North Dakota State Assessment for Science

2021–2022

Volume 1: Annual Technical Report

NORTH DAKOTA DEPARTMENT OF
PUBLIC INSTRUCTION

TABLE OF CONTENTS

1. INTRODUCTION 1

1.1 Background and Historical Context of Tests 1
1.2 Purpose and Intended Uses of the North Dakota State Assessment for Science 2
1.3 Participants in the Development and Analysis of the NDSA for Science 3
1.3.1 North Dakota Department of Public Instruction 3
1.3.2 North Dakota Educators 4
1.3.3 Cambium Assessment, Inc. 4
1.3.4 Caveon Test Security 4
1.4 Available Test Formats and Special Versions 4
1.5 Student Participation 5

2. OPERATIONAL PRACTICES AND PROCEDURES 6

2.1 Test Administration 6
2.2 Simulations 6
2.3 Universal Features, Designated Supports, and Accommodations 7

3. ITEM BANK AND TEST DESIGN 9

3.1 Shared Science Assessment Item Bank 9
3.2 Field Testing 10
3.2.1 2018 Field Test 10
3.2.2 2019 Field Test 17
3.2.3 2021 Field Test 24
3.2.4 2022 Field Test 35
3.3 Test Design 47

4. FIELD-TEST CLASSICAL ANALYSIS OVERVIEW 57

4.1 Item Discrimination 58
4.2 Item Difficulty 58
4.3 Response Time 58
4.4 Differential Item Functioning Analysis 58
4.5 Classical Analysis Results 61

5. ITEM CALIBRATION 64

5.1 Model Description 64
5.1.1 Latent Structure 64
5.1.2 Item Response Function 66
5.1.3 Multigroup Model 66
5.2 Item Calibration 67
5.2.1 Estimation 67
5.2.2 2018 Calibration Sequence 68
5.2.3 2019 Calibration Sequence 71

5.2.4 Linking the 2018 Scale to the 2019 Scale 75
5.2.5 Calibration of Field-Test Items in 2021 and Beyond 76
5.2.6 Overview of the Operational Bank 77
6. SCORING 79
6.1 Marginal Maximum Likelihood Function 79
6.2 Derivative 80
6.3 Extreme Case Handling 82
6.4 Standard Error of Estimate 82
6.5 Scoring Incomplete Tests 82
6.6 Student-Level Scale Score 83
6.7 Rules for Calculating Achievement Levels 84
6.7.1 Strengths and Weaknesses for Disciplines Relative to Proficiency Cut Score..85
6.8 Residual-Based Reporting at the Level of Disciplinary Core Idea-Level Reporting and Science and Engineering Practices 85
6.8.1 Relative to Overall Achievement 85
6.8.2 Relative to Proficiency Cut Score 86
7. QUALITY CONTROL PROCEDURES 87
7.1 Quality Assurance Reports 87
7.1.1 Item Analysis 87
7.1.2 Blueprint Match 87
7.1.3 Item Exposure Rates 88
7.1.4 Cheating Detection Analysis 88
7.2 Scoring Quality Check 88
8. REFERENCES 89

LIST OF TABLES

Table 1. Required Uses and Citations for the NDSA for Science 3
Table 2. Number of Students Participating in the NDSA for Science in Spring 2022 5
Table 3. Distribution of Demographic Characteristics of Student Population 5
Table 4. NDSA for Science Testing Windows 6
Table 5. Number of Testing Sessions with Allowed Designated Supports 8
Table 6. Number of Testing Sessions with Allowed Accommodations 8
Table 7. Number of Field-Test Items Administered in Spring 2018 11
Table 8. Number of Common Elementary School Field-Test Items Administered and Calibrated in Spring 2018 12
Table 9. Number of Common Middle School Field-Test Items Administered and Calibrated in Spring 2018 13
Table 10. Number of Common High School Field-Test Items Administered and Calibrated in Spring 2018 14
Table 11. Overview of Field-Test Item Administration, Rubric Validation, and Item Data Review in Spring 2018 16
Table 12. Overview of Shared Science Assessment Item Bank in Spring 2018 17
Table 13. Number of Field-Test Items Administered in Spring 2019 17
Table 14. Number of Common Elementary School Field-Test Items Administered and Calibrated in Spring 2019 19
Table 15. Number of Common Middle School Field-Test Items Administered and Calibrated in Spring 2019 20
Table 16. Number of Common High School Field-Test Items Administered and Calibrated in Spring 2019 21
Table 17. Overview of Field-Test Item Administration, Rubric Validation, and Item Data Review in Spring 2019 23
Table 18. Overview of Shared Science Assessment Item Bank in Spring 2019 24
Table 19. Number of Field-Test Items Administered in Spring 2021 25
Table 20. Number of Common Elementary School Field-Test Items Administered and Calibrated in Spring 2021 27
Table 21. Number of Common Middle School Field-Test Items Administered and Calibrated in Spring 2021 29
Table 22. Number of Common High School Field-Test Items Administered and Calibrated in Spring 2021 31
Table 23. Overview of Field-Test Item Administration, Rubric Validation, and Item Data Review in Spring 2021 34
Table 24. Overview of Shared Science Assessment Item Bank in Spring 2021 35
Table 25. Number of Field-Test Items Administered in Spring 2022 36
Table 26. Number of Common Elementary School Field-Test Items Administered and Calibrated in Spring 2022 38

Table 27. Number of Common Middle School Field-Test Items Administered and Calibrated in Spring 2022 40
Table 28. Number of Common High School Field-Test Items Administered and Calibrated in Spring 2022 42
Table 29. Overview of Field-Test Item Administration, Rubric Validation, and Item Data Review in Spring 2022 45
Table 30. Overview of Shared Science Assessment Item Bank in Spring 2022 46
Table 31. NDSA for Science Test Blueprint, Grade 4 48
Table 32. NDSA for Science Test Blueprint, Grade 8 50
Table 33. NDSA for Science Test Blueprint, Grade 10 54
Table 34. Thresholds for Flagging in Classical Item Analysis 57
Table 35. Differential Item Functioning Classification Rules 61
Table 36. Distribution of p-Values for Field-Test Items in Spring 2022 61
Table 37. Distribution of Item Biserial Correlations for Field-Test Items in Spring 2022 62
Table 38. Summary of Response Times for Field-Test Items Administered in Spring 2022 62
Table 39. Differential Item Functioning Classifications for Field-Test Items Administered in Spring 2022 63
Table 40. Groups per Grade Band for the Spring 2018 Core Calibration 68
Table 41. Spring 2018 State-Sharing Matrix 70
Table 42. Groups per Grade Band for the Spring 2019 Calibration of Operational Items 71
Table 43. Number of Common Elementary School Operational Items Administered and Calibrated in Spring 2019 72
Table 44. Number of Common Middle School Operational Items Administered and Calibrated in Spring 2019 73
Table 45. Number of Common High School Operational Items Administered and Calibrated in Spring 2019 74
Table 46. Groups per Grade Band for the Spring 2019 Calibration of Field-Test Items 74
Table 47. Estimated Latent Means and Number of Students per State 76
Table 48. Groups per Grade Band for the Spring 2021 Calibration of Field-Test Items 77
Table 49. Groups per Grade Band for the Spring 2022 Calibration of Field-Test Items 77
Table 50. Science Reporting Scale Linear Transformation Constants, Theta, and Corresponding Scaled-Score Limits for Extreme Ability Estimates (for 2021 θ Scale) 84
Table 51. Achievement-Level Cut Scores 85

LIST OF FIGURES

Figure 1. Directed Graph of the Science Item Response Theory Model 66
Figure 2. North Dakota State Assessment Item Difficulty and Student Proficiency Distributions, Grade 4 78
Figure 3. North Dakota State Assessment Item Difficulty and Student Proficiency Distributions, Grade 8 78
Figure 4. North Dakota State Assessment Item Difficulty and Student Proficiency Distributions, Grade 10 79

LIST OF APPENDICES

Appendix A. Distribution of Scale Scores and Achievement Levels for NDSA Science
Appendix B. Distribution of Scale Scores by Science Discipline for NDSA Science
Appendix C. Distribution of Scale Scores and Achievement Levels by Subgroup for NDSA Science

1. Introduction

The North Dakota State Assessment (NDSA) for Science is an assessment for grades 4, 8, and 10. The *North Dakota State Assessment for Science 2021–2022 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 is delivered as six separate, self-contained volumes, as listed below:

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 Achievement Standards.** This volume documents the methods and results of the standard-setting process used for the NDSA for Science.
4) **Evidence of Reliability and Validity.** This volume provides technical summaries of the test quality and special studies conducted to support the intended uses and interpretations of the test scores.
5) **Test Administration.** This volume describes the security protocols, accessibility features (including accommodations), methods used, and system characteristics developed to administer tests.
6) **Score Interpretation Guide.** This volume describes the score types reported and details the appropriate inferences that can be drawn from each reported score.

The North Dakota Department of Public Instruction (NDDPI) communicates the quality of the NDSA for Science by making these technical reports accessible to the public on the state's website.

1.1 Background and Historical Context of Tests

In February of 2019, North Dakota adopted three-dimensional science standards based on *A Framework for K–12 Science Education* (National Research Council, 2012) as the new North Dakota Science Content Standards. The three-dimensional science standards reflect the latest research and advances in modern science education. They describe specific performances that demonstrate what students know and can do. Each standard incorporates three dimensions: a science or engineering practice, a disciplinary core idea, and a crosscutting concept; for more information, refer to Volume 3, Section 3, The 2019 North Dakota Science Content Standards, and Volume 2, Test Development. The NDDPI and its assessment vendor, Cambium Assessment, Inc. (CAI), developed and administered a new online assessment to measure the new standards. The NDSA for Science was administered operationally for the first time in 2020–2021 and measured the science knowledge and skills of North Dakota students in grades 4, 8, and 10.

The NDDPI provides an overview of the NDSA for Science at: https://www.nd.gov/dpi/districtsschools/assessment/ndsa. Information about the 2019 North

Dakota Science Content Standards is available at https://www.nd.gov/dpi/districtsschools/k-12-education-content-standards.

