Connecticut Next Generation Science Standards Assessment

2020–2021

Volume 2
Test Development

CONNECTICUT STATE
DEPARTMENT OF EDUCATION

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

1. INTRODUCTION

Connecticut adopted the Next Generation Science Standards (NGSS) in 2015. The Connecticut State Department of Education (CSDE) 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 Connecticut NGSS Assessment was piloted in 2016–2017, field tested in 2017–2018, and administered operationally for the first time in 2018–2019. The Connecticut NGSS Assessment measures the science knowledge and skills of Connecticut students in grades 5, 8, and 11 as an online assessment, constructed with a linear-on-the-fly test (LOFT) design, making use of several technology-enhanced item types. 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, Annual Technical Report.

The interpretation, usage, and validity of test scores rely heavily upon the process of developing the test itself. This volume provides details on the test development process for the Connecticut NGSS Assessment that contributes to the validity of the test scores. Specifically, this volume provides evidence to support the following:

- The item specifications provided detailed guidance for item writers and reviewers to ensure that science items were aligned to the performance expectations (PEs) they were intended to measure (Appendix A, Item Writer Training Materials, and Appendix B, Item Review Checklist).
- The item development procedures employed for the Connecticut NGSS Assessment were consistent with industry standards.
- The development and maintenance of the Shared Science Assessment Item Bank, in which test items cover the range of measured PEs, grade-level difficulties, and levels of cognitive engagement through the use of both item clusters and stand-alone items.
- 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 (Appendix J, Adaptive Algorithm Design).

Note that for the science assessments, as outlined in Volume 1, Annual Technical Report, CAI works with a group of states that share common item development processes. In addition to developing items for each of those states, CAI develops and maintains the Independent College and Career Readiness (ICCR) item bank, which consists of items developed according to the same principles followed for the items owned by each of the states. Therefore, this volume of the annual technical report focuses on the general test development activities.

For the Connecticut NGSS Assessment, items are drawn from the Shared Science Assessment Item Bank that consists of ICCR items, items owned by Connecticut, and items owned by several other states that share a Memorandum of Understanding (MOU) to share content, leadership, and new ideas and methods. Specifically, all items developed under the MOU underwent the same development process. For the remainder of this volume, the term *item bank* will refer to all items developed under the MOU unless stated otherwise explicitly.

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 and four nationally recognized experts who helped author the NGSS. Two nationally recognized psychometricians also participated.

CAI staff and participating states collaborated to develop items and test specifications that would measure the three-dimensional science standards. The item specifications were generally accompanied by sample item clusters meeting 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. This platform ensures that every item flows through the correct sequence of reviews and captures every comment and change to the item.

CAI sought to ensure that the items measured the PEs in a fair and meaningful way by engaging educators and other stakeholders at each step of the process. Educators evaluated the alignment of

the items to the PEs and offered guidance and suggestions for improvement. They participated in the review of items for fairness and sensitivity. Following item field testing, educators engaged in rubric validation, a process that refines rule-based rubrics upon review of student responses.

Combined, these principles and the processes that support them 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 barriers to access. The details of these processes are described in this volume of the technical report.

1.3 ORGANIZATION OF THIS VOLUME

This volume is organized in 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 science item bank, the types of assessments the bank is designed to support, and methods for refreshing the bank
3. A description of the test construction process followed for the Connecticut NGSS 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

CAI developed the Shared Science Assessment Item Bank in collaboration with the states that were part of the Memorandum of Understanding (MOU) using a rigorous, structured process that engaged stakeholders at critical junctures. This process was managed by CAI's Item Tracking System (ITS), which is an auditable content-development tool that enforces rigorous workflow and captures each item change and comment. Reviewers, including internal CAI reviewers or 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 has a role in ensuring that the items can support the claims on which they will be based. Table 1 describes how each step contributes to these goals and describes each step in the process in more detail.

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

Developmental Steps	**Support Alignment to the Performance Expectations**	**Reduce Construct-Irrelevant Variance Through Universal Design**	**Expand 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 PEs and item specifications. Teaches item writers how to select 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 text-to-speech (TTS) for science, that reduce barriers.	Adds TTS, braille, American Sign Language (ASL), translations, 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 to identify 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.	Provides final, focused review on items flagged for differential item functioning (DIF).	

2.2 ITEM SPECIFICATIONS

CAI is working with a group of states, 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. Participating states included Connecticut, Hawaii, Idaho, Montana, Oregon, Rhode Island, Utah, Vermont, West Virginia, and Wyoming. New Hampshire, North Dakota and South Dakota participated in some activities. This collaboration has yielded item specifications for PEs, sample item clusters for some specifications, and hundreds of science item clusters and stand-alone items in various stages of development. Under this collaboration, utilizing guidelines for item specifications proposed by WestEd in collaboration with the Council of Chief State School Officers (CSSO), state members, and content experts (CCSSO, 2015), states 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 that together form coherent item clusters and stand-alone items. Table 2 provides a sample of the item specifications developed by content experts for a middle school Life Sciences PE. Item specifications in science include the following:

- **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. 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 placed on developing evidence that living things are made of cells, distinguishing between living and non-living things, and understanding that living things may be made of one cell or many varying cells.
Content Limits
<u>Students do not need to know the following</u>:
The structures or functions of specific organelles or different proteins
Systems of specialized cells
The mechanisms by which cells are alive
Specifics of DNA and proteins or of cell growth and division
Endosymbiotic theory
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, protists, archaea, nucleoid, plasmid, diatoms, cyanobacteria</td></tr>
<tr><td colspan="4">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.</td></tr>
</table>

	• Different parts of a frog's body (e.g., muscles, skin, tongue) are observed under a microscope, and are seen to be composed of cells. • One-celled organisms (e.g., bacteria, protists) perform the eight necessary functions of life, but nothing smaller has been seen to do this. • Swabs from the human cheek are observed under a microscope. Small cells can be seen.
This Performance Expectation and Associated Evidence Statements Support the Following Task Demands	
Task Demands	
1. Identify from a list the materials/tools, including distractors, needed for an investigation to find the smallest unit of life (cell).	
2. Identify the outcome data that should be collected in an investigation of the smallest unit of living things.	
3. Evaluate the sufficiency and limitations of data collected to explain that the smallest unit of living things is the cell.	
4. Make and/or record observations about whether the sample contains cells.[b]	
5. Interpret and/or communicate data from the investigation to determine if a specimen is alive.	
6. Construct a statement to describe the overall trend suggested by the observed data.	

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

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

2.3 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, Item Writer Training Materials. 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 uses 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 before client review and provide training and feedback for all content-area team members.

ICCR and MOU science items undergo a rigorous, multiple-level internal review process before they are sent to external review. Staff members are trained to review items for both content and accessibility throughout the process. A sample item review checklist that our test developers 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

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

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

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

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

2.4.2 Scoring Entry and Review

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

- The scoring works 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 way in which the scoring is set up is unambiguous and matches the questions asked (i.e., if we ask students to round a number to a certain decimal place, we score accordingly).

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

2.4.3 Content Review One

Content Review One is conducted by a senior content specialist who was not part of the Preliminary Review. This reviewer carefully examines each item based on the same criteria identified for Preliminary Review. 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 his or her expertise in test development.

2.4.4 Edit Review

During Edit Review, editors have four primary tasks:

1. Editors perform basic line editing for correct spelling, punctuation, grammar, and mathematical and scientific notation, ensuring consistency of style across the items.

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

2.4.5 Senior Review

By the time a science item arrives at Senior Review, both content reviewers and editors have thoroughly vetted it. Senior reviewers (in particular, senior content specialists) look at the item's entire review history, ensuring that all the issues identified in that item have been adequately addressed. Senior reviewers verify the overall content of each item, confirming its accuracy, alignment to the 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 undergo an exhaustive external review process. Items in the Shared Science Assessment Item Bank were reviewed by content experts in one or several states and reviewed and approved by multiple stakeholder committees that evaluated them for both content and bias/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 before 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. A CAI science content expert reviews all client-requested edits considering the science item specifications, other clients' requests, and existing items in the bank to determine whether the requested edits will be made. At this stage, clients 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. The states provide feedback on the ICCR items, and 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 either be set for field testing in a future administration or become a part of the locked operational pool.

2.5.2 Content Advisory Committee Reviews

During the Content Advisory Committee (CAC) reviews, items are reviewed for content validity, grade-level appropriateness, and alignment to the PE. CAC members are typically grade-level and subject-matter experts. During this review, educators also ensure that the scoring assertions clearly identify what is being scored as correct and give credit where they should (refer to Section 2.7.1, Rubric Validation). Before the CAC review begins, CAI provides a presentation on the three-dimensional science standards, the item development process, the CAI systems that will be used in the review, and how to review the items for content. Appendix K, Content Advisory Committee Review Training Slides, provides the slides used during the CAC review training.

