

Connecticut Next Generation Science Standards Assessment

2023–2024

Volume 1: Annual Technical Report



CONNECTICUT STATE
DEPARTMENT OF EDUCATION

TABLE OF CONTENTS

1.	INTRODUCTION	1
1.1	Background and Historical Context of Tests	1
1.2	Purpose and Intended Uses of the Connecticut NGSS Assessment	2
1.3	Participants in the Development and Analysis of the Connecticut NGSS Assessment	3
1.3.1	Connecticut State Department of Education	3
1.3.2	Connecticut Educators	3
1.3.3	Technical Advisory Committee	3
1.3.4	Cambium Assessment, Inc.	4
1.3.5	Caveon Test Security	4
1.4	Available Test Formats and Special Versions	4
1.5	Student Participation	4
2.	OPERATIONAL PRACTICES AND PROCEDURES	5
2.1	Testing Window	6
2.2	Test Administrators	6
2.3	Testing Environment	6
2.4	Simulations	6
2.5	Universal Tools, Designated Supports, and Accommodations	7
3.	ITEM BANK AND TEST DESIGN	8
3.1	Shared Science Assessment Item Bank	8
3.2	Field-Testing	10
3.2.1	2024 Field Tests	10
3.3	Test Design	21
4.	FIELD-TEST CLASSICAL ANALYSIS	22
4.1	Item Discrimination	23
4.2	Item Difficulty	23
4.3	Response Time	23
4.4	Differential Item Functioning	23
4.5	Classical Analysis Results	26
5.	ITEM CALIBRATION	29
5.1	Model Description	29
5.1.1	Latent Structure	29
5.1.2	Item Response Function	31
5.1.3	Multigroup Model	31
5.2	Estimation	32
5.3	Overview of the Operational Item Bank	33
6.	SCORING	35

6.1	Marginal Maximum Likelihood Function	35
6.2	Derivative	36
6.3	Extreme Case Handling.....	38
6.4	Standard Error of Measurement	38
6.5	Scoring Incomplete Tests	38
6.6	Student-Level Scale Score.....	39
6.7	Rules for Calculating Performance Levels.....	40
6.7.1	<i>Strengths and Weaknesses for Disciplines Relative to Proficiency Cut Score..</i>	<i>41</i>
6.8	Residual-Based Reporting at the Level of Disciplinary Core Ideas and Science and Engineering Practices	41
6.8.1	<i>Relative to Overall Performance.....</i>	<i>41</i>
6.8.2	<i>Relative to Proficiency Cut Score</i>	<i>42</i>
7.	QUALITY CONTROL PROCEDURES	43
7.1	Quality Assurance Reports.....	43
7.1.1	<i>Item Analysis</i>	<i>43</i>
7.1.2	<i>Blueprint Match.....</i>	<i>44</i>
7.1.3	<i>Item Exposure Rates</i>	<i>44</i>
7.1.4	<i>Cheating Detection Analysis</i>	<i>44</i>
7.2	Scoring Quality Check	45
8.	REFERENCES.....	46

LIST OF TABLES

Table 1. Required Uses and Citations for the Connecticut NGSS Assessment.....	3
Table 2. Number of Students Participating in the Connecticut NGSS Assessment, Spring 2024.	5
Table 3. Distribution of Demographic Characteristics of Student Population, Spring 2024.....	5
Table 4. Connecticut NGSS Assessment Testing Windows, Spring 2024.....	6
Table 5. Number of Test Sessions with Allowed Designated Supports, Spring 2024.....	7
Table 6. Number of Test Sessions with Allowed Accommodations, Spring 2024.....	8
Table 7. Number of Field-Test Items Administered, Spring 2024	11
Table 8. Number of Common Elementary School Field-Test Items Administered and Calibrated, Spring 2024.....	13
Table 9. Number of Common Middle School Field-Test Items Administered and Calibrated, Spring 2024.....	15
Table 10. Number of Common High School Field-Test Items Administered and Calibrated, Spring 2024.....	17
Table 11. Field-Test Item Administration, Rubric Validation, and Item Data Review, Spring 2024.....	20
Table 12. Shared Science Assessment Item Bank, Spring 2024.....	20
Table 13. Thresholds for Flagging in Classical Item Analysis	22
Table 14. DIF Classification Rules.....	26
Table 15. Distribution of p-Values for Field-Test Items, Spring 2024.....	26
Table 16. Distribution of Item Biserial Correlations for Field-Test Items, Spring 2024.....	27
Table 17. Summary of Response Times for Field-Test Items, Spring 2024.....	27
Table 18. Differential Item Functioning Classifications for Field-Test Items, Spring 2024.....	28
Table 19. Groups Per Grade Band for the Spring 2024 Calibration of Field-Test Items	33
Table 20. Science Reporting Scale Linear Transformation Constants, Theta, and Corresponding Scaled-Score Limits for Extreme Ability Estimates (for 2019 θ Scale).....	40
Table 21. Performance-Level Cut Scores	40

LIST OF FIGURES

Figure 1. Directed Graph of the Science IRT Model.....	31
Figure 2. Connecticut NGSS Assessment Assertion Difficulty and Student Proficiency Distributions, Grade 5, Spring 2024	34
Figure 3. Connecticut NGSS Assessment Assertion Difficulty and Student Proficiency Distributions, Grade 8, Spring 2024	34
Figure 4. Connecticut NGSS Assessment Assertion Difficulty and Student Proficiency Distributions, Grade 11, Spring 2024	35

LIST OF APPENDICES

Appendix 1-A. Caveon Test Security Overview	
Appendix 1-B. Shared Science Assessment Item Bank: Field-Testing	
Appendix 1-C. Calibration of the Shared Science Assessment Item Bank	
Appendix 1-D. Distribution of Scale Scores and Performance Levels	
Appendix 1-E. Distribution of Scale Scores by Science Discipline	
Appendix 1-F. Distribution of Scale Scores and Performance Levels by Subgroup	

1. INTRODUCTION

The Connecticut Next Generation Science Standards (NGSS) Assessment is a science assessment for grades 5, 8, and 11. The *Connecticut NGSS Assessment 2023–2024 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 the 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 Performance Standards.** This volume documents the methods and results of the Connecticut NGSS Assessment standard-setting process.
- 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 Connecticut State Department of Education (CSDE) communicates the quality of the Connecticut NGSS Assessment by making these technical reports accessible to the public on the state’s website.

1.1 BACKGROUND AND HISTORICAL CONTEXT OF TESTS

In 2015, Connecticut adopted three-dimensional science standards (the Next Generation Science Standards) based on *A Framework for K–12 Science Education* (National Research Council, 2012). The CSDE and its assessment vendor, Cambium Assessment, Inc. (CAI), developed and administered a new online assessment to measure these new standards. Piloted in 2016–2017, field-tested in 2017–2018, and administered operationally for the first time in 2018–2019, the Connecticut NGSS Assessment measures the science knowledge and skills of Connecticut students in grades 5, 8, and 11. The CSDE cancelled the spring 2020 administration of the Connecticut NGSS Assessment due to statewide school closures that followed the onset of the COVID-19 pandemic. Starting in spring 2021, the CSDE and CAI resumed administration of the Connecticut NGSS Assessment.

The CSDE provides an overview of the science assessment at: <https://portal.ct.gov/SDE/Student-Assessment/NGSS-Science/NGSS-Science>. Information about the NGSS is available online at: www.nextgenscience.org.

1.2 PURPOSE AND INTENDED USES OF THE CONNECTICUT NGSS ASSESSMENT

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

The three-dimensional NGSS reflect the latest research and advances in modern science education and differ from previous science standards in multiple ways. First, rather than describing general knowledge and skills that students should know and be able to do, they describe specific performances that demonstrate what students know and can do. The NGSS refer to such performed knowledge and skills as *performance expectations* (PEs). Second, the NGSS are intentionally multi-dimensional. Each performance expectation incorporates all three dimensions from the NGSS Framework—a science or engineering practice, a disciplinary core idea, and a crosscutting concept. Third, while traditional standards do not consider other subject areas, the NGSS connect to other subjects like the Common Core mathematics and English language arts (ELA) standards. Another unique feature of the NGSS 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 Connecticut NGSS Assessment supports 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 aggregate levels and to monitor improvement at the student and group levels over time.

The Connecticut NGSS Assessment draws items from an item bank that consists of Independent College and Career Readiness (ICCR) items and items owned by several other states and one U.S. territory that abide by a Memorandum of Understanding (MOU) to share content, leadership, and new ideas and methods. Full members of the MOU in 2024 were Arkansas, Connecticut, Hawaii, Idaho, Indiana, Montana, New Hampshire, Oregon, Rhode Island, Utah, West Virginia, and Wyoming. CAI had a supporting and coordinating role. 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 CSDE to ensure that the items in the tests constructed for all grades uniquely measured students' mastery of the three-dimensional NGSS.

Table 1 outlines the required uses and citations for the Connecticut NGSS Assessment based on the Connecticut General Statutes §10-14n and the federal *Every Student Succeeds Act* (ESSA) plan. The Connecticut NGSS Assessment fulfills all the requirements described in Table 1.

Table 1. Required Uses and Citations for the Connecticut NGSS Assessment

Required Use	Required Use Citation
Indicator of academic achievement and progress	ESSA § 1111(b)(2)(B)(ii)
Test administration frequency and grade levels	ESSA § 1111(b)(2)(B)(v)(II) Connecticut Statutes §10-14n. (a)(2) Connecticut Statutes §10-14n. (b)(3)
Disaggregation of test scores	ESSA § 1111(b)(3)(C)(xiii)
Publication of test scores	ESSA § 1111(b)(3)(C)(xii) Connecticut Statutes §10-14n. (g)

1.3 PARTICIPANTS IN THE DEVELOPMENT AND ANALYSIS OF THE CONNECTICUT NGSS ASSESSMENT

The CSDE manages the Connecticut state assessment programs with the assistance of several participants, including Connecticut educators, a Technical Advisory Committee (TAC), and vendors. The CSDE fulfills the diverse requirements of implementing Connecticut’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). To comply with the *Standards*, scale development, scoring, linking, and evaluation of differential item functioning are addressed in the current volume; item development, test design, and test blueprints are documented in Volume 2, Test Development; development of cut scores is summarized in Volume 3, Setting Performance Standards; evidence for validity and reliability/precision was collected and is reported in Volume 4, Evidence of Reliability and Validity; information on testing windows, test options, accommodations, training of test coordinators and administrators, and test security are provided in Volume 5, Test Administration; supporting documentation for tests, score uses and interpretation are included in Volume 6, Score Interpretation.

1.3.1 Connecticut State Department of Education

The Student Assessment, Performance Office manages test development, administration, scoring, and results reporting for the statewide comprehensive assessment programs, including coordinating with other CSDE offices, Connecticut public schools, and vendors.

1.3.2 Connecticut Educators

Connecticut educators participate in most aspects of the conceptualization and development of the Connecticut NGSS Assessment. Educators participate in clarifying how the standards are assessed, designing tests, and reviewing test questions and passages.

1.3.3 Technical Advisory Committee

The CSDE convenes an advisory committee panel twice each year to discuss psychometrics, test development, and administrative and policy issues relevant to the current and future Connecticut

assessments. This committee is comprised of several nationally recognized assessment experts and highly experienced practitioners from several Connecticut school districts.

1.3.4 Cambium Assessment, Inc.

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

1.3.5 Caveon Test Security

Caveon Test Security monitored web pages and social media during the spring 2024 test administration to ensure that secure testing materials such as items and prompts were not leaked. Details of Caveon Test Security are described in Appendix 1-A, Caveon Test Security Overview.

1.4 AVAILABLE TEST FORMATS AND SPECIAL VERSIONS

The Connecticut NGSS Assessment is administered online using an adaptive 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 them in various ways. The science test was an independent field test in spring 2018, and it became operational in spring 2019. Starting in 2021 and thereafter, additional items were field-tested to build out the item bank.

Students unable to participate in the online test administration have the option to use print-on-demand—a feature that provides the same items administered to students online in a paper format. Spanish versions of the Connecticut NGSS Assessment (developed to meet the same content standards as the English versions) are available for all tested grades. Students participating in the computer-based Connecticut NGSS Assessment can use standard online testing features in the Test Delivery System (TDS), which include a selection of font colors and sizes 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 Connecticut NGSS Assessment with or without accommodations or to take an alternate assessment. For additional information about testing features and accommodations, refer to Volume 5, Test Administration, of this technical report.

1.5 STUDENT PARTICIPATION

All students in Connecticut public schools are required to participate in statewide assessments. The Connecticut NGSS Assessment 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 analyses in this technical report (number reported).

Table 3 shows the demographic characteristics of the student population, by counts and percentages, in the spring administration of the 2023–2024 Connecticut NGSS Assessment. The subgroups reported are gender, ethnicity, students with limited English proficiency (LEP), special education students, and economically disadvantaged students.

Table 2. Number of Students Participating in the Connecticut NGSS Assessment, Spring 2024

Grade	Number Tested	Number Reported
5	36,481	36,451
8	37,119	37,097
11	37,355	37,288

Table 3. Distribution of Demographic Characteristics of Student Population, Spring 2024

Group	Grade 5		Grade 8		Grade 11	
	N	%	N	%	N	%
All Students	36,451	100.00	37,097	100.00	37,288	100.00
Female	17,865	49.01	18,164	48.96	18,223	48.87
Male	18,580	50.97	18,904	50.96	18,959	50.84
African American	4,316	11.84	4,756	12.82	4,601	12.34
American Indian/Native Alaskan	92	0.25	103	0.28	80	0.21
Asian	2,010	5.51	1,906	5.14	1,997	5.36
Hispanic	11,681	32.05	11,496	30.99	10,869	29.15
Multi-Racial	1,766	4.84	1,599	4.31	1,450	3.89
Pacific Islander	30	0.08	28	0.08	42	0.11
White	16,556	45.42	17,209	46.39	18,249	48.94
Limited English Proficiency	4,819	13.22	3,329	8.97	2,718	7.29
Special Education	6,180	16.95	6,183	16.67	5,108	13.70
Economically Disadvantaged	16,303	44.73	16,253	43.81	14,522	38.95

2. OPERATIONAL PRACTICES AND PROCEDURES

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

2.1 TESTING WINDOW

Table 4 shows the testing windows for the 2023–2024 Connecticut NGSS Assessment.

Table 4. Connecticut NGSS Assessment Testing Windows, Spring 2024

Tests	Grade	Start Date	End Date
NGSS Summative Assessments	11	2/5/2024	5/31/2024
	5, 8	3/25/2024	5/31/2024
NGSS Interim Assessments	5, 8, 11	8/29/2023	6/7/2024

2.2 TEST ADMINISTRATORS

The key personnel involved with test administration for the Connecticut State Department of Education (CSDE) included district test coordinators (DTCs), school test coordinators (STCs), and test administrators (TAs) who proctored the test. *Test Administration Manuals* (TAMs) (available at <https://ct.portal.cambiumast.com/resources>) were provided so that personnel involved with the statewide assessment administrations could maintain both standardized administration conditions and test security.

2.3 TESTING ENVIRONMENT

The Cambium Assessment, Inc. (CAI) Secure Browser was required to access the online Connecticut NGSS Assessment. 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 responses 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, of this technical report.

2.4 SIMULATIONS

CAI employs a simulation approach to all Connecticut NGSS tests before the operational testing window begins. For adaptive tests, simulations were conducted to configure the item selection algorithm settings, to evaluate whether individual tests adhered to the test blueprint and correlated highly with student ability, to monitor item exposure rates, and to verify the scores produced by CAI's scoring engine. Simulations were also conducted on fixed-form tests to quality check the scores. The simulation approaches and results are discussed in Volume 2, Test Development.

2.5 UNIVERSAL TOOLS, DESIGNATED SUPPORTS, AND ACCOMMODATIONS

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

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

Accommodations are changes in procedures or materials that ensure equitable access to instructional and assessment content and generate valid assessment results for students who need them. Embedded accommodations (e.g., text-to-speech [TTS]) are provided digitally through instructional or assessment technology, and non-embedded designated features (e.g., scribe) are non-digital. State-approved accommodations do not compromise the learning expectations, constructs, or grade-level standards. These accommodations help students with a documented need generate valid assessment outcomes that fully demonstrate what they know and are able to 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).

Connecticut TAs and SCs were responsible for ensuring that arrangements for accommodations were made before the test administration dates. The available accommodation options for eligible students included the following: braille, American Sign Language (ASL), closed captioning, streamline, abacus, assistive technology (e.g., adaptive keyboards, touch screens, switches), calculation device, print-on-demand, multiplication table, and scribe. Additional information about universal features, designated supports, and accommodations can be found in Volume 5, Test Administration, of this technical report.

Table 5 and Table 6 list the number of test sessions in which a student was provided with each designated support or accommodation during the spring 2024 test administration.

Table 5. Number of Test Sessions with Allowed Designated Supports, Spring 2024

Designated Supports	Grade		
	5	8	11
Embedded			
Color Contrast	9	14	12
Language-Spanish (Toggle)	1,005	947	845
Masking	243	158	70
Mouse Pointer	12	3	8

Designated Supports	Grade		
	5	8	11
Permissive Mode	35	19	2
Print Size	42	35	11
Streamlined Mode	282	302	116
Text-to-Speech: Stimuli and Items	9,344	6,611	1,543
Non-Embedded			
Bilingual Dictionary	89	332	518
Color Contrast	5	3	-
Color Overlay	2	2	2
Magnification	14	12	7
Medical Device	3	8	7
Native Language Reader Test Directions	34	59	49
Noise Buffer	15	17	8
Read Aloud: Stimuli and Items	180	116	89
Read Aloud: Stimuli and Items (Spanish)	24	23	3
Separate Setting	5,183	4,315	1,588

Table 6. Number of Test Sessions with Allowed Accommodations, Spring 2024

Accommodations	Grade		
	5	8	11
Non-Embedded			
Abacus	5	6	-
Alternate Response Options (Requires Permissive Mode)	6	4	13
Large Print	5	6	3
Sign Language for Test Items	3	4	3
Specialized Calculator	29	113	148
UEB Braille Booklet-Contracted + Nemeth Math	2	3	1

3. ITEM BANK AND TEST DESIGN

3.1 SHARED SCIENCE ASSESSMENT ITEM BANK

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

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

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

In 2024, the Shared Science Assessment Item Bank was used for operational tests in 14 states and one U.S. territory, including Connecticut.

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

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

A second set of cognitive lab studies was conducted in 2018 and 2019 to determine whether students using braille could understand the task demands of selected accommodated three-dimensional science-aligned item clusters. They also evaluated whether these students could navigate the interactive features of these item clusters in a manner that allowed them to fully display their knowledge and skills relative to the constructs of interest. In general, both the students who relied entirely on braille and/or Job Access With Speech (JAWS) and those who had some vision and were able to read the screen with magnification were able to find the information they needed to respond to the questions, navigate the various response formats, and finish within a reasonable amount of time (refer to Volume 4, Appendix 4-E, Braille Cognitive Lab Report, of this technical report).

In 2018, CAI field tested more than 540 item clusters and stand-alone items, of which 451 (including items from all sources) were accepted and made available as operational items in 2019 and future administrations. In 2019, 2021, 2022, and 2023, the numbers of items that were field tested were 347, 545, 471, and 348, while the numbers of items that were accepted and made available for future operational use were 268, 458, 403, 288, respectively. In 2024, 478 item clusters and stand-alone items were field tested, of which 386 were accepted and made available for operational use in future administrations. All these items follow the same specifications, test development processes, and review processes, summarized below:

- CAI staff and participating states collaborated to develop item specifications, which are documents designed to guide item writers as they craft test questions and stakeholders while they review items. The item specifications were generally accompanied by sample

items meeting those specifications. All specifications and sample items were reviewed by state content experts and committees of educators in at least one state.

- The specifications helped test developers create item clusters and stand-alone items that covered a range of difficulty, furthering the goal of measuring the full range of performance found in the population, but remaining at grade level. All item writers were trained in the principles of universal design, the appropriate use of item interactions, and the science item specifications.
- Items were reviewed by science experts in at least one state.
- Every item was reviewed by a content advisory committee (comprised of state educators) in at least one state or in a cross-state educator review process.
- Every item was reviewed by a committee of educators charged with evaluating language accessibility, bias, and sensitivity in at least one state or a cross-state educator review.
- Every item was field tested, all scoring protocols (i.e., rubrics) were validated using the field-test data, and items with questionable data were reviewed again by committees of educators.

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

3.2 FIELD-TESTING

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

3.2.1 2024 Field Tests

In 2024, field-test items were administered as unscored items embedded among operational items in 14 states and one U.S. territory (Arkansas, Connecticut, Hawaii, Idaho, Indiana, Montana, New Hampshire, North Dakota, Oregon, Rhode Island, South Dakota, U.S. Virgin Islands, Utah, West Virginia, and Wyoming). In total, 226 item clusters and 252 stand-alone items were administered as field-test items in the elementary, middle, and high school grade bands. Table 7 presents the number of field-test item clusters and stand-alone items administered in each grade band for each state. The numbers in parentheses in the column representing Connecticut show the number of field-test items owned by Connecticut.

Table 7. Number of Field-Test Items Administered, Spring 2024

Grade Band and Item Type	AR	CT	HI	ID	IN	MT	NH	ND	OR	RI	SD	USVI	UT	WV	WY	Total*
Elementary School	94	28 (21)	24	18	43	14	14	4	15	20	2	1	39	32	12	166
Cluster	21	8 (5)	8	7	20	6	6	4	12	6	1	1	39	12	4	69
Stand-Alone	73	20 (16)	16	11	23	8	8	0	3	14	1	0	0	20	8	97
Middle School	94	28 (21)	24	9	45	14	11	1	18	20	4	1	64	27	12	176
Cluster	33	13 (9)	9	4	22	6	3	1	15	5	1	1	64	11	4	90
Stand-Alone	61	15 (12)	15	5	23	8	8	0	3	15	3	0	0	16	8	86
High School	29	39 (29)	10	21	73	0	9	5	39	37	11	1	0	0	9	136
Cluster	17	15 (11)	6	13	37	0	5	5	20	15	3	1	0	0	5	67
Stand-Alone	12	24 (18)	4	8	36	0	4	0	19	22	8	0	0	0	4	69
Total	217	95 (71)	58	48	161	28	34	10	72	77	17	3	103	59	33	478

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

*The total count of field test items excludes 11 South Dakota legacy items and 32 Computer Science items in Indiana but includes several items that were moved to the comprehensive interim pool after rubric validation.

Two of the states (New Hampshire and Rhode Island) opted for a test in which operational items were grouped by science discipline. For these two states, the field-test items were presented together in a fourth group of items. The sequence of the four sets of items (corresponding to the three disciplines and a set of field-test items) was randomized across students. Twelve other states and one U.S. territory (Arkansas, Connecticut, Hawaii, Idaho, Indiana, Montana, North Dakota, Oregon, South Dakota, U.S. Virgin Islands, Utah, West Virginia, and Wyoming) opted for a test design in which the items were not grouped by discipline. In these 12 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 Connecticut NGSS Assessment is discussed in Section 3.3, Test Design.

A minimum sample size of 1,500 students per field-test item was targeted for any given state or territory. Most items were administered in two states or territory. All items met or exceeded the target sample size of 1,500 in at least one state.

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

Table 8. Number of Common Elementary School Field-Test Items Administered and Calibrated, Spring 2024

	State	AR	CT	HI	ID	IN	MT	NH	ND	OR	RI	SD	USVI	UT	WV	WY
Item Clusters	AR	0 (0)	2	3	2	1	2	4	0	2	0	0	0	1	3	0
	CT	2	0 (0)	0	0	2	0	1	0	0	0	0	0	6	0	0
	HI	3	0	0 (0)	2	1	0	0	0	0	0	0	0	2	0	0
	ID	2	0	2	0 (0)	3	0	0	0	0	0	0	0	2	0	0
	IN	1	2	1	3	0 (0)	3	0	0	0	2	0	0	14	4	2
	MT	3	0	0	0	3	0 (0)	0	0	0	0	0	0	2	0	1
	NH	4	1	0	0	0	0	0 (0)	1	0	0	0	1	0	0	0
	ND	0	0	0	0	0	0	2	0 (0)	0	2	0	1	2	0	0
	OR	2	0	0	0	0	0	0	0	0 (0)	0	0	0	10	0	0
	RI	0	0	0	0	2	0	0	2	0	0 (0)	0	0	3	1	0
	SD	0	0	0	0	0	0	1	1	0	0	0 (0)	0	0	0	0
	USVI	0	0	0	0	0	0	1	1	0	0	0	0 (0)	0	0	0
	UT	1	6	2	2	14	2	0	2	10	3	0	0	0 (0)	7	4
	WV	4	0	0	0	4	0	0	0	0	1	0	0	7	0 (0)	1
	WY	0	0	0	0	2	1	0	0	0	0	0	0	4	1	0 (0)
Stand-Alone Items	AR	0 (0)	15	8	7	13	4	7	0	3	6	0	0	0	13	8
	CT	15	0 (0)	0	0	1	4	0	0	0	0	0	0	0	0	0
	HI	8	0	0 (0)	4	4	0	0	0	0	0	0	0	0	0	0
	ID	7	0	4	0 (0)	4	0	0	0	0	0	0	0	0	0	0
	IN	13	1	4	4	0 (0)	0	4	0	0	6	0	0	0	2	0
	MT	4	4	0	0	0	0 (0)	0	0	0	0	0	0	0	0	0
	NH	7	0	0	0	4	0	0 (0)	0	0	0	1	0	0	0	0
	ND	0	0	0	0	0	0	0	0 (0)	0	0	0	0	0	0	0
	OR	3	0	0	0	0	0	0	0	0 (0)	0	0	0	0	0	0
	RI	6	0	0	0	6	0	0	0	0	0 (0)	0	0	0	5	0

	State	AR	CT	HI	ID	IN	MT	NH	ND	OR	RI	SD	USVI	UT	WV	WY
	SD	0	0	0	0	0	0	1	0	0	0	0 (0)	0	0	0	0
	USVI	0	0	0	0	0	0	0	0	0	0	0	0 (0)	0	0	0
	UT	0	0	0	0	0	0	0	0	0	0	0	0	0 (0)	0	0
	WV	13	0	0	0	2	0	0	0	0	5	0	0	0	0 (0)	0
	WY	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0 (0)
Total	AR	0 (0)	17	11	9	14	6	11	0	5	6	0	0	1	16	8
	CT	17	0 (0)	0	0	3	4	1	0	0	0	0	0	6	0	0
	HI	11	0	0 (0)	6	5	0	0	0	0	0	0	0	2	0	0
	ID	9	0	6	0 (0)	7	0	0	0	0	0	0	0	2	0	0
	IN	14	3	5	7	0 (0)	3	4	0	0	8	0	0	14	6	2
	MT	7	4	0	0	3	0 (0)	0	0	0	0	0	0	2	0	1
	NH	11	1	0	0	4	0	0 (0)	1	0	0	1	1	0	0	0
	ND	0	0	0	0	0	0	2	0 (0)	0	2	0	1	2	0	0
	OR	5	0	0	0	0	0	0	0	0 (0)	0	0	0	10	0	0
	RI	6	0	0	0	8	0	0	2	0	0 (0)	0	0	3	6	0
	SD	0	0	0	0	0	0	2	1	0	0	0 (0)	0	0	0	0
	USVI	0	0	0	0	0	0	1	1	0	0	0	0 (0)	0	0	0
	UT	1	6	2	2	14	2	0	2	10	3	0	0	0 (0)	7	4
	WV	17	0	0	0	6	0	0	0	0	6	0	0	7	0 (0)	1
	WY	8	0	0	0	2	1	0	0	0	0	0	0	4	1	0 (0)

Table 9. Number of Common Middle School Field-Test Items Administered and Calibrated, Spring 2024

	State	AR	CT	HI	ID	IN	MT	NH	ND	OR	RI	SD	USVI	UT	WV	WY
Item Clusters	AR	0 (0)	1	1	2	1	0	1	0	8	2	0	0	18	1	1
	CT	1	0 (0)	0	0	3	0	0	0	1	0	0	0	10	0	0
	HI	1	0	0 (0)	0	0	0	0	0	0	0	0	0	7	0	0
	ID	2	0	0	0 (0)	0	0	0	0	0	0	0	0	4	0	0
	IN	1	3	0	0	0 (0)	2	1	0	5	3	0	0	13	2	3
	MT	0	0	0	0	2	0 (0)	1	0	2	0	0	0	3	0	1
	NH	1	0	0	0	1	1	0 (0)	0	1	0	0	0	1	0	0
	ND	0	0	0	0	0	0	0	0 (0)	0	0	1	1	0	1	0
	OR	9	1	0	0	5	2	1	0	0 (0)	1	0	0	0	1	2
	RI	2	0	0	0	3	0	0	0	1	0 (0)	0	0	1	0	1
	SD	0	0	0	0	0	0	0	1	0	0	0 (0)	1	0	1	0
	USVI	0	0	0	0	0	0	0	1	0	0	1	0 (0)	0	1	0
	UT	18	11	8	4	13	3	1	0	0	1	0	0	5 (5)	8	0
	WV	1	0	0	0	2	0	0	1	1	0	1	1	8	0 (0)	0
	WY	1	0	0	0	3	1	0	0	2	1	0	0	0	0	0 (0)
Stand-Alone Items	AR	0 (0)	9	6	2	12	4	0	0	1	15	0	0	0	16	4
	CT	9	0 (0)	0	0	3	0	0	0	2	0	0	0	0	0	1
	HI	6	0	0 (0)	0	4	0	5	0	0	0	0	0	0	0	0
	ID	2	0	0	0 (0)	0	0	0	0	0	0	3	0	0	0	0
	IN	12	3	4	0	0 (0)	1	0	0	0	8	0	0	0	0	3
	MT	4	0	0	0	1	0 (0)	3	0	0	0	0	0	0	0	0
	NH	0	0	5	0	0	3	0 (0)	0	0	0	0	0	0	0	0
	ND	0	0	0	0	0	0	0	0 (0)	0	0	0	0	0	0	0
	OR	1	2	0	0	0	0	0	0	0 (0)	0	0	0	0	0	0
	RI	15	0	0	0	8	0	0	0	0	0 (0)	0	0	0	0	0

	State	AR	CT	HI	ID	IN	MT	NH	ND	OR	RI	SD	USVI	UT	WV	WY
	SD	0	0	0	3	0	0	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	0	0
	UT	0	0	0	0	0	0	0	0	0	0	0	0	0 (0)	0	0
	WV	16	0	0	0	0	0	0	0	0	0	0	0	0	0 (0)	0
	WY	4	1	0	0	3	0	0	0	0	0	0	0	0	0	0 (0)
Total	AR	0 (0)	10	7	4	13	4	1	0	9	17	0	0	18	17	5
	CT	10	0 (0)	0	0	6	0	0	0	3	0	0	0	10	0	1
	HI	7	0	0 (0)	0	4	0	5	0	0	0	0	0	7	0	0
	ID	4	0	0	0 (0)	0	0	0	0	0	0	3	0	4	0	0
	IN	13	6	4	0	0 (0)	3	1	0	5	11	0	0	13	2	6
	MT	4	0	0	0	3	0 (0)	4	0	2	0	0	0	3	0	1
	NH	1	0	5	0	1	4	0 (0)	0	1	0	0	0	1	0	0
	ND	0	0	0	0	0	0	0	0 (0)	0	0	1	1	0	1	0
	OR	10	3	0	0	5	2	1	0	0 (0)	1	0	0	0	1	2
	RI	17	0	0	0	11	0	0	0	1	0 (0)	0	0	1	0	1
	SD	0	0	0	3	0	0	0	1	0	0	0 (0)	1	0	1	0
	USVI	0	0	0	0	0	0	0	1	0	0	1	0 (0)	0	1	0
	UT	18	11	8	4	13	3	1	0	0	1	0	0	5 (5)	8	0
	WV	17	0	0	0	2	0	0	1	1	0	1	1	8	0 (0)	0
	WY	5	1	0	0	6	1	0	0	2	1	0	0	0	0	0 (0)

Table 10. Number of Common High School Field-Test Items Administered and Calibrated, Spring 2024

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

	State	AR	CT	HI	ID	IN	MT	NH	ND	OR	RI	SD	USVI	UT	WV	WY
	SD	0	3	0	0	2	-	2	0	0	1	0 (0)	0	-	-	0
	USVI	0	0	0	0	0	-	0	0	0	0	0	0 (0)	-	-	0
	UT	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	WV	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	WY	0	0	0	0	0	-	0	0	1	3	0	0	-	-	0 (0)
Total	AR	1 (1)	5	2	3	14	-	0	0	2	8	0	0	-	-	0
	CT	5	0 (0)	0	0	15	-	0	0	14	2	3	0	-	-	2
	HI	2	0	0 (0)	2	8	-	0	0	0	0	0	0	-	-	0
	ID	3	0	2	0 (0)	8	-	0	0	7	5	1	0	-	-	0
	IN	14	15	8	8	0 (0)	-	5	4	7	15	4	1	-	-	2
	MT	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	NH	0	0	0	0	5	-	0 (0)	0	0	2	2	0	-	-	0
	ND	0	0	0	0	5	-	0	0 (0)	0	0	0	1	-	-	0
	OR	2	14	0	8	7	-	0	0	0 (0)	4	0	0	-	-	4
	RI	8	2	0	5	15	-	2	0	4	0 (0)	2	0	-	-	3
	SD	0	3	0	1	4	-	2	0	0	2	0 (0)	0	-	-	0
	USVI	0	0	0	0	1	-	0	1	0	0	0	0 (0)	-	-	0
	UT	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	WV	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	WY	0	2	0	0	2	-	0	0	4	3	0	0	-	-	0 (0)

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

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

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

Table 11. Field-Test Item Administration, Rubric Validation, and Item Data Review, Spring 2024

Grade Band and Item Type	Number of Field-Test Items Administered	Number of Items Rejected Before/During Rubric Validation	Number of Items Sent to Data Review	Number of Items Rejected at Data Review	Number of Items Remaining
Elementary School	166 (21)	3 (0)	94 (16)	26 (5)	137 (16)
Cluster	69 (5)	3 (0)	11 (0)	5 (0)	61 (5)
Stand-Alone	97 (16)	0 (0)	83 (16)	21 (5)	76 (11)
Middle School	176 (21)	3 (1)	96 (15)	33 (6)	140 (14)
Cluster	90 (9)	3 (1)	35 (4)	20 (3)	67 (5)
Stand-Alone	86 (12)	0 (0)	61 (11)	13 (3)	73 (9)
High School	136 (29)	2 (0)	78 (17)	25 (3)	109 (26)
Cluster	67 (11)	2 (0)	22 (3)	11 (2)	54 (9)
Stand-Alone	69 (18)	0 (0)	56 (14)	14 (1)	55 (17)
Total	478 (71)	8 (1)	268 (48)	84 (14)	386 (56)

Note. Connecticut-owned items are indicated in the parentheses. Sound Dakota legacy items were excluded.

Table 12 summarizes the Shared Science Assessment Item Bank after adding the field-test items that were administered in 2024 and passed rubric validation and item data review. The numbers in parentheses present the number of items owned by Connecticut.

Table 12. Shared Science Assessment Item Bank, Spring 2024

Grade Band and Item Type	Science Discipline			Item Bank Total ^a
	Earth and Space Sciences	Life Sciences	Physical Sciences	
Elementary School	232 (25)	233 (30)	301 (38)	766 (93)
Cluster	128 (13)	113 (12)	154 (16)	395 (41)
Stand-Alone	104 (12)	120 (18)	147 (22)	371 (52)
Middle School	220 (24)	297 (44)	261 (27)	778 (95)
Cluster	107 (9)	146 (22)	125 (11)	378 (42)
Stand-Alone	113 (15)	151 (22)	136 (16)	400 (53)
High School	136 (35)	263 (48)	168 (49)	567 (132)
Cluster	59 (12)	119 (22)	68 (16)	246 (50)
Stand-Alone	77 (23)	144 (26)	100 (33)	321 (82)
Total	588 (84)	793 (122)	730 (114)	2111 (320)

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

^aCount excludes fourteen MOU items that do not align to the NGSS.

3.3 TEST DESIGN

The science tests were assembled under an adaptive design, with the exception of the braille and paper-pencil forms. Adaptive tests were assembled using CAI’s adaptive testing algorithm. The adaptive item selection algorithm selects items based on their content value and information value. At any given point during the test, the content value of an item is determined by its contribution to meeting the blueprint, given the content characteristics of the items that have already been administered. During the test, the content value increases for items that exhibit features that have not met their designated minimum as the end of the test approaches. Similarly, 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 the test blueprint is followed for each individual student. The Connecticut NGSS Assessment blueprints are presented in this technical report in Volume 2, Section 4.2, Test Blueprints. Details of CAI’s item selection algorithm are described in Volume 2, Appendix 2-L, Adaptive Algorithm Design.

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

The main characteristics of the blueprint were that any performance expectation (PE) could be tested only once (indicated by the values of 0 and 1 for the minimum and maximum values of the individual PEs in the test blueprints; see Section 4.2, Test Blueprints of Volume 2). In general, no more than one item cluster or two stand-alone items could be sampled from the same Disciplinary Core Idea (DCI), and no more than three total items could be sampled from the same DCI (as indicated by the minimum and maximum values in the rows representing DCIs). Some specific constraints for the Connecticut NGSS Assessment blueprint were that for grades 5 and 8, students would get two stand-alone items from the Earth Systems DCI (rather than one for other DCIs in the Earth and Space Sciences discipline) because it had the most PEs and was rated the highest in the district responses. In addition, three DCIs in grade 11—Motion and Stability, Waves, and Earth’s Place in the Universe—were constrained to not receive an item cluster due to low content priority ratings from districts.

A segmented test design was used for the 2018 independent field test; items were grouped by science discipline. In 2019, a non-segmented test design was used for the first operational test administration; items were no longer grouped by science discipline. Instead, students received items from different disciplines in random order. Embedded field-test items were randomly positioned in the test and randomly distributed among students. Every student received either one item cluster or five stand-alone items as field-test items throughout the test. Since 2021, a similar non-segmented test design with embedded field-test items was used. The only difference since 2021 was that every student received either one item cluster or four stand-alone items as field-test items throughout the test.

4. FIELD-TEST CLASSICAL ANALYSIS

As explained in Section 3, Item Bank and Test Design, science items administered as field-test items underwent rubric validation and data review. Items were flagged for data review based on business rules defined on classical item statistics. Except for response times, the classical item statistics are computed for individual assertions, whereas the business rules for flagging are defined at the item level.

In general, item statistics used to flag items for data review were computed using the student responses of the state that owned the items. However, for Independent College and Career Readiness (ICCR) items, the flagging rules were defined on the item statistics computed from the combined data of states that used ICCR items. In 2024, those states were Arkansas, Connecticut, Idaho, Indiana, New Hampshire, North Dakota, Rhode Island, South Dakota, U.S. Virgin Islands, Utah, and West Virginia. Furthermore, to compute the differential item functioning (DIF) statistics for the field-test items, the data from all states were combined to obtain a sufficient number of students for each demographic group.

The criteria for flagging and reviewing items are provided in Table 13, and the statistics are described in Section 4.1, Item Discrimination, through Section 4.4, Differential Item Functioning. Items flagged for data review were reviewed by a committee, as explained in Section 3, Item Bank and Test Design.

Table 13. Thresholds for Flagging in Classical Item Analysis

Analysis Type	Flagging Criteria
Item Discrimination	Average biserial correlation < 0.25 (across the assertions within an item)
	One or more assertions with a biserial correlation < 0.05
Item Difficulty (Clusters)	Average p -value < 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
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

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 is able to 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 from the item bank 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 13.

4.3 RESPONSE TIME

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

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

4.4 DIFFERENTIAL ITEM FUNCTIONING

DIF refers to items that appear to function differently across identifiable groups, typically across different demographic groups. Identifying DIF is important because it provides a statistical indicator that an item may contain cultural or other biases. 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, and various characteristics of the educational system may also lead to DIF.

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 statistic for balancing the stability and sensitivity of the DIF scoring category selection. For dichotomous items, the

following statistics were computed: the $MH\chi^2$ value, the conditional odds ratio, and the MH-delta. For polytomous items, the $GMH\chi^2$ and the standardized mean difference (SMD [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

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

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

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

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

where

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

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

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

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

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

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

DIF analysis was conducted for all field-test items with at least 200 responses per item in each subgroup (Zwick, 2012) to detect potential item bias for major demographic groups. Student responses from multiple states were combined to minimize the number of items with insufficient sample sizes for one or more demographic groups.

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

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

Similar to how the general MH statistic is used to classify items 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 14. Furthermore, assertions were categorized positively (i.e., +A, +B, or +C), signifying that an item favored the focal group (e.g., African American, Hispanic, or female), or negatively (i.e., –A, –B, or –C), signifying that an item favored the reference group (e.g., White or male).

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

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

Table 14. DIF Classification Rules

Assertions	
Category	Rule
C	$GMH\chi^2$ is significant at .05 and $\frac{ SMD }{\sigma} > .25$
B	$GMH\chi^2$ is significant at .05 and $.17 < \frac{ SMD }{\sigma} \leq .25$
A	$GMH\chi^2$ is not significant at .05 or $\frac{ SMD }{\sigma} \leq .17$

Note that, for the 2018 field test, a slightly less strict criterion was used for item clusters with 10 or more assertions (i.e., three or more assertions with “C” DIF in the same direction). The change was made taking into consideration the feedback received from several Technical Advisory Committees (TACs) and modified such that the rate of flagging items for DIF was similar for item clusters and stand-alone items (based on the flagging rates computed on items field tested in 2018).

4.5 CLASSICAL ANALYSIS RESULTS

This section presents a summary of results from classical item analysis of the field-test items administered in 2024. A total of 95 field-test items were administered in Connecticut; 94 passed rubric validation. Among these items, 11 items were flagged for item discrimination, 8 items were flagged for p -value, 48 items were flagged for response time, and 4 items were flagged for DIF according to the criteria used in 2024 (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. The total number of field-test items flagged and the total number of field-test items that passed item data review in 2024 were summarized in Table 11.

Table 15 and Table 16 provide the summary of the p -values and biserial correlations for the science field-test items administered in Connecticut in 2024 that passed rubric validation. The statistics were computed using Connecticut data only. The average values across the assertions within an item were used to compute percentiles and ranges.

Table 15. Distribution of p -Values for Field-Test Items, Spring 2024

Grade	Total FT Items	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
5	25	0.12	0.18	0.37	0.48	0.57	0.69	0.70
8	24	0.07	0.21	0.33	0.44	0.50	0.61	0.63
11	39	0.10	0.12	0.29	0.36	0.42	0.53	0.57

Table 16. Distribution of Item Biserial Correlations for Field-Test Items, Spring 2024

Grade	Total FT Items	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
5	25	0.31	0.33	0.43	0.47	0.54	0.63	0.65
8	24	0.14	0.33	0.41	0.47	0.57	0.69	0.73
11	39	-0.12	0.26	0.40	0.47	0.60	0.70	0.71

Table 17 presents the summary of the response times by item type (item cluster or stand-alone item) for field-test items administered in 2024.

Table 17. Summary of Response Times for Field-Test Items, Spring 2024

Grade	Item Type	Total FT Items	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
5	Cluster	5	7.50	7.78	8.90	9.70	11.10	12.86	13.30
	Stand-Alone	20	2.40	2.59	3.60	4.20	4.58	5.14	5.80
8	Cluster	9	8.20	9.04	10.60	11.60	12.90	15.30	15.50
	Stand-Alone	15	2.00	2.21	2.50	2.80	3.20	4.59	4.80
11	Cluster	15	6.50	6.57	7.25	8.20	9.10	11.76	15.40
	Stand-Alone	24	1.70	1.95	2.28	2.60	3.63	5.39	5.50

Table 18 presents the number of field-test items flagged for DIF for each item type and demographic group included in the DIF analyses in 2024.

