New Hampshire Statewide Assessment System

2018–2019

Volume 2 Part 2, Science Test Development

New Hampshire
Department of Education
Serving New Hampshire's Education Community

TABLE OF CONTENTS

LIST OF TABLES

LIST OF APPENDICES

1. INTRODUCTION

The New Hampshire Statewide Assessment System (NH SAS) for Science was first administered to students during spring 2018, replacing the New England Common Assessment Program in Science. The New Hampshire Statewide Assessment for Science was delivered to students in grades 5, 8, and 11 as an online assessment, constructed linearly on the fly, making use of several technology-enhanced item types.

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

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

- The test item specifications provided detailed guidance for item writers and reviewers to ensure that AIRCore science items were aligned to the performance expectations they were intended to measure.
- The item development procedures employed for NH SAS for science tests were consistent with industry standards.
- The development and maintenance of the AIRCore item pool plan established an item bank in which test items cover the range of measured performance expectations, 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 that were required on each test administration. This document was implemented in the item selection algorithm for science.

Note that for science assessments, as outlined in Volume 1, the American Institutes for Research (AIR) works with a group of states that share common item development processes. In addition to developing items for each of those states, AIR develops and maintains the AIRCore item bank, which consists of items that are developed according to the same principles that followed for the items owned by each of the states. Therefore, this volume focuses on the general test development activities, even though the NH SAS science tests draw exclusively from the AIRCore item bank. It is indicated in this volume which processes for the AIRCore bank deviate from the overall process and how they deviate.

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

1.1 Claim Structure

The goals, uses, and claims that the science item bank and subsequent tests would be designed to support were identified in a series of collaborative meetings over August 22–23, 2016, as an attempt to facilitate the transition from Next Generation Science Standards (NGSS) content standards to statewide summative assessments for science. AIR invited content and assessment leaders from 10 states as well as four nationally-recognized experts that helped co-author the NGSS standards. Two nationally-recognized psychometricians also participated.

AIR staff and participating states collaborated to develop items and test specifications to measure the NGSS. The item specifications were generally accompanied by sample item clusters meeting those specifications. All specifications and sample 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 AIRCore item bank for science was established using a highly structured, evidence-centered design. The process began with detailed item specifications. The specifications, discussed in Section 2.2, described the interaction types that can 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, translations, or assistive technologies. This goal is supported by the delivery of the items on AIR's test delivery platform, 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 items and item clusters 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.

AIR sought to ensure that the items were measuring the performance expectations in a fair and meaningful way by engaging educators and other stakeholders at each step of the process. Educators evaluated the alignment of items to the performance expectations and offered guidance and suggestions for improvement. They participated in the review of items for fairness and sensitivity. Following the field testing of items, 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 performance expectations with fidelity and does so in a way that minimizes construct-irrelevant variance and barriers to access. The details of these processes follow.

1.3 Organization of this Volume

This volume is organized in three subsequent sections:

- An overview of the science item development process that supports the validity of the claims that the science tests are designed to support
- An overview of the science item pool, the types of assessments the pool is designed to support, and methods for refreshing the pool
- A description of test construction for the New Hampshire Statewide Assessment System (NH SAS) for science, 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

AIR developed the AIRCore science item bank using a rigorous, structured process that engaged stakeholders at critical junctures. This process was managed by AIR'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 AIR 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 of these steps has a role in ensuring that the items can support the claims that will be based on them. Exhibit A describes how each step contributes to these goals. Each step in the process is discussed in more detail below.

Exhibit A: Summary of How Each Step of Development Supports the Validity of Claims

	Supports alignment to the performance expectations	**Reduces construct-irrelevant variance through universal design**	**Expands access through linguistic and other supports**
Item specifications	Specifies item interactions, content limits, and guidelines for meeting task demands and levels of	Avoids the use of any item interactions with accessibility constraints and provides language	

	Supports alignment to the performance expectations	Reduces construct-irrelevant variance through universal design	Expands access through linguistic and other supports
	cognitive engagement requirements and adjusting difficulty.	guidelines. Allows for multiple response modes to accommodate different styles.	
Selection and training of item writers	Ensures that item writers have the background to understand the performance expectations and specifications. Teaches item writers about selection of item interactions for measurement and accessibility.	Training in language accessibility, bias, and sensitivity helps item writers avoid unnecessary barriers.	
Writing and internal review of items	Checks content alignment and evaluates and improves overall quality.	Eliminates editorial issues, and flags and removes bias and accessibility issues.	
Markup for translation and accessibility features		Adds universal features, such as text-to-speech for science, that reduce barriers.	Adds text-to-speech, 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 check on quality and flags issues.	Flags items that appear to function differently for subsequent review for issues.	May reveal usability or implementation issues with markup.
Post-field-test reviews	Final, more focused check on flagged items. Rubric validation ensures that scoring reflects performance expectations.	Final, focused review on items flagged for differential item functioning.	

2.2 Item Specifications

AIR is working with a group of states, psychometricians, and science experts, including the authors of the NGSS, to develop powerful, innovative solutions to the challenges of measuring the NGSS. Participating states include Connecticut, Hawaii, Idaho, Oregon, Rhode Island, Utah, Vermont, West Virginia, and Wyoming. New Hampshire participates in some activities. This collaboration has yielded item specifications for NGSS performance expectations, sample item clusters for each specification, and hundreds of NGSS item clusters and stand-alone items in various stages of development. Under this collaboration, states have jointly developed item specifications.

Test item specifications are documents that are designed to guide the work of item writers as they craft test questions and the reviews of those items by stakeholders. 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 that comprise each NGSS, and which together properly structure into coherent items and item clusters. Exhibit B provides a sample of the item specifications developed by content experts for a middle school life sciences Performance Expectation (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 section 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.

Exhibit B: Sample Science Item Cluster Specifications for Middle School

Performance Expectation	**MS-LS1-1** Conduct an investigation to provide evidence that living things are made of cells; either one cell or many different numbers and types of cells.		
Dimensions	**Planning and Carrying Out Investigations** • Conduct an investigation to produce data to serve as the basis for evidence	**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	**Scale, Proportion, and Quantity** • Phenomena that can be observed at one scale may not be observable at another scale.

	that meets the goals of an investigation.	numbers and types of cells (multicellular).	

Clarifications and Content Limits	**Clarification Statements**
	• Emphasis is on developing evidence that living things are made of cells, distinguishing between living and non-living things, and understanding that living things may be made of one cell or many varying cells.
	
	Content Limits
	• <u>Students do not need to know the following</u>:
	  ○ 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		

Science Vocabulary Students are Expected to Know	Multicellular, unicellular, cell, tissue, organ, system, organism hierarchy, bacteria, colony, yeast, prokaryote, eukaryote, magnify, microscope, DNA, nucleus, cell wall, cell membrane, algae, chloroplasts, chromosome, cork		
Science Vocabulary Students are Not Expected to Know	Differentiation, mitosis, meiosis, genetics, cellular respiration, energy transfer, RNA, protozoa, amoeba, histology, Protista, archaea, nucleoid, plasmid, diatoms, cyanobacteria		
Phenomena			
Context/ Phenomena	Some example phenomena for MS-LS1-1: • Plant leaves and roots have tiny box-like structures that can be seen under a microscope. • Small creatures can be seen swimming in samples of pond water viewed through a microscope. • Different parts of a frog's body (e.g., muscles, skin, tongue) are observed under a microscope, and are seen to be composed of cells. • One-celled organisms (e.g., bacteria, protists) perform the eight necessary functions of life, but nothing smaller has been seen to do this. • Swabs from the human cheek are observed under a microscope. Small cells can be seen.		
This Performance Expectation and associated Evidence Statements support the following Task Demands.			
Task Demands			
1. Identify from a list, including distractors, the materials/tools needed for an investigation to find the smallest unit of life (cell).			
2. Identify the outcome data that should be collected in an investigation of the smallest unit of living things.			
3. Evaluate the sufficiency and limitations of data collected to explain that the smallest unit of living things is the cell.			
4. Make and/or record observations about whether the sample contains cells.			