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. In most cases, items are seen by multiple state committees before their field test or operational use.

In 2021, MOU states were all involved in a single CAC process where participants from multiple states reviewed items. The items were edited and then returned to the owning state for final approval.

A summary of the committee meetings is presented in Table 3, with additional details about the participants in Appendix C, Content Advisory Committee Participant Details.

Table 3. Summary of Content Advisory Committee Meetings

State/Item Bank	Meeting	Number of Committee Members	Number of Items Reviewed
Connecticut	February 2017	41	45
	May 2017	42	40
	October 2017	41	75
	November 2017	35	41
	January 2018	33	42
	October 2018	45	84
	November 2018	49	235
	December 2018	32	56
	January 2019	44	65
	September 2019	50	60
	July 2021	[c]	24

State/Item Bank	Meeting	Number of Committee Members	Number of Items Reviewed
Hawaii	July 2017	22	25
	September 2017	20	65
	October 2018	29	85
	February 2019	21	44
ICCR	March 2018	26	152
	July 2021	[c]	164
Idaho	December 2018	21	111
Montana	July 2021	[c]	41
MSSA[a]	January 2018	42	73
	March 2018	28	100
	January 2019	21	116
	July 2021	[c]	30
Oregon	August 2017	10	110
	August 2018	20	257
	December 2018	16	62
	July 2021	[c]	22
Utah	July 2017	23	55
	December 2017	36	48
	July 2021	[c]	65
West Virginia	January 2017	28[b]	39
	October 2018	10	191
	July 2019	12	50
	July 2021	[c]	12
Wyoming	December 2017	17	51
	October 2018	14	37
	July 2021	[c]	32

Note. [a]MSSA = Rhode Island and Vermont's Multi-State Science Assessment.
[b]Number of Committee Members includes total committee members for English language arts (ELA), mathematics, and science. The number for science only committee members is not available.
[c]Multi-State review occurred over two weeks, with participants from multiple states involved. Items were reviewed by at least four participants.

2.5.3 Language Accessibility, Bias, and Sensitivity Committee Reviews

During the bias and sensitivity reviews, stakeholders review items to check for issues that might unfairly impact students based on their background. For example, some states include representatives from student populations such as Special Education, low vision, and the hearing impaired. Further, diverse members of this committee represent students of various

ethnic and economic backgrounds to ensure that all items are free of bias and sensitivity concerns. Before the bias and sensitivity review begins, CAI provides a presentation on the three-dimensional science standards, the item development process, the CAI systems that will be used in the review, and how to review the items for fairness. Appendix L, Fairness Committee Review Training Slides, provides the slides used during the bias and sensitivity review training.

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

In 2021, MOU states were all involved in a single review process where 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 is presented in Table 4, with additional details about the participants in Appendix D, Fairness Committee Participant Details.

Table 4. Summary of Fairness Committee Meetings

State/Item Bank	Meeting	Number of Committee Members	Number of Items Reviewed	Number of Items Rejected
Connecticut	February 2017	6	45	1
	December 2017	9	75	N/A
	December 2017	10	41	N/A
	February 2018	3	42	N/A
	November 2018	11	319	38
	December 2018	10	56	N/A
	January 2019	9	65	N/A
	September 2019	9	48	N/A[a]
Hawaii	July 2017	22	25	2
	September 2017	20	65	13
	October 2018	29	85	6
	February 2019	21	44	0
ICCR	March 2018	13	152	N/A
	July 2021	[d]	124	5
Idaho	December 2018	15	111	1
Montana	July 2021	[d]	48	0
MSSA[b]	January 2018	21	73	14
	March 2018	11	100	24
	January 2019	14	116	18
	July 2021	N/A	31	0

State/Item Bank	Meeting	Number of Committee Members	Number of Items Reviewed	Number of Items Rejected
Oregon	August 2017	5	110	5
	August 2018	9	256	56
	December 2018	11	62	13
US Virgin Islands	October 2021	[e]	[e]	[e]
Utah	August 2017	6	44	2
	December 2017	6	48	1
	July/August 2021	[d]	56	2
West Virginia	January 2017	28[c]	34	N/A
	January 2019	10	191	N/A
	July 2021	[d]	12	1
Wyoming	December 2017	5	51	3
	October 2018	5	37	N/A
	July 2021	[d]	41	0

Note. [a]Number of rejected items has not been finalized through client resolution at the time of writing this report.
[b]MSSA = Rhode Island and Vermont's Multi-State Science Assessment.
[c]Number of Committee Members includes total committee members for ELA, mathematics, and science. The number for science only committee members is not available.
[d]Multi-State review occurred over two weeks, with participants from multiple states involved. Items were reviewed by at least four participants.
[e]U.S. Virgin Islands reviews were a review of previously accepted ICCR items by department staff.

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 for translations or TTS, follow similar processes. One trained expert enters the markup, 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 science field-test items was administered in the following nine states in spring 2018: Connecticut, Hawaii, New Hampshire, Oregon, Rhode Island, Utah, Vermont, West Virginia, and Wyoming. For 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 no scores were reported. In New Hampshire, Utah, Vermont, and West Virginia, an operational field test was administered.

In 2019, a second pool of field-test items was administered in the following nine states: Connecticut, Hawaii, Idaho, New Hampshire, Oregon, Rhode Island, Vermont, West Virginia, and Wyoming. For 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. Unscored field-test items were added as a separate segment to the operational (scored) legacy science test for Wyoming. 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.

CAI's field-test process is described in detail in Volume 1, Section 3.2, Annual Technical Report.

2.7 POST-FIELD-TEST REVIEW

Following the field test, items were subject to a substantial validation process. This included rubric validation and data review. These processes are described in Section 2.7.1, Rubric Validation, and Section 2.7.2, Data Review.

2.7.1 Rubric Validation

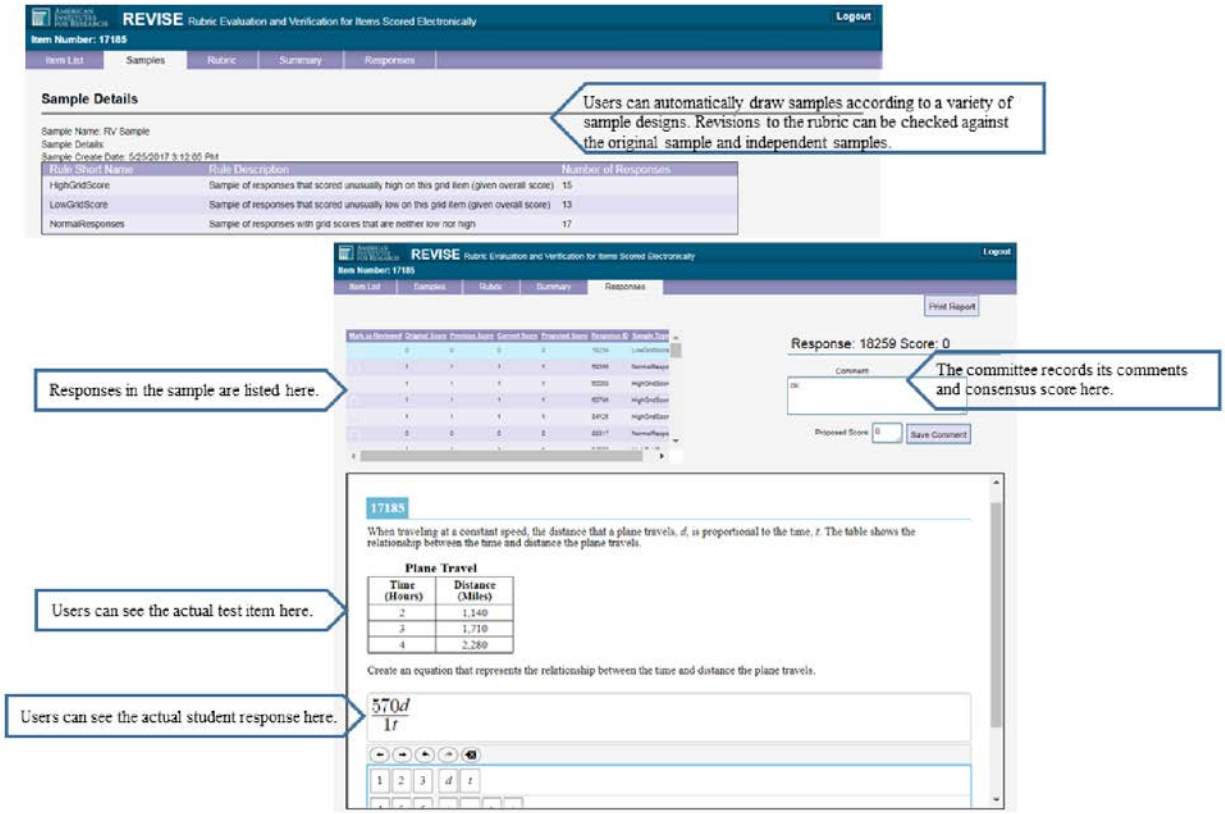
The validation process for the field-test items begins with rubric validation to verify and make any necessary revisions to the scoring rubrics. The rubric validation process occurs in two phases. During the first phase, CAI content experts work with the analysis team to prepare for the rubric validation meetings. 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: (1) low-scored responses from otherwise high-scoring students, and (2) high-scored responses from otherwise low-scoring students. This process allows CAI to identify any potential scoring concerns before the rubric validation meeting, such as unanticipated (but accurate) responses, equivalent responses that were not originally considered, and responses receiving credit but should not (based on the content and the item rubric). At this point, the rubrics may be adjusted, and responses rescored.