Table 18. Differential Item Functioning Classifications for Field-Test Items, Spring 2024

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 5											
Items Evaluated	Cluster	5	0	0	4	0	5	0	5	5	5
	Stand-Alone	20	0	1	17	0	20	1	20	20	20
Items Flagged C	Cluster	0	0	0	0	0	0	0	0	0	0
	Stand-Alone	0	0	0	1	0	0	0	0	0	0
% Items Flagged C	Cluster	0	NA	NA	0	NA	0	NA	0	0	0
	Stand-Alone	0	NA	0	5.88	NA	0	0	0	0	0
Grade 8											
Items Evaluated	Cluster	9	0	0	9	0	9	1	9	9	9
	Stand-Alone	15	0	0	15	0	15	5	14	15	15
Items Flagged C	Cluster	1	0	0	0	0	0	0	0	0	0
	Stand-Alone	1	0	0	1	0	0	0	0	0	0
% Items Flagged C	Cluster	11.11	NA	NA	0	NA	0	0	0	0	0
	Stand-Alone	6.67	NA	NA	6.67	NA	0	0	0	0	0
Grade 11											
Items Evaluated	Cluster	15	0	0	15	0	15	0	10	15	15
	Stand-Alone	24	0	0	24	0	24	0	9	24	24
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	NA	NA	0	NA	0	NA	0	0	0
	Stand-Alone	0	NA	NA	0	NA	0	NA	0	0	0

Note. Full DIF group names: ^aAmerican Indian/Alaskan Native; ^bHawaiian/Pacific Islander; ^cEconomically Disadvantaged vs. Non-Economically Disadvantaged

5. ITEM CALIBRATION

5.1 MODEL DESCRIPTION

In discussing item response theory (IRT) models for Connecticut, 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:

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

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

Connecticut NGSS Assessment items 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 Connecticut NGSS Assessment 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 using 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 dimensions¹. Whereas traditional unidimensional IRT models assume that all assertions (the basic units of analysis) are independent given a single underlying trait θ , we now assume the conditional independence of assertions, given the underlying latent trait θ and all nuisance dimensions:

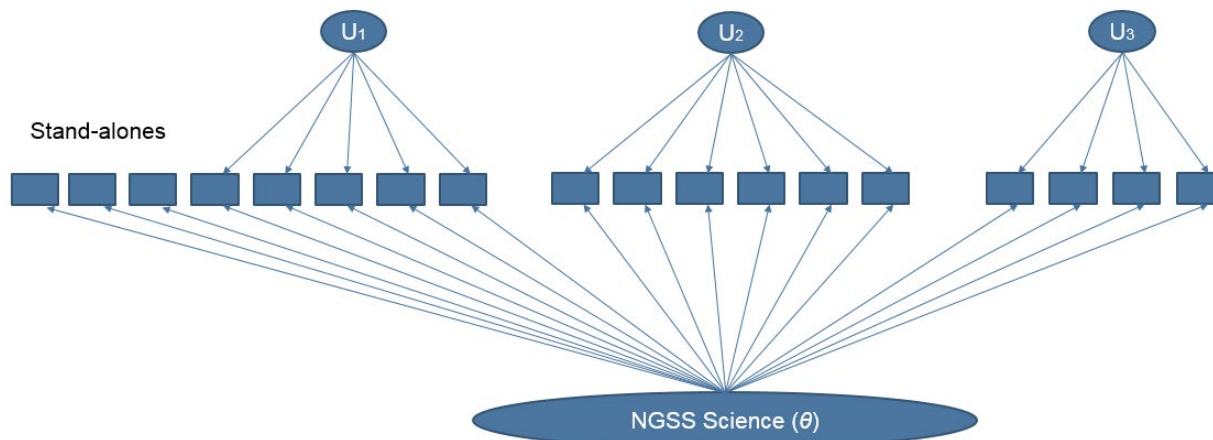
$$P(\mathbf{z}_j | \theta_j, \mathbf{u}_j) = \prod_{i \in \text{SA}} P(z_{ij} | \theta_j) \prod_{g=1}^G \prod_{i \in g} P(z_{ij} | \theta_j, u_{jg}), \quad (2)$$

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 $P(z_{ij} | \theta_j, u_{jg})$ 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; instead, it is sufficient that they are independent, given the general dimension θ .

The model structure of the IRT model for science is illustrated in Figure 1. Note that stand-alone items can be scored with more than one assertion. The assertions of stand-alone items with more than one assertion, but fewer than four assertions, are 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. The few stand-alone items with four or more assertions are treated as item clusters to take into account the conditional dependencies.

¹ The term *nuisance dimension* pertains to within-item local dependencies among scoring assertions and should not be confused with the three dimensions of the NGSS Framework.

Figure 1. Directed Graph of the Science IRT 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 Connecticut NGSS Assessment. For binary data, the Rasch testlet model is defined as:

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

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 incorporated in unidimensional models for polytomous data.

5.1.3 Multigroup Model

The Shared Science Assessment Item Bank is calibrated concurrently using all the items administered in any state that collaborates with CAI on their new science assessments. In the calibration, each state is treated as a population of students or a group. Overall group differences are taken into account by allowing a group-specific distribution of the overall proficiency variable

θ . Specifically, for every student j belonging to group k , $k = 1, \dots, K$, a normal distribution is assumed,

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

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

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

5.2 ESTIMATION

A separate IRT model is fit for each grade band. The parameters of the IRT model are estimated using the marginal maximum likelihood (MML) method. In the MML method, the latent proficiency variable θ_j and the vector of nuisance parameters \mathbf{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 \mathbf{z}_j for student j ,

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

where $\mathbf{\Sigma}$ 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 overall log likelihood equation is very high, issues with dimensionality can be avoided because the integration over the high-dimensional latent (θ, \mathbf{u}) space can be carried out as a sequence of computations in two-dimensional space (θ, \mathbf{u}_g) (Gibbons & Hedeker, 1992; Rijmen, 2010).

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

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

Starting in 2021, CAIRT (Cambium Assessment IRT) is used to calibrate item parameters. CAIRT was specifically developed by CAI to calibrate the multigroup Rasch model on very large data sets because estimation times in commercially available software (i.e., flexMIRT) became prohibitive. CAIRT relies on the same estimation methods as the Bayesian networks with the logistic regression (BNL; Rijmen, 2006), a suite of Matlab functions for estimating a wide variety of latent variable models. BNL uses an efficient expectation-maximization (EM) algorithm based on the graphical model theory (e.g., Rijmen, 2010). CAI has cross-validated parameter estimates from CAIRT with BNL and flexMIRT under various scenarios (Rijmen, Liao, & Lin, 2021). CAIRT is a web application that is available at no cost to members of the MOU. In 2024, field-test items were calibrated in CAIRT using the same procedure used in 2021.

Table 19 provides an overview of the groups per grade band for calibration of the 2024 field-test items. All items were calibrated on at least 1,500 student responses.

Table 19. Groups Per Grade Band for the Spring 2024 Calibration of Field-Test Items

Group	Elementary School	Middle School	High School
Arkansas	X	X	X
Connecticut	X	X	X
Hawaii	X	X	X
Idaho	X	X	X
Indiana	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
U.S. Virgin Islands	X	X	X
Utah	X	X	
West Virginia	X	X	
Wyoming	X	X	X

5.3 OVERVIEW OF THE OPERATIONAL ITEM BANK

Figure 2, Figure 3, and Figure 4 display the histogram of the difficulty parameters for grades 5, 8, and 11, respectively, for all assertions that are part of the Connecticut NGSS Assessment operational pool. The figures also display the student proficiency distributions. The distribution of the difficulty parameter overlaps well with the proficiency distribution in grade 5 and grade 8. The grade 11 assertions are slightly more difficult than the student proficiency in general.

Figure 2. Connecticut NGSS Assessment Assertion Difficulty and Student Proficiency Distributions, Grade 5, Spring 2024

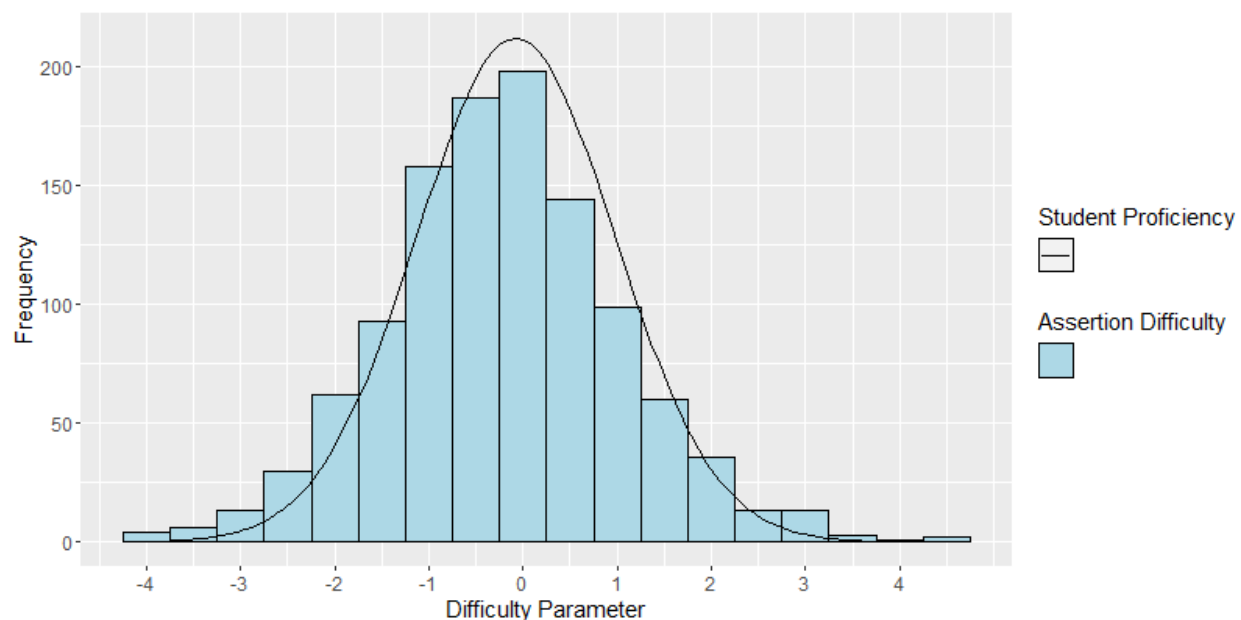


Figure 3. Connecticut NGSS Assessment Assertion Difficulty and Student Proficiency Distributions, Grade 8, Spring 2024

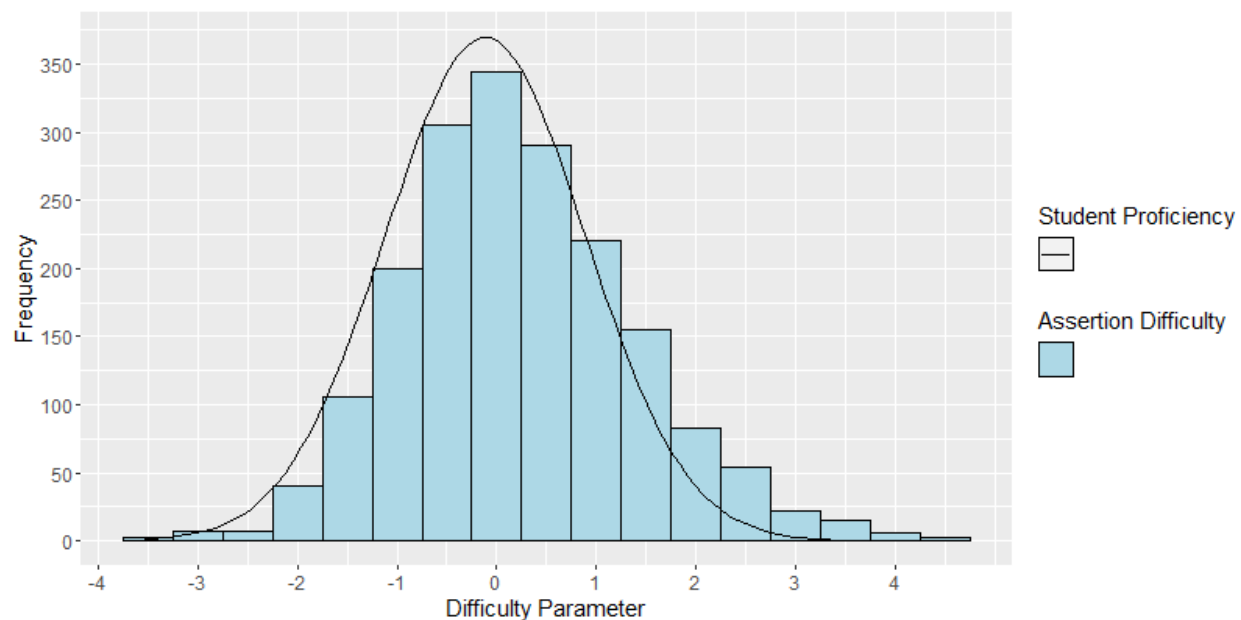
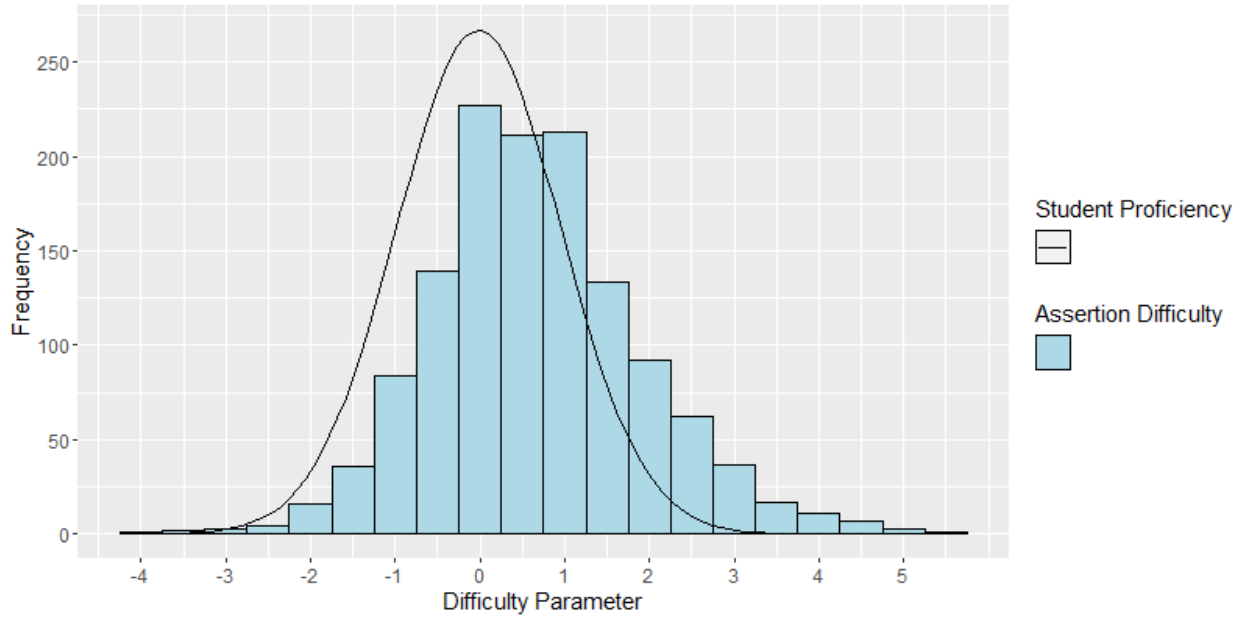


Figure 4. Connecticut NGSS Assessment Assertion Difficulty and Student Proficiency Distributions, Grade 11, Spring 2024



6. SCORING

6.1 MARGINAL MAXIMUM LIKELIHOOD FUNCTION

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

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

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. See Rijmen, Jiang, and Turhan (2018) for more details of the MMLE estimator and the validation study by Connecticut State Department of Education (2019) for the use of this estimator.

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

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

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

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

which is the likelihood under the unidimensional Rasch model.

6.2 DERIVATIVE

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

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

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

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

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

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

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

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

and

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

Therefore, we have,

$$\begin{aligned}
\frac{\frac{dp_i}{d\theta}}{p_i} &= q_i, \quad \frac{\frac{dq_i}{d\theta}}{q_i} = -p_i, \\
\frac{\frac{dp_{ig}}{d\theta}}{p_{ig}} &= q_{ig}, \quad \frac{\frac{dq_{ig}}{d\theta}}{q_{ig}} = -p_{ig},
\end{aligned}$$

$$\begin{aligned}\frac{\frac{d^2 p_i}{d\theta^2}}{p_i} - \left(\frac{\frac{dp_i}{d\theta}}{p_i}\right)^2 &= -p_i q_i, \\ \frac{\frac{d^2 q_i}{d\theta^2}}{q_i} - \left(\frac{\frac{dq_i}{d\theta}}{q_i}\right)^2 &= -p_i q_i, \\ \frac{\frac{d^2 p_{ig}}{d\theta^2}}{p_{ig}} - \left(\frac{\frac{dp_{ig}}{d\theta}}{p_{ig}}\right)^2 &= -p_{ig} q_{ig}, \text{ and} \\ \frac{\frac{d^2 q_{ig}}{d\theta^2}}{q_{ig}} - \left(\frac{\frac{dq_{ig}}{d\theta}}{q_{ig}}\right)^2 &= -p_{ig} q_{ig}.\end{aligned}$$

6.3 EXTREME CASE HANDLING

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

6.4 STANDARD ERROR OF MEASUREMENT

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

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

where $I(\hat{\theta}_{MMLE})$ is the observed information evaluated at $\hat{\theta}_{MMLE}$. The observed information is calculated as $I(\theta^2) = -\frac{d^2 l(\theta)}{d\theta^2}$, where $\frac{d^2 l(\theta)}{d\theta^2}$ is defined in Section 6.2, Derivative. Note that the calculation of the SEM depends on the unique set of items that each student answers and their estimate of θ . Different students have different SEM 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 SEM is set at the truncation value for the standard error.

6.5 SCORING INCOMPLETE TESTS

The Connecticut NGSS Assessment is assembled on the fly using an adaptive testing design. For science, a test is considered “attempted” if a student responded to at least one item (cluster or stand-alone). An attempted test is considered complete if the student responds to all the operational items. Otherwise, the test is “incomplete.”

Tests that are attempted but incomplete receive overall science scores. In order to receive a discipline score (i.e., Life Sciences, Physical Sciences, Earth and Space Sciences), a student must have attempted the corresponding discipline of the test. The MMLE is used to score the attempted incomplete tests, counting unanswered items as incorrect. If the identities of the unanswered items are unknown due to the test being assembled on the fly, the item parameters for a “typical” item are used. If a missing item is an item cluster, the simulated item parameters of the missing item are the item parameters of item cluster 4482 for grade 5, item cluster 3781 for grade 8, and item cluster 4350 for grade 11, which are operational item clusters that are typical for the Connecticut NGSS Assessment item pool used in Connecticut 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 4047 for grade 5, item 4529 for grade 8, and item 4555 for grade 11, which are operational stand-alone items that are typical for the Connecticut NGSS Assessment item pool used in Connecticut.

If the identities of items that have not been answered 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, scale scores are computed for

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

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

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

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

where a and b are the slope and intercept of the linear transformation that transforms $\hat{\theta}_{MMLE}$ to the reporting scale (refer to Table 20). The SEM for the estimated scale score is obtained as

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

In 2019, the slope a and intercept b were chosen so that the center of the reporting scale of each grade (500, 800, and 1100, respectively) is at the grade mean of the 2019 base-year and has a standard deviation of 28. Furthermore, for each grade, the reporting scale ranges approximately from the base-year mean minus 3.5 times the standard deviation to the base-year mean plus 3.5 times the standard deviation. Specifically, for grade 5, the slope and intercept were obtained as

$$\begin{aligned}
 SS &= 28\theta^* + 500 \\
 &= 28 \frac{\theta - \hat{\mu}_\theta}{\hat{\sigma}_\theta} + 500 \\
 &= \frac{28}{\hat{\sigma}_\theta} \theta + \left(500 - \frac{28\hat{\mu}_\theta}{\hat{\sigma}_\theta} \right),
 \end{aligned}$$

where the second line stems from standardizing theta, $\theta^* = \frac{\theta - \hat{\mu}_\theta}{\hat{\sigma}_\theta}$. For grades 8 and 11, the slope and intercept can also be derived similarly.

Per grade, Table 20 presents the intercept, slope, LOT, HOT, lowest obtainable scale score (LOSS), and highest obtainable scale score (HOSS) values used for the 2019 reporting scale. The scale score distribution is reported for overall science in Appendix 1-D, Distribution of Scale Scores and Performance Levels. The scale score distribution is reported for the science disciplines in Appendix 1-E, Distribution of Scale Scores by Science Discipline.

Table 20. Science Reporting Scale Linear Transformation Constants, Theta, and Corresponding Scaled-Score Limits for Extreme Ability Estimates (for 2019 θ Scale)

Grade	Slope (a)	Intercept (b)	Lowest of Theta (LOT)	Highest of Theta (HOT)	Lowest of Scale Score (LOSS)	Highest of Scale Score (HOSS)
5	31.684	500	–3.15	3.12	400	599
8	31.766	800	–3.14	3.11	700	899
11	30.792	1100	–3.24	3.21	1000	1199

6.7 RULES FOR CALCULATING PERFORMANCE LEVELS

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

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

Table 21. Performance-Level Cut Scores

Grade	Cut 1	Cut 2	Cut 3
5	468	498	535
8	772	798	842
11	1073	1099	1141

6.7.1 Strengths and Weaknesses for Disciplines Relative to Proficiency Cut Score

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

- if $(\hat{\theta}_{discipline} < \theta_{proficient} - 1.5 * SEM(\hat{\theta}_{discipline}))$, then performance is classified as *Below Standard*;
- if $(\theta_{proficient} - 1.5 * SEM(\hat{\theta}_{discipline}) \leq \hat{\theta}_{discipline} < \theta_{proficient} + 1.5 * SEM(\hat{\theta}_{discipline}))$, then performance is classified as *Approaching Standard*; and
- if $(\hat{\theta}_{discipline} \geq \theta_{proficient} + 1.5 * SEM(\hat{\theta}_{discipline}))$, then performance 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 IDEAS AND SCIENCE AND ENGINEERING PRACTICES

6.8.1 Relative to Overall Performance

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 Ideas (DCI). Starting in 2022, there is also reporting for aggregated units for four claims corresponding to Science and Engineering Practices (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,

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

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

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

and over students of the group on which is reported,

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

where n_{jDCI} is the number of assertions related to the DCI for student j , and n_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

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

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

We do not suggest direct reporting of the statistic $\bar{\delta}_{DCIm}$; instead, we recommend reporting in the aggregate whether a group of students performs 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 $\bar{\delta}_{DCIm} \leq -1.5 * SEM(\bar{\delta}_{DCIm})$, then performance is *worse than* on the overall test.
- If $\bar{\delta}_{DCIm} \geq 1.5 * SEM(\bar{\delta}_{DCIm})$, then performance is *better than* on the overall test.
- Otherwise, performance is *similar to* on the overall test.
- If $SEM(\bar{\delta}_{DCIm}) > 0.2$, data are insufficient.

6.8.2 Relative to Proficiency Cut Score

DCI-level scores for aggregated units can be computed using the same method as outlined in Section 6.8.1, Relative to Overall Performance, but with the expected score computed at the theta value corresponding to the proficiency cut score:

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

The following is reported for DCIs for aggregate units:

- If $\bar{\delta}_{DCIm} \leq -1.5 * SEM(\bar{\delta}_{DCIm})$, then performance is *below* the proficiency cut score.
- If $\bar{\delta}_{DCIm} \geq 1.5 * SEM(\bar{\delta}_{DCIm})$, then performance is *above* the proficiency cut score.

- Otherwise, performance is *approaching* the proficiency cut score.
- If $SEM(\bar{\delta}_{DCIm}) > 0.2$, data are insufficient.

7. QUALITY CONTROL PROCEDURES

CAI’s quality assurance (QA) procedures are built on two key principles: automation and 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 performs 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 Monitoring (QM) System that yields item statistics, blueprint match rates, item exposure rates, and cheating analysis 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 (i.e., proportion correct) and discrimination (i.e., biserial/polyserial correlation). Classical analysis indicators for assertions are also available. Section 4.1, Item Discrimination; Section 4.2, Item Difficulty; and Section 4.4, Differential Item Function, of this volume describe the statistical approaches used for item analysis.

In addition, the report provides item fit and cluster-based item drift (Cui, 2023) statistics based on the IRT model. The report is configurable and can be produced to flag only items with statistics that fall outside a specified range or to generate reports based on all items in the pool.

As a routine practice, CAI psychometricians monitor classical item statistics, item fit, and item drift periodically during the testing window. When a QA report flags items or assertions for poor performance, using the same criteria as evaluating FT items, a CAI psychometrician undertakes a systematic investigation to identify and address the issues and develops recommendations for each flagged item. Recommendations might include item revision, elimination, or further piloting.

7.1.2 Blueprint Match

As Section 2.4, Simulations of this volume discusses, test blueprints are evaluated before the testing window begins to identify potential blueprint violations. If a blueprint violation occurs during the Operational testing window, a CAI psychometrician undertakes a systematic investigation to identify and address the issues and develops a plan to remedy the violations.

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.

In Spring 2024, every test in all three grades met the blueprint specifications at the level of the science disciplines, which is the lowest content level at which scores for individual students are reported. Blueprint match is discussed in detail in this technical report in Volume 2, Test Development, for both simulated and operational test administrations.

7.1.3 Item Exposure Rates

As part of the QA procedures, item exposure reports are generated, allowing test items to be monitored for unexpectedly large exposure rates or unusually low item-pool usage throughout the testing window. As with other reports, it is possible to examine the exposure rate for all items or 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.

In Spring 2024, most of the test items were administered to 20% or fewer test takers in all grades. Only 1.56% of the items in grade 5, 1.56% of the items in grade 8, and 1.48% of the items in grade 11 were administered to 20% or more test takers at that grade. More details are discussed in Volume 2, Test Development, of this technical report.

7.1.4 Cheating Detection Analysis

As part of the QA procedures, a forensics report can also be provided to identify possible irregularities in test administration for further investigation. Unusual patterns of responding at the student level can be aggregated to the test session, test administrator, and school levels to identify possible group-level testing anomalies. CAI psychometricians can monitor testing anomalies throughout the testing window. Evidence can be evaluated with respect to item response times, and irregular item response patterns using the cluster-based person-fit index (Lin, Rijmen, Tao, & van Wamelen, 2021). 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 estimates scores separately using the procedures described within this report.

8. REFERENCES

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: Author.
- Cai, L. (2017). flexMIRT®: Flexible multilevel multidimensional item analysis and test scoring (version 3.51) [computer software]. Chapel Hill, NC: Vector Psychometric Group.
- Cui, M. (2023, July 25–28). *Item drift for item clusters* [Conference presentation]. The 88th Annual Meeting of the Psychometric Society, Maryland, United States.
- Connecticut State Department of Education. (2019). *Validating American Institutes for Research's calibration and scoring processes for science assessments* (Research Report). Hartford, CT: Author.
- Dorans, N. J., & Schmitt, A. P. (1991). *Constructed response and differential item functioning: A pragmatic approach* (ETS Research Report No. 91–47). Princeton, NJ: Educational Testing Service.
- Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, 57(3), 423–436.
- Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Koretz, D., & Hamilton, L. S. (2006). Testing for accountability in K–12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 531–578). Westport, CT: American Council on Education/Praeger.
- Lin, Z., Tao, J., Rijmen, F., & van Wamelen, P. (2024). Asymptotically correct person fit z-statistics for the Rasch testlet model. *Psychometrika*. <https://doi.org/10.1007/s11336-024-09997-y>
- National Center for Education Statistics. (2010). *Statistical methods for protecting personally identifiable information in aggregate reporting* (Statewide Longitudinal Data System Technical Brief, Brief 3). Retrieved from <https://nces.ed.gov/pubs2011/2011603.pdf>.
- National Research Council. (2012). *A framework for K–12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: The National Academies Press.
- Rijmen, F. (2006). *BNL: A Matlab toolbox for Bayesian networks with logistic regression nodes* (Technical Report). Amsterdam: VU University Medical Center.
- Rijmen, F. (2010). Formal relations and empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement*, 47(3), 361–372.

- Rijmen, F., Jiang, T., & Turhan, A. (2018, April). *An item response theory model for new science assessments*. Paper presented at the National Council on Measurement in Education, New York, NY.
- Rijmen, F., Liao, D., & Lin, Z. (2021). *The Rasch testlet model for the calibration of three-dimensional science assessments: A software comparison* [White paper]. Washington, DC: Cambium Assessment, Inc.
- Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, 28, 237–247.
- Somes, G. W. (1986). The generalized Mantel Haenszel statistic. *The American Statistician*, 40, 106–108.
- Wang, W. C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement*, 29(2), 126–149.
- Yen, W. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30, 187–213.
- Zwick, R. (2012). *A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement* (ETS Research Report No. 12–08). Princeton, NJ: Educational Testing Service.

Connecticut Next Generation Science Standards Assessment

2023–2024

Volume 2: Test Development



CONNECTICUT STATE
DEPARTMENT OF EDUCATION

TABLE OF CONTENTS

1.	INTRODUCTION	1
1.1	Claim Structure	2
1.2	Underlying Principles Guiding Development.....	2
1.3	Organization of This Volume	3
2.	ITEM DEVELOPMENT PROCESS THAT SUPPORTS VALIDITY OF CLAIMS	3
2.1	Overview	3
2.2	Item Specifications.....	11
2.3	Selection and Training of Item Writers	13
2.4	Internal Review	14
2.4.1	Preliminary Review.....	14
2.4.2	Scoring Entry and Review.....	15
2.4.3	Content Review One.....	16
2.4.4	Edit Review	16
2.4.5	Senior Review.....	17
2.5	Review by State Personnel and Stakeholder Committees	17
2.5.1	State Review.....	18
2.5.2	Content Advisory Committee Reviews	18
2.5.3	Language Accessibility, Bias, and Sensitivity Committee Reviews	19
2.5.4	Markup for Translation and Accessibility Features	20
2.6	Field-Testing.....	20
2.7	Post-Field-Test Review.....	21
2.7.1	Rubric Validation.....	21
2.7.2	Data Review.....	23
3.	SHARED SCIENCE ASSESSMENT ITEM BANK SUMMARY.....	26
3.1	Current Composition of the Shared Science Assessment Item Bank.....	27
3.2	Strategy for Item Bank Evaluation and Replenishment.....	33
4.	CONNECTICUT NGSS ASSESSMENT TEST CONSTRUCTION	33
4.1	Test Design	33
4.2	Test Blueprints	34
4.3	Online Test Construction	46
4.4	Paper-Pencil Accommodation Form Construction	51
5.	SIMULATION SUMMARY REPORT	51
5.1	Factors Affecting Simulation Results	52
5.2	Results of Simulated Test Administrations: English	52
5.2.1	Summary of Blueprint Match.....	52
5.2.2	Item Exposure	52
5.2.3	Precision	53

5.3	Results of Simulated Test Administrations: Spanish.....	54
5.3.1	Summary of Blueprint Match.....	54
5.3.2	Item Exposure	54
5.3.3	Precision	55
6.	OPERATIONAL TEST ADMINISTRATION SUMMARY REPORT	55
6.1	Blueprint Match	55
6.2	Item Exposure	56
7.	REFERENCES	57

LIST OF TABLES

Table 1. Summary of How Each Step of Development Supports the Validity of Claims	10
Table 2. Sample Science Item Cluster Specifications for Middle School Life Sciences Performance Expectation	12
Table 3. Summary of the 2023-2024 Content Advisory Committee Meetings	19
Table 4. Summary of the 2023-2024 Fairness Committee Meetings	20
Table 5. Summary of Data Review Committee Meetings	24
Table 6. Science Interaction Types and Descriptions	27
Table 7. Spring 2024 Shared Science Assessment Operational and Field-Test Item Bank	29
Table 8. Spring 2024 Shared Science Assessment Operational Item Bank	29
Table 9. Spring 2024 Shared Science Assessment Field-Test Item Bank	29
Table 10. Spring 2024 Shared Science Assessment Operational and Field-Test Item Bank by Science Discipline	30
Table 11. Spring 2024 Shared Science Assessment Operational and Field-Test Item Bank by Disciplinary Core Idea	31
Table 12. Science Test Blueprint, Grade 5	35
Table 13. Science Test Blueprint, Grade 8	38
Table 14. Science Test Blueprint, Grade 11	42
Table 15. Spring 2024 Connecticut NGSS Assessment 85th Percentile Testing Times by Grade	46
Table 16. Spring 2024 Connecticut NGSS Assessment Operational and Field-Test Item Pool..	46
Table 17. Spring 2024 Connecticut NGSS Assessment Operational Item Pool	47
Table 18. Spring 2024 Connecticut NGSS Assessment Field-Test Item Pool	47
Table 19. Spring 2024 Connecticut NGSS Assessment Operational and Field-Test Item Pool by Science Discipline	48
Table 20. Spring 2024 Connecticut NGSS Assessment Operational and Field-Test Item Pool by Disciplinary Core Idea	49
Table 21. Spring 2024 Item Exposure Rates by Grade: Percentage of Items by Exposure Rate, Across All English Online Simulation Sessions	53
Table 22. Spring 2024 Standard Errors of Ability Estimates, by Grade, Across All English Online Simulation Sessions	53
Table 23. Spring 2024 Spanish Operational Item Pool	54
Table 24. Spring 2024 Item Exposure Rates by Grade: Percentage of Items by Exposure Rate, Across All Spanish Simulation Sessions	54
Table 25. Spring 2024 Standard Errors of Ability Estimates, by Grade, Across All Spanish Online Simulation Sessions	55
Table 26. Spring 2024 Item Exposure Rates by Grade: Percentage of Items by Exposure Rate, Across All Test Administrations	56

LIST OF FIGURES

Figure 1. Structure of Three-Dimensional Item Clusters.....	4
Figure 2. Example of an NGSS Item Cluster.....	5
Figure 3. Example of NGSS Scoring Assertions.....	8
Figure 4. Features of the REVISE Software.....	22

LIST OF APPENDICES

Appendix 2-A. Item Writer Training Materials
Appendix 2-B. Item Specifications Grade 3 through High School
Appendix 2-C. Style Guide for Science Items
Appendix 2-D. Item Review Checklist
Appendix 2-E. Content Advisory Committee Review Training Slides
Appendix 2-F. Content Advisory Committee Participant Details
Appendix 2-G. Fairness Committee Review Training Slides
Appendix 2-H. Fairness Committee Participant Details
Appendix 2-I. Sample Data Review Training Materials
Appendix 2-J. Data Review Committee Participant Details
Appendix 2-K. Example Item Interactions
Appendix 2-L. Shared Science Assessment Item Bank
Appendix 2-M. Connecticut NGSS Assessment Item Pool
Appendix 2-N. Adaptive Algorithm Design

1. INTRODUCTION

Connecticut adopted the Next Generation Science Standards (NGSS) in 2015. The Connecticut State Department of Education (CSDE) and its assessment vendor, Cambium Assessment, Inc. (CAI; formerly the American Institutes for Research [AIR]), developed and administered a new online assessment to measure the new standards. The Connecticut NGSS Assessment was piloted in 2016–2017, field-tested in 2017–2018, and operationally administered for the first time in 2018–2019. The Connecticut NGSS Assessment measured the science knowledge and skills of Connecticut students in grades 5, 8, and 11 as an online adaptive assessment that made use of several technology-enhanced item types. The content measured the three-dimensional science standards based on the NGSS adopted by Connecticut in 2015.

Additional details on the implementation of the assessments can be found in Volume 1, Annual Technical Report.

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

- The item specifications, which provided detailed guidance for item writers and reviewers to ensure that science items were aligned to the performance expectations (PEs) they were intended to measure,
- The item development procedures employed for the Connecticut NGSS Assessment, which were consistent with industry standards,
- The development and maintenance of the Shared Science Assessment Item Bank, in which test items cover the range of measured PEs, grade-level difficulties, and levels of cognitive engagement by using both item clusters and stand-alone items, and
- The Test Design Summary/Blueprint, which stipulated the range of operational items from each item type and content category required for each test administration (this document was implemented using the item-selection algorithm for science).

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

For the Connecticut NGSS Assessment, items are drawn from the Shared Science Assessment Item Bank that consists of ICCR items, items owned by Connecticut, and items owned by several other states that share a Memorandum of Understanding (MOU) to share content, leadership, and new ideas and methods. Specifically, all items developed under the MOU underwent the same

development process. For the remainder of this volume, the term *item bank* will refer to all items developed under the MOU unless stated otherwise.

1.1 CLAIM STRUCTURE

The goals, uses, and claims that the Shared Science Assessment Item Bank and subsequent tests were designed to support were identified in a series of collaborative meetings held over August 22–23, 2016. The overarching goal of these meetings was to support the development of statewide summative assessments using science content that measures the three-dimensional science standards based on *A Framework for K–12 Science Education* (National Research Council, 2012).

To this end, CAI invited content and assessment leaders from 10 states, as well as four nationally recognized experts, Aneesha Badrinarayan, Rodger Bybee, Peter McLaren, and Brett Moulding, who helped author the NGSS. Two nationally recognized psychometricians, Laurie Wise, Ph.D. and Tom Hirsch, Ph.D., also participated.

CAI staff and participating states collaborated to develop items and test specifications that would measure the three-dimensional science standards. The item specifications were generally accompanied by sample item clusters that met those specifications. At the time, some standards did not have sample item clusters available. All specifications and sample item clusters were reviewed by state content experts and committees of educators in at least one of the states.

1.2 UNDERLYING PRINCIPLES GUIDING DEVELOPMENT

The Shared Science Assessment Item Bank was established using a highly structured, evidence-centered design. The process began with detailed item specifications. The specifications, discussed in Section 2.2, Item Specifications, described the interaction types that could be used, listed guidelines for targeting the appropriate cognitive engagement, offered suggestions for controlling item difficulty, and provided sample items.

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

Item development supported the goal of high-quality item clusters and stand-alone items through rigorous development processes managed and tracked by a content development platform. This platform ensures that every item flows through the correct sequence of reviews and captures every comment and change to the item.

CAI sought to ensure that the items measured the PEs in a fair and meaningful way by engaging educators, state officials, content experts, and fairness, bias, and sensitivity experts at each step of the process. Educators evaluated the alignment of the items to the PEs and offered guidance and suggestions for improvement. These educators participated in the review of items for fairness and

sensitivity. Following item field-testing, educators engaged in rubric validation, a process that refines rule-based rubrics upon review of student responses.

Combined, these principles and the processes that support them have been incorporated into an item bank that measures the PEs with fidelity and does so in a way that minimizes construct-irrelevant variance and barriers to access. The details of these processes are described in this volume of the annual technical report.

1.3 ORGANIZATION OF THIS VOLUME

This volume is organized into the following three sections:

1. An overview of the science item development process that supports the validity of the claims that science tests are designed to support
2. An overview of the Shared Science Assessment Item Bank, the types of assessments the item bank is designed to support, and methods for refreshing the item bank
3. A description of the test construction process for the Connecticut NGSS Assessment, including the blueprint, the test design, an evaluation of simulated test sessions, the operational blueprint match results, and the item exposure rates

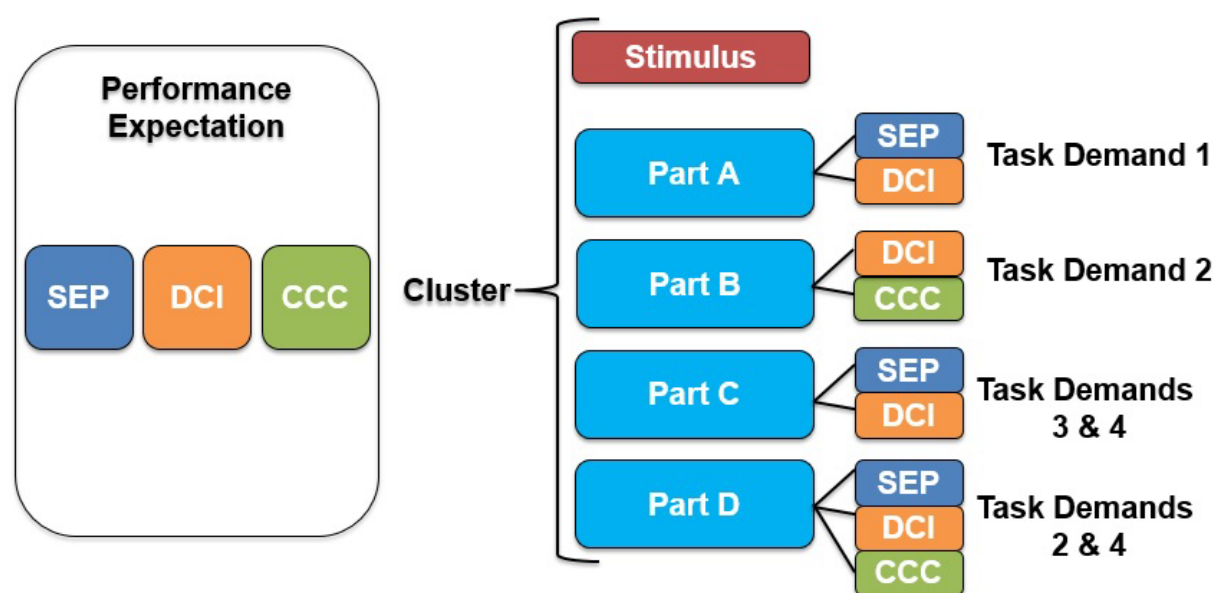
2. ITEM DEVELOPMENT PROCESS THAT SUPPORTS VALIDITY OF CLAIMS

2.1 OVERVIEW

Cambium Assessment, Inc. (CAI) developed the Shared Science Assessment Item Bank in collaboration with the states that were part of the Memorandum of Understanding (MOU) using a rigorous, structured process that engaged stakeholders at critical junctures.

A performance expectation is a point in a three-dimensional space formed by three dimensions of science learning: crosscutting concepts (CCCs), science and engineering practices (SEPs), and disciplinary core ideas (DCIs). That is, a performance expectation (PE) is characterized by a specific CCC, SEP, and DCI. When the MOU states first convened, many sessions were spent discussing how to assess these new three-dimensional standards. These group sessions are where the idea of an item cluster was conceived. An *item cluster* consists of a stimulus (scientific phenomenon) associated with multiple parts. Each of these parts contain questions that allow the student to explore the phenomenon. Each of the parts assess at least two dimensions, and the entire item cluster assesses a student on all three dimensions for a specific PE. Figure 1 is a visual representation of the structure of a three-dimensional cluster.

Figure 1. Structure of Three-Dimensional Item Clusters



Each part of an item cluster contains questions that require the student to interact with the item cluster. There are many different interactions that can be included in an item cluster. Section 3.1, Current Composition of the Shared Science Assessment Item Bank, describes and lists all of the different interactions available. The interactions used in an item cluster are chosen intentionally to best assess different aspects of the three-dimensional construct.

Figure 2 provides an example of an item cluster that has a phenomenon, five parts, and eight interactions; each part of an item cluster assesses multiple dimensions.

Figure 2. Example of an NGSS Item Cluster

A student rings a doorbell. When the person inside the house is on the main floor, he can easily hear the doorbell. When he is upstairs, though, he cannot so easily hear the doorbell.

Figure 1 shows the circuit of a simple doorbell when it is on (pressed) and off (not pressed).

Figure 1. Simple Doorbell Circuit

Table 1 shows the types of doorbell speakers available and their cost, in dollars (\$).

Table 1. Types of Speakers and Cost

Speaker	Cost (\$)
Bell	11
Buzzer	17
Chimes	25

Table 2 shows the types of batteries available based on their voltage (V), the amount of power each produces, and their cost.

Table 2. Types of Batteries and Cost

Battery (V)	Amount of Power	Cost (\$)
12	A lot	27
9	Average	3
1.5	A little	1

Table 3 shows the types of switches and their cost.

Table 3. Types of Switches and Cost

Switches	Cost (\$)
Rectangular	4
Circular	5
Lighted	11

Your Task

In the questions that follow, you will design a main-floor doorbell that can be heard from upstairs in a house.

Part A

Click on each blank box and select a phrase to describe what is happening to the energy at each part of the circuit when the doorbell is turned on.

Parts	Energy Pathway when Doorbell Is on
Battery	Energy is stored.
Wires	Energy is transferred.
Speaker	Electrical energy is converted to sound energy.

Part B

Use the simulation to select the materials necessary to conduct fair experiments and create a doorbell that can be heard from upstairs and costs less than \$40. The student can only hear a doorbell from upstairs if it is loud or very loud.

- Select the speaker, battery, and switch to determine the overall cost and loudness of the doorbell.
- Then click Run Trial.
- The cost of wire has already been included in the total cost.
- You must complete **two** trials.
- You may run up to **five** trials.
- Click the trash can icon if you want to delete a trial and generate new data.

Trial	Speaker	Battery (V)	Switch	Loudness	Cost (\$)
1	Bell	9.0	Rectangular	Loud	18
2	Bell	12.0	Rectangular	Very Loud	42
3	Bell	1.5	Rectangular	No Sound	16
4	Chimes	9.0	Lighted	Quiet	39
5	Bell	9.0	Lighted	Loud	25

Part C

Select **all** of the trials that meet the criteria for being heard upstairs and cost less than \$40.