5. Interpret and/or communicate data from the investigation to determine if a specimen is alive or not.
6. Construct a statement to describe the overall trend suggested by the observed data.

*Denotes those task demands which are deemed appropriate for use in stand-alone item development

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

2.3 Selection and Training of Item Writers

All item writers developing AIRCore science items at AIR 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 Next Generation Science Standards (NGSS) specifications.

Key materials are shown in Appendix A. These include

- AIR'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

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

AIRCore and MOU science items go through a rigorous, multiple-level internal review process before they are sent to external review. Staff members are trained to review items for both content and accessibility throughout the entire process. A sample item review checklist that our test developers use is included in Appendix B. The AIRCore internal review cycle includes the following phases:

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

2.4.1 Preliminary Review

Preliminary Review is conducted by team leads or senior content staff. Sometimes Preliminary Review is conducted in a group setting, led by a senior test developer. During the process, team leads or senior content staff analyze items to ensure 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 worthwhile in a reasonable way/it is a discrete observation that grounds a scenario which allows for the assessment of something worthwhile in a meaningful way).
- The item is properly aligned to the task demands.
- The vocabulary used in the item is appropriate for the grade and subject matter.
- The item considers language accessibility, bias, and sensitivity.
- The content is accurate and straightforward.
- The graphic and stimulus materials are necessary to answer the question.
- The stimulus is clear, concise, and succinct (i.e., it contains enough information to know what is being asked, it is stated positively, and it does not rely on negatives—such as *no, not, none, never*—unless absolutely necessary).

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

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

2.4.2 Scoring Entry and Review

At Scoring Entry level, the item writer inputs the machine scoring so that it can be reviewed by the team lead or senior staff that is reviewing the item prior to Content Review One. This step is kept separate from preliminary review so that the senior staff can suggest changes to the interaction at preliminary review without requiring the writer to overhaul scoring that they have already created. It also allows the senior staff to ensure that the scoring suggested by the writer at preliminary review is appropriate. This ensures the scoring is entered once, streamlining the process. At this level, the scoring is analyzed to ensure the following:

- The scoring works as it is 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 tell them they must round to a certain decimal place, we score them as such).

The senior staff approves the intent of the scoring at preliminary review. At scoring entry, the writer inputs this approved scoring, after which the senior staff checks the functionality of the scoring. Once the scoring is determined to be working correctly, the senior staff signs off on it 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 all the criteria identified for Preliminary Review. He or she also ensures that the revisions made during the Preliminary Review did not introduce errors or content inaccuracies. This reviewer approaches the item both from the perspective of potential clients as well as his or her own experience in test development.

2.4.4 Edit Review

During Edit Review, editors have four primary tasks:

1. Editors perform basic line editing for correct spelling, punctuation, grammar, and mathematical and scientific notation, ensuring consistency of style across the items.
2. Editors ensure that all items are accurate in content. Editors compare reading passages against the original publications to make sure that all information is internally consistent across stimulus materials and items, including names, facts, or cited lines of text that appear in the item. Editors ensure that the keys are correct and that all information in the item is correct. For items with mathematical tasks, editors perform all calculations to ensure accuracy.
3. Editors review all material for fairness and language accessibility issues.
4. Editors confirm that items reflect the accepted guidelines for good item construction. In all items, they look 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 and that the key accurately and correctly answers the question as posed, 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 an AIRCore science item arrives at Senior Review, it has been thoroughly vetted by both content reviewers and editors. Senior reviewers (in particular, senior content specialists) look back at the item's entire review history, making sure that all the issues identified in that item have been adequately addressed. Senior reviewers verify the overall content of each item, confirming its accuracy, alignment to the performance expectation, and consistency with the expectations for the highest quality. They check whether the scoring is working as intended and that the scoring assertions adequately address the evidence the student provides with each type of response.

2.5 Review by State Personnel and Stakeholder Committees

All AIRCore science items have been through an exhaustive external review process. Items in the science bank were reviewed by content experts in one or several states and reviewed and approved by multiple stakeholder committees to evaluate both content and bias/sensitivity.

2.5.1 State Review

After items have been developed for a state participating in the MOU, content experts from the state that owns the item review any eligible items prior to committee review. At this stage in the review process, clients can request edits, such as wording edits, scoring edits, alignment changes, or task demand updates. An AIR director for science reviews all client-requested edits in light of 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.

AIRCore items are reviewed by at least one or two states. The states provide feedback on the AIRCore items, and the AIR science leadership gathers suggestions and makes edits that improve the AIRCore item. Not all suggestions are implemented, as these items are owned by AIR. Further, most MOU states accept or reject AIRCore and MOU items (as they appear at the time) to be presented to their committees. Some clients skip this step and allow AIR to review all items with their committees before reviewing them. These items can be either set for field testing in a future administration or already at locked operational pool.

2.5.2 Content Advisory Committee Reviews

During the Content Advisory Committee (CAC) reviews, items are reviewed for content validity, grade-level appropriateness, and alignment to the performance expectation. CAC members are typically grade-level and subject-matter experts. During this review, educators also ensure that the scoring assertions make clear what is being scored as correct and give

credit where they should (see more information in the Rubric Validation section which follows).

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

A summary of the committee meetings appears in Exhibit C, with further details about the participants in Appendix C.

Exhibit C: Summary of Content Advisory Committee Meetings

Project	Meeting	Number of Committee Members	Number of Items Reviewed
AIRCore	March 2018	26	152
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
Hawaii	July 2017	22	25
	September 2017	20	65
	October 2018	29	85
	February 2019	21	44
Idaho	December 2018	21	111
MSSA	January 2018	42	73
	March 2018	28	100
	January 2019	21	116
Oregon	August 2017	10	110
	August 2018	18	256
	December 2018	16	62
Utah	July 2017	23	55
	December 2017	36	48
West Virginia	January 2017	28*	39
	January 2019	10	191
	July 2019	12	50
Wyoming	December 2017	17	51

Project	Meeting	Number of Committee Members	Number of Items Reviewed
	October 2018	14	37

* Number of Committee Members includes total committee members for ELA, mathematics, and science. The number for science only committee members is not available.

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.

A summary of the committee meetings appears in Exhibit D, with additional details about the participants in Appendix D.

Exhibit D: Summary of Fairness Committee Meetings

Project	Meeting	Number of Committee Members	Number of Items Reviewed	Number of Items Rejected
AIRCore	March 2018	13	152	N/A
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	*
Hawaii	July 2017	22	25	2
	September 2017	20	65	13
	October 2018	29	85	6
	February 2019	21	44	0
Idaho	December 2018	15	111	1
MSSA	January 2018	21	73	14
	March 2018	11	100	24
	January 2019	14	116	18
Oregon	August 2017	5	110	5
	August 2018	9	256	56

Project	Meeting	Number of Committee Members	Number of Items Reviewed	Number of Items Rejected
	December 2018	11	62	13
Utah	August 2017	6	44	2
	December 2017	6	48	1
West Virginia	January 2017	28**	34	N/A
	January 2019	10	191	N/A
Wyoming	December 2017	5	51	3
	October 2018	5	37	N/A

* Number of rejected items has not been finalized through client resolution at the time of writing this report.
** Number of committee members includes total committee members for ELA, mathematics, and science. The number for science only committee members is not available.

