The Rhode Island Next Generation Science Assessment

2022–2023

Volume 1: Annual Technical Report

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

1. INTRODUCTION

The Rhode Island Next Generation Science Assessment (RI NGSA) measures the achievement of science standards by students in grades 5, 8, and 11. The *2022–2023 RI NGSA Technical Report* is provided to document and make transparent all methods used in item development, test construction, psychometrics, standard setting, test administration, and score reporting, including summaries of student results and evidence and support for intended uses and interpretations of the test scores. The technical report comprises six separate, self-contained volumes:

1) **Annual Technical Report.** This volume is updated each year and provides a global overview of the tests administered to students each year.
2) **Test Development.** This volume summarizes the procedures used to construct test forms and provides summaries of the item bank and development process.
3) **Setting Performance Standards.** This volume documents the methods and results of the RI NGSA standard-setting process.
4) **Evidence of Reliability and Validity.** This volume provides technical summaries of the test quality and special studies to support the intended uses and interpretations of the test scores.
5) **Test Administration.** This volume describes the methods used to administer all tests, enforce security protocols, and ensure availability of modifications or accommodations.
6) **Score Interpretation Guide.** This volume describes the score types reported and details the inferences that can appropriately be drawn from each reported score.

The Rhode Island Department of Education (RIDE) communicates the quality of the RI NGSA by making these technical reports accessible to the public on the state's website.

1.1 BACKGROUND AND HISTORICAL CONTEXT OF TESTS

Rhode Island adopted the Next Generation Science Standards (NGSS) in 2013. The RIDE and the assessment vendor, Cambium Assessment, Inc. (CAI), developed and administered new online assessments to measure students' achievement in relation to the NGSS. The new RI NGSA was developed to measure the science knowledge and skills of Rhode Island students in grades 5, 8, and 11. In 2017–2018, the assessments were administered as an independent field test in Rhode Island. The RI NGSA was administered operationally for the first time in 2018–2019. The RIDE cancelled the spring 2020 administration of the RI NGSA due to statewide school closures that followed the onset of the COVID-19 pandemic. Starting in spring 2021, the RIDE and CAI resumed administration of the RI NGSA

The RIDE provides an overview of the RI NGSA at https://www.ride.ri.gov/InstructionAssessment/Assessment/NGSAAssessment.aspx and at https://ri.portal.cambiumast.com/index.html.

Information about the NGSS is available at: www.nextgenscience.org.

1.2 Purpose and Intended Uses of the Multi-State Science Assessment

The RI NGSA is a standard-referenced test that uses principles of evidence-centered design to yield overall and discipline-level test scores at the student level and other levels of aggregation that reflect student achievement. The three-dimensional science standards (i.e., the NGSS) establish a set of knowledge and skills that all students need to be prepared for a wide range of high-quality post-secondary opportunities, including higher education and entering the workplace.

The three-dimensional NGSS reflects the latest research and advances in modern science education and differ from previous science standards in multiple ways. First, rather than describing general knowledge and skills that students should know and be able to do, they describe specific performances that demonstrate what students know and can do. The NGSS refer to such performed knowledge and skills as *performance expectations* (PEs). Second, the NGSS are intentionally multidimensional. Each performance expectation incorporates all three dimensions from the NGSS Framework: a science or engineering practice, a disciplinary core idea, and a crosscutting concept. Third, whereas traditional standards do not consider other subject areas, the NGSS connect to standards for other subjects, such as the Common Core State Standards (CCSS) for mathematics and English language arts (ELA). Another unique feature of the NGSS is the assumption that students should learn all science disciplines rather than a select few, as is traditionally the expectation in many high schools, where students may elect, for example, to take biology and chemistry but not physics or astronomy.

The RI NGSA supports instruction and student learning by providing educators and parents with valuable feedback that can be used to remediate or enrich instruction. An array of reporting metrics is provided so that achievement can be evaluated at the student level and at aggregated levels and so that improvement over time can be monitored at both the student and group levels.

The RI NGSA draws items from an item bank comprised of Independent College and Career Readiness (ICCR) items and a pool of items owned by several other states and one U.S. territory that abide by a Memorandum of Understanding (MOU) to share content, leadership, and new ideas and methods. Full members of the MOU in 2023 were Connecticut, Hawaii, Idaho, Montana, Oregon, Rhode Island, Utah, West Virginia, and Wyoming. New Hampshire, North Dakota, South Dakota, and U.S. Virgin Islands observed and participated in some activities. CAI played a supporting and coordinating role, working with the RIDE to ensure that the items in the tests constructed for all grades uniquely measured students' mastery of the three-dimensional NGSS.

Table 1 outlines the required uses and citations for the RI NGSA based on §18-2E-5-(d)(3) and the federal *Every Student Succeeds Act* (ESSA) plan. The RI NGSA fulfills all the requirements described in Table 1.

Table 1. Required Uses and Citations for the RI NGSA

Required Use	Required Use Citation
Indicator of academic achievement and progress	ESSA Plan Section 1 A. i; ESSA Plan Section 4 4.1 A ; 16-97.111 (4)(i) and (iii)
Test administration frequency and grade levels	15.1-21-08.1; 16-97.1-1(8)(C)
Compilation of test scores	15.1-21-09

Required Use	Required Use Citation
Publication of test scores	15.1-21-10; 16-97.1-1; ESSA section 1111(b)(3)(c)(xii))
Requirement for alignment of test to academic content standards	15.1-21-11; 16-97.1-1 (a)(1)(i)

1.3 Participants in the Development and Analysis of the Multi-State Science Assessment

The Rhode Island Department of Education manages the RI NGSA with the assistance of several stakeholders, including Rhode Island educators, a Technical Advisory Committee (TAC), and vendors. RIDE fulfills the diverse requirements of implementing Rhode Island's statewide assessment while adhering to the guidelines established in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). To comply with the *Standards*, scale development, scoring, linking, and evaluation of differential item functioning are addressed in the current volume; item development, test design, and test blueprints are documented in Volume 2, Test Development; development of cut scores is summarized in Volume 3, Setting Performance Standards; evidence for validity and reliability/precision was collected and is reported in Volume 4, Evidence of Reliability and Validity; information on testing windows, test options, accommodations, training of test coordinators and administrators, and test security are provided in Volume 5, Test Administration; supporting documentation for tests, score uses and interpretation are included in Volume 6, Score Interpretation.

1.3.1 Rhode Island Department of Education

The Office of Instruction, Assessment, and Curriculum in the RIDE manage test development, administration, scoring, and reporting of results for their respective statewide comprehensive assessment programs, including coordinating with other RIDE offices, Rhode Island public schools, and vendors.

1.3.2 Rhode Island Educators

Rhode Island educators are involved in most aspects of the conceptualization and development of the RI NGSA. Educators participate in clarifying how the standards are assessed, designing test, and reviewing test items and passages.

1.3.3 Technical Advisory Committee

RIDE convenes an advisory committee panel several times each year to discuss psychometric, test development, administrative, and policy issues of relevance to current and future Rhode Island and Vermont assessments. This committee is comprised of several nationally recognized assessment experts and highly experienced practitioners from several school districts.

1.3.4 Cambium Assessment, Inc.

CAI (formerly the American Institutes for Research [AIR]) is the vendor that was selected through the state-mandated competitive procurement process. CAI is responsible for developing test content, building test forms, conducting psychometric analyses, administering and scoring test forms, and reporting test results for the RI NGSA. Additionally, CAI is responsible for developing and maintaining the ICCR item bank.

1.3.5 Caveon Test Security

Caveon Test Security monitored web pages and social media during the spring 2023 test administration to ensure that no secure testing materials such as items and prompts were leaked. Details of Caveon Test Security are described in Appendix 1-A, Caveon Test Security Overview.

1.4 Available Test Formats and Special Versions

The RI NGSA is administered online using a linear-on-the-fly (LOFT) test design. Science items focus on a scientific phenomenon and can consist of shorter (stand-alone) items or items with several parts (item clusters) that require the student to interact with the item in various ways. In Rhode Island, the assessment was administered as an independent field test in spring 2018 and as an operational test in spring 2019. Starting in 2021 and thereafter, additional items were field-tested to build out the item bank.

Students unable to participate in the online test administration have the option to use print-on-demand—a feature that provides the same items administered to students online in a paper format. Spanish versions of the RI NGSA (developed to meet the same content standards as the English versions) are available for all tested grades. Students participating in the computer-based RI NGSA can use standard online testing features in the Test Delivery System (TDS), which include a selection of font color and size and the ability to zoom in and zoom out or highlight text. In addition to the resources available to all students, options are available to accommodate students with an Individualized Education Program (IEP) or Section 504 Plan. These options include braille, American Sign Language (ASL), closed captioning, and large print. Students with disabilities have the option to take the RI NGSA with or without accommodations or to take an alternate assessment. For additional information about the testing feature and testing accommodations, refer to Volume 5, Test Administration, of this technical report.

1.5 Student Participation

All students in Rhode Island public schools are required to participate in the statewide assessments. The RI NGSA is administered in the spring. Table 2 shows the number of students who were tested (number tested) and the number of students whose scores were included in the analyses for this technical report (number reported). **Error! Reference source not found.**Table 3 shows the demographic characteristics of the student population, in counts and in percentages, in the spring administration of the 2022–2023 assessments. The subgroups reported here are gender, ethnicity, limited English proficiency (LEP), economic disadvantage, and eligibility for special education.

Table 2. Total Number of Students Participating in the RI NGSA, Spring 2023

Grade	Number Tested	Number Reported
5	9,816	9,813
8	10,093	10,089
11	9,225	9,212

Table 3. Distribution of Demographic Characteristics of Student Population

Group	Grade 5		Grade 8		Grade 11	
	N	*%*	*N*	*%*	*N*	*%*
All Students	9,813	100.00	10,089	100.00	9,212	100.00
Female	4,785	48.76	4,858	48.15	4,505	48.90
Male	5,023	51.19	5,221	51.75	4,695	50.97
Unspecified	5	0.05	10	0.10	12	0.13
African American	851	8.67	917	9.09	804	8.73
American Indian/Native Alaskan	79	0.81	70	0.69	59	0.64
Asian	337	3.43	338	3.35	292	3.17
Hispanic	2,935	29.91	2,962	29.36	2,548	27.66
Multi-Racial	508	5.18	504	5.00	343	3.72
Pacific Islander	13	0.13	11	0.11	14	0.15
White	5,090	51.87	5,287	52.40	5,152	55.93
Limited English Proficiency	1,743	17.76	1,754	17.39	937	10.17
Special Education	1,504	15.33	1,539	15.25	1,064	11.55
Economically Disadvantaged	4,715	48.05	4,403	43.64	3,375	36.64

2. Operational Practices and Procedures

This section outlines key elements of the operational administration, including testing window, test administrators, online testing environment, and simulations. Accessibility supports including universal tools, designated supports, and accommodations are also discussed, followed by the number of test sessions with allowed designated supports and accommodations for each test.

2.1 Test Window

Table 4 shows the testing window for the 2022–2023 Rhode Island Next Generation Science Assessment (RI NGSA).

Table 4. RI NGSA Testing Windows by State

Grades	Testing Window
5, 8, 11	April 24, 2023–May 26, 2023

2.2 Test Administrators

The key personnel involved with the Rhode Island test administration included the district test coordinators (DTCs), school test coordinators (STCs), and test administrators (TAs) who proctored the test. A *Test Administration Manual* (TAM) (available at https://ri.portal.cambiumast.com/resources was provided so that personnel involved with the statewide assessment administrations could maintain both standardized administration conditions and test security.

2.3 Testing Environment

The Cambium Assessment, Inc. (CAI) Secure Browser was required to access the online Rhode Island and Vermont tests. The online browser provided a secure environment for student testing by disabling the hot keys, copy, and screen-capture capabilities and preventing access to the desktop (Internet, email, and other files or programs installed on school machines). During the online assessment, students could pause a test, review previously answered questions and modify their response if the test had not been paused for more than 20 minutes. Students do not have a required time limit for each test session, but for planning purposes, schools were given approximate time estimates for how long most students would need to complete each test. For additional information about the test administration, refer to Volume 5, Test Administration, of this technical report.

2.4 Simulations

CAI employs a simulation approach to all RI NGSA tests. The test is delivered using the same item selection algorithm that CAI uses to deliver adaptive tests, except that only the blueprint of a test is considered during the item-selection process. Simulations were conducted to configure the item selection algorithm settings, to evaluate whether individual tests adhered to the test blueprint, monitor item exposure rates, and to verify the scores produced by CAI's scoring engine. Simulations were also conducted on fixed-form tests in order to quality check the scores. The simulation approaches and results are discussed in Volume 2, Test Development.

2.5 Universal Tools, Designated Supports, and Accommodations

Accessibility supports are available to students when needed to remove barriers during testing while maintaining the constructs that are measured by the RI NGSA. The accessibility supports discussed in this technical report include embedded (digitally provided) and non-embedded (non-digitally or locally provided) universal features that are available to all students as they access instructional or assessment content; designated features that are available to those students for whom the need has been identified by an informed educator or team of educators; and accommodations that are generally available for students for whom there is documentation on an Individualized Education Program (IEP) or Section 504 Plan. For English learners (ELs), Spanish language versions of the RI NGSA were available.

All educators making decisions about designated supports were trained on the process and understand the range of designated supports available.

Accommodations are changes in procedures or materials that ensure equitable access to instructional and assessment content and generate valid assessment results for students who need them. Embedded accommodations (e.g., text-to-speech [TTS]) are provided digitally through instructional or assessment technology, and non-embedded designated features (e.g., scribe) are non-digital. State-approved accommodations do not compromise the learning expectations, constructs, or grade-level standards. Such accommodations help students with a documented need generate valid assessment outcomes so that they can fully demonstrate what they know and can do. From the psychometric point of view, the purpose of providing accommodations is to "increase the validity of inferences about students with special needs by offsetting specific disability-related, construct-irrelevant impediments to performance" (Koretz & Hamilton, 2006, p. 562).

The TAs and STCs in Rhode Island were responsible for ensuring that arrangements for accommodations were made before the test administration dates. Some of the available accommodation options for eligible students are listed on the following pages.

Additional information about universal features, designated supports, and accommodations can be found in Volume 5, Test Administration, of this technical report.

Table 5 and Table 6 list the number of test sessions in which a student was provided with each designated support or accommodation during the spring 2023 test administration sessions in Rhode Island.

Table 5. Number of Testing Sessions with Accessibility Features

Accessibility Features	Grade		
	5	8	11
Embedded			
Color Choices	1	-	-
Masking	49	11	-
Mouse Pointer	-	2	-
Print Size	2	5	-
Text-to-Speech: Stimuli and Items	1,342	588	149
Non-Embedded			
Magnification	1	1	1

Table 6. Number of Testing Sessions with Allowed Accommodations

Accommodations	Grade		
	5	8	11
Non-Embedded			

Accommodations	Grade		
	5	8	11
AT/ACC Devices (Requires Permissive Mode)	-	-	-
Speech-to-Text (Requires Permissive Mode)	26	35	-

3. ITEM BANK AND TEST DESIGN

3.1 SHARED SCIENCE ASSESSMENT ITEM BANK

Cambium Assessment, Inc. (CAI) works with a group of states and one U.S. territory to develop science assessments to measure achievement of the Next Generation Science Standards (NGSS) and other standards influenced by the same science framework. Many of these states have signed a Memorandum of Understanding (MOU) to share item specifications and items. CAI has coordinated this group of states and one U.S. territory and holds contracts to develop and deliver the items for most of them.

CAI also built the Independent College and Career Readiness (ICCR) science item pool in partnership with these states and one U.S. territory. These CAI-owned items make up a substantial part of the item bank and are shared with partner states and one U.S. territory. Rhode Island signed the MOU, and therefore, the item pool available for Rhode Island includes items from the following three sources:

1. Items owned by Rhode Island
2. Items shared by other states/territory within the MOU collaboration
3. Items shared from the ICCR item bank

In 2023, the Shared Science Assessment Item Bank was used for operational tests in 12 states and one U.S. territory, including Rhode Island.

The goals, uses, and claims that the Shared Science Assessment Item Bank and resulting tests are designed to support were identified in a collaborative meeting on August 22–23, 2016, to facilitate the transition from a framework for three-dimensional science standards, specifically the NGSS, to statewide summative assessments for science. CAI invited content and assessment leaders from 10 states and four nationally recognized experts who helped co-author the NGSS. Two nationally recognized psychometricians also participated.

In 2017, cognitive lab studies were conducted to evaluate and refine the process of developing item clusters aligned to the three-dimensional science standards. The results of the cognitive lab studies confirmed the feasibility of the approach (refer to Volume 4, Appendix 4-D, Science Clusters Cognitive Lab Report, of this technical report).

A second set of cognitive lab studies was conducted in 2018 and 2019 to determine whether students using braille could understand the task demands of selected accommodated three-dimensional science-aligned item clusters. They also evaluated whether these students could

navigate the interactive features of these item clusters in a manner that allowed them to fully display their knowledge and skills relative to the constructs of interest. In general, both the students who relied entirely on braille and/or Job Access With Speech (JAWS) and those who had some vision and were able to read the screen with magnification were able to find the information they needed to respond to the questions, navigate the various response formats, and finish within a reasonable amount of time (refer to Volume 4, Appendix 4-E, Braille Cognitive Lab Report, of this technical report).

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 and future administrations. In 2019, 2021, and 2022, the numbers of items that were field tested were 347, 545, and 471, while the numbers of items that were accepted and made available for future operational use were 268, 458, and 403, respectively. In 2023, 348 item clusters and stand-alone items were field tested, of which 288 were accepted and made available for operational use in future administrations. All these items follow the same specifications, test development processes, and review processes, summarized below:

- CAI staff and participating states collaborated to develop item specifications, which are documents designed to guide item writers as they craft test questions and stakeholders while they review items. The item specifications were generally accompanied by sample items meeting those specifications. All specifications and sample items were reviewed by state content experts and committees of educators in at least one state.
- The specifications helped test developers create item clusters and stand-alone items that covered a range of difficulty, furthering the goal of measuring the full range of performance found in the population, but remaining at grade level. All item writers were trained in the principles of universal design, the appropriate use of item interactions, and the science item specifications.
- Items were reviewed by science experts in at least one state.
- Every item was reviewed by a content advisory committee (comprised of state educators) in at least one state or in a cross-state educator review process.
- Every item was reviewed by a committee of educators charged with evaluating language accessibility, bias, and sensitivity in at least one state or a cross-state educator review.
- Every item was field tested, all scoring protocols (i.e., rubrics) were validated using the field-test data, and items with questionable data were reviewed again by committees of educators.

A detailed description of the Shared Science Assessment Item Bank development process is included in Volume 2, Test Development.

3.2 FIELD-TESTING

All items that were part of the operational pool of the Shared Science Assessment Item Bank were field tested in prior years, which was documented in Appendix 1-B, Shared Science Assessment Item Bank: Field-Testing. Field-testing for the current administration is described in this section.

3.2.1 2023 Field Test

In 2023, items were field tested in 12 states and one U.S. territory. In all 12 states and one U.S. territory (Connecticut, Hawaii, Idaho, Montana, New Hampshire, North Dakota, Oregon, Rhode Island, South Dakota, U.S. Virgin Islands, Utah, West Virginia, and Wyoming), field-test items were administered as unscored items embedded among operational items. In total, 159 item clusters and 189 stand-alone items were administered as field-test items in the elementary, middle, and high school grade bands.. Table 7 presents the number of field-test item clusters and stand-alone items administered in each grade band for each state. The numbers in parentheses in the column representing RI show the number of field-test items owned by Rhode Island.

Table 7. Number of Field-Test Items Administered, Spring 2023

Grade Band and Item Type	CT	HI	ID	MT	ND	NH	OR	RI	SD	USVI	UT	WV	WY	Total
Elementary School	**41**	**19**	**20**	**12**	**12**	**11**	**28**	**35 (6)**	**10**	**1**	**32**	**25**	**7**	**126**
Cluster	16	7	12	4	4	7	12	12 (1)	4	1	32	7	3	60
Stand-Alone	25	12	8	8	8	4	16	23 (5)	6	0	0	18	4	66
Middle School	**36**	**24**	**21**	**9**	**11**	**11**	**41**	**29 (8)**	**7**	**1**	**49**	**28**	**7**	**136**
Cluster	6	8	5	5	4	7	10	9 (3)	5	1	49	8	3	59
Stand-Alone	30	16	16	4	7	4	31	20 (5)	2	0	0	20	4	77
High School	**37**	**8**	**20**	**0**	**12**	**12**	**31**	**36 (15)**	**6**	**1**	**0**	**0**	**10**	**86**
Cluster	21	4	8	0	3	1	25	12 (9)	4	1	0	0	2	40
Stand-Alone	16	4	12	0	9	11	6	24 (6)	2	0	0	0	8	46
Total	**114**	**51**	**61**	**21**	**35**	**34**	**100**	**100 (29)**	**23**	**3**	**81**	**53**	**24**	**348**

Note. RI-owned items are indicated in the parentheses.

Two of the states (New Hampshire and Rhode Island) opted for a test in which operational items were grouped by science discipline. For these two states, the field-test items were presented together in a fourth group of items. The sequence of the four sets of items (corresponding to the three disciplines and a set of field-test items) was randomized across students. Ten other states and one US territory (Connecticut, Hawaii, Idaho, Montana, North Dakota, Oregon, South Dakota, Utah, US Virgin Islands, West Virginia, and Wyoming) opted for a test design in which the items were not grouped by discipline. In these 10 states and one U.S. territory, field-test items were administered at random positions throughout the test. A student received either a field-test item cluster or a set of four field-test stand-alone items. The test design for the RI NGSA Assessment is discussed in Section **Error! Reference source not found.**, **Error! Reference source not found.**.

A minimum sample size of 1,500 students per field-test item was targeted for any given state or territory. Most items were administered in two states or territory. Approximately 85.5% of the items met or exceeded the target sample size of 1,500 in at least one state, and 100% of the items had a sample size of at least 1,350 (90% of the target) in at least one state.

Table 8 to Table 10 present the number of item clusters and stand-alone items that were shared between the field-test pools of any two states or territory. The numbers below the shaded cells represent the number of common field-test items between any two states, and the numbers above the shaded cells represent the number of common field-test items that survived rubric validation and were included in the calibration. In each of the shaded cells, the number outside the parentheses represents the number of unique field-test items administered only in the given state or territory, and the number in the parentheses represents the number of unique and/or common items that were calibrated with only the data from that state. Table 8 presents the results for elementary schools, Table 9 presents the results for middle schools, and Table 10 presents the results for high schools. The numbers of field-test items administered are slightly different from the numbers of field-test items at calibration because some items were rejected during rubric validation.