☒ Trial 1

☐ Trial 2

☐ Trial 3

☐ Trial 4

☒ Trial 5

☐ None

Part D

Click on the blank boxes and select words or phrases to predict what will happen to the loudness of the doorbell when the battery power increases.

The loudness of the doorbell will because

Part E

Select **two** trials that support the relationship between the loudness of the doorbell and the power of the battery.

☐ Trial 1

☒ Trial 2

☒ Trial 3

☐ Trial 4

☐ Trial 5

☐ Cannot be determined

This item cluster is aligned to the NGSS PE of 4-PS3-4: Apply scientific ideas to design, test, and refine a device that converts energy from one form to another. The PE uses the following three elements of the three-dimensional standards: (1) Constructing Explanations and Designing Solutions (i.e., SEP), (2) Conservation of Energy and Energy Transfer (i.e., DCI), and (3) Energy and Matter (i.e., CCC).

Part A requires students to demonstrate their knowledge of how energy is stored, transferred, or used within the system. In this item cluster, they must know how a battery, wires, and a speaker work within the circuit. This aligns with the DCI and the CCC.

Part B requires students to design and test designs that use electricity to produce a sound. This aligns with the DCI (how changes in current influence the production of sound) and the SEP (designing and testing solutions to a design problem).

Part C requires students to compare their designs with some criteria and constraints. This aligns with the SEP (designing and testing solutions) and the CCC (energy can be transferred in various ways and between objects). The answer for Part C is directly determined by how the student completes Part B. If all of the trials the student runs in Part B meet the given criteria, then all of those must be selected to be considered as correct in Part C. Therefore, there are multiple different ways to get this item correct.

Part D requires students to make a prediction from the evidence that they generated in Part B. This part is aligned to all three dimensions. The student has used their designs and information (representing SEP) from Part B to show how energy is transferred between objects (representing CCC) and specifically how increasing the current changes the volume (representing DCI).

Like Part C, Part E is dependent on Part B. Students are determining which trials support the prediction that they made in Part D. This part, combined with Part D and Part B, addresses all three dimensions of the PE.

The next big challenge for the MOU states was to properly score these item clusters so that all evidence of understanding the PEs and three dimensions could be collected. It was determined that scoring assertions would be the best way to capture and score student responses on item clusters. Scoring assertions were evidence statements that related specific features from the student response to skills and knowledge that were tested (of which they provide evidence). The use of these assertions in scoring created a direct linkage between what the student does and the inferences about the skills and knowledge that the student's response supports. This approach provided a physical embodiment of evidence-centered design, Mislevy and Haertel's well-regarded approach to cognitive measurement (Mislevy & Haertel, 2006). This also provided a structure for ensuring and reviewing alignment during test development and a clear explanation of what was measured, how it was measured, and why it was measured when tests were scored and reported.

By inspecting the student response for every meaningful piece of student input, more information about student skills and knowledge can be harvested than in a single interaction. In fact, evidence for some scoring assertions may derive from two or more interactions within an item cluster. This may happen if one interaction is dependent on another interaction, allowing for multiple solution








paths. This is one of the primary reasons that scoring assertions within item clusters can show deeper cognitive understanding and higher-order thinking that is required of the three-dimensional science standards.

Each of the parts in an item cluster likely has one or more scoring assertions where student skills and knowledge are being collected. The scoring mechanism has the capability to focus on one interaction, one part, or to focus across multiple interactions and parts as determined by the item writers, subject-matter expert (SME) reviewers and performance expectations. All permutations and combinations of measurable actions can be captured with scoring assertions.

The example item cluster from this section has seven assertions. Each scoring assertion is described in detail in Figure 3.

Figure 3. Example of NGSS Scoring Assertions

Your response earned **7** points of a possible **7**

Score Rationale	
When asked to describe what is happening to the energy for the battery when the doorbell is turned on, the student selected "energy is stored" or "energy is transferred." This provides some evidence of an ability to complete a causal chain explaining how energy can be transferred via electric current to produce light, sound, heat, and /or motion.	
When asked to describe what is happening to the energy of the wires when the doorbell is turned on, the student selected "energy is transferred." This provides some evidence of an ability to complete a causal chain explaining how energy can be transferred via electric current to produce light, sound, heat, and /or motion.	
When asked to describe what is happening to the energy of the speaker when the doorbell is turned on, the student selected "electrical energy is converted to sound energy." This provides some evidence of an ability to complete a causal chain explaining how energy can be transferred via electric current to produce light, sound, heat, and /or motion.	
The student ran at least two trials and ran at least one trial in which they selected components of a doorbell that produced "Loud" or "Very Loud" sound and that included components that cost less than \$40. This provides some evidence of an ability to select characteristics to be manipulated while gathering information to determine the loudest, cost-effective doorbell.	
When asked to select the trial that met the criteria for being heard upstairs and cost less than \$40, the student selected all trials from their simulation that produced "Loud" or "Very Loud" sound and cost less than \$40. This provides some evidence of an ability to use given information to design and test a device that converts energy from one form to another.	
When asked to predict what will happen to the sound of the doorbell if the battery power increases, the student selected "The loudness of the doorbell will increase because more energy is stored in the battery." This provides some evidence of an ability to use an explanation to predict how the sound of an object changes, given a change in the conversion of stored energy.	
When asked to select the trials that support the relationship between the loudness of the doorbell and the power of the battery, the student selected two trials from the simulation in which the loudness was higher for the trial with a battery with more power. This provides some evidence of an ability to use evidence to support an inference.	

Assertion texts like the one shown in Figure 3 are written for every assertion in every item. They describe the correct response and what evidence should be provided by the student's response.

In the example item cluster, Part A has three assertions. Each one "provides some evidence of an ability to complete a causal chain explaining how energy can be transferred via electric current to produce light, sound, heat, and/or motion." The student must know something about electrical energy (DCI) and how it is transferred or used (DCI and CCC) to correctly respond. One assertion corresponds to each row in the table (i.e., one for Battery, one for Wires, and one for Speaker).

Part B has two assertions. The first “provides some evidence of an ability to select characteristics to be manipulated while gathering information to determine the loudest, most cost-effective doorbell.” The second assertion “provides some evidence of an ability to use given information to design and test a device that converts energy from one form to another.” The student must use their knowledge of how electrical energy is used and transferred (DCI) and how to design and test a design of a device using electricity (SEP) to correctly interact with Part B.

Part C has one assertion, as the student’s selections are not independent of each other. The assertion “provides some evidence of an ability to use given information to design and test a device that converts energy from one form to another.” The student must be able use generated evidence to support a design decision (SEP) about the transfer of energy (CCC). This assertion is pulling responses from both Parts B and C. This is precisely how item clusters and assertions can assess multiple dimensions and higher levels of complexity, as students are running their own experiments and analyzing the outcomes, no matter what those outcomes are.

Part D has one assertion. The assertion “provides some evidence of an ability to use an explanation to predict how the sounds of an object changes, given a change in the conversion of stored energy.” This shows how the student must use elements from all three dimensions to respond correctly to this assertion. The student uses data from their generated designs and makes a prediction using that data to support their knowledge of energy and energy transformations.

Part E also has one assertion. The assertion “provides some evidence of an ability to use evidence to support an inference.” In this case, it is an inference about the relationship between the available battery power and the loudness of the bell. Again, this scoring assertion is pulling information from three different parts (Parts B, D, and E).

While each part of the item, each interaction within the item, or each assertion may not be three-dimensional, the item cluster as a whole represents all three dimensions. It also provides an organized flow of cognition from scaffolding (Part A), through the engineering process (Parts B and C), to a conclusion and evidentiary support of the conclusion (Parts D and E).

The assertion text explains how a student responded to a given task and what that task shows evidence of. This allows us to ensure that items allow each student an opportunity to show what they know and what their knowledge, skills, and abilities show about their understanding of science and engineering.

Once the item cluster, along with interactions and scoring assertions, came to fruition, CAI and the group of states were able to begin item and test development in earnest.

The item development process was managed by CAI’s Item Tracking System (ITS), which is an auditable content-development tool that enforces rigorous workflow and captures each item change and comment. Reviewers, including internal CAI reviewers or stakeholders in committee meetings, can review items in ITS as they will appear to the student, with all accessibility features and tools.

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

- selection and training of item writers;
- writing and internal review of items;

- review by state personnel and stakeholder committees;
- markup for translation and accessibility features;
- field-testing; and
- post-field-test reviews.

Each step has a role in ensuring that the items can support the claims on which they will be based. Table 1 describes how each step contributes to these goals and describes each step in the process in more detail.

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

Developmental Steps	Support Alignment to the Performance Expectations	Reduces Construct-Irrelevant Variance Through Universal Design	Expands Access Through Linguistic and Other Supports
Item specifications	Specifies item interactions, content limits, and guidelines for meeting task demands and levels of cognitive engagement requirements and adjusting difficulty.	Avoids the use of any item interactions with accessibility constraints and provides language guidelines. Allows for multiple response modes to accommodate different styles.	
Selection and training of item writers (some states select their own item writers who are teachers or administrators in the state)	Ensures that item writers have the background to understand the PEs and item specifications. Teaches item writers how to select item interactions for measurement and accessibility.	Training in language accessibility, and bias and sensitivity helps item writers avoid unnecessary barriers.	
Writing and internal review of items	Checks content alignment and evaluates and improves overall quality.	Eliminates editorial issues and flags and removes bias and accessibility issues.	
Markup that prepares items for translation and accessibility features		Adds universal features, such as text-to-speech (TTS) for science, that reduce barriers.	Adds TTS, braille, American Sign Language (ASL), translations, and glossaries.
Review by state personnel and stakeholder committees	Checks content and cognitive complexity alignment; evaluates and improves overall quality.	Flags sensitivity issues.	
Field testing	Provides statistical checks on quality and flags issues.	Flags items that appear to function differently for subsequent review to identify issues.	May reveal usability or implementation issues with markup.

Developmental Steps	Support Alignment to the Performance Expectations	Reduces Construct-Irrelevant Variance Through Universal Design	Expands Access Through Linguistic and Other Supports
Post-field-test reviews	Provides final, more focused checks on flagged items. Rubric validation ensures that scoring reflects PEs.	Provides final, focused review on items flagged for differential item functioning (DIF).	

2.2 ITEM SPECIFICATIONS

CAI collaborated with a group of states and one U.S. territory, psychometricians, and science experts, including the authors of the NGSS, to develop powerful innovative solutions to the challenges of measuring three-dimensional science standards based on the National Research Council’s *A Framework for K–12 Science Education* published in 2012. Participating states included Connecticut, Hawaii, Idaho, Montana, Oregon, Rhode Island, Utah, Vermont, West Virginia, and Wyoming. New Hampshire, North Dakota, South Dakota, and U.S. Virgin Islands participated in some activities. This collaboration yielded item specifications for PEs, sample item clusters for some specifications, and hundreds of science item clusters and stand-alone items in various stages of development. Under this collaboration, these states and the U.S. Virgin Islands jointly developed item specifications using the guidelines proposed by WestEd in conjunction with the Council of Chief State School Officers (CCSSO), state and territory members, and content experts (CCSSO, 2015).

Item specifications are documents designed to guide item writers as they craft test questions and stakeholders as they review those items. These specifications are intended to serve as a roadmap for writers to facilitate the creation of items that are properly aligned to the three dimensions comprising each science standard, and that together form coherent item clusters and stand-alone items. Table 2 provides a sample of the item specifications developed by content experts for a middle school Life Sciences PE. Item specifications in science include the following:

- **Performance Expectation.** The PE provides the unique identifier and the text for the PE.
- **Dimensions.** This identifies the CCCs, SEPs, and DCIs that the PE assesses.
- **Clarifications and Content Limits.** This section delineates the specific content that the PE measures and the parameters in which items must be developed to assess the PE accurately, including the lower and upper complexity limits of items. Specifically, content limits refine the intent of the PE and provide limits of what may be asked of test takers. For example, content limits may identify the specific formulae that students are expected to know or not know.
- **Science Vocabulary.** This section identifies the relevant technical words that students are expected to know, and related words that they are explicitly not expected to know.

These categories should not be considered exhaustive, as the boundaries of relevance are ambiguous and the list is limited by the imagination of the writers.

- **Content/Phenomena.** This section provides examples of the types of phenomena that would support the effective items related to the PE in question. In general, these are guideposts, and item writers seek comparable phenomena, rather than drawing on those within the documents.
- **Task Demands.** In this section, the PEs and associated evidence statements are broken down into specific task demands aligned to each PE. Task demands denote the specific ways in which students will provide evidence of their understanding of the concept or skill. Specifically, the task demands identify the types of interactions and activities that item writers should employ. Each item should be clearly linked to one or more of the task demands, and the verbs guide the types of interactions writers might employ to elicit the student response.

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

Performance Expectation	MS-LS1-1^a Conduct an investigation to provide evidence that living things are made of cells; either one cell or many different numbers and types of cells.		
Dimensions	Planning and Carrying Out Investigations <ul style="list-style-type: none"> Conduct an investigation to produce data to serve as the basis for evidence that meets the goals of an investigation. 	LS1.A: Structure and Function <ul style="list-style-type: none"> All living things are made up of cells, which is the smallest unit that can be said to be alive. An organism may consist of one single cell (unicellular) or many different numbers and types of cells (multicellular). 	Scale, Proportion, and Quantity <ul style="list-style-type: none"> Phenomena that can be observed at one scale may not be observable at another scale.
Clarifications and Content Limits	Clarification Statements <ul style="list-style-type: none"> Emphasis is placed on developing evidence that living things are made of cells, distinguishing between living and non-living things, and understanding that living things may be made of one cell or many varying cells. Content Limits <ul style="list-style-type: none"> <u>Students do not need to know the following:</u> <ul style="list-style-type: none"> The structures or functions of specific organelles or different proteins Systems of specialized cells The mechanisms by which cells are alive Specifics of DNA and proteins or of cell growth and division Endosymbiotic theory Histological procedures 		
Science Vocabulary Students Are Expected to Know	Multicellular, unicellular, cell, tissue, organ, system, organism hierarchy, bacteria, colony, yeast, prokaryote, eukaryote, magnify, microscope, DNA, nucleus, cell wall, cell membrane, algae, chloroplast(s), chromosome, cork		

Science Vocabulary Students Are Not Expected to Know	Differentiation, mitosis, meiosis, genetics, cellular respiration, energy transfer, RNA, protozoa, amoeba, histology, protists, archaea, nucleoid, plasmid, diatoms, cyanobacteria
Phenomena	
Context/ Phenomena	<p>Some example phenomena for MS-LS1-1 include the following:</p> <ul style="list-style-type: none"> • Plant leaves and roots have tiny, box-like structures that can be seen under a microscope. • Small creatures can be seen swimming in samples of pond water viewed through a microscope. • Different parts of a frog's body (e.g., muscles, skin, tongue) are observed under a microscope, and are seen to be composed of cells. • One-celled organisms (e.g., bacteria, protists) perform the eight necessary functions of life, but nothing smaller has been seen to do this. • Swabs from the human cheek are observed under a microscope. Small cells can be seen.
This Performance Expectation and Associated Evidence Statements Support the Following Task Demands	
Task Demands	
1. Identify from a list the materials/tools, including distractors, needed for an investigation to find the smallest unit of life (cell).	
2. Identify the outcome data that should be collected in an investigation of the smallest unit of living things.	
3. Evaluate the sufficiency and limitations of data collected to explain that the smallest unit of living things is the cell.	
4. Make and/or record observations about whether the sample contains cells. ^b	
5. Interpret and/or communicate data from the investigation to determine if a specimen is alive.	
6. Construct a statement to describe the overall trend suggested by the observed data.	

^aMS-LS1-1 is the PE code for Middle School Life Sciences 1-1.

^bDenotes task demands deemed appropriate for use in stand-alone item development

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

The assertions provide evidence that the student is completing specific aspects of the PEs. Each assertion is reviewed during development and at all state and committee reviews to ensure that students are doing what is stated in the assertion. The assertion links the student response to an interaction to the performance expectation.

2.3 SELECTION AND TRAINING OF ITEM WRITERS

All item writers developing science items at CAI have at least a bachelor's degree, and many bring teaching experience. All item writers are trained in

- the principles of universal design;
- the appropriate use of item interactions; and
- the science item specifications.

Key materials are shown in Appendix 2-A, Item Writer Training Materials, Appendix 2-B, Item Specifications Guide Grade 3 through High School, and Appendix 2-C, Style Guide for Science Items. These include

- CAI’s language accessibility, bias, and sensitivity (LABS) guidelines;
- a training (presented using Microsoft PowerPoint) for the appropriate use of item interactions;
- item specification for science for grades 3 through high school; and
- style guide for science items.

2.4 INTERNAL REVIEW

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

ICCR and MOU science items undergo a rigorous, multi-level internal review process before they are sent to external review. Staff members are trained to review items for both content and accessibility throughout the process. A sample item review checklist that CAI test developers use is included in Appendix 2-D, Item Review Checklist. The ICCR and MOU science internal review cycle includes the following phases:

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

2.4.1 Preliminary Review

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

- The item aligns with the PE, including the listed SEP, DCI, and CCC. The item matches the item specification for the skills and knowledge being assessed. The item

specification contains clarifying statements, content limits, and task demands, as well as knowledge, skills, and abilities that the PE is intended to assess.

- The item is based on a quality scientific phenomenon (i.e., it assesses something in a reasonable way, and it is a discrete observation that grounds a scenario, which allows for the assessment of something worthwhile in a meaningful way). A quality phenomenon is one that is natural, observable (even with instrumentation), and focused on a specific event, not a general category of similar events (e.g., the effects of Hurricane Katrina, not hurricanes in general).
- The item aligns appropriately with the task demands. Task demands are statements about what a student is expected to do with a phenomenon.
- The vocabulary used in the item is appropriate for the grade and subject matter. Most non-technical language is two grade levels below the testing grade to ensure that language is not a construct-irrelevant issue.
- The item considers language accessibility, bias, and sensitivity.
- The content is accurate and straightforward.
- The graphic and stimulus materials are necessary to answer the question. The phenomenon is described in the stimulus. Graphics are necessary and contain only the relevant information.
- The item follows the approved style guide.
- The stimulus is clear, concise, and succinct (i.e., it contains enough information to convey what is being asked, it is stated positively, and it does not rely on negatives—such as *no*, *not*, *none*, or *never*—unless necessary).

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

- as succinct and short as possible (e.g., without repeating text);
- parallel in structure, grammar, length, and content;
- sufficiently distinct from one another;
- all plausible (but with only one correct option); a plausible distractor is one that is related to the item, but contains a misconception, a logical error, or a pattern of thinking that a student might have, but is incorrect; and
- free of obvious or subtle cuing.

2.4.2 Scoring Entry and Review

During Scoring Entry, the item writer inputs the machine scoring for review by the team lead or senior staff before the Content-Review-One level. This step is separate from Preliminary Review to allow senior staff to suggest changes to the interaction at Preliminary Review without requiring

the writer to overhaul the scoring they already created, ensuring that the scoring is entered once, streamlining the process. This step also allows senior staff to ensure that the scoring suggested by the writer at Preliminary Review is appropriate. At this level, scoring is analyzed to ensure the following criteria:

- The scoring works as intended (i.e., the student gets a point for ALL correct responses and no points for ALL incorrect responses).
- The student receives a point for every unique piece of information they reveal about their understanding through their responses.
- Dependent scoring between and within interactions is captured.
- The way in which the scoring is set up is unambiguous and matches the questions asked (i.e., if we ask students to round a number to a certain decimal place, we score accordingly).

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

Senior staff are recruited based on experience and time in the assessment field. Senior staff are the reviewers of the intent of scoring because of their experience and knowledge of assessment, the expectations of the clients, and their understanding of student responses.

2.4.3 Content Review One

Content Review One is conducted by a senior content specialist who was not part of the Preliminary Review. This reviewer carefully examines each item based on the same criteria identified for Preliminary Review. They also ensure that the revisions made during the Preliminary Review did not introduce errors or content inaccuracies. This reviewer approaches the item by combining their expertise in test development while engaging from the perspective of potential clients.

2.4.4 Edit Review

During Edit Review, editors have the following four primary tasks:

1. Editors perform basic line editing for correct spelling, punctuation, grammar, and mathematical and scientific notation, ensuring consistency of style across the items.
2. Editors ensure that all items are accurate in content. Editors compare reading passages against the original publications to ensure that all information is internally consistent across stimulus materials and items, including names, facts, or cited lines of text that appear in the item. They ensure that the keys and all information in the item are correct. Keys are the correct answers to interactions. Information refers to the phenomena and the science content. For items with mathematical tasks, editors perform all calculations to ensure accuracy.
3. Editors review all material for fairness and language accessibility issues.

4. Editors confirm that items reflect the accepted guidelines for good item construction. They examine all items for language that is simple, direct, and free of ambiguity with minimal verbal difficulty. Editors confirm that a problem or task and its stem are clearly defined and concisely worded with no unnecessary information. For multiple-choice interactions, editors check that options are parallel in structure and fit logically and grammatically with the stem. They also ensure that the key (i.e., correct answer) answers the question posed accurately and correctly, is not inappropriately obvious, and is the only correct answer to an item among the distractors. For constructed-response interactions, editors review the rubrics for appropriate style and grammar.

2.4.5 Senior Review

By the time a science item arrives at Senior Review, both content reviewers and editors have thoroughly vetted it. Senior reviewers (in particular, senior content specialists) look at the item's entire review history, ensuring that all the issues identified in that item have been adequately addressed. Senior reviewers verify the overall content of each item, confirming its accuracy, alignment to the PE, and consistency with expectations for the highest quality. They check whether the scoring is working as intended and scoring assertions adequately address the evidence the student provides with each type of response.

Some examples of questions from the internal Review Checklist are listed below. These are the questions that reviewers ask of the item to ensure that it is three dimensional and properly aligned to the PE. A similar checklist is used at earlier stages.

- Is the phenomenon based on a specific real-world scenario and focused enough to get the student to investigate what the PE intends for them to investigate (i.e., the students' application of the Practice in the context of the DCI and CCC as intended by the PE is sufficient to make sense of the phenomena)?
- What information should the student already have before starting the cluster (DCI knowledge)?
- Cluster Task Statement: Does it align to the focus and intent of the PE?
- Does the interaction require the student to demonstrate the science practice and/or content that the PE is assessing them on?
- Do the interactions align with the task demands?

2.5 REVIEW BY STATE PERSONNEL AND STAKEHOLDER COMMITTEES

All science items undergo an exhaustive external review process. Items in the Shared Science Assessment Item Bank were reviewed by content experts in one or several states or territory and reviewed and approved by multiple stakeholder committees that evaluated them for both content and bias and sensitivity.

2.5.1 State Review

After items have been developed for a state participating in the MOU, content experts from the state that owns the item review any eligible items before committee review. At this stage in the review process, clients can request edits, such as wording edits, scoring edits, alignment changes, or task demand updates. A CAI science content expert reviews all client-requested edits concerning the science item specifications, other clients' requests, and existing items in the bank to determine whether the requested edits will be made. At this stage, clients have the option to present these items to the committee (based on the edits made) or withhold them from committee review.

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

2.5.2 Content Advisory Committee Reviews

During the Content Advisory Committee (CAC) reviews, items are reviewed for content accuracy, grade-level appropriateness, and alignment to the PE. CAC members are typically grade-level and subject-matter experts. During this review, educators also ensure that the scoring assertions clearly identify what is being scored as correct and give credit where they should (refer to Section 2.7.1, Rubric Validation). Before the CAC review begins, CAI provides a presentation on the three-dimensional science standards, the item development process, the CAI systems that will be used in the review, and how to review the items for content. Appendix 2-E, Content Advisory Committee Review Training Slides, provides the slides used during the CAC review training.

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

In 2024, the MOU states were all involved in a single CAC event where participants from multiple states reviewed items. The items were edited and then returned to the owning state for final approval.

A summary of the 2023–2024 committee meetings is presented in Table 3, with additional details about the participants in Appendix 2-F, Content Advisory Committee Participant Details. This appendix also contains detailed information about the participants of CAC meetings of previous years.

Table 3. Summary of the 2023-2024 Content Advisory Committee Meetings

State/ Item Bank	Meeting	Number of Committee Members	Number of Items Reviewed
Connecticut			
Hawaii			
ICCR			
Idaho			
Montana			
Oregon			
Rhode Island			
Utah			
West Virginia			
Wyoming			

^aItems were reviewed in a combined Content Advisory Committee Meeting that included ICCR and MOU state-owned items. Items reviewed in the combined meetings are displayed by their respective state or bank of ownership.

2.5.3 Language Accessibility, Bias, and Sensitivity Committee Reviews

During bias and sensitivity reviews, stakeholders review items to check for issues that might unfairly impact students based on the students' backgrounds. For example, some states include representatives from student populations such as Special Education, low vision, and the hearing impaired. Further, diverse members of this committee represent students of various ethnic and economic backgrounds to ensure that all items are free of bias and sensitivity concerns. States try to ensure that all demographics are represented when providing committee members. For example, if a state has a Native American population, they will try to ensure that the Bias and Sensitivity Committee has Native American representation on the committee. Before the bias and sensitivity review begins, CAI provides a presentation on the three-dimensional science standards, the item development process, the CAI systems that will be used in the review, and how to review the items for fairness. Appendix 2-G, Fairness Committee Review Training Slides, provides the slides used during the bias and sensitivity review training.

During 2020–2022, due to the COVID-19 pandemic, CAI reviewed items that contained references to virus, vaccine, bacteria, disease, infection, and related words and phrases. CAI content experts reviewed 65 items and rejected one item for sensitivity concerns.

In 2024, the MOU states were all involved in a single review process where participants from multiple states would review items. The items were edited and then returned to the owning state for final approval.

A summary of the 2023-2024 committee meetings is presented in Table 4, with additional details about the participants in Appendix 2-H, Fairness Committee Participant Details. This

appendix also contains detailed information about the participants of Fairness Committee meetings of previous years.

Table 4. Summary of the 2023-2024 Fairness Committee Meetings

State/ Item Bank	Meeting	Number of Committee Members	Number of Items Reviewed	Number of Items Rejected
Connecticut				
Hawaii				
ICCR				
Idaho				
Montana				
Oregon				
Rhode Island				
Utah				
West Virginia				
Wyoming				

^aItems were reviewed in a combined Fairness Committee Meeting that included ICCR and MOU state-owned items. Items reviewed in the combined meetings are displayed by their respective state or bank of ownership.

2.5.4 Markup for Translation and Accessibility Features

After all approved state/territory- and committee-recommended edits have been applied, the items are considered “locked” and ready for a portion of the accessibility tagging. Text-to-speech (TTS) tagging is applied prior to field-testing, while Spanish translations and braille are applied post-field-testing. Accessibility markup is embedded into each item as part of the item development process rather than as a *post-hoc* process applied to completed tests.

Accessibility markup, whether for translations or TTS, follows similar processes. One trained expert enters the markup, and then a second expert reviews the work and recommends changes, if necessary. If there is disagreement, a third expert is engaged to resolve the issue.

Currently, science items are tagged with TTS. Spanish translations, including Spanish TTS and braille, are available for a subset of items.

2.6 FIELD-TESTING

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

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

In 2021, a third wave of field-test items was administered in 12 states: Connecticut, Hawaii, Idaho, Montana, New Hampshire, North Dakota, Rhode Island, South Dakota, Utah, Vermont, West Virginia, and Wyoming. An independent field test, in which students were administered a full set of items, was conducted for Idaho and Montana. Unscored field-test items were added as a separate segment to the operational (scored) legacy science test for Wyoming. In the remaining nine states, field-test items were administered as unscored items embedded within the operational items.

In 2022, a fourth wave of field-test items was administered in 13 states and one U.S. territory: Connecticut, Hawaii, Idaho, Montana, New Hampshire, North Dakota, Oregon, Rhode Island, South Dakota, Utah, Vermont, West Virginia, Wyoming, and U.S. Virgin Islands. In all 13 states and the territory, field-test items were administered as unscored items embedded within the operational items.

In 2023, items were field-tested in 12 states and one U.S. territory: Connecticut, Hawaii, Idaho, Montana, New Hampshire, North Dakota, Oregon, Rhode Island, South Dakota, Utah, West Virginia, Wyoming, and U.S. Virgin Islands. Field-test items were administered as unscored items embedded within the operational items.

In 2024, items were field-tested in 14 states and one U.S. territory: Arkansas, Connecticut, Hawaii, Idaho, Indiana, Montana, New Hampshire, North Dakota, Oregon, Rhode Island, South Dakota, Utah, West Virginia, Wyoming, and U.S. Virgin Islands. Field-test items were administered as unscored items embedded within the operational items. CAI's field-test process is described in detail in Section 3.2, Field-Testing, in Volume 1 of this technical report.

2.7 POST-FIELD-TEST REVIEW

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

2.7.1 Rubric Validation

The validation process for the field-test items begins with rubric validation to verify and make any necessary revisions to the scoring rubrics. The rubric validation process occurs in two phases. During the first phase, CAI content experts work with the analysis team to prepare for the rubric-validation meetings. The CAI content experts use the Rubric Evaluation and Verification for Items Scored Electronically (REVISE) system to generate student responses that are scientifically sampled to overrepresent responses most likely to have been mis-scored. Specifically, the sample overrepresents low-scored responses from otherwise high-scoring students and

high-scored responses from otherwise low-scoring students. This process allows CAI to identify any potential scoring concerns before the rubric validation meeting, such as unanticipated (but accurate) responses, equivalent responses that were not originally considered, and responses receiving credit but should not (based on the content and the item rubric). At this point, the rubrics may be adjusted, and responses rescored.

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

Figure 4 illustrates the features provided by the REVISE system.

Figure 4. Features of the REVISE Software

The screenshot displays the REVISE software interface, which is used for Rubric Evaluation and Verification for Items Scored Electronically. The interface includes a top navigation bar with tabs for Item List, Samples, Rubric, Summary, and Responses. The main content area is divided into several sections:

- Sample Details:** This section provides information about the sample, including the Sample Name (RV Sample), Sample Details, and Sample Create Date (5/25/2017 3:12:05 PM). It also includes a table of Rule Short Name, Rule Description, and Number of Responses.
- Responses:** This section displays a table of responses for the sample. The table has columns for Mark as Reviewed, Original Score, Previous Score, Current Score, Proposed Score, Response ID, and Sample Score. The table shows several rows of data, including scores and response IDs.
- Test Item:** This section displays the actual test item, which is a math problem about plane travel. The problem asks the user to create an equation that represents the relationship between time and distance. The student's response is shown as $570d = 1t$.
- Committee Comments:** This section allows the committee to record their comments and consensus score. It includes a text area for comments and a dropdown menu for the consensus score.

Annotations in the image highlight specific features:

- "Users can automatically draw samples according to a variety of sample designs. Revisions to the rubric can be checked against the original sample and independent samples." (points to the Sample Details section)
- "Responses in the sample are listed here." (points to the Responses table)
- "The committee records its comments and consensus score here." (points to the Committee Comments section)
- "Users can see the actual test item here." (points to the Test Item section)
- "Users can see the actual student response here." (points to the student response $570d = 1t$)

After the rubric validation meetings, CAI staff apply the approved revisions to the rubrics, and any items rejected as part of the process are rejected in ITS. During this process, ITS archives critical information regarding the scoring certification completed during the rubric validation process. This includes any rubric changes made during the scoring decision meetings and the sign-off completed by the senior content expert once the rubric has been changed, rescoring the entire sample, and verifying that the final rubric functioned as intended.

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

2.7.2 Data Review

Following rubric validation, all items are rescored, and classical item statistics are computed for the scoring assertions, including item difficulty and item discrimination statistics, testing time, and differential item functioning (DIF) statistics. The states and U.S. territory established standards for the statistics, and any items violating these standards are flagged for a second educator review. Even though the scoring assertions are the basic units of analysis to compute classical item statistics, the business rules to flag items for additional educator review are established at the item level because assertions cannot be reviewed in isolation. A common set of business rules is defined for all the states and territory participating in the field test. The classical item statistics are computed on the data of the students testing in the state or territory that owned the item. For ICCR items administered in spring 2024, the data from students testing in Arkansas, Connecticut, Idaho, Indiana, New Hampshire, North Dakota, Rhode Island, South Dakota, U.S. Virgin Islands, Utah, and West Virginia were combined (states that administered ICCR items and used either an independent field test or operational test).

Section 4 of Volume 1, *Field-Test Classical Analysis*, describes the statistical flags that designate items for data review. The flags are designed to highlight potential content weaknesses, miskeys, or possible bias issues. Committee members are taught to interpret these flags and are given guidelines for examining the items for content or fairness issues.

For each of the states participating in the MOU, flagged items owned by the state are reviewed by a data review committee. The composition of the data review committees generally includes content experts from the state’s department of education or state educators (in this case, the state educators were science teachers) and are supported by CAI content experts. ICCR field-test items are taken to committee members from several states or territory participating in the MOU. The outcomes are decided by CAI science content leadership while taking the committees’ recommendations into consideration.

At the start of each state-owned item data-review meeting, CAI staff leads participants in a training session to familiarize them with the item development process, the purpose of the data review committee and the data review process, and the meaning of the various flags. Committee members are taught to interpret the various flags and are given guidelines for examining the items for content or fairness issues. The training includes a group review of item cards, which detail specific item attributes (e.g., grade level and alignment to the science PEs, the content and rubric of the item, and various item statistics). A sample of the training materials used for these data review meetings is presented in Appendix 2-I, *Sample Data Review Training Materials*. Participants use an online environment via laptop computers to review the items and interact with them in a manner similar to that of students, and to view the statistics associated with each item.

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

Table 5 summarizes the data review committee meetings. Details, including the composition of each committee, are presented in Appendix 2-J, Data Review Committee Participant Details.

Table 5. Summary of Data Review Committee Meetings

State/ Item Bank	Meeting	Number of Committee Members	Item Type	Number of Items Reviewed	Number of Items Rejected
Connecticut	August 2018	29	Cluster	7	5
			Stand-Alone	11	6
	August 2019	29	Cluster	14	6
			Stand-Alone	39	11
	August 2021	19	Cluster	8	2
			Stand-Alone	43	10
	August 2022	15	Cluster	5	4
			Stand-Alone	14	2
	September 2023	12	Cluster	11	4
			Stand-Alone	32	17
Hawaii	August 2018	18	Cluster	7	1
			Stand-Alone	25	2
	August 2019	18	Cluster	17	5
			Stand-Alone	20	8
	August 2021 ^a	25	Cluster	6	0
			Stand-Alone	20	8
	August 2022 ^a	12	Cluster	11	2
			Stand-Alone	38	6
	August 2023 ^a	15	Cluster	3	2
			Stand-Alone	23	3
ICCR	July 2018	18	Cluster	33	2
			Stand-Alone	51	6
	August 2019 ^b	–	Cluster	0	1
			Stand-Alone	43	2
	August 2021 ^a	25	Cluster	11	2
			Stand-Alone	64	4
	August 2022 ^a	20	Cluster	12	1
			Stand-Alone	56	13
	August 2023 ^a	19	Cluster	12	1
			Stand-Alone	42	8
Idaho	August 2019	10	Cluster	4	3
			Stand-Alone	8	3
	August 2021 ^a	25	Cluster	26	1
			Stand-Alone	34	4
	August 2022 ^a	8	Cluster	3	0
			Stand-Alone	1	0
Montana	August 2023 ^a	17	Cluster	2	0
			Stand-Alone	3	0
	September 2021	4	Cluster	3	2
			Stand-Alone	14	2
	September 2022	5	Cluster	5	2
			Stand-Alone	12	1
	August 2023 ^a	11	Cluster	2	1

State/ Item Bank	Meeting	Number of Committee Members	Item Type	Number of Items Reviewed	Number of Items Rejected
Multi-State Science Assessment (Rhode Island and Vermont)	August 2018	–	Stand-Alone	10	2
			Cluster	2	0
			Stand-Alone	7	6
	August 2019	–	Cluster	2	1
			Stand-Alone	12	3
			Cluster	4	4
	August 2021	–	Stand-Alone	14	5
			Cluster	1	1
September 2022	–	Stand-Alone	10	6	
		Cluster	28	5	
Oregon	September 2018	11	Stand-Alone	16	1
			Cluster	1	1
August 2019	4	Stand-Alone	7	6	
		Cluster	11	2	
August 2022 ^a	8	Stand-Alone	20	6	
		Legacy Stand-Alone	9	4	
		Cluster	9	1	
		Stand-Alone	3	1	
		Legacy Stand-Alone	24	11	
August 2023 ^a	16	Cluster	9	1	
		Stand-Alone	3	1	
		Legacy Stand-Alone	24	11	
Rhode Island	September 2023	–	Cluster	7	2
			Stand-Alone	10	4
			Legacy Stand-Alone	15	0
South Dakota	September 2021	–	Legacy Stand-Alone	4	1
	September 2022	–	Legacy Stand-Alone	6	2
	September 2023	–	Legacy Stand-Alone	6	2
Utah	August 2018	16	Cluster	40	6
	September 2021	6	Cluster	11	3
	September 2022	13	Cluster	11	6
	September 2023	20	Cluster	6	0
West Virginia	July 2018	4	Cluster	3	1
			Stand-Alone	0	0
	September 2019	4	Cluster	1	1
			Stand-Alone	6	5
	August 2021 ^a	25	Cluster	1	1
			Stand-Alone	6	2
	August 2022 ^a	9	Cluster	4	2
			Stand-Alone	6	2
	August 2023 ^a	11	Cluster	2	1
			Stand-Alone	10	2
Wyoming	October 2018	12	Cluster	6	1
			Stand-Alone	10	5
	August 2019	10	Cluster	4	3
			Stand-Alone	12	2
	August 2021 ^a	25	Cluster	3	1
			Stand-Alone	13	3
	August 2022 ^a	12	Cluster	2	0
			Stand-Alone	17	3
	August 2023 ^a	17	Cluster	3	0
		Stand-Alone	5	1	

Note. MSSA, Rhode Island, and South Dakota-owned items were reviewed by Rhode Island Department of Education and Vermont Agency of Education science content experts, the Rhode Island Department of Education, and the South Dakota Department of Education, respectively.

^aCombined Item Data Review Meetings were conducted for multiple states in 2021, 2022, and 2023 (184 items were reviewed in the combined meeting format for Hawaii, Idaho, West Virginia, Wyoming, and ICCR items in 2021; 181 items were reviewed in the combined meeting format for Hawaii, Idaho, Oregon, West Virginia, Wyoming, and ICCR items in 2022, and 129 items were reviewed in the combined meeting format for Hawaii, Idaho, Montana, Oregon, West Virginia, Wyoming, and ICCR items in 2023). In 2021, 25 committee members took part in the combined Item Data Review Meetings; in 2022, 38 committee members participated in the combined Item Data Review Meetings, and in 2023, 41 committee members participated in the combined Item Data Review Meetings. Items reviewed in the combined meetings are displayed by their respective state or bank of ownership.

3. SHARED SCIENCE ASSESSMENT ITEM BANK SUMMARY

Tests based on *A Framework for K–12 Science Education* (National Research Council, 2012) adopt a three-dimensional conceptualization of science understanding, including crosscutting concepts (CCCs), science and engineering practices (SEPs), and disciplinary core ideas (DCIs). Accordingly, the new science assessments are composed mostly of item clusters representing a series of interrelated student interactions directed towards describing, explaining, and predicting scientific phenomena. Some stand-alone items are added to increase the coverage of the test without increasing the testing time or testing burden.

CAI has built the Shared Science Assessment Item Bank in partnership with multiple states and one U.S. territory. The science item bank is robust and has been constructed to support multiple statewide science assessments. As described earlier, science items are written to the three-dimensional science standards. The Shared Science Assessment Item Bank comprises ICCR items and items developed for specific states, which are all shared with MOU partner states. These items follow the same specifications, test development processes, and review processes.

In 2018, CAI field-tested more than 540 item clusters and stand-alone items, of which 451 (including items from all sources) were accepted and made available as operational items in 2019. In 2019, 347 item clusters and stand-alone items were field-tested, of which 268 were accepted and made available as operational items in 2020. In 2021, CAI field-tested 545 item clusters and stand-alone items, of which 458 have passed rubric validation and item data review. In 2022, CAI field-tested 471 item clusters and stand-alone items, of which 403 have passed rubric validation and item data review. In 2023, CAI field-tested 348 item clusters and stand-alone items (excluding 90 legacy items, 9 Hawaii research items, and 3 items for interim pool), of which 288 have passed rubric validation and item data review. In 2024, CAI field-tested 478 item clusters and stand-alone items, of which 386 have passed rubric validation and item data review.

Each state or territory using the Shared Science Assessment Item Bank selects items that are appropriately aligned and have passed required reviews (as described in Section 2, Item Development Process That Supports Validity of Claims) for use on its statewide assessment. The Shared Science Assessment Item Bank continues to grow as participating states and territory

continue to field-test new items. Participating states and territory collectively share the items and agree to field-test new items each year.

3.1 CURRENT COMPOSITION OF THE SHARED SCIENCE ASSESSMENT ITEM BANK

The Shared Science Assessment Item Bank contains item clusters and stand-alone items. Item clusters represent a series of interrelated student interactions directed toward describing, explaining, and predicting scientific phenomena. Item clusters can comprise several item parts requiring the student to interact with the item in various ways. In addition, shorter items (stand-alone items) are included to increase the coverage of the assessments without also increasing testing time or testing burden.

Within each item (including both item clusters and stand-alone items), a series of explicit assertions is made about the knowledge and skills that a student has demonstrated based on specific features of the student’s responses across multiple interactions. For example, a student may correctly graph data points indicating that they can construct a graph showing the relationship between two variables, but they may make an incorrect inference about the relationship between the two variables, therefore not supporting the assertion that the student can interpret relationships expressed graphically.

Each year, before development begins, CAI reviews the bank for the state and for the MOU as a whole against the state operational blueprint. The number of eligible items (stand-alone and cluster items) are compared to the requirements of the blueprint. The target numbers are five times the maximum number of clusters and 10 times the maximum number of stand-alone items per Performance Expectation (PE) in the blueprint. If the blueprint has a maximum of one cluster per PE and one stand-alone item per PE then the total target is 15 (i.e., five clusters and 10 stand-alone items). This comparison provides information on areas that need further development. In addition to this target analysis, there are other conditions that might warrant additional item development. These things may include difficulty of the items, releasing operational items to a practice or interim test, a changed blueprint, or overexposed items. CAI creates an Item Development Plan (IDP) to address these areas, as possible, within the limits of the development for that cycle. The IDP is reviewed and approved by the state partner.

Table 6 lists and describes the science interaction types. Examples of various interaction types can be found in Appendix 2-K, Example Item Interactions.

Table 6. Science Interaction Types and Descriptions

Interaction Type	Associated Sub-Types	Description
Choice	Multiple-Choice	Traditional multiple-choice interaction allows students to select a single option from a list of possible answer options.
	Multi-Select	Traditional multi-select interaction (checkboxes) allows students to select one or more options from a list of possible answer choices.
Text Entry	Simple Text Entry	Students type a response in a text box.

Interaction Type	Associated Sub-Types	Description
	Embedded Text Entry	Students type their response in one or more text boxes that are embedded in a section of read-only text.
	Natural Language	Students are directed to provide a short, written response.
	Extended-Response	Students are directed to provide a longer, written response in the form of an essay.
Table	Table Match	Interaction allows students to check a box to indicate if the information from a column header matches information from a row header.
	Table Input	Interaction solicits students to complete tabular data.
Edit Task	Edit Task	Students click a word and replace it with another word that they type to revise a sentence.
	Edit Task with Choice	Students click a word or phrase and select the replacement from several options.
	Edit Task Inline Choice	Drop-down menus are placed throughout the text, and students select an option to complete the text.
Hot-Text	Selectable	Selectable hot-text interactions require students to select one or more text elements in the response area.
	Re-orderable	Re-orderable hot-text interactions require students to click and drag hot-text elements into a different order.
	Drag-from-Palette	Drag-from-Palette hot-text interactions require students to drag elements from a palette into the available blank table cells or “gaps” (text boxes) in the response area.
	Custom	Custom hot-text interactions combine the functionality of the other hot-text interaction sub-types. Students responding to a custom hot-text interaction may need to select text elements, rearrange text elements, and/or drag text elements from a palette to blank table cells or drop targets in the response area.
Equation	N/A	Equation interactions require students to enter a response into input boxes. These boxes may stand alone, or they may be in line with text or embedded in a table. The equation interaction may have an on-screen keypad which may consist of special mathematic characters. Students may also enter their response via a physical keyboard.
Grid	Grid	Grid interactions require students to enter a response by interacting with a grid area in the answer space. The student may be required to draw a line or shape, plot a point, or create a graph. The student may also drag and drop or click on selectable hot-spots.
	Hot-Spot	Hot-spot interaction sub-types facilitate grid interactions with specific hot-spot functionality. These interactions require students to select hot-spot regions in the grid area.
	Graphic Gap Match	Graphic gap match interactions facilitate grid interactions with specific drag-and-drop functionality. These interactions require students to drag image objects from a palette to specified regions (gaps) in the grid area.
Simulation	N/A	Simulation interactions allow students to investigate a phenomenon by selecting variables to get output data. Some simulations are accompanied by animations.