The second phase of rubric validation involves committees of educators in each state. The committees review the response samples generated by CAI to make recommendations to change or to confirm the rubrics of each item. The committee recommendations are then discussed with

the state of ownership to resolve any inconsistencies. The rubric is 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, and any items rejected as part of the process are rejected in ITS. ITS archives critical information regarding the scoring certification completed during the rubric validation process. This includes any rubric changes made during the scoring decision meetings and the sign-off completed by the senior content expert once the rubric has been changed, rescoring the entire sample, and the verification that the final rubric functioned as intended.

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

2.7.2 Data Review

Following rubric validation, all items are rescored and classical item statistics are computed for the scoring assertions, including item difficulty and item discrimination statistics, testing time, and differential item functioning (DIF) statistics. The states established standards for the statistics, and any items violating these standards are flagged for a second educator review. Even though the scoring assertions were the basic units of analysis 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, Hawaii, Idaho, Montana, New Hampshire, North Dakota, 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, Annual Technical Report, describes the statistical flags that designate items for data review. The flags are designed to highlight potential content weaknesses, miskeys, or possible bias issues. Committee members are taught to interpret these flags and are given guidelines for examining the items for content or fairness issues.

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. The 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 the data review committee and the data review process, and the meaning of the various flags. 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 PEs, the content and rubric of the item, and various item statistics). A sample of the training materials used for these data review meetings is presented in Appendix 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 (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 data review committee meetings. Details, including the composition of each committee, are presented 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[d]	25	**Total**	**51**	**12**
			Cluster	8	2
			Stand-Alone	43	10
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[d]	25	**Total**	**26**	**8**
			Cluster	6	0
			Stand-Alone	20	8
ICCR	July 2018	18	**Total**	**84**	**8**
			Cluster	33	2
			Stand-Alone	51	6
	August 2019	N/A[a]	**Total**	**43**	**3**
			Cluster	0	1
			Stand-Alone	43	2
	August 2021[d]	25	**Total**	**75**	**6**
			Cluster	11	2
			Stand-Alone	64	4
Idaho	August 2019	10	**Total**	**12**	**6**
			Cluster	4	3
			Stand-Alone	8	3
	August 2021[d]	25	**Total**	**60**	**5**
			Cluster	26	1
			Stand-Alone	34	4
Montana	September 2021	4	**Total**	**17**	**4**
			Cluster	3	2
			Stand-Alone	14	2
MSSA[b]	August 2018	2[c]	**Total**	**9**	**6**
			Cluster	2	0
			Stand-Alone	7	6

Owner	Meeting	Number of Committee Members	Item Type	Number of Items Reviewed	Number of Items Rejected
	August 2019	2[c]	**Total**	**14**	**4**
			Cluster	2	1
			Stand-Alone	12	3
	August 2021[d]	25	**Total**	**18**	**9**
			Cluster	4	4
			Stand-Alone	14	5
Oregon	September 2018	11	**Total**	**44**	**6**
			Cluster	28	5
			Stand-Alone	16	1
	August 2019	4	**Total**	**8**	**7**
			Cluster	1	1
			Stand-Alone	7	6
South Dakota[e]	September 2021	N/A[f]	**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
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[d]	25	**Total**	**7**	**3**
			Cluster	1	1
			Stand-Alone	6	2
Wyoming	October 2018	19	**Total**	**16**	**6**
			Cluster	6	1
			Stand-Alone	10	5
	August 2019	10	**Total**	**16**	**5**
			Cluster	4	3
			Stand-Alone	12	2
	August 2021[d]	25	**Total**	**16**	**4**
			Cluster	3	1
			Stand-Alone	13	3

Note. [a]In summer 2019, ICCR field-test items were taken to Connecticut, Hawaii, and Idaho for committee review.
[b]MSSA = Rhode Island and Vermont's Multi-State Science Assessment.
[c]Conducted by Rhode Island Department of Education and Vermont Agency of Education science content experts.
[d]Cross-state committee item data review.
[e]Legacy field-test items only.
[f]State Department of Education review only.

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 increasing the testing time or testing burden.

CAI has built the Shared Science Assessment Item Bank in partnership with multiple states. 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.

Each state 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 continue to field test new items. Participating states 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 he or she can construct a graph showing the relationship between two variables, but he or she 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 and describes the science interaction types. Examples of various interaction types can be found in Appendix G, Example Item Interactions.

Table 6. Science Interaction Types and Descriptions

Interaction Type	Associated Sub-Types	Description
Choice	Multiple-Choice	Traditional multiple-choice interaction allows students to select a single option from a list of possible answer options.
	Multi-Select	Traditional multi-select interaction (checkboxes) allows students to select one or more options from a list of possible answer choices.
Text Entry	Simple Text Entry	Students type a response in a text box.
	Embedded Text Entry	Students type their 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.
	Table Input	Interaction solicits students to complete tabular data.
Edit Task	Edit Task	Students click a word and replace it with another word that they type to revise a sentence.
	Edit Task with Choice	Students click a word or phrase and select the replacement from several options.
	Edit Task Inline Choice	Drop-down menus are placed through the text, and students select an option to complete the text.
Hot-Text	Selectable	Selectable hot-text interactions require students to select one or more text elements in the response area.
	Re-orderable	Re-orderable hot-text interactions require students to click and drag hot-text elements into a different order.
	Drag-from-Palette	Drag-from-Palette hot-text interactions require students to drag elements from a palette into the available blank table cells or "gaps" (text boxes) in the response area.
	Custom	Custom hot-text interactions combine the functionality of the other hot-text interaction sub-types. Students responding to a custom hot-text interaction may need to select text elements, rearrange text elements, and/or drag text elements from a palette to blank table cells or drop targets in the response area.
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

Interaction Type	Associated Sub-Types	Description
		mathematics 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. The student may be required to draw a line or shape, plot a point, or create a graph. The student may also drag and drop or click on selectable hot-spots.
	Hot-Spot	Hot-spot interaction sub-types facilitate grid interactions with specific hot-spot functionality. These interactions require students to select hot-spot regions in the grid area.
	Graphic Gap Match	Graphic gap match interactions facilitate grid interactions with specific drag-and-drop functionality. These interactions require students to drag image objects from a palette to specified regions (gaps) in the grid area.
Simulation	N/A	Simulation interactions allow students to investigate a phenomenon by selecting variables to get output data. Some simulations are accompanied by animations.

Table 7 through Table 11 present the number of items in the Shared Science Assessment Item Bank available for use in the spring 2021 statewide assessments. Appendix H, Shared Science Assessment Item Bank, provides the items available within the bank by grade band, PE, and origin.