Table 8. Number of Common Elementary School Field-Test Items Administered and Calibrated, Spring 2023

	State	CT	HI	ID	MT	NH	ND	OR	RI	SD	USVI	UT	WV	WY
Item Clusters	CT	0 (0)	0	2	2	2	0	5	4	0	0	1	0	0
	HI	0	0 (0)	1	0	0	0	0	0	0	0	6	0	0
	ID	2	1	0 (0)	0	0	0	0	2	0	0	6	0	0
	MT	2	0	0	0 (0)	0	0	0	0	0	0	2	0	0
	NH	2	0	0	0	0 (0)	0	0	0	0	0	2	3	0
	ND	0	0	0	0	0	0 (0)	3	0	0	0	1	0	0
	OR	5	0	0	0	0	3	0 (0)	0	0	0	4	0	0
	RI	4	0	2	0	0	0	0	0 (0)	0	0	4	2	0
	SD	0	0	0	0	0	0	0	0	0 (0)	1	2	2	0
	USVI	0	0	0	0	0	0	0	0	1	0 (0)	1	0	0
	UT	1	6	7	2	2	1	4	4	2	1	0 (0)	0	3
	WV	0	0	0	0	3	0	0	2	2	0	0	0 (0)	0
	WY	0	0	0	0	0	0	0	0	0	0	3	0	0 (0)
Stand-Alone Items	CT	0 (0)	0	6	3	0	0	2	9	0	0	0	3	0
	HI	0	0 (0)	0	0	0	0	8	0	0	0	0	0	4
	ID	6	0	0 (0)	0	0	0	0	2	0	0	0	0	0
	MT	4	0	0	0 (0)	0	0	0	4	0	0	0	0	0
	NH	0	0	0	0	0 (0)	0	0	0	0	0	0	4	0
	ND	0	0	0	0	0	0 (0)	6	2	0	0	0	0	0
	OR	2	8	0	0	0	6	0 (0)	0	0	0	0	0	0
	RI	10	0	2	4	0	2	0	0 (0)	0	0	0	5	0
	SD	0	0	0	0	0	0	0	0	0 (0)	0	0	6	0
	USVI	0	0	0	0	0	0	0	0	0	0 (0)	0	0	0
	UT	0	0	0	0	0	0	0	0	0	0	0 (0)	0	0
	WV	3	0	0	0	4	0	0	5	6	0	0	0 (0)	0

	State	CT	HI	ID	MT	NH	ND	OR	RI	SD	USVI	UT	WV	WY
	WY	0	4	0	0	0	0	0	0	0	0	0	0	0 (0)
Total	**CT**	0 (0)	0	8	5	2	0	7	13	0	0	1	3	0
	HI	0	0 (0)	1	0	0	0	8	0	0	0	6	0	4
	ID	8	1	0 (0)	0	0	0	0	4	0	0	6	0	0
	MT	6	0	0	0 (0)	0	0	0	4	0	0	2	0	0
	NH	2	0	0	0	0 (0)	0	0	0	0	0	2	7	0
	ND	0	0	0	0	0	0 (0)	9	2	0	0	1	0	0
	OR	7	8	0	0	0	9	0 (0)	0	0	0	4	0	0
	RI	14	0	4	4	0	2	0	0 (0)	0	0	4	7	0
	SD	0	0	0	0	0	0	0	0	0 (0)	1	2	8	0
	USVI	0	0	0	0	0	0	0	0	1	0 (0)	1	0	0
	UT	1	6	7	2	2	1	4	4	2	1	0 (0)	0	3
	WV	3	0	0	0	7	0	0	7	8	0	0	0 (0)	0
	WY	0	4	0	0	0	0	0	0	0	0	3	0	0 (0)

Table 9. Number of Common Middle School Field-Test Items Administered and Calibrated, Spring 2023

	State	CT	HI	ID	MT	NH	ND	OR	RI	SD	US VI	UT	WV	WY
Item Clusters	CT	0 (0)	0	0	0	0	0	0	1	0	1	4	1	0
	HI	0	0 (0)	0	0	0	0	0	0	0	0	8	0	0
	ID	0	0	0 (0)	0	0	0	0	1	0	0	3	2	0
	MT	0	0	0	0 (0)	0	0	0	0	0	0	5	0	0
	NH	0	0	0	0	0 (0)	0	0	0	0	0	7	0	0
	ND	0	0	0	0	0	0 (0)	0	1	0	0	3	0	0
	OR	0	0	0	0	0	0	0 (0)	1	0	0	8	1	0
	RI	1	0	1	0	0	1	1	0 (0)	2	0	2	2	0
	SD	0	0	0	0	0	0	0	2	0 (0)	0	3	0	0
	USVI	1	0	0	0	0	0	0	0	0	0 (0)	1	0	0
	UT	4	8	3	5	7	3	8	2	3	1	0 (0)	3	2
	WV	1	0	2	0	0	0	1	2	0	0	3	0 (0)	0
	WY	0	0	0	0	0	0	0	0	0	0	3	0	0 (0)
Stand-Alone Items	CT	0 (0)	4	10	0	4	0	7	0	0	0	0	4	0
	HI	4	0 (0)	5	0	0	0	3	0	0	0	0	4	0
	ID	10	5	0 (0)	0	0	0	0	1	0	0	0	0	0
	MT	0	0	0	0 (0)	0	0	4	0	0	0	0	0	0
	NH	4	0	0	0	0 (0)	0	0	0	0	0	0	0	0
	ND	0	0	0	0	0	0 (0)	0	7	0	0	0	0	0
	OR	8	3	0	4	0	0	0 (0)	6	0	0	0	7	3
	RI	0	0	1	0	0	7	6	0 (0)	2	0	0	4	0
	SD	0	0	0	0	0	0	0	2	0 (0)	0	0	0	0
	USVI	0	0	0	0	0	0	0	0	0	0 (0)	0	0	0
	UT	0	0	0	0	0	0	0	0	0	0	0 (0)	0	0
	WV	4	4	0	0	0	0	7	4	0	0	0	0 (0)	1
	WY	0	0	0	0	0	0	3	0	0	0	0	1	0 (0)
Total	CT	0 (0)	4	10	0	4	0	7	1	0	1	4	5	0
	HI	4	0 (0)	5	0	0	0	3	0	0	0	8	4	0
	ID	10	5	0 (0)	0	0	0	0	2	0	0	3	2	0
	MT	0	0	0	0 (0)	0	0	4	0	0	0	5	0	0
	NH	4	0	0	0	0 (0)	0	0	0	0	0	7	0	0
	ND	0	0	0	0	0	0 (0)	0	8	0	0	3	0	0
	OR	8	3	0	4	0	0	0 (0)	7	0	0	8	8	3
	RI	1	0	2	0	0	8	7	0 (0)	4	0	2	6	0
	SD	0	0	0	0	0	0	0	4	0 (0)	0	3	0	0
	USVI	1	0	0	0	0	0	0	0	0	0 (0)	1	0	0

	State	CT	HI	ID	MT	NH	ND	OR	RI	SD	US VI	UT	WV	WY
	UT	4	8	3	5	7	3	8	2	3	1	0 (0)	3	2
	WV	5	4	2	0	0	0	8	6	0	0	3	0 (0)	1
	WY	0	0	0	0	0	0	3	0	0	0	3	1	0 (0)

Table 10. Number of Common High School Field-Test Items Administered and Calibrated, Spring 2023

	State	CT	HI	ID	MT	NH	ND	OR	RI	SD	US VI	UT	WV	WY
Item Clusters	CT	0 (0)	1	1	-	0	1	10	6	0	0	-	-	1
	HI	1	0 (0)	1	-	0	0	2	0	0	0	-	-	0
	ID	1	1	0 (0)	-	0	0	5	0	0	0	-	-	0
	MT	-	-	-	-	-	-	-	-	-	-	-	-	-
	NH	0	0	0	-	0 (0)	0	0	0	1	0	-	-	0
	ND	1	0	0	-	0	0 (0)	0	0	1	1	-	-	1
	OR	10	2	6	-	0	0	0 (0)	3	2	0	-	-	0
	RI	7	0	0	-	0	0	5	0 (0)	0	0	-	-	0
	SD	0	0	0	-	1	1	2	0	0 (0)	0	-	-	0
	USVI	0	0	0	-	0	1	0	0	0	0 (0)	-	-	1
	UT	-	-	-	-	-	-	-	-	-	-	-	-	-
	WV	-	-	-	-	-	-	-	-	-	-	-	-	-
	WY	1	0	0	-	0	1	0	0	0	1	-	-	0 (0)
Stand-Alone Items	CT	0 (0)	0	5	-	0	0	0	10	0	0	-	-	0
	HI	0	0 (0)	0	-	0	2	2	0	0	0	-	-	0
	ID	5	0	0 (0)	-	3	0	4	0	0	0	-	-	0
	MT	-	-	-	-	-	-	-	-	-	-	-	-	-
	NH	0	0	3	-	0 (0)	5	0	3	0	0	-	-	0
	ND	0	2	0	-	5	0 (0)	0	1	0	0	-	-	1
	OR	0	2	4	-	0	0	0 (0)	0	0	0	-	-	0
	RI	11	0	0	-	3	1	0	0 (0)	2	0	-	-	7
	SD	0	0	0	-	0	0	0	2	0 (0)	0	-	-	0
	USVI	0	0	0	-	0	0	0	0	0	0 (0)	-	-	0
	UT	-	-	-	-	-	-	-	-	-	-	-	-	-
	WV	-	-	-	-	-	-	-	-	-	-	-	-	-
	WY	0	0	0	-	0	1	0	7	0	0	-	-	0 (0)
Total	CT	0 (0)	1	6	-	0	1	10	16	0	0	-	-	1
	HI	1	0 (0)	1	-	0	2	4	0	0	0	-	-	0
	ID	6	1	0 (0)	-	3	0	9	0	0	0	-	-	0
	MT	-	-	-	-	-	-	-	-	-	-	-	-	-
	NH	0	0	3	-	0 (0)	5	0	3	1	0	-	-	0
	ND	1	2	0	-	5	0 (0)	0	1	1	1	-	-	2
	OR	10	4	10	-	0	0	0 (0)	3	2	0	-	-	0
	RI	18	0	0	-	3	1	5	0 (0)	2	0	-	-	7
	SD	0	0	0	-	1	1	2	2	0 (0)	0	-	-	0
	USVI	0	0	0	-	0	1	0	0	0	0 (0)	-	-	1

	State	CT	HI	ID	MT	NH	ND	OR	RI	SD	US VI	UT	WV	WY
	UT	-	-	-	-	-	-	-	-	-	-	-	-	-
	WV	-	-	-	-	-	-	-	-	-	-	-	-	-
	WY	1	0	0	-	0	2	0	7	0	1	-	-	0 (0)

Following the administration, field-test items went through a substantial validation process. The process began with rubric validation. Rubric validation is a process in which a committee of state educators reviews student responses and the proposed scoring of those responses. The process is described in Volume 2, Section 2.7.1, Rubric validation, of this technical report.

After rubric validation, classical item statistics were computed for the scoring assertions, including item difficulty and item discrimination statistics, testing time, and differential item functioning (DIF) statistics. The MOU established common standards for the statistics. Any items violating these standards were flagged for a second educator review. Even though the scoring assertions were the basic units of analysis used to compute classical item statistics, the business rules to flag items for another educator review were established at the item level because assertions cannot be reviewed in isolation. The statistics and business rules for flagging items are described in Section 4, Field-Test Classical Analysis. For each state, a data review committee consisting of educators (i.e., science teachers) supported by CAI content experts reviewed the items that were owned by the state and flagged for data review according to the established business rules. For ICCR, cross-state review committees were established.

Table 11 presents the number of field-test items administered in Rhode Island, or another state or territory, the number of items rejected before or during rubric validation, the number of items sent for data review, and the number of items rejected during data review. The numbers in parentheses present the number of field-test items owned by Rhode Island.

Table 11. Number of Field-Test Item Administration, Rubric Validation, and Item Data Review, Spring 2023

Grade Band and Item Type	Number of Field-Test Items Administered	Number of Items Rejected Before/During Rubric Validation	Number of Items Sent to Data Review	Number of Items Rejected at Data Review	Number of Items Remaining
Elementary School	**126 (6)**	**3 (1)**	**71 (3)**	**13 (2)**	**110 (3)**
Cluster	60 (1)	1 (0)	17 (0)	1 (0)	58 (1)
Stand-Alone	66 (5)	2 (1)	54 (3)	12 (2)	52 (2)
Middle School	**136 (8)**	**2 (0)**	**80 (6)**	**20 (2)**	**114 (6)**
Cluster	59 (3)	1 (0)	21 (2)	5 (1)	53 (2)
Stand-Alone	77 (5)	1 (0)	59 (4)	15 (1)	61 (4)

Grade Band and Item Type	Number of Field-Test Items Administered	Number of Items Rejected Before/During Rubric Validation	Number of Items Sent to Data Review	Number of Items Rejected at Data Review	Number of Items Remaining
High School	**86 (15)**	**5 (4)**	**44 (8)**	**17 (2)**	**64 (9)**
Cluster	40 (9)	4 (3)	19 (5)	6 (1)	30 (5)
Stand-Alone	46 (6)	1 (1)	25 (3)	11 (1)	34 (4)
Total	**348 (29)**	**10 (5)**	**195 (17)**	**50 (6)**	**288 (18)**

Note. RI-owned items are indicated in the parentheses.

Table 12 summarizes the Shared Science Assessment Item Bank after adding the field-test items that were administered in 2023 and passed rubric validation and item data review. The numbers in parentheses present the number of items owned by Rhode Island.

Table 12. Summary of Shared Science Assessment Item Bank, Spring 2023

Grade Band and Item Type	Science Discipline			Item Bank Total[a]
	Earth and Space Sciences	*Life Sciences*	*Physical Sciences*	
Elementary School	**205 (13)**	**202 (10)**	**254 (12)**	**661 (35)**
Cluster	112 (6)	102 (6)	131 (6)	345 (18)
Stand-Alone	93 (7)	100 (4)	123 (6)	316 (17)
Middle School	**185 (8)**	**262 (11)**	**215 (12)**	**662 (31)**
Cluster	88 (4)	129 (4)	106 (5)	323 (13)
Stand-Alone	97 (4)	133 (7)	109 (7)	339 (18)
High School	**110 (10)**	**207 (7)**	**151 (8)**	**468 (25)**
Cluster	45 (5)	89 (4)	62 (4)	196 (13)
Stand-Alone	65 (5)	118 (3)	89 (4)	272 (12)
Total	**500 (31)**	**671 (28)**	**620 (32)**	**1791 (91)**

Note. RI-owned items are indicated in the parentheses.
[a]Count excludes nine MOU items that do not align to the NGSS.

3.3 TEST DESIGN

The science tests were assembled under a LOFT design, with the exception of the braille and paper-pencil forms. Tests were assembled using CAI's adaptive testing algorithm. 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. Vice versa, the content value decreases for items with content features for which the minimum has been met. 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.

Under a LOFT design, operational items are selected on the fly solely based on their contributions to meeting the blueprint by assigning a weight of zero to the information value of an item with respect to the underlying proficiency. The RI NGSA blueprints are presented in this technical report in Volume 2, Section 4.2, Test Blueprints. Details for CAI's adaptive testing algorithm are described in Appendix 2-L, Adaptive Algorithm Design of Volume 2, Test Development.

The braille and paper-pencil tests were accommodated fixed forms. Form construction of the accommodated form is discussed in Volume 2, Section 4.4, Paper-Pencil Accommodation Form Construction.

The main characteristics of the blueprint were that any performance expectation (PE) could be tested only once (indicated by the values of 0 and 1 for the minimum and maximum values of the individual PEs in the test blueprints; see Section 4.2, Test Blueprints of Volume 2). In general, no more than one item cluster or two stand-alone items could be sampled from the same Disciplinary Core Idea (DCI), and no more than three total items could be sampled from the same DCI (as indicated by the minimum and maximum values in the rows representing DCIs). For both the 2018 and 2019 test administrations, a segmented test design was used; items were administered grouped in four segments. The segments corresponded to each of the three science disciplines and an additional field-test segment that could contain items from all three science disciplines.

In 2018, the order of the segments corresponding to the science disciplines was randomized over students. The additional field-test segment consisted of one cluster and was always presented at the end of the test (segment 4). The primary purpose was to collect additional student responses for the item clusters that had low exposure in the first three segments.

Starting from 2019, the scored operational part of the test consisted of the three segments corresponding to science disciplines. The embedded field-test segment consisted of two item clusters and four stand-alone items. In order to ensure that every student received exactly two item clusters and four stand-alone items as field-test items, the embedded field-test segment was split into two segments: one for field-test item clusters, and one for field-test stand-alone items. The test was taken over two days. On the first day, half of the students received two operational segments, chosen at random from the three operational segments. The other half received one randomly chosen operational segment and the embedded field-test segments. The remaining segments were administered on the second day. Within a day, the order of the segments was randomized, with the restriction that the field-test segments for item clusters and stand-alone items were always administered right after each other.

4. FIELD TEST CLASSICAL ANALYSIS

As explained in Section 3, **Error! Reference source not found.**, science items administered as field-test items underwent rubric validation and data review. Items were flagged for data review based on business rules defined on classical item statistics. Except for response times, the classical item statistics are computed for individual assertions, whereas the business rules for flagging are defined at the item level.

In general, item statistics used to flag items for data review were computed using the student responses of the state that owned the items. However, for Independent College and Career

Readiness (ICCR) items, the flagging rules were defined on the item statistics computed from the combined data of states that used ICCR items. In 2023, those states were Connecticut, Hawaii, Idaho, New Hampshire, North Dakota, Oregon, Rhode Island, South Dakota, Utah, U.S. Virgin Islands, West Virginia, and Wyoming. Furthermore, to compute the differential item functioning (DIF) statistics for the field-test items, the data from all states were combined to obtain a sufficient number of students for each demographic group.

The criteria for flagging and reviewing items are provided in Table 13 and the statistics are described in Section 4.1, Item Discrimination through Section 4.4, Differential Item Functioning . Items that were flagged for data review were reviewed by a committee, as explained in Section 0, Item Bank and Test Design.

Table 13. Thresholds for Flagging in Classical Item Analysis

Analysis Type	Flagging Criteria
Item Discrimination	Average biserial correlation < 0.25 (across the assertions within an item)
	One or more assertions with a biserial correlation < 0.05
Item Difficulty (Clusters)	Average *p*-value < .30 or > 0.85 (across the assertions within an item cluster)
Item Difficulty (Stand-Alone items)	Average *p*-value < .15 or > 0.95 (across the assertions within a stand-alone item)
Timing (Clusters)	Percentile 80* > 15 minutes
Timing (Stand-Alone items)	Percentile 80* > 3 minutes
Timing	Assertions per (percentile 80) minute < 0.5
DIF (Clusters)	Two or more assertions show 'C' DIF in the same direction
DIF (Stand-Alone items)	One or more assertions show 'C' DIF in the same direction

Note. *A percentile 80 of *x* minutes: 80% of the students spent *x* minutes or less on the item.

4.1 ITEM DISCRIMINATION

The item discrimination index indicates the extent to which each item differentiated between those test takers who possessed the skills being measured and those who did not. Generally, the higher the value, the better the item can differentiate between high- and low-achieving students.

For each assertion within an item, the discrimination index was calculated as the biserial correlation between the assertion score and the ability estimate for students. The average biserial correlation was then calculated across the assertions within an item.

4.2 ITEM DIFFICULTY

Items that are either very difficult or very easy are flagged for review but are not necessarily removed if they are grade-level appropriate and aligned with the test specifications. For science, both the *p*-value for individual assertions and the average across all assertions of an item are calculated. Acceptable item *p*-values are summarized in Table 13.

4.3 RESPONSE TIME

Given that the science clusters consist of multiple student interactions, they require more time for students to complete. To ensure a good balance between the amount of information an item provides, and the time students spend on the item, item response time was recorded and analyzed. Specifically, the statistic "percentile 80" was computed for each item. A percentile 80 of *x* minutes means that 80% of the students spent *x* minutes or fewer on the item. An item was flagged for review when the

- percentile 80 > 15 minutes, if the item is an item cluster;
- percentile 80 > 3 minutes, if the item is a stand-alone item; or
- assertions per (percentile 80) minute < 0.5.

4.4 DIFFERENTIAL ITEM FUNCTIONING

DIF refers to items that appear to function differently across identifiable groups, typically across different demographic groups. Identifying DIF is important because it provides a statistical indicator that an item may contain cultural or other bias. DIF-flagged items are further examined by content experts who are asked to re-examine each flagged item to decide whether the item should be excluded from the pool due to bias. Not all items that exhibit DIF are biased; characteristics of the educational system may also lead to DIF.

Cambium Assessment, Inc. (CAI) uses a generalized Mantel-Haenszel (MH) procedure to calculate DIF. The generalizations include (1) adaptation to polytomous items, and (2) improved variance estimators to render the test statistics valid under complex sample designs. With this procedure, each student's estimated theta score on the operational items on a given test is used as the ability-matching variable. That score is divided into 10 intervals to compute the MH chi-square ($MH\chi^2$) DIF statistic for balancing the stability and sensitivity of the DIF scoring category selection. For dichotomous items, the following statistics were computed the $MH\chi^2$ value, the conditional odds ratio, and the MH-delta. For polytomous items, the $GMH\chi^2$ and the standardized mean difference (SMD) were computed.

The MH chi-square statistic (Holland & Thayer, 1988) is calculated as:

$$MH\chi^2 = \frac{(|\sum_k n_{R1k} - \sum_k E(n_{R1k})| - 0.5)^2}{\sum_k var(n_{R1k})}$$

where $k = \{1, 2, \ldots K\}$ for the strata, $n_{R1k}$is the number of students with correct responses for the reference group in stratum k, and 0.5 is a continuity correction. The expected value is calculated as

$$E(n_{R1k}) = \frac{n_{+1k} n_{R+k}}{n_{++k}}$$

where n_{+1k} is the number of students with correct responses, n_{R+k} is the number of students in the reference group, and n_{++k} is the number of students in stratum k, and the variance is calculated as

$$var(n_{R1k}) = \frac{n_{R+k}n_{F+k}n_{+1k}n_{+0k}}{n_{++k}^2(n_{++k}-1)}.$$

n_{F+k} is the number of students in the focal group, n_{+1k} is the number of students with correct responses, and n_{+0k} is the number of students with incorrect responses in stratum k.

The MH conditional odds ratio is calculated as

$$\alpha_{MH} = \frac{\sum_k n_{R1k}n_{F0k}/n_{++k}}{\sum_k n_{R0k}n_{F1k}/n_{++k}}.$$

The MH-delta (Δ_{MH}, Holland & Thayer, 1988) is then defined as

$$\Delta_{MH} = -2.35\ln(\alpha_{MH}).$$

The generalized MH statistic generalizes the MH statistic to polytomous items (Somes, 1986), and is defined as

$$GMH\chi^2 = \left(\sum_k \boldsymbol{a}_k - \sum_k E(\boldsymbol{a}_k)\right)' \left(\sum_k var(\boldsymbol{a}_k)\right)^{-1} \left(\sum_k \boldsymbol{a}_k - \sum_k E(\boldsymbol{a}_k)\right),$$

where $\boldsymbol{a}_k$ is a $(T-1)\times 1$ vector of item response scores and $E(\boldsymbol{a}_k)$ is a $(T-1)\times 1$ mean vector, both corresponding to the T response categories of a polytomous item (excluding one response); $var(\boldsymbol{a}_k)$ is a $(T-1)\times(T-1)$ covariance matrix calculated analogously to the corresponding elements in $MH\chi^2$ in stratum k.

The SMD (Dorans & Schmitt, 1991) is defined as

$$SMD = \sum_k p_{FK}m_{Fk} - \sum_k p_{FK}m_{Rk},$$

where

$$p_{FK} = \frac{n_{F+k}}{n_{F++}}$$

is the proportion of the focal group students in stratum k,

$$m_{Fk} = \frac{1}{n_{F+k}}\left(\sum_t a_t n_{Ftk}\right)$$

is the mean item score for the focal group in stratum k, and

$$m_{Rk} = \frac{1}{n_{R+k}}\left(\sum_t a_t n_{Rtk}\right)$$

is the mean item score for the reference group in stratum k.