Table 7–Table 11 present the number of items in the Shared Science Assessment Item Bank available for use in the spring 2024 statewide assessments. Appendix 2-L, Shared Science Assessment Item Bank, provides the items available within the bank by grade band, PE, and origin.

Table 7. Spring 2024 Shared Science Assessment Operational and Field-Test Item Bank

Grade Band and Item Type	ICCR Items	Connecticut Items	MOU Items ^a	Total Bank Items
Elementary School	149	98	548	795
Cluster	60	41	302	403
Stand-Alone	89	57	246	392
Middle School	172	102	553	827
Cluster	68	46	298	412
Stand-Alone	104	56	255	415
High School	156	135	304	595
Cluster	60	52	148	260
Stand-Alone	96	83	156	335
Total	477	335	1405	2217

^aOther MOU states include Arkansas, Hawaii, Idaho, Indiana, Montana, Oregon, Rhode Island, Utah, West Virginia, and Wyoming.

Table 8. Spring 2024 Shared Science Assessment Operational Item Bank

Grade Band and Item Type	ICCR Operational Items	Connecticut Operational Items	MOU Operational Items ^a	Total Bank Operational Items
Elementary School	141	77	411	629
Cluster	55	36	243	334
Stand-Alone	86	41	168	295
Middle School	168	81	402	651
Cluster	67	37	218	322
Stand-Alone	101	44	184	329
High School	140	106	213	459
Cluster	52	41	100	193
Stand-Alone	88	65	113	266
Total	449	264	1026	1739

^aOther MOU states include Hawaii, Idaho, Montana, Oregon, Rhode Island, Utah, West Virginia, and Wyoming.

Table 9. Spring 2024 Shared Science Assessment Field-Test Item Bank

Grade Band and Item Type	ICCR Field-Test Items	Connecticut Field-Test Items	MOU Field-Test Items ^a	Total Bank Field-Test Items
Elementary School	8	21	137	166
Cluster	5	5	59	69

Grade Band and Item Type	ICCR Field-Test Items	Connecticut Field-Test Items	MOU Field-Test Items ^a	Total Bank Field-Test Items
Stand-Alone	3	16	78	97
Middle School	4	21	151	176
Cluster	1	9	80	90
Stand-Alone	3	12	71	86
High School	16	29	91	136
Cluster	8	11	48	67
Stand-Alone	8	18	43	69
Total	28	71	379	478

^aOther MOU states include Arkansas, Hawaii, Idaho, Indiana, Montana, Oregon, Rhode Island, Utah, West Virginia, and Wyoming.

Table 10. Spring 2024 Shared Science Assessment Operational and Field-Test Item Bank by Science Discipline

Grade Band	Science Discipline	Item Type	ICCR Items	Connecticut Items	MOU Items ^a	Total Bank Items ^b
Elementary School	Earth and Space Sciences	Cluster	21	13	98	132
		Stand-Alone	24	14	77	115
	Life Sciences	Cluster	19	12	83	114
		Stand-Alone	32	20	73	125
	Physical Sciences	Cluster	20	16	121	157
		Stand-Alone	33	23	96	152
Middle School	Earth and Space Sciences	Cluster	20	10	85	115
		Stand-Alone	26	15	75	114
	Life Sciences	Cluster	26	22	105	153
		Stand-Alone	46	24	87	157
	Physical Sciences	Cluster	22	14	97	133
		Stand-Alone	32	17	91	140
High School	Earth and Space Sciences	Cluster	14	13	34	61
		Stand-Alone	25	23	32	80
	Life Sciences	Cluster	27	22	76	125
		Stand-Alone	42	26	83	151
	Physical Sciences	Cluster	19	17	37	73
		Stand-Alone	29	34	41	104
Total			477	335	1391	2203

^aOther MOU states include Arkansas, Hawaii, Idaho, Indiana, Montana, Oregon, Rhode Island, Utah, West Virginia, and Wyoming. ^bCount excludes fourteen MOU items that do not align to the NGSS and excludes South Dakota legacy items.

Table 11. Spring 2024 Shared Science Assessment Operational and Field-Test Item Bank by Disciplinary Core Idea

Grade Band	Science Discipline	Disciplinary Core Idea	ICCR Items	Connecticut Items	MOU Items ^a	Total Bank Items ^b
Elementary School	Earth and Space Sciences	ESS1	10	8	47	65
		ESS2	16	11	76	103
		ESS3	19	8	52	79
	Life Sciences	LS1	18	10	63	91
		LS2	5	4	34	43
		LS3	7	7	20	34
		LS4	21	11	39	71
	Physical Sciences	PS1	12	8	48	68
		PS2	20	12	43	75
		PS3	15	11	79	105
		PS4	6	8	47	61
Middle School	Earth and Space Sciences	ESS1	15	8	55	78
		ESS2	17	9	52	78
		ESS3	14	8	53	75
	Life Sciences	LS1	24	19	66	109
		LS2	24	11	63	98
		LS3	6	4	20	30
		LS4	18	12	43	73
	Physical Sciences	PS1	17	8	61	86
		PS2	7	8	52	67
		PS3	19	11	42	72
		PS4	11	4	33	48
High School	Earth and Space Sciences	ESS1	12	7	22	41
		ESS2	14	16	26	56
		ESS3	13	13	18	44
	Life Sciences	LS1	19	14	46	79
		LS2	23	21	50	94
		LS3	9	6	22	37
		LS4	18	7	41	66
	Physical Sciences	PS1	19	18	24	61

Grade Band	Science Discipline	Disciplinary Core Idea	ICCR Items	Connecticut Items	MOU Items ^a	Total Bank Items ^b
		PS2	11	14	21	46
		PS3	10	11	17	38
		PS4	8	8	16	32
Total			477	335	1391	2203

^aOther MOU states include Arkansas, Hawaii, Idaho, Indiana, Montana, Oregon, Rhode Island, Utah, West Virginia, and Wyoming.

^bCount excludes fourteen MOU items that do not align to the NGSS and excludes South Dakota legacy items.

3.2 STRATEGY FOR ITEM BANK EVALUATION AND REPLENISHMENT

CAI and the participating MOU states continue to develop items to replenish and grow the Shared Science Assessment Item Bank. The general strategy for targeting item development gathers information from the following three sources:

1. Characteristics of released items to be replaced
2. Characteristics of items that are overused
3. Tabulations of content coverage and ranges of difficulty (to identify gaps in the bank)

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

4. CONNECTICUT NGSS ASSESSMENT TEST CONSTRUCTION

4.1 TEST DESIGN

The Connecticut NGSS Assessment was administered online to students in grades 5, 8, and 11 using an adaptive test design in spring 2024. Appendix 2-M, Connecticut NGSS Assessment Item Pool, provides the 2024 item pool by grade band, performance expectation (PE), and origin. In an adaptive test, operational items are selected on the fly based on the performance of a student on past items while ensuring the test blueprint is followed for each individual student. An advantage of adaptive testing is that it can provide more precise scores for students with lower and higher proficiencies, in contrast to fixed forms and linear-on-the fly tests (LOFTs) that are typically targeted to provide the best precision for students with medium proficiencies. Also, as opposed to a fixed form and a LOFT, every student has the potential to see a different set of items that adapt to the student’s ability, thus offering a better testing experience.

Items are selected by an item-selection algorithm based on the content and information value. At any given point during the test, the content value of an item is determined by its contribution to meeting the blueprint, given the content characteristics of the items that have already been administered. During the test, the content value increases for items that exhibit features that have not met their designated minimum as the end of the test approaches. Conversely, the content value decreases for items with content features that met the minimum. The information value of an item is based on the item information function evaluated at the estimated proficiency. The proficiency estimate is updated throughout the test.

The adaptive item-selection algorithm is the same algorithm CAI uses to deliver ELA and mathematics tests, but with some modifications to make it suitable for using item clusters. Specifically, the proficiencies that are estimated during the test are computed under an IRT model that incorporates cluster effects. In order to avoid over-selection of items with many scoring

assertions, the information of an item at an estimated proficiency level is normalized by the number of assertions in the item (similar to how information is computed for item sets in ELA and Mathematics assessments). Details for CAI’s adaptive testing algorithm are described in Appendix 2-M, Adaptive Algorithm Design.

A non-segmented test design was used for the Connecticut NGSS Assessment, meaning students received items from different disciplines in a random order (in contrast to a segmented test design, in which separate parts of the test are administered in a fixed order). In an adaptive test, the use of a non-segmented test design provides more freedom when selecting items targeting a current best estimate of proficiency. Embedded field-test items were randomly positioned in the test and randomly distributed across students. Every student received either one item cluster or four stand-alone items as field-test items throughout the test.

4.2 TEST BLUEPRINTS

Test blueprints provide the following guidelines:

- Length of the test
- Science disciplines to be covered and the acceptable number of items across PEs within each science discipline and Disciplinary Core Idea (DCI)

The blueprint for science is presented in Table 12–Table 14.

Table 12. Science Test Blueprint, Grade 5

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

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

Grade 5	Min Item Clusters	Max Item Clusters	Min Stand-Alone Items	Max Stand-Alone Items	Min Item Clusters + Min Stand-Alone Items	Max Item Clusters + Max Stand-Alone Items
3-ESS2-2: Weather and Climate	0	1	0	1	0	1
4-ESS2-1: Earth's Systems and Processes	0	1	0	1	0	1
4-ESS2-2: Earth's Systems and Processes	0	1	0	1	0	1
5-ESS2-1: Earth's Systems	0	1	0	1	0	1
5-ESS2-2: Earth's Systems	0	1	0	1	0	1
DCI—Earth and Human Activity	0	1	0	1	0	2
3-ESS3-1: Weather and Climate*	0	1	0	1	0	1
4-ESS3-2: Earth's Systems and Processes*	0	1	0	1	0	1
4-ESS3-1: Energy	0	1	0	1	0	1
5-ESS3-1: Earth's Systems	0	1	0	1	0	1
DCI—Earth's Place in the Universe	0	1	0	1	0	2
4-ESS1-1: Earth's Systems and Processes	0	1	0	1	0	1
5-ESS1-1: Space Systems	0	1	0	1	0	1
5-ESS1-2: Space Systems	0	1	0	1	0	1
PE Total = 42	6	6	12	12	18	18

*These PEs have an engineering component.

Table 13. Science Test Blueprint, Grade 8

Grade 8	Min Item Clusters	Max Item Clusters	Min Stand-Alone Items	Max Stand-Alone Items	Min Item Clusters + Min Stand-Alone Items	Max Item Clusters + Max Stand-Alone Items
Discipline—Physical Sciences, PE Total = 19	2	2	4	4	6	6
DCI—Matter and Its Interactions	0	1	0	2	0	3
MS-PS1-1: Structure and Properties of Matter	0	1	0	1	0	1
MS-PS1-2: Chemical Reactions	0	1	0	1	0	1
MS-PS1-3: Structure and Properties of Matter	0	1	0	1	0	1
MS-PS1-4: Structure and Properties of Matter	0	1	0	1	0	1
MS-PS1-5: Chemical Reactions	0	1	0	1	0	1
MS-PS1-6: Chemical Reactions*	0	1	0	1	0	1
DCI—Motion and Stability: Forces and Interactions	0	1	0	2	0	3
MS-PS2-1: Forces and Interactions*	0	1	0	1	0	1
MS-PS2-2: Forces and Interactions	0	1	0	1	0	1
MS-PS2-3: Forces and Interactions	0	1	0	1	0	1
MS-PS2-4: Forces and Interactions	0	1	0	1	0	1
MS-PS2-5: Forces and Interactions	0	1	0	1	0	1
DCI—Energy	0	1	0	2	0	3
MS-PS3-1: Energy	0	1	0	1	0	1
MS-PS3-2: Energy	0	1	0	1	0	1
MS-PS3-3: Energy*	0	1	0	1	0	1
MS-PS3-4: Energy	0	1	0	1	0	1
MS-PS3-5: Energy	0	1	0	1	0	1
DCI—Waves and Their Applications in Technologies for Information Transfer	0	1	0	2	0	3
MS-PS4-1: Waves and Electromagnetic Radiation	0	1	0	1	0	1

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

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

Grade 8	Min Item Clusters	Max Item Clusters	Min Stand-Alone Items	Max Stand-Alone Items	Min Item Clusters + Min Stand-Alone Items	Max Item Clusters + Max Stand-Alone Items
MS-ESS3-2: Human Impacts	0	1	0	1	0	1
MS-ESS3-3: Human Impacts*	0	1	0	1	0	1
MS-ESS3-4: Human Impacts	0	1	0	1	0	1
MS-ESS3-5: Weather and Climate	0	1	0	1	0	1
PE Total = 55	6	6	12	12	18	18

*These PEs have an engineering component.

Table 14. Science Test Blueprint, Grade 11

Grade 11	Min Item Clusters	Max Item Clusters	Min Stand-Alone Items	Max Stand-Alone Items	Min Item Clusters + Min Stand-Alone Items	Max Item Clusters + Max Stand-Alone Items
Discipline—Physical Sciences, PE Total = 24	2	2	4	4	6	6
DCI—Matter and Its Interactions	0	1	0	2	0	3
HS-PS1-1: Structure and Properties of Matter	0	1	0	1	0	1
HS-PS1-2: Structure and Properties of Matter	0	1	0	1	0	1
HS-PS1-3: Structure and Properties of Matter	0	1	0	1	0	1
HS-PS1-4: Chemical Reactions	0	1	0	1	0	1
HS-PS1-5: Chemical Reactions	0	1	0	1	0	1
HS-PS1-6: Chemical Reactions*	0	1	0	1	0	1
HS-PS1-7: Chemical Reactions	0	1	0	1	0	1
HS-PS1-8: Nuclear Processes	0	1	0	1	0	1
DCI—Motion and Stability: Forces and Interactions	0	0	0	2	0	2
HS-PS2-1: Forces and Motion	0	0	0	1	0	1
HS-PS2-2: Forces and Motion	0	0	0	1	0	1
HS-PS2-3: Forces and Motion*	0	0	0	1	0	1
HS-PS2-4: Types of Interactions	0	0	0	1	0	1
HS-PS2-5: Types of Interactions	0	0	0	1	0	1
HS-PS2-6: Chemical Reactions*	0	0	0	1	0	1
DCI—Energy	0	1	0	2	0	3
HS-PS3-1: Energy	0	1	0	1	0	1
HS-PS3-2: Energy	0	1	0	1	0	1
HS-PS3-3: Energy*	0	1	0	1	0	1
HS-PS3-4: Energy	0	1	0	1	0	1

Grade 11	Min Item Clusters	Max Item Clusters	Min Stand-Alone Items	Max Stand-Alone Items	Min Item Clusters + Min Stand-Alone Items	Max Item Clusters + Max Stand-Alone Items
HS-PS3-5: Energy	0	1	0	1	0	1
DCI—Waves and Their Applications in Technologies for Information Transfer	0	0	0	2	0	2
HS-PS4-1: Wave Properties	0	0	0	1	0	1
HS-PS4-2: Wave Properties	0	0	0	1	0	1
HS-PS4-3: Wave Properties/Electromagnetic Radiation	0	0	0	1	0	1
HS-PS4-4: Electromagnetic Radiation	0	0	0	1	0	1
HS-PS4-5: Electromagnetic Radiation*	0	0	0	1	0	1
Discipline—Life Sciences, PE Total = 24	2	2	4	4	6	6
DCI—From Molecules to Organisms: Structures and Processes	0	1	0	2	0	3
HS-LS1-1: Structure and Function	0	1	0	1	0	1
HS-LS1-2: Structure and Function	0	1	0	1	0	1
HS-LS1-3: Structure and Function	0	1	0	1	0	1
HS-LS1-4: Growth and Development of Organisms	0	1	0	1	0	1
HS-LS1-5: Organization for Matter and Energy Flow in Organisms	0	1	0	1	0	1
HS-LS1-6: Organization for Matter and Energy Flow in Organisms	0	1	0	1	0	1
HS-LS1-7: Organization for Matter and Energy Flow in Organisms	0	1	0	1	0	1
DCI—Ecosystems: Interactions, Energy, and Dynamics	0	1	0	2	0	3
HS-LS2-1: Interdependent Relationships in Ecosystems	0	1	0	1	0	1
HS-LS2-2: Interdependent Relationships in Ecosystems	0	1	0	1	0	1

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

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

*These PEs have an engineering component.

The main characteristics of the blueprint were that any PE could be tested only once (indicated by the values of 0 and 1 for the minimum and maximum values of the individual PEs in Table 12–Table 14). In general, no more than one item cluster or two stand-alone items could be sampled from the same DCI, and no more than three total items could be sampled from the same DCI (as indicated by the minimum and maximum values in the rows representing DCIs). Some specific constraints for the Connecticut blueprint stated that for grades 5 and 8, the Earth Systems DCI had to have two stand-alone items (rather than one for other DCIs in the Earth and Space Sciences Discipline) because it had the most PEs and was rated the highest in the district responses. In addition, three DCIs in grade 11 (Motion and Stability, Waves, and Earth’s Place in the Universe) were constrained to not receive an item cluster due to low content priority ratings from districts.

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

*Table 15. Spring 2024 Connecticut NGSS Assessment
85th Percentile Testing Times by Grade*

Subject	Grade	85th Percentile Testing Time
Science	5	131.13
	8	116.28
	11	92.00

4.3 ONLINE TEST CONSTRUCTION

During fall 2023, CAI psychometricians and content experts worked with CSDE content specialists and leadership to build item pools for the spring 2024 administration. The Connecticut NGSS Assessment test construction used a structured test construction plan, explicit blueprints, and active collaborative participation from all parties.

CAI test developers built the 2024 Connecticut NGSS Assessment test item pools to match items exactly to the detailed test blueprints. Operational items were selected from ten item banks (ICCR, Connecticut, Hawaii, Idaho, Montana, Oregon, Rhode Island, Utah, West Virginia, and Wyoming) to fulfill the blueprint for each grade. Table 16–Table 20 summarize the 2024 Connecticut NGSS Assessment item pool. Appendix 2-M, Connecticut NGSS Assessment Item Pool, provides the 2024 item pool by grade band, PE, and origin.

*Table 16. Spring 2024 Connecticut NGSS Assessment Operational and Field-Test
Item Pool*

Grade and Item Type	ICCR Items	Connecticut Items	MOU Items ^a	Total Pool Items
Grade 5	89	98	102	289
Cluster	31	41	51	123
Stand-Alone	58	57	51	166

Grade and Item Type	ICCR Items	Connecticut Items	MOU Items ^a	Total Pool Items
Grade 8	97	102	155	354
Cluster	35	46	94	175
Stand-Alone	62	56	61	179
Grade 11	107	131	74	312
Cluster	32	48	35	115
Stand-Alone	75	83	39	197
Total	293	331	331	955

^aOther MOU states include Arkansas, Hawaii, Idaho, Indiana, Montana, Oregon, Rhode Island, Utah, West Virginia, and Wyoming.

Table 17. Spring 2024 Connecticut NGSS Assessment Operational Item Pool

Grade and Item Type	ICCR Operational Items	Connecticut Operational Items	MOU Operational Items ^a	Total Operational Pool Items
Grade 5	88	77	96	261
Cluster	30	36	49	115
Stand-Alone	58	41	47	146
Grade 8	97	81	148	326
Cluster	35	37	90	162
Stand-Alone	62	44	58	164
Grade 11	104	102	67	273
Cluster	32	37	31	100
Stand-Alone	72	65	36	173
Total	289	260	311	860

^aOther MOU states include Hawaii, Idaho, Montana, Oregon, Rhode Island, Utah, West Virginia, and Wyoming.

Table 18. Spring 2024 Connecticut NGSS Assessment Field-Test Item Pool

Grade and Item Type	ICCR Field-Test Items	Connecticut Field-Test Items	MOU Field-Test Items ^a	Total Field-Test Pool Items
Grade 5	1	21	6	28
Cluster	1	5	2	8
Stand-Alone	0	16	4	20
Grade 8	0	21	7	28
Cluster	0	9	4	13
Stand-Alone	0	12	3	15
Grade 11	3	29	7	39

Grade and Item Type	ICCR Field-Test Items	Connecticut Field-Test Items	MOU Field-Test Items ^a	Total Field-Test Pool Items
Cluster	0	11	4	15
Stand-Alone	3	18	3	24
Total	4	71	20	95

^aOther MOU states include Arkansas, Indiana, Montana, Oregon, Rhode Island, Utah, and Wyoming.

Table 19. Spring 2024 Connecticut NGSS Assessment Operational and Field-Test Item Pool by Science Discipline

Grade	Science Discipline	Item Type	ICCR Items	Connecticut Items	MOU Items ^a	Total Pool Items
5	Earth and Space Sciences	Cluster	12	13	15	40
		Stand-Alone	14	14	20	48
	Life Sciences	Cluster	9	12	17	38
		Stand-Alone	22	20	13	55
	Physical Sciences	Cluster	10	16	19	45
		Stand-Alone	22	23	18	63
8	Earth and Space Sciences	Cluster	9	10	29	48
		Stand-Alone	16	15	21	52
	Life Sciences	Cluster	13	22	35	70
		Stand-Alone	29	24	17	70
	Physical Sciences	Cluster	13	14	30	57
		Stand-Alone	17	17	23	57
11	Earth and Space Sciences	Cluster	8	12	5	25
		Stand-Alone	21	23	7	51
	Life Sciences	Cluster	13	22	23	58
		Stand-Alone	35	26	20	81
	Physical Sciences	Cluster	11	14	7	32
		Stand-Alone	19	34	12	65
Total			293	331	331	955

^aOther MOU states include Hawaii, Idaho, Montana, Oregon, Rhode Island, Utah, West Virginia, and Wyoming.

Table 20. Spring 2024 Connecticut NGSS Assessment Operational and Field-Test Item Pool by Disciplinary Core Idea

Grade	Science Discipline	Disciplinary Core Idea	ICCR Items	Connecticut Items	MOU Items ^a	Total Pool Items
5	Earth and Space Sciences	ESS1	8	8	12	28
		ESS2	9	11	15	35
		ESS3	9	8	8	25
	Life Sciences	LS1	9	10	10	29
		LS2	4	4	6	14
		LS3	4	7	6	17
		LS4	14	11	8	33
	Physical Sciences	PS1	9	8	10	27
		PS2	10	12	11	33
		PS3	11	11	11	33
		PS4	2	8	5	15
8	Earth and Space Sciences	ESS1	7	8	16	31
		ESS2	9	9	16	34
		ESS3	9	8	18	35
	Life Sciences	LS1	17	19	19	55
		LS2	11	11	18	40
		LS3	4	4	5	13
		LS4	10	12	10	32
	Physical Sciences	PS1	10	8	23	41
		PS2	3	8	13	24
		PS3	11	11	11	33
		PS4	6	4	6	16
11	Earth and Space Sciences	ESS1	7	6	1	14
		ESS2	11	16	8	35
		ESS3	11	13	3	27
	Life Sciences	LS1	16	14	16	46
		LS2	10	21	11	42
		LS3	8	6	5	19
		LS4	14	7	11	32

Grade	Science Discipline	Disciplinary Core Idea	ICCR Items	Connecticut Items	MOU Items ^a	Total Pool Items
	Physical Sciences	PS1	15	18	12	45
		PS2	5	11	2	18
		PS3	6	11	4	21
		PS4	4	8	1	13
Total			293	331	331	955

^aOther MOU states include Hawaii, Idaho, Montana, Oregon, Rhode Island, Utah, West Virginia, and Wyoming.

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

4.4 PAPER-PENCIL ACCOMMODATION FORM CONSTRUCTION

Student scores should not depend on the mode of administration or type of test form. Because the Connecticut NGSS Assessment was primarily administered in an online test delivery system in spring 2024, only 7 students took the paper-pencil form in grade 5, 10 students took it in grade 8, and 4 in grade 11. Scores obtained via alternate modes of administration were established as comparable to scores obtained through online testing. This section outlines the overall test development plans that ensured the comparability of online and paper-pencil tests.

To build paper-pencil forms, content specialists began with the online pool and removed any items that could not be rendered on paper. Next, content specialists constructed fixed forms adhering to the test blueprint. All overall and discipline-level (reporting category) blueprint requirements were met; however, due to the availability of items in paper-pencil forms, some blueprint requirements at the DCI level were violated. For the grade 5 paper-pencil test, the blueprint requires two stand-alone items for Earth and Space Science (ESS) 2, but the form had one stand-alone item for ESS2.

5. SIMULATION SUMMARY REPORT

This section describes the results of the simulated test administrations used to configure and evaluate the adequacy of the item-selection algorithm, which was used to administer the 2023–2024 Connecticut NGSS Assessment for grades 5, 8, and 11. Simulations were conducted to configure the settings of the algorithm and to evaluate whether individual tests adhered to the test blueprint.

Some important settings included “Select Candidate Set 1” (cset1) and “Select Candidate Set 2” (cset2), which represent subsets of the item pool that were eligible for item selection. Refer to Appendix 2-N, Adaptive Algorithm Design, for more details of the current item selection algorithm. In spring 2024, cset1 and cset2 values were set to 10 and 5 for both online English and Spanish tests in all grades except for the Spanish test in grade 11, in which the values were set to 5 and 1. Psychometricians reviewed the simulation results and configured settings based on some key diagnostics, including the following:

- **Match-to-Test Blueprint.** This diagnostic determines whether the tests have the correct number of test items overall and the appropriate proportion by content categories at each level of the content hierarchy, as specified in the test blueprints for every science grade.
- **Item Exposure Rate.** This diagnostic evaluates the utility of item pools and identifies overexposed and underexposed items.
- **Precision.** This diagnostic determines whether the size of the standard error of measurement is within the acceptable range and whether there is any possible bias in the estimates of student ability.

These diagnostics are interrelated. For example, if the test pool for a particular content category is limited (i.e., there are only a few test items available), achieving a 100% match to the blueprint for this content level will lead to a high item exposure rate, which means that a large number of students are sharing items. The software system that performs the simulation allows adjustments to the setting parameters in order to attain the best possible balance among these diagnostics. The simulation involves an iterative process that reviews initial results, adjusts these system parameters, runs new simulations, reviews the new results, and repeats the exercise until an optimal balance is achieved. The final setting would then be applied for the operational tests.

5.1 FACTORS AFFECTING SIMULATION RESULTS

There are several factors that may influence simulation results for an adaptive administration. These factors include the following:

- *The proportional relationship between the pool and the constraints to be met.* Proportionally distributed pools tend to make better use of the pool (i.e., more uniform item exposure) and make it easier to meet blueprint and other constraints. For example, if the specifications call for at least one item cluster per Disciplinary Core Idea (DCI), but the pool has no item cluster for some DCIs, it may be impossible to meet this constraint.
- *The correlational structure between constraints.* It is easier to satisfy a constraint if there are instances of the constraint at all levels of another constraint. For example, if stand-alone items within a discipline are associated only with a specific DCI, it may be difficult to meet both the desired distribution of content and the desired distribution of item type.
- *Whether or not there is a strict maximum on a given constraint.* This means that the requirement must be met exactly in each test administration.

5.2 RESULTS OF SIMULATED TEST ADMINISTRATIONS: ENGLISH

This section presents the simulation results for the English online tests, which is the test taken by most students (97.46%). Simulations were evaluated for all content areas using 15,000 simulated cases for grades 5 and 8, and 5,000 for grade 11.

5.2.1 Summary of Blueprint Match

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

5.2.2 Item Exposure

The simulator output also reports the degree to which the constraints set forth in the blueprints may yield greater exposure of items to students. This is reported by examining the percentage of test administrations in which an item appears. For instance, in a fixed paper-based form, 100% of the items appear on 100% of the test administrations because every test taker takes the same form. In an adaptive test or a LOFT with a sufficiently large item pool, it is expected that most of the items would appear on a relatively small percentage of the test administrations.

When this condition holds, it suggests that test administrations between students are more or less unique. Therefore, the item exposure rate was calculated for each item by dividing the total number of test administrations in which an item appears by the total number of tests administered. Then, the distribution of the item exposure rate (r) is reported in eight bins. The bins are $r = 0\%$ (unused), $0\% < r \leq 1\%$, $1\% < r \leq 5\%$, $5\% < r \leq 20\%$, $20\% < r \leq 40\%$, $40\% < r \leq 60\%$, $60\% < r \leq 80\%$, and $80\% < r \leq 100\%$. If global item exposure is minimal, it is expected that the largest proportion of items would appear in the bins of $0\% < r \leq 20\%$, which is an indication that most of the items appear on a very small percentage of the test forms.

Table 21 presents the percentage of items that fell into each exposure bin for all grades. Most test items were administered in 1%–20% of the test administrations. No items had an exposure rate higher than 60% in all three grades.

Table 21. Spring 2024 Item Exposure Rates by Grade: Percentage of Items by Exposure Rate, Across All English Online Simulation Sessions

Grade	Total Items	[0,0]%	(0,1]%	(1,5]%	(5,20]%	(20,40]%	(40,60]%	(60,80]%	(80,100]%
5	257	0	0	30.35	68.09	1.17	0.39	0	0
8	321	0	11.53	40.5	46.73	1.25	0	0	0
11	270	0.37	7.41	41.85	48.52	1.85	0	0	0

5.2.3 Precision

Each simulated record includes a true score and an ability estimate based on the adaptive test administration. The correlations between the true score and estimated ability for grades 5, 8, and 11 were 0.941, 0.955, and 0.955, respectively. Correlations between the estimated ability and the true score were nearly one, indicating that the adaptive test administrations reliably estimated student ability.

The mean bias, which is the average of the biases of the estimated abilities across all students, was 0.002, 0.001, and -0.016 for grades 5, 8, and 11, respectively. In all cases, the mean bias of the estimated abilities was very small, providing further evidence that the true score was adequately recovered in the observed score.

Table 22 shows the mean standard errors of the ability estimate across all simulated test administrations and the standard error at the 5th, 25th, 75th, and 95th percentiles of the ability distribution. For all English tests, the standard error was lowest at the low end of the ability spectrum, indicating a greater precision of measurement of lower performing students. Conversely, the standard error was highest for higher ability students.

Table 22. Spring 2024 Standard Errors of Ability Estimates, by Grade, Across All English Online Simulation Sessions

Grade	Overall Mean	5th Percentile	25th Percentile	75th Percentile	95th Percentile
5	0.354	0.316	0.334	0.366	0.413
8	0.311	0.283	0.296	0.320	0.351
11	0.317	0.284	0.297	0.325	0.379

5.3 RESULTS OF SIMULATED TEST ADMINISTRATIONS: SPANISH

This section presents the simulation results for the Spanish tests. The Spanish item pool consisted of a subset of ICCR items and some MOU items that had Spanish translations available. Table 22 presents the number of items available for the Spanish tests in spring 2024.

Table 23. Spring 2024 Spanish Operational Item Pool

Grade	Item Type	Number of Items
5	Cluster	33
	Stand-Alone	47
8	Cluster	26
	Stand-Alone	55
11	Cluster	32
	Stand-Alone	64
Total		257

Simulations were evaluated for all content areas using 5,000 simulated cases per grade.

5.3.1 Summary of Blueprint Match

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

5.3.2 Item Exposure

Table 24 presents the percentage of items that fell into each exposure bin for all grades. Most items were administered in 1%–40% of the test administrations. Some items had an exposure rate higher than 60% because of the limited Spanish item pool for some content categories.

Table 24. Spring 2024 Item Exposure Rates by Grade: Percentage of Items by Exposure Rate, Across All Spanish Simulation Sessions

Grade	Total Items	[0,0]%	(0,1]%	(1,5]%	(5,20]%	(20,40]%	(40,60]%	(60,80]%	(80,100]%
5	80	0	0	0	51.25	42.50	5.00	1.25	0
8	81	0	0	0	53.09	41.98	2.47	0	2.47

Grade	Total Items	[0,0]%	(0,1]%	(1,5]%	(5,20]%	(20,40]%	(40,60]%	(60,80]%	(80,100]%
11	97	1.03	2.06	20.62	42.27	21.65	8.25	4.12	0

5.3.3 Precision

For the Spanish tests, the correlations between the true score and estimated ability for grades 5, 8, and 11 were 0.944, 0.953, and 0.953, respectively. Correlations between the estimated ability and the true score were nearly one, indicating that the adaptive test administrations reliably estimated student ability.

The mean bias was 0.001, 0.006, and -0.002 for grades 5, 8, and 11, respectively. In all cases, the mean bias of the estimated abilities was very small, providing further evidence that the true score was adequately recovered in the observed score.

Table 25 shows the mean standard errors of the ability estimate across all simulated test administrations and the standard error at the 5th, 25th, 75th, and 95th percentiles of the ability distribution. For all Spanish tests, the standard error was lowest at the low end of the ability spectrum, indicating a greater precision of measurement of lower performing students. Conversely, the standard error was highest for higher ability students.

Table 25. Spring 2024 Standard Errors of Ability Estimates, by Grade, Across All Spanish Online Simulation Sessions

Grade	Overall Mean	5th Percentile	25th Percentile	75th Percentile	95th Percentile
5	0.348	0.308	0.326	0.360	0.409
8	0.322	0.291	0.303	0.331	0.378
11	0.330	0.302	0.313	0.335	0.387

6. OPERATIONAL TEST ADMINISTRATION SUMMARY REPORT

This section presents the blueprint match reports and item exposure rates for the spring 2024 operational test administrations.

6.1 BLUEPRINT MATCH

The English online tests in all grades met the blueprint specifications with a 100% match at all content levels. For the Spanish tests, all tests met the blueprint specifications with a 100% match at all content levels, except for two students in grade 8. The student with blueprint violation in grade 8 received two stand-alone items from the “Earth’s Place in the Universe” (MS-ESS1) DCI, while the blueprint requires at most one stand-alone item from each of these DCIs.

These types of violations did not happen during simulations. The reason they occurred in the operational test administrations is that these students had seen the items that were designed to meet the blueprint requirements before the test administration. The item selection algorithm automatically filtered out the items they had already seen so the students would not see the same items twice. There are two possible scenarios for this type of violation to occur. First, the students saw the items in a previous attempt in the current school year. Second, these students took the science test at the same grade in previous test administrations. Therefore, the pool became shallower for these students. At the end of the test, the algorithm did not have the option to select an item that would satisfy the blueprint requirement, and it could only select an item that caused blueprint violations. Note that these violations were all below the discipline level. Both students with blueprint violations in the spring 2024 Spanish tests took an item that would have satisfied the blueprint requirements in another English online test.

6.2 ITEM EXPOSURE

Table 26 presents the item exposure rates for the spring 2024 test administration. The exposure rates were very similar to the simulation results described in Section 5.2.2, Item Exposure, for the English test administrations. For the Spanish tests, more items had high exposure rates as compared to the English tests because of a smaller item pool. Also, the operational exposure rates were slightly different from the simulation results due to small population sizes in all three grades. In spring 2024, around 800 to 1000 students took the Spanish test across grades.

Table 26. Spring 2024 Item Exposure Rates by Grade: Percentage of Items by Exposure Rate, Across All Test Administrations

Grade	Total Items	0%	(0,1]%	(1,5]%	(5,20]%	(20,40]%	(40,60]%	(60,80]%	(80,100]%
English									
5	257	0	0	32.68	65.76	1.17	0.39	0	0
8	321	0	9.97	42.37	46.11	1.56	0	0	0
11	270	0.37	6.3	43.7	48.15	1.48	0	0	0
Spanish									
5	80	0	0	1.25	45	47.5	5	1.25	0
8	81	0	0	8.64	40.74	43.21	4.94	0	2.47
11	97	1.03	4.12	41.24	29.9	7.22	4.12	5.15	7.22

7. REFERENCES

- Council of Chief State School Officers. (2015). *Science Assessment Item Collaborative (SAIC) assessment framework for the Next Generation Science Standards*. Washington, DC: Council of Chief State School Officers. Retrieved from https://ccsso.org/sites/default/files/2017-12/SAICAssessmentFramework_FINAL.pdf
- Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, 25(4), 6–20.
- National Research Council. (2012). *A framework for K–12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: The National Academies Press.

Connecticut Next Generation Science Standards Assessment

2023–2024

Volume 3: Setting Performance Standards



CONNECTICUT STATE
DEPARTMENT OF EDUCATION

TABLE OF CONTENTS

1.	EXECUTIVE SUMMARY	1
1.1	Standard-Setting Workshop	2
1.1.1	Overall Structure of the Workshop	2
1.1.2	Results of the Standard-Setting Workshop.....	3
2.	INTRODUCTION	5
3.	THE NEXT GENERATION SCIENCE STANDARDS	6
4.	CONNECTICUT’S NGSS SCIENCE ASSESSMENT	7
4.1	Item Clusters and Stand-alone Items	7
4.2	Scoring Assertions	7
5.	STANDARD SETTING	8
5.1	The Assertion-Mapping Procedure.....	9
5.2	Workshop Structure	11
5.3	Participants and Roles.....	11
5.3.1	Connecticut Department of Education Staff.....	11
5.3.2	American Institutes for Research Staff	12
5.3.3	Room Facilitators	12
5.3.4	Educator Participants.....	13
5.3.5	Table Leaders.....	16
5.4	Materials	16
5.4.1	Performance-Level Descriptors.....	16
5.4.2	Ordered Scoring Assertion Booklets.....	17
5.5	Workshop Technology.....	19
5.6	Events.....	19
5.6.1	Table Leader Orientation	20
5.6.2	Large-Group Introductory Training.....	21
5.6.3	Confidentiality and Security	21
5.6.4	Take the Test	21
5.6.5	Range Performance-Level Descriptor Review.....	22
5.6.6	Create Threshold Performance-Level Descriptors.....	22
5.6.7	Ordered Scoring Assertion Booklet Review.....	22
5.6.8	Assertion-Mapping Training.....	22
5.6.9	Practice Quiz	24
5.6.10	Practice Round.....	24
5.6.11	Readiness Assertion	25
5.7	Assertion Mapping.....	25
5.7.1	Calculating Cut Scores from the Assertion Mapping	25
5.7.2	Feedback and Impact Data	26
5.7.3	Context Data	26
5.7.4	Benchmark Data	26

5.8	Workshop Results	27
5.8.1	Round 1	27
5.8.2	Round 2	28
5.9	Post Workshop Refinements	30
5.10	Workshop Evaluations	32
5.10.1	Workshop Participant Feedback	35
6.	VALIDITY EVIDENCE	35
6.1	Evidence of Adherence to Professional Standards and Best Practices	36
6.2	Evidence in Terms of Peer Review Critical Elements	37
7.	REFERENCES	39

LIST OF TABLES

Table 1. Performance Standards Recommended for Science	3
Table 2. Percentage of Students Reaching or Exceeding Each Recommended Science Performance Standard in 2019	3
Table 3. Percentage of Students Classified Within Each Science Performance Level in 2019.....	4
Table 4. Table Assignments.....	11
Table 5. Panelist Characteristics	13
Table 6. Panelist Qualifications	15
Table 7. Standard-Setting Agenda Summary.....	20
Table 8. Round 1 Results	27
Table 9. Round 2 Results	28
Table 10. Percentage of Students Classified Within Each Recommended Science Performance Level in 2019	29
Table 11. Post-Standard-Setting Workshop: Final Cut Scores (Change from Workshop Recommendation) and Impact Data.....	31
Table 12. Post-Standard-Setting Workshop: Percentage of Students Classified Within Each Science Performance Level in 2019	31
Table 13. Evaluation Results: Clarity of Materials and Process.....	32
Table 14. Evaluation Results: Appropriateness of Process	33
Table 15. Evaluation Results: Importance of Materials.....	33
Table 16. Evaluation Results: Understanding Processes and Tasks	34
Table 17. Evaluation Results: Student Expectations	35

LIST OF FIGURES

Figure 1. Percentage of Students Reaching or Exceeding Each Recommended Science Performance Standard in 2019.....	4
Figure 2. Percentage of Students Classified Within Each Science Performance Level in 2019	5
Figure 3. Structure of NGSS Performance Expectations.....	6
Figure 4. Example NGSS Item Cluster and Scoring Assertions.....	8
Figure 5. Three Performance Standards Defining Connecticut’s Four Performance Levels	9
Figure 6. Workshop Panels per Room	11
Figure 7. Ordered Scoring Assertion Booklet (OSAB)	18
Figure 8. Example Features in Standard-Setting Tool.....	19
Figure 9. Example of Assertion Mapping.....	24
Figure 10. Percentage of Students Reaching or Exceeding Each Recommended Science Performance Standard in 2019.....	29
Figure 11. Percentage of Students Classified Within Each Recommended Science Performance Level in 2019	30
Figure 12. Post-Standard-Setting Workshop: Percentage of Students Reaching or Exceeding Each Science Performance Standard in 2019	31
Figure 13. Post-Standard-Setting Workshop: Percentage of Students Classified Within Each Science Performance Level in 2019	32

LIST OF APPENDICES

Appendix 3-A. Standard-Setting Panelist Characteristics	
Appendix 3-B. Development of Range Performance-Level Descriptors	
Appendix 3-C. Standard-Setting Workshop Agenda	
Appendix 3-D. Standard-Setting Training Slides	
Appendix 3-E. Standard-Setting Practice Quiz	
Appendix 3-F. Standard-Setting Readiness Forms	

1. EXECUTIVE SUMMARY

In November 2015, the Connecticut State Department of Education (CSDE) adopted the Next Generation Science Standards (NGSS). The new standards employ a three-dimensional conceptualization of science understanding, including science and engineering practices, crosscutting concepts, and disciplinary core ideas. With the adoption of the NGSS standards in science, and the development of new statewide assessments to measure student achievement relative to those standards, the CSDE convened a standard-setting workshop to recommend a system of performance standards for determining whether students have met the learning goals defined by the NGSS science standards.

Under contract to CSDE, the American Institutes for Research (AIR; currently Cambium Assessment, Inc. [CAI]) conducted the standard-setting workshop to recommend performance standards for Connecticut’s Next Generation Science Standards (NGSS) Assessment at grades 5, 8, and 11. The workshop was conducted July 31–August 1, 2019, at the Red Lion Hotel Cromwell, 100 Berlin Road, Cromwell, CT.

Connecticut’s NGSS Assessments are designed to measure attainment of the NGSS adopted by CSDE. The assessments are comprised of item clusters and stand-alone items. Item clusters represent a series of interrelated student interactions directed toward describing, explaining, and predicting scientific phenomena. Stand-alone items are added to increase the test’s coverage of the standards while limiting increases in testing time and burden on students and schools. Test items were developed by AIR (now CAI) in conjunction with a group of states working to implement three-dimensional NGSS. Test items were developed to ensure that each student is administered a test meeting all elements of Connecticut’s NGSS Assessment blueprint, which was constructed to align to the NGSS.

Connecticut science educators, serving as standard-setting panelists, followed a rigorous standardized procedure to recommend performance standards demarcating each performance level. To recommend performance standards for the new science assessments, panelists participated in the Assertion-Mapping Procedure, an adaptation of the Item-Descriptor (ID) Matching procedure (Ferrara & Lewis, 2012). Consistent with ordered-item procedures generally (e.g., Mitzel, Lewis, Patz, & Green, 2001), workshop panelists reviewed and recommended performance standards using an ordered set of scoring assertions¹ derived from student interactions within items. Because the new science items—specifically the item clusters—represent multiple, interdependent interactions through which students engage in scientific phenomena, scoring assertions cannot be meaningfully evaluated independently of the item interactions from which they are derived. Thus, panelists were presented ordered scoring assertions for each item separately rather than for the test overall. Panelists mapped each scoring assertion to the most apt performance-level descriptor (PLD).

Panelists reviewed PLDs describing the degree to which students have achieved Connecticut’s NGSS. Range PLDs were reviewed and revised by CSDE prior to the standard-setting workshop.

¹ Scoring assertions articulate the evidence the student provides as a means to infer a specific skill or concept, which is aligned to content standards. In other words, scoring assertions capture each measurable action of an item and articulate what evidence the student has provided to infer a specific skill or concept.

After reviewing the range PLDs, standard-setting panelists worked to identify knowledge and skills characteristic of students just qualifying for entry into each performance level.