1.2 Purpose and Intended Uses of the North Dakota State Assessment for Science

The NDSA for Science is a standard-referenced[1] 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 2019 North Dakota Science Content Standards) established 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 North Dakota Science Content Standards reflect the latest research and advances in modern science and differ from previous science standards in multiple ways. First, rather than describe 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 North Dakota Science Content Standards refer to such performed knowledge and skills as *performance standards (standards)*[2]. Second, while unidimensionality is a typical goal of standards (and the items that measure them), the North Dakota Science Content Standards are intentionally multi-dimensional. Each performance standard incorporates all three dimensions from *A Framework for K–12 Science Education* (National Research Council, 2012)—a science or engineering practice, a disciplinary core idea, and a crosscutting concept. Another unique feature of the North Dakota Science Content Standards is the assumption that students should learn all science disciplines, rather than a select few, as is traditionally done in many high schools, where students may elect, for example, to take biology and chemistry but not physics or astronomy.

The NDDPI supervises the development, implementation, and evaluation of the NDSA, the statewide assessment that measures student performance against the state's challenging content and achievement standards in select academic subjects and grades. The primary purpose of the NDSA for Science is to yield accurate information on student achievement accessed by the North Dakota Science Content Standards. The NDSA for Science measures the science knowledge and skills of North Dakota students in grades 4, 8, and 10. The NDSA for Science test scores are useful indicators for understanding individual students' academic achievement of the North Dakota Science Content Standards and for evaluating whether students' performance is improving over time. Both individual and aggregated scores on the NDSA for Science support instruction and student learning by providing valuable feedback to educators and parents, which can be used to form instructional strategies to remediate or enrich instruction. An array of reporting metrics is provided so that achievement can be evaluated at the student level and at aggregate levels and to monitor improvement at the student and group levels over time.

The NDSA for Science tests draw all items from the Independent College and Career Readiness (ICCR) item pool, which is part of the Shared Science Assessment Item Bank that consists of items

[1] The term "standard-referenced" is used throughout this technical report, as suggested by the NDSA Technical Advisory Committee. Note that the term "criterion-referenced" is used by NDDPI on the official website of the State of North Dakota; for example, see https://www.nd.gov/dpi/districtsschools/assessment/ndsa.

[2] North Dakota educators refer to the academic standards as performance standards. Achievement standards, on the other hand, refer to cut scores set during standard setting in this technical report.

owned by several other states and one U.S. territory. Each of these states and one U.S. territory has signed a Memorandum of Understanding (MOU) to share content, leadership, and new ideas and methods; for more information, refer to Volume 2, Test Development. Full members of the MOU in 2022 were Connecticut, Hawaii, Idaho, Montana, Oregon, Rhode Island, Utah, Vermont, 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 NDDPI to ensure that items in the tests constructed for all grades uniquely measured students' mastery of the three-dimensional North Dakota Science Content Standards.

Table 1 outlines the required uses and citations for the NDSA for Science based on the North Dakota Century Code Title 15.1. Elementary and Secondary Education (www.legis.nd.gov/cencode/t15-1c21.pdf) and the federal *Every Student Succeeds Act* (ESSA) plan. The NDSA for Science fulfills all the requirements described in Table 1.

Table 1. Required Uses and Citations for the NDSA for Science

Required Use	Required Use Citation
Required Use	**Required Use Citation**
Indicator of academic achievement and progress	ESSA section 1111(b)(2)(B)(ii)
Test administration frequency and grade levels	ESSA section 1111(b)(2)(B)(v)(II) North Dakota Century Code §15.1-21-08.1
Disaggregation of test scores	ESSA section 1111(b)(2)(B)(xi) North Dakota Century Code §15.1-21-09
Publication of test scores	ESSA section 1111(b)(2)(B)(x) North Dakota Century Code §15.1-21-10
Requirement of the alignment of test to academic content standards	North Dakota Century Code §15.1-21-11

1.3 PARTICIPANTS IN THE DEVELOPMENT AND ANALYSIS OF THE NDSA FOR SCIENCE

The NDDPI manages the NDSA for Science with the assistance of several participants, including North Dakota educators, a Technical Advisory Committee (TAC), and vendors. The NDDPI fulfills the diverse requirements of implementing North Dakota's statewide assessments while meeting or exceeding the guidelines established in the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014).

1.3.1 North Dakota Department of Public Instruction

The Office of Assessment manages test development, administration, scoring, and reporting of results for the NDSA program, including coordinating with other NDDPI offices, North Dakota public schools, and vendors.

1.3.2 North Dakota Educators

North Dakota educators participate in most aspects of the conceptualization and development of the NDSA for Science. Science panels established by NDDPI assist in the various activities that surround the academic content standards, such as the development of the academic standards, the clarification of how these standards are assessed, the test design, and the review of test items and passages.

1.3.3 Cambium Assessment, Inc.

CAI 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 NDSA for Science described in this report. Additionally, CAI is responsible for developing and maintaining the ICCR item bank.

1.3.4 Caveon Test Security

Caveon Test Security monitored web pages and social media during the spring 2022 test administration to ensure that any secure testing materials, such as items and prompts, were not disclosed. No science content breaches were detected before, during, or after the spring 2022 NDSA for Science test administration.

1.4 Available Test Formats and Special Versions

The NDSA for Science is administered online using an adaptive test design (see details of the adaptive test design in Section 3.3, Test Design). Science items are centered on a scientific phenomenon. They can consist of shorter (stand-alone) items or items with several parts (referred to as *item clusters*) that require the student to interact with the item in various ways. The NDSA for Science was administered as an operational test in spring 2022. Starting in 2021 and thereafter, additional items will be field-tested to build out the item bank.

Students unable to participate in the online administration have the option to use print-on-demand—a feature that prints the same items administered to students online so that students may read them in a paper format; student responses are still made in the online test. Spanish versions of the NDSA for Science (developed to meet the same content standards as the English versions, as well as being online and adaptive) are available for all tested grades as a designated support for students with qualifying needs. Students participating in the computer-based NDSA for Science 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 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 include braille, American Sign Language (ASL), closed captioning, and large print. Students with disabilities have the option to take the NDSA for Science with or without accommodations or to take an alternate assessment. For additional information about the test features and accommodations, refer to Volume 5, Test Administration.

1.5 Student Participation

The NDSA for Science 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 for the analyses in this technical report (number reported). The number of students reported excludes students who tested but did not have a valid score (e.g., the student opened the test and viewed the first item but abandoned the test without responding to any item). Table 3 shows the demographic characteristics of the student population, in counts and in percentages, in the spring administration of the 2021–2022 NDSA for Science. The subgroups reported here are gender, ethnicity, students with limited English proficiency (LEP), students in special education programs, and students from disadvantaged socioeconomic backgrounds.

Table 2. Number of Students Participating in the NDSA for Science in Spring 2022

Grade	Number Tested	Number Reported
4	9,132	9,131
8	8,743	8,736
10	7,882	7,873

Table 3. Distribution of Demographic Characteristics of Student Population

Group	Grade 4		Grade 8		Grade 10	
	N	*%*	*N*	*%*	*N*	*%*
All Students	9,131	100.00	8,736	100.00	7,873	100.00
Female	4,471	48.97	4,300	49.22	3,834	48.70
Male	4,660	51.03	4,436	50.78	4,039	51.30
African American	466	5.10	415	4.75	377	4.79
American Indian/Native Alaskan	692	7.58	728	8.33	588	7.47
Asian	128	1.40	113	1.29	93	1.18
Hispanic	522	5.72	470	5.38	384	4.88
Multi-Racial	494	5.41	431	4.93	320	4.06
Pacific Islander	23	0.25	26	0.30	18	0.23
White	6,806	74.54	6,553	75.01	6,093	77.39
Limited English Proficiency	427	4.68	276	3.16	192	2.44
Special-Education	1,281	14.03	1,066	12.20	882	11.20
Economically Disadvantaged	2,263	24.78	2,051	23.48	1,554	19.74

2. Operational Practices and Procedures

2.1 Test Administration

Table 4 shows the testing windows for the 2021–2022 NDSA for Science.

Table 4. NDSA for Science Testing Windows

Assessment	Grades	Testing Window
Science Online	4, 8, and 10	03/14/22–05/06/22
Braille Science	4, 8, and 10	03/14/22–05/06/22
Spanish Science	4, 8, and 10	03/14/22–05/06/22

The key personnel involved with the NDSA for Science test administration for the North Dakota Department of Public Instruction (NDDPI) are district administrators (DAs), district test coordinators (DTCs), school test coordinators (SCs), and test administrators (TAs). A *Test Administration Manual* (available at https://ndsa.portal.cambiumast.com/resources/administration-resources/test-administration-manual) was provided so that personnel involved with the statewide assessment administrations could maintain both standardized administration conditions and test security.

A secure browser developed by Cambium Assessment, Inc. (CAI) was required to access the online NDSA for Science. 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 did not have a fixed 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.

2.2 Simulations

CAI delivers the NDSA for Science under an adaptive test design and employs a simulation approach to all NDSA for Science tests. Simulations are performed before the operational testing window begins. The test is delivered using an item-selection algorithm in which operational items are selected on the fly on the basis of a student's performance on past items while ensuring that the test blueprint is followed for each individual student. Simulations were carried out to configure the algorithm settings, to evaluate whether individual tests adhered to the test blueprint, and to monitor item exposure rates. The simulation approaches and results are discussed in Volume 2, Test Development.

2.3 Universal Features, Designated Supports, and Accommodations

The NDSA for Science provides embedded (digitally provided) and non-embedded (non-digitally or locally provided) universal features for all students as they access instructional or assessment content. In addition to those universal features, designated supports and accommodations are provided for students who have special needs.

The accessibility supports discussed in this document 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 supports 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), Individualized Learning Plan (ILP), or Section 504 Plan. For English learners (ELs), Spanish-language versions of the NDSA for Science are available.

Scores achieved by students using designated supports are included for federal accountability purposes. All educators making these decisions 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 North Dakota SCs and TAs are responsible for ensuring that arrangements for accommodations are made before the test administration dates. The available accommodation options for eligible students include the following: braille, American Sign Language (ASL), closed captioning, streamlined mode, abacus, assistive technology (e.g., adaptive keyboards, touch screen, and switches), calculation device, print-on-demand, multiplication table, scribe, and speech-to-text.

Descriptions of each of these designated supports and accommodations can be found in Volume 5, Test Administration. For more details on the accessibility supports, refer to the *North Dakota Accessibility Manual* at https://ndsa.portal.cambiumast.com/resources/administration-resources/nddpi-accessibility-manual.

Table 5 and Table 6 list the number of testing sessions in which a student was provided with each designated support or accommodation during the 2021−2022 test administration.