DIF analysis was conducted for all field-test items with at least 200 responses per item in each subgroup (Zwick, 2012) to detect potential item bias for major demographic groups. Student

responses from multiple states were combined to minimize the number of items with insufficient sample sizes for one or more demographic groups.

DIF statistics were calculated at the assertion level and were performed for the following groups (some items had insufficient sample sizes for DIF analyses in some groups):

- Female vs. Male
- American Indian/Alaskan Native vs. White
- Asian vs. White
- African American vs. White
- Hawaiian/Pacific Islander vs. White
- Hispanic vs. White
- Multi-Racial vs. White
- English Learner (EL) vs. Non-EL
- Special Education (SPED) vs. Non-SPED
- Economically Disadvantaged vs. Non-Economically Disadvantaged

Similar to how the general MH statistic is used to classify items of traditional tests, assertions were classified into three categories (A, B, or C) for DIF, ranging from no evidence of DIF to severe DIF. The classification rules are shown in Table 14. Furthermore, assertions were categorized positively (i.e., +A, +B, or +C), signifying that an item favored the focal group (e.g., African American or female), or negatively (i.e., –A, –B, or –C), signifying that an item favored the reference group (e.g., White or male).

An item was flagged for data review according to the following criteria:

- **Item Clusters.** Two or more assertions showed "C" DIF in the same direction.
- **Stand-Alone Items.** One or more assertions showed "C" DIF.

Table 14. DIF Classification Rules

Assertions

Category	*Rule*
C	MH_{X^2} is significant and $\lvert SMD\rvert/\lvert SD\rvert \geq 0.25$.
B	MH_{X^2} is significant and $\lvert SMD\rvert/\lvert SD\rvert < 0.25$.
A	MH_{X^2} is not significant.

Note that for the 2018 field test, a slightly less strict criterion was used for item clusters with 10 or more assertions (i.e., three or more assertions with "C" DIF in the same direction). The change was made taking into consideration the feedback received from several Technical Advisory

Committees (TACs) and modified such that the rate of flagging items for DIF was similar for item clusters and stand-alone items (based on the flagging rates computed on items field tested in 2018).

4.5 CLASSICAL ANALYSIS RESULTS

This section presents a summary of results from classical item analysis of the 2023 field-test items administered in RI NGSA. Table 15 and Table 16 provide summaries of the *p*-values and biserial correlations for the science field-test items administered in Rhode Island in 2023. The *p*-values, biserials, and response times were computed using Rhode Island data. The DIF statistics are computed using data from all MOU states that administered those items. The average values across the assertions within an item were used in the computation of the percentiles and ranges.

Table 15. Distribution of p-*Values for Field-Test Items, 2023*

Grade	Total FT Items	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
5	34	0.15	0.19	0.34	0.48	0.54	0.76	0.83
8	32	0.07	0.09	0.19	0.32	0.44	0.57	0.57
11	29	0.14	0.20	0.31	0.39	0.45	0.63	0.74

Table 16. Distribution of Item Biserial Correlations for Field-Test Items, 2023

Grade	Total FT Items	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
5	34	0.19	0.25	0.39	0.44	0.51	0.64	0.67
8	32	-0.02	0.08	0.29	0.36	0.49	0.60	0.64
11	29	0.14	0.17	0.42	0.46	0.53	0.61	0.71

Table 17 presents respective summaries of response times by item type (item cluster or stand-alone item) for Rhode Island field-test items administered in 2022.*Table 17. Summary of Response Times for Field-Test Items Administered Spring 2023*

Grade	Item Type	Total FT Items	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
5	Cluster	12	5.90	6.95	9.65	11.50	13.20	15.69	16.90
	Stand-Alone	22	1.80	2.20	2.93	3.45	4.28	5.39	5.70
8	Cluster	9	4.50	5.30	7.00	7.80	8.10	8.88	9.00
	Stand-Alone	23	1.00	1.41	1.75	2.10	2.45	3.74	4.20

Grade	Item Type	Total FT Items	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
11	Cluster	9	7.80	8.52	9.70	9.80	14.20	14.80	14.80
	Stand-Alone	20	1.80	1.80	2.40	2.95	3.65	4.12	4.40

Table 18 presents, for each item type, the number of field-test items flagged for DIF for each demographic group included in the 2023 DIF analyses for Rhode Island.

Table 18. Differential Item Functioning Classifications for Field-Test Items Administered, Spring 2023

DIF Flag	Item Type	Female/ Male	American Indian[a]/ White	Asian/ White	African American /White	Hawaiian[b] /White	Hispanic /White	Multi-Racial/ White	EL/Non-EL	SPED/ Non-SPED	Econ-Disad/ Non-Econ-Disad[c]
Grade 5											
Items Evaluated	Cluster	12	0	0	5	0	12	2	12	12	12
	Stand-Alone	22	2	0	13	0	22	0	22	22	22
Items Flagged C	Cluster	0	0	0	0	0	0	0	0	0	0
	Stand-Alone	0	0	0	0	0	0	0	0	1	0
% Items Flagged C	Cluster	0	-	-	0	-	0	0	0	0	0
	Stand-Alone	0	0	-	0	-	0	-	0	4.55	0
Grade 8											
Items Evaluated	Cluster	9	1	0	9	0	9	5	9	9	9
	Stand-Alone	20	2	0	20	0	20	4	20	20	20
Items Flagged C	Cluster	0	0	0	0	0	0	0	0	0	0
	Stand-Alone	0	0	0	0	0	0	0	0	0	0
% Items Flagged C	Cluster	0	0	-	0	-	0	0	0	0	0
	Stand-Alone	0	0	-	0	-	0	0	0	0	0
Grade 11											
Items Evaluated	Cluster	9	0	0	6	0	9	0	0	9	9
	Stand-Alone	23	0	0	11	0	23	0	0	20	23
Items Flagged C	Cluster	0	0	0	0	0	0	0	0	0	0
	Stand-Alone	1	0	0	1	0	0	0	0	0	0
% Items Flagged C	Cluster	0	-	-	0	-	0	-	-	0	0
	Stand-Alone	4.35	-	-	9.09	-	0	-	-	0	0

Note. Full DIF Group names: [a]American Indian/Alaskan Native; [b]Hawaiian/Pacific Islander; [c]Economically Disadvantaged vs. Non-Economically Disadvantaged

In 2023, 100 field-test items were administered in Rhode Island. Of those, 95 items passed rubric validation, 13 were flagged for item discrimination, 16 were flagged for *p*-value, 42 were flagged for response time, and 3 were flagged for DIF according to the criteria (as described in Section **Error! Reference source not found.**, Item Discrimination, through Section **Error! Reference source not found.**, Differential Item Functioning). Some items were flagged for multiple reasons. Flagged field-test items were reviewed by educators during data review. The total number of field-test items flagged and the total number of field-test items that passed item data review in 2023 are summarized in Table 11.

5. ITEM CALIBRATION

5.1 MODEL DESCRIPTION

In discussing item response theory (IRT) models for Rhode Island and Vermont, we distinguish between the underlying latent structure of a model and the parameterization of the item response function conditional on that assumed latent structure. Subsequently, we discuss how group effects are considered.

5.1.1 Latent Structure

Most operational assessment programs rely on a unidimensional IRT model for item calibration and computing scores for students. These models assume a single underlying trait and that items are independent given the value of that underlying trait. In other words, the models assume that given the value of the underlying trait, knowing the response to one item provides no information about responses to other items. This assumption of conditional independence implies that the conditional probability of a pattern of *I* item responses takes the relatively simple form of a product over items for a single student:

$$P(\mathbf{z_j}|\theta_j) = \prod_{i=1}^{I} P(z_{ij}|\theta_j),$$

where z_{ij} represents the scored response of student j (j = 1, …, N) to item i (i = 1, …, I), $\mathbf{z_j}$ represents the pattern of scored item responses for student j, and θ_j represents student j's proficiency. Unidimensional IRT models differ with respect to the functional relation between the proficiency θ_j and the probability of obtaining a score z_{ij} on item i.

The items in the RI NGSA are more complex than traditional item types. A single item may contain multiple parts, and each part may contain multiple student interactions. For example, a student may be asked to select a term from a set of terms at several places in a single item. Instead of receiving a single score for each item, multiple inferences are made about the knowledge and skills that a student has demonstrated based on specific features of the student's responses to the item. These scoring units are called *assertions* and are the basic unit of analysis in our IRT analysis. That is, they fulfill the role of items in traditional assessments. However, for the RI NGSA items,

multiple assertions are typically developed around a single item so that assertions are clustered within items.

One approach is to apply one of the traditional IRT models to the scored assertions. However, a substantial complexity that arises from the use of this new item type is that local dependencies exist between assertions pertaining to the same stimulus (item or item cluster). The local dependencies between the assertions pertaining to the same stimulus constitute a violation of the assumption that a single latent trait can explain all dependencies between assertions. Fitting a unidimensional model in the presence of local dependencies may result in biased item parameters and standard errors of measurement (SEMs). In particular, it is well documented that ignoring local item dependencies leads to an overestimation of the amount of information conveyed by a set of responses and an underestimation of the SEM (e.g., Sireci, Thissen, & Wainer, 1991; Yen, 1993).

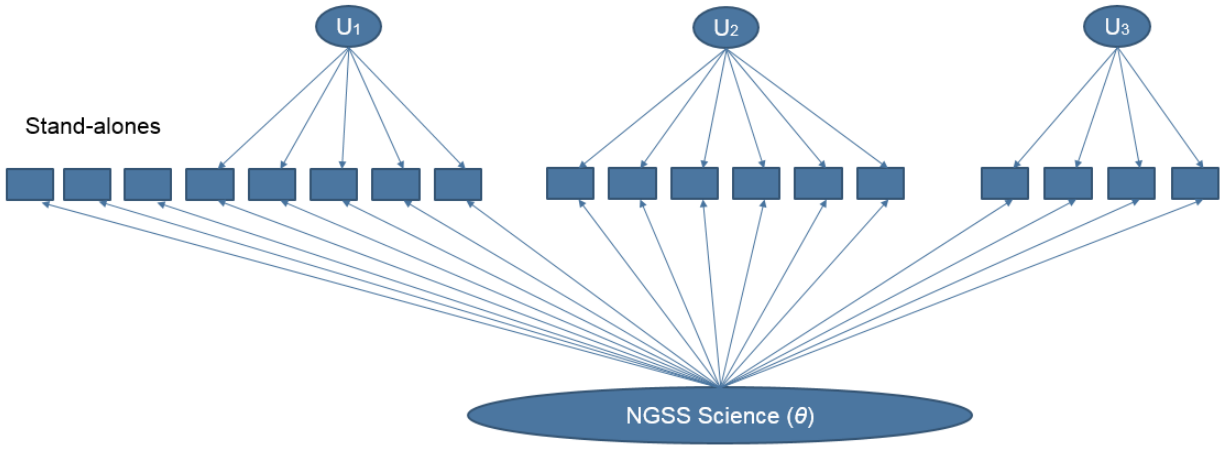
The effects of groups of assertions developed around a common stimulus can be accounted for by including additional dimensions corresponding to those groupings in the IRT model. These dimensions are considered to be nuisance dimensions. Whereas traditional unidimensional IRT models assume that all assertions (the basic units of analysis) are independent given a single underlying trait θ, we now assume the conditional independence of assertions given the underlying latent trait θ and all nuisance dimensions:

$$P(\mathbf{z_j}|\theta_j, \mathbf{u}_j) = \prod_{i \in \mathrm{SA}} P(z_{ij}|\theta_j) \prod_{g=1}^{G} \prod_{i \in g} P(z_{ij}|\theta_j, u_{jg}),$$

where SA indicates stand-alone assertions, u_g indicates the nuisance dimension for assertion group g (with the position of student j on that dimension denoted as u_{jg}), and u is the vector of all G nuisance dimensions. It can be seen that the conditional probability $P(z_{ij}|\theta_j, u_{jg})$ now becomes a function of two latent variables: the latent trait θ, representing a student's proficiency in science (the underlying trait of interest), and the nuisance dimension u_g, accounting for the conditional dependencies between assertions of the same group. Furthermore, we assume that the nuisance dimensions are all uncorrelated with one another and with the general dimension. It is important to point out that even though every group of assertions introduces an additional dimension, models with this latent structure do not suffer from the curse of dimensionality like other multidimensional IRT models because one can take advantage of this special structure during model calibration (Gibbons & Hedeker, 1992). In this regard, Rijmen (2010) showed that it is unnecessary to assume that all nuisance dimensions are uncorrelated; rather, it is sufficient that they are independent, given the general dimension θ.

The model structure of the IRT model for science is illustrated in Figure 1. Note that stand-alone items can be scored with more than one assertion. The assertions of stand-alone items with more than one assertion but fewer than four assertions were also modeled as stand-alone assertions. Even though these assertions are likely to exhibit conditional dependencies, the variance of the nuisance dimension cannot be reliably estimated if it is based on a very small number of assertions. The few stand-alone items with four or more assertions were treated as item clusters to consider the conditional dependencies.

Figure 1. Directed Graph of the Science IRT Model



5.1.2 Item Response Function

The item response functions of the stand-alone assertions are modeled with a unidimensional model. For the grouped assertions, like in unidimensional models, different parametric forms can be assumed for the conditional probability of obtaining a score of z_{ij}. For binary data, the Rasch testlet model (Wang & Wilson, 2005) is defined as:

$$P(z_{ij}|\theta_j, u_{jg}; b_i) = \frac{\exp(\theta_j + u_{jg} - b_i)}{1 + \exp(\theta_j + u_{jg} - b_i)}.$$

The item response function of the Rasch testlet model is the probability of a correct answer (i.e., a true assertion), as a function of the overall proficiency θ, the nuisance dimension u_g, and the item (i.e., assertion) difficulty b_i. The Rasch testlet model does not include item discrimination parameters; however, the same model structure as presented in Figure 1 could be employed with discrimination parameters included in Equations **Error! Reference source not found.** and **Error! Reference source not found.**. Furthermore, only models for binary data are considered. Assertions are always binary because they are either true or false. Nevertheless, the model could easily accommodate polytomous responses by using the same response function incorporated in unidimensional models for polytomous data.

5.1.3 Multigroup Model

The Share Science Assessment Item Bank is calibrated concurrently using all the items administered in any of the state or territory that collaborate with Cambium Assessment, Inc. (CAI) on their new science assessments. In the calibration, each state or territory is treated as a population of students or group. Overall group differences are taken into account by allowing a group-specific distribution of the overall proficiency variable θ. Specifically, for every student j belonging to group k, $k = 1, \ldots, K$, a normal distribution is assumed,

$$\theta_j \sim N(\mu_k, \sigma_k^2),$$

where μ_k and σ_k^2 are the mean and variance of a normal distribution. The mean of the reference distribution ($k = 1$) is set to 0 to identify the model (for free calibrations, where there are no anchor items with their location parameters set to specific values). For each of the nuisance variables u_g, a common variance parameter across groups was assumed, and the means are set to 0 in order to identify the model,

$$u_{jg} \sim N\left(0, \sigma_{u_g}^2\right).$$

5.2 ESTIMATION

A separate IRT model was fit for each grade band. The parameters of the IRT model were estimated using the marginal maximum likelihood (MML) method. In the MML method, the latent proficiency variable θ_j and the vector of nuisance parameters $\boldsymbol{u}_j$ for each student j are treated as random effects and integrated out to obtain the marginal log likelihood corresponding to the observed response pattern $\boldsymbol{z}_j$ for student j,

$$\ell_j = \log \int \int P\left(\boldsymbol{z}_j | \theta_j, \boldsymbol{u}_j\right) N\left(\theta_j | \mu_k, \sigma_k^2\right) N\left(\boldsymbol{u}_j | \boldsymbol{0}, \boldsymbol{\Sigma}\right) d\boldsymbol{u}_j d\theta_j,$$

where $\boldsymbol{\Sigma}$ is a diagonal matrix with diagonal elements $\sigma_{u_k}^2$, denoting nuisance variance for group k. Across all students and groups, the overall log likelihood to be maximized with respect to the vector $\boldsymbol{\gamma}$ of all model parameters (item difficulty parameters, and the mean and variance parameters of the latent variables) is

$$\ell(\boldsymbol{\gamma}) = \sum_k \sum_{j \in k} \ell_j.$$

Even though the number of latent variables in the overall log likelihood equation is very high, the curse of dimensionality can be avoided because the integration over the high-dimensional latent $(\theta, \boldsymbol{u})$ space can be carried out as a sequence of computations in two-dimensional space $(\theta, \boldsymbol{u_g})$ (Gibbons & Hedeker, 1992; Rijmen, 2010).

The Shared Science Assessment Item Bank was calibrated freely in 2018 after the 2018 science test administrations concluded, and it was recalibrated in 2019 following the 2019 test administrations. Following 2019, field-test items are calibrated onto the scale of the Shared Science Assessment Item Bank by anchoring the operational items to their bank. In the anchored calibrations, the mean and variance of the overall science dimension are also estimated for each group.

Appendix 1-C, Calibration of the Shared Science Assessment Item Bank, contains a detailed description of the 2018 and 2019 calibration processes as well as a description of how the 2018 and 2019 scales were linked.

Starting in 2021, CAIRT (Cambium Assessment IRT) is used to calibrate item parameters. CAIRT was specifically developed by CAI to calibrate the multigroup Rasch model on very large data sets because estimation times in commercially available software (i.e., flexMIRT) became prohibitive. CAIRT relies on the same estimation methods as the Baysian networks with the logistic regression

(BNL; Rijmen, 2006), a suite of Matlab functions for estimating a wide variety of latent variable models. BNL uses an efficient expectation-maximization (EM) algorithm based on graphical model theory (Rijmen, 2010). CAI has cross-validated parameter estimates from CAIRT with BNL and flexMIRT under various scenarios (Rijmen, Liao, & Lin, 2021). CAIRT is a web application that is available at no cost to members of the MOU.

Table 19 provides an overview of the groups per grade band for calibration of the 2023 field-test items. All but 81 items were calibrated on at least 1,500 student responses, and all but 5 items had a sample size larger than 1,200. The 5 items with fewer than 1,200 responses were either Oregon legacy items or Hawaii items for a research study.

Table 19. Groups Per Grade Band for the Spring 2023 Calibration of Field-Test Items

Group	Elementary School	Middle School	High School
Connecticut	X	X	X
Hawaii	X	X	X
Idaho	X	X	X
Montana	X	X	
North Dakota	X	X	X
New Hampshire	X	X	X
Oregon	X	X	X
Rhode Island	X	X	X
South Dakota	X	X	X
U.S. Virgin Islands	X	X	X
Utah	X	X	
West Virginia	X	X	
Wyoming	X	X	X

5.3 Overview of the Operational Bank

Figure 2 through Figure 4**Error! Reference source not found.** display the histogram of the difficulty parameters for grades 5, 8, and 11 for all items that are part of the Rhode Island operational pool. The figures also display the student proficiency distributions. The grade 5 items are slightly easier compared to the student proficiency level. The distribution of the difficulty parameter overlaps well with the proficiency distribution in grade 8. The grade 11 items are slightly more difficult than the student proficiency in general.

Figure 2. Assertion Difficulty and Student Proficiency Distributions, Grade 5

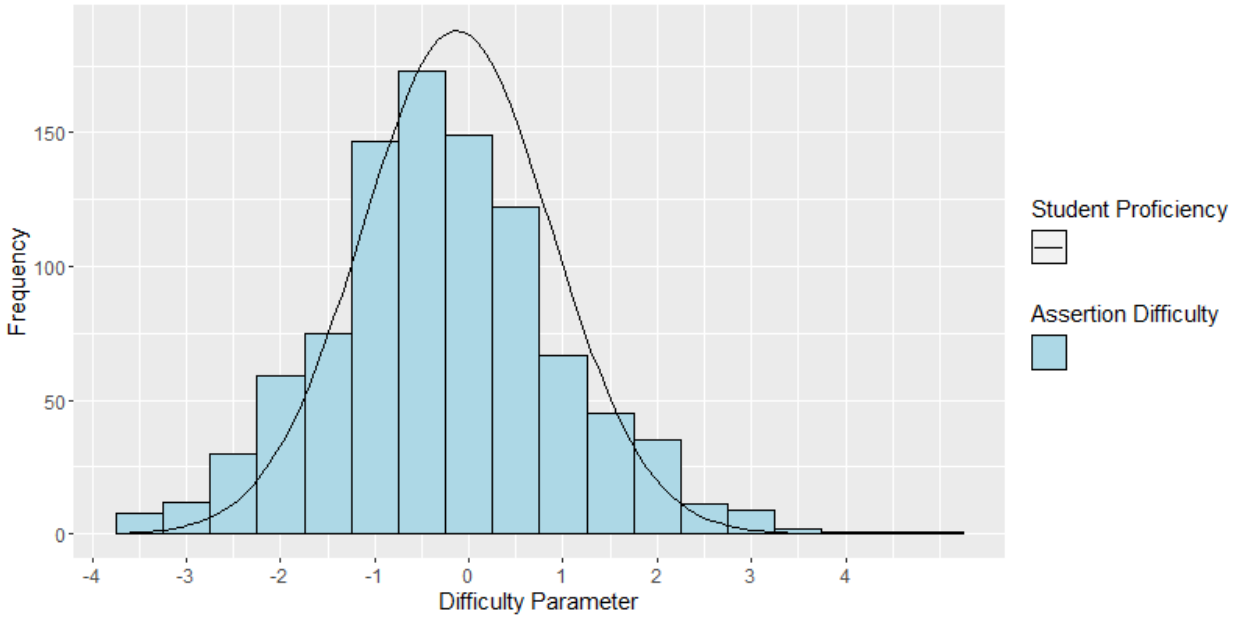


Figure 3. Assertion Difficulty and Student Proficiency Distributions, Grade 8

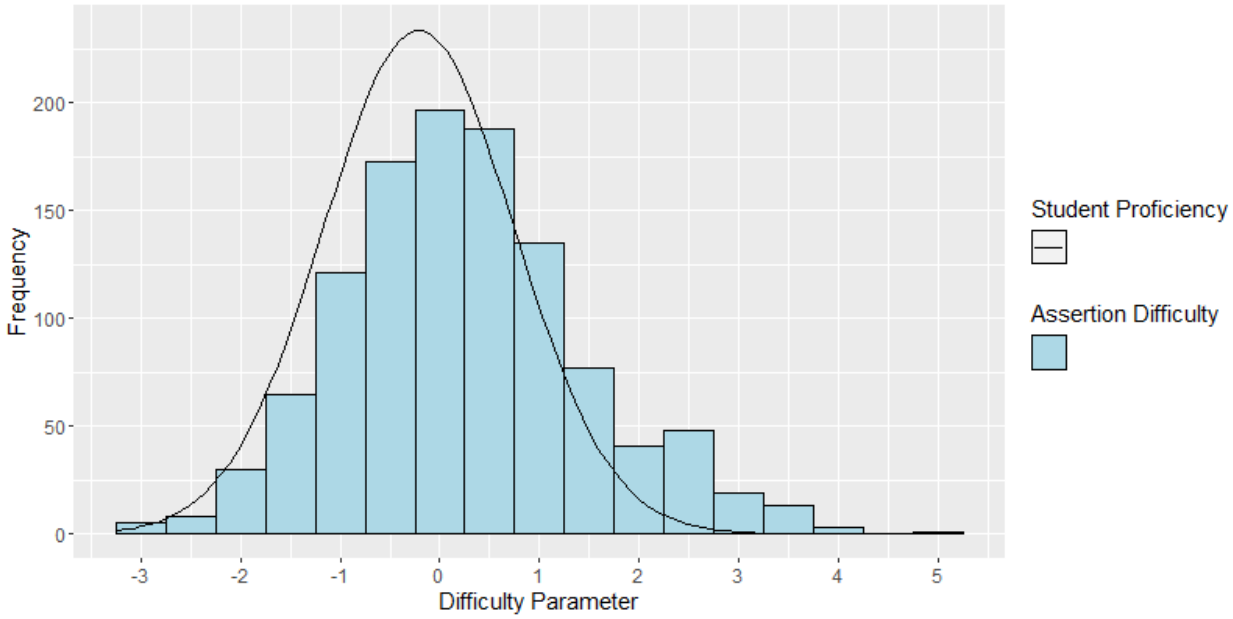
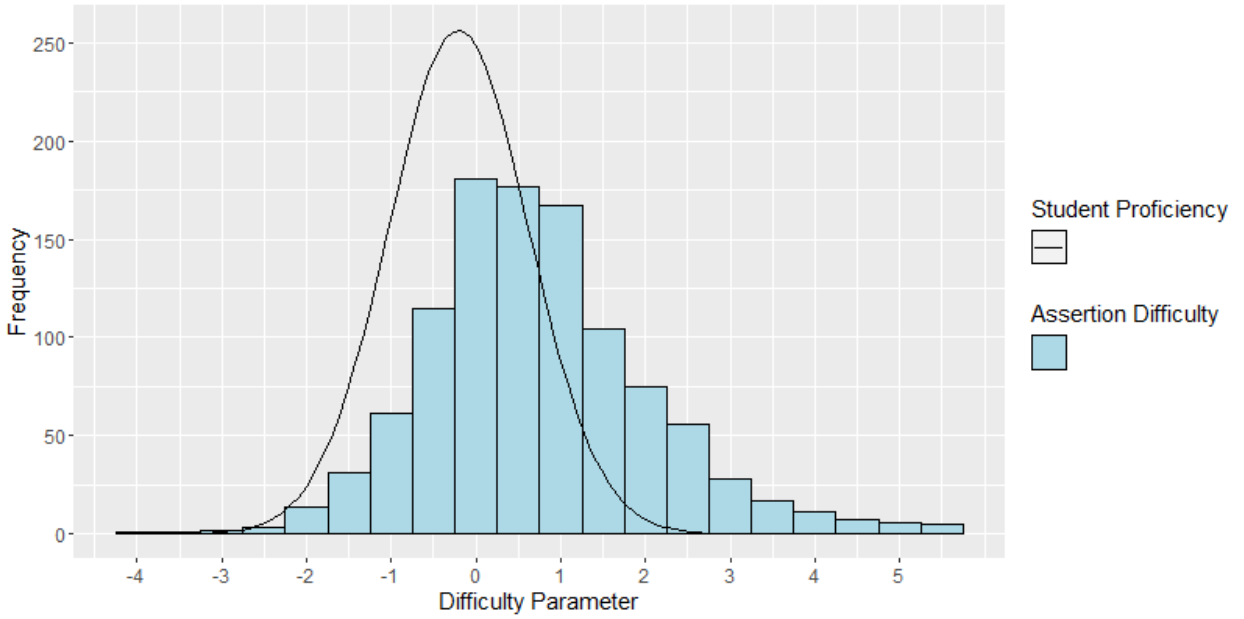