Working through the ordered scoring assertions for each item, panelists mapped each assertion into one of the four performance levels—Does Not Meet, Approaching, Meets, and Exceeds. The panelists performed the assertion mapping in two rounds of standard setting during the two-day workshop. Panelists’ mapping of the scoring assertions was used to identify the location of the three performance standards used to classify student achievement—Approaching, Meets, and Exceeds. Mapping of scoring assertions in Round 1 was based only on consideration of test content. Following Round 1, panelists were provided with feedback about the mappings of their fellow panelists and discussed their mappings as a group. Panelists were then provided additional contextual information, including the percentage of students who performed at or above the proficiency level associated with each individual assertion, as well as the projected achievement level of the National Assessment of Educational Progress (NAEP) science, Smarter Balanced English language arts (ELA) and mathematics for elementary and middle school grades, and SAT evidence-based reading and writing and mathematics college-ready indicators for grade 11 for each assertion.

Forty-two Connecticut science educators² served as science standard-setting panelists, with 15 participants each for the grade 5 and grade 11 panels, and 12 participants in the grade 8 panel. The panelists represented a group of experienced teachers and curriculum specialists, as well as district administrators and other stakeholders. The composition of the panel ensured that a diverse range of perspectives contributed to the standard-setting process. The panel was also representative in terms of gender, race/ethnicity, and region of the state.

1.1 STANDARD-SETTING WORKSHOP

1.1.1 Overall Structure of the Workshop

The key features of the workshops included the following:

- The standard-setting procedure produced three recommended performance standards (Approaching, Meets, and Exceeds) that will be used to classify student science performance on the Connecticut NGSS Assessment.
- Panelists recommended performance standards in two rounds.
- Context data, including the percentage of students who performed at or above the proficiency level associated with each individual assertion, and approximate benchmark locations for NAEP science performance standards, Smarter Balanced ELA and mathematics performance standards for elementary and middle school grades, and SAT evidence-based reading and writing and mathematics college-ready indicators, were provided to panelists following the first round of recommending performance standards.
- The standard-setting workshops were conducted online using AIR’s online standard-setting tool. A laptop computer was provided for each panelist at the workshop.

² See Section 5.3.4, Educator Participants for more information on the panelists.

1.1.2 Results of the Standard-Setting Workshop

Table 1 displays the performance standards recommended by the standard-setting panelists.³

Table 1. Performance Standards Recommended for Science

Grade	Level 2 Approaching	Level 3 Meets	Level 4 Exceeds
5	465	493	525
8	783	798	842
11	1078	1099	1141

Table 2 indicates the percentage of students that will reach each of the performance standards in 2019.

Table 2. Percentage of Students Reaching or Exceeding Each Recommended Science Performance Standard in 2019

Grade	Level 2 Approaching	Level 3 Meets	Level 4 Exceeds
5	87	60	22
8	69	52	9
11	74	48	11

Figure 1 represents those values graphically.

³ Following the standard-setting workshop, the Connecticut State Department of Education (CSDE) reviewed and made some refinements to the final panelist-recommended performance standards. More information on this is available in Section 5.9, Post Workshop Refinement, and the post-standard-setting workshop final cut scores are presented in Table 11.

Figure 1. Percentage of Students Reaching or Exceeding Each Recommended Science Performance Standard in 2019

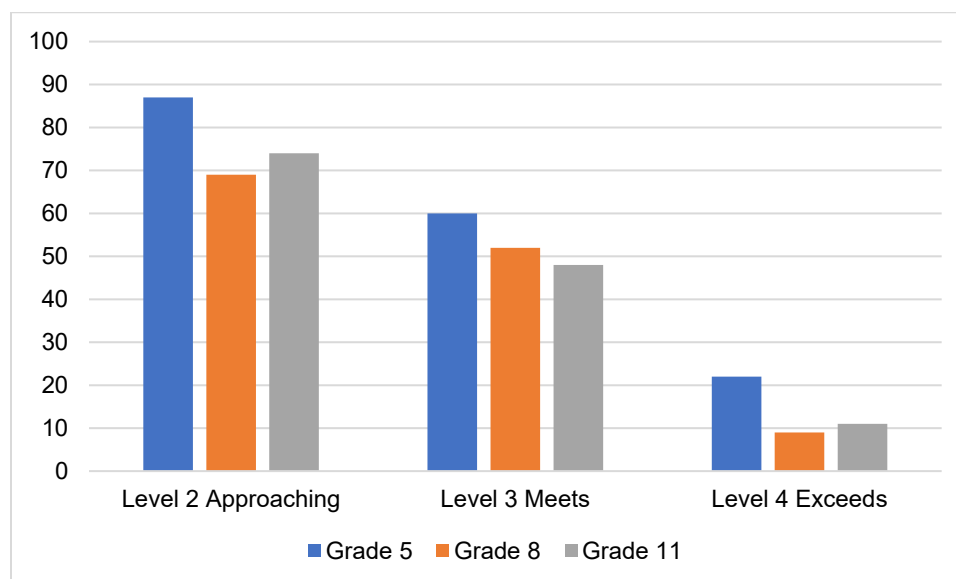
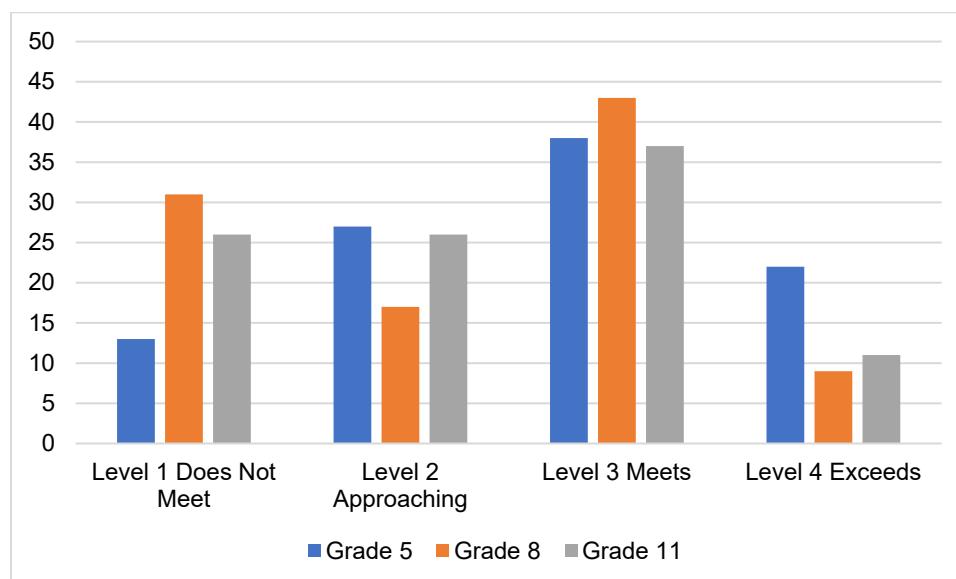


Table 3 indicates the percentage of students classified within each of the performance levels in 2019. The values are displayed graphically in Figure 2.

Table 3. Percentage of Students Classified Within Each Science Performance Level in 2019

Grade	Level 1 Does Not Meet	Level 2 Approaching	Level 3 Meets	Level 4 Exceeds
5	13	27	38	22
8	31	17	43	9
11	26	26	37	11

Figure 2. Percentage of Students Classified Within Each Science Performance Level in 2019



2. INTRODUCTION

Connecticut adopted the Next Generation Science Standards (NGSS) in 2015. The Connecticut State Department of Education (CSDE) and its assessment vendor, the American Institutes for Research (AIR; now Cambium Assessment, Inc. [CAI]) developed and administered a new assessment to measure the new standards. Piloted in 2016–2017, field-tested in 2017–2018, and administered operationally for the first time in 2018–2019, the new Connecticut NGSS measure the science knowledge and skills of Connecticut students in grades 5, 8, and 11.

CSDE provides an overview of the science assessment at: <https://portal.ct.gov/SDE/Student-Assessment/NGSS-Science/NGSS-Science>.

New tests require new performance standards to link performance on the test to the content standards. The CSDE contracted with AIR (now CAI) to establish cut scores for the new tests. To fulfill this responsibility, AIR implemented an innovative, defensible, valid, and technically sound method; provided training on standard setting to all participants; oversaw the process; computed real-time feedback data to inform the process; and produced a technical report documenting the method, approach, process, and outcomes. Performance standards were recommended for grades 5, 8, and 11 science in July 2019.

The purpose of this documentation is to detail the standard-setting process for NGSS science and resulting performance standard recommendations.

3. THE NEXT GENERATION SCIENCE STANDARDS

The Connecticut Next Generation Science Standards (NGSS) Assessment assesses the learning objectives described by the NGSS, adopted in 2015. Information about the NGSS is available at: www.nextgenscience.org.

The three-dimensional science standards (i.e., the NGSS), based on *A Framework for K–12 Science Education* (National Research Council, 2012), reflect the latest research and advances in modern science education 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 NGSS refer to these performed knowledge and skills as *performance expectations* (PEs). Second, the NGSS are intentionally multi-dimensional. Each performance expectation incorporates all three dimensions from the NGSS framework—a science or engineering practice, a disciplinary core idea, and a crosscutting concept. Third, while traditional standards do not consider other subject areas, the NGSS connect to other subjects like the Common Core mathematics and English language arts (ELA) standards. Another unique feature of the NGSS is the assumption that students should learn all science disciplines, rather than select a few, as is traditionally done in many high schools, where students may elect to take biology and chemistry but not physics or astronomy.

Figure 3 shows the structure of the NGSS for a single grade 5 PE, 5-PS1-1.

Figure 3. Structure of NGSS Performance Expectations

Students who demonstrate understanding can: 5-PS1-1. Develop a model to describe that matter is made of particles too small to be seen. [Clarification Statement: Examples of evidence supporting a model could include adding air to expand a basketball, compressing air in a syringe, dissolving sugar in water, and evaporating salt water.] [Assessment Boundary: Assessment does not include the atomic-scale mechanism of evaporation and condensation or defining the unseen particles.]		
The performance expectation above was developed using the following elements from the NRC document <i>A Framework for K–12 Science Education</i> :		
Science and Engineering Practices Developing and Using Models Modeling in 3–5 builds on K–2 experiences and progresses to building and revising simple models and using models to represent events and design solutions. • Use models to describe phenomena.	Disciplinary Core Ideas PS1.A: Structure and Properties of Matter • Matter of any type can be subdivided into particles that are too small to see, but even then the matter still exists and can be detected by other means. A model showing that gases are made from matter particles that are too small to see and are moving freely around in space can explain many observations, including the inflation and shape of a balloon and the effects of air on larger particles or objects.	Crosscutting Concepts Scale, Proportion, and Quantity • Natural objects exist from the very small to the immensely large.
Connections to other DCIs in fifth grade: N/A		
Articulation of DCIs across grade-levels: 2.PS1.A ; MS.PS1.A		
Common Core State Standards Connections:		
ELA/Literacy - RI.5.7 Draw on information from multiple print or digital sources, demonstrating the ability to locate an answer to a question quickly or to solve a problem efficiently. (5-PS1-1)		
Mathematics - MP.2 Reason abstractly and quantitatively. (5-PS1-1) MP.4 Model with mathematics. (5-PS1-1) 5.NBT.A.1 Explain patterns in the number of zeros of the product when multiplying a number by powers of 10, and explain patterns in the placement of the decimal point when a decimal is multiplied or divided by a power of 10. Use whole-number exponents to denote powers of 10. (5-PS1-1) 5.NF.B.7 Apply and extend previous understandings of division to divide unit fractions by whole numbers and whole numbers by unit fractions. (5-PS1-1) 5.MD.C.3 Recognize volume as an attribute of solid figures and understand concepts of volume measurement. (5-PS1-1) 5.MD.C.4 Measure volumes by counting unit cubes, using cubic cm, cubic in, cubic ft, and improvised units. (5-PS1-1)		

* The performance expectations marked with an asterisk integrate traditional science content with engineering through a Practice or Disciplinary Core Idea.

Source. <https://www.nextgenscience.org/pe/5-ps1-1-matter-and-its-interactions>.

4. CONNECTICUT’S NGSS SCIENCE ASSESSMENT

Due to the unique features of the three-dimensional Next Generation Science Standards (NGSS), items and tests based on the NGSS, such as Connecticut’s test, must also incorporate similarly unique features. The most impactful of these changes is that NGSS tests are multi-dimensional and are thus comprised mostly of item clusters representing a series of interrelated student interactions directed toward describing, explaining, and predicting scientific phenomena.

4.1 ITEM CLUSTERS AND STAND-ALONE ITEMS

There are two types of items: item clusters and stand-alone items. An item cluster includes a phenomenon-based stimulus and a series of interactions that allow the student to demonstrate their mastery of the performance expectation (PE) by explaining the phenomenon or designing a solution to a presented engineering problem. The expectation is that item clusters will take students approximately 10 to 12 minutes to complete. Each stimulus ends with a task statement that provides the goal or understanding the student should reach. For example, “In the questions that follow, you will analyze what happens to the train when the brakes are applied.” The student may explain, model, investigate, and/or create designs using the knowledge, skills, and abilities described by the PE. For example, in Figure 3, proficiency in this single PE requires activities that demonstrate the ability to analyze and evaluate data, the knowledge of properties and purposes of different forms of matter, and the application of experimental cause and effect. All interactions within an item cluster address the phenomenon presented in the stimulus. Item clusters contain between four and eight interactions.

Most states also utilize stand-alone items. Stand-alone items increase the number of covered PEs per student while being much quicker to complete than item clusters. Incorporating stand-alone items allows the blueprint to cover a greater number of PEs within a limited time. Stand-alone items are also phenomenon-based, contain only one or two interactions, and take students one to three minutes to complete in general.

Both item types may use any of the available interaction types, including selected response, multi-select, table match, external copy, edit in-line choice, grids, and/or simulations of scientific investigations. For additional information on interaction types, refer to Volume 2, Appendix 2-C, Style Guide for Science Items, of this technical report.

4.2 SCORING ASSERTIONS

Each item cluster and stand-alone item assumes a series of explicit assertions about the knowledge and skills that a student demonstrates based on specific features of the student’s responses across multiple interactions. *Scoring assertions* capture each measurable moment and articulate what evidence the student has provided as a means to infer a specific skill or concept. Some stand-alone items have more than one scoring assertion, while all item clusters have multiple scoring assertions.

Figure 4 illustrates an item cluster and associated scoring assertions. CSDE provides sample items at: <https://ct.portal.cambiumast.com/>.

Figure 4. Example NGSS Item Cluster and Scoring Assertions

Stimulus and phenomenon →

Item Cluster

Cluster task statement →

Scoring Assertions

Score Rationale

Score Rationale	
The student selected "wheels" for the first blank and "brakes" or "rails" for the second blank showing an understanding of the interactions in the system and the effects of that energy flow.	✗
The student selected "wheels" for the third blank and "less" for the fourth blank showing an understanding of the interactions in the system and the effects of that energy flow.	✗
The student selected "The surroundings gain energy," showing an understanding of how the energy of the wheels change and is distributed throughout the system.	✗
The student selected "Sound is produced," providing evidence of how the energy of the surroundings has changed.	✗
The student selected "Light is produced," providing evidence of how the energy of the surroundings has changed.	✗
The student selected "Heat is produced," providing evidence of how the energy of the surroundings has changed.	✗
The student selected "The brakes make a screeching sound," which shows an understanding of how the energy changed throughout the system and that those changes serve as evidence that the Kinetic Energy of the wheels transfers out of the wheels/system when the brakes are applied.	✗
The student selected "The sparks that fly off the wheels give off light," which shows an understanding of how the energy changed throughout the system and that those changes serve as evidence that the Kinetic Energy of the wheels transfers out of the wheels/system when the brakes are applied.	✗
The student selected "The brakes give off energy as heat," which shows an understanding of how the energy changed throughout the system and that those changes serve as evidence that the Kinetic Energy of the wheels transfers out of the wheels/system when the brakes are applied.	✗

Part A
Click on each blank box to select the word or phrase that completes each sentence, constructing an argument about what happens when the train's brakes are applied.
Applying the brakes causes the _____ to transfer kinetic energy to the _____. This causes the _____ to slow down and have _____ kinetic energy, which slows the train.

Part B
When the train applies its brakes, what happens to _____?
☐ The surroundings gain energy.
☐ The surroundings lose energy.
☐ The surroundings do not gain or lose energy.
☐ There is not enough information to determine if the surroundings gain or lose energy.

Part C
Which **three** statements support your choice in part B?
☐ The train maintains its speed.
☐ Sound is produced.
☐ Sound is consumed.
☐ Light is produced.
☐ Light is consumed.
☐ Heat is produced.
☐ Heat is consumed.

Table 1. Properties of the Train System

Before Brakes Are Applied	After Brakes Applied
No sparks	Sparks fly off the wheels and brake pads
Brake pads make no sound	Brake pads make sound

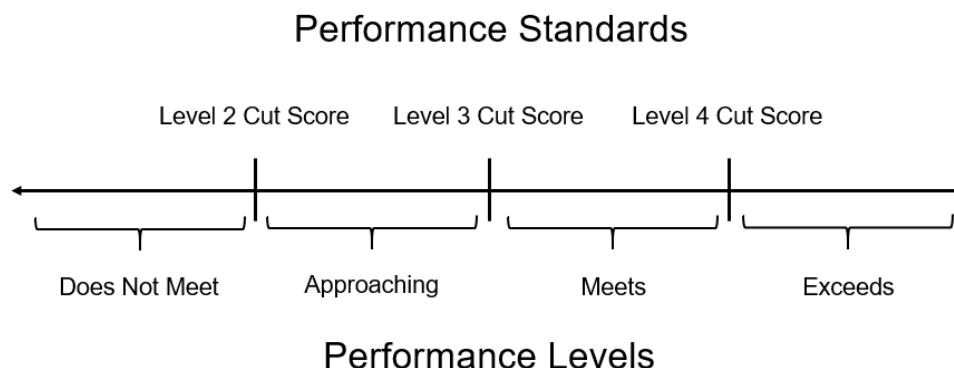
Your Task
In the questions that follow, you will analyze what happens to the train when the brakes are applied.

5. STANDARD SETTING

Forty-two educators from Connecticut convened at the Red Lion in Cromwell, CT, from July 31 through August 1, 2019, to complete two rounds of standard setting to recommend three performance standards for the new Connecticut Next Generation Science Standards (NGSS) Assessment.

Standard setting is the process used to define performance on the test. Performance levels are defined by performance standards, or *cut scores*, that specify how much of the performance expectations (PEs) students must know and be able to do in order to meet the minimum for each performance level. As shown in Figure 5, three performance standards are sufficient to define Connecticut's four performance levels.

Figure 5. Three Performance Standards Defining Connecticut’s Four Performance Levels



The cut scores are derived from the knowledge and skills measured by the test items that students at each performance level are expected to be able to answer correctly.

5.1 THE ASSERTION-MAPPING PROCEDURE

A modification of traditional approaches to setting performance standards is necessary for tests based on the Next Generation Science Standards (NGSS) due to the structure of the performance expectations, and subsequently, the structure of test items assessing the performance expectations. While traditional tests and measurement models assume unidimensionality, tests based on the NGSS adopt a three-dimensional conceptualization of science understanding. Each item cluster or stand-alone item aligns to a science practice, one or more crosscutting concepts, and one disciplinary core idea. Accordingly, the new science assessments are comprised mostly of item clusters representing a series of interrelated student interactions directed toward describing, explaining, and predicting scientific phenomena. Some stand-alone items are added to increase the test’s coverage of the standards without also increasing testing time or testing burden.

Within each item, a series of explicit assertions are made about the knowledge and skills that a student has demonstrated based on specific features of the student’s responses across multiple interactions. For example, a student may correctly graph data points indicating that they can construct a graph showing the relationship between two variables but may make an incorrect inference about the relationship between the two variables, thereby not supporting the assertion that the student can interpret relationships expressed graphically.

While some other assessments, especially English language arts (ELA), comprise items probing a common stimulus, the degree of interdependence among such items is limited; and student performance on such items can be evaluated independently of student performance on other items within the stimulus set. This is not the case with the new science items, which may, for example, involve multiple steps in which students interact with products of previous steps. However, unlike with traditional stimulus- or passage-based items, the conditional dependencies between the interactions and resulting assertions of an item cluster are too substantial to ignore because those item interactions and assertions are more intrinsically related to each other. The interdependence of student interactions within items has consequences both for scoring and recommending performance standards.

To account for the cluster-specific variation of related item clusters, additional dimensions can be added to the item response theory (IRT) model. Typically, these are nuisance dimensions unrelated to student ability. Examples of IRT models that follow this approach are the bi-factor model (Gibbons & Hedeker, 1992) and the testlet model (Bradlow, Wainer, & Wang, 1999). The testlet model is a special case of the bi-factor model (Rijmen, 2010).

Because the item clusters represent performance tasks, the Body of Work (BoW) method (Kingston, Kahl, Sweeny, & Bay, 2001) could be appropriate for recommending performance standards. However, the BoW method is manageable only with small numbers of performance tasks and quickly becomes onerous when the number of item clusters approaches 10 or more.

Skaggs, Hein, & Awuor (2007) proposed a standard setting method called the Single-Passage Bookmark method to address challenges presented by passage-based assessments. This method is a variation of the traditional Bookmark method (e.g., Mitzel, Lewis, Patz, & Green, 2001) in which individual ordered item booklets (OIBs) are created for each set of items associated with a passage. Items within each OIB are arranged in order of difficulty. The task of the panelists is to place a bookmark in each OIB as opposed to a single OIB in the traditional Bookmark method. Even though this method showed promise, one limitation and concern expressed by the authors is whether this method can be applied to derive two or more standards.

To address these challenges, AIR (now CAI) psychometricians designed a new method for setting performance standards on new tests of the NGSS. AIR implemented this method for the New Hampshire, Utah, and West Virginia state assessments in 2018.

The test-centered Assertion-Mapping Procedure (AMP) is an adaptation of the Item-Descriptor (ID) Matching procedure (Ferrara & Lewis, 2012) that preserves the integrity of the item clusters while also taking advantage of ordered-item procedures such as the Bookmark procedure used frequently for other accountability tests (Rijmen et al., 2018).

The main distinction between AMP and the Single-Passage Bookmark method is that the panelists evaluate scoring assertions rather than individual items. Scoring assertions are not test items, but inferences that are supported (or not supported) by students' responses in one or more interactions within an item cluster or stand-alone item. Because item clusters represent multiple, interdependent interactions through which students engage in scientific phenomena, scoring assertions cannot be meaningfully evaluated independently of the item from which they are derived. Therefore, the scoring assertions from the same item cluster or stand-alone item are always presented together. Within each item cluster or stand-alone item, scoring assertions are ordered by difficulty (i.e., the IRT difficulty parameter) consistent with the Single-Passage Bookmark method. One can think of the resulting booklet as consisting of different chapters, where each chapter represents an item cluster or stand-alone item. Within each chapter, the (ordered) pages represent scoring assertions. As in ID matching, panelists are asked to map each scoring assertion to the most apt performance-level descriptor (PLD) during two rounds of standard setting. Like the Bookmark method, assertion mappings are made independently with the goal of convergence over two rounds of rating, rather than consensus.⁴

⁴ AIR (now CAI) historically implements two rounds of standard setting as best practice in the Bookmark method and extends this practice to the AMP method. In addition to lessening the panelists' burden of needing to repeat a cognitively demanding task for a third time, using two rounds introduces significant cost efficiency by reducing the

5.2 WORKSHOP STRUCTURE

One large meeting room served as an all-participant training room. This room broke into three separate working rooms, one for each set of grade-level panels, after the all-group orientation. As shown in Figure 6, three separate panels set performance standards for each grade.

Figure 6. Workshop Panels Per Room

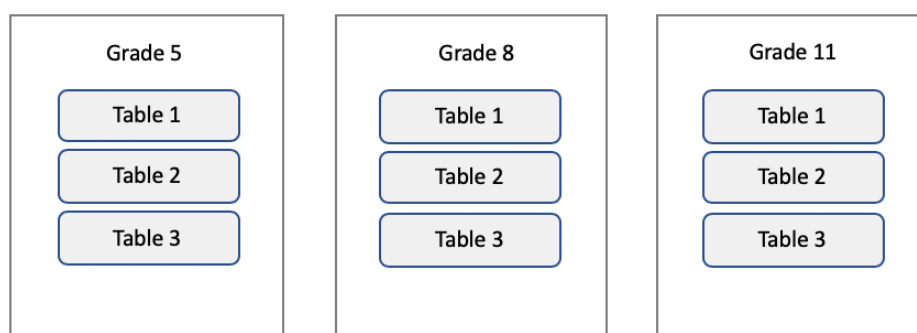


Table 4 summarizes the composition of the tables and the number of facilitators and panelists assigned to each. The 42 standard-setting participants included table leaders and panelists who taught in the content area and grade level for which standards were being set.

Table 4. Table Assignments

Room	Grade	Tables & Table Leaders (One Per Table)	Panelists (Per Table)	Facilitator	Facilitator Assistant
1	5	3	5/5/5	Jim McCann	Matt Davis
2	8	3	4/4/4	Kevin Dwyer	Kam Mangis de Mark
3	11	3	5/5/5	Meg McMahon	Heather MacRae

Note. The Connecticut State Department of Education (CSDE) recruited 15 panelists per grade, but three were unable to attend at the last minute so the total number of panelists was 42.

5.3 PARTICIPANTS AND ROLES

5.3.1 Connecticut Department of Education Staff

Staff from the Connecticut State Department of Education (CSDE) were present throughout the process and provided overall policy context and answered any policy questions that arose. They included:

- Abe Krisst, Performance Office, Bureau Chief

number of days needed for standard setting. Panels typically converge in Round 2, and panelists completing two rounds report levels of confidence in the outcomes that are similar to the confidence expressed by panelists participating in three rounds. Psychometric evaluation of the reliability and variability in results from two and three rounds are generally consistent. AIR has used two rounds in standard setting in more than 12 states and 30 assessments, beginning in 2001 with the enactment of the No Child Left Behind (NCLB) Act.

- Janet Stuck, Special Populations
- Jeff Greig, Science
- Michelle Rosado, Connecticut SAT School Day
- Pei-Hsuan Chiu, Psychometrics Team
- Mohamed Dirir, Psychometrics Team
- Michael Sabados, Data Team

5.3.2 American Institutes for Research Staff

AIR (now CAI) facilitated the workshop and each of the content-area rooms, provided psychometric and statistical support, and oversaw technical set-up and logistics. AIR team members included the following personnel:

- Dr. Stephan Ahadi, Managing Director of Psychometrics, facilitated and oversaw all AMP processes and tasks. He provided training to participants, including the facilitators and table leaders.
- Jennifer Chou, Program Director, oversaw the project and managed processes and logistics throughout the meeting.
- Dr. Frank Rijmen, Director of Psychometrics, supervised all psychometric analyses conducted during and after the workshop.
- Dr. Dandan Liao, Psychometrician, provided psychometric analyses.
- Alesha Ballman, Psychometric Project Coordinator, oversaw analytics technology and psychometrics.
- Patrick Kozak, Psychometric Support Manager, and Azza Hussein, Psychometric Support Assistant, provided support.
- Drew Azar and Dotun Adebayo set up, tested, and troubleshooted technology during the workshop.

5.3.3 Room Facilitators

An AIR room facilitator and assistant facilitator guided the process in each room. Facilitators were content experts experienced in leading standard-setting processes, had led standard-setting processes before, and could answer any questions about the workshop or about the items or what the items were intended to measure. They also monitored time and motivated panelists to complete tasks within the scheduled time.

- Jim McCann, assisted by Matt Davis, facilitated the grade 5 panel.
- Kevin Dwyer, assisted by Kam Mangis de Mark, facilitated the grade 8 panel.
- Meg McMahon, assisted by Heather MacRae, facilitated the grade 11 panel.

Each facilitator was trained to be extensively knowledgeable of the constructs, processes, and technologies used in standard setting.

5.3.4 Educator Participants

To establish performance standards, CSDE recruited a diverse set of participants from across the state. Panelists included science teachers, administrators, and representatives from other stakeholder groups (e.g., higher education) to ensure that a diverse range of perspectives contributed to the standard-setting process and product. In recruiting panelists, CSDE targeted the recruitment of participants to be representative of the gender and geographic representation of the teacher population found in Connecticut and the diversity of the students they serve. All participants also had to be familiar with the NGSS content and test.

Overall, panelists were 24% male and 29% non-white. They included teachers, administrators, and other stakeholder groups who worked in schools (48%), districts (29%), both schools and districts (17%) and elsewhere (7%). Panelists represented suburban districts (45%), urban districts (33%), and rural districts (19%) that were most often medium in size (45%), followed by small (31%) and large (21%). Ninety percent taught science and a third (33%) taught both elementary school and middle school students. Table 5 summarizes characteristics of the panels.

Table 5. Panelist Characteristics

	Percentage of Panelists by Panel			
	Grade 5	Grade 8	Grade 11	Overall
Characteristics				
Male	7%	17%	47%	24%
Non-White	20%	42%	27%	29%
Stakeholder Group				
Administrator	13%	25%	13%	17%
Coach	7%	17%	0%	7%
Coach, Administrator	0%	8%	0%	2%
Coach, Other	7%	0%	0%	2%
Other	7%	0%	7%	5%
Professor	0%	0%	7%	2%
Specialist	13%	0%	7%	7%
Teacher	33%	17%	47%	33%
Teacher, Administrator, Other	0%	0%	7%	2%
Teacher, Coach	0%	17%	0%	5%
Teacher, Coach, Other	7%	8%	0%	5%
Teacher, Other	7%	0%	13%	7%
Teacher, Specialist	7%	8%	0%	5%
Current Position				

	Percentage of Panelists by Panel			
	Grade 5	Grade 8	Grade 11	Overall
School	33%	50%	60%	48%
District	33%	42%	13%	29%
School, District	27%	8%	13%	17%
Other	7%	0%	13%	7%
District Size				
Large	20%	17%	27%	21%
Medium	47%	50%	40%	45%
Small	27%	33%	33%	31%
Not applicable	7%	0%	0%	2%
District Urbanicity				
Urban	33%	42%	27%	33%
Suburban	60%	42%	33%	45%
Rural	0%	17%	40%	19%
Not applicable	7%	0%	0%	2%
Primary Grades Taught				
ES (grades 1–5)	33%	8%	0%	14%
MS (grades 6–8)	13%	17%	13%	14%
HS (grades 9–12)	0%	8%	60%	24%
ES and MS (grades 1–8)	40%	50%	13%	33%
MS and HS (grades 6–12)	0%	8%	13%	7%
N/A (Non-educators)	13%	8%	0%	7%
Subjects Taught				
Science	87%	92%	93%	90%
Other (including N/A)	13%	8%	7%	10%

Note. Number of participants = 42 (grade 5 participants = 15, grade 8 participants = 12, and grade 11 participants = 15). Other stakeholder groups included the department chair, consultant, adjunct professor, and curriculum coordinator.

For results of any judgment-based method to be valid, the judgments must be made by individuals who are qualified to make them. Participants in the Connecticut NGSS standard-setting workshop were highly qualified. They brought a variety of experience and expertise. All held a master's degree or higher and nearly a third had taught for more than 20 years. Over half (52%) had taught in their assigned panel's grade and subject for 1–10 years, while 19% had taught it for more than 20 years. Most (67%) had professional experience outside the classroom and over half (between 55 and 60%) were experienced in teaching special student populations. Table 6 summarizes the qualifications of the panels.

Table 6. Panelist Qualifications

	Percentage of Panelists by Panel			
	Grade 5	Grade 8	Grade 11	Overall
Highest Degree				
Bachelors	0%	0%	0%	0%
Masters	47%	75%	53%	57%
Doctorate	20%	17%	20%	19%
Sixth Year/Education Specialist	33%	8%	27%	24%
Years teaching experience				
0 years	0%	0%	0%	0%
1–5 years	13%	8%	20%	14%
6–10 years	27%	25%	27%	26%
11–15 years	20%	17%	7%	14%
16–20 years	0%	25%	20%	14%
21+ years	40%	25%	27%	31%
Years teaching experience in assigned grade/subject				
0 years	13%	17%	13%	14%
1–5 years	33%	33%	13%	26%
6–10 years	27%	17%	33%	26%
11–15 years	13%	17%	0%	10%
16–20 years	0%	0%	13%	5%
21+ years	13%	17%	27%	19%
Other professional experience in education	67%	83%	53%	67%
Years professional experience in education				
0 years	33%	17%	47%	33%
1–5 years	33%	33%	13%	26%
6–10 years	13%	25%	13%	17%
11–15 years	13%	0%	7%	7%
16–20 years	0%	8%	13%	7%
21+ years	7%	17%	7%	10%
Experience teaching special student populations				
Students receiving free/reduced price lunch	40%	50%	80%	57%
English Language Learners (ELLs)	47%	42%	73%	55%
Students on an Individualized Education Plan (IEP)	47%	42%	87%	60%

Note. Number of participants = 42 (grade 5 participants = 15, grade 8 participants = 12, and grade 11 participants = 15). Other professional experience in education included positions such as science coordinator, department chair or dean, committee member, coach, or specialist.

Appendix 3-A, Standard-Setting Panelist Characteristics, provides additional information about the individuals participating in the standard-setting workshop.

5.3.5 Table Leaders

CSDE pre-selected table leaders from the participant pool for their specialized knowledge or experience with the assessment, items, or NGSS. Table leaders also served as panelists and set individual cut scores or assigned assertions.

Table leaders trained as a group early in the morning of the first day to ensure that each table leader was knowledgeable of the constructs, processes, and technologies used in standard setting and was able to adhere to a standardized process across the grade/subject committees. Training consisted of an overview of their responsibilities and some process guidance.

Table leaders provided the following support throughout the workshop:

- Lead table discussions
- Helped panelists see the ‘big picture’
- Monitored security of materials
- Monitored panelist understanding and reported issues or misunderstandings to room facilitators
- Maintained a supportive atmosphere of professionalism and respect

5.4 MATERIALS

5.4.1 Performance-Level Descriptors

With the adoption of the new standards in science, and the development of new statewide assessments to assess performance of those standards, CSDE must adopt a similar system of performance, or performance standards, to determine whether students have met the learning goals defined by the new standards in science.

Determining the nature of the categories into which students are classified is a prerequisite to standard setting. These categories, or performance levels, are associated with performance-level descriptors (PLDs) that define the content-area knowledge, skills, and processes that students at each performance level can demonstrate.

PLDs link the content standards (NGSS performance expectations) to the performance standards. There are four types of PLDs:

1. Policy PLDs: These are brief descriptions of each performance level that do not vary across grade or content area.
2. Range PLDs: Provided to panelists to review and endorse during the workshop, these detailed grade- and content-area-specific descriptions communicate exactly what students performing at each level know and can do.

3. **Threshold PLDs:** Typically created during and used for standard setting only, these describe what a student just barely scoring into each performance level knows and can do. They may also be called Target PLDs or Just Barely PLDs.
4. **Reporting PLDs:** These are much-abbreviated PLDs (typically 350 or fewer characters) created following state approval of the performance standards used to describe student performance on score reports.

Connecticut uses four performance levels to describe student performance: “Does Not Meet,” “Approaching,” “Meets,” and “Exceeds.”

Science Range Performance-Level Descriptor Development

AIR and staff from participating states’ departments of education reviewed existing range PLDs from several states’ assessments based on three-dimensional science standards. States selected the range PLDs based on standards drafted by the Washington Office of Superintendent of Public Instruction (OSPI) as a starting point. Subsequently, AIR, state department of education staff, and educators from multiple states using AIR’s science assessment item bank convened in May of 2018 to review and refine the draft range PLDs. The panels created policy PLDs and reviewed and identified refinements to the range PLDs to describe observable evidence for what student performance looks like in science at each performance level and grade. AIR and one of the NGSS authors reviewed and applied recommendations to the PLDs. They ensured consistency, coherence, and articulation across grades and levels. Appendix 3-B, *Development of Range Performance-Level Descriptors*, provides additional information about the development of the range PLDs prior to the standard-setting workshop.

The CSDE then reviewed the PLDs to ensure that the language accurately represented the goals and policies of the state. AIR worked with them to make revisions where necessary. The Connecticut State Science Assessment Advisory Committee also reviewed the PLDs and made revisions where necessary.

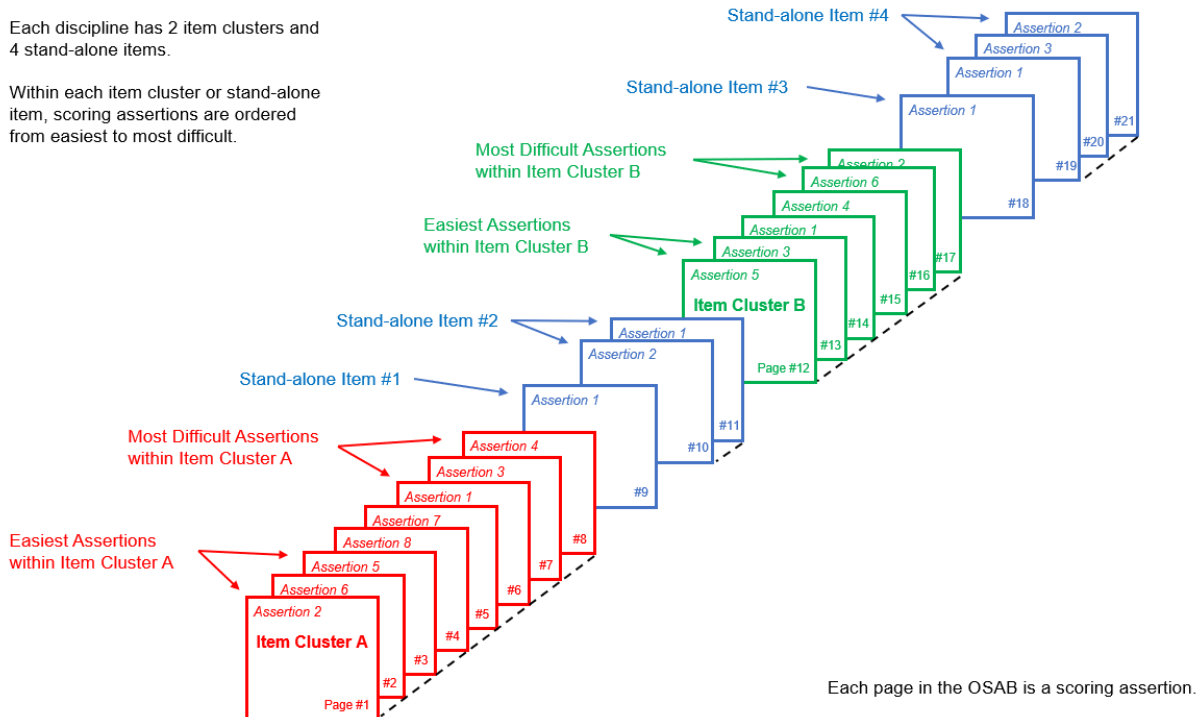
5.4.2 Ordered Scoring Assertion Booklets

Like the Bookmark method used for establishing performance standards for the Connecticut Alternate Science tests (CTAS), the AMP method uses booklets of ordered test materials for setting standards. Instead of test items, the AMP uses scoring assertions presented in grade-specific booklets called ordered scoring assertion booklets (OSABs). Each OSAB represents one possible testing instance resulting from applying the test blueprints to the item pool.

The OSABs were assembled using a mixed integer programming approach. The objective function that was minimized was the number of gaps between the impact values of the assertions across the entire OSAB. A gap was defined as a difference of 3% or more between the impact values of two consecutive assertions ordered by difficulty. The linear constraints of the mixed integer problem represented the constraints implied by the blueprint. In addition, the total number of assertions was not allowed to exceed 85. A set of feasible solutions was further evaluated based on the distribution of the impact values of assertions across the OSAB. The candidate solution was then reviewed internally by content experts and by the CSDE and approved without any changes for all three grades.

Figure 7 describes the structure of the OSAB.

Figure 7. Ordered Scoring Assertion Booklet (OSAB)



For the operational test, the order of the items was randomized over students. For the OSABs, Earth and Space Sciences items were presented first, then Life Sciences items, and then Physical Sciences items. Two item clusters and four stand-alone items represent each discipline. Within a discipline, item clusters and stand-alone items were presented in order of average difficulty. Within each item cluster or stand-alone item, scoring assertions were also ordered by difficulty. Easier assertions are those that the most students were able to demonstrate, and difficult assertions are those that the fewest students were able to demonstrate. Across all items, this was generally not the case; for example, the most difficult assertion of an item presented early on in the OSAB was typically more difficult than the easiest assertion of the next item in the OSAB. That is, the order of assertions in Figure 7 represents the order of presentation to the panelists, but assertions were not ordered by overall difficulty across all items.

Not all items have assertions that will map onto all performance levels. For example, an item cluster may have assertions that map onto “Does Not Meet Standard,” “Approaches Standard,” and “Meets Standard,” but not “Exceeds Standard.” Item clusters may have as few as four assertions or as many as 20 assertions. Each assertion is worth one score-point.

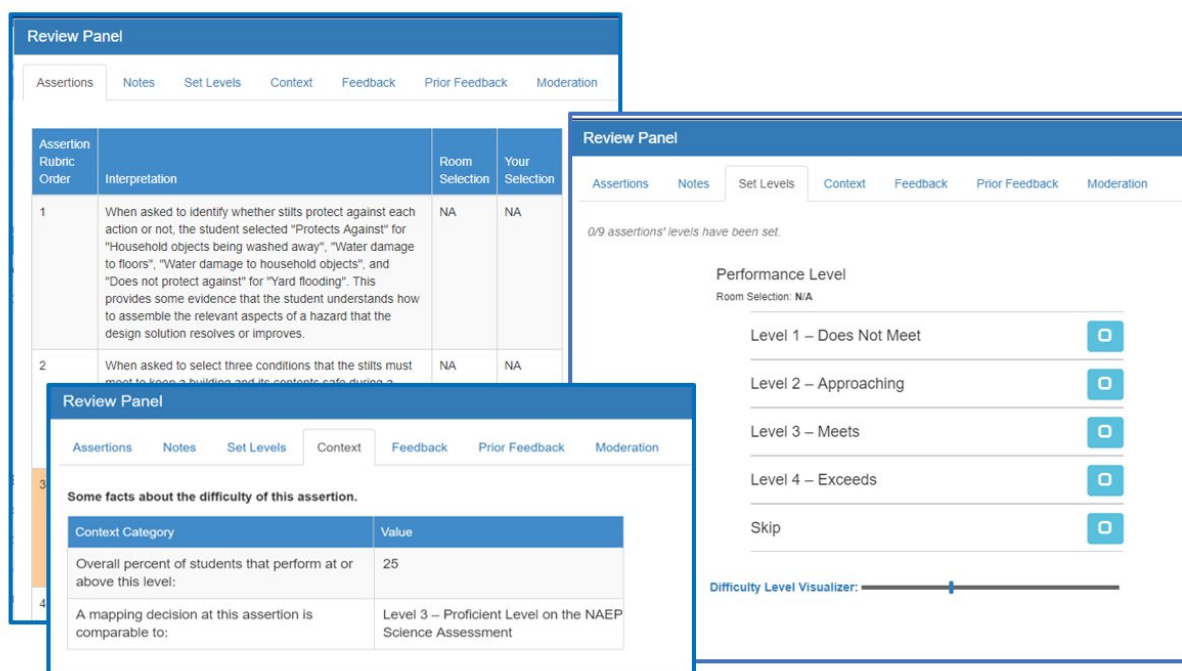
Each OSAB contains three disciplines and 18 items (item clusters and stand-alone items). The grade 5 OSAB contained 70 assertions, the grade 8 OSAB contained 76 assertions, and the grade 11 OSAB contained 80 assertions. Each OSAB was comprised of six item clusters and 12 stand-alone items.

5.5 WORKSHOP TECHNOLOGY

The standard-setting panelists used AIR’s online application for standard setting. Each panelist used an AIR laptop or Chromebook on which they took the test, reviewed item clusters and stand-alone items and ancillary materials, and mapped assertions to performance levels.

Using tabs in the review panel of the tool (see Figure 8), panelists could review the items and scoring assertions, they could determine the relative difficulty of assertions to other assertions in the same item, examine the content alignment of each item (via the alignment of the assertions within an item, which all align to the same performance expectation), assign assertions to performance levels, add notes and comments on the assertions as they reviewed them, and review context and benchmark data. Additionally, they had access to a difficulty visualizer, a graphic representation of the difficulty of each assertion relative to the all other assertions in the OSAB (not just within the item). Panelists also reviewed their own assertion placement, their table’s placement, the other tables’ placement, and the overall placement for all tables.