Table 7. Spring 2021 Shared Science Assessment Operational and Field-Test Item Bank

Grade Band and Item Type	ICCR Items	Connecticut Items	MOU Items[a]	Total Bank Items
Elementary School	**130**	**66**	**243**	**439**
Cluster	41	30	148	219
Stand-Alone	89	36	95	220
Middle School	**115**	**62**	**268**	**445**
Cluster	32	32	158	222
Stand-Alone	83	30	110	223
High School	**122**	**83**	**165**	**370**
Cluster	43	30	72	145
Stand-Alone	79	53	93	225
Total	**367**	**211**	**676**	**1254**

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

Table 8. Spring 2021 Shared Science Assessment Operational Item Bank

Grade Band and Item Type	ICCR Operational Items	Connecticut Operational Items	MOU Operational Items[a]	Total Bank Operational Items
Elementary School	**79**	**38**	**108**	**225**
Cluster	32	19	62	113
Stand-Alone	47	19	46	112
Middle School	**68**	**38**	**180**	**286**
Cluster	24	18	120	162
Stand-Alone	44	20	60	124
High School	**79**	**34**	**85**	**198**
Cluster	28	19	41	88
Stand-Alone	51	15	44	110
Total	**226**	**110**	**373**	**709**

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

Table 9. Spring 2021 Shared Science Assessment Field-Test Item Bank

Grade Band and Item Type	ICCR Field-Test Items	Connecticut Field-Test Items	MOU Field-Test Items[a]	Total Bank Field-Test Items
Elementary School	**51**	**28**	**135**	**214**
Cluster	9	11	86	106
Stand-Alone	42	17	49	108
Middle School	**47**	**24**	**88**	**159**
Cluster	8	14	38	60
Stand-Alone	39	10	50	99
High School	**43**	**49**	**80**	**172**
Cluster	15	11	31	57
Stand-Alone	28	38	49	115
Total	**141**	**101**	**303**	**545**

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

Table 10. Spring 2021 Shared Science Assessment Operational and Field-Test Item Bank by Science Discipline

Grade Band	Science Discipline	Item Type	ICCR Items	Connecticut Items	MOU Items[a]	Total Bank Items[b]
		Cluster	14	10	43	67

Grade Band	Science Discipline	Item Type	ICCR Items	Connecticut Items	MOU Items[a]	Total Bank Items[b]
Elementary School	Earth and Space Sciences	Stand-Alone	28	11	37	76
	Life Sciences	Cluster	14	9	46	69
		Stand-Alone	30	10	26	66
	Physical Sciences	Cluster	13	11	59	83
		Stand-Alone	31	15	32	78
Middle School	Earth and Space Sciences	Cluster	11	5	45	61
		Stand-Alone	23	8	31	62
	Life Sciences	Cluster	10	17	55	82
		Stand-Alone	38	12	38	88
	Physical Sciences	Cluster	11	10	52	73
		Stand-Alone	22	10	40	72
High School	Earth and Space Sciences	Cluster	9	8	13	30
		Stand-Alone	12	16	17	45
	Life Sciences	Cluster	20	9	38	67
		Stand-Alone	49	11	47	107
	Physical Sciences	Cluster	14	13	20	47
		Stand-Alone	18	26	29	73
Total			**367**	**211**	**668**	**1246**

Note. [a]Other MOU states include Hawaii, Idaho, Montana, MSSA (Rhode Island and Vermont), Oregon, Utah, West Virginia, and Wyoming. [b]Count excludes eight MOU items that do not align to the NGSS.

Table 11. Spring 2021 Shared Science Assessment Operational and Field-Test Item Bank by Disciplinary Core Idea

Grade Band	Science Discipline	Disciplinary Core Idea	ICCR Items	Connecticut Items	MOU Items[a]	Total Bank Items[b]
Elementary School	Earth and Space Sciences	ESS1	12	6	23	41
		ESS2	13	9	36	58
		ESS3	17	6	21	44
	Life Sciences	LS1	17	7	34	58
		LS2	5	3	13	21
		LS3	4	3	9	16
		LS4	18	6	16	40
	Physical Sciences	PS1	12	6	29	47
		PS2	11	8	17	36
		PS3	17	7	31	55
		PS4	4	5	14	23
Middle School	Earth and Space Sciences	ESS1	15	4	20	39
		ESS2	9	5	28	42
		ESS3	10	4	28	42
	Life Sciences	LS1	10	10	35	55
		LS2	20	6	29	55
		LS3	4	4	10	18
		LS4	14	9	19	42
	Physical Sciences	PS1	9	4	31	44
		PS2	3	4	26	33
		PS3	14	7	20	41
		PS4	7	5	15	27
High School	Earth and Space Sciences	ESS1	7	6	12	25
		ESS2	7	8	11	26
		ESS3	7	10	7	24
	Life Sciences	LS1	18	5	28	51
		LS2	20	7	27	54
		LS3	10	4	10	24
		LS4	21	4	20	45
	Physical Sciences	PS1	14	15	20	49

Grade Band	Science Discipline	Disciplinary Core Idea	ICCR Items	Connecticut Items	MOU Items[a]	Total Bank Items[b]
		PS2	8	9	11	28
		PS3	6	10	11	27
		PS4	4	5	7	16
Total			**367**	**211**	**668**	**1246**

Note. [a]Other MOU states include Hawaii, Idaho, Montana, MSSA (Rhode Island and Vermont), Oregon, Utah, West Virginia, and Wyoming. [b]Count excludes eight 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 a linear-on-the-fly test (LOFT) 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 in two years, once replacements have been introduced into the operational bank.

4. CONNECTICUT NGSS ASSESSMENT TEST CONSTRUCTION

4.1 TEST DESIGN

The Connecticut NGSS Assessment was administered online to students in grades 5, 8, and 11 using a linear-on-the-fly test (LOFT) design. Contrary to a fixed form, every student potentially sees a different set of items. Items are selected by an item selection algorithm to ensure that the blueprint is met whenever possible. The algorithm used is the same algorithm that CAI uses for the administration of adaptive tests. The adaptive item-selection algorithm selects items based on their content value 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. By assigning a weight of zero to the information value of an item with respect to the underlying proficiency, the items are selected solely based on their contributions to meeting the blueprint. Details for CAI's adaptive testing algorithm are described in Appendix J, Adaptive Algorithm Design.

A non-segmented test design was used for the Connecticut NGSS Assessment. Students received items from different disciplines in a random order. The choice of design was partially motivated by a possible move to a fully adaptive test in future administrations. 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
- Science disciplines to be covered and the acceptable number of items across performance expectations (PEs) within each science discipline and DCI

The blueprint for science is presented 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	**2**	**2**	**4**	**4**	**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	**2**	**2**	**4**	**4**	**6**	**6**
DCI—From Molecules to Organisms: Structure and Function	**0**	**1**	**0**	**2**	**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**	**2**	**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**	**2**	**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**	**2**	**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	**2**	**2**	**4**	**4**	**6**	**6**
DCI—Earth's Systems	**0**	**1**	**0**	**2**	**0**	**3**
3-ESS2-1: Weather and Climate	0	1	0	1	0	1

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**	**1**	**0**	**2**
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**	**1**	**0**	**2**
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	**6**	**6**	**12**	**12**	**18**	**18**

Note. *These PEs have an engineering component.

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	**2**	**2**	**4**	**4**	**6**	**6**
DCI—Matter and Its Interactions	**0**	**1**	**0**	**2**	**0**	**3**
MS-PS1-1: Structure and Properties of Matter	0	1	0	1	0	1
MS-PS1-2: 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	**2**	**2**	**4**	**4**	**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	**2**	**2**	**4**	**4**	**6**	**6**
DCI—Earth's Place in the Universe	**0**	**1**	**0**	**1**	**0**	**2**
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**	**1**	**0**	**2**
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	**6**	**6**	**12**	**12**	**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**	**0**	**0**	**2**	**0**	**2**
HS-PS2-1: Forces and Motion	0	0	0	1	0	1
HS-PS2-2: Forces and Motion	0	0	0	1	0	1
HS-PS2-3: Forces and Motion*	0	0	0	1	0	1
HS-PS2-4: Types of Interactions	0	0	0	1	0	1
HS-PS2-5: Types of Interactions	0	0	0	1	0	1
HS-PS2-6: Chemical Reactions*	0	0	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**	**0**	**0**	**2**	**0**	**2**
HS-PS4-1: Wave Properties	0	0	0	1	0	1
HS-PS4-2: Wave Properties	0	0	0	1	0	1
HS-PS4-3: Wave Properties/Electromagnetic Radiation	0	0	0	1	0	1
HS-PS4-4: Electromagnetic Radiation	0	0	0	1	0	1
HS-PS4-5: Electromagnetic Radiation*	0	0	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**	**1**	**0**	**2**
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**	**0**	**0**	**1**	**0**	**1**
HS-ESS1-1: The Universe and Its Stars	0	0	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	0	0	1	0	1
HS-ESS1-3: The Universe and Its Stars	0	0	0	1	0	1
HS-ESS1-4: Earth and the Solar System	0	0	0	1	0	1
HS-ESS1-5: The History of Planet Earth	0	0	0	1	0	1
HS-ESS1-6: The History of Planet Earth	0	0	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.

The main characteristics of the blueprint were that any 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 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). Some specific constraints for the Connecticut blueprint included that for grades 5 and 8; the Earth Systems DCI would have two stand-alone items (rather than one for other DCIs in the Earth and Space Sciences Discipline) because it had the most PEs and was rated the highest in the district responses. In addition, three DCIs in grade 11 (Motion and Stability, Waves, and Earth's Place in the Universe) were constrained to not receive an item cluster due to low content priority ratings from districts.