Figure 4. Assertion Difficulty and Student Proficiency Distributions, Grade 11



6. SCORING

6.1 MARGINAL MAXIMUM LIKELIHOOD FUNCTION

Student scores are obtained by marginalizing out the nuisance dimensions u_j from the likelihood of the observed response pattern z_j for student *j*,

$$\ell_i(\theta_j) = log \int_{u_j} P(z_j|\theta_j, u_j)\, N(u_j|0, \Sigma) du_j,$$

and maximizing this marginalized likelihood function for θ_j. The marginal maximum likelihood estimation (MMLE) estimator is a hybrid between the expected a posteriori (EAP) estimator (by marginalizing out the nuisance dimensions) and the maximum likelihood estimation (MLE) estimator (by maximizing the resulting marginal likelihood for θ). The marginal likelihood is maximized with respect to θ using the Newton Raphson method. See Rijmen, Jiang, and Turhan (2018) for more details of the MMLE estimator and the validation study by Connecticut State Department of Education (2019) for the use of this estimator.

The proposed model reduces to the unidimensional Rasch model when the nuisance variances are zero for all *g*. Likewise, the proposed MMLE is equivalent to the MLE of the unidimensional Rasch model when all the nuisance variances are zero. This can be shown by using the variable transformation $v = \boldsymbol{\Sigma}\, \boldsymbol{\Sigma}^{-\frac{1}{2}} u$. Then we have

$$\int_{u_j} P(z_j|\theta_j, u_j)\, N(u_j|0, \Sigma) du_j = \int_{v_j} P\left(z_j \left| \theta_j, \Sigma^{\frac{1}{2}} v_j\right.\right) N(v_j|0, I) dv_j.$$

If $\sigma_{u_g}^2 = 0$ for all g, then

$$\int_{u_j} P\left(z_j \middle| \theta_j, u_j\right) N\left(u_j \middle| 0, \Sigma\right) du_j = P\left(z_j \middle| \theta_j\right),$$

which is the likelihood under the unidimensional Rasch model.

6.2 DERIVATIVE

The marginal log likelihood function based on the item response theory (IRT) model with one overall dimension and one nuisance dimension for each grouping of assertions can be written as

$$l(\theta) = \sum_{i \in SA} log\left(P(z_i|\theta)\right) + \sum_{g=1}^{G} log\left\{\int Exp\left[\sum_{i \in g} log\left(P\left(z_{ig}|\theta, u_g\right)\right)\right] N\left(u_g \middle| 0, \sigma_{u_g}^2\right) du_g\right\}.$$

The first derivative of the marginal log likelihood function with respect to θ is

$$\frac{dl(\theta)}{d\theta} = \sum_{i \in \mathrm{SA}} \frac{\frac{dP(z_i|\theta)}{d\theta}}{P(z_i|\theta)} + \sum_{g=1}^{G} \frac{\int \left\{ \mathrm{Exp}\left[\sum_{i \in g} \log\left(P\left(z_{ig}|\theta, u_g\right)\right)\right] \left(\sum_{i \in g} \frac{\frac{dP\left(z_{ig}|\theta, u_g\right)}{d\theta}}{P\left(z_{ig}|\theta, u_g\right)} \right) N\left(u_g \middle| 0, \sigma_{u_g}^2\right) \right\} du_g}{\int \left\{ \mathrm{Exp}\left[\sum_{i \in g} \log\left(P\left(z_{ig}|\theta, u_g\right)\right)\right] N\left(u_g \middle| 0, \sigma_{u_g}^2\right) \right\} du_g}$$

and the second derivative of the marginal log likelihood function with respect to θ is

$$\frac{d^2 l(\theta)}{d\theta^2}$$

$$= \sum_{i \in \mathrm{SA}} \left[\frac{\frac{d^2 P(z_i|\theta)}{d\theta^2}}{P(z_i|\theta)} - \left(\frac{\frac{d P(z_i|\theta)}{d\theta}}{P(z_i|\theta)} \right)^2 \right]$$

$$+ \sum_{g=1}^{G} \frac{\int \mathrm{Exp}\left[\sum_{i \in g} \log\left(P\left(z_{ig}|\theta, u_g\right)\right)\right] \left(\sum_{i \in g} \frac{\frac{d P\left(z_{ig}|\theta, u_g\right)}{d\theta}}{P\left(z_{ig}|\theta, u_g\right)} \right)^2 N\left(u_g \middle| 0, \sigma_{u_g}^2\right) du_g}{\int \left\{ \mathrm{Exp}\left[\sum_{i \in g} \log\left(P\left(z_{ig}|\theta, u_g\right)\right)\right] N\left(u_g \middle| 0, \sigma_{u_g}^2\right) \right\} du_g}$$

$$+ \sum_{g=1}^{G} \frac{\int \mathrm{Exp}\left[\sum_{i \in g} \log\left(P\left(z_{ig}|\theta, u_g\right)\right)\right] \left(\sum_{i \in g} \left[\frac{\frac{d^2 P\left(z_{ig}|\theta, u_g\right)}{d\theta^2}}{P\left(z_{ig}|\theta, u_g\right)} - \left(\frac{\frac{d P\left(z_{ig}|\theta, u_g\right)}{d\theta}}{P\left(z_{ig}|\theta, u_g\right)} \right)^2 \right] \right) N\left(u_g \middle| 0, \sigma_{u_g}^2\right) du_g}{\int \left\{ \mathrm{Exp}\left[\sum_{i \in g} \log\left(P\left(z_{ig}|\theta, u_g\right)\right)\right] N\left(u_g \middle| 0, \sigma_{u_g}^2\right) \right\} du_g}$$

$$- \sum_{g=1}^{G} \left\{ \frac{\int \mathrm{Exp}\left[\sum_{i \in g} \log\left(P\left(z_{ig}|\theta, u_g\right)\right)\right] \left(\sum_{i \in g} \frac{\frac{d P\left(z_{ig}|\theta, u_g\right)}{d\theta}}{P\left(z_{ig}|\theta, u_g\right)} \right) N\left(u_g \middle| 0, \sigma_{u_g}^2\right) du_g}{\int \left\{ \mathrm{Exp}\left[\sum_{i \in g} \log\left(P\left(z_{ig}|\theta, u_g\right)\right)\right] N\left(u_g \middle| 0, \sigma_{u_g}^2\right) \right\} du_g} \right\}^2$$

Based on the above equations, we need only to define the ratios of the first and second derivatives of the item response probabilities with respect to θ to the response probabilities. For the Rasch testlet model, these are obtained as

$$p_i = P(z_i = 1|\theta) = \frac{\mathrm{Exp}(\theta - b_i)}{1 + \mathrm{Exp}(\theta - b_i)}, q_i = P(z_i = 0|\theta) = 1 - p_i,$$

and

$$p_{ig} = P\left(z_{ig} = 1|\theta, u_g\right) = \frac{\mathrm{Exp}\left(\theta + u_g - b_i\right)}{1 + \mathrm{Exp}\left(\theta + u_g - b_i\right)}, q_{ig} = P\left(z_{ig} = 0|\theta, u_g\right) = 1 - p_{ig}.$$

Therefore, we have,

$$\frac{\frac{dp_i}{d\theta}}{p_i} = q_i \ , \ \frac{\frac{dq_i}{d\theta}}{q_i} = -p_i,$$

$$\frac{\frac{dp_{ig}}{d\theta}}{p_{ig}} = q_{ig} \ , \ \frac{\frac{dq_{ig}}{d\theta}}{q_{ig}} = -p_{ig},$$

$$\frac{\frac{d^2 p_i}{d\theta^2}}{p_i} - \left(\frac{\frac{d p_i}{d\theta}}{p_i}\right)^2 = -p_i q_i,$$

$$\frac{\frac{d^2 q_i}{d\theta^2}}{q_i} - \left(\frac{\frac{d q_i}{d\theta}}{q_i}\right)^2 = -p_i q_i,$$

$$\frac{\frac{d^2 p_{ig}}{d\theta^2}}{p_{ig}} - \left(\frac{\frac{d p_{ig}}{d\theta}}{p_{ig}}\right)^2 = -p_{ig} q_{ig}, \text{ and}$$

$$\frac{\frac{d^2 q_{ig}}{d\theta^2}}{q_{ig}} - \left(\frac{\frac{d q_{ig}}{d\theta}}{q_{ig}}\right)^2 = -p_{ig} q_{ig}.$$

6.3 Extreme Case Handling

As with the MLE, the MMLE is not defined for zero and perfect scores. These cases are handled by assigning the lowest obtainable theta (LOT) scores and highest obtainable theta (HOT) scores, respectively. Table 20 contains the LOT and HOT values for each grade.

6.4 Standard Errors of Measurement

The standard error of measurement (SEM) of the MMLE score estimate is:

$$SEM(\hat{\theta}_{MMLE}) = \frac{1}{\sqrt{I(\hat{\theta}_{MMLE})}}$$

where $I(\hat{\theta}_{MMLE})$ is the observed information evaluated at $\hat{\theta}_{MMLE}$. The observed information is calculated as $I(\theta^2) = -\frac{d^2 l(\theta)}{d\theta^2}$, where $\frac{d^2 l(\theta)}{d\theta^2}$ is defined in the Section 6.2, Derivative. Note that the calculation of the standard error of estimate depends on the unique set of items that each student answers and their estimate of θ. Different students have different standard errors of measurement values, even if they have the same raw score and/or theta estimate. Standard errors are truncated at 1 for the overall science scores and truncated at 1.4 for the discipline scores.

Standard errors for MMLE estimates truncated at the LOT and HOT are computed by evaluating the observed information at the MMLE before truncation. For all incorrect or all correct answers, the reported standard errors are set at the truncation value for the standard error.

6.5 Scoring Incomplete Tests

The RI NGSA is assembled on-the-fly using a matrix design. For science, a test is considered "attempted" if a student responded to at least one item (cluster or stand-alone). An attempted test is considered complete if the student responds to all the operational items. Otherwise, the test is "incomplete".

Tests that are attempted but incomplete receive overall science scores. In order to receive a discipline score (i.e., Life Sciences, Physical Sciences, Earth and Space Sciences), a student must have attempted the corresponding discipline of the test. The MMLE is used to score the attempted incomplete tests, counting unanswered items as incorrect. If the identities of the unanswered items are unknown due to the test being assembled on the fly, the item parameters for a "typical" item are used. If a missing item is a cluster, the simulated item parameters of the missing item are the item parameters of item cluster 139 for grade 5, item cluster 119 for grade 8, and item cluster 345 for grade 11, which are operational clusters that are typical for the item bank used in RI NGSA in terms of the number of assertions and estimated parameters. Likewise, if a missing item is a stand-alone, the simulated item parameters of the missing item are the item parameters of stand-alone item 55 for grade 5, item 109 for grade 8, and item 171 for grade 11, which are operational stand-alone items that are typical for the item bank used in RI NGSA in terms of the number of assertions and estimated parameters.

If the identities of items that have not been answered to are known because they have already been lined up through the pre-fetch process, the item parameters of the lined-up items are used. Similarly, for the accommodated forms that are fixed forms, the item parameters of the unanswered items on the form are used.

6.6 STUDENT-LEVEL SCALE SCORE

At the student level, scale scores are computed for

1. Overall Science;
2. Life Sciences;
3. Physical Sciences;
4. Earth and Space Sciences.

Scores are computed using the MMLE method outlined in this report, with all items for overall science or only items within the given discipline. Scores are truncated on the "theta" scale at the LOT and HOT values specified in Table 20, which correspond to values of the estimated mean minus/plus four times the estimated standard deviation of θ.

The reporting scales will be a linear transformation of the theta scales:

$$SS = a * \hat{\theta}_{MMLE} + b$$

Where *a* and *b* are the slope and intercept of the linear transformation that transforms $\hat{\theta}_{MMLE}$ to the reporting scale (see Table 20). The standard error of estimate for the estimated scale score is obtained as:

$$SEM_{SS} = a * SEM_{\hat{\theta}_{MMLE}}.$$

In 2019, the reporting scale had a range of 120 points, from 1 to 120. The slope *a* and intercept *b* were chosen so that the center of the reporting scale of each grade (SS = 60) is centered at the proficiency cut and has a standard deviation of 15. Because a scale was required during standard setting, before the proficiency cut was known, the scale is established in two steps. In the first step,

the scale was established based on a tentative cut where 40% of the population would be proficient, corresponding to how proficiency cuts were set in New Hampshire and West Virginia across grades in 2018. Specifically, for grade 5, the slope *a* is obtained as:

$$\begin{aligned} SS &= 15\theta^* + b \\ &= 15\frac{\theta}{\hat{\sigma}_\theta} + b \\ &= a\theta + b, \end{aligned}$$

where the second line stems from transforming theta into a variable with a standard deviation of 1, $\theta^* = \frac{\theta}{\hat{\sigma}_\theta}$. Subsequently, the intercept *b* is obtained by equating the center of the scale (SS = 60) to the linear transformation of the tentative cut score on the theta scale,

$$SS = 60 = a\hat{\theta}_{tentative\_cut} + b$$

$$b = 60 - a\hat{\theta}_{tentative\_cut}$$

For grades 8 and 11, the slope and intercept can also be derived in a similar fashion.

After the 2019 standard setting, the final proficiency cut was set at 63 on the proposed scale for all three grades (detailed standard-setting results are presented in Volume 3 of this technical report). In order to center the reporting scale around the final cut, the scale was translated by minus 3, the difference between the tentative and final cuts expressed on the reporting scale. Per grade, Table 20 presents the intercept, slope, LOT, HOT, lowest obtainable scale score (LOSS), and highest obtainable scale score (HOSS) values that were used for the final reporting scale. The scale score distribution for overall science is reported in Appendix 1-C, Distribution of Scale Scores and Achievement Levels, and for the disciplines in Appendix 1-D, Distribution of Scale by Science Discipline.

Table 20. Reporting Scale Linear Transformation Constants and Theta and Corresponding Scaled-Score Limits for Extreme Ability Estimates (for 2021 θ scale)

Grade	Slope (*a*)	Intercept (*b*)	Lowest of Theta (LOT)	Highest of Theta (HOT)	Lowest of Scale Score (LOSS)	Highest of Scale Score (HOSS)
5	16.677	52.196	-3.06	4.06	1	120
8	17.001	53.266	-3.07	3.92	1	120
11	18.084	57.041	-3.09	3.48	1	120

6.7 RULES FOR CALCULATING ACHIEVEMENT LEVELS

Achievement levels and corresponding cut scores were set during standard setting in summer 2019. Students are classified into one of four achievement levels, based on their total score. The distribution of achievement levels is summarized in Appendix 1-C, Distribution of Scale Scores and Achievement Levels. Further, the distribution of scale scores and achievement levels for subgroups described in Section 4.4, Differential Item Functioning are presented in Appendix 1-F, Distribution of Scale Scores and Achievement Levels by Subgroup.

Table 21 lists the cut scores on the reporting scale metrics for each grade.

Table 21. Achievement-Level Cut Scores

Grade	Cut 1	Cut 2	Cut 3
5	37	60	72
8	38	60	74
11	36	60	71

6.7.1 Strengths and Weaknesses for Disciplines Relative to Proficiency Cut Score

Discipline-level classifications are computed to classify student achievement levels for each of the science disciplines. The classification rules are:

- if $(\hat{\theta}\hat{\theta}_{discipline} < \theta_{proficient} - 1.5 * SEM(\hat{\theta}_{discipline}))$, then achievement is classified as *Below Mastery*;
- if $(\theta_{proficient} - 1.5 * SEM(\hat{\theta}_{discipline}) \leq \hat{\theta}\hat{\theta}_{discipline} < \theta_{proficient} + 1.5 * SEM(\hat{\theta}_{discipline}))$, then achievement is classified as *At/Near Mastery*; and

- if $(\hat{\theta}\hat{\theta}_{discipline} \geq \theta_{proficient} + 1.5 * SEM(\hat{\theta}_{discipline}))$, then achievement is classified as *Above Mastery*,

where $\theta_{Proficient}$ is the proficiency cut score of the overall test. Standard errors are truncated at 1.4. The LOT is always classified as *Below Mastery*, and the HOT is always classified as *Above Mastery*.

6.8 Residual-Based Reporting at the Level of Disciplinary Core Ideas and Science and Engineering Practices

6.8.1 Relative to Overall Achievement

For aggregated units (classrooms, schools, and districts), there is reporting at levels below the science discipline level. In 2021–2022, reports were provided at the level of Disciplinary Core Ideas (DCI). The method for reporting at levels below the science discipline level is based on the use of residuals. The equations are presented first for DCIs.

For each assertion *i*, the residual between observed and expected score for each student *j* is defined as

$$\delta_{ij} = z_{ij} - E(z_{ij}).$$

The expected score is computed for a student's estimated overall ability. For the assertions clustered within an item, the expected score is marginalized over the nuisance dimensions for the assertions clustered within an item,

$$E(z_{ijg} = 1; \theta_{j,overall}, \tau_i) = \int P(z_{ijg} = 1|u_{jg};\ \theta_{j,overall}, \tau_i) N(u_{jg}) du_{jg},$$

where τ_i is the vector of parameters for assertion *i* (e.g., for the Rasch testlet model, $\tau_i = b_i$), and $P(z_{ijg} = 1|u_{jg};\ \theta_{j,overall}, \tau_i)$ is defined in Section 6.2, Derivative. Next, residuals are aggregated over assertions within each student,

$$\delta_{jDCI} = \frac{\sum_{i \in DCI} \delta_{ij}}{n_{jDCI}},$$

and over students of the group on which is reported,

$$\bar{\delta}_{DCIm} = \frac{1}{n_m} \sum_{j \in m} \delta_{jDCI},$$

where n_{jDCI} is the number of assertions related to the DCI for student *j,* and n_g is the number of students in a group assessed on the DCI. If a student did not see any items on a DCI, the student is not included in the n_g count for the aggregate. The standard error of the average residual is computed as

$$SEM(\bar{\delta}_{DCIm}) = \sqrt{\frac{1}{n_m(n_m-1)} \sum_{j \in \mathrm{m}} (\delta_{jDCI} - \bar{\delta}_{DCIm})^2}.$$

A statistically significant difference from zero in these aggregates is evidence that a class, teacher, school, or district is more effective (if $\bar{\delta}_{DCIm}$ is positive) or less effective (negative $\bar{\delta}_{DCIm}$) in teaching a given DCI.

We do not suggest the direct reporting of the statistic $\bar{\delta}_{DCIm}$; instead, we recommend reporting whether, in the aggregate, a group of students performs better, worse, or as expected on this DCI. In some cases, sufficient information is not available, and that will be indicated, as well.

For target-level strengths/weakness, the following is reported:

- If $\bar{\delta}_{DCIm} \leq -1.5 * SEM(\bar{\delta}_{DCIm})$, then achievement is *worse than* on the overall test.
- If $\bar{\delta}_{DCIm} \geq 1.5 * SEM(\bar{\delta}_{DCIm})$, then achievement is *better than* on the overall test.
- Otherwise, achievement is *similar to* the overall test.
- If $SEM(\bar{\delta}_{DCIm}) > 0.2$, data are insufficient.

6.8.2 Relative to Proficiency Cut Score

DCI-level scores for aggregated units can be computed using the same method as outlined in Section 6.8.1, Strengths and Weaknesses for Disciplines Relative to Proficiency Cut Score, but with the expected score computed at the theta value corresponding to the proficiency cut score:

$$E(z_{ijg} = 1; \theta_{proficiency}, \tau_i) = \int P(z_{ijg} = 1 | u_{jg};\ \theta_{proficiency}, \tau_i) N(u_{jg}) du_{jg}.$$

The following is reported for DCIs for aggregate units:

- If $\bar{\delta}_{DCIm} \leq -1.5 * SEM(\bar{\delta}_{DCIm})$, then achievement is *below* the proficiency cut score.
- If $\bar{\delta}_{DCIm} \geq 1.5 *\ SEM(\bar{\delta}_{DCIm})$, then achievement is *above* the proficiency cut score.
- Otherwise, achievement is *near* the proficiency cut score.
- If $SEM(\bar{\delta}_{DCIm}) > 0.2$, data are insufficient.

7. QUALITY CONTROL PROCEDURES

Cambium Assessment, Inc.'s (CAI) quality assurance (QA) procedures are built on two key principles: (1) automation and (2) replication. Certain procedures can be automated, which removes the potential for human error. Procedures that cannot be reasonably automated are replicated by two independent analysts at CAI.