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. Text-to-speech tagging is applied prior to field testing while Spanish translations and braille are applied post field test. Accessibility markup is embedded into each item as part of the item development process rather than as a post-hoc process applied to completed tests.

Accessibility markup, whether translations or for text-to-speech, 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, AIRCore science items are tagged with text-to-speech. Spanish translations, including Spanish text-to-speech, and braille are available for a subset of items.

2.6 Field Testing

A large pool of items was field tested in nine states in spring 2018 for science. 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 wave of items was field tested in nine states. For Hawaii, Idaho elementary school, and Wyoming, unscored field-test items were added as a separate segment to the operational (scored) legacy science test. For a sample of Idaho middle schools, an independent field test in which students were administered a full set of items was conducted. In Connecticut, New Hampshire, Oregon, Rhode Island, Vermont, and West Virginia, field-test items were administered as unscored items embedded within the operational items. AIR's field-testing process is described in detail in Volume 1, Section 3.1.4.

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 below.

2.7.1 Rubric Validation

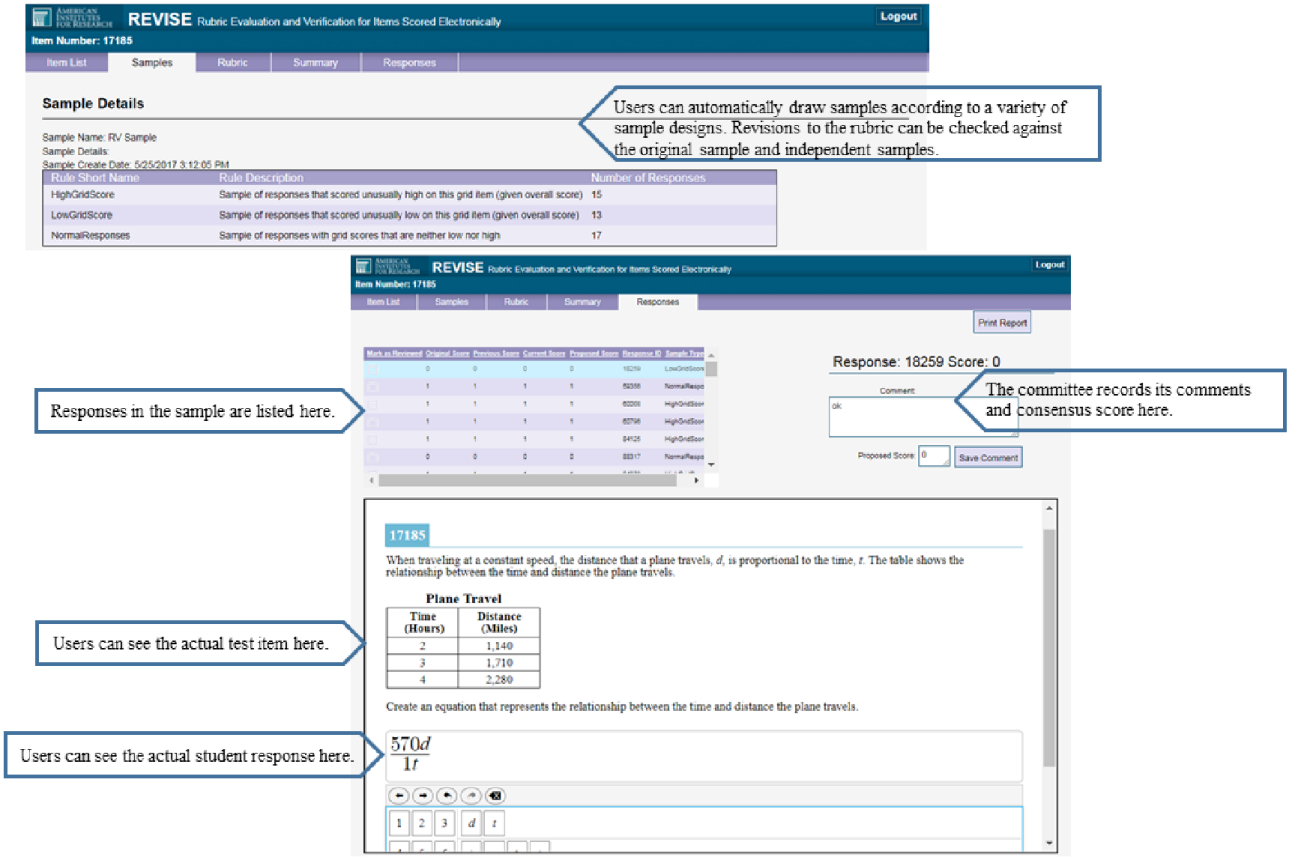
The validation process of field-test items begins with rubric validation to verify and make any necessary revisions to the scoring rubrics. The rubric validation process occurs in two phases.

During the first phase, AIR content experts work with the analysis team to prepare for the rubric validation meetings. The AIR content experts use the REVISE system to generate student responses that are scientifically sampled to overrepresent responses most likely to have been mis-scored. Specifically, the sample overrepresents: (a) low-scored responses from otherwise high-scoring students, and (b) high-scored responses from otherwise low-scoring students. This process allows AIR to identify any potential scoring concerns before the rubric validation meeting, such as unanticipated (but accurate) responses, equivalent responses that were not originally considered, and responses that are getting credit but should not (based on the content and the item rubric). The rubrics may be adjusted, and responses rescored at this point.

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

Exhibit E shows some features from REVISE.

Exhibit E: Features of the REVISE Software



After the rubric validation meetings, AIR staff apply the approved revisions to the rubrics, and any items rejected as part of the process are rejected in the Item Tracking System (ITS). ITS archives critical information regarding the scoring certification completed during the rubric validation process. This includes any rubric changes made during the scoring decision meetings and the sign-off completed by the senior content expert once the rubric has been changed, rescoring the entire sample has been completed, and it has been verified that the scoring used 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 statistics. The states established standards for the testing 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 AIRCore items, the data from students testing in Connecticut, Idaho grade 8, New Hampshire, Rhode Island, Vermont, Oregon, and West Virginia were combined (states that administered AIRCore items and utilized either an independent or operational test).

Volume 1, Section 4, describes in detail the statistical flags that send items to data review. The flags are designed to highlight potential content weaknesses, miskeys, or possible bias issues. Committee members are taught to interpret these flags and are given guidelines for examining the items for content or fairness issues.

For each of the states participating in the MOU, flagged items owned by the state were reviewed by a data review committee. The composition of the data review committees generally consisted of content experts from the state's department of education or state educators (in this case, the state educators were science teachers) and were supported by AIR content experts. AIRCore items were distributed over the data review committees of states participating in the MOU. In summer 2018, AIRCore field-test items were reviewed in webinars with committee members from several states in each session. Outcomes were decided by AIR science content leadership. In summer 2019, AIRCore field-test items were taken to Connecticut, Hawaii, and Idaho for committee review. Outcomes were decided by AIR science content leadership, taking the committees' recommendations into consideration.

At the start of each state-owned item data review meeting, AIR staff leads participants in a training session to familiarize them with the item development process, the purpose of data review, the meaning of the various flags, and the purpose of the data review committee. Committee members are taught to interpret the various flags and are given guidelines for examining the items for content or fairness issues. The training includes a group review of item cards which detail specific item attributes (including grade level and alignment to the science performance expectations, the content and rubric of the item, and the various item statistics). A sample of the training materials used for these data review meetings appears in **Error! Reference source not found.**. Participants use an online environment via laptop computers to review the items in order to interact with them in a manner similar to that of students, and also to view all statistics associated with each item.

