	(5,20]%	(20,40]%	(40,60]%	(60,80]%	(80,100]%
English									
5	358	0	23.18	47.77	24.58	3.63	0.84	0	0
8	285	0	34.04	26.67	32.98	4.91	1.40	0	0
11	286	0	30.07	33.57	29.72	4.90	1.40	0.35	0
Spanish									
5	61	0	0	8.20	39.34	22.85	14.75	13.11	1.64
8	53	0	0	7.55	28.30	20.75	30.19	9.43	3.77
11	51	5.88	1.96	15.69	21.57	23.53	1.96	15.69	13.73

7. References

Council of Chief State School Officers (CCSSO). (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.

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