Table 5. Number of Testing Sessions with Allowed Designated Supports

Designated Supports	Grade		
	4	*8*	*10*
Embedded			
Language-Spanish	3	5	5
Masking	59	33	6
Spanish Glossary	8	16	13
Streamlined Mode	5	24	15
Text-to-Speech Tracking	124	38	36
Text-to-Speech: Constructed Response	117	77	51
Text-to-Speech: Items	-		
Text-to-Speech: Stimuli and Items	-		
Non-Embedded			
Color Contrast	3	1	4
Color Overlay	8	5	4
Magnification	5	3	2
Noise Buffer	42	16	14
Read Aloud: Items	414	301	276
Read Aloud: Stimuli	111	31	49
Read Aloud: Stimuli and Items	-		
Read Aloud: Stimuli and Items (Spanish)	5	-	4
Separate Setting	834	733	610
Simplified Test Directions	87	70	15
Translated Test Directions	9	37	15

Table 6. Number of Testing Sessions with Allowed Accommodations

Accommodations	Grade		
	4	*8*	*10*
Embedded			
Embedded Speech-to-Text	78	78	25
Permissive Mode	1	3	3
Non-Embedded			
Alternate Response Options (Requires Permissive Mode)	5	5	2
Braille Test Booklet	-	-	1
Calculator	70	340	387

Accommodations	Grade		
	4	8	10
Print on Request	8	13	4
Speech-to-Text (Requires Permissive Mode)	190	160	122

3. ITEM BANK AND TEST DESIGN

3.1 SHARED SCIENCE ASSESSMENT ITEM BANK

Cambium Assessment, Inc. (CAI) has built and maintained the Independent College and Career Readiness (ICCR) science item pool in partnership with several states and one U.S. territory[3]. These CAI-owned items make up a substantial part of the item bank and are shared with partner states and one U.S. territory. A detailed description of the Shared Science Assessment Item Bank development process is included in Volume 2, Test Development. All these items follow the same specifications, test development processes, and review processes. In 2018, CAI field tested more than 540 item clusters and stand-alone items, of which 451 (including items from all sources) were accepted and made available as operational items in 2019. In 2019, 347 item clusters and stand-alone items were field tested, of which 268 were accepted and made available for operational use in future years. In 2021, 545 item clusters and stand-alone items were field tested, of which 458 were accepted and will be made available for operational use in future years. In 2022, 471 item clusters and stand-alone items were field tested, of which 403 were accepted and will be made available for operational use in future years.

The NDSA for Science uses only ICCR items, but because the ICCR science items are part of the Shared Science Assessment Item Bank, the latter will be described in this technical report. The Shared Science Assessment Item Bank was used for operational accountability tests in 13 states and one U.S. territory in 2022.

CAI's process for developing and field testing science items is detailed in Volume 2, Test Development. Here, note that the following best practices have been implemented at every turn:

- The goals, uses, and claims that resulting tests would be designed to support were identified in a collaborative meeting on August 22–23, 2016, as an attempt to facilitate the transition from a framework for three-dimensional science standards, specifically the Next Generation Science Standards (NGSS), to statewide summative assessments for science. CAI invited content and assessment leaders from 10 states, as well as four nationally recognized experts who helped co-author the NGSS. Two nationally recognized psychometricians also participated.

[3] CAI also works with a group of states and one U.S. territory, who have signed a Memorandum of Understanding (MOU) to share content, leadership, and new ideas and methods, to develop science assessments to measure a three-dimensional conceptualization of science understanding and other standards influenced by the same science framework. Many of these participants also share item specifications and items. CAI has coordinated this group and holds contracts to develop and deliver the items for most of them. Without being a full member of the MOU, North Dakota observed and participated in some MOU activities in 2022.

- CAI staff and participating states collaborated to develop items, as well as item specifications (documents that are designed to guide the work of item writers as they craft test questions and the reviews of those items by stakeholders). 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.
- Items were reviewed by science experts in at least one state.
- Every item was reviewed by a content advisory committee (composed 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 and bias and sensitivity in at least one state or a cross-state educator review.
- Every item was field tested, and items with questionable data were reviewed again by committees of educators.
- All scoring protocols (i.e., rubrics) were validated.
- In 2017, cognitive lab studies were carried out to evaluate and refine the process of developing item clusters aligned to three-dimensional science standards. Results of the cognitive lab studies confirmed the feasibility of the approach (see Volume 4, Section 6.1, Cognitive Laboratory Studies).
- A second set of cognitive lab studies was carried out in 2018 and 2019 to determine whether students using braille can understand the task demands of selected accommodated three-dimensional science-aligned item clusters and navigate the interactive features of these item clusters in a manner that allows students 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 (see Volume 4, Section 6.1, Cognitive Laboratory Studies).

3.2 FIELD TESTING

All items that are part of the operational pool were field tested in 2018, 2019, 2021, and 2022 as described in Section 3.2.1, 2018 Field Test; Section 3.2.2, 2019 Field Test, Section 3.2.3, 2021 Field Test, and Section 3.2.4, 2022 Field Test.

3.2.1 2018 Field Test

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 (New Hampshire, Utah, Vermont, and

West Virginia), an operational field test was administered, meaning tests consisted of 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, 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 7 presents the number of item clusters and stand-alone items administered in each grade band for each state.

Table 7. Number of Field-Test Items Administered in Spring 2018

Grade Band and Item Type	CT	HI	MSSA[a]	NH	OR	UT	WV	WY	Entire Bank
Elementary School	**135**	**24**	**69**	**58**	**26**	**–**	**91**	**14**	**153 (65)**
Cluster	78	13	40	34	20	–	56	6	86 (34)
Stand-Alone	57	11	29	24	6	–	35	8	67 (31)
Middle School	**174**	**27**	**56**	**55**	**28**	**98**	**123**	**17**	**241 (59)**
Cluster	115	13	26	30	22	98	90	5	171 (31)
Stand-Alone	59	14	30	25	6	–	33	12	70 (28)
High School	**149**	**23**	**75**	**60**	**38**	**–**	**–**	**14**	**151 (63)**
Cluster	81	14	34	33	30	–	–	6	83 (34)
Stand-Alone	68	9	41	27	8	–	–	8	68 (29)
Total	**458**	**74**	**200**	**173**	**92**	**98**	**214**	**45**	**545 (187)**

Note. ICCR items are indicated in the parentheses.
[a]MSSA = Rhode Island and Vermont's Multi-State Science Assessment.

For the states with a separate field-test segment (states with a legacy science test), including Hawaii, 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 spiraled across students. For the independent and operational field tests (except for Utah), items were administered using a linear-on-the-fly (LOFT) test design. The difference between the test design for the independent field tests and operational field tests depended on the test blueprint. For the independent field tests, the only blueprint constraint imposed was that students received four stand-alone items and two item clusters for each of the three science disciplines, whereas a full blueprint was implemented for the states with an operational field test.

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. Table 8, Table 9, and Table 10 present the number of item clusters and stand-alone items that were in common between the item pools of any two states. The numbers below the shaded diagonal elements represent the numbers for all the field-test items, and the numbers above the shaded diagonal elements represent the number of common items at the time of the 2018 calibration. The shaded diagonal elements represent the number of items that were administered only in the given state, with the number of unique items at the time of calibration provided in parentheses. Table 8 presents the results for elementary school, Table 9 presents the results for middle school, and Table 10 presents the results for high school. The numbers at field testing are slightly different from the numbers at calibration for a variety of reasons, such as items not passing rubric validation and versioning issues for some items in some states.

Table 8. Number of Common Elementary School Field-Test Items Administered and Calibrated in Spring 2018

	State	CT	HI	MSSA[a]	NH	OR	UT	WV	WY
Cluster	CT	**3 (3)**	9	36	28	16	–	49	6
	HI	10	**0 (0)**	7	8	5	–	12	1
	MSSA	36	8	**0 (2)**	15	12	–	26	2
	NH	30	8	17	**1 (3)**	5	–	22	2
	OR	17	5	13	5	**1 (1)**	–	5	1
	UT	–	–	–	–	–	**–**	–	–
	WV	49	12	27	25	5	–	**0 (4)**	2
	WY	6	1	2	2	1	–	2	**0 (0)**
Stand-Alone	CT	**1 (3)**	5	25	22	2	–	33	7
	HI	5	**6 (6)**	0	0	0	–	4	0
	MSSA	26	0	**0 (1)**	10	4	–	13	3
	NH	24	0	11	**0 (2)**	0	–	15	2
	OR	2	0	4	0	**1 (1)**	–	0	0
	UT	–	–	–	–	–	**–**	–	–
	WV	35	4	14	17	0	–	**0 (2)**	1
	WY	8	0	3	3	0	–	2	**0 (1)**
Grade Band Total	CT	**4 (6)**	14	61	50	18	–	82	13
	HI	15	**6 (6)**	7	8	5	–	16	1
	MSSA	62	8	**0 (3)**	25	16	–	39	5
	NH	54	8	28	**1 (5)**	5	–	37	4
	OR	19	5	17	5	**2 (2)**	–	5	1
	UT	–	–	–	–	–	**–**	–	–
	WV	84	16	41	42	5	–	**0 (6)**	3
	WY	14	1	5	5	1	–	4	**0 (1)**

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

Table 9. Number of Common Middle School Field-Test Items Administered and Calibrated in 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)**
Grade Band 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)**

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

Table 10. Number of Common High School Field-Test Items Administered and Calibrated in Spring 2018

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

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

The common item design was used to calibrate all the items on a common science scale. The calibration model is explained in detail in Section 5, Item Calibration.

Following the test administration, items went through a substantial validation process. The process began with rubric validation. In the science test, *scoring assertions* capture each measurable action of an item and articulate what evidence the student has provided to infer a specific skill or concept, while *rubrics* establish criteria, including rules, principles and illustrations, to communicate expectations of students' success in providing this evidence (see Section 4.2 of Volume 3, Setting Achievement Standards, for a sample scoring assertion and rubric). Rubric validation is a process in which a committee of state educators reviews student responses and the proposed scoring of those responses. The responses reviewed are scientifically sampled to overrepresent responses that were most likely to have been mis-scored. Specifically, the sample overrepresents two types of responses: (1) low-scored responses from otherwise high-scoring students, and (2) high-scored responses from otherwise low-scoring students.

During rubric validation, educators recommend revisions to rubrics where necessary. CAI staff revise the rubrics and rescore the entire sample to ensure that the rubric changes have all and only the intended effects.