Although the quality of any test is monitored as an ongoing activity, several sources of CAI's quality control system are described here. First, QA reports are routinely generated and evaluated throughout the testing window to ensure that each test is performing as anticipated. Second, the quality of scores is ensured by employing a second independent scoring verification system.

7.1 QUALITY ASSURANCE REPORTS

Test monitoring occurs while tests are administered in a live environment to ensure that item behavior is consistent with expectations. This is accomplished using CAI's Quality Monitor (QM) System that yields item statistics, blueprint match rates, and item exposure rate reports.

7.1.1 Item Analysis

The item analysis report is a key check for the early detection of potential problems with item scoring, including incorrect designation of a keyed response or other scoring errors, as well as potential breaches of test security that may be indicated by changes in the difficulty of test items. To examine the performance of test items, this report generates classical item analysis indicators of difficulty and discrimination, including proportion correct and biserial/polyserial correlation, as well as item fit statistics based on the item response theory (IRT) model. The report is configurable and can be produced to flag only items with statistics that fall outside a specified range or to generate reports based on all items in the pool.

7.1.2 Blueprint Match

As part of the QA procedures, Blueprint Match reports are generated at the content-standards level and for other content requirements such as strand and affinity groups for science. For each blueprint element, the report indicates the minimum and maximum number of items specified in the blueprint, the number of test administrations in which those specifications were met, the number of administrations in which the blueprint requirements were not met, and, for administrations in which specifications were not met, the number of items by which the requirement was not met.

In Spring 2023, every test in all three grades met the blueprint specifications at the level of the science disciplines, which is the lowest content level at which scores for individual students are reported. Blueprint match is discussed in detail in this technical report in Volume 2, Test Development, for both simulated and operational test administrations.

7.1.3 Item Exposure Rates

As part of the QA procedures, item exposure reports are generated, allowing test items to be monitored for unexpectedly large exposure rates or unusually low item-pool usage throughout the

testing window. As with other reports, it is possible to examine the exposure rate for all items or flag items with exposure rates that exceed an acceptable range. Often, item overexposure indicates a blueprint element or combination of blueprint elements that are underrepresented in the item pool and should be targeted for future item development. Such item overexposure is also usually anticipated in the simulation studies used to configure the adaptive algorithm.

In Spring 2023, most of the test items were administered to 20% or fewer test takers in all grades. Only 1.44 % of the items in grade 5, 0.94% of the items in grade 8, and 1.49% of the items in grade 11 were administered to 20% or more English test takers at that grade. More details are discussed in Volume 2, Test Development, of this technical report.

7.1.4 Cheating Detection Analysis

As part of the QA procedures, a forensics report can also be provided to identify possible irregularities in test administration for further investigation. Unusual patterns of responding at the student level can be aggregated to the test session, test administrator, and school levels to identify possible group-level testing anomalies. CAI psychometricians can monitor testing anomalies throughout the testing window. Evidence can be evaluated with respect to item response times and irregular item response patterns using the person-fit index. The flagging criteria used for these analyses are configurable and can be changed by the user. The analyses used to detect the testing anomalies can be run anytime within the testing window.

7.2 SCORING QUALITY CHECK

All student test scores are produced using CAI's scoring engine. A second score verification system is used to verify that all test scores match with 100% agreement in all tested grades. This second system is independently constructed and maintained from the main scoring engine and estimates scores separately using the procedures described within this report.

8. REFERENCES

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

Cai, L. (2017). flexMIRT®: Flexible multilevel multidimensional item analysis and test scoring (version 3.51) [computer software]. Chapel Hill, NC: Vector Psychometric Group.

Connecticut State Department of Education. (2019). *Validating American Institutes for Research's calibration and scoring processes for science assessments* (Research Report). Hartford, CT: Author.

Dorans, N. J., & Schmitt, A. P. (1991). *Constructed response and differential item functioning: A pragmatic approach* (ETS Research Report No. 91–47). Princeton, NJ: Educational Testing Service.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*(3), 423–436.

Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Koretz, D., & Hamilton, L. S. (2006). Testing for accountability in K–12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 531–578). Westport, CT: American Council on Education/Praeger.

National Center for Education Statistics. (2010). *Statistical methods for protecting personally identifiable information in aggregate reporting* (Statewide Longitudinal Data System Technical Brief, Brief 3). Retrieved from https://nces.ed.gov/pubs2011/2011603.pdf.

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

Rijmen, F. (2006). *BNL: A Matlab toolbox for Bayesian networks with logistic regression nodes* (Technical Report). Amsterdam: VU University Medical Center.

Rijmen, F. (2010). Formal relations and empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement*, *47*(3), 361–372.

Rijmen, F., Jiang, T., & Turhan, A. (2018, April). *An item response theory model for new science assessments*. Paper presented at the National Council on Measurement in Education, New York, NY.

Rijmen, F., Liao, D., & Lin, Z. (2021). *The Rasch testlet model for the calibration of three-dimensional science assessments: A software comparison* [White paper]. Washington, DC: Cambium Assessment, Inc.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28,* 237–247.

Somes, G. W. (1986). The generalized Mantel Haenszel statistic. *The American Statistician*, *40*, 106–108.

Wang, W. C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement*, *29*(2), 126–149.

Yen, W. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*, 187–213.

Zwick, R. (2012). *A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement* (ETS Research Report No. 12–08). Princeton, NJ: Educational Testing Service.

Appendix 1-A

Caveon Test Security Overview

Caveon Test Security Overview

TEST ADMINISTRATION SECURITY – CAVEON

The Cambium Assessment, Inc (CAI) utilizes the Caveon Web Patrol™ service to support test security compliance. Caveon is recognized as the only full-service test security organization that has national experience and expertise in this area. Caveon has been successfully providing Web Patrol monitoring services to influential clients since 2003 and has been delivering Web Patrol services on behalf of State Education Agencies since 2005. Caveon currently provides full-scope Web Patrol services in twenty-nine (29) states plus the WIDA consortium, the Smarter Balanced Assessment Consortium, and nearly fifty (50) certification and licensure programs.

By scouring the Internet and public-facing social media sites for breaches in test security, Caveon can systematically find and track threats to the testing program.

Web Patrol leverages the best of both automated technologies and the human capacity to judge and analyze. The result of this unique combination is a service that continually and systematically finds and tracks threats to the testing program.

DESCRIPTION

Caveon Web Patrol leverages technology tools and human expertise to identify, prioritize, and monitor websites, discussion forums, peer-to-peer servers, etc., where sensitive test information may be disclosed or at risk of disclosure.

Patrolling efforts routinely find and evaluate "brain-dumps" (websites where test questions have been posted, supposedly by individuals who memorized them and/or where disclosed test content may be inexpensively resold), test preparation training/education sites that may use actual (operational) test questions in the training, online auctions and classifieds such as eBay and Craigslist, and other social media channels and forums in which actual test items may be revealed or proxy test-takers offer their services. Regular update reports are generated that categorize identified threats by level of actual or potential risk to the testing program based on the representations made on the websites, or actual analysis of the proffered content. Websites and Internet extracts are ranked from CLEARED (Lowest risk but should be monitored) to SEVERE (Highest risk). The reports contain specific URLs and other content extractions that represent and depict the categorized threat. Additionally, the reports include overall and specific threat analysis, with actionable recommendations as well as any anticipated mitigation strategies for the Rhode Island State Department of Education (RIDE) to follow in minimizing and removing the dangers.

COMPREHENSIVE, CONSISTENT MONITORING

In conducting web patrol operations, Caveon utilizes a team of specialists who spend days and evenings continually trolling the Internet for intellectual property, the team leverages numerous search technologies, some licensed and some publicly accessible (e.g., "Open Source"), to ensure comprehensive, consistent, and continual monitoring of the web.

VERIFYING AND MANAGING THREATS

Casting such a broad net across the web means the team must cull through thousands of search results (each is a possible threat) and dig deeper to explore whether a result is benign or a legitimate worry. Team members have, after years of service, become experts at quickly reviewing a search hit and discerning a level of risk. Despite technology innovations in other aspects of the service, this work requires human judgment and is vitally necessary to take action against real threats to test security.

Once a threat is verified, CAI and Caveon coordinate with RIDE to systematically work through the steps necessary to have infringing content removed.

An escalation path of legal remedies is available. That path begins with formal "bystander" notifications and cease and desist letters. The path ends when the website operators remove copyrighted material and/or cease operations, either voluntarily or by compulsion. CAI endeavors to complement existing activities of RIDE, including issuing formal notices under existing U.S. copyright laws to offending website owners, ISPs, search engines, etc. Keys to successful threat removal include the following:

Timeliness of Notification

By continually, systematically patrolling for new threats and monitoring existing ones, Caveon Web Patrol quickly ascertains when a breach has or may occur. When a breach has been discovered, CAI will immediately notify the RIDE.

Assistance Taking Down Material

Immediate notification of dangerous threats to the testing program is only half the solution. With direction and support from the RIDE, CAI provides quick front line support through various means to take the next step, neutralizing the hazard. There are multiple options at RIDE disposal to help protect its IP. CAI has experience with:

- DMCA Takedown Request Letters

DMCA Takedown Request Letters can be sent immediately to website operators upon threat detection by Caveon. In most cases, simply alerting operators that copyrighted materials may be published on their websites is enough to get it removed.

Caveon Web Patrol service begins one week prior to the opening of the administration window and continues for one week after the test administration window closes.

Appendix 1-B

Shared Science Assessment Item Bank: Field Testing

TABLE OF CONTENTS

LIST OF TABLES

The Shared Science Assessment Item Bank is the product of a collaboration between Cambium Assessment, Inc. (CAI), multiple states and one U.S. territory that share a Memorandum of Understanding (MOU). Every participant of the MOU contributes items to the Shared Science Assessment Item Bank that underwent the same development process. The portion of the bank contributed by CAI is part of the Independent College and Career Readiness (ICCR) bank. Since the start of the 2017−2018 school year, items have been field-tested annually. This appendix describes how field tests were conducted from 2017−2018 through 2021−2022 across the states relying on the Shared Science Assessment Item Bank, or the ICCR portion thereof, for their three-dimensional science assessments.

1. 2018 FIELD TESTS

In 2018, a large pool of items was field tested in nine states. For three states (Hawaii, Oregon, and Wyoming), unscored field-test items were added as an additional segment to the operational (scored) legacy science test. Two other states (Connecticut and Rhode Island) conducted an independent field test in which all students participated and were administered a full set of items, but no scores were reported. In the remaining four states that field-tested items from the Shared Science Assessment Item Bank (New Hampshire, Utah, Vermont, and West Virginia), an operational field test was administered, meaning tests consisted of scored field-test items. Items became operational and were scored after the test administration if they were not rejected during rubric validation or item data review, as described later in this section. In total, 340 item clusters and 205 stand-alone items were administered in the elementary, middle, and high school grade bands. Table 1 presents the number of item clusters and stand-alone items administered in each grade band for each state.

Table 1. Number of Field-Test Items Administered, Spring 2018

Grade Band and Item Type	CT	HI	MSSA[a]	NH	OR	UT	WV	WY	Total
Elementary School	**135**	**24**	**69 (10)**	**58**	**26**	**–**	**91**	**14**	**153**
Cluster	78	13	40 (5)	34	20	–	56	6	86
Stand-Alone	57	11	29 (5)	24	6	–	35	8	67
Middle School	**174**	**27**	**56 (5)**	**55**	**28**	**98**	**123**	**17**	**241**
Cluster	115	13	26 (3)	30	22	98	90	5	171
Stand-Alone	59	14	30 (2)	25	6	–	33	12	70
High School	**149**	**23**	**75 (12)**	**60**	**38**	**–**	**–**	**14**	**151**
Cluster	81	14	34 (5)	33	30	–	–	6	83
Stand-Alone	68	9	41 (7)	27	8	–	–	8	68
Total	**458**	**74**	**200**	**173**	**92**	**98**	**214**	**45**	**545**

Note. The numbers in parentheses indicate MSSA-owned items.

For the states with a separate field-test segment (states with a legacy science test) and one of the states with an operational field test (Utah), fixed field-test forms were constructed (using a

balanced incomplete design except for Utah) and randomly assigned so that the group of students administered one form was comparable to the groups of students that were assigned other forms.

For the independent and operational field tests (except in Utah), including Rhode Island and Vermont, items were administered using a linear-on-the-fly (LOFT) test design in which items are selected on the fly, resulting in a unique test form for each student. The difference between the test design for the independent field tests and operational field tests depended on the test blueprint. The only blueprint constraint imposed on the independent field tests was that students received four stand-alone items and two item clusters for each of the three science disciplines. In contrast, a full blueprint was implemented for the states with an operational field test. The blueprint for the MSSA is discussed in Section **Error! Reference source not found.**, **Error! Reference source not found.**.

For any given state, there was a target of a minimum sample size of 1,500 students per item. Most items were administered in two or more states so that the item pools for all individual states were linked through common items. Approximately 98.3% of the items met or exceeded the target sample size of 1,500 in at least one state, while 98.8% of the items had a sample size of at least 1,350 (10% of the target) in at least one state. The common item design was used to calibrate all the items on a common science scale.

Table 2, Table 3, and Table 4 present the numbers of item clusters and stand-alone items that were in common between the item pools of any two states. The numbers below the shaded cell on a diagonal represent the number of common field-test items between any two states, and the numbers above the shaded cells represent the number of common items that survived rubric validation and were included in the 2018 calibration. In each of the shaded cells, the number outside the parentheses represents the number of unique items administered only in the given state, and the number provided in parentheses represents the number of unique and/or common items that were calibrated with the data from that state only. Table 2 presents the results for elementary school, Table 3 presents the results for middle school, and Table 4 presents the results for high school. The numbers at field-testing differ slightly from the numbers at calibration for various reasons, such as items not passing rubric validation and versioning issues for some items in some states.

Table 2. Number of Common Elementary School Field-Test Items Administered and Calibrated, Spring 2018

	State	CT	HI	MSSA[a]	NH	OR	UT	WV	WY
Cluster	CT	**3 (3)**	9	36	28	16	0	49	6
	HI	10	**0 (0)**	7	8	5	0	12	1
	MSSA	36	8	**0 (2)**	15	12	0	26	2
	NH	30	8	17	**1 (3)**	5	0	22	2
	OR	17	5	13	5	**1 (1)**	0	5	1
	UT	0	0	0	0	0	**0 (0)**	0	0
	WV	49	12	27	25	5	0	**0 (4)**	2
	WY	6	1	2	2	1	0	2	**0 (0)**

	State	CT	HI	MSSA[a]	NH	OR	UT	WV	WY
Stand-Alone	**CT**	**1 (3)**	5	25	22	2	0	33	7
	HI	5	**6 (6)**	0	0	0	0	4	0
	MSSA	26	0	**0 (1)**	10	4	0	13	3
	NH	24	0	11	**0 (2)**	0	0	15	2
	OR	2	0	4	0	**1 (1)**	0	0	0
	UT	0	0	0	0	0	**0 (0)**	0	0
	WV	35	4	14	17	0	0	**0 (2)**	1
	WY	8	0	3	3	0	0	2	**0 (1)**
Total	**CT**	**4 (6)**	14	61	50	18	0	82	13
	HI	15	**6 (6)**	7	8	5	0	16	1
	MSSA	62	8	**0 (3)**	25	16	0	39	5
	NH	54	8	28	**1 (5)**	5	0	37	4
	OR	19	5	17	5	**2 (2)**	0	5	1
	UT	0	0	0	0	0	**0 (0)**	0	0
	WV	84	16	41	42	5	0	**0 (6)**	3
	WY	14	1	5	5	1	0	4	**0 (1)**

[a]MSSA = Rhode Island and Vermont's Multi-State Science Assessment

Table 3. Number of Common Middle School Field-Test Items Administered and Calibrated, Spring 2018

	State	CT	HI	MSSA[a]	NH	OR	UT	WV	WY
Cluster	**CT**	**2 (6)**	12	22	26	19	44	77	5
	HI	11	**1 (0)**	3	6	6	0	9	1
	MSSA	23	3	**0 (1)**	9	1	7	22	2
	NH	26	6	10	**1 (2)**	7	0	17	3
	OR	19	6	1	7	**2 (2)**	0	5	1
	UT	48	0	7	0	0	**48 (52)**	43	0
	WV	83	10	21	18	6	48	**1 (9)**	2
	WY	5	1	2	3	1	0	2	**0 (0)**
Stand-Alone	**CT**	**2 (3)**	6	27	25	3	0	33	12
	HI	6	**8 (8)**	2	0	0	0	2	0
	MSSA	27	2	**0 (0)**	18	3	0	20	2
	NH	25	0	18	**0 (0)**	0	0	21	3
	OR	3	0	3	0	**0 (0)**	0	0	0
	UT	0	0	0	0	0	**0 (0)**	0	0
	WV	33	2	20	21	0	0	**0 (0)**	2
	WY	12	0	2	3	0	0	2	**0 (0)**
Total	**CT**	**4 (9)**	18	49	51	22	44	110	17
	HI	17	**9 (8)**	5	6	6	0	11	1
	MSSA	50	5	**0 (1)**	27	4	7	42	4
	NH	51	6	28	**1 (2)**	7	0	38	6
	OR	22	6	4	7	**2 (2)**	0	5	1
	UT	48	0	7	0	0	**48 (52)**	43	0
	WV	116	12	41	39	6	48	**1 (9)**	4
	WY	17	1	4	6	1	0	4	**0 (0)**

[a]MSSA = Rhode Island and Vermont's Multi-State Science Assessment

Table 4. Number of Common High School Field-Test Items Administered and Calibrated, Spring 2018

	State	CT	HI	MSSA[a]	NH	OR	UT	WV	WY
Cluster	CT	**10 (16)**	13	30	29	30	0	0	5
	HI	13	**0 (0)**	7	7	8	0	0	1
	MSSA	32	7	**0 (2)**	13	12	0	0	1
	NH	32	7	14	**0 (3)**	12	0	0	3
	OR	30	8	12	12	**0 (0)**	0	0	1
	UT	0	0	0	0	0	**0 (0)**	0	0
	WV	0	0	0	0	0	0	**0 (0)**	0
	WY	6	1	1	3	1	0	0	**0 (1)**
Stand-Alone	CT	**4 (4)**	9	40	27	8	0	0	8
	HI	9	**0 (0)**	4	0	0	0	0	0
	MSSA	39	4	**0 (1)**	20	3	0	0	1
	NH	25	0	20	**0 (0)**	0	0	0	1
	OR	8	0	3	0	**0 (0)**	0	0	0
	UT	0	0	0	0	0	**0 (0)**	0	0
	WV	0	0	0	0	0	0	**0 (0)**	0
	WY	7	0	1	1	0	0	0	**0 (0)**
Total	CT	**14 (20)**	22	70	56	38	0	0	13
	HI	22	**0 (0)**	11	7	8	0	0	1
	MSSA	71	11	**0 (3)**	33	15	0	0	2
	NH	57	7	34	**0 (3)**	12	0	0	4
	OR	38	8	15	12	**0 (0)**	0	0	1
	UT	0	0	0	0	0	**0 (0)**	0	0
	WV	0	0	0	0	0	0	**0 (0)**	0
	WY	13	1	2	4	1	0	0	**0 (1)**

[a]MSSA = Rhode Island and Vermont's Multi-State Science Assessment

Following the (operational) field test administration, items went through rubric validation and item data review, as described in Volume 2, Section, 2.7.1, Rubric Validation, and Section 2.7.2, Data Review.

Table 5 presents the number of items field tested in Rhode Island and Vermont, the number of items that were rejected before or during rubric validation, the number of items that were sent out for data review, and the number of items that were rejected during data review. The numbers in parentheses present the number of items owned by Rhode Island and Vermont.

Table 5. Number of Field-Test Item Administration, Rubric Validation, Item Data Review, Spring 2018

Grade Band and Item Type	Number of Field-Test Items Administered	Number of Items Rejected Before/During Rubric Validation	Number of Items Sent to Data Review	Number of Items Rejected at Data Review[a]	Number of Items Remaining
Elementary School	**153 (10)**	**3 (0)**	**65 (4)**	**13 (3)**	**137 (7)**
Cluster	86 (5)	3 (0)	24 (1)	5 (0)	78 (5)
Stand-Alone	67 (5)	0 (0)	41 (3)	8 (3)	59 (2)
Middle School	**241 (5)**	**16 (0)**	**102 (0)**	**24 (0)**	**201 (5)**
Cluster	171 (3)	12 (0)	65 (0)	15 (0)	144 (3)
Stand-Alone	70 (2)	4 (0)	37 (0)	9 (0)	57 (2)
High School	**151 (12)**	**10 (2)**	**80 (6)**	**13 (3)**	**128 (7)**
Cluster	83 (5)	8 (2)	35 (1)	4 (0)	71 (3)
Stand-Alone	68 (7)	2 (0)	45 (5)	9 (3)	57 (4)
Total	**545 (27)**	**29 (2)**	**247 (10)**	**50 (6)**	**466 (19)**

Note. MSSA-owned items are indicated in the parentheses.
[a]Including three middle school clusters rejected after item data review.

Table 6 summarizes the operational Shared Science Assessment Item Bank for each of the three science disciplines after adding the 2018 field-test items that passed rubric validation and item data review. The numbers in parentheses present the number of items owned by MSSA.

Table 6. Summary of Shared Science Assessment Item Bank, Spring 2018

Grade Band and Item Type	Science Discipline			Total[a]
	Earth and Space Sciences	*Life Sciences*	*Physical Sciences*	
Elementary School	**41 (2)**	**47 (3)**	**49 (2)**	**137 (7)**
Cluster	23 (1)	29 (2)	26 (2)	78 (5)
Stand-Alone	18 (1)	18 (1)	23 (0)	59 (2)
Middle School	**56 (1)**	**72 (2)**	**70 (2)**	**198 (5)**
Cluster	41 (1)	49 (1)	51 (1)	141 (3)
Stand-Alone	15 (0)	23 (1)	19 (1)	57 (2)
High School	**37 (4)**	**53 (1)**	**38 (2)**	**128 (7)**
Cluster	19 (2)	32 (0)	20 (1)	71 (3)
Stand-Alone	18 (2)	21 (1)	18 (1)	57 (4)
Total	**134 (7)**	**172 (6)**	**157 (6)**	**463 (19)**

[a]Totals exclude three Utah-owned middle school clusters that do not align to the NGSS.

2. 2019 FIELD TESTS

In 2019, a second wave of items was field tested in nine states. For three states (Hawaii, Idaho [elementary school only], 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 the remaining six states (Connecticut, New Hampshire, Oregon, Rhode Island, Vermont, and West Virginia), field-test items were administered as unscored items embedded among the operational items. In total, 123 item clusters and 224 stand-alone items were administered as field-test items in the elementary, middle, and high school grade bands. Table 7 presents the number of field-test item clusters and stand-alone items administered in each grade band for each state. The numbers in parentheses in the column representing MSSA indicate the number of items owned by MSSA.