Items are then reviewed by participants who are most familiar with the particular grade (band) level and content domain of these items. AIR content specialists, who are also well versed in item statistics, facilitate the discussion in each room with AIR psychometricians available to answer questions as they arise. At the end of each meeting day, AIR content specialists meet with the state content specialists to review the committee recommendations and decide whether to accept the item for inclusion in the operational pool or reject the item from the operational pool. Items that were rejected are potentially eligible for changes to the item and an additional field test.

Exhibit F summarizes the data review committee meetings. Details, including the composition of each committee, appear in **Error! Reference source not found.**.

Exhibit F: Summary of Data Review Committee Meetings

Owner and Item Type	Meeting	Number of Committee Members	Number of Items Reviewed	Number of Items Rejected
AIRCore	July 2018	18	**84**	**8**
Cluster			33	2
Stand-alone			51	6
AIRCore	August 2019	*	**43**	**3**
Cluster			0	1
Stand-alone			43	2
Connecticut	August 2018	29	**18**	**11**
Cluster			7	5
Stand-alone			11	6
Connecticut	August 2019	29	**53**	**20**
Cluster			14	6
Stand-alone			39	14
Hawaii	August 2018	18	**32**	**3**
Cluster			7	1
Stand-alone			25	2
Hawaii	August 2019	18	**37**	**13**
Cluster			17	5
Stand-alone			20	8
Idaho	August 2019	10	**12**	**6**
Cluster			4	3
Stand-alone			8	3
MSSA	August 2018	2[a]	**9**	**6**
Cluster			2	0
Stand-alone			7	6
MSSA	August 2019	2[a]	**14**	**4**
Cluster			2	1
Stand-alone			12	3
Oregon	September 2018	11	**44**	**6**
Cluster			28	5
Stand-alone			16	1
Oregon	August 2019	4	**8**	**7**
Cluster			1	1
Stand-alone			7	6
Utah	August 2018	16	**40**	**6**
Cluster			40	6
Stand-alone			0	0
West Virginia	July 2018	4	**3**	**1**

Owner and Item Type	Meeting	Number of Committee Members	Number of Items Reviewed	Number of Items Rejected
Cluster			3	1
Stand-alone			0	0
West Virginia	September 2019	4	**7**	**6**
Cluster			1	1
Stand-alone			6	5
Wyoming	October 2018	19	**16**	**6**
Cluster			6	1
Stand-alone			10	5
Wyoming	August 2019	10	**16**	**5**
Cluster			4	3
Stand-alone			12	2

[a]Conducted by RIDE and VT AOE science content experts.
* In summer 2019, AIRCore field-test items were taken to Connecticut, Hawaii, and Idaho for committee review.
** Number of committee members unavailable at the time of writing this report

3. AIRCORE SCIENCE ITEM BANK SUMMARY

Tests based on or inspired by the NGSS framework, such as the NH SAS science assessment, 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 phenomenon. Some stand-alone items are added to increase the coverage of the test without also increasing the testing time or testing burden.

AIR Assessment has built the science 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 Next Generation Science Standards (NGSS). The science item bank comprises AIR-owned items, which are shared with partner states. These items follow the same specifications, test development processes, and review processes. In 2018, AIR 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 265 have passed rubric validation and item data review.

Each state using the science 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 form. The science 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. The New Hampshire science assessments draw exclusively from the AIRCore science item bank because its items are part of the larger across-state science item bank, this item bank is described below.

3.1 Current Composition of the Science Item Bank

The New Hampshire Statewide Assessment System (NH SAS) science assessments are composed of stand-alone items and item clusters. 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 are made about the knowledge and skills that a student has demonstrated based on specific features of the student's responses across multiple interactions. For example, a student may correctly graph data points indicating that they can construct a graph showing the relationship between two variables, but they may make an incorrect inference about the relationship between the two variables, therefore not supporting the assertion that the student can interpret relationships expressed graphically. Table 1 lists the science interaction types. Examples of various interaction types can be found in **Error! Reference source not found.**.

Table 1. Science Interaction Types and Descriptions

Interaction Type	Associated Sub-Types	Description
Choice	Multiple Choice	Traditional multiple-choice interaction allows the student to select a single option from a list of possible answer options.
	Multiple 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 a student to complete tabular data.
Edit Task	Edit Task	A student clicks a word and replaces it with another word that they type to revise a sentence.
	Edit Task with Choice	A student clicks a word or phrase and chooses the replacement from a number of options.
	Edit Task Inline Choice	Drop-down menus are placed through the text, and a student chooses the right option to complete the text.
Hot Text	Selectable	Selectable hot text interactions require students to select one or more text elements in the response area.

Interaction Type	Associated Sub-Types	Description
	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 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 allow you to create grid interactions with specific hot spot functionality. These interactions require students to select hot spot regions in the grid area.
	Graphic Gap Match	Graphic gap match interactions allow you to create grid interactions with specific drag-and-drop functionality. These interactions require students to drag image objects from a palette to specified regions (gaps) in the grid area.
Simulation	n/a	Simulation interactions allow the student to investigate a phenomenon by selecting variables to get output data. Some simulations are accompanied by animations.

Table 2 through Table 6 provide the number of items in the across-state science item bank available for use in the spring 2019 statewide assessments. Appendix H provides the across-state science item bank available by grade band, performance expectation (PE), and origin.

Table 2. Across-State Science Bank Spring 2019 Operational and Field-Test Item Pool

Grade Band	Item Type	Total Number of AIRCore Items	Total Number of MOU Items[a]	Total Number of Items
Elementary School	Cluster	32	94	126
	Stand-alone	47	79	126
Middle School	Cluster	29	146	175
	Stand-alone	50	96	146
High School	Cluster	30	71	101
	Stand-alone	54	70	124
Total		**242**	**556**	**798**

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

Table 3. Across-State Science Bank Spring 2019 Operational Item Pool

Grade Band	Item Type	Sp19 AIRCore OP Items	Sp19 MOU OP Items[a]	Total Sp19 OP Items
Elementary School	Cluster	32	44	76
	Stand-alone	29	30	59
Middle School	Cluster	25	112	137
	Stand-alone	26	31	57
High School	Cluster	28	38	66
	Stand-alone	27	29	56
Total		**167**	**284**	**451**

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

Table 4. Across-State Science Bank Spring 2019 Field-Test Item Pool

Grade Band	Item Type	Sp19 AIRCore FT Items	Sp19 MOU FT Items[a]	Total Sp19 FT Items
Elementary School	Cluster	0	50	50
	Stand-alone	18	49	67
Middle School	Cluster	4	34	38
	Stand-alone	24	65	89
High School	Cluster	2	33	35
	Stand-alone	27	41	68
Total		**75**	**272**	**347**

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

Table 5. Across-State Science Bank Spring 2019 Item Pool by Grade Band, Science Discipline, and Origin

Grade Band	Science Discipline	Item Type	AIRCore Items	MOU Items[a]	Total Items
Elementary School	Earth and Space Sciences	Cluster	11	26	37
		Stand-alone	12	27	39
	Life Sciences	Cluster	11	34	45
		Stand-alone	17	25	42
	Physical Sciences	Cluster	10	34	44
		Stand-alone	18	27	45
Middle School	Earth and Space Sciences	Cluster	9	39	48
		Stand-alone	17	30	47
	Life Sciences	Cluster	8	59	67
		Stand-alone	23	30	53
	Physical Sciences	Cluster	12	47	59
		Stand-alone	10	36	46
	Engineering, Technology, and Applications of Science	Cluster	0	1	1
		Stand-alone	0	0	0
High School		Cluster	6	15	21