Following 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 states established 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. A common set of business rules was defined for all the states participating in the operational field test, although some states decided to include additional items for data review. The item statistics were computed on the basis of student data for the students testing in the state that owned the item. For Rhode Island and Vermont, which share their item development, the statistics were computed on the basis of combined data. For ICCR items, the data from Connecticut, New Hampshire, Rhode Island, Vermont, and West Virginia (states that used ICCR items and with either an independent or operational field test) were combined. For each state, a data review committee consisting of educators (i.e., science teachers) and 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, 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 ICCR items.

Table 11. Overview of Field-Test Item Administration, Rubric Validation, and Item Data Review in 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 (65)**	**3 (0)**	**65 (26)**	**13 (3)**	**137 (62)**
Cluster	86 (34)	3 (0)	24 (7)	5 (1)	78 (33)
Stand-Alone	67 (31)	0 (0)	41 (19)	8 (2)	59 (29)
Middle School	**241 (59)**	**16 (0)**	**102 (26)**	**24 (3)**	**201 (56)**
Cluster	171 (31)	12 (0)	65 (11)	15 (1)	144 (30)
Stand-Alone	70 (28)	4 (0)	37 (15)	9 (2)	57 (26)
High School	**151 (63)**	**10 (2)**	**80 (31)**	**13 (2)**	**128 (59)**
Cluster	83 (34)	8 (2)	35 (14)	4 (0)	71 (32)
Stand-Alone	68 (29)	2 (0)	45 (17)	9 (2)	57 (27)
Total	**545 (187)**	**29 (2)**	**247 (83)**	**50 (8)**	**466 (177)**

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

Table 12 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 ICCR items.

Table 12. Overview of Shared Science Assessment Item Bank in Spring 2018

Grade Band and Item Type	Science Discipline			Total[a]
	Earth and Space Sciences	*Life Sciences*	*Physical Sciences*	
Elementary School	**41 (19)**	**47 (21)**	**49 (22)**	**137 (62)**
Cluster	23 (11)	29 (11)	26 (11)	78 (33)
Stand-Alone	18 (8)	18 (10)	23 (11)	59 (29)
Middle School	**56 (16)**	**72 (19)**	**70 (21)**	**198 (56)**
Cluster	41 (9)	49 (7)	51 (14)	141 (30)
Stand-Alone	15 (7)	23 (12)	19 (7)	57 (26)
High School	**37 (19)**	**53 (23)**	**38 (17)**	**128 (59)**
Cluster	19 (8)	32 (15)	20 (9)	71 (32)
Stand-Alone	18 (11)	21 (8)	18 (8)	57 (27)
Total	**134 (54)**	**172 (63)**	**157 (60)**	**463 (177)**

Note. ICCR items are indicated in the parentheses.
[a] Excluding three Utah-owned middle school item clusters aligned to Utah-specific standards.

3.2.2 2019 Field Test

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 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 the entire bank presents the number of ICCR items.

Table 13. Number of Field-Test Items Administered in Spring 2019

Grade Band and Item Type	CT	HI	ID	MSSA[a]	NH	OR	WV	WY	Entire Bank
Elementary School	**47**	**31**	**53**	**42**	**18**	**27**	**18**	**16**	**117 (18)**

Grade Band and Item Type	CT	HI	ID	MSSA[a]	NH	OR	WV	WY	Entire Bank
Cluster	18	19	20	17	0	16	10	5	50 (0)
Stand-Alone	29	12	33	25	18	11	8	11	67 (18)
Middle School	**56**	**23**	**53**	**46**	**28**	**26**	**26**	**15**	**127 (28)**
Cluster	14	9	17	10	4	9	8	5	38 (4)
Stand-Alone	42	14	36	36	24	17	18	10	89 (24)
High School	**69**	**21**	**–**	**37**	**29**	**28**	**–**	**25**	**103 (29)**
Cluster	25	14	–	18	2	13	–	2	35 (2)
Stand-Alone	44	7	–	19	27	15	–	23	68 (27)
Total	**172**	**75**	**106**	**125**	**75**	**81**	**44**	**56**	**347 (75)**

Note. ICCR items are indicated in the parentheses.
[a]MSSA = Rhode Island and Vermont's Multi-State Science Assessment.

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 spiraled across students. 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 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. 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.

For any given state, a minimum sample size of 1,500 students per field-test item was targeted. Most items were administered in two or more states. Tables 14–16 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 diagonal elements represent the numbers for all administered field-test items, and the numbers above the shaded diagonal elements represent the number of common field-test items at the time of calibration. The shaded diagonal elements represent the number of field-test items that were administered only in the given state (with the number of unique field-test items at the time of calibration in parentheses). 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 in 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)**
Grade Band 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)**

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

Table 15. Number of Common Middle School Field-Test Items Administered and Calibrated in 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)**
Grade Band 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)**

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

Table 16. Number of Common High School Field-Test Items Administered and Calibrated in 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)**
Grade Band 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)**

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

The calibration and linking of the field-test items in 2019 is explained in detail in Section 5.2, Item Calibration.

Following essentially the same process explained in Section 3.2.1, 2018 Field Test, items went through a substantial validation process. The following are minor modifications to the process followed in 2018:

- In 2018, all the item statistics were computed on the basis of student data for the students testing in the state that owned the item. In 2019, all item statistics were computed on the basis of student data for the students testing in the state that owned the item, *except for the statistics related to DIF*. Following recommendations of several Technical Advisory

Committees (TACs), the data of states were combined in the calculation of DIF statistics whenever possible (i.e., for states with an independent field test or an operational test for which the relevant demographic variable was available).

- In 2018, for ICCR items, the data from Connecticut, New Hampshire, Rhode Island, Vermont, and West Virginia (i.e., the states that used ICCR items and with either an independent or operational field test) were combined. In 2019, these states were Connecticut, Idaho (for middle school only), New Hampshire, Oregon, Rhode Island, Vermont, and West Virginia.
- The business rule to flag an item cluster for DIF was slightly modified by making it more liberal following recommendations of several TACs. The modification is discussed in Section 4.4, Differential Item Functioning Analysis.

Table 17 presents the number of field-test items administered, 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 ICCR items.

Table 17. Overview of Field-Test Item Administration, Rubric Validation, and Item Data Review in Spring 2019

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	**117 (18)**	**2 (0)**	**72 (16)**	**24 (0)**	**91 (18)**
Cluster	50 (0)	1 (0)	16 (0)	10 (0)	39 (0)
Stand-Alone	67 (18)	1 (0)	56 (16)	14 (0)	52 (18)
Middle School	**127 (28)**	**6 (8)**	**66 (15)**	**21 (1)**	**97 (19)**
Cluster	38 (4)	1 (4)	12 (0)	5 (0)	29 (0)
Stand-Alone	89 (24)	5 4()	54 (15)	16 (1)	68 (19)
High School	**103 (29)**	**6 (4)**	**52 (12)**	**15 (1)**	**80 (24)**
Cluster	35 (2)	2 (2)	15 (0)	5 (0)	26 (0)
Stand-Alone	68 (27)	4 (2)	37 (12)	10 (1)	54 (24)
Total	**347 (75)**	**14 (12)**	**190 (43)**	**60 (2)**	**268 (61)**

Note: ICCR items are indicated in the 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 18 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 ICCR items.

Table 18. Overview of Shared Science Assessment Item Bank in Spring 2019

Grade Band and Item Type	Science Discipline			Total
	Earth and Space Sciences	*Life Sciences*	*Physical Sciences*	
Elementary School	**68 (23)**	**77 (28)**	**80 (28)**	**225 (79)**
Cluster	33 (11)	40 (11)	40 (10)	113 (32)
Stand-Alone	35 (12)	37 (17)	40 (18)	112 (47)
Middle School	**83 (23)**	**108 (26)**	**92 (20)**	**287 (69)**
Cluster	44 (8)	62 (6)	53 (11)	163 (25)
Stand-Alone	39 (15)	46 (20)	39 (9)	124 (44)
High School	**39 (17)**	**108 (46)**	**53 (16)**	**200 (79)**
Cluster	18 (6)	48 (14)	24 (8)	90 (28)
Stand-Alone	21 (11)	60 (32)	29 (8)	110 (51)
Total	**190 (63)**	**293 (100)**	**225 (64)**	**712 (227)**

Note. ICCR items are indicated in the parentheses.

3.2.3 2021 Field Test

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, Utah, Vermont, 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 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 present the number of ICCR field-test items.

Table 19. 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	Entire Bank
Elementary School	**36**	**22**	**140**	**55**	**21**	**11**	**19**	**8**	**54**	**19**	**17**	**214 (51)**
Cluster	16	6	58	18	7	3	3	3	54	7	5	106 (9)
Stand-Alone	20	16	82	37	14	8	16	5	0	12	12	108 (42)
Middle School	**33**	**19**	**129**	**54**	**20**	**11**	**18**	**11**	**45**	**19**	**20**	**159 (47)**
Cluster	17	6	44	18	7	3	2	2	45	7	4	60 (8)
Stand-Alone	16	13	85	36	13	8	16	9	0	12	16	99 (39)
High School	**49**	**17**	**156**	**49**	**–**	**11**	**12**	**8**	**–**	**–**	**20**	**172 (43)**
Cluster	11	5	54	16	–	3	4	3	–	–	3	57 (15)
Stand-Alone	38	12	102	33	–	8	8	5	–	–	17	115 (28)
Total	**118**	**58**	**425**	**158**	**41**	**33**	**49**	**27**	**99**	**38**	**57**	**545 (141)**

Note. ICCR items are indicated in the parentheses.

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

For the state with a separate field-test segment (i.e., Wyoming), field-test forms were constructed using a balanced incomplete design and spiraled across students. For the independent field test, items were administered under a LOFT design, where the only blueprint constraint imposed was the discipline level constraint for stand-alone items and item clusters.

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. 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 NDSA for Science is discussed in Section 3.3, Test Design.