Figure 8. Example Features in Standard-Setting Tool



Two full-time AIR information technology specialists oversaw laptop setup and testing, answered questions, and ensured that technological processes ran smoothly and without interruption throughout the meeting.

5.6 EVENTS

The standard-setting workshop occurred over a period of two days. Table 7 summarizes each day’s events, and this section describes each event listed in greater detail. Appendix 3-C, Standard-Setting Workshop Agenda, provides the full workshop agenda.

Table 7. Standard-Setting Agenda Summary

Day 1: Wednesday, July 31, 2019
<ul style="list-style-type: none"> • Table leader orientation • Large-group introductory training • Take the test • PLD review • Create threshold PLDs • OSAB review
Day 2: Thursday, August 1, 2019
<ul style="list-style-type: none"> • OSAB review (continued) • Assertion-mapping training • Round 1 assertion mapping; feedback, context data, benchmark data, and articulation review and discussion • Round 2 assertion mapping; feedback, context data, benchmark data, and articulation review and discussion • Workshop evaluation and debrief

5.6.1 Table Leader Orientation

Table leaders met as a group early in the morning of the first day for briefing on the constructs, processes, and technologies used in standard setting. The objective of the training was to ensure everyone followed a standardized process across all grade panels.

Table leaders provided the following throughout the workshop:

- Help panelists see the “big picture”
- Lead table discussions
- Support panelists with tasks
- Monitor security of materials
- Monitor panelist understanding and reported issues or misunderstandings to room facilitators
- Maintain a supportive atmosphere of professionalism and respect

In addition to these responsibilities, table leaders also served as panelists and set individual cut scores.

Appendix 3-D, Standard-Setting Training Slides, provides the slides used during the table leader orientation.

5.6.2 Large-Group Introductory Training

Abe Krisst from the CSDE welcomed panelists to the workshop and provided context and background. He outlined the roles and responsibilities of the three groups of participants at the workshop: panelists, AIR staff, and CSDE personnel. Dr. Ahadi then oriented participants to the workshop by describing the purpose and objectives of the meeting, explaining the process to be implemented to meet those objectives, and outlining the events that would happen each day. He explained that panelists were selected because they were experts, and how the process to be implemented over the two days was designed to elicit and apply their expertise to recommend new cut scores. Finally, he described how standard setting works and what would happen once the panelists had finalized their recommendations. Appendix 3-D, Standard-Setting Training Slides, provides the slides used during the large-group training.

5.6.3 Confidentiality and Security

Workshop leaders and room facilitators addressed confidentiality and security during orientation and again in each room. Standard setting uses live science test items from the operational NGSS test, requiring confidentiality to maintain their security. Participants were not to do any of the following during or after the workshop:

- Discuss the test items outside of the meeting
- Remove any secure materials from the room on breaks or at the end of the day
- Discuss judgments or cut scores (their own or others') with anyone outside of the meeting
- Discuss secure materials with non-participants
- Use cell phones in the meeting rooms
- Take notes on anything other than provided materials
- Bring any other materials into the workshop

Participants could have general conversations about the process and days' events, but workshop leaders warned them against discussing details, particularly those involving test items, cut scores, and any other confidential information.

5.6.4 Take the Test

Following the large-group introductory training, participants broke out into their separate grade-level rooms. As their introduction to the standard-setting process, panelists took a form of the test that students took in 2019, in the grade level to which they would be setting performance standards. They took the tests online via the same tool used to deliver operational tests to students, and the testing environment closely matched that of students when they took the test.

Taking the same test as students take provides the opportunity to interact with and become familiar with the test items and the look and feel of the student experience while testing. They could score their responses and had 90 minutes to interact with the test.

5.6.5 Range Performance-Level Descriptor Review

After taking the test, panelists completed a thorough review of the PLDs for their assigned grade. Panelists identified key words describing the skills necessary for performance at each level and discussed the skills and knowledge that differentiated performance in each of the four levels. Tables discussed separately at first and then joined for an all-grade discussion.

Reviewing the PLDs ensured that participants understood what students in Connecticut should know and be able to do and how much knowledge and skill students are expected to demonstrate at each level of performance.

5.6.6 Create Threshold Performance-Level Descriptors

After reviewing and discussing the range PLDs, panelists worked in their grade-level groups to draft threshold PLDs that describe the skills that students just barely able to score in one performance level have but that students scoring just below the performance level do not have. Looking at each PLD, panelists identified the skills needed to just barely perform at that level and noted this in a worksheet. The following two questions guided the process:

- What skills and knowledge must the student demonstrate to qualify for entrance into this performance level?
- How does this differ from the upper range of the adjacent (lower) performance level?

After each table drafted threshold PLDs, panelists discussed them across all tables.

The point of this exercise was for panelists to consider and define the knowledge and skills that differentiate the bottom of each performance level from the top of the previous performance level. Panelists, working across tables, drafted descriptions for “Meets,” “Exceeds,” and “Approaching.”

5.6.7 Ordered Scoring Assertion Booklet Review

After completing the threshold PLDs, panelists independently reviewed the item clusters, stand-alone items, and assertions in the OSAB. They took notes on each assertion to document the interactions required by each and described why an assertion might be more or less difficult than the previous assertion within the item. They also noted how each assertion related to the PLDs.

After reviewing the item interactions and scoring assertions individually, panelists engaged in discussion with table members about the skills required and relationships among the reviewed test materials and performance levels. This process ensured that panelists built a solid understanding of how the scoring assertions relate to the item interactions and how the items relate to the PLDs, and also helped to facilitate a common understanding among workshop panelists.

5.6.8 Assertion-Mapping Training

After reviewing the entire OSAB, facilitators described the processes for mapping assertions and determining cut scores. They explained that the objective of standard setting is aspirational; to identify what all students should know and be able to do, and not to describe what they currently know and can do.

Panelists were to match each assertion to the performance level best supported by the assertion using the PLDs, the difficulty visualizer (described in Section 5.5, Workshop Technology), their notes from the OSAB review, and their professional judgments. Figure 9 graphically describes the assertion-mapping process.

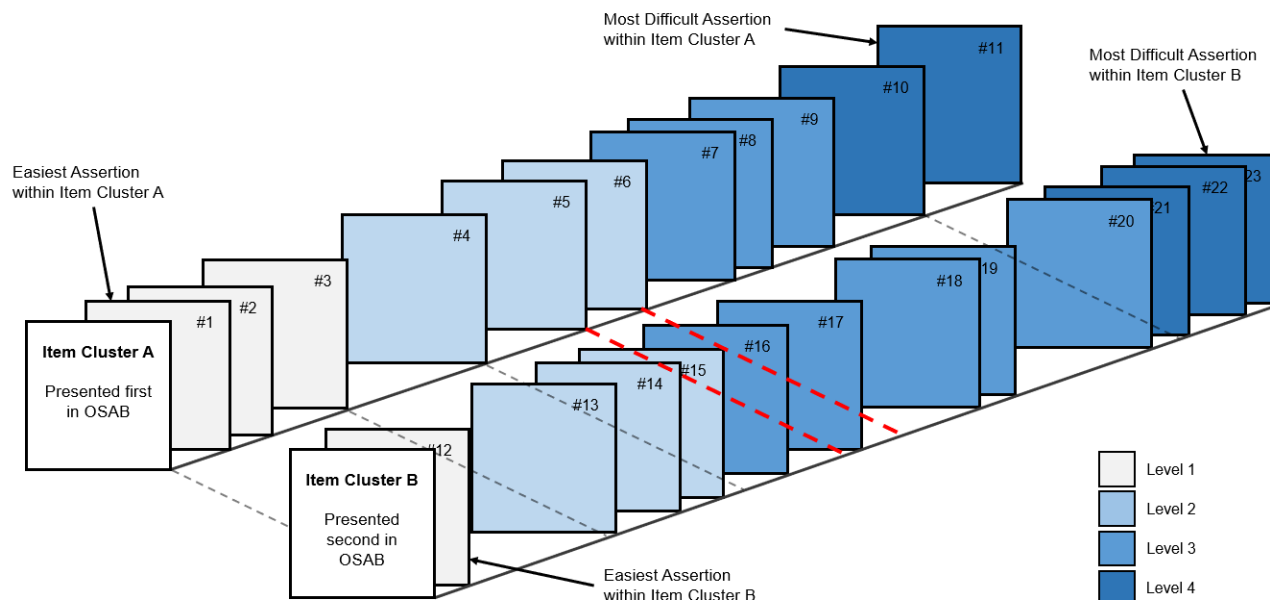
Facilitators provided the following process to guide the mapping of assertions onto PLDs:

1. How does the student interaction give rise to the assertion? Did they plot, select, or write something?
2. Why is this assertion more difficult to achieve than the previous one?
3. Which PLD most ably describes this assertion?

It was emphasized that assertions within an item were ordered by difficulty, and therefore, the assigned performance levels should be ordered, as well. Within each item, panelists were not allowed to place an assertion into a lower performance level than the level at which the previous assertions had been placed. If panelists felt very strongly that an assertion was out of order in the OSAB, they were asked to skip (not assign any performance level to) the assertion. However, this was to be used as a last resort.

Because the assertion mapping was conducted separately for each item, it was possible that there was no perfect ordering of the assigned levels of the assertions across all items as a function of assertion difficulty. It was allowed (and it occurred frequently) that an assertion of one item had a higher difficulty but lower assigned performance level than another assertion from a different item. For example, in Figure 9, the difficulty of the assertion on page 6 of item cluster A (“Level 2”) has a higher difficulty than the assertion on page 17 of item cluster B (“Level 3”). However, it was expected for the higher performance levels to be assigned more frequently with increasing assertion difficulty across items. Appendix 3-D, Standard-Setting Training Slides, provides the training slides used during the breakout room training.

Figure 9. Example of Assertion Mapping



Note. Figure 9 describes scoring assertion mapping across two item clusters, where the assertions on pages 1, 2, 3, and 12 are mapped onto level 1; the assertions on pages 4, 5, 6, 13, 14, and 15 are mapped onto level 2; the assertions on pages 7, 8, 9, 16, 17, 18, 19, and 20 are mapped onto level 3; and the assertions on pages 10, 11, 21, 22, and 23 are mapped onto level 4.

5.6.9 Practice Quiz

Panelists completed a practice quiz prior to beginning a practice round. The quiz assessed panelists' understanding in multiple ways. They must be able to

- Describe where “Just Barely” students fall on a performance scale.
- Indicate on a diagram how performance standards define performance levels.
- Identify more- and less-difficult scoring assertions in the OSAB.
- Answer questions about the assertion-mapping process and online application.

Room facilitators reviewed the quizzes with the panelists and provided additional training for incorrect responses on the quiz. Appendix 3-E, Standard-Setting Practice Quiz, provides the quiz that panelists completed.

5.6.10 Practice Round

Following the practice quiz, panelists practiced mapping assertions to PLDs in a short practice OSAB. The purpose of the practice round was to ensure that panelists were comfortable with the technology, items, item interactions, and scoring assertions prior to mapping any assertions in the OSAB. Panelists asked questions, and the room facilitators provided clarifications and further instructions until everyone had successfully completed the practice round.

5.6.11 Readiness Assertion

After completing the practice round, and prior to mapping assertions in Round 1, panelists completed a readiness assertion form. On this form, panelists asserted that their training was sufficient for them to understand the following concepts and tasks:

- The concept of a student who just barely meets the criteria described in the PLDs
- The structure, use, and importance of the OSAB
- The process to determine and map assertions to PLDs in the standard-setting tool
- Readiness to begin the Round 1 task

The readiness form for Round 2 focused on affirming understanding of the context and benchmark data supplied after Round 1. On this form, all panelists affirmed the following:

- Understanding the context data
- Understanding the feedback data
- Understanding the Round 2 task
- Readiness to complete the Round 2 task

Room facilitators reviewed the readiness forms and provided additional training to panelists not asserting understanding or readiness. However, every panelist affirmed readiness before mapping assertions in both rounds of the workshop. Appendix 3-F, Standard-Setting Readiness Forms, provides the forms that panelists completed.

5.7 ASSERTION MAPPING

Panelists mapped assertions independently, using the PLDs, their notes from reviewing each assertion, and the difficulty visualizer to place each of the assertions into one of the four performance levels.

5.7.1 Calculating Cut Scores from the Assertion Mapping

A propriety algorithm utilized RP67 (for grades 5 and 8) and RP50 (for grade 11) to minimize misclassifications to calculate cut scores based on the assertion mappings.⁵ Each cut score was defined as the score point that minimized the weighted number of discrepancies between the mappings implied by the cut score and the observed mappings. The weights were defined as the inverse of the observed frequencies of each level. For each cut score, only the assertion mappings for the two adjacent levels were considered (e.g., for the second cut, only the assertions that were

⁵ Typically, the probability used in standard setting is .67 (“RP67” [Huynh, 1994]). RP67 is the assertion difficulty point where 67% of the students would earn the score point. The reason to adopt RP50 for grade 11 was because the difficulty of most items exceeded students’ abilities. RP50 better aligned with the PLD and therefore led to more appropriate performance cut scores. Using the RP50 prevented panelists from mapping the first cut score onto the lowest-difficulty assertions on the test. This approach has been taken by other high-stakes tests, such as the Smarter Balanced Assessments (see Cizek & Koons, 2014).

mapped onto the levels “Approaching” and “Meets” were used). Specifically, let n_k be the number of assertions put at performance level k , t_k be the cut to be estimated, d_i be the assigned performance level, and θ_i be the RP value of the i th assertion. For each assertion placed at levels k and $k + 1$, define the misclassification indicator as

$$z_{ik}|t_k = \begin{cases} 1 & \text{if } (d_i = k \text{ and } t_k \leq \theta_i) \text{ or } (d_i = k + 1 \text{ and } t_k > \theta_i) \\ 0 & \text{otherwise} \end{cases}.$$

The cut t_k is then estimated by minimizing a loss function based on the weighted number of misclassifications

$$\arg \min_{t_k} \left(\frac{1}{n_k} \sum_{i \in \{d_i=k\}} z_{ik}|t_k + \frac{1}{n_{k+1}} \sum_{i \in \{d_i=k+1\}} z_{ik}|t_k \right).$$

Unlike the Bookmark method, the cut scores for a table or room were not the median value of the cut scores of the individual panelists. Instead, cut scores at the table and grade level were computed using the same method but taking into account the assigned levels of all the raters at the table and in the room, respectively. Applying these cut scores to the 2019 test data created data describing the percentage of students falling into each performance level. This algorithm calculated cut scores from the assertion maps by panelist, table, and for the room.

5.7.2 Feedback and Impact Data

Feedback included the cut scores corresponding to the assertion mappings for each panelist, for each table, and for the room overall (across all three tables). In addition, panelists were shown impact data based on the cut scores resulting from their assertion mappings. Impact data were defined for panelists as the percentages of students who would reach or exceed each of the performance standards given the assertion mappings. Percentages were calculated using the student data from the 2019 NGSS administration. This information allowed panelists to compare their mappings to other panelist’s mappings to evaluate the impact they might have.

Feedback also included review of a variance monitor, part of AIR’s online standard-setting tool, that color-codes the variance of assertion classifications. For all assertions, the variance monitor shows the performance level to which each panelist assigned the assertion. The tool highlights assertions that panelists have assigned to different performance levels. Room facilitators and panelists reviewed and discussed the assertions with the most variable mappings.

5.7.3 Context Data

Panelists were provided with additional context data to inform their Round 2 assertion mappings. Context data included the percentage of students who performed at or above the level associated for each of the assertions in the OSAB. Specifically, the context data for an assertion is defined as the percentage of students who performed at or above the specified RP value associated with the assertion.

5.7.4 Benchmark Data

To be adoptable, performance standards for a statewide system must be coherent across grades and subjects. There should be no irregular peaks and valleys and they should be orderly across subjects

with no dramatic differences in expectation. The following are characteristics of well-articulated standards:

- The cut scores for each performance level increase smoothly with each increasing grade.
- The cut scores should result in a reasonable percentage of students at each performance level; reasonableness can be determined by the percentage of students in the performance levels on historical tests, or contemporaneous tests measuring the same or similar content.
- Barring significant content standard changes (e.g., major changes in rigor), the percentage proficient on new tests should not be radically different from the percentage proficient on historical tests.

Panelists used benchmark data to ensure their recommendations were well articulated. The 2018 grades 5 and 8 Smarter Balanced Assessment, the 2018 SAT (for grade 11), and the 2015 National Assessment of Educational Progress (NAEP) science scores provided benchmark data.⁶ By comparing the results of each round against the percentage proficient on the benchmark tests, it was possible to judge the reasonableness of the proposed performance standards.

Comparing the results of Round 1 against the benchmark data, panelists could see how the proposed standards for the NGSS science assessment compare to those for the existing mathematics and ELA assessments and judge the reasonableness and rigor of the proposed performance standards for the new test. Panelists discussed this information and the impact that the Round 1 cut scores may have on Connecticut students before mapping the Round 2 assertions.

5.8 WORKSHOP RESULTS

The AIR online standard-setting tool automatically computes the results and impact data for each round, and then AIR room facilitators and psychometricians present the Round 1 results for each grade.

5.8.1 Round 1

Table 8 presents the performance standards and associated impact data from Round 1.

Table 8. Round 1 Results

Grade and Table	Cut Scores			Impact Data		
	<i>A</i>	<i>M</i>	<i>E</i>	<i>A</i>	<i>M</i>	<i>E</i>
Grade 5	465	493	522	87	60	25
Table 1	478	502	519	76	49	28
Table 2	478	499	522	76	52	25
Table 3	465	484	522	87	70	25
Grade 8	783	798	842	69	52	9

⁶ The National Assessment of Educational Progress (NAEP) provides state-level benchmark data in science for grade 8; benchmark data for grade 5 are interpolated, and for grade 11, they are extrapolated.

Grade and Table	Cut Scores			Impact Data		
	A	M	E	A	M	E
Table 1	783	798	826	69	52	21
Table 2	781	800	836	71	50	13
Table 3	776	806	842	77	43	9
Grade 11	1064	1099	1132	90	48	16
Table 1	1081	1098	1132	70	49	16
Table 2	1061	1092	1127	93	56	20
Table 3	1069	1104	1141	85	42	11

Note. The grade-level row summarizes the room data (all of the mappings across the three tables). Impact data describes the percentage of students falling at or above each of the performance levels based on the recommended Round 1 cut scores. Performance level: Approaching (A), Meets (M), and Exceeds (E).

After reviewing the feedback and impact data, workshop facilitators provided panelists with additional instructions for completing Round 2. They described the goal of Round 2 as one of convergence but not consensus on a common performance standard. Each table then spent time reviewing and discussing assertion mappings. After completing these discussions, panelists again worked through the OSAB, mapping assertions for Round 2.

5.8.2 Round 2

Table 9 presents the recommended performance standards and associated impact data for Round 2.

Table 9. Round 2 Results

Grade and Table	Cut Scores			Impact Data		
	A	M	E	A	M	E
Grade 5	465	493	525	87	60	22
Table 1	478	502	521	76	49	26
Table 2	480	499	527	74	52	20
Table 3	465	495	522	87	57	25
Grade 8	783	798	842	69	52	9
Table 1	783	798	842	69	52	9
Table 2	779	798	842	73	52	9
Table 3	779	806	842	73	43	9
Grade 11	1078	1099	1141	74	48	11
Table 1	1081	1101	1141	70	45	11
Table 2	1078	1099	1132	74	48	16
Table 3	1077	1107	1141	75	38	11

Note. The grade-level row summarizes the room data (across the three tables). Impact data describe the percentage of students falling at or above each of the performance levels based on the recommended Round 2 cut scores. Performance level: A = Approaching, M = Meets, and E = Exceeds.

Given the recommended Round 2 cut scores, for all grades, between 48% and 60% of students would meet the recommended standard, between 9% and 22% would exceed the standard, and between 69% and 87% would approach the standard. Figure 10 represents those values graphically.

Figure 10. Percentage of Students Reaching or Exceeding Each Recommended Science Performance Standard in 2019

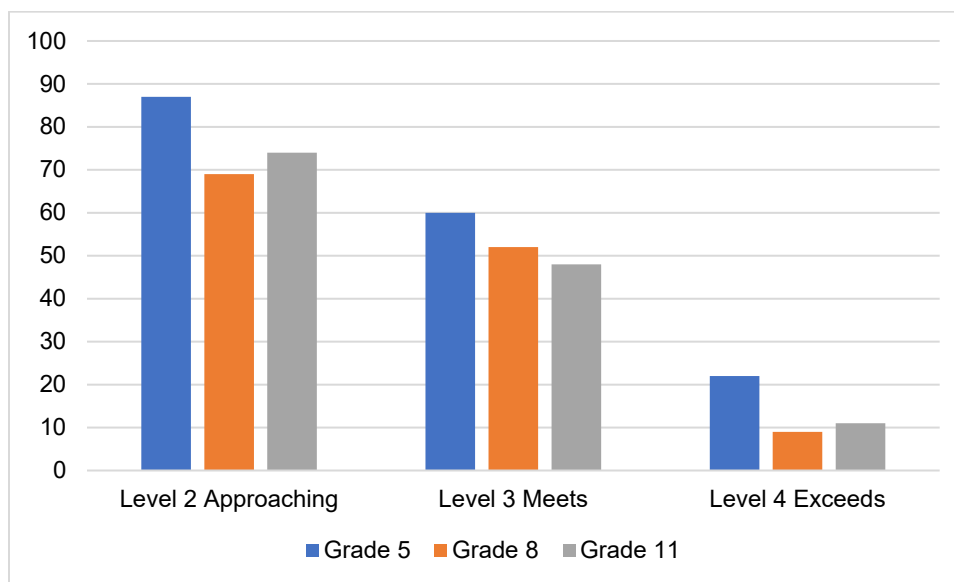
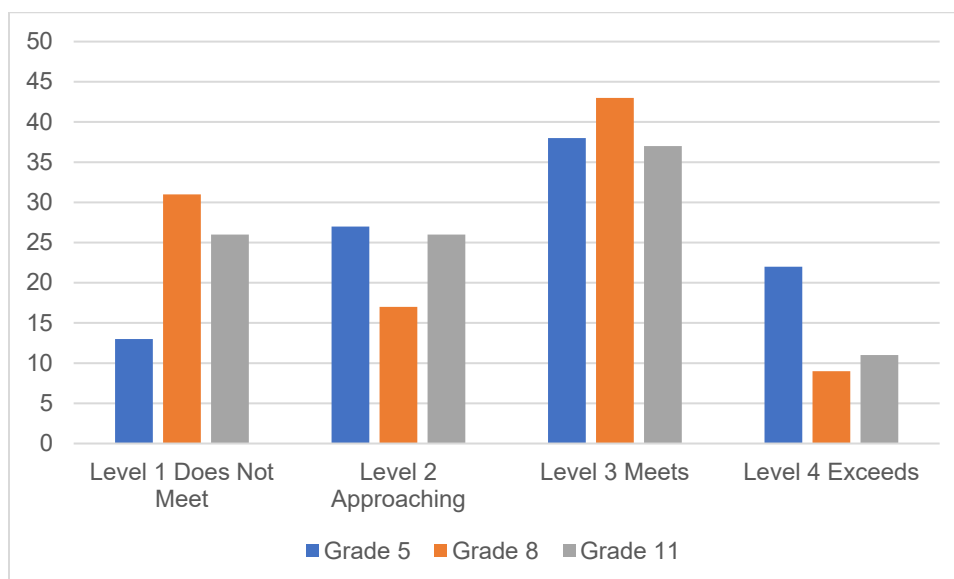


Table 10 indicates the percentage of students classified within each of the performance levels in 2019. The values are displayed graphically in Figure 11.

Table 10. Percentage of Students Classified Within Each Recommended Science Performance Level in 2019

Grade	Level 1 Does Not Meet	Level 2 Approaching	Level 3 Meets	Level 4 Exceeds
5	13	27	38	22
8	31	17	43	9
11	26	26	37	11

Figure 11. Percentage of Students Classified Within Each Recommended Science Performance Level in 2019



5.9 POST WORKSHOP REFINEMENTS

Following the workshop, CSDE reviewed and made some refinements to five of the nine the workshop recommendations. These refinements were all less than the mean of one standard error of measurement for students achieving in the four performance levels for each of the three grades. These refinements were conducted to:

- Ensure greater comparability in the distribution of student performance in each of the four levels, across the three grades;
- Maintain reasonableness and alignment of student performance with other performance data for those same students (i.e., Smarter Balanced and CT SAT School Day) in both English language arts and mathematics; and
- Facilitate alignment and communication of results within the context of the state’s Next Generation Accountability System.

Table 11 presents the final performance standards that were presented to the CSDE’s Technical Advisory Committee for discussion and input and subsequently accepted by the CSDE. These final academic performance standards were formalized through their inclusion in the Online Reporting System (ORS) portal and in a reporting FAQ that was communicated to all educators at the time of results release through the [October 2019 issue of the Student Assessment Newsletters](#). Figure 12 represents those values graphically.

Table 11. Post-Standard-Setting Workshop: Final Cut Scores (Change from Workshop Recommendation) and Impact Data

Grade	Cut Scores (Revision)			Impact Data		
	A	M	E	A	M	E
5	468 (+3)	498 (+5)	535 (+10)	85	54	13
8	772 (–11)	798	842	81	52	9
11	1073 (–5)	1099	1141	80	48	11

Note. Performance level: A = Approaching, M = Meets, and E = Exceeds.

Figure 12. Post-Standard-Setting Workshop: Percentage of Students Reaching or Exceeding Each Science Performance Standard in 2019

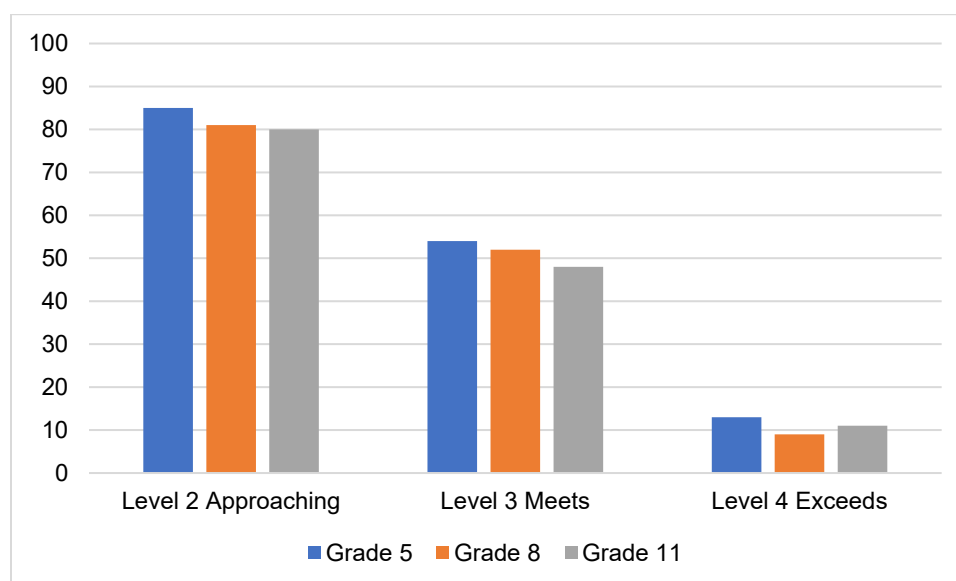
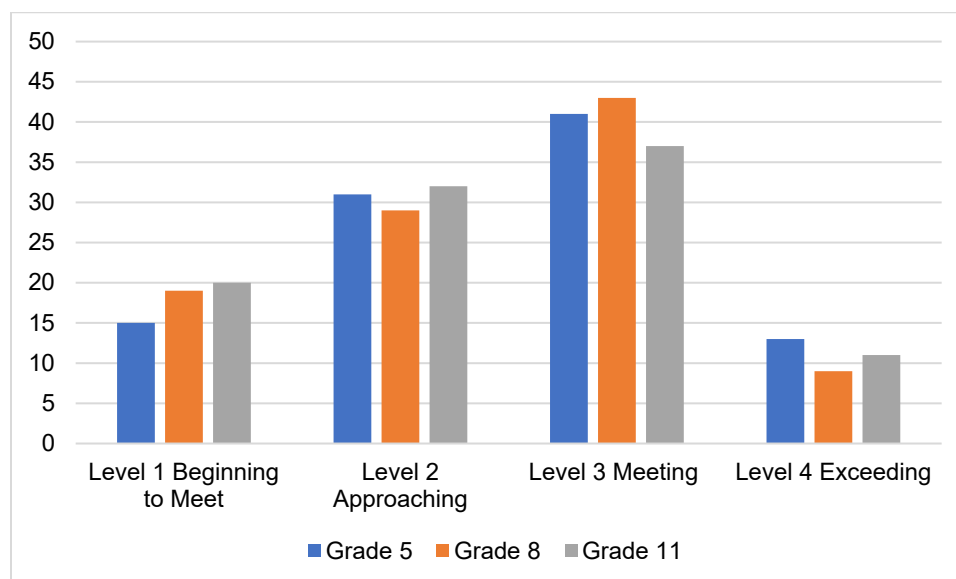


Table 12 indicates the percentage of students classified within each of the performance levels in 2019 proceeding from CSDE refinements to the recommended performance standards. The values are displayed graphically in Figure 13.

Table 12. Post-Standard-Setting Workshop: Percentage of Students Classified Within Each Science Performance Level in 2019

Grade	Level 1 Does Not Meet	Level 2 Approaching	Level 3 Meets	Level 4 Exceeds
5	15	31	41	13
8	19	29	43	9
11	20	32	37	11

Figure 13. Post-Standard-Setting Workshop: Percentage of Students Classified Within Each Science Performance Level in 2019



5.10 WORKSHOP EVALUATIONS

After finishing all activities, panelists completed online meeting evaluations independently, in which they described and evaluated their experience taking part in the standard setting. Tables 13 through 17 summarize the results of the evaluations. Evaluation items endorsed by fewer than 90% of panelists are discussed in text, and the least endorsed items are discussed in terms of the number and type of response. Three panelists left the workshop without completing an evaluation, so while the number of panelists was 42, the number of responses to the evaluation was 39.

Workshop participants indicated clarity in the instructions, materials, data, and process (see Table 13). A few grade 5 and grade 8 panelists reported some lack of clarity with the PLDs, while some grade 8 and grade 11 panelists reported the same with the context data.

Table 13. Evaluation Results: Clarity of Materials and Process

Please rate the clarity of the following components of the workshop.	Percentage “Somewhat Clear” or “Very Clear”			
	Grade 5	Grade 8	Grade 11	Overall
Instructions provided by the workshop leader	100%	100%	100%	100%
Performance-level descriptors (PLDs)	79%	83%	100%	87%
Ordered Scoring Assertion Booklet (OSAB)	100%	100%	100%	100%
Panelist agreement data	100%	100%	100%	100%
Context data (percentage of students who would reach any standard you select)	100%	83%	85%	90%

Note. Number of responses = 39 (grade 5 responses = 14, grade 8 responses = 12, and grade 11 responses = 13). Evaluation options included “Very Unclear,” “Somewhat Unclear,” “Somewhat Clear,” and “Very Clear.”

Participants felt they had sufficient time to complete all activities. In fact, some indicated having too much time to complete some tasks (see Table 14). Some panelists (n=14) indicated that the large group training was too long and that there was both too much (n=9) and too little (n=6) time devoted to PLD review, too much (n=1) and too little (n=4) time to experience the test, and too much (n=7) and too little (n=2) time to review the OSABs. Three panelists each indicated having too much and too little time to map their assertions, and in grade 8, one panelist indicated wanting more time for the Round 1 discussion, while another indicated wanting less time.

Table 14. Evaluation Results: Appropriateness of Process

How appropriate was the amount of time you were given to complete the following components of the standard-setting process?	Percentage responding “About Right”			
	Grade 5	Grade 8	Grade 11	Overall
Large-group orientation	57%	67%	69%	64%
Experiencing the online assessment	86%	100%	77%	87%
Reviewing the Performance-Level Descriptors (PLDs)	71%	58%	54%	62%
Reviewing the Ordered Scoring Assertion Booklet (OSAB)	71%	75%	85%	77%
Mapping your scoring assertions to performance levels in each round	79%	83%	92%	85%
Round 1 discussion	93%	83%	92%	90%

Note. Number of responses = 39 (grade 5 responses = 14, grade 8 responses = 12, and grade 11 responses = 13). Evaluation options included “Too Little,” “Too Much,” and “About Right.”

Participants appreciated the importance of the multiple factors contributing to assertion mapping, with nearly all participants rating each factor as important or very important (see Table 15). Two grade 5 panelists indicated the PLDs were not important, while two grade 8 panelists reported that their perception of item difficulty was not important.

Table 15. Evaluation Results: Importance of Materials

How important were each of the following factors in your mapping of scoring assertions to performance levels?	Percentage responding “Somewhat Important” or “Very Important”			
	Grade 5	Grade 8	Grade 11	Overall
Performance-Level Descriptors (PLDs)	86%	92%	100%	92%
Your perception of the difficulty of the scoring assertions and items in general	93%	83%	100%	92%
Your experience with students	100%	92%	100%	97%
Discussions with other panelists	100%	100%	100%	100%
External benchmark data	100%	92%	100%	97%
Room agreement data (room, table, and individual cuts)	100%	100%	92%	97%
Context data (percentage of students who would reach any standard you select)	100%	100%	100%	100%

Note. Number of responses = 39 (grade 5 responses = 14, grade 8 responses = 12, and grade 11 responses = 13). Evaluation options included “Not Important,” “Somewhat Important,” and “Very Important.”

Participant understanding of the workshop processes and tasks was high (see Table 16). The least agreed with statement in Table 16 related to the expectations described by the PLDs. A total of nine panelists disagreed with this statement.

Table 16. Evaluation Results: Understanding Processes and Tasks

At the end of the workshop, please rate your agreement with the following statements.	Percentage “Agree” or “Strongly Agree”			
	Grade 5	Grade 8	Grade 11	Overall
I understood the purpose of this standard-setting workshop.	100%	100%	100%	100%
The procedures used to recommend performance standards were fair and unbiased.	100%	92%	100%	97%
The training provided me with the information I needed to recommend performance standards.	100%	100%	100%	100%
Taking the online assessment helped me to better understand what students need to know and be able to do to answer each question.	100%	100%	100%	100%
The Performance-Level Descriptors (descriptions of what students within each performance level are expected to know and be able to do) provided a clear picture of expectations for student performance at each level.	86%	67%	77%	77%
I understood how to review each assertion in the Ordered Scoring Assertion Booklet (OSAB) to determine what students must know and be able to do to answer each assertion correctly.	100%	100%	100%	100%
I understood how to map assertions to the most apt performance level.	100%	100%	100%	100%
I found the benchmark data and discussions helpful in my decisions about the assertions I mapped to performance levels.	100%	100%	100%	100%
I found the context data (percentage of students that would achieve at the level indicated by the assertion difficulty) and discussions helpful in my decisions about the assertions I mapped to performance levels.	100%	83%	92%	92%
I found the panelist agreement data (room, table, and individual cuts) and discussion helpful in my decisions about assertions I mapped to performance levels.	100%	100%	100%	100%
I felt comfortable expressing my opinions throughout the workshop.	100%	100%	100%	100%
Everyone was given the opportunity to express their opinions throughout the workshop.	100%	100%	100%	100%

Note. Number of responses = 39 (grade 5 responses = 14, grade 8 responses = 12, and grade 11 responses = 13). Evaluation options included “Strongly Disagree,” “Disagree,” “Agree,” and “Strongly Agree.”

Participants agreed that the standards set during the workshop reflected the intended grade-level expectations (see Table 17); however, three grade 8 panelists did not agree with the Level 2 statement.

Table 17. Evaluation Results: Student Expectations

Please read the following statement carefully and indicate your response.	Percentage Indicating “Agree” or “Strongly Agree”			
	Grade 5	Grade 8	Grade 11	Overall
A student performing at Level 2 is approaching the performance expectations for the grade.	100%	75%	100%	92%
A student performing at Level 3 meets the performance expectations for the grade.	100%	100%	100%	100%
A student performing at Level 4 exceeds the performance expectations for the grade.	100%	100%	100%	100%

Note. Number of responses = 39 (grade 5 responses = 14, grade 8 responses = 12, and grade 11 responses = 13). Evaluation options included “Strongly Disagree,” “Disagree,” “Agree,” and “Strongly Agree.”

5.10.1 Workshop Participant Feedback

Finally, panelists responded to two open-ended questions: “What suggestions do you have to improve the training or standard-setting process?” and “Do you have any additional comments? Please be specific.”

Thirty-three participants responded to the first question, and 21 to the second. Most responses indicated the training was effective and the process was clear. Participants provided minor suggestions, such as shortening the time allocated for some tasks, lessening the emphasis on Just Barely students adjusting the time allocated to some tasks, revising the PLDs, halting off-topic discussions, or introducing the assertion-mapping task earlier. Many commented on the value of discussions, the helpfulness of the facilitators and table leaders, and the positive interactions with other panelists. Many appreciated the opportunity and indicated it was a useful learning experience for them.

Additional participant comments included:

“Overall this was effective, collaborative, and productive, thanks for that!”

“Very useful training and work. I know this can help me in my district.”

“Thank you to the AIR/CSDE team for their detailed overview and organization of the process.”

6. VALIDITY EVIDENCE

Validity evidence for standard setting is established in multiple ways. First, standard setting should adhere to the standards established by appropriate professional organizations and be consistent with the recommendations for best practices in the literature and established validity criteria. Second, the process should provide the evidence required of states to meet federal peer review requirements. We describe each of these in the following sections.

6.1 EVIDENCE OF ADHERENCE TO PROFESSIONAL STANDARDS AND BEST PRACTICES

The Next Generation Science Standards (NGSS) standard-setting workshop was designed and executed consistent with established practices and best-practice principles (Hambleton & Pitoniak, 2006; Hambleton, Pitoniak, & Copella, 2012; Kane, 2001). The process also adhered to the following professional standards recommended 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) related to standard setting:

Standard 5.21: When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be documented clearly.

Standard 5.22: When cut scores defining pass-fail or proficiency levels are based on direct judgments about the adequacy of item or test performances, the judgmental process should be designed so that the participants providing the judgments can bring their knowledge and experience to bear in a reasonable way.

Standard 5.23: When feasible and appropriate, cut scores defining categories with distinct substantive interpretations should be informed by sound empirical data concerning the relation of test performance to the relevant criteria.

The sections of this documentation discussing the rationale and procedures used in the standard-setting workshop address Standard 5.21. The Assertion-Mapping Procedure (AMP) standard-setting procedure is appropriate for tests of this type—with interrelated sets of three-dimensional item clusters and scaled using item response theory (IRT). Section 5.1, The Assertion-Mapping Procedure, provides the justification for and the additional benefits of selecting the AMP method to establish the cut scores; and Section 5.6, Events, through Section 5.7.1, Calculating Cut Scores from the Assertion Mapping, document the process followed to implement the method.

The design and implementation of the AMP procedure address Standard 5.22. The method directly leverages the subject-matter expertise of the panelists placing assertions into performance levels and incorporates multiple, iterative rounds of ratings in which panelists modify their judgments based on feedback and discussion. Panelists apply their expertise in multiple ways throughout the process by

- understanding the test, test items, and scoring assertions (from an educator and student perspective);
- describing the knowledge and skills measured by the test;
- identifying the skills associated with each test item scoring assertion;
- describing the skills associated with student performance in each performance level;
- identifying which test item scoring assertions students at each performance level should be able to answer correctly; and

- evaluating and applying feedback and reference data to the Round 2 recommendations and considering the impact of the recommended cut scores on students.

Panelists' understanding of the AMP was assessed with a quiz prior to the practice round. Additionally, panelists' readiness evaluations provided evidence of a successful orientation to the process and understanding of the process, while their workshop evaluations provide evidence of confidence in the process and resulting recommendations.

The recruitment process resulted in panels that were representative of important regional and demographic groups who were knowledgeable about the subject area and students' developmental level. Section 5.3.4, Educator Participants, summarizes details about the panel demographics and qualifications.

The provision of benchmark and context data to panelists after Round 1 addresses Standard 5.23 (see Section 5.7.3, Context Data, and Section 5.7.4, Benchmark Data). This set of empirical data provides necessary and additional context describing student performance given the recommended standards.

6.2 EVIDENCE IN TERMS OF PEER REVIEW CRITICAL ELEMENTS

The U.S. Department of Education (USDOE) provides guidance for the peer review of state assessment systems. This guidance is intended to support states in meeting statutory and regulatory requirements under Title I of the Elementary and Secondary Education Act of 1965 (ESEA; USDOE, 2015). The following critical elements are relevant to standard setting; evidence supporting each element immediately follows.

Critical Element 1.5: Meaningful consultation in the development of challenging state standards and assessments.

Connecticut educators played a critical role in establishing performance levels for the NGSS tests. They created the item clusters, reviewed and revised the PLDs, mapped assertions to performance levels to delineate performance at each performance level, considered benchmark data and the impact of their recommendations, and formally recommended performance standards.

Many subject-matter experts contributed to developing Connecticut's performance standards. Contributing educators were subject-matter experts in their content area, in the content standards and curriculum that they teach, and in the developmental and cognitive capabilities of their students. AIR's facilitators were subject-matter experts in the subjects tested and in facilitating effective standard-setting workshops. The psychometricians performing the analyses and calculations throughout the meeting were subject-matter experts in the measurement and statistics principles required of the standard-setting process.

Critical Element 6.2: Achievement standards setting. The state used a technically sound method and process that involved panelists with appropriate experience and expertise for setting its academic performance standards.

Evidence to support this critical element includes:

- 1) The rationale for and technical sufficiency of the AMP method selected to establish performance standards (Section 5.1, The Assertion-Mapping Procedure).

- 2) Documentation that the method used for setting cut scores allowed panelists to apply their knowledge and experience in a reasonable manner and supported the establishment of reasonable and defensible cut scores (Section 5.6, Events; Section 5.6.2, Large-Group Introductory Training; Section 5.7, Assertion Mapping; Section 5.8, Workshop Results; and Section 6.1, Evidence of Adherence to Professional Standards and Best Practices).
- 3) Panelists' self-reported readiness to undertake the task (Section 5.6.11, Readiness Assertion) and confidence in the workshop process and outcomes (Section 5.10, Workshop Evaluations) supporting the validity of the process.
- 4) The standard-setting panels consisted of panelists with appropriate experience and expertise, including content experts with experience teaching Connecticut's science content standards, and individuals with experience and expertise teaching special population and general education students in Connecticut (Section 5.3.4, Educator Participants; and Appendix 3-A, Standard-Setting Panelist Characteristics).

7. REFERENCES

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: Author.
- Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, 64, 153–168.
- Cizek, G. J., & Koons, H. (2014). Observation and Report on Smarter Balanced Standard Setting: October 12–20, 2014. Accessed from <https://portal.smarterbalanced.org/library/en/standard-setting-observation-and-report.pdf>.
- Ferrara, S., & Lewis, D. M. (2012). The item-descriptor (ID) matching method. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 255–282). New York: Routledge.
- Gibbons, R. D., & Hedeker, D. R. (1992). Full-information bi-factor analysis. *Psychometrika*, 57, 423–436.
- Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433–470). Westport, CT: Praeger.
- Hambleton, R. K., Pitoniak, M. J., & Copella, J. M. (2012). Essential steps in setting performance standards on educational tests and strategies for assessing the reliability of results. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 47–76). New York: Routledge.
- Huynh, H. (1994, October). Some technical aspects in standard setting. In *Proceedings of the Joint Conference on Standard Setting for Large Scale Assessment Programs* (co-sponsored by National Assessment Governing Board and National Center for Education Statistics), Washington, DC, October 5–7, 1994, pp. 75–91.
- Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 53–88). Mahwah, NJ: Lawrence Erlbaum.
- Kingston, N. M., Kahl, S. R., Sweeney, K. P., & Bay, L. (2001). Setting performance standards using the body of work method. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 219–248). Mahwah, NJ: Lawrence Erlbaum Associates.
- Mitzel, H. C., Lewis, D. M., Patz, R. J., & Greene, D. R. (2001). The Bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249–281). Mahwah, NJ: Lawrence Erlbaum Associates.

- 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. (2010). Formal relations and an empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement*, 47, 361–372.
- Rijmen, F., Cohen, J., Butcher, T., & Farley, D. (2018, June). Scoring and reporting for assessments developed for the new science standards [Symposium]. National Conference on Student Assessment, San Diego, CA.
- Skaggs, G., Hein, S. F., & Awuor, R. (2007). Setting passing scores on passage-based tests: A comparison of traditional and single-passage bookmark methods. *Applied Measurement in Education*, 20, 405–426.
- U. S. Department of Education. (2015). *Non-Regulatory Guidance for States for Meeting Requirements of the Elementary and Secondary Education Act of 1965, as amended*. Washington, D.C. Accessed from <https://www2.ed.gov/policy/elsec/guid/assessguid15.pdf>.