Grade Band	Science Discipline	Item Type	AIRCore Items	MOU Items[a]	Total Items
	Earth and Space Sciences	Stand-alone	11	14	25
	Life Sciences	Cluster	16	37	53
		Stand-alone	35	31	66
	Physical Sciences	Cluster	8	19	27
		Stand-alone	8	25	33
Total			**242**	**556**	**798**

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

Table 6. Across-State Science Bank Spring 2019 Item Pool by Grade Band, Disciplinary Core Idea, and Origin

Grade Band	Science Discipline	Disciplinary Core Idea	AIRCore Items	MOU Items[a]	Total Items
Elementary School	Earth and Space Sciences	ESS1: Earth's Place in the Universe	7	15	22
		ESS2: Earth's Systems	10	28	38
		ESS3: Earth and Human Activity	6	10	16
	Life Sciences	LS1: From Molecules to Organisms: Structure and Function	10	23	33
		LS2: Ecosystems: Interactions, Energy, and Dynamics	4	11	15
		LS3: Heredity: Inheritance and Variation of Traits	2	9	11
		LS4: Biological Evolution: Unity and Diversity	12	16	28
	Physical Sciences	PS1: Matter and Its Interactions	6	15	21
		PS2: Motion and Stability: Forces and Interactions	7	19	26
		PS3: Energy	13	17	30
		PS4: Waves and Their Applications in Technologies for Information Transfer	2	10	12
Middle School	Earth and Space Sciences	ESS1: Earth's Place in the Universe	12	17	29
		ESS2: Earth's Systems	5	28	33
		ESS3: Earth and Human Activity	9	24	33
	Life Sciences	LS1: From Molecules to Organisms: Structure and Function	5	33	38
		LS2: Ecosystems: Interactions, Energy, and Dynamics	15	24	39
		LS3: Heredity: Inheritance and Variation of Traits	2	10	12

Grade Band	Science Discipline	Disciplinary Core Idea	AIRCore Items	MOU Items[a]	Total Items
		LS4: Biological Evolution: Unity and Diversity	9	22	31
	Physical Sciences	PS1: Matter and Its Interactions	6	33	39
		PS2: Motion and Stability: Forces and Interactions	3	22	25
		PS3: Energy	8	17	25
		PS4: Waves and Their Applications in Technologies for Information Transfer	5	11	16
	Engineering, Technology, and Applications of Science	ETS1: Engineering Design	0	1	1
High School	Earth and Space Sciences	ESS1: Earth's Place in the Universe	6	12	18
		ESS2: Earth's Systems	5	11	16
		ESS3: Earth and Human Activity	6	6	12
	Life Sciences	LS1: From Molecules to Organisms: Structure and Function	11	24	35
		LS2: Ecosystems: Interactions, Energy, and Dynamics	15	22	37
		LS3: Heredity: Inheritance and Variation of Traits	8	4	12
		LS4: Biological Evolution: Unity and Diversity	17	18	35
	Physical Sciences	PS1: Matter and Its Interactions	8	17	25
		PS2: Motion and Stability: Forces and Interactions	4	11	15
		PS3: Energy	4	11	15
		PS4: Waves and Their Applications in Technologies for Information Transfer	0	5	5
Total			**242**	**556**	**798**

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

3.2 Strategy for Pool Evaluation and Replenishment

AIR and MOU states continue to develop items to replenish and grow the science item pool. Our 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 pool

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

4. NH SAS Test Construction

The NH SAS science assessment was administered online to students in grades 5, 8, and 11 using a linear-on-the-fly (LOFT) test design. Contrary to a fixed form, every student potentially sees a different set of items. Items are selected by an item selection algorithm so that the blueprint is met whenever possible. The algorithm that was used is the same algorithm that AIR uses for the administration of adaptive tests. The adaptive item-selection algorithm selects items based on their content value and information value. By assigning weights of zero to the information value of an item with respect to the underlying latent variable, the items are solely selected based on their contribution to meeting the blueprint.

4.1 Test Design

The main characteristics of the test design were as follows. There were four segments on the test, each with its own item pool. The segments and respective item pools included:

- Life Sciences
- Earth and Space Sciences
- Physical Sciences
- Embedded field-test segment (all three disciplines)

For the three segments corresponding to science disciplines, which constituted the operational segments of the test, a student received two clusters and four stand-alone items of the respective discipline (see also the Min and Max cluster values of the blueprint in Table 7 through Table 9 at the discipline level). The fourth segment was an unscored embedded field-test (EFT) segment consisting of either one cluster or a set of five stand-alone items from the AIRCore field-test pool. The order of the four segments was randomized across students.

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 within each science discipline and DCI

The blueprint for science is given in Table 7 through Table 9.

Table 7. Science Test Blueprint, Grade 5 Science

Grade 5	Min Clusters	Max Clusters	Min Stand-Alone Items	Max Stand-Alone Items	Min Clusters + Min Stand-Alone Items	Max Clusters + Max Stand-Alone Items
Discipline – Physical Science, 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 & 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 & Properties of Matter	0	1	0	1	0	1
5-PS1-2: Structure & Properties of Matter	0	1	0	1	0	1
5-PS1-3: Structure & Properties of Matter	0	1	0	1	0	1
5-PS1-4: Structure & Properties of Matter	0	1	0	1	0	1
Discipline – Life Science, 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 & 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 & Energy	0	1	0	1	0	1

Grade 5	Min Clusters	Max Clusters	Min Stand-Alone Items	Max Stand-Alone Items	Min Clusters + Min Stand-Alone Items	Max Clusters + Max Stand-Alone Items
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 Science, PE Total = 13	**2**	**2**	**4**	**4**	**6**	**6**
DCI – Earth's Systems	**0**	**1**	**0**	**2**	**0**	**3**
3-ESS2-1: Weather & Climate	0	1	0	1	0	1
3-ESS2-2: Weather & Climate	0	1	0	1	0	1
4-ESS2-1: Earth's Systems & Processes	0	1	0	1	0	1
4-ESS2-2: Earth's Systems & 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**	**2**	**0**	**3**
3-ESS3-1: Weather & Climate*	0	1	0	1	0	1
4-ESS3-2: Earth's Systems & 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**	**2**	**0**	**3**
4-ESS1-1: Earth's Systems & 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 8. Science Test Blueprint, Grade 8 Science

Grade 8	Min Clusters	Max Clusters	Min Stand-Alone Items	Max Stand-Alone Items	Min Clusters + Stand-Alone Items	Max Clusters + Stand-Alone Items
Discipline – Physical Science, PE Total = 19	**2**	**2**	**4**	**4**	**6**	**6**
DCI – Matter and Its Interactions	**0**	**1**	**0**	**2**	**0**	**3**
MS-PS1-1: Structure & Properties of Matter	0	1	0	1	0	1
MS-PS1-2: Chemical Reactions	0	1	0	1	0	1