For any given state, a minimum sample size of 1,500 students per field-test item was targeted. Most items were administered in two or more states. Tables 20–22 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 diagonal elements represent the numbers for all administered field-test items, and the numbers above the shaded diagonal elements represent the number of common field-test items at the time of calibration. The shaded diagonal elements represent the number of field-test items that were administered only in the given state (with the number of unique field-test items at the time of calibration in parentheses). 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 in Spring 2021

	State	CT	HI	ID	MSSA[a]	MT	ND	NH	SD	UT	WV	WY
Cluster	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	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)

	State	CT	HI	ID	MSSA[a]	MT	ND	NH	SD	UT	WV	WY
Grade Band Total	CT	6 (6)	0	27	2	0	0	0	0	0	0	1
	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)

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

Table 21. Number of Common Middle School Field-Test Items Administered and Calibrated in Spring 2021

	State	CT	HI	ID	MSSA[a]	MT	ND	NH	SD	UT	WV	WY
Cluster	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	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)

	State	CT	HI	ID	MSSA[a]	MT	ND	NH	SD	UT	WV	WY
Grade Band Total	CT	2 (2)	0	21	4	0	0	0	3	10	0	2
	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)

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

Table 22. Number of Common High School Field-Test Items Administered and Calibrated in Spring 2021

	State	CT	HI	ID	MSSA[a]	MT	ND	NH	SD	UT	WV	WY
Cluster	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	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
Grade Band 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)

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

The calibration and linking of the field-test items in 2021 is explained in detail in Section 5.2, Item Calibration.

Table 23 presents the number of field-test items administered in North Dakota, 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 ICCR field-test items.

Table 23. Overview 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 (51)**	**7 (2)**	**100 (32)**	**19 (2)**	**188 (47)**
Cluster	106 (9)	5 (0)	24 (3)	7 (0)	94 (9)
Stand-Alone	108 (42)	2 (2)	76 (29)	12 (2)	94 (38)
Middle School	**159 (47)**	**15 (4)**	**87 (27)**	**13 (0)**	**129 (43)**
Cluster	60 (8)	10 (3)	22 (2)	5 (0)	43 (5)
Stand-Alone	99 (39)	5 (1)	65 (25)	8 (0)	86 (38)
High School	**172 (43)**	**9 (3)**	**94 (16)**	**22 (4)**	**141 (36)**
Cluster	57 (15)	6 (2)	27 (6)	4 (2)	47 (11)
Stand-Alone	115 (28)	3 (1)	67 (10)	18 (2)	94 (25)
Total	**545 (141)**	**31 (9)**	**281 (75)**	**54 (6)**	**458 (126)**

Note: ICCR items are indicated in the parentheses.

[a] Two Hawaii-owned items were not shared to the Shared Science Assessment Item bank.

Table 24 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 ICCR items.

Table 24. Overview 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 (42)**	**128 (41)**	**149 (43)**	**413 (126)**
Cluster	65 (14)	66 (14)	76 (13)	207 (41)
Stand-Alone	71 (28)	62 (27)	73 (30)	206 (85)
Middle School	**114 (32)**	**156 (46)**	**137 (34)**	**407 (112)**
Cluster	55 (9)	76 (9)	67 (12)	198 (30)
Stand-Alone	59 (23)	80 (37)	70 (22)	209 (82)
High School	**68 (21)**	**163 (66)**	**106 (27)**	**337 (114)**
Cluster	27 (9)	64 (18)	42 (11)	133 (38)
Stand-Alone	41 (12)	99 (48)	64 (16)	204 (76)
Total	**318 (95)**	**447 (153)**	**392 (104)**	**1,157 (352)**

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

3.2.4 2022 Field Test

In 2022, a fourth wave of items was field tested in 13 states and one U.S. territory (Connecticut, Hawaii, Idaho, Montana, New Hampshire, North Dakota, Oregon, Rhode Island, South Dakota, Utah, Vermont, West Virginia, Wyoming, and U.S. Virgin Islands,). 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. Most of the items were filed-tested in two states (97%), nine items (2%) were administered in one state only, and three items were administered in three states. Table 25 presents the number of field-test item clusters and stand-alone items administered in each grade band for each state. The numbers in parentheses present the number of ICCR field-test items.

Table 25. Number of Field-Test Items Administered in Spring 2022

Grade Band and Item Type	CT	HI	ID	MSSA[a]	MT	ND	NH	OR	SD	UT	WV	WY	USVI	Entire Bank
Elementary School	**34**	**28**	**22**	**66**	**12**	**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	8	8 **(8)**	12	26	6	-	8	8	0	82
Middle School	**40**	**30**	**35**	**64**	**12**	**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	8	8 **(8)**	12	23	6	-	28	8	0	102
High School	**46**	**14**	**14**	**58**	**-**	**12 (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	-	8 **(8)**	12	27	6	-	-	8	0	70
Total	**120**	**72**	**71**	**188**	**24**	**36 (36)**	**50**	**123**	**29**	**138**	**52**	**30**	**3**	**471**

Note. ICCR items are indicated in the parentheses.

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

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, West Virginia, Wyoming, and U.S. Virgin Islands) 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 NDSA for Science is discussed in Section 3.3, Test Design.

For any given state or territory, a minimum sample size of 1,500 students per field-test item was targeted. Most items were administered in two states or territory. Tables 26–28 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 diagonal elements represent the numbers for all administered field-test items, and the numbers above the shaded diagonal elements represent the number of common field-test items at the time of calibration. The shaded diagonal elements represent the number of field-test items that were administered only in the given state or territory (with the number of unique field-test items at the time of calibration in parentheses). Table 26 presents the results for elementary schools, Table 27 presents the results for middle schools, and Table 28 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 26. Number of Common Elementary School Field-Test Items Administered and Calibrated in Spring 2022

	State	CT	HI	ID	MSSA[a]	MT	ND	NH	OR	SD	UT	WV	WY	USVI
Cluster	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	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	0	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	0	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

	State	CT	HI	ID	MSSA[a]	MT	ND	NH	OR	SD	UT	WV	WY	USVI
	WY	0	0	0	8	0	0	0	0	0	0	0	0 (0)	0
	USVI	0	0	0	0	0	0	0	0	0	0	0	0	0 (0)
Grade Band 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	0	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	0	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)

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

Table 27. Number of Common Middle School Field-Test Items Administered and Calibrated in Spring 2022

	State	CT	HI	ID	MSSA[a]	MT	ND	NH	OR	SD	UT	WV	WY	USVI
Cluster	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	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

	State	CT	HI	ID	MSSA[a]	MT	ND	NH	OR	SD	UT	WV	WY	USVI
	WY	0	0	0	8	0	0	0	0	0	0	0	0(0)	0
	USVI	0	0	0	0	0	0	0	0	0	0	0	0	0(0)
Grade Band 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	0	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	0	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)

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

Table 28. Number of Common High School Field-Test Items Administered and Calibrated in Spring 2022

	State	CT	HI	ID	MSSA[a]	MT	ND	NH	OR	SD	UT	WV	WY	USVI
Cluster	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	0	0	0	0	0(0)	0	0	0	0	0	0	0	0
	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	0	0	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)	0	0
	WY	1	0	1	0	-	0	0	0	0	-	-	0(0)	0
	USVI	1	0	0	0	-	0	0	1	0	-	-	0	0(0)
Stand-Alone	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	0	0	0	0	0(0)	0	0	0	0	0	0	0	0
	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	0	0	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)	0	0

	State	CT	HI	ID	MSSA[a]	MT	ND	NH	OR	SD	UT	WV	WY	USVI
	WY	0	0	0	0	-	0	0	0	0	-	-	0(0)	0
	USVI	0	0	0	0	-	0	0	0	0	-	-	0	0(0)
Grade Band 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	0	0	0	0	0(0)	0	0	0	0	0	0	0	0
	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	0	0	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)	0	0
	WY	1	0	1	8	-	0	0	0	0	-	-	0(0)	0
	USVI	1	0	0	0	-	0	0	1	0	-	-	0	0(0)

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

The calibration and linking of the field-test items in 2022 is explained in detail in Section 5.2, Item Calibration.

Table 29 presents the number of field-test items administered in North Dakota, or another state or territory, 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 ICCR field-test items.

Table 29. Overview of Field-Test Item Administration, Rubric Validation, and Item Data Review in 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[a]
Elementary School	**170 (32)**	**3 (1)**	**85 (2)**	**14 (5)**	**153 (26)**
Cluster	88 **(9)**	1 **(0)**	18 **(0)**	4 **(0)**	83 **(9)**
Stand-Alone	82 **(23)**	2 **(1)**	67 **(2)**	10 **(5)**	70 **(17)**
Middle School	**190 (62)**	**4 (2)**	**99 (9)**	**26 (5)**	**160 (55)**
Cluster	88 **(26)**	3 **(1)**	26 **(3)**	13 **(1)**	72 **(24)**
Stand-Alone	102 **(36)**	1 **(1)**	73 **(6)**	13 **(4)**	88 **(31)**
High School	**111 (42)**	**2 (0)**	**64 (8)**	**19 (4)**	**90 (38)**
Cluster	41 **(13)**	2 **(0)**	21 **(2)**	3 **(0)**	36 **(13)**
Stand-Alone	70 **(29)**	0 **(0)**	43 **(6)**	16 **(4)**	54 **(25)**
Total	**471 (136)**	**9 (3)**	**248 (19)**	**59 (14)**	**403 (119)**

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

Table 30 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 ICCR items.

Table 30. Overview of Shared Science Assessment Item Bank in Spring 2022

Grade Band and Item Type	Science Discipline			Total[a]
	Earth and Space Sciences	*Life Sciences*	*Physical Sciences*	
Elementary School	**180 (46)**	**162 (44)**	**214 (52)**	**556 (142)**
Cluster	96 (18)	82 (14)	111 (17)	289 (49)
Stand-Alone	84 (28)	80 (30)	103 (35)	267 (93)
Middle School	**149 (43)**	**220 (65)**	**187 (48)**	**556 (156)**
Cluster	70 (15)	110 (21)	90 (17)	270 (53)
Stand-Alone	79 (28)	110 (44)	97 (31)	286 (103)
High School	**91 (34)**	**194 (68)**	**129 (39)**	**414 (141)**
Cluster	35 (13)	78 (20)	53 (17)	166 (50)
Stand-Alone	56 (21)	116 (48)	76 (22)	248 (91)
Total	**420 (123)**	**576 (177)**	**530 (139)**	**1 526 (439)**

Note. ICCR items are indicated in the parentheses.

[a]Two Hawaii-owned items were not shared to the Shared Science Assessment Item bank.

3.3 TEST DESIGN

The science tests were assembled under an adaptive test design, with the exception of braille. 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, an item's content value 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 an adaptive test design, operational items are selected on the fly based on the performance of a student on past items while ensuring that the test blueprint is followed for each individual student. The NDSA for Science blueprints are given in Tables 31 through Table 33. Details of CAI's item selection algorithm are described in Volume 2, Test Development, and its Appendix J, Adaptive Algorithm Design. The braille tests are accommodated fixed forms. Form construction of the accommodated forms is discussed in Volume 2, Section 4.4, Paper-Based Accommodation Form Construction.