Table 7. Number of Field-Tested Items Administered in Spring 2019

Grade Band and Item Type	CT	HI	ID	MSSA[a]	NH	OR	WV	WY	Total
Elementary School	**47**	**31**	**53**	**42 (10)**	**18**	**27**	**18**	**16**	**117**
Cluster	18	19	20	17 (4)	0	16	10	5	50
Stand-Alone	29	12	33	25 (6)	18	11	8	11	67
Middle School	**56**	**23**	**53**	**46 (8)**	**28**	**26**	**26**	**15**	**127**

Grade Band and Item Type	CT	HI	ID	MSSA[a]	NH	OR	WV	WY	Total
Cluster	14	9	17	10 (3)	4	9	8	5	38
Stand-Alone	42	14	36	36 (5)	24	17	18	10	89
High School	**69**	**21**	**–**	**37 (6)**	**29**	**28**	**–**	**25**	**103**
Cluster	25	14	–	18 (3)	2	13	–	2	35
Stand-Alone	44	7	–	19 (3)	27	15	–	23	68
Total	**172**	**75**	**106**	**125 (24)**	**75**	**81**	**44**	**56**	**347**

Note. MSSA-owned items are indicated in parentheses.

For the three states with a separate field-test segment (i.e., states with a legacy science test), field-test forms were constructed using a balanced incomplete design and randomly assigned so that the group of students administered one form was comparable to the groups of students that were assigned other forms. For the independent field test, items were administered under a LOFT design, where the only blueprint constraint imposed was that students received four stand-alone items and two item clusters for each of the three science disciplines.

In three states with an operational test, field-test items were embedded within the operational test. Some of the states with an operational test (New Hampshire, Rhode Island, and Vermont) opted for a test in which operational items were grouped by science discipline. For these three states, the field-test items were presented together in a fourth group of items. The sequence of the four sets of items (corresponding to the three disciplines and a set of field-test items) was randomized across students Three other states (Connecticut, Oregon, and West Virginia) opted for a test design in which the items were not grouped by discipline. In these three states, field-test items were administered at random positions throughout the test. A student received either a field-test item cluster or a set of five field-test stand-alone items. The test design for the MSSA is discussed in Section **Error! Reference source not found.**, **Error! Reference source not found.**, of this volume.

A minimum sample size of 1,500 students per field-test item was targeted for any given state. Most items were administered in two or more states. Approximately 88.8% of the items met or exceeded the target sample size of 1,500 in at least one state, while 96.4% of the items had a sample size of at least 1,350 (10% of the target) in at least one state.

Table 8 to Table 10 present the numbers of cluster items and stand-alone items that were shared between the field-test pools of any two states. The numbers below the shaded cell on a diagonal represent the number of common field-test items between any two states, and the numbers above the shaded cells represent the number of common items that survived rubric validation and were included in the 2018 calibration. In each of the shaded cells, the number outside the parentheses represents the number of unique field-test items administered only in the given state, and the number provided in parentheses represents the number of unique and/or common items that were calibrated with only the data from that state.

Table 8 presents the results for elementary schools, Table 9 presents the results for middle schools, and Table 10 presents the results for high schools. The numbers at field testing are slightly different from the numbers at calibration because some items were rejected during rubric validation.

Table 8. Number of Common Elementary School Field-Test Items Administered and Calibrated, Spring 2019

	State	CT	HI	ID	MSSA[a]	NH	OR	WV	WY
Cluster	**CT**	**2 (2)**	2	10	3	0	2	1	4
	HI	2	**0 (0)**	3	8	0	14	2	0
	ID	10	3	**4 (4)**	0	0	1	3	3
	MSSA	3	8	0	**3 (3)**	0	9	4	1
	NH	0	0	0	0	**0 (0)**	0	0	0
	OR	2	14	1	9	0	**1 (1)**	0	0
	WV	1	2	3	4	0	0	**1 (0)**	1
	WY	4	0	3	1	0	0	1	**0 (0)**
Stand-Alone	**CT**	**5 (5)**	1	13	1	9	0	0	2
	HI	1	**0 (0)**	10	6	0	6	0	0
	ID	13	11	**1 (1)**	12	1	9	2	4
	MSSA	1	7	13	**3 (3)**	5	8	5	6
	NH	9	0	1	5	**2 (3)**	0	0	6
	OR	0	7	10	9	0	**1 (1)**	0	0
	WV	0	0	2	5	0	0	**1 (1)**	0
	WY	2	0	4	6	7	0	0	**0 (0)**
Total	**CT**	**7 (7)**	3	23	4	9	2	1	6
	HI	3	**0 (0)**	13	14	0	20	2	0
	ID	23	14	**5 (5)**	12	1	10	5	7
	MSSA	4	15	13	**6 (6)**	5	17	9	7
	NH	9	0	1	5	**2 (3)**	0	0	6
	OR	2	21	11	18	0	**2 (2)**	0	0
	WV	1	2	5	9	0	0	**2 (1)**	1
	WY	6	0	7	7	7	0	1	**0 (0)**

[a]MSSA = Rhode Island and Vermont's Multi-State Science Assessment

Table 9. Number of Common Middle School Field-Test Items Administered and Calibrated, Spring 2019

	State	CT	HI	ID	MSSA[a]	NH	OR	WV	WY
Cluster	**CT**	**5 (5)**	3	4	2	0	2	1	0
	HI	3	**0 (0)**	4	4	0	5	1	0
	ID	4	4	**2 (2)**	4	0	4	3	3
	MSSA	2	4	4	**1 (1)**	0	2	3	1
	NH	0	0	1	0	**3 (0)**	0	0	0
	OR	2	5	4	2	0	**1 (1)**	1	2
	WV	1	1	3	3	0	1	**0 (0)**	2
	WY	0	0	3	1	0	2	2	**0 (0)**
Stand-Alone	**CT**	**10 (9)**	2	13	9	10	3	6	0
	HI	2	**0 (0)**	9	9	0	6	3	0
	ID	13	9	**2 (2)**	11	1	12	6	5
	MSSA	9	9	11	**1 (1)**	6	11	9	7
	NH	10	0	2	6	**3 (1)**	0	0	2
	OR	3	6	12	11	0	**0 (0)**	2	7
	WV	6	3	6	9	1	2	**0 (0)**	0
	WY	0	0	5	7	2	7	0	**0 (0)**
Total	**CT**	**15 (14)**	5	17	11	10	5	7	0
	HI	5	**0 (0)**	13	13	0	11	4	0
	ID	17	13	**4 (4)**	15	1	16	9	8
	MSSA	11	13	15	**2 (2)**	6	13	12	8
	NH	10	0	3	6	**6 (1)**	0	0	2
	OR	5	11	16	13	0	**1 (1)**	3	9
	WV	7	4	9	12	1	3	**0 (0)**	2
	WY	0	0	8	8	2	9	2	**0 (0)**

[a]MSSA = Rhode Island and Vermont's Multi-State Science Assessment

Table 10. Number of Common High School Field-Test Items Administered and Calibrated, Spring 2019

	State	CT	HI	ID	MSSA[a]	NH	OR	WV	WY
Cluster	CT	**9 (9)**	10	-	11	0	8	-	1
	HI	11	**0 (0)**	-	8	0	11	-	0
	ID	-	-	-	-	-	-	-	-
	MSSA	12	9	-	**3 (2)**	0	7	-	2
	NH	0	0	-	0	**1 (0)**	1	-	0
	OR	8	11	-	7	1	**1 (1)**	-	0
	WV	-	-	-	-	-	-	-	-
	WY	1	0	-	2	0	0	-	**0 (0)**
Stand-Alone	CT	**14 (13)**	7	-	7	6	13	-	13
	HI	7	**0 (0)**	-	0	0	6	-	0
	ID	-	-	-	-	-	-	-	-
	MSSA	8	0	-	**3 (3)**	6	5	-	12
	NH	8	0	-	6	**10 (10)**	0	-	7
	OR	14	6	-	6	0	**0 (1)**	-	8
	WV	-	-	-	-	-	-	-	-
	WY	14	0	-	13	7	9	-	**0 (0)**
Total	CT	**23 (22)**	17	-	18	6	21	-	14
	HI	18	**0 (0)**	-	8	0	17	-	0
	ID	-	-	-	-	-	-	-	-
	MSSA	20	9	-	**6 (5)**	6	12	-	14
	NH	8	0	-	6	**11 (10)**	1	-	7
	OR	22	17	-	13	1	**1 (1)**	-	8
	WV	-	-	-	-	-	-	-	-
	WY	15	0	-	15	7	9	-	**0 (0)**

[a]MSSA = Rhode Island and Vermont's Multi-State Science Assessment

Following the operational field test administration, items underwent rubric validation and item data review, as described in Volume 2, Section, 2.7.1, Rubric Validation, and Section 2.7.2, Data Review. Table 11 presents the number of items field-test items administered in Rhode Island and Vermont (or another state), the number of items rejected before or during rubric validation, the number of items sent out to data review, and the number of items rejected during data review. The numbers in parentheses present the number of items owned by Rhode Island and Vermont.

Table 11. Number of Field-Test Item Administration, Rubric Validation, and Item Data Review in Spring 2019

Grade Band and Item Type	Number of Items Field Tested	Number of Items Rejected Before/During Rubric Validation	Number of Items Sent to Data Review	Number of Items Rejected at Data Review	Number of Items Remaining[a]
Elementary School	**117 (10)**	**2 (0)**	**72 (5)**	**24 (0)**	**91 (10)**
Clusters	50 (4)	1 (0)	16 (0)	10 (0)	39 (4)
Stand-Alone	67 (6)	1 (0)	56 (5)	14 (0)	52 (6)
Middle School	**127 (8)**	**6 (0)**	**66 (5)**	**21 (2)**	**97 (6)**
Clusters	38 (3)	1 (0)	12 (1)	5 (1)	29 (2)
Stand-Alone	89 (5)	5 (0)	54 (4)	16 (1)	68 (4)
High School	**103 (6)**	**6 (0)**	**52 (4)**	**15 (2)**	**80 (3)**
Clusters	35 (3)	2 (1)	15 (1)	5 (0)	26 (2)
Stand-Alone	68 (3)	4 (0)	37 (3)	10 (2)	54 (1)
Total	**347 (24)**	**14 (0)**	**190 (14)**	**60 (4)**	**268 (19)**

Note. MSSA-owned items are indicated in parentheses.
[a]Number of items remaining excludes five AI scoring items (four ICCR and one MSSA-owned) field tested in spring 2019 that were not brought to item data review.

Table 12 summarizes the Shared Science Assessment Item Bank after adding the field-test items that were administered in 2019 and passed rubric validation and item data review The numbers in parentheses present the number of items owned by Rhode Island and Vermont.

Table 12. Summary of Shared Science Assessment Item Bank in Spring 2019

Grade Band and Item Type	Science Discipline			Total[a]
	Earth and Space Sciences	*Life Sciences*	*Physical Sciences*	
Elementary School	**68 (7)**	**225 (17)**	**80 (4)**	**225 (17)**
Cluster	33 (3)	113 (9)	40 (3)	113 (9)
Stand-Alone	35 (4)	112 (8)	40 (1)	112 (8)
Middle School	**83 (2)**	**287 (11)**	**92 (4)**	**287 (11)**
Cluster	44 (1)	163 (5)	53 (2)	163 (5)
Stand-Alone	39 (1)	124 (6)	39 (2)	124 (6)
High School	**39 (4)**	**200 (9)**	**53 (3)**	**200 (9)**
Cluster	18 (2)	90 (4)	24 (1)	90 (4)
Stand-Alone	21 (2)	110 (5)	29 (2)	110 (5)
Total	**190 (13)**	**712 (37)**	**225 (11)**	**712 (37)**

Note. MSSA-owned items are indicated in the parentheses.

3. 2021 FIELD TESTS

In 2021, a third wave of items was field tested in 12 states. For one state (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 in Idaho and Montana. In the remaining nine states (Connecticut, Hawaii, New Hampshire, North Dakota, Rhode Island, South Dakota, Vermont, Utah, and West Virginia), field-test items were administered as unscored items embedded among the operational items. In total, 223 item clusters and 322 stand-alone items were administered as field-test items in the elementary, middle, and high school grade bands. Table 13 presents the number of field-test item clusters and stand-alone items administered in each grade band for each state. The numbers in parentheses in the column representing MSSA presents the number of field-test items owned by MSSA.

Table 13. Number of Field-Test Items Administered in Spring 2021

Grade Band and Item Type	CT	HI	ID	MSSA[a]	MT	ND	NH	SD	UT	WV	WY	Total
Elementary School	**36**	**22**	**140**	**55 (7)**	**21**	**11**	**19**	**8**	**54**	**19**	**17**	**214**
Cluster	16	6	58	18 (4)	7	3	3	3	54	7	5	106
Stand-Alone	20	16	82	37 (3)	14	8	16	5	0	12	12	108
Middle School	**33**	**19**	**129**	**54 (12)**	**20**	**11**	**18**	**11**	**45**	**19**	**20**	**159**
Cluster	17	6	44	18 (6)	7	3	2	2	45	7	4	60
Stand-Alone	16	13	85	36 (6)	13	8	16	9	0	12	16	99
High School	**49**	**17**	**156**	**49 (7)**	**0**	**11**	**12**	**8**	**0**	**0**	**20**	**172**
Cluster	11	5	54	16 (2)	0	3	4	3	0	0	3	57
Stand-Alone	38	12	102	33 (5)	0	8	8	5	0	0	17	115
Total	**118**	**58**	**425**	**158 (26)**	**41**	**33**	**49**	**27**	**99**	**38**	**57**	**545**

Note. MSSA-owned items are indicated in the parentheses.

For the state with a separate field-test segment (i.e., Wyoming), field-test forms were constructed using a balanced incomplete design and randomly assigned so that the group of students administered one form was comparable to the groups of students that were assigned other forms. For the independent field test, items were administered under a LOFT design, where the only blueprint constraint imposed was that students received four stand-alone items and two item clusters for each of the three science disciplines.

For the states with an operational test, field-test items were embedded within the test. Three states with an operational test (New Hampshire, Rhode Island, and Vermont) chose to administer a test in which operational items were grouped by science discipline. For these states, the field-test items were presented together as a fourth group of items. The sequence of the four sets of items (corresponding to the three disciplines and a set of field-test items) was randomized across students. Six other states—Connecticut, Hawaii, North Dakota, South Dakota, Utah, and West Virginia—opted for a test design in which the items were not grouped by discipline. In these states, field-test items were administered at random positions throughout the test. A student received either a field-test item cluster or a set of four field-test stand-alone items. The test design for the MSSA is discussed in Section **Error! Reference source not found.**, **Error! Reference source not found.**.

A minimum sample size of 1,500 students per field-test item was targeted for any given state. Most items were administered in two or more states. Approximately 96.7% of the items met or exceeded the target sample size of 1,500 in at least one state, while 99.1% of the items had a sample size of at least 1,350 (10% of the target) in at least one state.

Table 14 to Table 16 present the number of item clusters and stand-alone items that were shared between the field-test pools of any two states. The numbers below the shaded cell on a diagonal represent the number of common field-test items between any two states, and the numbers above the shaded cells represent the number of common items that survived rubric validation and were included in the 2018 calibration. In each of the shaded cells, the number outside the parentheses represents the number of unique field-test items administered only in the given state, and the number in parentheses represents the number of unique and/or common items that were calibrated with only the data from that state. Table 14 presents the results for elementary schools, Table 15 presents the results for middle schools, and Table 16 presents the results for high schools. The numbers of field-test items administered are slightly different from the numbers of field-test items at calibration because some items were rejected during rubric validation.

Table 14. Number of Common Elementary School Field-Test Items Administered and Calibrated, Spring 2021

	State	CT	HI	ID	MSSA[a]	MT	ND	NH	SD	UT	WV	WY
Item Clusters	CT	3 (3)	0	13	0	0	0	0	0	0	0	0
	HI	0	1 (1)	3	0	0	0	0	0	0	1	0
	ID	13	4	3 (2)	5	5	2	0	2	20	1	4
	MSSA	0	0	6	2 (2)	2	0	0	0	7	0	0
	MT	0	0	5	2	0 (0)	0	0	0	0	0	0
	ND	0	0	2	0	0	0 (0)	0	1	0	1	0
	NH	0	0	0	0	0	0	0 (0)	0	0	3	0
	SD	0	0	2	0	0	1	0	0 (0)	0	2	0
	UT	0	0	20	8	0	0	0	0	25 (24)	0	2
	WV	0	1	1	0	0	1	3	2	0	1 (1)	0
	WY	0	0	4	0	0	0	0	0	2	0	0 (0)
Stand-Alone Items	CT	3 (3)	0	14	2	0	0	0	0	0	0	1
	HI	0	0 (0)	12	1	0	0	2	3	0	1	0
	ID	14	12	3 (3)	30	13	4	3	3	0	4	9
	MSSA	2	1	30	0 (0)	12	0	3	1	0	0	0
	MT	0	0	13	12	0 (0)	0	0	0	0	0	0
	ND	0	0	4	0	0	0 (0)	2	0	0	0	1
	NH	0	2	4	3	0	2	0 (0)	2	0	3	1
	SD	0	3	3	1	0	0	2	0 (0)	0	0	0
	UT	0	0	0	0	0	0	0	0	0 (0)	0	0
	WV	0	1	4	0	0	1	3	0	0	3 (3)	0
	WY	1	0	9	0	0	1	1	0	0	0	0 (0)
Total	CT	6 (6)	0	27	2	0	0	0	0	0	0	1

	State	CT	HI	ID	MSSA[a]	MT	ND	NH	SD	UT	WV	WY
	HI	0	1 (1)	15	1	0	0	2	3	0	2	0
	ID	27	16	6 (5)	35	18	6	3	5	20	5	13
	MSSA	2	1	36	2 (2)	14	0	3	1	7	0	0
	MT	0	0	18	14	0 (0)	0	0	0	0	0	0
	ND	0	0	6	0	0	0 (0)	2	1	0	1	1
	NH	0	2	4	3	0	2	0 (0)	2	0	6	1
	SD	0	3	5	1	0	1	2	0 (0)	0	2	0
	UT	0	0	20	8	0	0	0	0	25 (24)	0	2
	WV	0	2	5	0	0	2	6	2	0	4 (4)	0
	WY	1	0	13	0	0	1	1	0	2	0	0 (0)

[a]MSSA = Rhode Island and Vermont's Multi-State Science Assessment

Table 15. Number of Common Middle School Field-Test Items Administered and Calibrated, Spring 2021

	State	CT	HI	ID	MSSA[a]	MT	ND	NH	SD	UT	WV	WY
Item Clusters	CT	0 (0)	0	9	2	0	0	0	0	10	0	0
	HI	0	0 (0)	2	3	0	0	0	0	3	1	0
	ID	11	2	1 (1)	10	6	2	1	1	31	0	4
	MSSA	4	3	11	0 (0)	0	2	0	0	9	1	1
	MT	0	0	6	0	1 (1)	0	1	1	4	0	0
	ND	0	0	3	2	0	0 (0)	0	0	2	0	0
	NH	0	0	1	0	1	0	0 (0)	1	0	1	0
	SD	0	0	1	0	1	0	1	0 (0)	0	0	0
	UT	14	3	36	11	4	3	0	1	0 (0)	2	2
	WV	0	1	1	1	0	0	1	1	5	0 (0)	0
	WY	0	0	4	1	0	0	0	0	2	0	0 (0)
Stand-Alone Items	CT	2 (2)	0	12	2	0	0	0	3	0	0	2
	HI	0	0 (0)	10	1	0	0	0	0	0	2	0
	ID	13	10	2 (2)	29	10	6	12	7	0	5	15
	MSSA	2	1	29	0 (0)	10	2	1	1	0	2	4
	MT	0	0	12	10	0 (0)	0	0	0	0	0	0
	ND	0	0	7	2	0	0 (0)	1	0	0	0	0
	NH	0	0	12	1	0	1	0 (0)	2	0	1	3
	SD	3	0	7	1	0	0	2	0 (0)	0	3	4
	UT	0	0	0	0	0	0	0	0	0 (0)	0	0
	WV	0	2	6	3	0	1	1	3	0	0 (0)	0
	WY	2	0	15	4	0	0	3	4	0	0	0 (0)
Total	CT	2 (2)	0	21	4	0	0	0	3	10	0	2

	State	CT	HI	ID	MSSA[a]	MT	ND	NH	SD	UT	WV	WY
	HI	0	0 (0)	12	4	0	0	0	0	3	3	0
	ID	24	12	3 (3)	39	16	8	13	8	31	5	19
	MSSA	6	4	40	0 (0)	10	4	1	1	9	3	5
	MT	0	0	18	10	1 (1)	0	1	1	4	0	0
	ND	0	0	10	4	0	0 (0)	1	0	2	0	0
	NH	0	0	13	1	1	1	0 (0)	3	0	2	3
	SD	3	0	8	1	1	0	3	0 (0)	0	3	4
	UT	14	3	36	11	4	3	0	1	0 (0)	2	2
	WV	0	3	7	4	0	1	2	4	5	0 (0)	0
	WY	2	0	19	5	0	0	3	4	2	0	0 (0)

[a]MSSA = Rhode Island and Vermont's Multi-State Science Assessment.

Table 16. Number of Common High School Field-Test Items Administered and Calibrated, Spring 2021

	State	CT	HI	ID	MSSA[a]	MT	ND	NH	SD	UT	WV	WY
Item Clusters	CT	1 (1)	0	8	0	0	0	0	0	0	0	0
	HI	0	0 (0)	5	0	0	0	0	0	0	0	0
	ID	10	5	16 (15)	12	0	2	2	3	0	0	3
	MSSA	0	0	15	0 (0)	0	0	1	2	0	0	0
	MT	0	0	0	0	0 (0)	0	0	0	0	0	0
	ND	0	0	2	0	0	0 (0)	1	0	0	0	0
	NH	0	0	2	1	0	1	0 (0)	0	0	0	0
	SD	0	0	3	2	0	0	0	0 (0)	0	0	0
	UT	0	0	0	0	0	0	0	0	0 (0)	0	0
	WV	0	0	0	0	0	0	0	0	0	0 (0)	0
	WY	0	0	3	0	0	0	0	0	0	0	0 (0)
Stand-Alone Items	CT	3 (3)	0	31	3	0	0	0	0	0	0	1
	HI	0	0 (0)	11	1	0	0	0	0	0	0	0
	ID	31	11	9 (8)	24	0	7	4	5	0	0	14
	MSSA	3	1	25	0 (0)	0	0	3	4	0	0	1
	MT	0	0	0	0	0 (0)	0	0	0	0	0	0
	ND	0	0	7	0	0	0 (0)	1	0	0	0	0
	NH	0	0	4	3	0	1	0 (0)	0	0	0	0
	SD	0	0	5	4	0	0	0	0 (0)	0	0	1
	UT	0	0	0	0	0	0	0	0	0 (0)	0	0
	WV	0	0	0	0	0	0	0	0	0	0 (0)	0
	WY	1	0	15	1	0	0	0	1	0	0	0 (0)

	State	CT	HI	ID	MSSA[a]	MT	ND	NH	SD	UT	WV	WY
Total	CT	4 (4)	0	39	3	0	0	0	0	0	0	1
	HI	0	0 (0)	16	1	0	0	0	0	0	0	0
	ID	41	16	25 (23)	36	0	9	6	8	0	0	17
	MSSA	3	1	40	0 (0)	0	0	4	6	0	0	1
	MT	0	0	0	0	0 (0)	0	0	0	0	0	0
	ND	0	0	9	0	0	0 (0)	2	0	0	0	0
	NH	0	0	6	4	0	2	0 (0)	0	0	0	0
	SD	0	0	8	6	0	0	0	0 (0)	0	0	1
	UT	0	0	0	0	0	0	0	0	0 (0)	0	0
	WV	0	0	0	0	0	0	0	0	0	0 (0)	0
	WY	1	0	18	1	0	0	0	1	0	0	0 (0)

[a]MSSA = Rhode Island and Vermont's Multi-State Science Assessment

Following the (operational) field test administration, items went through rubric validation and item data review, as described in Volume 2, Section, 2.7.1, Rubric Validation, and Section 2.7.2, Data Review. Table 17 presents the number of field-test items administered in MSSA, or another state, the number of items rejected before or during rubric validation, the number of items sent out to data review, and the number of items rejected during data review. The numbers in parentheses present the number of field-test items owned by MSSA.