Connecticut Next Generation Science Standards Assessment

2023–2024

Volume 4: Evidence of Reliability and Validity



CONNECTICUT STATE
DEPARTMENT OF EDUCATION

TABLE OF CONTENTS

1.	INTRODUCTION AND OVERVIEW OF RELIABILITY AND VALIDITY EVIDENCE ...	1
1.1	Reliability.....	2
1.2	Validity	3
2.	PURPOSE OF THE CONNECTICUT NEXT GENERATION SCIENCE STANDARDS ASSESSMENT.....	5
3.	RELIABILITY	6
3.1	Standard Error of Measurement.....	7
3.2	Reliability of Performance Classification.....	9
3.2.1	Classification Accuracy	9
3.2.2	Classification Consistency.....	10
3.3	Precision at Cut Scores	11
4.	EVIDENCE OF CONTENT VALIDITY.....	12
4.1	Content Standards	12
4.2	Independent Alignment Study	12
5.	EVIDENCE OF INTERNAL-EXTERNAL STRUCTURE	13
5.1	Correlations Among Discipline Scores.....	13
5.2	Convergent and Discriminant Validity	14
5.3	Cluster Effects.....	17
5.4	Confirmatory Factor Analysis.....	20
5.4.1	Results.....	24
5.4.2	Conclusion	28
6.	FAIRNESS IN CONTENT.....	29
6.1	Cognitive Laboratory Studies	29
6.2	Statistical Fairness in Item Statistics.....	30
7.	SUMMARY.....	30
8.	REFERENCES	31

LIST OF TABLES

Table 1. Spring 2024 Assessment Modes	1
Table 2. Spring 2024 Marginal Reliability Coefficients.....	7
Table 3. Spring 2024 Classification Accuracy Index	10
Table 4. Spring 2024 Classification Consistency Index	11
Table 5. Spring 2024 Performance Levels and Associated Conditional Standard Error of Measurement.....	11
Table 6. Number of Spring 2024 Items for Each Discipline	12
Table 7. Spring 2024 Correlations Among Disciplines	13
Table 8. Spring 2024 Correlations Across Subjects, Grade 5.....	15
Table 9. Spring 2024 Correlations Across Subjects, Grade 8.....	16
Table 10. Spring 2024 Correlations Across ELA, Mathematics, and Science Scores.....	17
Table 11. Numbers of Forms, Clusters per Discipline (Range Across Forms), Assertions per Form (Range Across Forms), and Students per Form (Range Across Forms)	20
Table 12. Guidelines for Evaluating Goodness of Fit.....	24
Table 13. Fit Measures per Model and Form, Grade 6	25
Table 14. Fit Measures per Model and Form, Grade 7	25
Table 15. Fit Measures per Model and Form, Grade 8	26
Table 16. Fit Measures per Model and Form, Grade 6, with One Cluster Removed	27
Table 17. Model-Implied Correlations per Form for the Disciplines in Model 4.....	27

LIST OF FIGURES

Figure 1. Spring 2024 Conditional Standard Errors of Measurement	7
Figure 2. Cluster Variance Proportion for Operational Items in Elementary School.....	18
Figure 3. Cluster Variance Proportion for Operational Items in Middle School.....	19
Figure 4. Cluster Variance Proportion for Operational Items in High School	19
Figure 5. One-Factor Structural Model (Assertions-Overall): “Model 1”	22
Figure 6. Second-Order Structural Model (Assertions-Disciplines-Overall): “Model 2”	22
Figure 7. Second-Order Structural Model (Assertions-Clusters-Overall): “Model 3”	23
Figure 8. Third-Order Structural Model (Assertions-Clusters-Disciplines-Overall): “Model 4”.	23

LIST OF APPENDICES

Appendix 4-A. Student Demographics and Reliability Coefficients
Appendix 4-B. Conditional Standard Error of Measurement
Appendix 4-C. Classification Accuracy and Consistency Indices by Subgroups
Appendix 4-D. Science Clusters Cognitive Lab Report
Appendix 4-E. Braille Cognitive Lab Report
Appendix 4-F. Independent Alignment Study Report

1. INTRODUCTION AND OVERVIEW OF RELIABILITY AND VALIDITY EVIDENCE

The state of Connecticut implemented the Connecticut Next Generation Science Standards (NGSS) Assessment for operational use starting in the 2018–2019 school year. The Connecticut NGSS Assessment replaced the Connecticut Mastery Test (CMT) in science (administered to students in grades 5 and 8) and the Connecticut Academic Performance Test (CAPT) in science (administered to students in grade 10). The Connecticut NGSS Assessment is administered online to grades 5, 8, and 11 using an adaptive test design. Accommodated versions are available for each grade, including braille and large print Data Entry Interface (DEI) forms. Spanish-language versions of the tests are also available. Table 1 shows the complete list of tests for the operational test administration in spring 2024.

Table 1. Spring 2024 Assessment Modes

Language/Format	Assessment Mode	Grade
English	Online	5, 8, & 11
Spanish	Online	5, 8, & 11
English/DEI	Paper	5, 8, & 11
English/braille	Online, Paper	5, 8, & 11

Given the intended uses of these tests, both reliability evidence and validity evidence are necessary to support appropriate inferences of student academic achievement from the Connecticut NGSS Assessment scores. The analyses to support reliability and validity evidence that are reported in this volume were conducted on the basis of test results for students whose scores were reported, including those taking the online English-language version and the accommodated versions of the Connecticut NGSS Assessment.

The purpose of this report is to provide empirical evidence that can subsequently be used to support a validity argument for the uses of and inferences from the Connecticut NGSS Assessment. This volume addresses the following five topics:

1. **Reliability.** The reliability estimates are presented by grade and demographic subgroup. This section also includes conditional standard errors of measurement (CSEM) and classification accuracy and consistency results by grade.
2. **Content Validity.** This section presents evidence showing that all students' tests were constructed to measure the NGSS with a sufficient number of items targeting each area of the test blueprint.
3. **Internal Structure Validity.** Evidence is provided regarding the internal relationships among the subscale scores to support their use and to justify the item response theory (IRT) measurement model. This type of evidence includes observed and disattenuated Pearson correlations among discipline scores per grade. As explained in detail in Volume 1, Annual Technical Report, the IRT model is a multidimensional model, with an overall dimension representing proficiency in science and nuisance dimensions that consider within-item

local dependencies among scoring assertions. In this volume, evidence is provided with respect to the presence of item cluster effects. Additionally, confirmatory factor analysis was used to evaluate the fit of the IRT model and to compare it with alternative models, including models with a simpler internal structure (e.g., unidimensional models) and models with a more elaborate internal structure.

4. **Relationship of Test Scores to External Variables.** Evidence of convergent and discriminant validity is provided using observed and disattenuated subscore correlations both within and across subjects.
5. **Test Fairness.** Fairness is an explicit concern during item development. Items are developed following the principles of universal design. Universal design removes barriers to provide access for the widest range of students possible. Test fairness is further monitored statistically using differential item functioning (DIF) analysis in tandem with content reviews by specialists.

1.1 RELIABILITY

The term *reliability* refers to consistency in test scores. Reliability can be defined as the degree to which individuals' deviation scores remain relatively consistent over repeated administrations of the same test or alternate test forms (Crocker & Algina, 1986). For example, if a person takes the same or parallel tests repeatedly, they should receive consistent results. The reliability coefficient refers to the ratio of true score variance to observed score variance:

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2}.$$

Another way to view reliability is to consider its relationship with the standard errors of measurement (SEM)—the smaller the standard error, the higher the precision of the test scores. For example, classical test theory assumes that an observed score (X) of an individual can be expressed as a true score (T) plus some error (E), $X = T + E$. The variance of X can be shown to be the sum of two orthogonal variance components:

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2.$$

Returning to the definition of reliability as the ratio of true score variance to observed score variance, we can arrive at the following theorem:

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_X^2 - \sigma_E^2}{\sigma_X^2} = 1 - \frac{\sigma_E^2}{\sigma_X^2}.$$

As the fraction of error variance to observed score variance tends to zero, the reliability then tends to 1. The classical test theory SEM, which assumes a homoscedastic error, is derived from the classical notion expressed above as $\sigma_X \sqrt{1 - \rho_{XX'}}$, where σ_X is the standard deviation of the scaled score, and $\rho_{XX'}$ is a reliability coefficient. Based on the definition of reliability, this formula can be derived as follows:

$$\rho_{XX'} = 1 - \frac{\sigma_E^2}{\sigma_X^2},$$

$$\frac{\sigma_E^2}{\sigma_X^2} = 1 - \rho_{XX'},$$

$$\sigma_E^2 = \sigma_X^2(1 - \rho_{XX'}), \text{ and}$$

$$\sigma_E = \sigma_X \sqrt{(1 - \rho_{XX'})}.$$

In general, the SEM is relatively constant across samples, as the group dependent term, σ_X , can be shown to cancel out:

$$\sigma_E = \sigma_X \sqrt{(1 - \rho_{XX'})} = \sigma_X \sqrt{(1 - (1 - \frac{\sigma_E^2}{\sigma_X^2}))} = \sigma_X \sqrt{\frac{\sigma_E^2}{\sigma_X^2}} = \sigma_X \times \frac{\sigma_E}{\sigma_X} = \sigma_E.$$

This shows that the SEM in the classical test theory is assumed to be a homoscedastic error, irrespective of the standard deviation of a group.

In contrast, the SEMs in IRT vary over the ability continuum. These heterogeneous errors are a function of a test information function (TIF) that provides different information about examinees depending on their estimated abilities.

Because the TIF indicates the amount of information provided by the test at different points along the ability scale, its inverse indicates the lack of information at different points along the ability scale. This lack of information is the uncertainty, or the measurement error, of the score at various score points. See Section 3, Reliability, for the derivation of heterogeneous measurement errors in IRT and a discussion of how these errors are aggregated over the score distribution to obtain a single, marginal, IRT-based reliability coefficient.

1.2 VALIDITY

The term *validity* refers to the degree to which “evidence and theory support the interpretations of test scores entailed by proposed uses of tests” (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014). Messick (1989) defines validity as “an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment” (p.13). Both definitions emphasize evidence and theory to support inferences and interpretations of test scores. The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) suggest five sources of validity evidence that can be used in evaluating a proposed interpretation of test scores. When validating test scores, these sources of evidence should be carefully considered.

The first source of evidence for validity is the relationship between the test content and the intended test construct (see Section 4, Evidence of Content Validity). For test score inferences to support a validity claim, the items should be representative of the content domain, and the content domain should be relevant to the proposed interpretation of test scores. To determine content representativeness, diverse panels of content experts conduct alignment studies, in which experts review individual items and rate them based on how well they match the test specifications or cognitive skills required for a construct (see Section 4.2, Independent Alignment Study, for the

results of an independent alignment study; and Volume 2, Test Development, for details on the item development process).

Technology-enhanced items should be examined to ensure that no construct-irrelevant variance is introduced. If some aspect of the technology impedes or advantages a student in their responses to items, this could affect item responses and inferences regarding abilities on the measured construct (see Volume 2, Test Development).

The second source of validity evidence is based on “the fit between the construct and the detailed nature of performance or response actually engaged in by examinees” (AERA, APA, & NCME, 2014, p. 12). This evidence is collected by surveying test takers about their performance strategies or responses to specific items. Because items are developed to measure specific constructs and intellectual processes, evidence that examinees have engaged in relevant performance strategies to correctly answer the items supports the validity of the test scores.

The third source of evidence for validity is based on *internal structure*: the degree to which the relationships among test items and test components relate to the construct on which the proposed test scores are interpreted. Possible analyses to examine internal structure are dimensionality assessment, goodness-of-model-fit to data, and reliability analysis (see Section 3, Reliability; and Section 5, Evidence of Internal-External Structure, for details). In addition, it is important to assess the degree to which the statistical relation between items and test components is invariant across groups. DIF analysis can be used to assess whether specific items function differently for subgroups of test takers (see Volume 1, Annual Technical Report).

The fourth source of evidence for validity is the relationship of test scores to external variables. The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) divides this source of evidence into three parts: (1) convergent and discriminant evidence; (2) test-criterion relationships; and (3) validity generalization. Convergent evidence supports the relationship between the test and other measures intended to assess similar constructs. Conversely, discriminant evidence delineates the test from other measures intended to assess different constructs. To analyze both convergent and discriminant evidence, a multitrait-multimethod matrix can be used. Additionally, test-criterion relationships indicate how accurately test scores predict criterion performance. The degree of accuracy depends mainly on the test’s purpose, such as classification, diagnosis, or selection. Test-criterion evidence is also used to investigate predictions of favoring different groups. Due to construct underrepresentation or construct-irrelevant components, the relation of test scores to a relevant criterion may differ from one group to another. Furthermore, validity generalization is related to whether the evidence is situation-specific or can be generalized across different settings and times. For example, sampling errors or range restriction may need to be considered in order to determine whether the conclusions of a test can be assumed for the larger population. Convergent and discriminant validity evidence are discussed in Section 5.2, Convergent and Discriminant Validity.

The fifth source of validity evidence is the intended and unintended consequences of test use, which should be included in the test-validation process. Determining the validity of the test should depend upon evidence directly related to the test; this process should not be influenced by external factors. For example, if an employer administers a test to determine hiring rates for different groups of people, an unequal distribution of skills related to the measurement construct does not necessarily imply a lack of validity for the test. However, if the unequal distribution of scores is in

fact due to an unintended, confounding aspect of the test, this *would* interfere with the test’s validity. As described in Volume 1, Annual Technical Report, and in this volume, test use should align with the intended purpose of the test.

Supporting a validity argument requires multiple sources of validity evidence. This enables one to evaluate whether sufficient evidence has been presented to support the intended uses and interpretations of the test scores. Thus, determining the validity of a test first requires an explicit statement regarding the intended uses of the test scores and, subsequently, evidence that the scores can be used to support these inferences.

2. PURPOSE OF THE CONNECTICUT NEXT GENERATION SCIENCE STANDARDS ASSESSMENT

The primary purpose of Connecticut’s Summative Assessment System is to yield accurate information on students’ achievement of Connecticut’s education standards. The Connecticut NGSS Assessment measures the science knowledge and skills of Connecticut students in grades 5, 8, and 11. The Connecticut State Department of Education (CSDE) provides an overview of the science assessment at <https://portal.ct.gov/SDE/Student-Assessment/NGSS-Science/NGSS-Science>. Information about the NGSS is available at www.nextgenscience.org.

The Connecticut NGSS Assessment supports instruction and student learning by measuring growth in student achievement. Assessments can be used as indicators to determine whether students in Connecticut are ready with the knowledge and skills that are essential for college education and careers.

Connecticut’s educational assessments also provide evidence for the requirements of state and federal accountability systems. Test scores can be employed to evaluate students’ learning progress and to help teachers to improve their instruction, which in turn has a positive effect on students’ learning over time.

The tests are constructed to measure student proficiency in accordance with best practice as described in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014). Item development adheres to the principles of universal design in order to ensure that all students have access to the test content. Volume 2, Test Development, describes in more detail the Connecticut NGSS Assessment standards and test blueprints. Additional evidence of content validity can also be found in Section 4, Evidence of Content Validity. The Connecticut NGSS Assessment test scores are useful indicators for understanding individual students’ academic achievement of the Connecticut content standards and for evaluating whether students’ performance is improving over time. Additionally, both individual and aggregated scores can be used for measuring reliability of the test. A discussion of test score reliability can be found in Section 3, Reliability.

The Connecticut NGSS Assessment is a criterion-referenced test that is designed to measure student performance on the NGSS in Connecticut schools. As a comparison, norm-referenced tests are designed to rank or compare all students with one another. The Connecticut NGSS Assessment standards and test blueprints are discussed in Volume 2, Test Development.

The scale score and relative strengths and weaknesses at the discipline level are provided for each student to indicate student strengths and weaknesses in different content areas of the test, relative to the other areas and to the district and state. These scores serve as useful feedback that teachers can use to tailor their instruction. To support their practical use across the state, we must examine the reliability coefficients for and the validity of these test scores.

3. RELIABILITY

Classical test-theory-based reliability indices are not appropriate for science assessments for two reasons. First, in spring 2024, the science test was administered under an adaptive test design. Potentially, each student received a unique set of items, whereas classical test-theory-based reliability indices require that the same set of items be administered to a (large) group of students. Second, since item response theory (IRT) methods are used for calibration and scoring, the measurement error of ability estimates is not constant across the ability range, even for the same set of items. The reliability of science tests is computed as follows:

$$\bar{\rho} = [\sigma^2 - \left(\frac{\sum_{i=1}^N CSEM_i^2}{N}\right)]/\sigma^2,$$

where N is the number of students; $CSEM_i$ is the conditional standard errors of measurement (CSEM) of the overall ability estimate for student i ; and σ^2 is the variance of the overall ability estimates. The higher the reliability coefficient, the greater the precision of the test.

The marginal reliability of science for the overall sample is reported by grade in Table 2. The overall reliability ranges from 0.90 to 0.91. Due to the new structure of the test, Cambium Assessment, Inc. (CAI) has also explored the relationships between reliability and other important factors, such as the effect of nuisance dimensions (see Section 5 of Volume 1, Annual Technical Report). It was found that if the local dependencies among assertions pertaining to the same item are ignored, the marginal reliability typically increases to 0.90 or above. Ignoring local dependencies can be achieved either by computing the maximum likelihood estimation (MLE) ability estimates under the unidimensional Rasch model or by setting the variance parameters to zero for all item clusters when computing the marginal maximum likelihood estimation (MMLE) ability estimates under the one-parameter logistic (1PL) bifactor model (see Section 6.1 of Volume 1, Annual Technical Report).

By ignoring the local dependencies, which are substantial for many item clusters, the reliability coefficient is overestimating the true reliability of the test. Note, however, that local dependencies are also present to some degree in traditional assessments that make use of item groups (e.g., a set of items relating to the same reading passage). Local dependencies are typically not accounted for by traditional assessments, and hence reported reliability coefficients may be overestimating to some degree the true reliability of these tests. The reliability coefficients are also reported for demographics subgroups in Appendix 4-A, Student Demographics and Reliability Coefficients.

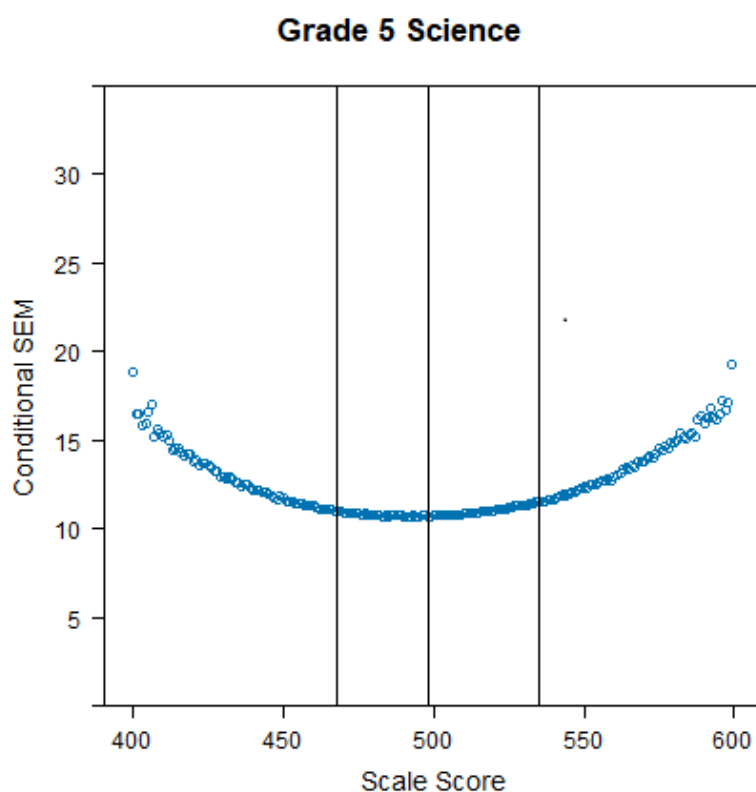
Table 2. Spring 2024 Marginal Reliability Coefficients

Grade	Sample Size	Reliability
5	36,451	0.90
8	37,097	0.91
11	37,288	0.90

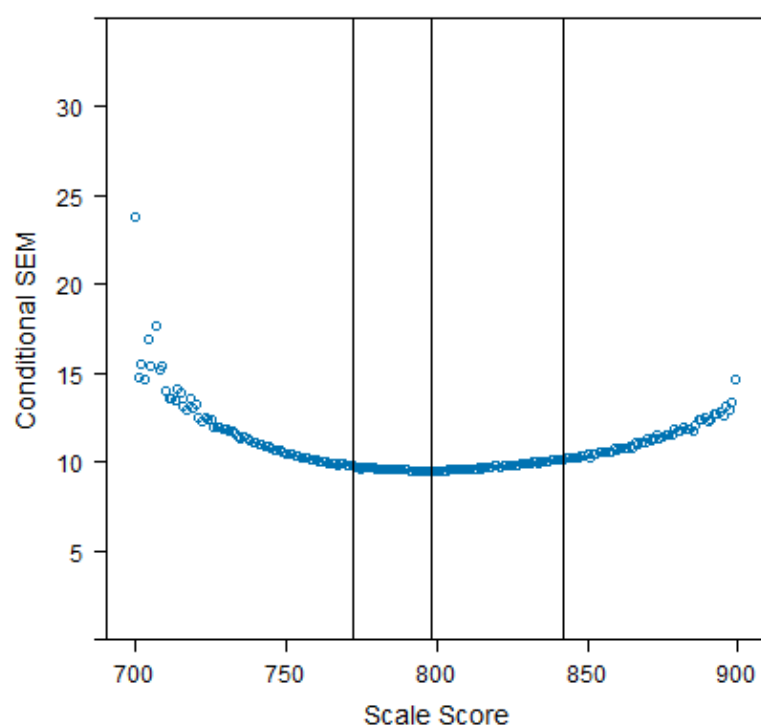
3.1 STANDARD ERROR OF MEASUREMENT

The computation method of CSEM has been described in Section 6.4 of Volume 1, Annual Technical Report. Figure 1 presents the average CSEM for each scale score. The lowest standard errors are observed near the proficiency cut (the middle vertical line) for all grades, which is a desirable test property. The CSEM at each scale score is reported in Appendix 4-B, Conditional Standard Error of Measurement.

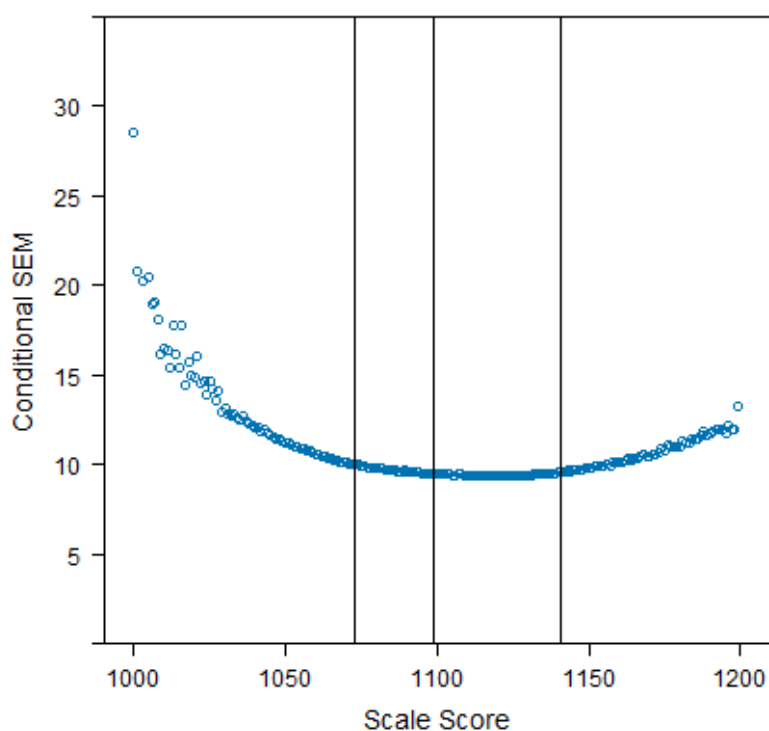
Figure 1. Spring 2024 Conditional Standard Errors of Measurement



Grade 8 Science



Grade 11 Science



3.2 RELIABILITY OF PERFORMANCE CLASSIFICATION

When student performance is reported in terms of performance levels, the reliability of classifying students into a specific level can be computed in terms of the likelihood of accurate and consistent classification as specified in Standard 2.16 of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014).

The reliability of performance classification can be examined in terms of *classification accuracy* (CA) and *classification consistency* (CC). CA refers to the agreement between the classifications based on the form taken and the classifications that would be made based on the students' true scores if hypothetically they could be obtained. CC refers to the agreement between the classifications based on the form taken and the classifications that would be made based on an alternate, equivalently constructed test form.

In reality, the true ability is unknown, and students are not administered an alternate, equivalent form. Therefore, CA and CC are estimated on the basis of students' item scores, the item parameters, and the assumed latent ability distribution as described in the following sections. The true score is an expected value of the test score with measurement error.

For student j , the student's estimated ability is $\hat{\theta}_j$ with a standard error of measurement (SEM) of $se(\hat{\theta}_j)$; and the estimated ability is distributed as $\hat{\theta}_j \sim N(\theta_j, se^2(\hat{\theta}_j))$, assuming a normal distribution, where θ_j is the unknown true ability of student j . The probability of the true score at performance level l ($l = 1, \dots, L$) is estimated as

$$p_{jl} = p(c_{Ll} \leq \theta_j < c_{Ul}) = p\left(\frac{c_{Ll} - \hat{\theta}_j}{se(\hat{\theta}_j)} \leq \frac{\theta_j - \hat{\theta}_j}{se(\hat{\theta}_j)} < \frac{c_{Ul} - \hat{\theta}_j}{se(\hat{\theta}_j)}\right) = p\left(\frac{\hat{\theta}_j - c_{Ul}}{se(\hat{\theta}_j)} < \frac{\hat{\theta}_j - \theta_j}{se(\hat{\theta}_j)} \leq \frac{\hat{\theta}_j - c_{Ll}}{se(\hat{\theta}_j)}\right) = \Phi\left(\frac{\hat{\theta}_j - c_{Ll}}{se(\hat{\theta}_j)}\right) - \Phi\left(\frac{\hat{\theta}_j - c_{Ul}}{se(\hat{\theta}_j)}\right),$$

where c_{Ll} and c_{Ul} denote the score corresponding to the lower and upper limits of performance level l , respectively.

3.2.1 Classification Accuracy

Using p_{jl} , an $L \times L$ matrix \mathbf{E}_A can be calculated. Each element E_{Akl} of matrix \mathbf{E}_A represents the expected number of students to score at level l (based on their true scores) given students from observed level k , and can be calculated as

$$E_{Akl} = \sum_{pl_j \in k} p_{jl},$$

where pl_j is the j th student's observed performance level. The CA at level l is estimated as

$$CA_l = \frac{E_{Akl}}{N_k},$$

where N_k is the observed number of students scoring in performance level k .

The CA for the p th cut is estimated by forming square partitioned blocks of the matrix \mathbf{E}_A and taking the summation over all elements within the block as follows:

$$CAC = (\sum_{k=1}^p \sum_{l=1}^p E_{Akl} + \sum_{k=p+1}^L \sum_{l=p+1}^L E_{Akl}) / N,$$

where N is the total number of students.

The overall CA is estimated from the diagonal elements of the matrix:

$$CA = \frac{tr(E_A)}{N}.$$

Table 3 provides the overall CA and the CA for the individual cuts. The overall CA of the test ranges from 78.02% to 80.52%. The individual cut accuracy rates are high across all grades and forms, with the minimum value being 90.97% for grade 5 level 3 cut. It denotes that more than 90% of the time we can accurately differentiate students above and below each cut score in the spring 2023 Connecticut NGSS Assessment. The CA for demographic subgroups is presented in Appendix 4-C, Classification Accuracy and Consistency Indices by Subgroups.

Table 3. Spring 2024 Classification Accuracy Index

Grade	Overall Accuracy (%)	Cut Accuracy (%)		
		Level 2 Cut	Level 3 Cut	Level 4 Cut
5	78.02	93.21	90.97	93.81
8	80.52	92.59	91.64	96.26
11	78.56	91.06	91.34	96.13

3.2.2 Classification Consistency

Assuming the test is administered twice independently to the same group of students, similarly to accuracy, a $L \times L$ matrix E_C can be constructed. The element of E_C is populated by

$$E_{Ckl} = \sum_{j=1}^N p_{jl} p_{jk},$$

where p_{jl} is the probability of the true score at performance level l in the first administration, and p_{jk} is the probability of the true score at performance level k in the second administration for the j th student. The classification consistency index for the cuts (CCC) and overall CC were estimated in a way similar to the classification accuracy for the cuts and CA.

$$CCC = (\sum_{k=1}^p \sum_{l=1}^p E_{Ckl} + \sum_{k=p+1}^L \sum_{l=p+1}^L E_{Ckl}) / N,$$

and

$$CC = \frac{tr(E_C)}{N}.$$

Table 4 provides the overall CC and the CC for the cuts. The overall CC of the test ranges from 69.40% to 72.80 %. The individual cut consistency rates are high across all grades and forms, with the minimum value being 87.33 % for grade 5 level 3 cut. In all performance levels, CA is slightly higher than CC. CC rates can be lower than CA; the consistency is based on two tests with measurement errors, but the accuracy is based on one test with a measurement error and the true

score. The accuracy and consistency rates for each performance level are higher for the levels with a smaller standard error. The CC for demographic subgroups is presented in Appendix 4-C, Classification Accuracy and Consistency Indices by Subgroups.

Table 4. Spring 2024 Classification Consistency Index

Grade	Overall Consistency (%)	Cut Consistency (%)		
		Level 2 Cut	Level 3 Cut	Level 4 Cut
5	69.40	90.41	87.33	91.30
8	72.80	89.54	88.26	94.70
11	70.26	87.51	87.81	94.55

3.3 PRECISION AT CUT SCORES

Table 5 presents the mean CSEM at each performance level by grade. The table also includes performance level cut scores and associated CSEM. The CSEM at each scale score is reported in Appendix 4-B, Conditional Standard Error of Measurement.

Table 5. Spring 2024 Performance Levels and Associated Conditional Standard Error of Measurement

Grade	Performance Level	Mean CSEM	Cut Score (Scale Score)	CSEM at Cut Score
5	1	11.91	-	-
	2	10.80	468	10.97
	3	10.99	498	10.74
	4	12.64	535	11.51
8	1	10.48	-	-
	2	9.62	772	9.79
	3	9.74	798	9.49
	4	10.87	842	10.18
11	1	10.83	-	-
	2	9.73	1,073	10.01
	3	9.44	1,099	9.49
	4	10.22	1,141	9.60

4. EVIDENCE OF CONTENT VALIDITY

This section demonstrates that the knowledge and skills assessed by the Connecticut NGSS Assessment are representative of the content standards of the larger knowledge domain. We describe the content standards for the Connecticut NGSS Assessment and discuss the test development process and mapping Connecticut NGSS Assessment tests to the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014). A complete description of the test development process can be found in Volume 2, Test Development.

4.1 CONTENT STANDARDS

The Connecticut NGSS Assessment was aligned to the NGSS, adopted by Connecticut in 2015. The standards are available for review at the following URL: <https://portal.ct.gov/SDE/Science/Science-Standards-and-Resources>. Blueprints were developed to ensure that the test and the items were aligned to the standards that they were intended to measure. A complete description of the blueprint and test construction process can be found in Volume 2, Test Development.

Table 6 presents the disciplines by grade, as well as the number of operational items administered that measured each discipline.

Table 6. Number of Spring 2024 Items for Each Discipline

Grade	Reporting Category	Item Cluster	Stand-Alone Item
5	Earth and Space Sciences	37	42
	Life Sciences	37	49
	Physical Sciences	41	55
8	Earth and Space Sciences	44	47
	Life Sciences	68	64
	Physical Sciences	50	53
11	Earth and Space Sciences	22	43
	Life Sciences	50	74
	Physical Sciences	28	56

4.2 INDEPENDENT ALIGNMENT STUDY

While it is critically important to develop and strictly enforce an item development process that works to ensure alignment of test items to content standards, it is also important to independently verify the alignment of test items to content standards. The WebbAlign team of the not-for-profit Wisconsin Center for Education Products and Services (WCEPS) conducted an alignment study in July 2019. The study comprised two components. The first component addressed the alignment of the Memorandum of Understanding (MOU) item bank, shared by all states that are part of the MOU. In a second component, alignment was investigated for each state participating in the study, in the context of their state-specific blueprint and item bank, which is a particular state-vetted subset of items from the shared MOU item bank (see Volume 2, Test Development).

The results of the alignment study are presented in Appendix 4-F, Independent Alignment Study Report.

5. EVIDENCE OF INTERNAL-EXTERNAL STRUCTURE

In this section, the internal structure of the assessment is explored using the scores provided at the discipline level. The relationship between the discipline scores is just one indicator of the test dimensionality. The Connecticut NGSS Assessment is calibrated with the Rasch testlet model (Wang & Wilson, 2005). The testlet model is a high-dimensional model that incorporates a nuisance dimension for each item cluster (and stand-alone items with four or more assertions) in addition to an overall dimension representing overall proficiency. This approach is innovative and quite different from the traditional approach of ignoring local dependencies. Validity evidence for the internal structure will focus on the presence of cluster effects and how substantial they are. Additionally, confirmatory factor analysis is used to evaluate the fit of the IRT model and to compare the model with alternative models, including those with a simpler internal structure (i.e., unidimensional models without cluster effects) and models with a more elaborate internal structure (refer to Section 5.4, Confirmatory Factor Analysis).

Another pathway is to explore observed correlations between the discipline scores. However, as each discipline is measured with a small number of items, the standard errors of the observed scores within each discipline are typically larger than the standard error of the total test score. Disattenuating for measurement error could offer some insight into the theoretical true score correlations. Both observed correlations and disattenuated correlations are provided in the following section.

5.1 CORRELATIONS AMONG DISCIPLINE SCORES

Table 7 presents the observed and disattenuated correlation matrix of the discipline scores. The observed correlations range from 0.72 to 0.76, and disattenuated correlations range from 0.98 to 0.99.

In some instances, the observed correlations were lower than one might expect. However, as previously noted, the correlations were subject to a large amount of measurement error at the discipline level due to the limited number of items from which the scores were derived. Consequently, interpretation of these correlations, as either high or low, should be made cautiously. After correcting for measurement error, the correlations between the discipline scores become very high. The disattenuated correlations are close to 1, supporting the use of a psychometric model that does not include a separate dimension for each of the three disciplines.

Table 7. Spring 2024 Correlations Among Disciplines

Grade	Reporting Category	Earth and Space Sciences (ESS)	Life Sciences (LS)	Physical Sciences (PS)
5	ESS	0.75*	0.99	0.98
	LS	0.74	0.74*	0.98
	PS	0.72	0.72	0.72*

Grade	Reporting Category	Earth and Space Sciences (ESS)	Life Sciences (LS)	Physical Sciences (PS)
8	ESS	0.76*	0.98	0.98
	LS	0.75	0.77*	0.98
	PS	0.75	0.76	0.77*
11	ESS	0.74*	0.99	0.98
	LS	0.75	0.77*	0.99
	PS	0.73	0.75	0.74*

Note. *Diagonal value represents marginal reliability for each discipline. Observed correlations are below the diagonal, and disattenuated are above. Disattenuated correlations larger than 1 were truncated to 1.

5.2 CONVERGENT AND DISCRIMINANT VALIDITY

Collectively, Standard 1.16 through Standard 1.19 of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) emphasize practices to provide evidence of convergent and discriminant validity. It is a part of validity evidence demonstrating that assessment scores are related as expected with criteria and other variables for all student groups. However, a second, independent test measuring the same science construct as the Connecticut NGSS Assessment, which could easily permit for a cross-test set of correlations, was not available. Alternatively, the correlations between subscores were examined. The *a priori* expectation is that subscores within the same subject (e.g., correlation of science disciplines within science) will correlate more positively than subscores across subjects (e.g., correlation of science disciplines with reporting categories within mathematics). These correlations are based on a small number of items; consequently, the observed score correlations will be smaller in magnitude as a result of the larger measurement error at the subscore level. For this reason, both the observed score and the disattenuated correlations are provided.

Observed and disattenuated subscore correlations were calculated both within and across subjects. The pattern was generally consistent with the *a priori* expectation that subscores within a test correlate higher than correlations between tests measuring a different construct. The correlations between reporting categories from science, English language arts (ELA), and mathematics are presented in Table 8 and Table 9. On the diagonal, the reliability coefficient of the reporting category is shown. Correlations across subjects are presented only for grades 5 and 8 since ELA and mathematics assessments are administered only in grades 3–8 in Connecticut.

Table 8. Spring 2024 Correlations Across Subjects, Grade 5

Subject	Number of Students	Reporting Category	Science			English Language Arts (ELA)			Mathematics		
			ESS	LS	PS	R	L	WR	CP	PS	CR
Science	36,363	Earth and Space Sciences (ESS)	0.75*	0.99	0.98	0.93	0.92	0.95	0.89	0.97	0.94
		Life Sciences (LS)	0.74	0.74*	0.98	0.94	0.93	0.97	0.87	0.95	0.92
		Physical Sciences (PS)	0.72	0.71	0.72*	0.91	0.90	0.94	0.87	0.94	0.92
ELA		Reading (R)	0.72	0.72	0.70	0.81*	0.98	1.00	0.84	0.92	0.90
		Listening (L)	0.73	0.73	0.71	0.81	0.85*	1.00	0.87	0.94	0.91
		Writing and Research (WR)	0.66	0.67	0.64	0.72	0.74	0.64*	0.88	0.96	0.92
Mathematics		Concepts and Procedures (CP)	0.73	0.71	0.70	0.72	0.76	0.67	0.90*	1.00	0.99
		Problem Solving, Modeling, and Data Analysis (PS)	0.69	0.68	0.66	0.68	0.72	0.64	0.80	0.69*	1.00
		Communicating and Reasoning (CR)	0.69	0.67	0.66	0.69	0.72	0.63	0.80	0.74	0.72*

Note. *Diagonal value represents the reliability coefficient of the reporting category. Observed correlations are below the diagonal, and disattenuated are above. Disattenuated correlations larger than 1 were truncated to 1.

Table 9. Spring 2024 Correlations Across Subjects, Grade 8

Subject	Number of Students	Reporting Category	Science			English Language Arts (ELA)			Mathematics		
			ESS	LS	PS	R	L	WR	CP	PS	CR
Science	36,744	Earth and Space Sciences (ESS)	0.76*	0.98	0.98	0.89	0.89	0.92	0.89	0.99	0.93
		Life Sciences (LS)	0.75	0.77*	0.98	0.91	0.90	0.92	0.89	0.99	0.93
		Physical Sciences (PS)	0.75	0.75	0.77*	0.89	0.89	0.91	0.89	0.99	0.93
ELA		Reading (R)	0.69	0.71	0.69	0.78*	1.00	1.00	0.86	0.97	0.90
		Listening (L)	0.70	0.71	0.70	0.79	0.80*	1.00	0.88	0.98	0.92
		Writing and Research (WR)	0.64	0.64	0.63	0.72	0.71	0.63*	0.89	0.99	0.92
Mathematics		Concepts and Procedures (CP)	0.74	0.74	0.74	0.72	0.75	0.67	0.90*	1.00	1.00
		Problem Solving, Modeling, and Data Analysis (PS)	0.69	0.69	0.69	0.68	0.69	0.62	0.80	0.63*	1.00
		Communicating and Reasoning (CR)	0.67	0.68	0.68	0.66	0.68	0.61	0.79	0.73	0.69*

Note. *Diagonal value represents the reliability coefficient of the reporting category. Observed correlations are below the diagonal, and disattenuated are above. Disattenuated correlations larger than 1 were truncated to 1.

Additionally, the correlation was computed among the overall scores for the three tested subjects: ELA, mathematics, and science. Correlations are presented in Table 10 and are relatively high, between 0.81 and 0.85.

Table 10. Spring 2024 Correlations Across ELA, Mathematics, and Science Scores

Grade	N	English Language Arts (ELA) & Mathematics	ELA & Science	Mathematics & Science
5	36,363	0.82	0.85	0.82
8	36,744	0.81	0.82	0.84

5.3 CLUSTER EFFECTS

The Connecticut NGSS Assessment is calibrated with the Rasch testlet model (Wang & Wilson, 2005). The testlet model is a high-dimensional model that incorporates a nuisance dimension for each item cluster in addition to an overall dimension representing overall proficiency. Section 5.1 of Volume 1, Annual Technical Report, presents a detailed description of the IRT model. The internal (latent) structure of the model is presented in Figure 7. The psychometric approach for the assessment is innovative and quite different from the traditional approach of ignoring local dependencies. The validity evidence for the internal structure presented in this section relates to the presence of cluster effects (i.e., nuisance dimensions) and how substantial they are.

Simulation studies conducted by Rijmen, Jiang, and Turhan (2018) confirmed that both the item difficulty parameters and the cluster variances are recovered well for the Rasch testlet model under a variety of conditions. Cluster effects with a range of magnitudes were recovered well. The results obtained by Rijmen et al. (2018) confirmed earlier findings reported in the literature (e.g., Bradlow, Wainer, & Wang, 1999) under conditions that were chosen to closely resemble the assessment. For example, in one of the studies, the item location parameters and cluster variances used to simulate data were based on the results of a pilot study.

CAI examined the distribution of cluster variances obtained from the 2019 IRT calibrations for the entire bank used across all states that participate in the Memorandum of Understanding (MOU) item-sharing agreement and the states that rely on the science Independent College and Career Readiness (ICCR) item pool.

For elementary school, the estimated value of the cluster variances of all operational, scored items ranged from 0 to 5.13, with a median value of 0.57 and a mean value of 0.92. As a comparison, the estimated variance parameter of the overall dimension for Connecticut elementary school in 2019 was $\hat{\sigma}_{\theta CT}^2 = 0.78$.

For middle school, the estimated value of the cluster variances of all operational, scored items ranged from 0 to 4.63, with a median value of 0.46 and a mean value of 0.68. The estimated variance parameter of the overall dimension for Connecticut middle school in 2019 was $\hat{\sigma}_{\theta CT}^2 = 0.78$.

For high school, the estimated value of the cluster variances of all operational, scored items ranged from 0.11 to 7.75, with a median value of 0.45 and a mean value of 0.65. The estimated variance parameter of the overall dimension for Connecticut high school in 2019 was $\hat{\sigma}_{\theta CT}^2 = 0.83$.

Figure 2 through Figure 4 present the histograms of the cluster variances expressed as the proportion of the systematic variance due to the cluster variance for each cluster (computed as $\eta_g = \frac{\hat{\sigma}_g^2}{\hat{\sigma}_{\theta CT}^2 + \hat{\sigma}_g^2}$), where $\hat{\sigma}_{\theta CT}^2$ is the variance estimate of the overall proficiency of Connecticut students. The variance proportion shows the relative magnitude of the variance of a cluster compared to the variance of the overall dimension. For instance, if the variance proportion of a cluster is larger than 0.5, then the cluster variance is larger than the overall variance; otherwise, the cluster variance is smaller than the overall variance. For all three grade bands, a wide range of cluster variances is observed. These results indicate that, for all grades, cluster effects can be substantial and provide evidence for the appropriateness of a psychometric model that explicitly takes local dependencies among the assertions of an item cluster into account.