Table 31. NDSA for Science Test Blueprint, Grade 4

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

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

Note. *These performance standards have an engineering component.

Table 32. NDSA for Science Test Blueprint, Grade 8

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

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

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

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

*These performance standards have an engineering component.

Table 33. NDSA for Science Test Blueprint, Grade 10

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

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

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

*These performance standards have an engineering component.

The main characteristics of the blueprint were that any performance standard (standards) could be tested only once (indicated by the values of 0 and 1 for the minimum and maximum values of the individual performance standards in Tables 31–33); in general, no more than one item cluster or two stand-alone items could be sampled from the same DCI, and no more than three total items could be sampled from the same DCI (as indicated by the minimum and maximum values in the rows representing DCIs).

In 2021 and 2022, a non-segmented test design was used. Students received items from different disciplines in random order. Embedded field-test items were randomly positioned in the test and randomly distributed across students. Every student received either one item cluster or four stand-alone items as field-test items throughout the test.

4. Field-Test Classical Analysis Overview

As explained in Section 3, Item Bank and Test Design, science items administered as field-test items in 2018, 2019, 2021, and 2022 in North Dakota or any of the states that signed the Memorandum of Understanding (MOU) for item sharing underwent rubric validation and data review. Items were flagged for data review on the basis of 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 responses from students in the state that owned the items; however, for Independent College and Career Readiness (ICCR) items, the flagging rules were defined by the item statistics computed from the combined data of states or territory that used ICCR items and that administered either an independent or operational test. In 2022, they were Connecticut, Idaho, New Hampshire, North Dakota, Oregon, Rhode Island, South Dakota, Utah, Vermont, and West Virginia. Furthermore, for the computation differential item functioning (DIF) statistics, the data of all states and territory with an operational or independent field test were combined to obtain a sufficient number of students for each demographic group. The criteria for flagging and reviewing items are provided in Table 34, and the statistics are described below 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 3, Item Bank and Test Design.

Table 34. 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 < 0.30 or > 0.85 (across the assertions within a cluster)
Item Difficulty (Stand-Alone Items)	Average *p*-value < 0.15 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

Analysis Type	Flagging Criteria
Timing	Assertions per 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. 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 34.

4.3 RESPONSE TIME

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

4.4 DIFFERENTIAL ITEM FUNCTIONING ANALYSIS

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 adaptation to polytomous items, and 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 statistics 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 [Dorans & Schmitt, 1991]) 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 $k = \{1, 2, \dots K\}$ 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. The variance is calculated as

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

where 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

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

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):

- Male vs. Female
- American Indian/Alaskan Native vs. White
- Native Hawaiian/Pacific Islander vs. White
- Asian vs. White
- Black or African American vs. White
- Hispanic or Latino 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 on traditional tests, assertions were classified into three categories (i.e., A, B, or C) for DIF, ranging from "no evidence of DIF" to "severe DIF". The classification rules are shown in Table 35. 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 in the same direction.

Table 35. Differential Item Functioning Classification Rules

Assertions	
Category	*Rule*
C	MH_{χ^2} is significant and $|SMD|/|SD| \geq 0.25$
B	MH_{χ^2} is significant and $|SMD|/|SD| < 0.25$
A	MH_{χ^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 for items field tested in 2018).

4.5 Classical Analysis Results

This section presents a summary of results from classical item analysis of the 2022 field-test items. Table 36 and Table 37 provide the summary of the p-values and biserial correlations for the ICCR science field-test items administered in North Dakota in 2022. The statistics were computed using North Dakota data only. The average values across the assertions within an item were used in the computation of the percentiles and ranges.

Table 36. Distribution of p-*Values for Field-Test Items in Spring 2022*

Grade	Total FT Items	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
4	12	0.17	0.19	0.27	0.35	0.40	0.44	0.44
8	12	0.24	0.28	0.33	0.40	0.48	0.55	0.60
10	12	0.02	0.09	0.18	0.22	0.36	0.52	0.53

Table 37. Distribution of Item Biserial Correlations for Field-Test Items in Spring 2022

Grade	Total FT Items	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
4	12	0.35	0.37	0.40	0.44	0.53	0.55	0.58
8	12	0.19	0.21	0.24	0.32	0.37	0.54	0.63
10	12	0.30	0.33	0.36	0.41	0.44	0.52	0.56

Table 38 presents the summary of the response times by item type (item cluster or stand-alone item) for field-test items administered in 2022.

Table 38. Summary of Response Times for Field-Test Items Administered in Spring 2022

Grade	Item Type	Total FT Items	Min	25th Percentile	50th Percentile	75th Percentile	Max
4	Cluster	4	6.50	7.63	8.50	9.30	10.20
	Stand-Alone	8	2.70	2.78	3.35	3.63	5.10
8	Cluster	4	5.70	7.50	9.00	10.53	12.40
	Stand-Alone	8	1.40	1.88	2.00	2.13	3.20
10	Cluster	4	5.80	6.55	7.75	8.73	8.80
	Stand-Alone	8	1.20	1.80	2.60	3.58	4.90

Table 39 presents, for each item type, the number of field-test items flagged for DIF for each demographic group included in the DIF analyses in 2022.

Table 39. Differential Item Functioning Classifications for Field-Test Items Administered in Spring 2022

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	Low Income/ Non-Low Income[c]
Grade 4											
Items Evaluated	Cluster	4	0	0	0	0	0	0	0	4	4
	Stand-Alone	8	0	0	0	0	5	0	3	8	8
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
	Stand-Alone	0	-	-	-	-	0	-	0	0	0
Grade 8											
Items Evaluated	Cluster	4	0	0	0	0	4	0	3	4	4
	Stand-Alone	8	0	0	0	0	0	0	0	1	8
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
	Stand-Alone	0	-	-	-	-	-	-	-	0	0
Grade 10											
Items Evaluated	Cluster	4	0	0	2	0	4	0	0	4	4
	Stand-Alone	8	0	0	4	0	8	0	0	8	8
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
	Stand-Alone	0	-	-	0	-	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 2022, 36 ICCR field-test items were administered in North Dakota; all items passed rubric validation. Among them, four were flagged for item discrimination, none was flagged for p-value, 12 items were flagged for response time, and no item was flagged for DIF according to the criteria used in 2022 (as described in Section 4.1, Item Discrimination, through Section 4.4, Differential Item Functioning). Flagged field-test items were reviewed by educators during data review; five items were rejected in the end, with one having relatively low discrimination and four with concerns on potential speededness. The total number of field-test items flagged and the total number of field-test items that passed item data review in 2022 are summarized in Table 29.

5. Item Calibration

5.1 Model Description

In discussing item response theory (IRT) models for North Dakota, 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 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 as shown below:

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

where z_{ij} represents the scored response of student j $(j = 1, \ldots, N)$ to item i $(i = 1, \ldots, I)$, $\boldsymbol{z}_j$ 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 of the NDSA for Science 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 NDSA for Science 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 (i.e., 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[4]. 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:

$$P\left(\boldsymbol{z_j}\middle|\theta_j, \mathbf{u}_j\right) = \prod_{i \in \mathrm{SA}} P\left(z_{ij}\middle|\theta_j\right) \prod_{g=1}^{G} \prod_{i \in g} P\left(z_{ij}\middle|\theta_j, u_{jg}\right), \tag{2}$$

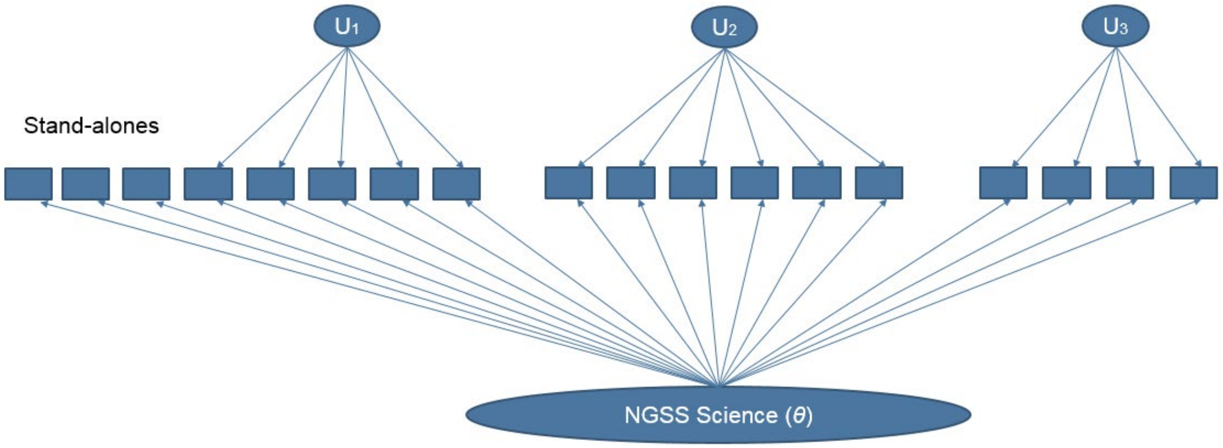
where SA indicates stand-alone item assertions, u_g indicates the nuisance dimension for assertion group g (with the position of student j on that dimension denoted as u_{jg}), and $\mathbf{u}$ is the vector of all G nuisance dimensions. It can be seen that the conditional probability $P\left(z_{ij}\middle|\theta_j, u_{jg}\right)$ 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 complications 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 all nuisance dimensions are uncorrelated; rather, it is sufficient that they are independent, given the general dimension $\boldsymbol{\theta}$.

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 item 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.

[4] The term *nuisance dimension* here pertains to within-item local dependencies among scoring assertions and should not be confused with the three dimensions of *A Framework for K–12 Science Education*.

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 Item Response Theory Model



5.1.2 Item Response Function

The item response functions of the stand-alone item 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}. The Rasch testlet model (Wang & Wilson, 2005) is adopted as the IRT model for the NDSA for Science. For binary data, the Rasch testlet model is defined as:

$$P\left(z_{ij}|\theta_j, u_{jg}; b_i\right) = \frac{\exp(\theta_j + u_{jg} - b_i)}{1 + \exp(\theta_j + u_{jg} - b_i)}. \tag{3}$$

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 (2) and (3). 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 that is incorporated in unidimensional models for polytomous data.