Table 17. Number of Field-Test Item Administration, Rubric Validation, and Item Data Review in Spring 2021

Grade Band and Item Type	Number of Field-Test Items Administered	Number of Items Rejected Before/During Rubric Validation	Number of Items Sent to Data Review	Number of Items Rejected at Data Review	Number of Items Remaining[a]
Elementary School	**214 (7)**	**7 (0)**	**100 (3)**	**19 (0)**	**188 (7)**
Cluster	106 (4)	5 (0)	24 (0)	7 (0)	94 (4)
Stand-Alone	108 (3)	2 (0)	76 (3)	12 (0)	94 (3)
Middle School	**159 (12)**	**15 (1)**	**87 (9)**	**13 (5)**	**129 (6)**
Cluster	60 (6)	10 (1)	22 (3)	5 (3)	43 (2)
Stand-Alone	99 (6)	5 (0)	65 (6)	8 (2)	86 (4)
High School	**172 (7)**	**9 (0)**	**94 (6)**	**22 (4)**	**141 (3)**
Cluster	57 (2)	6 (0)	27 (1)	4 (1)	47 (1)
Stand-Alone	115 (5)	3 (0)	67 (5)	18 (3)	94 (2)
Total	**545 (26)**	**31 (1)**	**281 (18)**	**54 (9)**	**458 (16)**

Note: MSSA-owned items are indicated in the parentheses.
[a]Two Hawaii-owned items were not shared to the Shared Science Assessment Item bank.

Table 18 summarizes the Shared Science Assessment Item Bank after adding the field-test items that were administered in 2021 and passed rubric validation and item data review. The numbers in parentheses present the number of items owned by MSSA.

Table 18. Summary of Shared Science Assessment Item Bank in Spring 2021

Grade Band and Item Type	Science Discipline			Total[a]
	Earth and Space Sciences	*Life Sciences*	*Physical Sciences*	
Elementary School	**136 (10)**	**128 (7)**	**149 (7)**	**413 (24)**
Cluster	65 (4)	66 (4)	76 (5)	207 (13)
Stand-Alone	71 (6)	62 (3)	73 (2)	206 (11)
Middle School	**114 (4)**	**156 (6)**	**137 (7)**	**407 (17)**
Cluster	55 (2)	76 (2)	67 (3)	198 (7)
Stand-Alone	59 (2)	80 (4)	70 (4)	209 (10)
High School	**68 (6)**	**163 (3)**	**106 (3)**	**337 (12)**
Cluster	27 (3)	64 (1)	42 (1)	133 (5)
Stand-Alone	41 (3)	99 (2)	64 (2)	204 (7)
Total	**318 (20)**	**447 (16)**	**392 (17)**	**1157 (53)**

Note. MSSA-owned items are indicated in the parentheses.
[a]Two Hawaii-owned items were not shared to the Shared Science Assessment Item bank.

4. 2022 FIELD TESTS

In 2022, a fourth wave of items was field tested in 13 states and one U.S. territory. In all of these locations—Connecticut, Hawaii, Idaho, Montana, New Hampshire, North Dakota, Oregon, Rhode Island, South Dakota, the U.S. Virgin Islands, Utah, Vermont, West Virginia, and Wyoming—the field-test items were administered as unscored items embedded among the operational items. In total, 217 item clusters and 254 stand-alone items were administered as field-test items in the elementary, middle, and high school grade bands. Table 19 presents the number of field-test item clusters and stand-alone items administered in each grade band for each state. The numbers in parentheses in the column representing MSSA presents the number of field-test items owned by MSSA.

Table 19. Number of Field-Test Items Administered in Spring 2022

Grade Band and Item Type	CT	HI	ID	MSSA[a]	MT	ND	NH	OR	SD	USVI	UT	WV	WY	Total
Elementary School	**34**	**28**	**22**	**66 (9)**	**12**	**12**	**17**	**41**	**10**	**62**	**19**	**10**	**1**	**170**
Cluster	22	8	11	22 (4)	4	4	5	15	4	62	11	2	1	88
Stand-Alone	12	20	11	44 (5)	8	8	12	26	6	-	8	8	0	82
Middle School	**40**	**30**	**35**	**64 (9)**	**12**	**12**	**17**	**39**	**10**	**76**	**33**	**10**	**1**	**190**
Cluster	20	10	7	21 (4)	4	4	5	16	4	76	5	2	1	88
Stand-Alone	20	20	28	43 (5)	8	8	12	23	6	-	28	8	0	102
High School	**46**	**14**	**14**	**58 (9)**	**-**	**12**	**16**	**43**	**9**	**-**	**-**	**10**	**1**	**111**
Cluster	18	6	10	19 (4)	-	4	4	16	3	-	-	2	1	41
Stand-Alone	28	8	4	39 (5)	-	8	12	27	6	-	-	8	0	70
Total	**120**	**72**	**71**	**188 (27)**	**24**	**36**	**50**	**123**	**29**	**138**	**52**	**30**	**3**	**471**

Note. MSSA-owned items are indicated in the parentheses.

In the spring 2022 administrations, for the states with an operational test, field-test items were embedded within the operational test. Some of the states with an operational test (New Hampshire, Rhode Island, and Vermont) opted for a test in which operational items were grouped by science discipline. For these three states, the field-test items were presented together in a fourth group of items. The sequence of the four sets of items (corresponding to the three disciplines and a set of field-test items) was randomized across students. Ten other states and one U.S. territory (Connecticut, Hawaii, Idaho, Montana, North Dakota, Oregon, South Dakota, Utah, U.S. Virgin Islands, West Virginia, and Wyoming) opted for a test design in which the items were not grouped by discipline. In these ten states and one US territory, field-test items were administered at random positions throughout the test. A student received either a field-test item cluster or a set of four field-test stand-alone items. The test design for the MSSA is discussed in Section **Error! Reference source not found.**, **Error! Reference source not found.**.

A minimum sample size of 1,500 students per field-test item was targeted for any given state. Most items were administered in two states or territory. Approximately 61.6% of the items met or exceeded the target sample size of 1,500 in at least one state, while 88.0% of the items had a sample size of at least 1,350 (10% of the target) in at least one state. In addition, 98.3% of the items had a sample size of at least 1,200 in at least one state.

Table 20 to Table 22 present the number of item clusters and stand-alone items that were shared between the field-test pools of any two states or territory. The numbers below the shaded cells represent the number of common field-test items between any two states, and the numbers above the shaded cells represent the number of common field-test items that survived rubric validation and were included in the calibration. In each of the shaded cells, the number outside the parentheses represents the number of unique field-test items administered only in the given state or territory, and the number in parentheses represents the number of unique and/or common items that were calibrated with only the data from that state. Table 20 presents the results for elementary schools, Table 21 presents the results for middle schools, and Table 22 presents the results for high schools. The numbers of field-test items administered are slightly different from the numbers of field-test items at calibration because some items were rejected during rubric validation.

Table 20. Number of Common Elementary School Field-Test Items Administered and Calibrated, Spring 2022

	State	CT	HI	ID	MSSA[a]	MT	ND	NH	OR	SD	UT	WV	WY	USVI
Item Clusters	CT	0 (0)	0	3	1	0	0	0	3	0	15	0	0	0
	HI	0	0(0)	0	0	0	0	0	5	0	2	0	0	0
	ID	3	0	0(0)	3	0	0	0	0	0	5	0	0	0
	MSSA	1	0	3	0(0)	0	0	0	5	1	12	0	0	0
	MT	0	0	0	0	0(0)	0	0	0	0	4	0	0	0
	ND	0	0	0	0	0	0(0)	4	0	0	0	0	0	0
	NH	0	0	0	0	0	4	0(0)	0	0	1	0	0	1
	OR	3	6	0	5	0	0	0	0(0)	0	1	0	0	0
	SD	0	0	0	1	0	0	0	0	0 (0)	3	0	0	0
	UT	15	2	5	12	4	0	1	1	3	6 (6)	11	2	1
	WV	0	0	0	0	0	0	0	0	0	11	0(0)	0	0
	WY	0	0	0	0	0	0	0	0	0	2	0	0(0)	0
	USVI	0	0	0	0	0	0	0	0	0	0	0	0	0(0)
Stand-Alone Items	CT	0(0)	2	0	4	4	0	0	0	2	0	0	0	0
	HI	2	0(0)	3	7	0	0	0	8	0	0	0	0	0
	ID	0	3	0(0)	1	1	4	0	2	0	0	0	0	0
	MSSA	4	7	1	0(0)	3	0	1	7	4	0	8	8	0
	MT	4	0	1	3	0(0)	0	0	0	0	0	0	0	0
	ND	0	0	4	0	0	0(0)	3	1	0	0	0	0	0
	NH	0	0	0	1	0	3	1(0)	7	0	0	0	0	0
	OR	0	8	2	8	0	1	7	0(0)	0	0	0	0	0
	SD	2	0	0	4	0	0	0	0	0 (0)	0	0	0	0
	UT	0	0	0	0	0	0	0	0	0	0(0)	0	0	0
	WV	0	0	0	8	0	0	0	0	0	0	0(0)	0	0
	WY	0	0	0	8	0	0	0	0	0	0	0	0 (0)	0

	State	CT	HI	ID	MSSA[a]	MT	ND	NH	OR	SD	UT	WV	WY	USVI
	USVI	0	0	0	0	0	0	0	0	0	0	0	0	0 (0)
Total	**CT**	0(0)	2	3	5	4	0	0	3	2	15	0	0	0
	HI	2	0(0)	3	7	0	0	0	13	0	2	0	0	0
	ID	3	3	0(0)	4	1	4	0	2	0	5	0	0	0
	MSSA	5	7	4	0(0)	3	0	1	12	5	12	8	8	0
	MT	4	0	1	3	0(0)	0	0	0	0	4	0	0	0
	ND	0	0	4	0	0	0(0)	7	1	0	0	0	0	0
	NH	0	0	0	1	0	7	1(0)	7	0	1	0	0	1
	OR	3	14	2	13	0	1	7	0(0)	0	1	0	0	0
	SD	2	0	0	5	0	0	0	0	0(0)	3	0	0	0
	UT	15	2	5	12	4	0	1	1	3	6(6)	11	2	1
	WV	0	0	0	8	0	0	0	0	0	11	0(0)	0	0
	WY	0	0	0	8	0	0	0	0	0	2	0	0(0)	0
	USVI	0	0	0	0	0	0	1	0	0	1	0	0	0 (0)

[a]MSSA = Rhode Island and Vermont's Multi-State Science Assessment

Table 21. Number of Common Middle School Field-Test Items Administered and Calibrated, Spring 2022

	State	CT	HI	ID	MSSA[a]	MT	ND	NH	OR	SD	UT	WV	WY	USVI
Item Clusters	CT	0(0)	1	1	0	0	0	0	1	0	17	0	0	0
	HI	1	0(0)	0	1	0	0	0	1	0	5	0	0	0
	ID	1	0	0(0)	0	0	0	0	0	1	5	0	0	0
	MSSA	0	1	0	0(0)	0	0	0	2	0	18	0	0	0
	MT	0	0	0	0	0(0)	0	0	0	0	4	0	0	0
	ND	0	0	0	0	0	0(0)	0	0	0	4	0	0	0
	NH	0	0	0	0	0	0	0(0)	3	0	2	0	0	0
	OR	1	2	0	2	0	0	3	0(0)	0	8	0	0	0
	SD	0	0	1	0	0	0	0	0	0(0)	2	0	0	1
	UT	17	6	5	18	4	4	2	8	3	2(2)	5	2	1
	WV	0	0	0	0	0	0	0	0	0	5	0(0)	0	0
	WY	0	0	0	0	0	0	0	0	0	2	0	0(0)	0
	USVI	0	0	0	0	0	0	0	0	1	1	0	0	0(0)
Stand-Alone Items	CT	0(0)	0	0	12	0	0	0	4	1	0	3	0	0
	HI	0	0(0)	8	5	0	0	0	6	0	0	1	0	0
	ID	0	8	0(0)	5	8	0	0	3	4	0	0	0	0
	MSSA	12	5	5	0(0)	0	0	0	4	0	0	9	8	0
	MT	0	0	8	0	0(0)	0	0	0	0	0	0	0	0
	ND	0	0	0	0	0	0(0)	0	0	0	0	8	0	0
	NH	0	0	0	0	0	0	0(0)	6	0	0	5	0	0
	OR	4	6	3	4	0	0	6	0(0)	0	0	0	0	0
	SD	1	0	4	0	0	0	0	0	0(0)	0	1	0	0
	UT	0	0	0	0	0	0	0	0	0	0(0)	0	0	0
	WV	3	1	0	9	0	8	6	0	1	0	0(0)	0	0
	WY	0	0	0	8	0	0	0	0	0	0	0	0(0)	0

	State	CT	HI	ID	MSSA[a]	MT	ND	NH	OR	SD	UT	WV	WY	USVI
	USVI	0	0	0	0	0	0	0	0	0	0	0	0	0(0)
Total	**CT**	0(0)	1	1	12	0	0	0	5	1	17	3	0	0
	HI	1	0(0)	8	6	0	0	0	7	0	5	1	0	0
	ID	1	8	0(0)	5	8	0	0	3	5	5	0	0	0
	MSSA	12	6	5	0(0)	0	0	0	6	0	18	9	8	0
	MT	0	0	8	0	0(0)	0	0	0	0	4	0	0	0
	ND	0	0	0	0	0	0(0)	0	0	0	4	8	0	0
	NH	0	0	0	0	0	0	0(0)	9	0	2	5	0	0
	OR	5	8	3	6	0	0	9	0(0)	0	8	0	0	0
	SD	1	0	5	0	0	0	0	0	0(0)	2	1	0	1
	UT	17	6	5	18	4	4	2	8	3	2(2)	5	2	1
	WV	3	1	0	9	0	8	6	0	1	5	0(0)	0	0
	WY	0	0	0	8	0	0	0	0	0	2	0	0(0)	0
	USVI	0	0	0	0	0	0	0	0	1	1	0	0	0(0)

[a]MSSA = Rhode Island and Vermont's Multi-State Science Assessment.

Table 22. Number of Common High School Field-Test Items Administered and Calibrated, Spring 2022

	State	CT	HI	ID	MSSA[a]	MT	ND	NH	OR	SD	UT	WV	WY	USVI
Item Clusters	CT	0(0)	0	2	6	-	2	1	5	1	-	-	1	1
	HI	0	0(0)	3	0	-	0	0	2	0	-	-	0	0
	ID	2	3	0(0)	2	-	0	0	2	0	-	-	1	0
	MSSA	6	1	2	0(0)	-	2	1	4	2	-	-	0	0
	MT	-	-	-	-	-	-	-	-	-	-	-	-	-
	ND	2	0	0	2	-	0(0)	0	0	0	-	-	0	0
	NH	1	0	0	1	-	0	0 (0)	2	0	-	-	0	0
	OR	5	2	2	5	-	0	2	0(0)	0	-	-	0	1
	SD	1	0	0	2	-	0	0	0	0(0)	-	-	0	0
	UT	-	-	-	-	-	-	-	-	-	-	-	-	-
	WV	-	-	-	-	-	-	-	-	-	-	-	-	-
	WY	1	0	1	0	-	0	0	0	0	-	-	0(0)	0
	USVI	1	0	0	0	-	0	0	1	0	-	-	0	0(0)

	State	CT	HI	ID	MSSA[a]	MT	ND	NH	OR	SD	UT	WV	WY	USVI
Stand-Alone Items	**CT**	0(0)	0	1	19	-	6	0	1	1	-	-	0	0
	HI	0	0(0)	1	1	-	0	0	6	0	-	-	0	0
	ID	1	1	0(0)	1	-	0	0	1	0	-	-	0	0
	MSSA	19	1	1	0(0)	-	2	0	5	3	-	-	8	0
	MT	-	-	-	-	-	-	-	-	-	-	-	-	-
	ND	6	0	0	2	-	0(0)	0	0	0	-	-	0	0
	NH	0	0	0	0	-	0	0(0)	12	0	-	-	0	0
	OR	1	6	1	5	-	0	12	0(0)	2	-	-	0	0
	SD	1	0	0	3	-	0	0	2	0(0)	-	-	0	0
	UT	-	-	-	-	-	-	-	-	-	-	-	-	-
	WV	-	-	-	-	-	-	-	-	-	-	-	-	-
	WY	0	0	0	8	-	0	0	0	0	-	-	0(0)	0
	USVI	0(0)	0	1	19	-	6	0	1	1	-	-	0	0
Total	**CT**	0(0)	0	3	25	-	8	1	6	2	-	-	1	1
	HI	0	0(0)	4	1	-	0	0	8	0	-	-	0	0
	ID	3	4	0(0)	3	-	0	0	3	0	-	-	1	0
	MSSA	25	2	3	0(0)	-	4	1	9	5	-	-	8	0
	MT	-	-	-	-	-	-	-	-	-	-	-	-	-
	ND	8	0	0	4	-	0(0)	0	0	0	-	-	0	0
	NH	1	0	0	1	-	0	0(0)	14	0	-	-	0	0
	OR	6	8	3	10	-	0	14	0(0)	2	-	-	0	1
	SD	2	0	0	5	-	0	0	2	0(0)	-	-	0	0
	UT	-	-	-	-	-	-	-	-	-	-	-	-	-
	WV	-	-	-	-	-	-	-	-	-	-	-	-	-
	WY	1	0	1	8	-	0	0	0	0	-	-	0(0)	0
	USVI	1	0	0	0	-	0	0	1	0	-	-	0	0(0)

[a]MSSA = Rhode Island and Vermont's Multi-State Science Assessment

Following the (operational) field test administration, items went through rubric validation and item data review, as described in Volume 2, Section, 2.7.1, Rubric Validation, and Section 2.7.2, Data Review.

Table 23 presents the number of field-test items administered in MSSA, or another state, the number of items rejected before or during rubric validation, the number of items sent out to data review, and the number of items rejected during data review. The numbers in parentheses present the number of field-test items owned by MSSA

Table 23. Number of Field-Test Item Administration, Rubric Validation, and Item Data Review, Spring 2022

Grade Band and Item Type	Number of Field-Test Items Administered	Number of Items Rejected Before/During Rubric Validation	Number of Items Sent to Data Review	Number of Items Rejected at Data Review	Number of Items Remaining
Elementary School	**170 (9)**	**3 (0)**	**82 (3)**	**14 (1)**	**153 (8)**
Cluster	88 **(4)**	1 **(0)**	18 **(0)**	4 **(0)**	83 **(4)**
Stand-Alone	82 **(5)**	2 **(0)**	64 **(3)**	10 **(1)**	70 **(4)**
Middle School	**190 (9)**	**4 (0)**	**94 (3)**	**26 (1)**	**160 (8)**
Cluster	88 **(4)**	3 **(0)**	26 **(0)**	13 **(0)**	72 **(4)**
Stand-Alone	102 **(5)**	1 **(0)**	68 **(3)**	13 **(1)**	88 **(4)**
High School	**111 (9)**	**2 (0)**	**63 (5)**	**19 (5)**	**90 (4)**
Cluster	41 **(4)**	2 **(0)**	21 **(1)**	3 **(1)**	36 **(3)**
Stand-Alone	70 **(5)**	0 **(0)**	42 **(4)**	16 **(4)**	54 **(1)**
Total	**471 (27)**	**9 (0)**	**239 (11)**	**59 (7)**	**403 (20)**

Note: MSSA-owned items are indicated in the parentheses.

Table 24 summarizes the Shared Science Assessment Item Bank after adding the field-test items that were administered in 2022 and passed rubric validation and item data review. The numbers in parentheses present the number of items owned by MSSA.

Table 24. Summary of Shared Science Assessment Item Bank, Spring 2022

Grade Band and Item Type	Science Discipline			Total[a]
	Earth and Space Sciences	*Life Sciences*	*Physical Sciences*	
Elementary School	**180 (13)**	**162 (9)**	**214 (10)**	**556 (32)**
Cluster	96 (6)	82 (5)	111 (6)	289 (17)
Stand-Alone	84 (7)	80 (4)	103 (4)	267 (15)
Middle School	**150 (6)**	**220 (9)**	**187 (10)**	**557 (25)**
Cluster	70 (3)	110 (4)	90 (4)	270 (11)
Stand-Alone	80 (3)	110 (5)	97 (6)	287 (14)
High School	**91 (7)**	**194 (4)**	**129 (5)**	**414 (16)**
Cluster	35 (4)	78 (2)	53 (2)	166 (8)
Stand-Alone	56 (3)	116 (2)	76 (3)	248 (8)
Total	**421 (26)**	**576 (22)**	**530 (25)**	**1 527 (73)**

Note. MSSA-owned items are indicated in the parentheses.
[a]Count excludes nine MOU items that do not align to the NGSS.

Appendix 1-C
Calibration of the Shared Science Assessment Item Bank

TABLE OF CONTENTS

LIST OF TABLES

The Shared Science Assessment Item Bank was first calibrated in 2018 after the 2018 science test administrations concluded, and it was recalibrated in 2019 following the 2019 test administrations. The calibration sequences are documented in this appendix, which also includes details on scale linking and the creation of the anchor scale in 2019. The calibration of field-test items in 2021 and beyond as well as the calibration software are addressed.

1. 2018 CALIBRATION SEQUENCE

Table 1 provides an overview of the groups per grade for the 2018 calibration.

Table 1. Groups per Grade Band for the Spring 2018 Core Calibration

Group	Elementary School	Middle School	High School
Connecticut	X	X	X
Hawaii	X	X	X
New Hampshire	X	X	X
Rhode Island	X	X	X
Utah Grade 6		X	
Utah Grade 7		X	
Utah Grade 8		X	
Vermont	X	X	X
West Virginia	X	X	

Items were calibrated in three steps for two reasons. First, the rubric validations for some states took place at a later date, and the student responses for the items owned by those states could not be included in the first round of calibrations without jeopardizing the reporting schedule of the two states with operational field tests. (i.e., those two states did not have any of the items with late rubric validation in their item pool.) Second, to divide the large set of items and assertions into more manageable pieces, a separate calibration was carried out for two states with many items administered only in those states. Specifically, the following sequence of calibrations was carried out:

1. **Core Calibration.** The core calibration was performed on the following:
    a. All the item responses of New Hampshire and West Virginia. These states administered items from the following (as described in the state-sharing matrix in Table 2):
        i. ICCR item bank
        ii. Connecticut
        iii. Hawaii
        iv. Rhode Island
        v. Vermont

vi. Utah

vii. West Virginia

A more detailed overlap of the common items at the time of the 2018 calibration was given in Section 1 of Appendix 1-B, Shared Science Assessment Item Bank: Field Testing (see Table 2 through Table 4).

b. All the item responses of Connecticut, Rhode Island, and Vermont, excluding responses to Wyoming and Oregon items. These states administered items from the following sources:

i. ICCR item bank

ii. Connecticut

iii. Hawaii

iv. Rhode Island

v. Vermont

vi. Utah

vii. West Virginia

viii. Wyoming (items were treated as "not administered"; responses were replaced by missing code)

ix. Oregon (items were treated as "not administered"; responses were replaced by missing code)

c. Item responses from Hawaii to items also administered in another state (Hawaii items were used in Hawaii, Connecticut, Rhode Island, Vermont, and West Virginia) underwent core calibration.

d. Item responses from Utah to items also administered in another state (Utah items were used in Utah, Connecticut, Rhode Island, Vermont, and West Virginia) underwent core calibration. Utah tested only middle school students. One-third of students were selected at random to balance the large population size for Utah.