Grade 8	Min Clusters	Max Clusters	Min Stand-Alone Items	Max Stand-Alone Items	Min Clusters + Stand-Alone Items	Max Clusters + Stand-Alone Items
MS-PS1-3: Structure & Properties of Matter	0	1	0	1	0	1
MS-PS1-4: Structure & 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 & Interactions*	0	1	0	1	0	1
MS-PS2-2: Forces & Interactions	0	1	0	1	0	1
MS-PS2-3: Forces & Interactions	0	1	0	1	0	1
MS-PS2-4: Forces & Interactions	0	1	0	1	0	1
MS-PS2-5: Forces & 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 & Electromagnetic Radiation	0	1	0	1	0	1
MS-PS4-2: Waves & Electromagnetic Radiation	0	1	0	1	0	1
MS-PS4-3: Waves & Electromagnetic Radiation	0	1	0	1	0	1
Discipline – Life Science, 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 & Energy	0	1	0	1	0	1
MS-LS1-7: Matter & 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 & Energy	0	1	0	1	0	1
MS-LS2-2: Interdependent Relationships in Ecosystems	0	1	0	1	0	1
MS-LS2-3: Matter & Energy	0	1	0	1	0	1

Grade 8	Min Clusters	Max Clusters	Min Stand-Alone Items	Max Stand-Alone Items	Min Clusters + Stand-Alone Items	Max Clusters + Stand-Alone Items
MS-LS2-4: Matter & 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
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 & Adaptation	0	1	0	1	0	1
MS-LS4-2: Natural Selection & Adaptation	0	1	0	1	0	1
MS-LS4-3: Natural Selection & Adaptation	0	1	0	1	0	1
MS-LS4-4: Natural Selection & Adaptation	0	1	0	1	0	1
MS-LS4-5: Growth, Development, Reproduction	0	1	0	1	0	1
MS-LS4-6: Natural Selection & Adaptation	0	1	0	1	0	1
Discipline – Earth and Space Science, PE Total = 15	**2**	**2**	**4**	**4**	**6**	**6**
DCI – Earth's Place in the Universe	**0**	**1**	**0**	**2**	**0**	**3**
MS-ESS1-1: Space Systems	0	1	0	1	0	1
MS-ESS1-2: Space Systems	0	1	0	1	0	1
MS-ESS1-3: Space Systems	0	1	0	1	0	1
MS-ESS1-4: History of Earth	0	1	0	1	0	1
DCI – Earth's Systems	**0**	**1**	**0**	**2**	**0**	**3**
MS-ESS2-1: Earth's Systems	0	1	0	1	0	1
MS-ESS2-2: History of Earth	0	1	0	1	0	1
MS-ESS2-3: History of Earth	0	1	0	1	0	1
MS-ESS2-4: Earth's Systems	0	1	0	1	0	1
MS-ESS2-5: Weather & Climate	0	1	0	1	0	1
MS-ESS2-6: Weather & Climate	0	1	0	1	0	1
DCI – Earth and Human Activity	**0**	**1**	**0**	**2**	**0**	**3**
MS-ESS3-1: Earth's Systems	0	1	0	1	0	1
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 & Climate	0	1	0	1	0	1
Total PE = 55	**6**	**6**	**12**	**12**	**18**	**18**

*Note: These PEs have an engineering component.

Table 9. Science Test Blueprint, Grade 11 Science

Grade 11	Min Clusters	Max Clusters	Min Stand-Alone Items	Max Stand-Alone Items	Min Clusters + Stand-Alone Items	Max Clusters + Stand-Alone Items
Discipline – Physical Science, PE Total = 24	**2**	**2**	**4**	**4**	**6**	**6**
DCI – Matter and Its Interactions	**0**	**1**	**0**	**2**	**0**	**3**
HS-PS1-1: Structure and Properties of Matter	0	1	0	1	0	1
HS-PS1-2: Structure and Properties of Matter	0	1	0	1	0	1
HS-PS1-3: Structure and Properties of Matter	0	1	0	1	0	1
HS-PS1-4: Chemical Reactions	0	1	0	1	0	1
HS-PS1-5: Chemical Reactions	0	1	0	1	0	1
HS-PS1-6: Chemical Reactions*	0	1	0	1	0	1
HS-PS1-7: Chemical Reactions	0	1	0	1	0	1
HS-PS1-8: Nuclear Processes	0	1	0	1	0	1
DCI – Motion and Stability: Forces and Interactions	**0**	**1**	**0**	**2**	**0**	**3**
HS-PS2-1: Forces and Motion	0	1	0	1	0	1
HS-PS2-2: Forces and Motion	0	1	0	1	0	1
HS-PS2-3: Forces and Motion*	0	1	0	1	0	1
HS-PS2-4: Types of Interactions	0	1	0	1	0	1
HS-PS2-5: Types of Interactions	0	1	0	1	0	1
HS-PS2-6: Chemical Reactions*	0	1	0	1	0	1
DCI – Energy	**0**	**1**	**0**	**2**	**0**	**3**
HS-PS3-1: Energy	0	1	0	1	0	1
HS-PS3-2: Energy	0	1	0	1	0	1
HS-PS3-3: Energy*	0	1	0	1	0	1
HS-PS3-4: Energy	0	1	0	1	0	1
HS-PS3-5: Energy	0	1	0	1	0	1
DCI – Waves and Their Applications in Technologies for Information Transfer	**0**	**1**	**0**	**2**	**0**	**3**
HS-PS4-1: Wave Properties	0	1	0	1	0	1
HS-PS4-2: Wave Properties	0	1	0	1	0	1
HS-PS4-3: Wave Properties/Electromagnetic Radiation	0	1	0	1	0	1
HS-PS4-4: Electromagnetic Radiation	0	1	0	1	0	1
HS-PS4-5: Electromagnetic Radiation*	0	1	0	1	0	1
Discipline – Life Science, 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

Grade 11	Min Clusters	Max Clusters	Min Stand-Alone Items	Max Stand-Alone Items	Min Clusters + Stand-Alone Items	Max Clusters + Stand-Alone Items
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
HS-LS2-3: Cycles of Matter and Energy Transfer in Ecosystems	0	1	0	1	0	1
HS-LS2-4: Cycles of Matter and Energy Transfer in Ecosystems	0	1	0	1	0	1
HS-LS2-5: Cycles of Matter and Energy Transfer in Ecosystems	0	1	0	1	0	1
HS-LS2-6: Ecosystem Dynamics, Functioning, and Resilience	0	1	0	1	0	1
HS-LS2-7: Ecosystem Dynamics, Functioning, and Resilience*	0	1	0	1	0	1
HS-LS2-8: Social Interactions and Group Behavior	0	1	0	1	0	1
DCI – Heredity: Inheritance and Variation of Traits	**0**	**1**	**0**	**2**	**0**	**3**
HS-LS3-1: Structure and Function	0	1	0	1	0	1
HS-LS3-2: Variation of Traits	0	1	0	1	0	1
HS-LS3-3: Variation of Traits	0	1	0	1	0	1
DCI – Biological Evolution: Unity and Diversity	**0**	**1**	**0**	**2**	**0**	**3**
HS-LS4-1: Evidence of Common Ancestry and Diversity	0	1	0	1	0	1
HS-LS4-2: Natural Selection	0	1	0	1	0	1
HS-LS4-3: Natural Selection	0	1	0	1	0	1
HS-LS4-4: Adaptation	0	1	0	1	0	1
HS-LS4-5: Adaptation	0	1	0	1	0	1
HS-LS4-6: Adaptation*	0	1	0	1	0	1
Discipline – Earth and Space Science, PE Total = 19	**2**	**2**	**4**	**4**	**6**	**6**
DCI – Earth's Place in the Universe	**0**	**1**	**0**	**2**	**0**	**3**
HS-ESS1-1: The Universe and Its Stars	0	1	0	1	0	1
HS-ESS1-2: The Universe and Its Stars	0	1	0	1	0	1
HS-ESS1-3: The Universe and Its Stars	0	1	0	1	0	1
HS-ESS1-4: Earth and the Solar System	0	1	0	1	0	1
HS-ESS1-5: The History of Planet Earth	0	1	0	1	0	1
HS-ESS1-6: The History of Planet Earth	0	1	0	1	0	1
DCI – Earth's Systems	**0**	**1**	**0**	**2**	**0**	**3**