While tests are not timed, the Connecticut State Department of Education (CSDE) published estimated testing times for the Connecticut NGSS Assessment. The 85th percentile of the testing times is presented in Table 15.

Table 15. Connecticut NGSS Assessment 85th Percentile Testing Times by Grade

Subject	Grade	85th Percentile Testing
Science	5	112.82
	8	106.18
	11	88.47

4.3 ONLINE TEST CONSTRUCTION

During fall 2020, CAI psychometricians and content experts worked with CSDE content specialists and leadership to build item pools for the spring 2021 administration. The Connecticut NGSS Assessment test construction used a structured test construction plan, explicit blueprints, and active collaborative participation from all parties.

The 2021 Connecticut NGSS Assessment test item pools were built by CAI test developers to match items exactly to the detailed test blueprints. Operational items were selected from nine item banks (ICCR, Connecticut, Hawaii, Idaho, MSSA [Rhode Island and Vermont], Oregon, Utah, West Virginia, and Wyoming) to fulfill the blueprint for each grade. Table 16 through Table 20 summarize the 2021 Connecticut NGSS Assessment item pool. Appendix I, Connecticut NGSS Assessment Item Pool, provides the 2021 item pool by grade band, PE, and origin.

Table 16. Spring 2021 Connecticut NGSS Assessment Operational and Field-Test Item Pool

Grade and Item Type	ICCR Items	Connecticut Items	MOU Items[a]	Total Pool Items
Grade 5	**73**	**66**	**100**	**239**
Cluster	26	30	54	110

Grade and Item Type	ICCR Items	Connecticut Items	MOU Items[a]	Total Pool Items
Stand-Alone	47	36	46	129
Grade 8	**65**	**62**	**151**	**278**
Cluster	24	32	92	148
Stand-Alone	41	30	59	130
Grade 11	**67**	**79**	**62**	**208**
Cluster	19	26	28	73
Stand-Alone	48	53	34	135
Total	**205**	**207**	**313**	**725**

Note. [a]Other MOU state items administered include Hawaii, Idaho, MSSA (Rhode Island and Vermont), Oregon, Utah, West Virginia, and Wyoming.

Table 17. Spring 2021 Connecticut NGSS Assessment Operational Item Pool

Grade and Item Type	ICCR Operational Items	Connecticut Operational Items	MOU Operational Items[a]	Total Operational Pool Items
Grade 5	**72**	**38**	**93**	**203**
Cluster	26	19	49	94
Stand-Alone	46	19	44	109
Grade 8	**62**	**38**	**145**	**245**
Cluster	24	18	89	131
Stand-Alone	38	20	56	114
Grade 11	**67**	**30**	**62**	**159**
Cluster	19	15	28	62
Stand-Alone	48	15	34	97
Total	**201**	**106**	**300**	**607**

Note. [a]Other MOU state operational items administered include Hawaii, Idaho, MSSA (Rhode Island and Vermont), Oregon, Utah, West Virginia, and Wyoming.

Table 18. Spring 2021 Connecticut NGSS Assessment Field-Test Item Pool

Grade and Item Type	ICCR Field-Test Items	Connecticut Field-Test Items	MOU Field-Test Items[a]	Total Field-Test Pool Items
Grade 5	**1**	**28**	**7**	**36**
Cluster	0	11	5	16
Stand-Alone	1	17	2	20
Grade 8	**3**	**24**	**6**	**33**
Cluster	0	14	3	17
Stand-Alone	3	10	3	16

Grade and Item Type	ICCR Field-Test Items	Connecticut Field-Test Items	MOU Field-Test Items[a]	Total Field-Test Pool Items
Grade 11	**0**	**49**	**0**	**49**
Cluster	0	11	0	11
Stand-Alone	0	38	0	38
Total	**4**	**101**	**13**	**118**

Note. [a]Other MOU state field-test items administered include Idaho and MSSA (Rhode Island and Vermont).

Table 19. Spring 2021 Connecticut NGSS Assessment Operational and Field-Test Item Pool by Science Discipline

Grade	Science Discipline	Item Type	ICCR Items	Connecticut Items	MOU Items[a]	Total Pool Items
5	Earth and Space Sciences	Cluster	9	10	15	34
		Stand-Alone	11	11	19	41
	Life Sciences	Cluster	9	9	20	38
		Stand-Alone	17	10	14	41
	Physical Sciences	Cluster	8	11	19	38
		Stand-Alone	19	15	13	47
8	Earth and Space Sciences	Cluster	8	5	29	42
		Stand-Alone	12	8	19	39
	Life Sciences	Cluster	6	17	35	58
		Stand-Alone	20	12	17	49
	Physical Sciences	Cluster	10	10	28	48
		Stand-Alone	9	10	23	42
11	Earth and Space Sciences	Cluster	4	7	5	16
		Stand-Alone	11	16	5	32
	Life Sciences	Cluster	10	9	19	38
		Stand-Alone	29	11	18	58
	Physical Sciences	Cluster	5	10	4	19
		Stand-Alone	8	26	11	45
Total			**205**	**207**	**313**	**725**

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

Table 20. Spring 2021 Connecticut NGSS Assessment Operational and Field-Test Item Pool by Disciplinary Core Idea

Grade	Science Discipline	Disciplinary Core Idea	ICCR Items	Connecticut Items	MOU Items[a]	Total Pool Items
5	Earth and Space Sciences	ESS1	7	6	10	23
		ESS2	9	9	18	36
		ESS3	4	6	6	16
	Life Sciences	LS1	10	7	12	29
		LS2	4	3	7	14
		LS3	1	3	6	10
		LS4	11	6	9	26
	Physical Sciences	PS1	6	6	8	20
		PS2	8	8	9	25
		PS3	12	7	11	30
		PS4	1	5	4	10
8	Earth and Space Sciences	ESS1	8	4	14	26
		ESS2	5	5	16	26
		ESS3	7	4	18	29
	Life Sciences	LS1	6	10	19	35
		LS2	10	6	17	33
		LS3	3	4	6	13
		LS4	7	9	10	26
	Physical Sciences	PS1	5	4	23	32
		PS2	2	4	12	18
		PS3	7	7	9	23
		PS4	5	5	7	17
11	Earth and Space Sciences	ESS1	5	5	1	11
		ESS2	4	8	6	18
		ESS3	6	10	3	19
	Life Sciences	LS1	8	5	16	29
		LS2	8	7	9	24
		LS3	7	4	3	14
		LS4	16	4	9	29

Grade	Science Discipline	Disciplinary Core Idea	ICCR Items	Connecticut Items	MOU Items[a]	Total Pool Items
	Physical Sciences	PS1	7	15	8	30
		PS2	2	6	2	10
		PS3	4	10	4	18
		PS4	0	5	1	6
Total			**205**	**207**	**313**	**725**

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

More information about *p*-values, biserial correlations, and item response theory (IRT) parameters can be found in Volume 1, Annual Technical Report. The details on calibration, equating, and scoring of the Connecticut NGSS Assessment can also be found in Volume 1.

4.4 PAPER-PENCIL ACCOMMODATION FORM CONSTRUCTION

Student scores should not depend upon the mode of administration or type of test form. Because the Connecticut NGSS Assessment was primarily administered in an online test delivery system in spring 2021, only six students took the paper-pencil form in grade 5, two in grade 8, and three in grade 11. Scores obtained via alternate modes of administration must be established as comparable to scores obtained through online testing. This section outlines the overall test development plans that ensured the comparability of online and paper-pencil tests.

To build paper-pencil forms, content specialists began with the online pool and removed any items that could not be rendered on paper. Next, content specialists constructed fixed forms adhering to the test blueprint. All overall and discipline (reporting category)-level blueprint requirements were met; however, due to the availability of items in paper-pencil forms, some blueprint requirements at the DCI level were violated. For the grade 11 paper-pencil test, the blueprint for high school specified that the two item clusters for the Physical Sciences discipline could be taken from only two DCIs: "Matter and Its Interactions" and "Energy". However, none of the item clusters in the "Energy" DCI in the Connecticut pool were amenable to being rendered on paper because they contained either a simulation or animation. Furthermore, the blueprint specified that the two item clusters could not pertain to the same DCI. The two Earth and Space Sciences item clusters were selected from the "Earth and Human Activity" DCI, and two Physical Sciences item clusters were selected from the "Matter and its Interactions" DCI, violating the blueprint constraint that no more than one item cluster could be selected from the same DCI.

Future item development will focus on developing items that are amenable to paper delivery for those instances where the blueprint was not met.