Table 2. Spring 2018 State-Sharing Matrix

Source Bank	CT	HI	MSSA[a]	NH	OR	UT	WV	WY
ICCR	X	X	X	X	X		X	X
Connecticut	X		X				X	
Hawaii	X	X	X				X	
MSSA	X		X				X	
Oregon	X		X		X			
Utah	X		X			X	X	
West Virginia	X		X				X	
Wyoming	X		X					X

Note. The core calibration provided parameters for all items used in New Hampshire and West Virginia.

2. **Calibration of State-Specific Items.** In terms of the calibration of state-specific items, both Hawaii and Utah had a substantial proportion of items that were administered only in Hawaii and Utah, respectively. Hawaii had both Hawaii and ICCR items in common with the states involved in the core calibration (Hawaii administered Hawaii and ICCR items only); whereas Utah had only Utah items in common (Utah administered Utah items only). The parameters for the unique Hawaii items depended on responses from Hawaii students only, and the parameters for the unique Utah items depended on responses from Utah students only. For both states, the state-specific items were calibrated through a separate calibration based on the state data only, with the items in common with the core states mentioned in Step 1 anchored to the estimates from Step 1. These calibrations were performed separately for each group under a single-group item response theory (IRT) model. The mean and variance of the groups were fixed to the estimated mean and variance from the core calibration.

3. **Calibration of States with Late Rubric Validation.** Oregon and Wyoming items were administered in some of the states involved in the core calibration (Connecticut, Rhode Island, and Vermont) but could not be calibrated in Step 1 because of their late rubric validation dates. In a later stage, items from Oregon and Wyoming were calibrated by

    a. adding Oregon and Wyoming student responses to the core calibration;

    b. keeping the responses from Connecticut, Rhode Island, and Vermont to Oregon and Wyoming items (as opposed to treating them as missing in Step 1);

    c. removing the responses from Hawaii, New Hampshire, Utah, and West Virginia, who did not administer Oregon or Wyoming items (as the item parameters for the Oregon and Wyoming items did not depend on the students from these states); and

    d. fixing the parameters of all other items to the values obtained in Step 1, and the group means and standard deviations that were estimated in Step 1.

2. 2019 CALIBRATION

Calibration was performed in two steps. First, CAI calibrated all items in operational use in 2019 for which 1,000 or more student responses were available (among these, there were 1,500 or more student responses for all but three items). In this step, only the data of states with an operational test were included. Table 3 provides an overview of the groups per grade band for this first calibration. All students who attempted the test were included in the calibration. The assertions of skipped items were scored as incorrect. Note that only Rhode Island allowed students to skip items. Out of a total of 438 items, there were nine items administered as operational items in 2019, for which the sample size was smaller than 1,000.

Table 4–Table 6 present the number of operational item clusters and stand-alone items that were shared between the item pools of any two states. The numbers below the shaded cells represent the number of all the operational items administered, and the numbers above the shaded cells represent the number of common operational items at the time of the 2019 calibration. The shaded cells represent the number of operational items administered only in the given state (the number of unique operational items at the time of calibration are provided in parentheses). Since the items that were administered but not calibrated were administered only in one state, the numbers above the diagonal are the same as the numbers below the diagonal.

Table 4 presents the results for elementary schools, Table 5 presents the results for middle schools, and Table 6 presents the results for high schools. The numbers at the operational administration are slightly different from the numbers at the calibration because items with a sample size smaller than 1,000 students were excluded from the calibration.

Table 3. Groups per Grade Band for the Spring 2019 Calibration of Operational Items

Group	Elementary School	Middle School	High School
Connecticut	X	X	X
New Hampshire	X	X	X
Oregon	X	X	X
Rhode Island	X	X	X
Vermont	X	X	X
West Virginia	X	X	

Table 4. Number of Common Elementary School Operational Items Administered and Calibrated, Spring 2019

	State	CT	MSSA[a]	NH	OR	WV
Cluster	CT	**1 (1)**	44	24	42	55
	MSSA	44	**0 (0)**	17	37	41
	NH	24	17	**0 (0)**	14	27
	OR	42	37	14	**0 (0)**	41
	WV	55	41	27	41	**1 (1)**
Stand-Alone	CT	**3 (3)**	34	26	30	47
	MSSA	34	**0 (0)**	20	23	32
	NH	26	20	**0 (0)**	14	25
	OR	30	23	14	**0 (0)**	25
	WV	47	32	25	25	**1 (1)**
Total	CT	**4 (4)**	78	50	72	102
	MSSA	78	**0 (0)**	37	60	73
	NH	50	37	**0 (0)**	28	52
	OR	72	60	28	**0 (0)**	66
	WV	102	73	52	66	**2 (2)**

[a]MSSA = Rhode Island and Vermont's Multi-State Science Assessment

Table 5. Number of Common Middle School Operational Items Administered and Calibrated, Spring 2019

	State	CT	MSSA[a]	NH	OR	WV
Cluster	**CT**	**3 (3)**	26	24	54	92
	MSSA	26	**0 (0)**	11	14	21
	NH	24	11	**1 (1)**	9	18
	OR	54	14	9	**2 (2)**	56
	WV	92	21	18	56	**12 (4)**
Stand-Alone	**CT**	**0 (0)**	42	26	34	50
	MSSA	42	**0 (0)**	25	30	37
	NH	26	25	**0 (0)**	16	21
	OR	34	30	16	**1 (0)**	29
	WV	50	37	21	29	**0 (0)**
Total	**CT**	**3 (3)**	68	50	88	142
	MSSA	68	**0 (0)**	36	44	58
	NH	50	36	**1 (1)**	25	39
	OR	88	44	25	**3 (2)**	85
	WV	142	58	39	85	**12 (4)**

[a]MSSA = Rhode Island and Vermont's Multi-State Science Assessment

Table 6. Number of Common High School Operational Items Administered and Calibrated, Spring 2019

	State	CT	MSSA[a]	NH	OR	WV
Cluster	CT	**5 (5)**	33	22	30	0
	MSSA	33	**0 (0)**	20	31	0
	NH	22	20	**2 (2)**	15	0
	OR	30	31	15	**1 (1)**	0
	WV	0	0	0	0	**0 (0)**
Stand-Alone	CT	**0 (0)**	39	27	40	0
	MSSA	39	**2 (2)**	23	32	0
	NH	27	23	**0 (0)**	20	0
	OR	40	32	20	**4 (4)**	0
	WV	0	0	0	0	**0 (0)**
Total	CT	**5 (5)**	72	49	70	0
	MSSA	72	**2 (2)**	43	63	0
	NH	49	43	**2 (2)**	35	0
	OR	70	63	35	**5 (5)**	0
	WV	0	0	0	0	**0 (0)**

[a]MSSA = Rhode Island and Vermont's Multi-State Science Assessment

In Step 2, the field-test items were calibrated. The calibration included the operational items that were calibrated in Step 1, and the field-test items across all states in which they were administered. All students who attempted at least one field-test item were included in the calibration. Table 7 provides an overview of the groups per grade band for calibration of the field-test items.

Table 7. Groups per Grade Band for the Spring 2019 Calibration of Field-Test Items

Group	Elementary School	Middle School	High School
Connecticut	X	X	X
Hawaii	X	X	X
Idaho	X	X	
New Hampshire	X	X	X
Oregon	X	X	X
Rhode Island	X	X	X
Vermont	X	X	X
West Virginia	X	X	
Wyoming	X	X	X

3. LINKING THE 2018 SCALE TO THE 2019 SCALE

The item parameter estimates obtained from the 2018 student responses were highly correlated with the item parameters obtained from the 2019 student responses. For item difficulties, the correlation between the 2018 and 2019 estimates was 0.993 for elementary school, 0.986 for middle school, and 0.994 for high school. For the standard deviations of the clusters, these correlations were 0.971 for elementary school, 0.972 for middle school, and 0.964 for high school. These high correlations indicate that items functioned similarly in 2018 and 2019. Nevertheless, item parameters from separate calibrations cannot be directly compared because the scale of an IRT model was not determined. In the multigroup Rasch testlet model, the only scale indeterminacy was the origin of the scale. The models can be identified by setting the mean of the overall proficiency variable θ to zero for the reference distribution. As a result, the 2018 and 2019 variable θ and item parameters were on the same scale except for an overall shift parameter B. Specifically, the 2018 scale can be linked to the 2019 scale as follows:

$$P\left(z_{ij}|\theta_{j\,2018}, u_{jg}; b_{i\,2018}\right) = \frac{\exp\left(\theta_{j\,2018} + u_{jg} - b_{i\,2018}\right)}{1 + \exp\left(\theta_{j\,2018} + u_{jg} - b_{i\,2018}\right)}$$

$$= \frac{exp\left(\theta_{j\,2018} + B + u_{jg} - b_{i\,2018} - B\right)}{1 + exp\left(\theta_{j\,2018} + B + u_{jg} - b_{i\,2018} - B\right)}$$

$$= \frac{exp\left(\theta_{j\,2019} + u_{jg} - b_{i\,2019}\right)}{1 + exp\left(\theta_{j\,2019} + u_{jg} - b_{i\,2019}\right)}.$$

Because $\theta_{j\,2019} = \theta_{j\,2018} + B$, the population means of θ must be transformed accordingly,

$$\theta_{j\,2019} \sim N\left(\mu_{k\,2018} + B, \sigma_k^2\right) \text{ and}$$

$$\theta_{j\,2018} \sim N\left(\mu_{k\,2018}, \sigma_k^2\right).$$

Item parameters based on 2018 student responses were expressed on the 2019 scale by adding the constant B to the 2018 item parameter. The 2018 parameters were expressed on the 2019 scale for items that were part of the pool in both 2018 and 2019 but were not administered in any states in 2019 (13 items), and for items that were administered in 2019 but the number of student responses from the 2019 assessments was lower than 1,000 (nine items). Therefore, the linking process was performed for 22 items only.

All items that were operational in 2019 were also administered in 2018. Therefore, the shift parameter B was estimated from a separate calibration of the 2019 operational items using the 2019 student responses (from the six operational states), but with the item parameters fixed to the estimates obtained from the 2018 calibrations. By fixing a subset of the item parameters, the model is identified so that the means and variances of θ can be estimated for all groups. Parameter B can be obtained by equating the overall mean of θ across all groups for the 2019 student response data from the free calibration (i.e., the 2019 overall mean expressed on the 2019 scale) to the overall mean of θ across all groups for the 2019 student response data from the calibration with items anchored to their 2018 parameters values (i.e., the 2019 overall mean expressed on the 2018 scale):

$$\frac{1}{K}\sum_{k=1}^{K}\mu_{k\ 2019} = \frac{1}{K}\sum_{k=1}^{K}(\mu_{k\ 2018} + B),$$

Therefore, an estimate of parameter *B* can be obtained as

$$\hat{B} = \frac{1}{K}\sum_{k=1}^{K}(\hat{\mu}_{k\ 2019} - \hat{\mu}_{k\ 2018}).$$

Table 8 presents the estimated means of θ under both the free and anchored calibrations, and the number of students per state. Table 8 also presents the overall means and estimated shift in parameter *B*. Note that the parameters for three items were not anchored, but instead were freely estimated together with the means and variances in the anchored calibration. The reason for not treating these items as common items across the 2018 and 2019 administrations is that they had an omit rate of 4% or higher for the last item interaction in the 2018 administration in at least one state. In 2019, these interactions could no longer be omitted because all interactions of an item needed to be responded to in states where skipping was not allowed (i.e., all states excluding Rhode Island). Therefore, out of an abundance of caution, these three items are not anchored to their 2018 parameter values.

Table 8. Estimated Latent Means and Number of Students Per State

Group	Elementary School			Middle School			High School		
	$\hat{\mu}_{k\ 2019}$	$\hat{\mu}_{k\ 2018}$	*N*	$\hat{\mu}_{k\ 2019}$	$\hat{\mu}_{k\ 2018}$	*N*	$\hat{\mu}_{k\ 2019}$	$\hat{\mu}_{k\ 2018}$	*N*
Connecticut	0.0000	0.0518	38,549	0.0000	0.0234	39,347	0.0000	0.1443	37,616
New Hampshire	0.0631	0.1083	13,187	0.0940	0.1108	12,060	0.0798	0.2278	11,385
Oregon	-0.0101	0.0096	44,989	0.0028	0.0156	42,043	-0.0383	0.1030	41,630
Rhode Island	-0.0312	0.0142	10,751	-0.1044	-0.0692	10,306	-0.2261	-0.0879	9,612
Vermont	0.1069	0.1504	6,017	0.0781	0.1133	5,894	0.0179	0.1545	5,332
West Virginia	-0.1970	-0.1529	19,540	-0.3012	-0.2783	19,043	–	–	–
	$\frac{1}{K}\sum_{k=1}^{K}\hat{\mu}_{k\ 2019}$	$\frac{1}{K}\sum_{k=1}^{K}\hat{\mu}_{k\ 2018}$	$\hat{B}$	$\frac{1}{K}\sum_{k=1}^{K}\hat{\mu}_{k\ 2019}$	$\frac{1}{K}\sum_{k=1}^{K}\hat{\mu}_{k\ 2018}$	$\hat{B}$	$\frac{1}{K}\sum_{k=1}^{K}\hat{\mu}_{k\ 2019}$	$\frac{1}{K}\sum_{k=1}^{K}\hat{\mu}_{k\ 2018}$	$\hat{B}$
Overall	-0.0114	0.0303	-0.0416	-0.0385	-0.0141	-0.0244	-0.0333	0.1083	-0.1417

4. CALIBRATION OF FIELD-TEST ITEMS IN 2021 AND BEYOND

Starting in 2021, field-test items were calibrated with one multigroup calibration per grade band. In each calibration, the parameters of the operational items were fixed to their bank values (anchor items) and the item parameters of the field-test items, as well as the mean and variance of each group, were estimated using the marginal maximum likelihood (MML) method. The calibration included the field-test items across all states in which the items were administered. All students who attempted at least one field-test item were included in the calibration. In 2021 and 2022, the

same groups were included for each grade band for the field-test calibration. Refer to Table 9 and Table 10 for an overview of the groups per grade band for calibration of the field-test items in 2021 and 2022, respectively.

In 2021, all but 12 items were calibrated on at least 1,500 student responses, and all but one item had a sample size larger than 1,200. The item with fewer than 1,200 responses had a sample size of 981 and was an interim item. In 2022, all but 64 items were calibrated on at least 1,500 student responses, and all but nine items were calibrated on at least 1,200 responses. The nine items with fewer than 1,200 responses were either Oregon Legacy items or interim items.

Table 9. Groups per Grade Band for the Spring 2021 Calibration of Field-Test Items

Group	Elementary School	Middle School	High School
Connecticut	X	X	X
Hawaii	X	X	X
Idaho	X	X	
Montana	X	X	
North Dakota	X	X	X
New Hampshire	X	X	X
Oregon	X	X	X
Rhode Island	X	X	X
South Dakota	X	X	X
Utah	X	X	
Vermont	X	X	X
West Virginia	X	X	
Wyoming	X	X	X

Table 10.Groups per Grade Band for the Spring 2022 Calibration of Field-Test Items

Group	Elementary School	Middle School	High School
Connecticut	X	X	X
Hawaii	X	X	X
Idaho	X	X	
Montana	X	X	
North Dakota	X	X	X
New Hampshire	X	X	X
Oregon	X	X	X
Rhode Island	X	X	X
South Dakota	X	X	X
Utah	X	X	

Group	Elementary School	Middle School	High School
Vermont	X	X	X
West Virginia	X	X	
Wyoming	X	X	X

5. CALIBRATION SOFTWARE

In 2018 and 2019, the IRT models were fitted using the Bayesian networks with the logistic regression (BNL) suite of Matlab functions (Rijmen, 2006) and flexMIRT (Cai, 2017). The resulting parameters from BNL were used as starting values for flexMIRT to reduce the estimation time for flexMIRT. The flexMIRT estimates were taken to be the operational parameters, except for the middle-school items calibrated in 2018 during the core calibration. For the 2018 core calibration of middle-school items, flexMIRT did not converge after several weeks, and the estimates obtained from BNL were used as operational parameters. Note that the parameters estimates were very similar across software packages.

Starting in 2021, Cambium Assessment IRT (CAIRT) was used for all calibrations because the estimation time in flexMIRT became prohibitive. CAIRT was specifically developed by CAI to calibrate the multigroup Rasch model on very large data sets. It relies on the same estimation methods as BNL. CAI has cross-validated parameter estimates from CAIRT with BNL and flexMIRT under various scenarios (Rijmen, Liao, & Lin, 2021). In 2023, field-test items were calibrated in CAIRT using the same procedure used in 2021.

6. REFERENCES

Cai, L. (2017). flexMIRT®: Flexible multilevel multidimensional item analysis and test scoring (version 3.51) [computer software]. Chapel Hill, NC: Vector Psychometric Group.

Rijmen, F. (2006). *BNL: A Matlab toolbox for Bayesian networks with logistic regression nodes* (Technical Report). Amsterdam: VU University Medical Center.

Rijmen, F., Liao, D., & Lin, Z. (2021). *The Rasch testlet model for the calibration of three-dimensional science assessments: A software comparison* [White paper]. Washington, DC: Cambium Assessment, Inc.

Appendix 1-D

Distribution of Scale Scores and Achievement Levels

Distribution of Scale Scores and Achievement Levels

Table D-1. Scale Score Mean and Standard Deviation by Grade

	Grade		
	5	**8**	**11**
Number of Students	9,813	10,089	9,212
Mean Scale Score	50.15	49.49	52.86
SD of Scale Score	18.44	17.37	16.64

Table D-2. Proportion of Students in Each Achievement Level by Grade

Achievement Level	**Grade**		
	5	**8**	**11**
Number of Students	9,813	10,089	9,212
Level 1	0.25	0.27	0.12
Level 2	0.43	0.45	0.56
Level 3	0.19	0.19	0.17
Level 4	0.13	0.09	0.15

Appendix 1-E

Distribution of Scale Scores by Science Discipline

Distribution of Scale Scores by Science Discipline

Table E-1. Science Disciplines

Grade	Discipline
5, 8, 11	Physical Sciences (PS) Life Sciences (LS) Earth & Space Sciences (ESS)

Table E-2. Overall Discipline Score Mean and Standard Deviation, Grade 5 Science

N	Scale Score	Discipline		
		PS	LS	ESS
9,813	Mean	50.54	50.48	50.41
	SD	20.55	22.70	21.78

Table E-3. Overall Discipline Score Mean and Standard Deviation, Grade 8 Science

N	Scale Score	Discipline		
		PS	LS	ESS
10,089	Mean	49.84	49.40	49.46
	SD	19.64	21.51	19.80

Table E-4. Overall Discipline Score Mean and Standard Deviation, Grade 11 Science

N	Scale Score	Discipline		
		PS	LS	ESS
9,212	Mean	52.49	52.24	52.76
	SD	19.43	20.61	19.76

Appendix 1-F

Distribution of Scale Scores and Achievement Levels by Subgroup

Distribution of Scale Scores and Achievement Levels by Subgroup

Table F-1. Mean and Standard Deviation of Scale Scores by Subgroup

Group	Grade 5			Grade 8			Grade 11		
	N	*Mean*	*SD*	*N*	*Mean*	*SD*	*N*	*Mean*	*SD*
All Students	9,813	50.15	18.44	10,089	49.49	17.37	9,212	52.86	16.64
Female	4,785	50.04	17.75	4,858	49.57	16.62	4,505	53.09	15.34
Male	5,023	50.23	19.04	5,221	49.38	18.01	4,695	52.61	17.81
Unspecified	*	*	*	10	71.55	22.98	12	57.95	14.97
African American	851	42.88	16.04	917	40.96	15.48	804	45.80	13.35
American Indian/Native Alaskan	79	38.86	15.94	70	40.54	13.18	59	45.01	13.37
Asian	337	56.78	19.21	338	54.72	18.25	292	59.98	18.22
Hispanic	2,935	42.83	16.04	2,962	42.46	15.06	2,548	46.06	13.22
Multi-Racial	508	48.45	18.03	504	47.91	15.84	343	50.98	16.31
Pacific Islander	13	45.24	24.11	11	46.99	13.47	14	48.73	15.66
White	5,090	55.50	18.13	5,287	54.85	16.96	5,152	57.14	17.01
Limited English Proficiency	1,743	40.85	16.67	1,754	40.92	15.71	937	39.73	9.72
Special Education	1,504	34.90	15.03	1,539	36.86	13.29	1,064	41.65	11.99
Economically Disadvantaged	4,715	43.20	16.40	4,403	42.60	14.91	3,375	46.62	13.45

**Note. Subgroup is not reported due to small sample size (sample size <10).*

Table F-2. Percentage of Achievement Level by Subgroup

Group	Grade 5					Grade 8					Grade 11				
	N	*L1*	*L2*	*L3*	*L4*	*N*	*L1*	*L2*	*L3*	*L4*	*N*	*L1*	*L2*	*L3*	*L4*
All Students	9,813	0.25	0.43	0.19	0.13	10,089	0.27	0.45	0.19	0.09	9,212	0.12	0.56	0.17	0.15
Female	4,785	0.24	0.45	0.19	0.12	4,858	0.25	0.48	0.19	0.09	4,505	0.1	0.58	0.18	0.13
Male	5,023	0.26	0.42	0.18	0.14	5,221	0.29	0.42	0.19	0.1	4,695	0.14	0.55	0.15	0.16
Unspecified	*	*	*	*	*	10	0.10	0.10	0.40	0.40	12	-	0.58	0.25	0.17
African American	851	0.39	0.45	0.12	0.05	917	0.46	0.42	0.09	0.03	804	0.18	0.68	0.09	0.05
American Indian/Native Alaskan	79	0.47	0.38	0.13	0.03	70	0.50	0.37	0.13	-	59	0.25	0.61	0.10	0.03
Asian	337	0.15	0.38	0.24	0.23	338	0.16	0.47	0.22	0.15	292	0.08	0.43	0.22	0.27
Hispanic	2,935	0.38	0.46	0.12	0.04	2,962	0.41	0.45	0.11	0.04	2,548	0.18	0.67	0.09	0.05
Multi-Racial	508	0.28	0.45	0.16	0.11	504	0.28	0.49	0.16	0.07	343	0.15	0.61	0.12	0.12
Pacific Islander	13	0.46	0.08	0.31	0.15	11	0.27	0.55	0.18	-	14	0.14	0.64	0.14	0.07
White	5,090	0.15	0.42	0.24	0.19	5,287	0.16	0.45	0.25	0.14	5,152	0.08	0.50	0.22	0.21
Limited English Proficiency	1,743	0.45	0.4	0.10	0.05	1,754	0.48	0.40	0.08	0.04	937	0.3	0.67	0.02	0.00
Special Education	1,504	0.59	0.35	0.05	0.02	1,539	0.58	0.36	0.05	0.01	1,064	0.28	0.66	0.03	0.03
Economically Disadvantaged	4,715	0.38	0.46	0.12	0.05	4,403	0.40	0.47	0.11	0.03	3,375	0.17	0.68	0.10	0.05

**Note. Subgroup is not reported due to small sample size (sample size <10).*