5.1.3 Multigroup Model

The Shared Science Assessment Item Bank was calibrated concurrently using all the items administered in any of the states or territory that collaborate with Cambium Assessment, Inc. (CAI) on their new science assessments. In the calibration, each state or territory was treated as a population of students or group. Overall group differences were considered 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 was 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 ($k = 1$) was set to 0 to identify the model. For each of the nuisance variables u_g, a common variance parameter across groups was assumed, and the means were set to 0 in order to identify the model,

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

5.2 Item Calibration

5.2.1 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,

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

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 (i.e., 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 equation above is very high, issues with 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 in 2018 after the 2018 science test administrations concluded, and it was recalibrated in 2019 following the 2019 test administrations. The 2019 parameters were used for the 2021 and future test administration. Because the calibration sequence was somewhat different between 2018 and 2019, the calibration sequences are presented in detail below for both years.

In 2018 and 2019, the IRT models were fitted using the Bayesian networks with 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 (see Section 5.2.2, 2018 Calibration Sequence). 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.

In 2021, field-test items were calibrated with one multi-group 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 MML method. Because the estimation time in flexMIRT became prohibitive, CAIRT (Cambium Assessment IRT) was used. 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 a variety of scenarios (Rijmen, Liao, Lin, 2021). In 2022, field-test items were calibrated in CAIRT using the same procedure as 2021.

5.2.2 2018 Calibration Sequence

Table 40 provides an overview of the groups per grade band for the 2018 calibration.

Table 40. 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 in those states only. Specifically, the following sequence of calibrations was carried out:

1. **Core Calibration.** The core calibration was performed on the following:
    a. All item responses for New Hampshire and West Virginia. These states administered items from the following sources (as described in the state-sharing matrix in Table 41):
        i. Independent College and Career Readiness (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 3.2.1, 2018 Field Test (see Tables 8–10).

    b. All item responses from Connecticut, Rhode Island, and Vermont except for the responses to Wyoming and Oregon items. These states administered items from the following sources:
        i. ICCR
        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).
    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). Utah tested only middle school students but included every grade in middle school. One third of students were selected at random to balance the large population size for Utah.

Table 41. Spring 2018 State-Sharing Matrix

Source Bank	CT	HI	MSSA	NH	OR	UT	WV	WY
ICCR	X	X	X	X	X		X	X
Connecticut	X		X				X	
Hawaii	X	X	X				X	
MSSA[a]	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.
[a]MSSA = Rhode Island and Vermont's Multi-State Science Assessment.

2. **Calibration of State-Specific Items.** Both Hawaii and Utah had a substantial proportion of items that were only administered in Hawaii and Utah, respectively. Hawaii has both Hawaii and ICCR items in common with the states of the core calibration (Hawaii administered only Hawaii and ICCR items); Utah has only Utah items in common (Utah only administered Utah items). The parameters for the unique Hawaii items depended only on responses from Hawaii students, and the parameters for the unique Utah items depended only on responses from Utah students. 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 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 from 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, as well as the group means and standard deviations that were estimated in Step 1.

5.2.3 2019 Calibration Sequence

The 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 42 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. There were nine items administered as operational items in 2019 for which the sample size was smaller than 1,000, out of a total of 438 items.

Tables 43 through Table 45 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 diagonal elements represent the numbers for all the operational items administered, and the numbers above the shaded diagonal elements represent the number of common operational items at the time of the 2019 calibration. The shaded diagonal elements represent the number of operational items that were 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 only administered in one state, the numbers above the diagonal are the same as the numbers below the diagonal.

Table 43 presents the results for elementary schools, Table 44 presents the results for middle schools, and Table 45 presents the results for high schools. The numbers at the operational administration are slightly different from the numbers at the calibration because items with sample sizes smaller than 1,000 were excluded from the calibration.

Table 42. 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 43. Number of Common Elementary School Operational Items Administered and Calibrated in 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)**
Grade Band 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)**

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

Table 44. Number of Common Middle School Operational Items Administered and Calibrated in 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)**
Grade Band 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)**

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

Table 45. Number of Common High School Operational Items Administered and Calibrated in Spring 2019

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

Note. [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 46 provides an overview of the groups per grade band for calibration of the field-test items.

Table 46. 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

5.2.4 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 the 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 is not determined. In the multigroup Rasch testlet model, the only scale indeterminacy is 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 are on the same scale except for an overall shift parameter B. Specifically, the 2018 scale can be linked to the 2019 scale as follows:

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

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

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

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

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

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

Item parameters based on 2018 student responses can be 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 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). So, 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 can be estimated from a separate calibration of the items operational in 2019 using the 2019 student responses (of 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 (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 47 presents the estimated means of θ under both the free and anchored calibrations, as well as the number of students per state. Table 47 also presents the overall means and estimated shift in parameter B. Note that the parameters for three items were not anchored but 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 was 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 (all states except Rhode Island). Therefore, out of an abundance of caution, these three items were not anchored to their 2018 parameter values.

Table 47. 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

5.2.5 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 MML method. The calibration included 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 48 and Table 49 provides an overview of the groups per grade band for calibration of the field-test items in 2021 and 2022, respectively.

Table 48. 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	X
Montana	X	X	
New Hampshire	X	X	X
North Dakota	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 49. 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	X
Montana	X	X	
New Hampshire	X	X	X
North Dakota	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

5.2.6 Overview of the Operational Bank

Figure 2, Figure 3, and Figure 4 display the histogram of the difficulty parameters for grades 4, 8, and 10, respectively, for all items that are part of the NDSA for Science operational pool. The figures also display the student proficiency distributions. For all grades, items are slightly more difficult than the student proficiency in general.

Figure 2. North Dakota State Assessment Item Difficulty and Student Proficiency Distributions, Grade 4

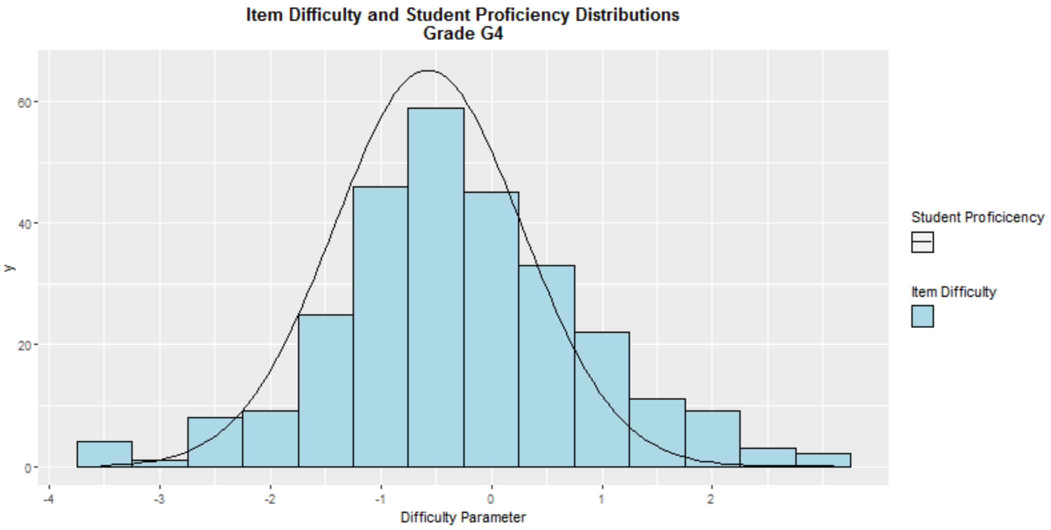


Figure 3. North Dakota State Assessment Item Difficulty and Student Proficiency Distributions, Grade 8

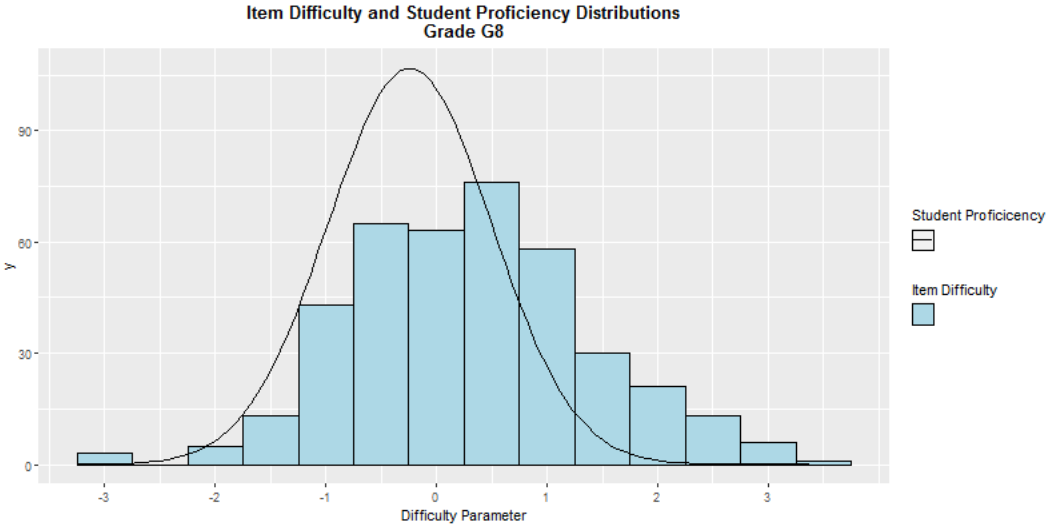
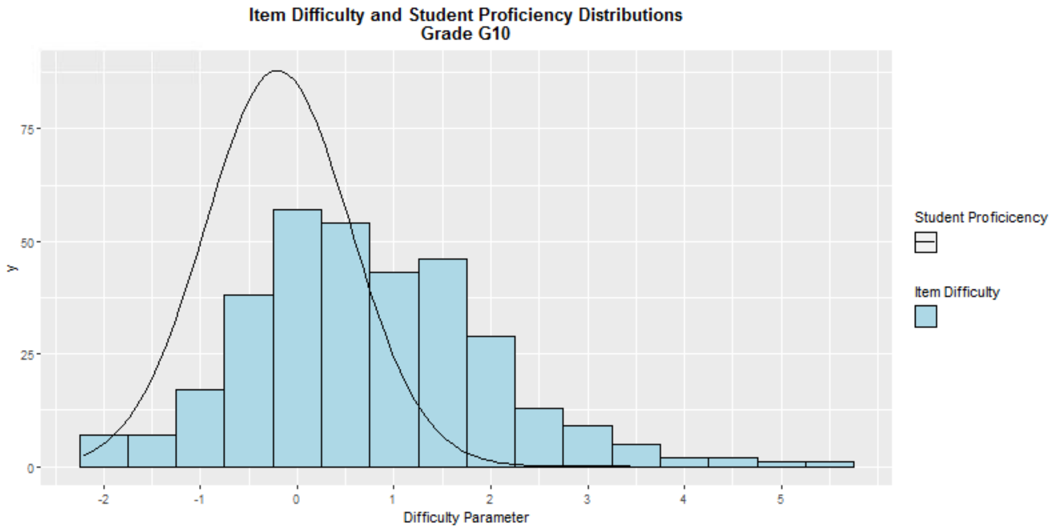


Figure 4. North Dakota State Assessment Item Difficulty and Student Proficiency Distributions, Grade 10



6. Scoring

6.1 Marginal Maximum Likelihood Function

Student scores are obtained by marginalizing out the nuisance dimensions $\boldsymbol{u}_j$ from the likelihood of the observed response pattern $\boldsymbol{z}_j$ for student j,

$$\ell_i(\theta_j) = log \int_{\boldsymbol{u}_j} P(\boldsymbol{z}_j|\theta_j, \boldsymbol{u}_j)\, N(\boldsymbol{u}_j|\boldsymbol{0}, \boldsymbol{\Sigma})d\boldsymbol{u}_j,$$

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.