Grade 11	Min Clusters	Max Clusters	Min Stand-Alone Items	Max Stand-Alone Items	Min Clusters + Stand-Alone Items	Max Clusters + Stand-Alone Items
HS-ESS2-1: Earth Materials and Systems	0	1	0	1	0	1
HS-ESS2-2: Earth Materials and Systems	0	1	0	1	0	1
HS-ESS2-3: Earth Materials and Systems	0	1	0	1	0	1
HS-ESS2-4: Weather and Climate	0	1	0	1	0	1
HS-ESS2-5: The Roles of Water in Earth's Surface Processes	0	1	0	1	0	1
HS-ESS2-6: Weather and Climate	0	1	0	1	0	1
HS-ESS2-7: Weather and Climate	0	1	0	1	0	1
DCI – Earth and Human Activity	**0**	**1**	**0**	**2**	**0**	**3**
HS-ESS3-1: Natural Resources	0	1	0	1	0	1
HS-ESS3-2: Natural Resources*	0	1	0	1	0	1
HS-ESS3-3: Human Impacts on Earth Systems	0	1	0	1	0	1
HS-ESS3-4: Human Impacts on Earth Systems*	0	1	0	1	0	1
HS-ESS3-5: Global Climate Change	0	1	0	1	0	1
HS-ESS3-6: Global Climate Change*	0	1	0	1	0	1
PE Total = 67	**6**	**6**	**12**	**12**	**18**	**18**

*Note: These PEs have an engineering component.

Main characteristics of the blueprint were that any performance expectation (PE) could be tested only once (indicated by the values of 0 and 1 for the Min and Max values of the individual PEs in Table 7 through Table 9); no more than one item cluster or two stand-alone items could be sampled from the same disciplinary core idea (DCI); and no more than three items in total could be sampled from the same DCI (as indicated by the Min and Max values in the rows representing DCIs).

While tests are not timed, the New Hampshire Department of Education (NHDOE) published estimated testing times for the NH SAS science assessment. Percentile 85 of testing times are presented in Table 10.

Table 10. NH SAS Science Percentile 85 Testing Times by Grade

Subject	Grade	85th Percentile Testing
Science	5	109.63
	8	100.78
	11	85.31

4.3 Test Construction

During fall 2018, AIR psychometricians and content experts worked with NHDOE content specialists and leadership to build item pools for the spring 2019 administration. The New

Hampshire Statewide Assessment System (NH SAS) for Science test construction utilizes a structured test construction plan, explicit blueprints, and active collaborative participation from all parties.

The 2019 NH SAS test item pools were built by AIR test developers to exactly match items to the detailed test blueprints. Operational items were selected from the AIRCore item bank to fulfill the blueprint for that grade. For 2019, the NH SAS science item pool was identical to the AIRCore pool (as described in Table 2 through Table 6), except for two elementary school items. These two items were excluded because the 2018 parameters were no longer valid due to a change in the way students could interact with the item (these items had an omit rate of 4% or higher in 2018 for their last interaction, and in 2019, a response to every interaction was required for all items in order to move to the next item).

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

5. Simulation Summary Report

This section describes the results of simulated test administrations used to configure and evaluate the adequacy of the item selection algorithm used to administer the NH SAS 2018–2019 assessments for science grades 5, 8, and 11. Simulations were carried out to configure the settings of the algorithm and to evaluate whether individual tests adhered to the test blueprint.

Psychometricians reviewed the simulation results for the following key diagnostics:

- Match-to-test blueprint: Determines that the tests have the correct number of test items overall and the appropriate proportion by content categories at each level of the content hierarchy, as specified in the test blueprints for every science grade.
- Item exposure rate: Evaluates the utility of item pools and identifies overexposed and underexposed items.

These diagnostics are interrelated. For example, if the test pool for a particular content level 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 the adjustment of setting parameters 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 a number of factors that may influence simulation results for a linear-on-the-fly (LOFT) test administration. These include:

1. 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 cluster per DCI, but the pool has no item for some DCIs, it may be impossible to meet this constraint.
2. 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 only associated with a specific DCI, it may be difficult to meet both the desired distribution of content and the desired distribution of item type.
3. 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 almost all students (99.91%). 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 form, 100% of the items appear on 100% of the test administrations because every test taker sees the same items. In an adaptive test or a LOFT with a sufficiently large item pool, we would expect that most of the items would appear on only a relatively small percentage of the test administrations.

When this condition holds, it suggests that test administrations between students are more or less unique. Therefore, we calculated the item exposure rate for each item across by dividing the total number of test administrations in which an item appears by the total number of tests administered. Then we report the distribution of the item exposure rate (r) in six 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, we would expect the largest proportion of items to 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 11 presents the percentage of items that fall into each exposure bin for all grades. Most test items (90% or more) were administered in 5–60% of the test administrations. No item had an exposure rate less than 5%, which means that there was a sufficiently large sample for each item for item calibration. A few items had an exposure rate of 100% due to the limitation of the current

pool for some content categories. Only those items were available to satisfy the blueprint constraints. In addition, field-test items were administered in the embedded field-test segment. - The item exposure rate for field-test items ranged from 10%–25% for all grades.

Table 11. Item Exposure Rates by Grade: Percentage of Items by Exposure Rate, Across All English Simulation Sessions

Grade	Total Items	[0,0] %	[0,1] %	[1,5] %	[5,20] %	[20,40] %	[40,60] %	[60,80] %	[80,100] %
					Science				
5	59	-	-	-	40.68	37.29	13.56	3.39	5.08
8	51	-	-	-	19.61	50.98	19.61	3.92	5.88
11	55	-	-	-	30.91	40	23.64	1.82	3.64

5.3 Results of Simulated Test Administrations: Spanish

This section presents the simulation results for the Spanish tests. The Spanish item pool is smaller than the AIRCore item pool because only a subset of AIRCore items has a Spanish translation available. Table 12 presents the numbers of items available for the Spanish tests.

Table 12. Spring 2019 Spanish Operational Item Pool

Grade Band	Item Type	Total Number of Items
Elementary School	Cluster	8
	Stand-alone	19
Middle School	Cluster	6
	Stand-alone	19
High School	Cluster	6
	Stand-alone	19
Total		**77**

Simulations were evaluated for all content areas using 500 simulated cases per grade.

5.3.1 Summary of Blueprint Match

There was no blueprint violation at the discipline level for all three grades. However, due to the limitation of the Spanish item pool, blueprint violations were observed in grades 8 and 11 for content levels below the discipline level. For both grades, students always received two clusters from the same DCI (ESS2), but the blueprint required no more than one cluster from each DCI. The reason is that there was no cluster available in two DCIs (ESS1 and ESS3). In grade 8, ESS1 and ESS3 also have a limited number of standalone items. Among the 500 simulated cases, 5.6%

of students received two standalone items from ESS2, therefore students received four items from the same DCI—one more item than the blueprint requirement.

5.3.2 Item Exposure

Table 11 presents the percentage of items that fall into each exposure bin for all grades. All test items were administered in more than 20% of the test administrations. Some items had an exposure rate of 100% due to the limited Spanish item pool. Only those items were available to satisfy the blueprint constraints.