4.5 REMOTE TESTING FORMS

For the 2020–2021 school year, remote testing forms were constructed to assess science among students taking the test remotely. They were 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 Connecticut NGSS 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 LOFT forms for the Connecticut NGSS 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 spring 2020 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
	Avg. 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
	Avg. 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
	Avg. 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 to Figure 5 illustrate the minimum and maximum boundaries of the TIF per grade.

Figure 2. Test Information Function, Selected Elementary School Form and Boundaries

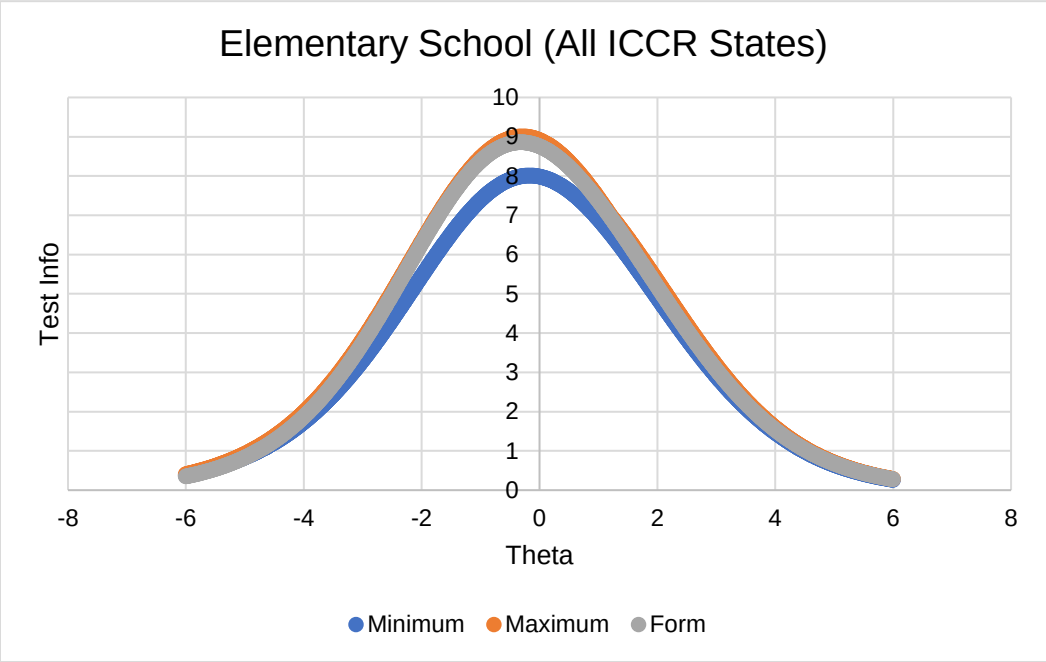


Figure 3. Test Information Function, Selected Middle School Form and Boundaries

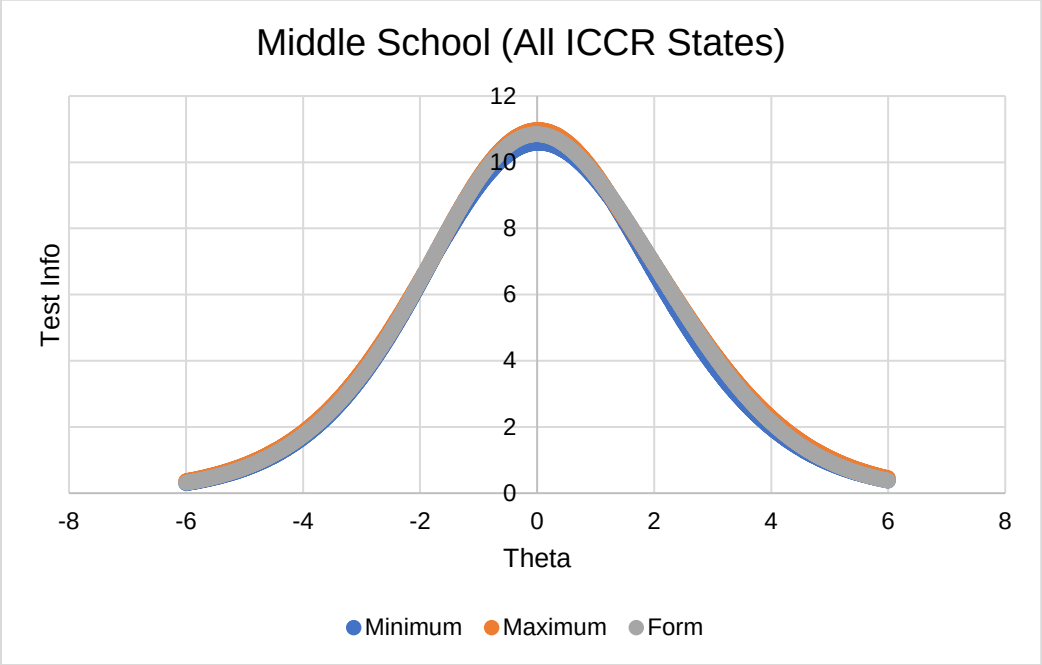


Figure 4. Test Information Function, Selected High School Form (New Hampshire and South Dakota) and Boundaries

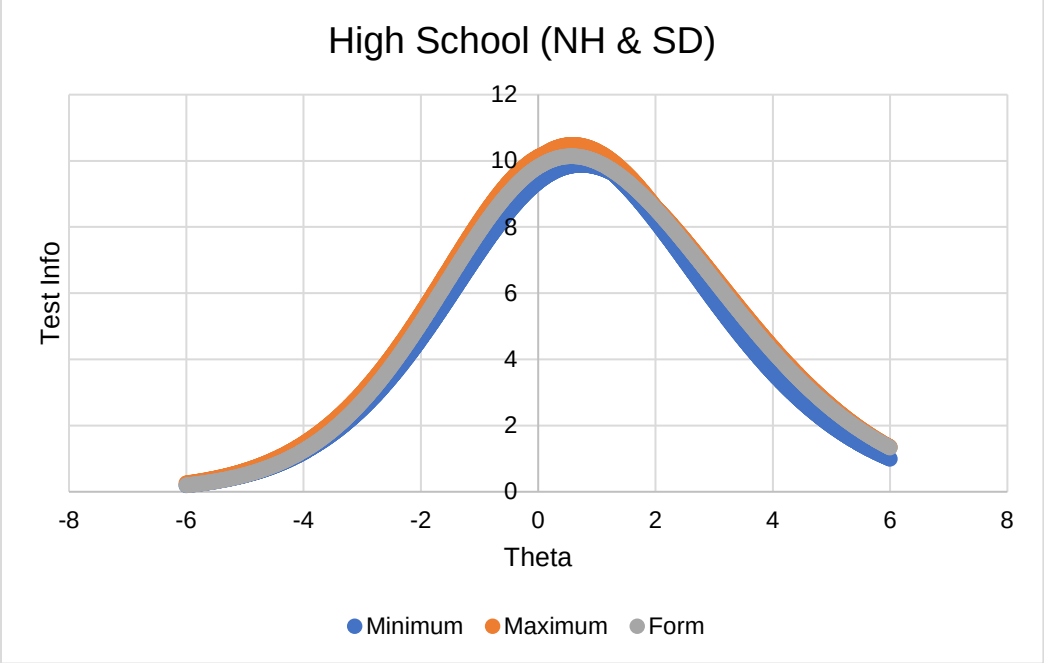
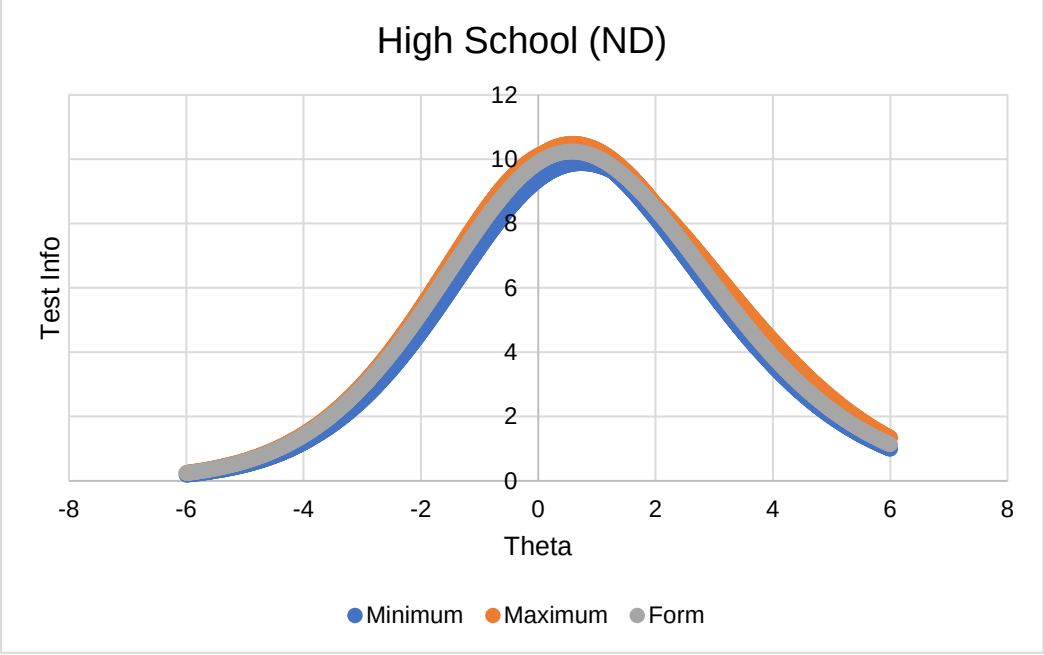


Figure 5. Test Information Function, Selected High School Form (North Dakota) and Boundaries



Among all possible forms adhering to the blueprint for the pool of eligible items, the forms were ranked according to how well their TIF mimicked the target information function (fell within the boundaries). Subsequently, the other psychometric criteria were 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 Connecticut 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 grade 11.