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 $\mathbf{v} = \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{u}$. Then we have

$$\int_{\boldsymbol{u}_j} P(\boldsymbol{z}_j|\theta_j, \boldsymbol{u}_j)\, N(\boldsymbol{u}_j|\boldsymbol{0}, \boldsymbol{\Sigma})d\boldsymbol{u}_j = \int_{\boldsymbol{v}_j} P\left(\boldsymbol{z}_j\left|\theta_j, \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{v}_j\right.\right) N(\boldsymbol{v}_j|\boldsymbol{0}, \boldsymbol{I})d\boldsymbol{v}_j.$$

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

$$\int_{\boldsymbol{u}_j} P(\boldsymbol{z}_j|\theta_j, \boldsymbol{u}_j)\, N(\boldsymbol{u}_j|\boldsymbol{0}, \boldsymbol{\Sigma})d\boldsymbol{u}_j = P(\boldsymbol{z}_j|\theta_j),$$

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

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

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

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

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

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

Based on the above equations, we need to define only 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

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

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 50 contains the LOT and HOT values for each grade.

6.4 Standard Error of Estimate

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 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 NDSA for Science are assembled on-the-fly using an adaptive testing design. Tests are considered complete if students respond to all the operational items. Otherwise, the tests are "incomplete". Tests that are incomplete but attempted are scored. In order to receive a discipline score (i.e., Life Sciences, Physical Sciences, and 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 an item cluster, the simulated item parameters of the missing item are the item parameters of item cluster 3119 for grade 4, 2476 for grade 8, and 2985 for grade 10, which are operational Independent College and Career Readiness (ICCR) item clusters that are typical for the item pool used in North Dakota in terms of the number of assertions and estimated parameters. Likewise, if a missing item is a stand-alone item, the simulated item parameters of the missing item are the item parameters of stand-alone item 2938 for grade 4, 2905 for grade 8, and 3002 for grade 10, which are operational ICCR stand-alone items that are typical for the item pool used in North Dakota 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 will be used. Similarly, for the accommodated forms that are fixed forms, the item parameters of the unanswered items on the form will be used.

6.6 STUDENT-LEVEL SCALE SCORE

At the student level in grades 4, 8, and 10, scale scores are computed for

1. Overall Science;
2. Life Sciences;
3. Physical Sciences; and
4. Earth and Space Sciences (only grades 4 and 8[5]).

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 50, which correspond to values of the estimated mean minus/plus four times of 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 50). The standard error of estimate for the estimated scale score is obtained as

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

In 2021, the slope a and intercept b were chosen so that the center of the reporting scale of each grade (400, 800, and 1000, respectively) at the grade mean of the 2021 base-year and has a standard deviation of 25. Furthermore, for each grade, the reporting scale ranges from the base-year mean

[5] The NDSA for Science in grade 10 does not include the Earth and Space Sciences discipline; therefore, scale scores are computed for Overall Science, Life Sciences, and Physical Sciences in grade 10.

minus 4 times the standard deviation to the base-year mean plus 4 times the standard deviation. Specifically, for grade 4, the slope and intercept were obtained as

$$\begin{aligned} SS &= 25\theta^* + 400 \\ &= 25\frac{\theta - \hat{\mu}_\theta}{\hat{\sigma}_\theta} + 400 \\ &= \tfrac{25}{\hat{\sigma}_\theta}\theta + \left(400 - \tfrac{25\hat{\mu}_\theta}{\hat{\sigma}_\theta}\right), \end{aligned}$$

where the second line stems from standardizing theta, $\theta^* = \frac{\theta - \hat{\mu}_\theta}{\hat{\sigma}_\theta}$. For grades 8 and 10, the slope and intercept can also be derived in a similar fashion.

Per grade, Table 50 presents the intercept, slope, LOT, HOT, lowest obtainable scale score (LOSS), and highest obtainable scale score (HOSS) values that were used for the 2021 reporting scale. The scale score distribution is reported for overall science in Appendix A, Distribution of Scale Scores and Achievement Levels for NDSA Science. The scale score distribution is reported for the science disciplines in Appendix B, Distribution of Scale Scores by Science Discipline for NDSA Science.

Table 50. Science Reporting Scale Linear Transformation Constants, Theta, and Corresponding Scaled-Score Limits for Extreme Ability Estimates (for 2021 θ Scale)

Grade	Slope (*a*)	Intercept (*b*)	Lowest Obtainable Theta (LOT)	Highest Obtainable Theta (HOT)	Lowest Obtainable Scale Score (LOSS)	Highest Obtainable Scale Score (HOSS)
4	30.55	415.705	-3.78	2.75	300	500
8	38.482	807.416	-2.79	2.40	700	900
10	36.167	1005.242	-2.90	2.62	900	1100

6.7 Rules for Calculating Achievement Levels

Achievement levels and corresponding cut scores were set during standard setting in summer 2021. Students are classified into one of four achievement levels, based on their total score. The distribution of achievement levels is summarized in Appendix A, Distribution of Scale Scores and Achievement Levels for NDSA Science. Further, the distribution of scale scores and achievement levels for subgroups described in Section 4.4, Differential Item Functioning, are presented in Appendix C, Distribution of Scale Scores and Achievement Levels by Subgroup for NDSA Science.

Table 51 lists the cut scores on the reporting scale metrics for each grade.

Table 51. Achievement-Level Cut Scores

Grade	Cut 1 Partially Proficient	Cut 2 Proficient	Cut 3 Advanced
4	380	407	431
8	775	802	835
10	973	1000	1035

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/areas of science. The following are the classification rules:

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

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 Standard*, and the HOT is always classified as *Above Standard*.

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

6.8.1 Relative to Overall Achievement

For aggregated units (i.e., classrooms, schools, and districts), there is residual-based reporting at more fine-grained levels. Before 2022, reports were provided at the level of Disciplinary Core Idea (DCI). Starting in 2022, there is also reporting for aggregated units for four claims corresponding to Science and Engineering Practice (SEP).

The method for reporting on these additional categories for aggregated units is based on the use of residuals. The equations are presented for DCIs but can be computed in a similar way for SEPs. For future reporting categories, the equations will be obtained in an analogous way.

For each assertion i, the residual between the 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,

$$E\left(z_{ijg} = 1; \theta_{j,overall}, \boldsymbol{\tau}_i\right) = \int P\left(z_{ijg} = 1 | u_{jg};\ \theta_{j,overall}, \boldsymbol{\tau}_i\right) N\left(u_{jg}\right) du_{jg},$$

where $\boldsymbol{\tau}_i$ is the vector of parameters for assertion i (e.g., for the Rasch testlet model, $\boldsymbol{\tau}_i = b_i$), and $P\left(z_{ijg} = 1 | u_{jg};\ \theta_{j,overall}, \boldsymbol{\tau}_i\right)$ 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_m 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_m count for the aggregate. The standard error of the average residual is computed as

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

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

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

- If $\bar{\delta}_{DCIm} \leq -1.5 * SEM\left(\bar{\delta}_{DCIm}\right)$, 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* on the overall test.
- If $SEM\left(\bar{\delta}_{DCIm}\right) > 0.2$, 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, Relative to Overall Achievement, but with the expected score computed at the theta value corresponding to the proficiency cut score:

$$E\left(z_{ijg}=1;\theta_{proficiency},\boldsymbol{\tau}_i\right)=\int P\left(z_{ijg}=1|u_{jg};\ \theta_{proficiency},\boldsymbol{\tau}_i\right)N\left(u_{jg}\right)du_{jg}.$$

The following is reported for DCIs for aggregate units:

- If $\bar{\delta}_{DCIm}\leq-1.5*SEM\left(\bar{\delta}_{DCIm}\right)$, then achievement is *below* the proficiency cut score.
- If $\bar{\delta}_{DCIm}\geq1.5*\ SEM\left(\bar{\delta}_{DCIm}\right)$, then achievement is *above* the proficiency cut score.
- Otherwise, achievement is *near* the proficiency cut score.
- If $SEM\left(\bar{\delta}_{DCIm}\right)>0.2$, 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 the incorrect designation of a keyed response or other scoring errors and 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, biserial/polyserial correlation, and 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 falling outside a specified range or to generate reports based on all items in the pool. For science, statistics reports at the assertion level (which are the units of analysis for science) are currently not yet available; however, as a routine and continuing practice, CAI psychometricians compute and monitor classical item statistics and item fit (i.e., item drift) at the end of each testing window.

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 group 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.

For all three grades, every test 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. Some violations did occur at lower content levels. Blueprint match is discussed in detail 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 flagged 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. A total of 43.75% of the items in grade 4, 33.33% of items in grade 8, and 42.47% of items in grade 10 were administered to 20% or more test takers at that grade in the online English version of the test. More details are discussed in Volume 2, Test Development.

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. Before releasing any scores, a second score verification system is used to verify that all test scores match with 100% agreement in all tested grades. The second system is independently constructed and maintained from the main scoring engine and separately estimates scores 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.

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., 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.