Table 13 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] %
					Science				
5	27	-	-	-	-	18.52	29.63	14.81	37.04
8	25	-	-	-	-	4	32	28	36
11	25	-	-	-	-	16	16	24	44

6. Operational Test Administration Summary Report

This section presents the blueprint match reports and item exposure rates for the spring 2019 operational test administrations.

6.1 Blueprint Match

Table 15 presents the percentages of the spring 2019 tests that aligned with the blueprint requirement. Across all grades, every English test, except for one student in grade 11, met the blueprint specifications with a 100% match at all content levels. That one student received two items (one cluster and one standalone) from the same performance expectation (PE), while the blueprint requires no more than one item from the same PE. This type of violation did not happen during simulations. The reason it happened in operational test administrations is because the student took the same test twice. The first test was reset, and the student took the same test again on the following day. For the second test, the items the student saw at the first trial were removed from the pool so that the student would not see the same item twice in two consecutive days. Therefore, the pool became shallower for this student. When it came to the last item, the algorithm had no item left to satisfy the blueprint requirement and picked the one that caused the violation.

For Spanish tests, blueprint violations were observed for some content levels in grade 8 due to the limited item pool, similar to the findings during simulations. No blueprint violation was observed for grade 5. No student took the grade 11 Spanish test in spring 2019.

Table 14 Spring 2019 Blueprint Match for Test Delivered, Science

Grade	Content Level	MinItems	MaxItems	% of Cases Meeting BP	% of Cases Violating BP			
					1	2	-1	-2
English								
5	Discipline	6	6	100	-	-	-	-
	Discipline – Cluster	2	2	100	-	-	-	-
	Discipline – Standalone	4	4	100	-	-	-	-
	DCI	0	3	100	-	-	-	-
	DCI – Cluster	0	1	100	-	-	-	-
	DCI – Standalone	0	2	100	-	-	-	-
	PE	0	1	100	-	-	-	-
8	Discipline	6	6	100	-	-	-	-
	Discipline – Cluster	2	2	100	-	-	-	-
	Discipline – Standalone	4	4	100	-	-	-	-
	DCI	0	3	100	-	-	-	-
	DCI – Cluster	0	1	100	-	-	-	-
	DCI – Standalone	0	2	100	-	-	-	-
	PE	0	1	100	-	-	-	-
11	Discipline	6	6	100	-	-	-	-
	Discipline – Cluster	2	2	100	-	-	-	-
	Discipline – Standalone	4	4	100	-	-	-	-
	DCI	0	3	100	-	-	-	-
	DCI – Cluster	0	1	100	-	-	-	-
	DCI – Standalone	0	2	100	-	-	-	-
	PE	0	1	99.99	0.001	-	-	-
Spanish								
5	Discipline	6	6	100	-	-	-	-
	Discipline – Cluster	2	2	100	-	-	-	-
	Discipline – Standalone	4	4	100	-	-	-	-
	DCI	0	3	100	-	-	-	-
	DCI – Cluster	0	1	100	-	-	-	-
	DCI – Standalone	0	2	100	-	-	-	-
	PE	0	1	100	-	-	-	-
8	Discipline	6	6	100	-	-	-	-
	Discipline – Cluster	2	2	100	-	-	-	-
	Discipline – Standalone	4	4	100	-	-	-	-
	DCI	0	3	94.74	5.26	-	-	-
	DCI – Cluster	0	1	-	100	-	-	-
	DCI – Standalone	0	2	100	-	-	-	-
	PE	0	1	100	-	-	-	-

6.2 Item Exposure

Table 16 presents the item exposure rates of the spring 2019 test administration. As is consistent with the simulation results described in Section 5.2.2 and 5.3.2, most test items (90% or more) were administered in 5–60% of the English test administrations. A few items had an exposure rate of 100% due to the limitation of the item pool in some content areas. The item exposure rate for field-test items ranged from 10%–25% for all grades. For Spanish tests, more items had an

exposure rate of 100% compared to the English tests due to a smaller item pool. Also, the operational exposure rates were slightly different from the simulation results because of small population sizes in both grades. In spring 2019, only 11 students in grade 5 and 19 students in grade 8 took the Spanish tests.

Table 15 Item Exposure Rates by Grade: Percentage of Items by Exposure Rate, Across All Spring 2019 Test Administrations

Grade	Total Items	[0,0] %	[0,1] %	[1,5] %	[5,20] %	[20,40] %	[40,60] %	[60,80] %	[80,100] %
					English				
5	59	-	-	-	40.37	35.59	13.56	1.69	6.78
8	51	-	-	-	21.57	49.02	19.61	3.92	5.88
11	55	-	-	-	30.91	40	23.64	1.82	3.64
					Spanish				
5	27	-	-	-	7.41	7.41	33.33	7.41	44.44
8	25	-	-	-	4	4	24	28	40

REFERENCES

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014*). Standards for educational and psychological testing*.

Calisir, F., & Gurel, Z. (2003). Influence of text structure and prior knowledge of the learner on reading comprehension, browsing and perceived control. *Computers in Human Behavior, 19*(2), 135–145.

Fisher, D., Frey, N., & Lapp, D. (2012). *Text complexity: Raising rigor in reading*. Newark, DE: International Reading Association.

Freebody, P., & Anderson, R. C. (1983). Effects on Text Comprehension of Differing Proportions and Locations of Difficult Vocabulary. *Journal of Reading Behavior, 15*(3), 19–39.

Gillioz, C., Gygax, P., & Tapiero, I. (2012). Individual differences and emotional inferences during reading comprehension. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 66*(4), 239–250.

Kucer, S. B. (2010). Going beyond the author: What retellings tell us about comprehending narrative and expository texts. *Literacy,* 45(2), 62–69.

Long, D. L., & De Ley, L. (2000). Implicit causality and discourse focus: The interaction of text and reader characteristics in pronoun resolution. *Journal of Memory and Language, 42*(4), 545–570.

McConaughy, S. (1985). Good and Poor Readers' Comprehension of Story Structure across Different Input and Output Modalities. *Reading Research Quarterly, 20*(2), 219–232. doi:10.2307/747757.

Rapp, D. N., & Mensink, M. C. (2011). Focusing effects from online and offline reading tasks. In M. T. McCrudden, J. P. Magliano, & G. Schraw (Eds.), *Text relevance and learning from text* (pp. 141–164). Charlotte, NC, US: IAP Information Age Publishing.

Rich, S. S., & Taylor, H. A. (2000). Not all narrative shifts function equally. *Memory & Cognition, 28*(7), 1257–1266.

Riding, R. J., & Taylor, E. M. (1976). Imagery performance and prose comprehension in seven-year-old children. *Educational Studies, 2*(1), 21–2.

Rommers, J., Dijkstra, T., & Bastiaansen, M. (2013). Context-dependent semantic processing in the human brain: Evidence from idiom comprehension. *Journal of Cognitive Neuroscience, 25*(5), 762–776.

Sadoski, M., Goetz, E. T., & Fritz, J. B. (1993). A causal model of sentence recall: Effects of familiarity, concreteness, comprehensibility, and interestingness. *Journal of Reading Behavior, 25*(1), 5–16.

Schmitt, N., Jiang, X., & Grabe, W. (2011). The Percentage of Words Known in a Text and Reading Comprehension. *Modern Language Journal, 95*(1), 26–43.

Sparks, J. R., & Rapp, D. N. (2011). Readers reliance on source credibility in the service of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(1), 230–247.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved February 15, 2012, from http://www.cehd.umn.edu/NCEO/onlinepubs/Synthesis44.html.

Webb, N. L. (2002). *Alignment study in language arts, mathematics, science, and social studies of state standards and assessments for four states*. Washington, DC: Council of Chief State School Officers.