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

Grade	N. Assertions	Avg. b	Expected Time
5	67	-0.214	102.0
8	75	0.154	93.5
11	86	0.653	75.3

Figure 6. Test Information Function, Grade 5 Remote Form

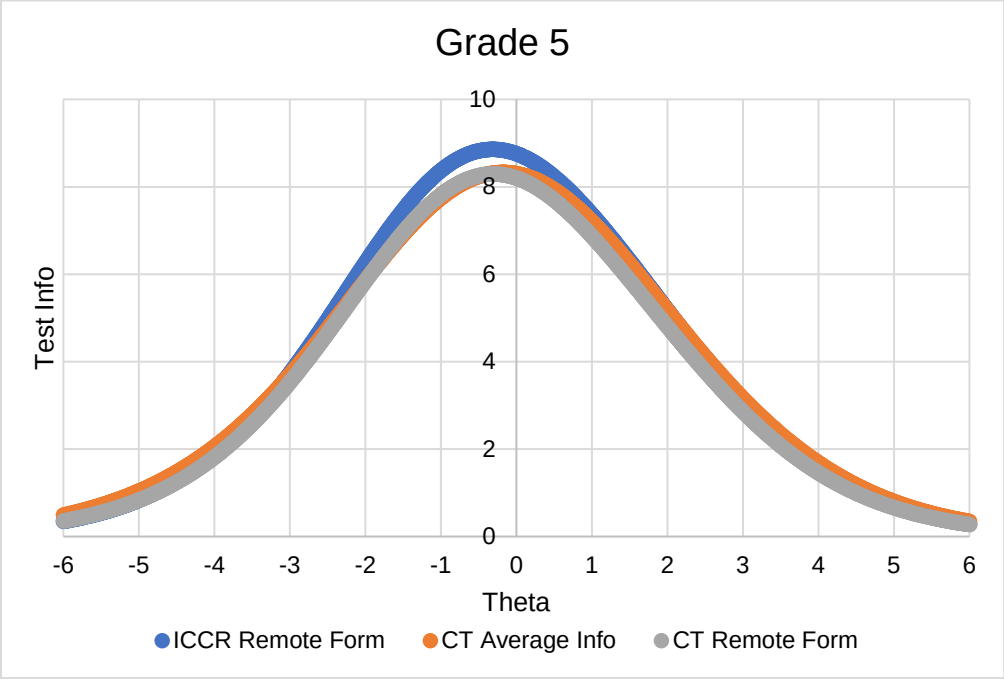


Figure 7. Test Information Function, Grade 8 Remote Form

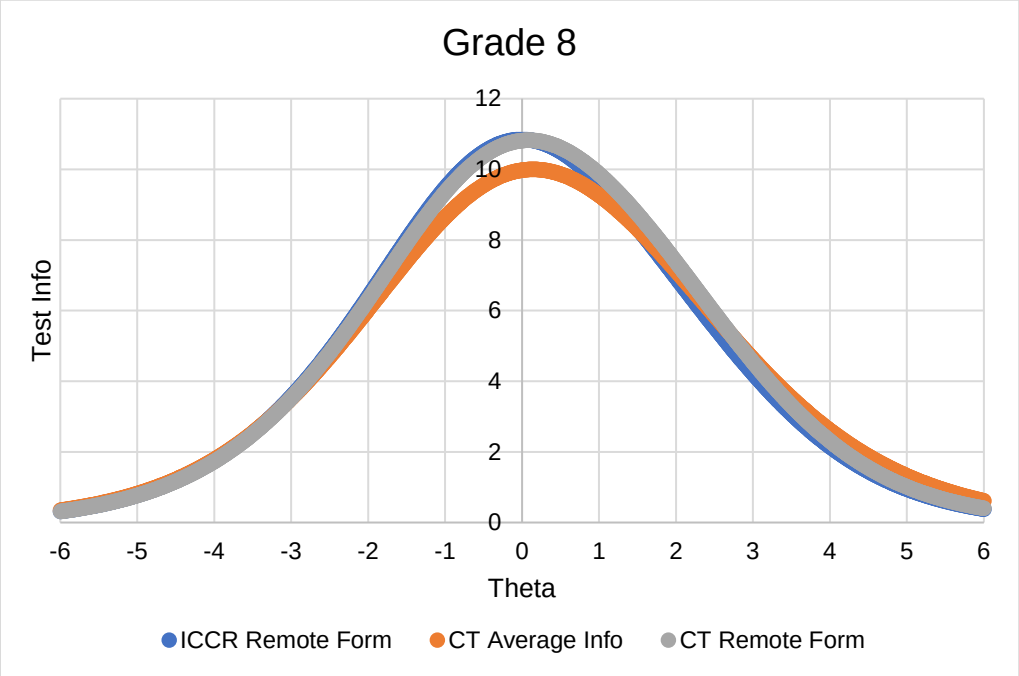
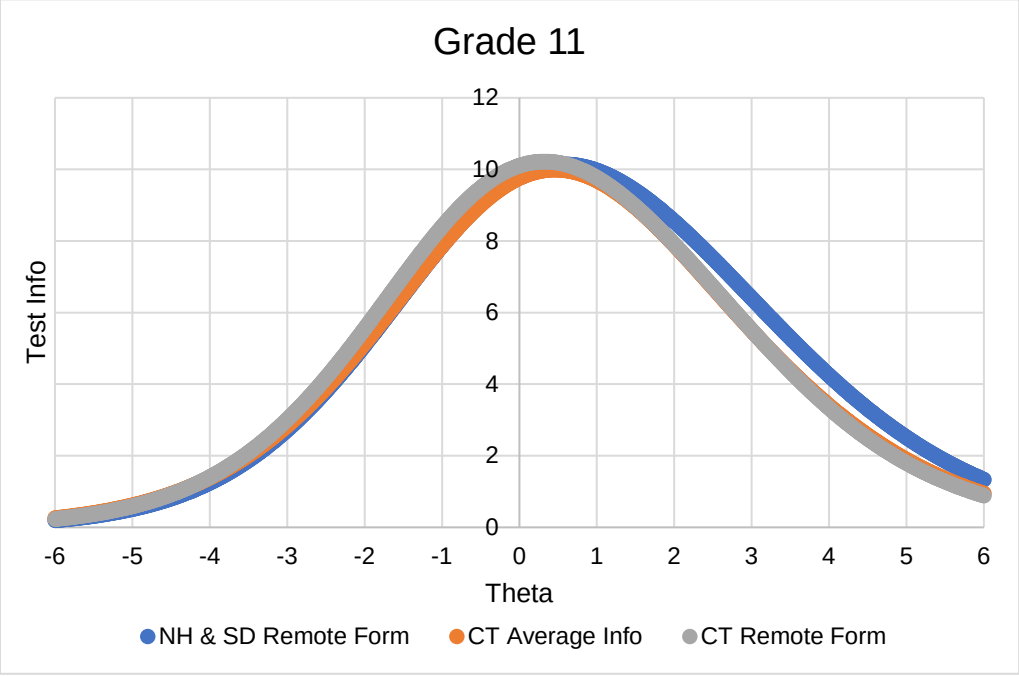


Figure 8. Test Information Function, Grade 11 Remote Form



Adaptations were made for the selected remote forms for Connecticut NGSS Assessment based on the ICCR remote forms. All adaptations were made to fulfill the blueprints for the Connecticut NGSS Assessment and to closely resemble the psychometric characteristics of the average

simulated forms in Connecticut. For grade 5, three ICCR items rejected from the online English test were removed, including one Earth and Space Sciences (ESS2) stand-alone item, and two Physical Sciences (PS) item clusters. Correspondingly, one ESS2 stand-alone item and two PS item clusters owned by Connecticut were added to meet the blueprint. To assess all the DCIs in PS, an ICCR PS2 stand-alone item was replaced with a Connecticut-owned PS4 stand-alone item.

For grade 8, an ICCR ESS3 stand-alone item was rejected and removed. Additionally, there were two ESS1 stand-alone items on the ICCR remote form, whereas the Connecticut grade 8 blueprint required only one ESS1 stand-alone item. As a result, one of the ICCR ESS1 stand-alone items was removed. Two Connecticut-owned ESS2 stand-alone items were then added to meet the blueprint.

For grade 11, the New Hampshire and South Dakota remote form was chosen as a starting point for the Connecticut-specific adaptations because their blueprints were more similar to Connecticut's. Three rejected ICCR items were removed first, including an ESS1 item cluster, a Life Sciences (LS2) item cluster, and a PS2 item cluster. Then, one of the two ICCR ESS1 stand-alone items was removed because the Connecticut grade 11 blueprint permits only one ESS1 stand-alone item. For better blueprint coverage, one of the two PS1 stand-alone items was removed. Five items were then added to fulfill the grade 11 blueprint: an ESS2 item cluster, a LS2 item cluster, a PS3 item cluster, an ESS2 stand-alone item, and a PS2 stand-alone item. All items were Connecticut-owned except for the first ESS2 item cluster. This item was owned by Oregon and used on the Connecticut grade 11 remote form with Oregon's prior consent. There was no other Connecticut-owned item that could fulfill the grade 11 blueprint.

5. SIMULATION SUMMARY REPORT

This section describes the results of the simulated test administrations used to configure and evaluate the adequacy of the item-selection algorithm used to administer the 2020–2021 Connecticut NGSS Assessment for grades 5, 8, and 11. Simulations were conducted to configure the settings of the algorithm and to evaluate whether individual tests adhered to the test blueprint.

Some important settings included cset1 and 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 2021, cset1 and cset2 values were set to 10 and 1. Psychometricians reviewed the simulation results and configured settings based on some key diagnostics, including:

- **Match-to-Test Blueprint.** This diagnostic determines whether the tests 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.** This diagnostic 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 are sharing items. The software system that performs the simulation allows adjustments to the setting parameters in order to attain the best possible balance among these diagnostics. The simulation involves an iterative process that reviews initial results, adjusts these 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

There are several factors that may influence simulation results for a linear-on-the-fly test (LOFT) administration. These factors 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 DCI, but the pool has no item cluster for some DCIs, it may be impossible to meet this constraint.
- *The correlational structure between constraints.* It is easier to satisfy a constraint if there are instances of the constraint at all levels of another constraint. For example, if stand-alone items within a discipline are associated only with a specific DCI, it may be difficult to meet both the desired distribution of content and the desired distribution of item type.
- *Whether or not there is a strict maximum on a given constraint.* This means that the requirement must be met exactly in each test administration.

5.2 RESULTS OF SIMULATED TEST ADMINISTRATIONS: ENGLISH

This section presents the simulation results for the English online tests, which is the test taken by most students (87.21%). 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 paper-pencil 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 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) is reported 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 global item exposure is minimal, it is expected that the largest proportion of items would appear in the bins of 0%<r<=20%, 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 fell into each exposure bin for all grades. Most test items were administered in 1%–20% of the test administrations. No item had an exposure rate less than 1%, and the minimum exposure rate was 1.49% in grade 5. A few items had an exposure rate higher than 60% because of the limitation of the current pool for some content categories.

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	199	0	0	28.14	68.34	3.02	0	0.50	0
8	245	0	0	37.14	59.18	3.27	0.41	0	0
11	159	0	0	11.95	76.1	10.69	0.63	0	0.63

5.3 RESULTS OF SIMULATED TEST ADMINISTRATIONS: SPANISH

This section presents the simulation results for the Spanish tests. The Spanish item pool consisted of a subset of ICCR items and some MOU items that had Spanish translations available. Table 25 presents the number of items available for the Spanish tests in spring 2021.

Table 25. Spring 2021 Spanish Operational Item Pool

Grade	Item Type	Number of Items
5	Cluster	9
	Stand-Alone	21
8	Cluster	11
	Stand-Alone	21
11	Cluster	8
	Stand-Alone	23
Total		**93**

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

5.3.1 Summary of Blueprint Match

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

5.3.2 Item Exposure

Table 26 presents the percentage of items that fell into each exposure bin for all grades. Most items were administered in more than 20% of the test administrations. Some items had an exposure rate of 100% because of the limited Spanish item pool. Only those items were available to satisfy the blueprint constraints.

Table 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	30	0	0	0	3.33	13.33	46.67	10.00	26.67
8	32	0	0	0	0	31.25	40.62	3.12	25.00
11	31	0	0	0	9.68	35.48	12.90	6.45	35.48

6. Operational Test Administration Summary Report

This section presents the blueprint match reports and item exposure rates for the spring 2021 operational test administrations.

6.1 Blueprint Match

Table 27 presents the percentages of the spring 2021 tests that violated the blueprint requirement. The English online tests in all grades met the blueprint specifications with a 100% match at all content levels. For the Spanish tests, all tests met the blueprint specifications with a 100% match at all content levels, except for three students in grade 5 and one student in grade 8.

Among the three students with blueprint violations in grade 5, one student received two stand-alone items and three items in total from the "Earth's Place in the Universe" (ESS1) DCI, two students received two stand-alone items and three items in total from the "Earth and Human Activity" (ESS3) DCI, while the blueprint requires at most one stand-alone item and two items in total from each of these DCIs. The one student with blueprint violations in grade 8 received two stand-alone items and three items in total from the "Earth and Human Activity" (ESS3) DCI, while the blueprint requires at most one stand-alone item and two items in total from this DCI.

These types of violations did not happen during simulations. The reason they occurred in the operational test administrations is that these students had seen the items that were designed to meet the blueprint requirements before the test administration. The item selection algorithm automatically filtered out the items they had already seen so the students would not see the same items twice. There are two possible scenarios for this type of violation to occur. First, the students saw the items in a previous attempt in the current school year. Second, these students took the science test at the same grade in previous test administrations. Therefore, the pool became shallower for these students. At the end of the test, the algorithm did not have the option to select an item that would satisfy the blueprint requirement, and it could only select an item that caused

blueprint violations. Note that these violations were all below the discipline level. The four students with blueprint violations in the spring 2021 Spanish tests took an item that would have satisfied the blueprint requirements in another English online test.

Table 27. Spring 2021 Blueprint Match for Test Delivered

Grade	Content Level	Min Items	Max Items	% of Cases Violating BP			
				+1	*+2*	*-1*	*-2*
Spanish							
5	DCI – ESS1	0	2	0.31%	–	–	–
	DCI – ESS3	0	2	0.61%	–	–	–
	DCI – Stand-Alone – ESS1	0	1	0.31%	–	–	–
	DCI – Stand-Alone – ESS3	0	1	0.61%	–	–	–
8	DCI – ESS3	0	2	0.25%	–	–	–
	DCI – Stand-Alone – ESS3	0	1	0.25%	–	–	–

6.2 ITEM EXPOSURE

Table 28 presents the item exposure rates for the spring 2021 test administration. The exposure rates were very similar to the simulation results described in Section 5.2.2, Item Exposure, for the English test administrations. The item exposure rate for field-test items ranged from 4% to 5% for all three grades. For the Spanish tests, more items had high exposure rates compared to the English tests because of a smaller item pool. Also, the operational exposure rates were slightly different from the simulation results because of small population sizes in all three grades. In spring 2021, fewer than 500 students took the Spanish test in each grade.

Table 28. Item Exposure Rates by Grade: Percentage of Items by Exposure Rate, Across All Spring 2021 Test Administrations

Grade	Total Items	[0,0]%	[0,1]%	[1,5]%	[5,20]%	[20,40] %	[40,60] %	[60,80] %	[80, 100]%
English									
5	199	0	0	29.15	67.34	3.02	0	0.50	0
8	245	0	0	37.55	58.78	3.27	0.41	0	0
11	159	0	0	11.96	75.47	11.31	0.63	0	0.63
Spanish									
5	30	0	0	0	0	16.67	46.67	6.67	30.00
8	32	0	0	0	0	34.38	34.38	6.25	25.00
11	31	0	0	0	6.45	35.48	16.13	6.45	35.48

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.