Figure 2. Cluster Variance Proportion for Operational Items in Elementary School

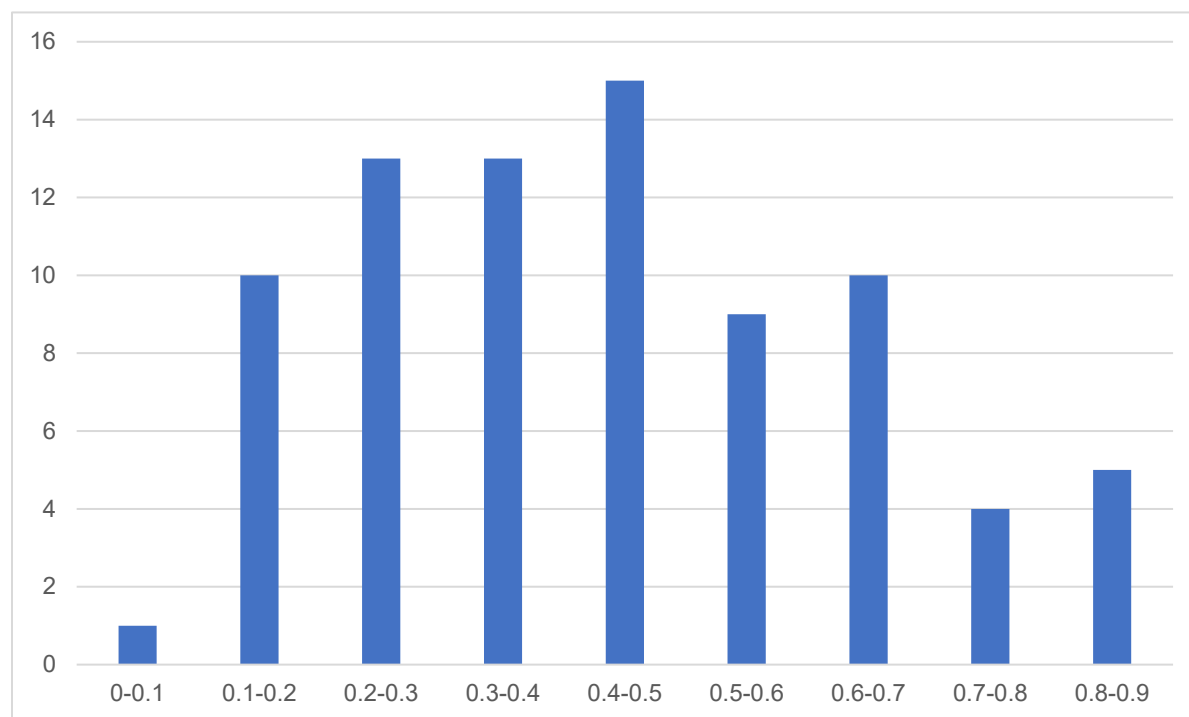
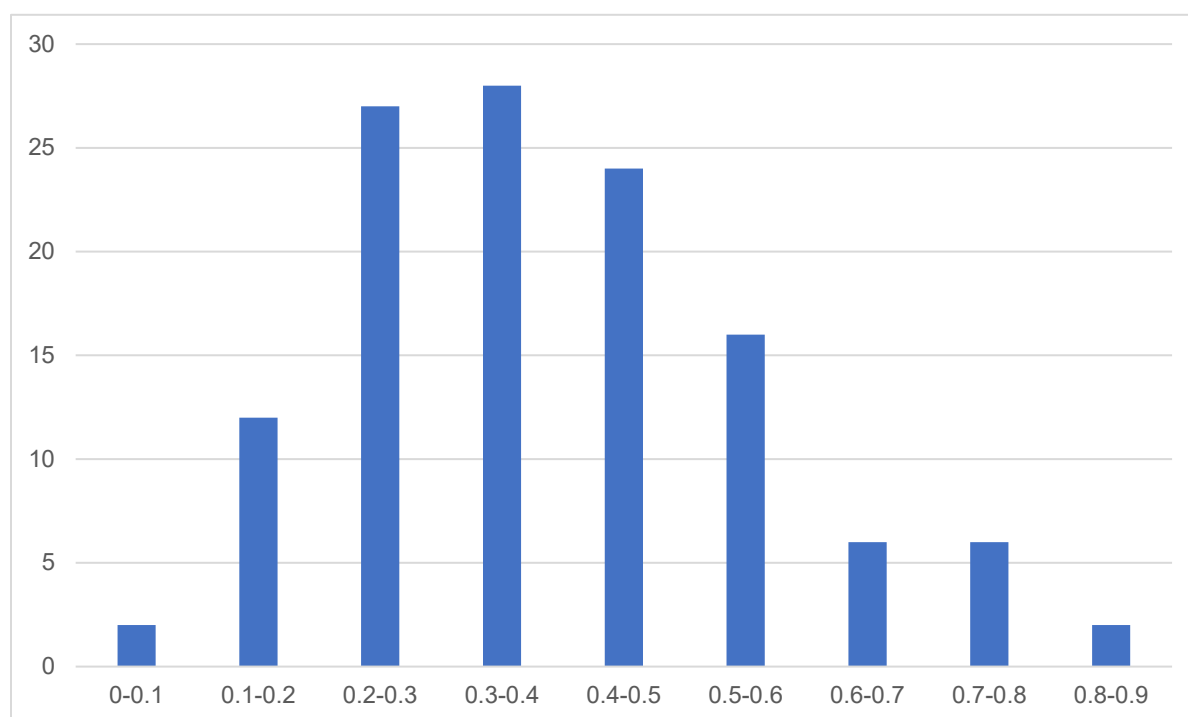
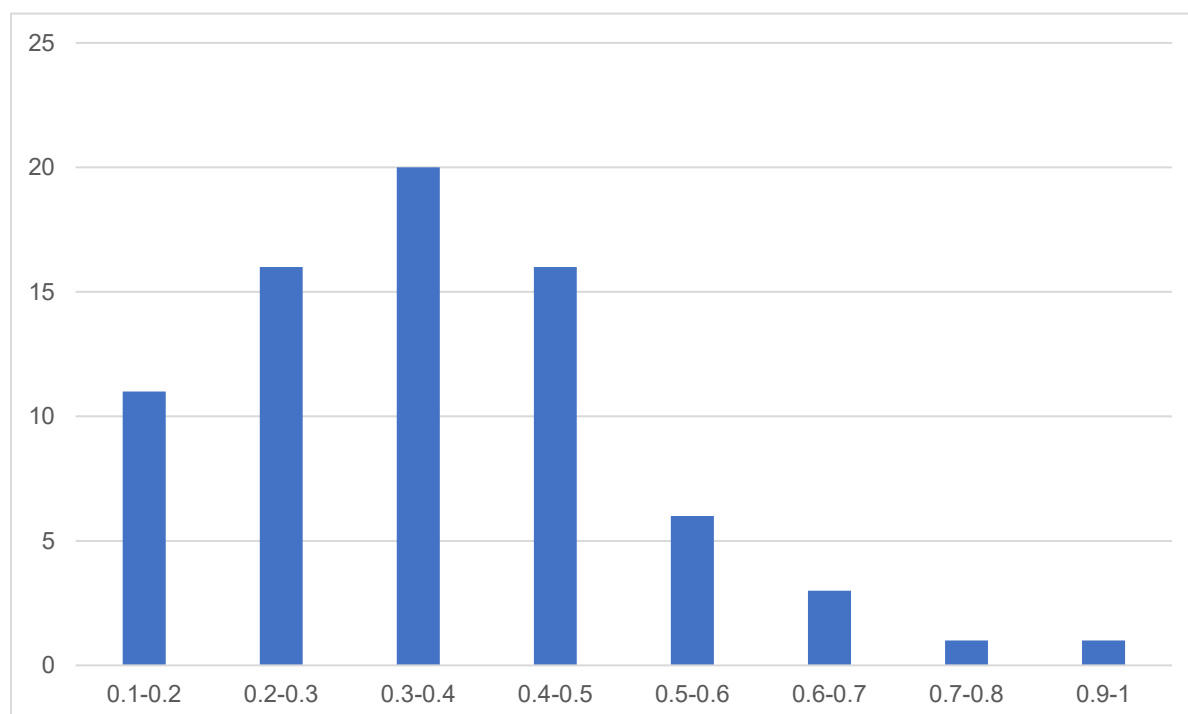


Figure 3. Cluster Variance Proportion for Operational Items in Middle School*Figure 4. Cluster Variance Proportion for Operational Items in High School*

5.4 CONFIRMATORY FACTOR ANALYSIS

In Section 5.3, Cluster Effects, evidence is presented for the existence of substantial cluster effects. In this section, the internal structure of the IRT model used for calibrating the item parameters is further evaluated using confirmatory factor analysis. In addition, alternative models are considered, including models with a simpler internal structure (e.g., unidimensional models) and models with a more elaborate internal structure.

Estimation methods for confirmatory factor analysis for discrete observed variables are not well suited for incomplete data collection designs where each case has data only on a subset of the set of observed variables. The linear-on-the-fly (LOFT) test design results in sparse data matrices. Every student is only responding to a small number of items relative to the size of the item pool, so data are missing on most of the manifest variables for any given student. In 2018 and 2019, a LOFT design was used for all operational science assessments inspired by the Next Generation Science Standards (NGSS) framework, except for Utah. As a result, the student responses of these other states are not readily amenable for the application of confirmatory factor analysis techniques.

The 2018 Utah operational field test for science made use of a set of fixed-form tests for each grade. Therefore, the data for each fixed-form test are complete, and the fixed-form tests are amenable to confirmatory factor analysis. The Utah science standards, even though the standards are grade-specific for middle school, were developed under a framework similar to the one developed for the NGSS, and a crosswalk is available between both sets of standards. Utah is part of the MOU, and many of the other states that take part in the MOU also use the middle school items developed for and owned by Utah. Taken together, analyzing the fixed science forms that were administered in Utah in 2018 can provide evidence with respect to the internal structure of the Connecticut NGSS Assessment.

In 2018, Utah’s science assessments comprised a set of fixed-form tests per grade, and all items in these forms were clusters. The number of fixed-form tests varied by grade, but within each grade the total number of clusters was the same across forms. However, some items were rejected during the rubric validation or data review and were removed from this analysis. All students with a “completed” status were included in the factor analysis. The percentage of students per grade that had a status other than “completed” was less than 0.85%. Table 11 summarizes the number of forms included in this analysis, the number of clusters per discipline (range across forms), the number of assertions (range across forms), and the number of students (range across forms) for each of the grades.

Table 11. Numbers of Forms, Clusters per Discipline (Range Across Forms), Assertions per Form (Range Across Forms), and Students per Form (Range Across Forms)

Grade	Number of Fixed Forms	Number of Clusters per Discipline in Each Form			Number of Assertions per Form	Number of Students per Form
		<i>Physical Sciences</i>	<i>Earth and Space Sciences</i>	<i>Life Sciences</i>		
6	3	2	2–3	2–3	74–83	6,804–6,881
7	6	2	2	5	83–89	3,822–3,890

Grade	Number of Fixed Forms	Number of Clusters per Discipline in Each Form			Number of Assertions per Form	Number of Students per Form
		<i>Physical Sciences</i>	<i>Earth and Space Sciences</i>	<i>Life Sciences</i>		
8	3	6–7	2	2	93–100	5,061–5,104

The factor structure of a testlet model, which is the model used for calibration, is formally equivalent to a second-order model. Specifically, the testlet model is the model obtained after a Schmid–Leiman transformation of the second-order model (Li, Bolt, & Fu, 2006; Rijmen, 2009; Yung, Thissen, & McLeod, 1999). In the corresponding second-order model, the group of assertions related to a cluster are indicators of the cluster, and each cluster is an indicator of overall science performance. Because assertions are not pure indicators of a specific factor, each assertion has a corresponding error component. Similarly, clusters include an error component indicating they are not pure indicators of the overall science performance.

CAI used confirmatory factor analysis to evaluate the fit of the second-order model described above to student data from spring 2018. Three additional structural models were included in the analysis as well. In the first model, only one factor represented overall science performance. All assertions are indicators of this overall proficiency factor. The first model was a testlet model where all cluster variances were zero. In the second model, assertions were indicators of the corresponding science discipline, and each discipline was an indicator of the overall science performance. This was a second-order model with science disciplines rather than clusters as first-order factors. This model did not take the cluster effects into account. In the last, most general model, assertions were indicators of the corresponding cluster, and clusters were indicators of the corresponding science discipline, with disciplines being indicators of the overall science performance.

For the sake of simplicity, the models in the analysis are here referred to as

- Model 1–Assertions-Overall Science (one factor model)
- Model 2–Assertions-Disciplines-Overall Science (second-order model)
- Model 3–Assertions-Clusters-Overall Science (second-order model)
- Model 4–Assertions-Clusters-Disciplines-Overall Science (third-order model)

Figure 5 through Figure 8 illustrate these four structural models. Model 1 is nested within Models 2, 3, and 4. Also, Models 2 and 3 are nested within Model 4. The paths from the factors to the assertions represent the first-order factor loadings. Note that all four models include factor loadings for the assertions, which differs from the calibration model where all the discrimination parameters of the assertions were set to 1. All models were estimated using the lavaan package in R (Rosseel, 2012), with the diagonally weighted least squares (DWLS) method for parameter estimation, the recommended approach for binary data (Flora & Curran, 2004).

Figure 5. One-Factor Structural Model (Assertions-Overall): “Model 1”

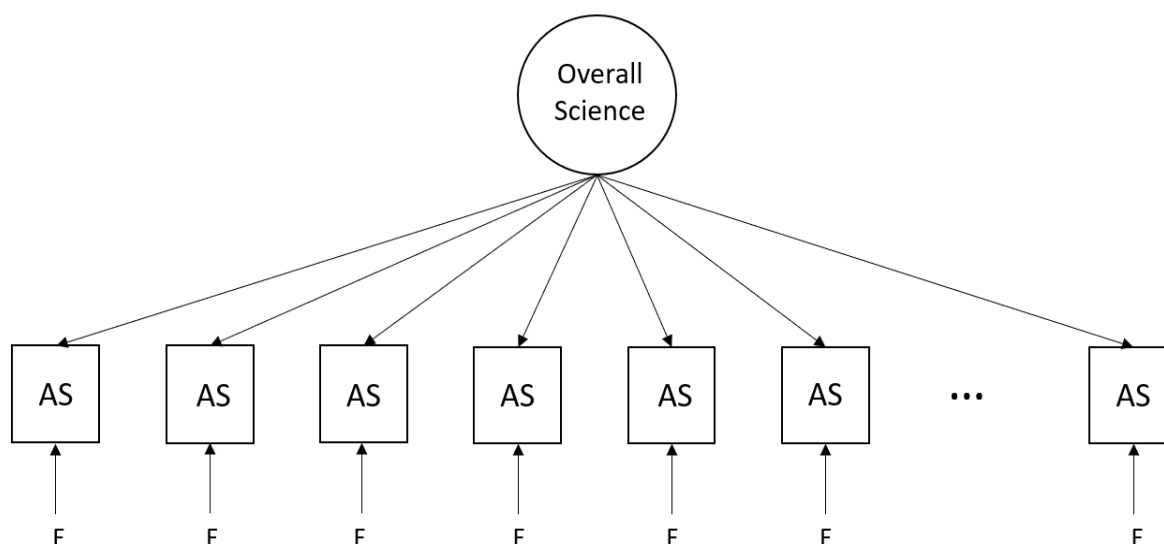


Figure 6. Second-Order Structural Model (Assertions-Disciplines-Overall): “Model 2”

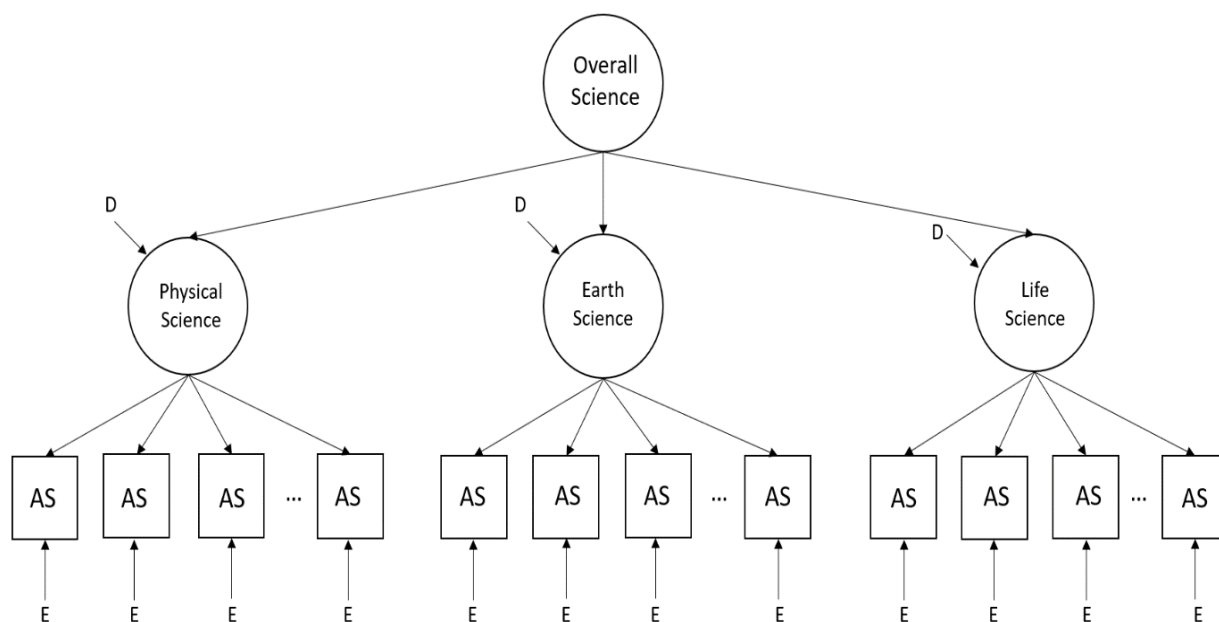


Figure 7. Second-Order Structural Model (Assertions-Clusters-Overall): “Model 3”

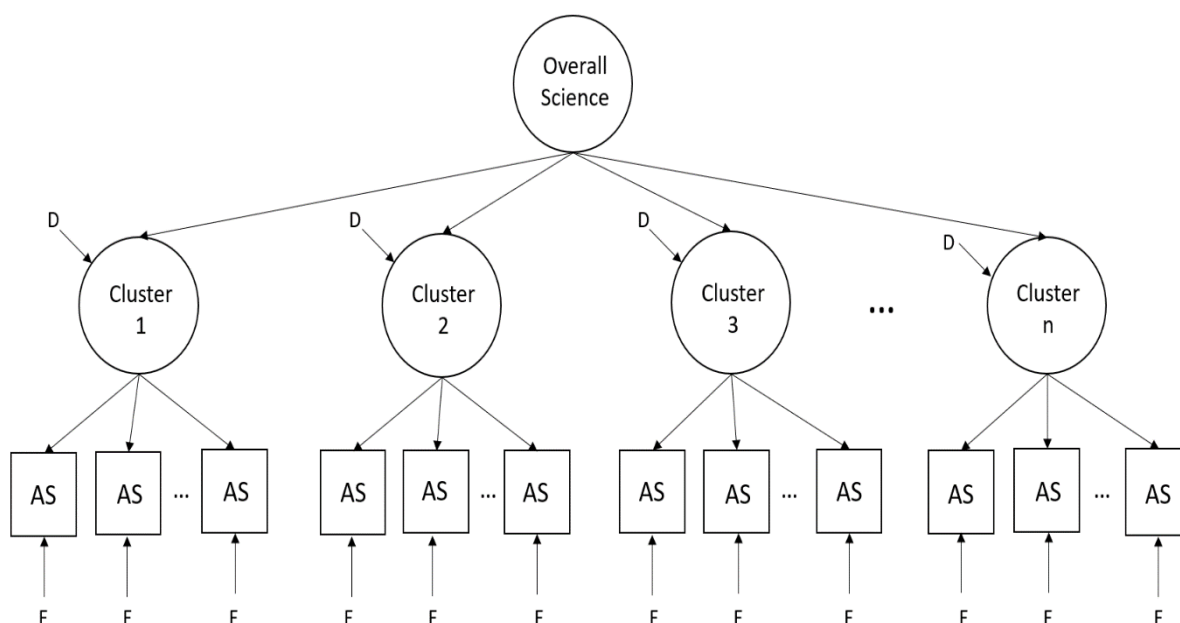
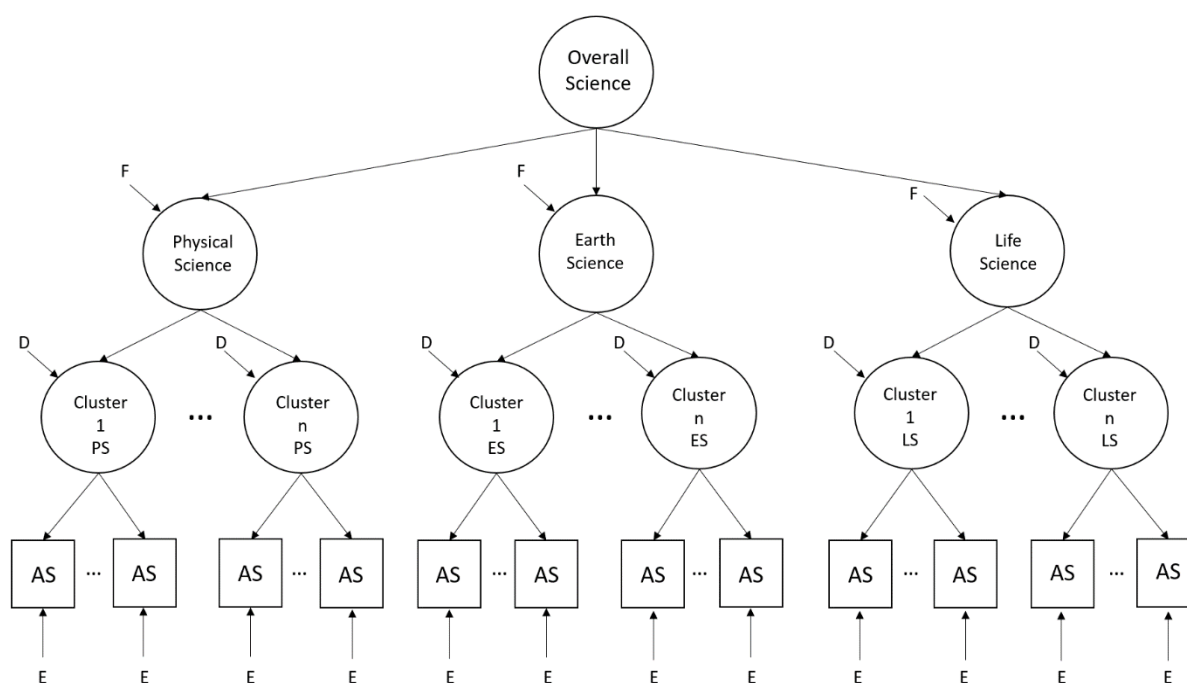


Figure 8. Third-Order Structural Model (Assertions-Clusters-Disciplines-Overall): “Model 4”



5.4.1 Results

For each test form, fit measures were computed for each of the four models. The fit measures used to evaluate goodness-of-fit were the comparative fit index (CFI), the Tucker-Lewis index (TLI), the root mean square error of approximation (RMSEA), and the standardized root mean residual (SRMR). CFI and TLI are relative fit indices, meaning they evaluate model fit by comparing the model of interest to a baseline model. RMSEA and SRMR are indices of absolute fit. Table 12 provides a list of these measures along with the corresponding thresholds indicating a good fit.

Table 12. Guidelines for Evaluating Goodness of Fit

Goodness-of-Fit Measure*	Indication of Good Fit
CFI	≥ 0.95
TLI	≥ 0.95
RMSEA	≤ 0.06
SRMR	≤ 0.08

*Brown, 2015; Hu & Bentler, 1999

Table 13 through Table 15 show the goodness-of-fit statistics for grades 6–8, respectively.¹ Numbers in bold indicate those indices that did not meet the criteria established in Table 12. Across all grades and models, the following conclusions can be drawn:

6. Model 1 shows the most misfit across grades and forms.
7. Across forms, Model 3 generally shows more improvement in model fit relative to Model 1 than Model 2 does (i.e., higher values for CFI and TLI and lower values for RMSEA and SRMR). This means that accounting for the clusters resulted in a higher improvement in model fit over a single factor model than accounting for disciplines.
8. Model 4 does not show improvement in model fit over Model 3. Fit measures remained the same (or had a difference of 0.001 or smaller in very few cases) across forms for Models 3 and 4. Thus, when clusters were taken into account, incorporating disciplines into the model did not improve model fit.
9. Overall model fit for Models 3 and 4 decreases with decreasing grades. For grade 8, all fit indices for Models 3 and 4 indicate good model fit for all three forms. For grade 7, all fit indices for Models 3 and 4 indicate good fit for two out of the six forms, and the degree of misfit for the other four forms is small. For grade 6, all three forms have fit indices above the threshold values for at least one of the absolute fit indices for Models 3 and 4. The amount of misfit is small for the RMSEA but more substantial for the SRMR for two out of the three forms.

¹ For very few assertions per form and models, some error variances were slightly below 0. For grade 6, 1–2 assertions per form and model had error variance below 0, with the lowest error variance being -0.027. For grade 7, Forms 1, 2, 5, and 6 had one negative error variance for one assertion in Models 3 and 4, with the lowest error variance being -0.099. Form 4 had 1–2 assertions with negative error variance in each model, and the lowest error variance was -0.102. For grade 8, there were no assertions with negative error variances for any of the forms and models.

Table 13. Fit Measures per Model and Form, Grade 6

Model	Form	CFI	TLI	RMSEA	SRMR
Model 1 Assertions-Overall (one-factor model)	1	0.995	0.995	0.106	0.163
	2	0.997	0.997	0.093	0.148
	3	0.995	0.995	0.109	0.161
Model 2 Assertions-Disciplines-Overall (second-order model)	1	0.996	0.996	0.089	0.144
	2	0.998	0.998	0.078	0.128
	3	0.997	0.997	0.087	0.135
Model 3 Assertions-Clusters-Overall (second-order model)	1	0.998	0.998	0.065	0.107
	2	0.999	0.999	0.056	0.095
	3	0.998	0.998	0.067	0.104
Model 4 Assertions-Clusters-Disciplines-Overall (third-order model)	1	0.998	0.998	0.065	0.107
	2	0.999	0.999	0.056	0.095
	3	0.998	0.998	0.067	0.104

Note. Numbers in bold do not meet the criteria for goodness of fit.

Table 14. Fit Measures per Model and Form, Grade 7

Model	Form	CFI	TLI	RMSEA	SRMR
Model 1 Assertions-Overall (one-factor model)	1	0.892	0.889	0.060	0.074
	2	0.938	0.936	0.083	0.109
	3	0.940	0.939	0.052	0.065
	4	0.937	0.936	0.068	0.114
	5	0.939	0.937	0.093	0.119
	6	0.898	0.895	0.056	0.071
Model 2 Assertions-Disciplines-Overall (second-order model)	1	0.908	0.906	0.055	0.073
	2	0.962	0.961	0.065	0.088
	3	0.950	0.949	0.048	0.063
	4	0.955	0.954	0.058	0.094
	5	0.959	0.957	0.077	0.103
	6	0.906	0.903	0.054	0.070
Model 3 Assertions-Clusters-Overall (second-order model)	1	0.938	0.937	0.046	0.072
	2	0.974	0.973	0.054	0.082
	3	0.967	0.966	0.039	0.055
	4	0.977	0.976	0.041	0.072
	5	0.975	0.974	0.060	0.089
	6	0.932	0.930	0.046	0.072
	1	0.939	0.937	0.045	0.072

Model	Form	CFI	TLI	RMSEA	SRMR
Model 4 Assertions-Clusters-Disciplines-Overall (third-order model)	2	0.974	0.973	0.054	0.082
	3	0.967	0.966	0.039	0.055
	4	0.977	0.976	0.041	0.072
	5	0.975	0.974	0.060	0.089
	6	0.932	0.930	0.046	0.072

Note. Numbers in bold do not meet the criteria for goodness of fit.

Table 15. Fit Measures per Model and Form, Grade 8

Model	Form	CFI	TLI	RMSEA	SRMR
Model 1 Assertions-Overall (one-factor model)	1	0.929	0.927	0.043	0.060
	2	0.959	0.958	0.042	0.056
	3	0.943	0.941	0.052	0.074
Model 2 Assertions-Disciplines-Overall (second-order model)	1	0.934	0.932	0.041	0.060
	2	0.963	0.963	0.040	0.056
	3	0.950	0.949	0.049	0.072
Model 3 Assertions-Clusters-Overall (second-order model)	1	0.953	0.952	0.034	0.057
	2	0.974	0.973	0.034	0.054
	3	0.970	0.969	0.038	0.064
Model 4 Assertions-Clusters-Disciplines-Overall (third-order model)	1	0.953	0.952	0.034	0.057
	2	0.974	0.974	0.033	0.053
	3	0.970	0.969	0.038	0.064

Note. Numbers in bold do not meet the criteria for goodness of fit.

For Models 3 and 4, grade 6 showed some degree of misfit across all three forms according to the measures of absolute model fit, especially for the SRMR. Further examination indicated that the lack of fit could be attributed to a single item that was common to all three grade 6 forms that were part of this factor analysis study. After removing this item, there were only two forms that had two or more clusters per discipline. The fit for both forms improved drastically in Models 3 and 4, with all fit measures except the SRMR for one form meeting the criteria for model fit. The SRMR value that exceeded the threshold value did so barely, with a value of 0.083. Table 16 shows the fit measures for grade 6 after removal of the item causing misfit. Note that, unlike Models 3 and 4, Models 1 and 2 still did not meet the criteria of model fit after removing the item.

Table 16. Fit Measures per Model and Form, Grade 6, with One Cluster Removed²

Model	Form	CFI	TLI	RMSEA	SRMR
Model 1 Assertions-Overall (one-factor model)	1	0.977	0.976	0.094	0.130
	2	0.974	0.973	0.082	0.118
Model 2 Assertions-Disciplines-Overall (second-order model)	1	0.986	0.986	0.072	0.106
	2	0.985	0.984	0.062	0.094
Model 3 Assertions-Clusters-Overall (second-order model)	1	0.992	0.991	0.057	0.083
	2	0.991	0.991	0.048	0.072
Model 4 Assertions-Clusters-Disciplines-Overall (third-order model)	1	0.992	0.991	0.057	0.083
	2	0.991	0.991	0.048	0.072

Note. Numbers in bold do not meet the criteria for goodness of fit.

Table 17 shows the estimated correlations among disciplines for Model 4 (third-order model). The correlations are all very high, ranging between 0.913 and 1. The high correlations between the disciplines in Model 4 indicate that, after considering the cluster effects, the disciplines do not add much to the model. This may explain why Model 4 did not show an improvement in fit compared to Model 3. Overall, the findings support the IRT model used for calibration.

Table 17. Model-Implied Correlations per Form for the Disciplines in Model 4

Grade	Form	Discipline	Earth and Space Sciences	Life Sciences
6	1	Physical Sciences	0.999	0.941
		Earth and Space Sciences	–	0.940
	2	Physical Sciences	1.000	0.964
		Earth and Space Sciences	–	0.964
	3	Physical Sciences	0.975	0.923
		Earth and Space Sciences	–	0.947
7	1	Physical Sciences	0.983	0.947
		Earth and Space Sciences	–	0.937
	2	Physical Sciences	0.978	0.972
		Earth and Space Sciences	–	0.951
	3	Physical Sciences	0.955	0.936
		Earth and Space Sciences	–	0.966
	4	Physical Sciences	0.938	0.913
		Earth and Space Sciences	–	0.973

² One assertion per model in form 1 and one assertion on three of the models in form 2 had error variance below 0, with the lowest error variance being -0.027.

Grade	Form	Discipline	Earth and Space Sciences	Life Sciences
8	5	Physical Sciences	0.931	0.944
		Earth and Space Sciences	–	0.965
	6	Physical Sciences	0.941	0.928
		Earth and Space Sciences	–	0.967
	1	Physical Sciences	0.971	0.971
		Earth and Space Sciences	–	0.970
	2	Physical Sciences	0.956	0.958
		Earth and Space Sciences	–	0.935
	3	Physical Sciences	0.966	0.978
		Earth and Space Sciences	–	0.988

5.4.2 Conclusion

The models with no cluster effects provided the highest degrees of misfit across forms and grades (Models 1 and 2), indicating that the cluster effects need to be taken into account as additional latent variables. On the other hand, once the cluster effects are accounted for, a single science dimension is sufficient (Model 3): including additional dimensions for the science disciplines (Life Science, Physical Science, Earth and Space Sciences) did not improve model fit and the correlations among those three dimensions are very high (Model 4). Model 3, with a single overall dimension for Science and additional latent variables to account for the effect of item clusters, provided the best balance between model fit and parsimony.

Overall, the findings support the use of the Rasch testlet model as the IRT calibration model and the reporting of an overall score directly computed from all the items a student took. Because there are enough items within each discipline in the test blueprint, discipline subscores can be reported at the individual level although they may not provide much unique information from the total score for most students. However, many stakeholders often desire information about student performance in addition to a single overall score. Note that it is not uncommon to provide subscores at the individual level even when the assessment is essentially unidimensional in a psychometric sense. For example, based on the dimensionality analyses for the Smarter Balanced Assessment, there is evidence suggesting “no consistent and pervasive multidimensionality was demonstrated” (Smarter Balanced Assessment Consortium, 2016, p.182) yet individual claim scores are routinely reported in addition to overall ELA and Mathematics scores.

6. FAIRNESS IN CONTENT

The principles of universal design of assessments provide guidelines for test design to minimize the impact of construct-irrelevant factors in assessing student achievement. Universal design removes barriers to provide access for the widest range of students possible. Seven principles of universal design are applied in the process of test development (Thompson, Johnstone, & Thurlow, 2002):

1. Inclusive assessment population
2. Precisely defined constructs
3. Accessible, non-biased items
4. Amenable to accommodations
5. Simple, clear, and intuitive instructions and procedures
6. Maximum readability and comprehensibility
7. Maximum legibility

Test development specialists have received extensive training on the principles of universal design and applied them in the development of all test materials. In the review process, adherence to the principles of universal design is verified by Connecticut educators and stakeholders. More details on how to reduce construct-irrelevant variance through universal design and on training on the principles of universal design are described in Section 2, Item Development Process That Supports Validity of Claims, as well as Appendix 2-C, Style Guide for Science Items, of Volume 2 of this technical report.

6.1 COGNITIVE LABORATORY STUDIES

In 2017, when the development of item clusters for the states that are part of the Memorandum of Understanding (MOU) began, cognitive lab studies were carried out to evaluate and refine the process of developing item clusters aligned to the Next Generation Science Standards (NGSS). Results of the cognitive lab studies confirmed the feasibility of the approach. Item clusters were completed within 12 minutes on average, and students reported being familiar with the format conventions and online tools used in the item clusters. They appeared to easily navigate the item clusters' interactive features and response formats. In general, students who received credit on a given item displayed a reasoning process that aligned with the skills that the item was intended to measure.

A second set of cognitive lab studies were carried out in 2018 and 2019 to determine if students using braille can understand the task demands of selected accommodated three-dimensional science standards-aligned item clusters and can navigate the interactive features of these clusters in a manner that allows them to fully display their knowledge and skills relative to the constructs of interest. In general, both the students who relied entirely on braille and/or the Job Access With Speech (JAWS) screen-reading software 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. The clusters

were clearly different from (and more complex than) other tests with which the students were familiar, however; and the study recommended that students be given adequate time to practice with at least one sample cluster before taking the summative test. The study also resulted in tool-specific recommendations for accessibility for visually impaired students. The reports of both sets of cognitive lab studies are presented in Appendix 4-D, Science Clusters Cognitive Lab Report, and Appendix 4-E, Braille Cognitive Lab Report.

6.2 STATISTICAL FAIRNESS IN ITEM STATISTICS

Differential item functioning (DIF) analyses were conducted with other states that field-tested the items for the initial item bank. A thorough content review was performed in those states. The details surrounding this review of items for bias is further described in Section 4.4 of Volume 1, Annual Technical Report, along with the DIF analysis process for the Connecticut NGSS Assessment.

7. SUMMARY

This volume is intended to provide a collection of reliability and validity evidence to support appropriate inferences from the observed test scores. The overall results can be summarized as follows:

- **Reliability.** Various measures of reliability are provided at the aggregate and subgroup levels, showing that the reliability of all tests is in line with acceptable industry standards.
- **Content Validity.** Evidence is provided to support the assertion that content coverage on each test was consistent with the test specifications of the blueprint across testing modes.
- **Internal Structural Validity.** Evidence is provided to support the selection of the measurement model, the tenability of model assumptions, and the reporting of an overall score and subscores at the reporting category levels.
- **Relationship of Test Scores to External Variables.** Evidence of convergent and discriminant validity is provided to support the relationship between the test and other measures intended to assess similar constructs, as well as between the test and other measures intended to assess different constructs.
- **Test Fairness.** Items are developed following the principles of universal design, which removes barriers to provide access for the widest range of students possible. Evidence of test fairness is provided statistically using DIF analysis in tandem with content reviews by specialists.

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.
- Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, 64, 153–168.
- Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York: The Guilford Press.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart, and Winston.
- Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, 9(4), 466–491.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1–55.
- Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement*, 30, 3–21.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13–103). New York: Macmillan.
- 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>
- Rijmen, F. (2009). *Three multidimensional models for testlet-based tests: Formal relations and an empirical comparison*. Educational Testing Service (ETS) Research Rep. No. RR–09–37, Princeton, NJ: ETS.
- Rijmen, F., Jiang, T., & Turhan, A. (2018, April). An item response theory model for new science assessments. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.
- Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, 48(2), 1–36.
- Smarter Balanced Assessment Consortium. (2016). *2013-2014 Technical Report*. Retrieved from <https://portal.smarterbalanced.org/library/en/2013-14-technical-report.pdf>

- Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments*. (Synthesis Report 44). Minneapolis: University of Minnesota, National Center on Educational Outcomes. Retrieved from <http://education.umn.edu/NCEO/OnlinePubs/Synthesis44.html>.
- Wang, W. C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement*, 29(2), 126–149.
- Yung, Y. F., Thissen, D., & McLeod, L. D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika*, 64, 113–128.

Connecticut Next Generation Science Standards Assessment

2023–2024

Volume 5: Test Administration



CONNECTICUT STATE
DEPARTMENT OF EDUCATION

TABLE OF CONTENTS

1.	TEST INTERVALS, OPTIONS, AND ADMINISTRATIVE ROLES	1
1.1	Testing Windows	1
1.2	Test Options and Administrative Roles	1
1.2.1	<i>Administrative Roles</i>	1
1.2.2	<i>Online Test Administration</i>	4
1.2.3	<i>Paper-Pencil Test Administration</i>	5
1.2.4	<i>Braille Test Administration</i>	5
2.	TRAINING AND INFORMATION FOR TEST COORDINATORS AND ADMINISTRATORS	6
2.1	Online Training	6
2.1.1	<i>TA Certification Course</i>	6
2.1.2	<i>Office Hour Webinars</i>	6
2.1.3	<i>Practice Site</i>	6
2.1.4	<i>Manuals and User Guides</i>	6
2.1.5	<i>Brochures and Quick Guides</i>	7
2.2	District Test Coordinator Training Workshops	8
3.	TEST SECURITY	9
3.1	Student-Level Testing Confidentiality	9
3.2	System Security	10
3.3	Security of the Testing Environment	10
3.3.1	<i>Duties of Testing Personnel</i>	10
3.3.2	<i>Room Preparation</i>	11
3.3.3	<i>Seating Arrangements</i>	11
3.3.4	<i>After the Test</i>	11
3.4	Test Security Violations	12
4.	STUDENT PARTICIPATION	12
4.1	Eligibility	12
4.2	Homeschooled Students	12
4.3	Exempt Students	12
5.	ONLINE TESTING FEATURES AND TESTING ACCOMMODATIONS	13

LIST OF TABLES

Table 1. Summary of Tests and Testing Options in 2023–2024	1
--	---

LIST OF APPENDICES

Appendix 5-A. Test Coordinator Manual
Appendix 5-B. Test Administration Manual
Appendix 5-C. Assessment Guidelines

1. TEST INTERVALS, OPTIONS, AND ADMINISTRATIVE ROLES

1.1 TESTING WINDOWS

The 2023–2024 Connecticut Next Generation Science Standards (NGSS) Assessment testing window spanned approximately two and a half months for the summative assessments and eight months for the interim assessments. The paper-pencil fixed-form tests for summative assessments were administered concurrently during the two-and-a-half-month online summative window.

1.2 TEST OPTIONS AND ADMINISTRATIVE ROLES

The Connecticut NGSS Assessments are administered primarily online. To ensure that all eligible students in the tested grades were given the opportunity to take the Connecticut NGSS Assessments, a number of assessment options were available for the 2023–2024 administration to accommodate students' needs. Table 1 lists the testing options that were offered in 2023–2024. Once a testing option was selected, it applied to all tests in the content area.

Table 1. Summary of Tests and Testing Options in 2023–2024

Assessments	Test Options	Test Mode
Summative Assessments	English	Online
	Braille	Online
	Spanish (toggle)	Online
	Paper-Pencil Large-Print Fixed-Form Test*	Paper-Pencil
	Paper-Pencil Braille Fixed-Form Test*	Paper-Pencil
Interim Assessments	English Braille Spanish (toggle)	Online

*For the paper-pencil fixed-form tests, all student responses on the paper-pencil tests were entered in the Data Entry Interface (DEI) by test administrators.

To ensure standardized administration conditions, teachers (TEs) and test administrators (TAs) followed procedures outlined in the NGSS *Test Administration Manual* (TAM, see Appendix 5-B). TEs and TAs reviewed the TAM before testing to ensure that the testing room was prepared appropriately (e.g., removing certain classroom posters, arranging desks). Make-up procedures were established for any students who were absent on the day(s) of testing. TEs and TAs followed required administration procedures and directions and read the boxed directions verbatim to students, ensuring standardized administration conditions.

1.2.1 Administrative Roles

The key personnel involved with the test administration for the Connecticut State Department of Education (CSDE) are District Administrators (DAs), District Test Coordinators (DCs), School Test Coordinators (SCs), Teachers (TEs), and Test Administrators (TAs). The main responsibilities of these key personnel are described in the following subsections. More detailed

descriptions can be found in the TAM provided online at <https://ct.portal.cambiumast.com/resources>.

District Administrator

The District Administrator (DA) may add users with District Test Coordinator (DC) roles in the Test Information Distribution Engine (TIDE). For example, a director of special education may need DC privileges in TIDE to access district-level data for the purposes of verifying test settings for designated supports and accommodations. DAs have the same test administration responsibilities as DCs. Their primary responsibility is to coordinate the administration of the Connecticut NGSS Assessment in the district.

District Test Coordinator

The District Test Coordinator (DC) is primarily responsible for coordinating the administration of the Connecticut NGSS Assessment at the district level.

DCs are responsible for the following:

- Reviewing all NGSS policies and test administration documents
- Reviewing scheduling and test requirements with SCs, TEs, and TAs
- Working with SCs and Technology Coordinators (TCs) to ensure that all systems, including the secure browser, are properly installed and functional
- Importing users (including SCs, TEs, and TAs) into TIDE
- Verifying all student information and eligibility in TIDE
- Scheduling and administering training sessions for all SCs, TEs, TAs, and TCs
- Ensuring that all personnel are trained in proper administration of the Connecticut NGSS Assessment
- Monitoring the secure administration of the tests
- Investigating and reporting all testing improprieties, irregularities, and breaches reported by the TEs and TAs
- Attending to any secure material according to CSDE policies

School Test Coordinator

The School Test Coordinator (SC) is primarily responsible for coordinating the administration of the Connecticut NGSS Assessment at the school level and ensuring that testing in their school is conducted in accordance with the test procedures and security policies established by the CSDE. SC responsibilities include the following:

- Based on test administration windows, establishing a testing schedule with DCs, TEs, and TAs

- Working with technology staff to ensure timely computer setup and installation
- Working with TEs and TAs to review student information in TIDE to ensure that student information and test settings for designated supports and accommodations are correctly applied
- Identifying students who may require designated supports and test accommodations and ensuring that procedures for testing these students follow CSDE policies
- Attending all district trainings and reviewing all CSDE policies and test administration documents
- Ensuring that all TEs and TAs attend school or district trainings and review online training modules posted on the portal
- Establishing secure and separate testing rooms if needed
- Downloading and planning the administration of the classroom activity with TEs and TAs
- Monitoring secure administration of the tests
- Monitoring testing progress during the testing window and ensuring that all students participate, as appropriate
- Investigating and reporting all testing improprieties, irregularities, and breaches reported by the TEs and TAs
- Attending to any secure material according to CSDE policies

Teacher

A teacher (TE) who is responsible for administering the Connecticut NGSS Assessment must have the same qualifications as a Test Administrator (TA). TEs also have the same test administration responsibilities as TAs. TEs can view their own students' results that are rostered to them when they are made available. This role may also be assigned to teachers who do not administer the test but will need access to student results.

Test Administrator

A Test Administrator (TA) is primarily responsible for administering the Connecticut NGSS Assessment. The TA's role does not allow access to student results and is designed for TAs, such as technology staff, who administer tests but do not have access to student results.

TAs are responsible for the following:

- Completing NGSS test administration training
- Reviewing all CSDE policy and test administration documents before administering any Connecticut NGSS Assessments

- Viewing student information before testing to ensure that a student receives the proper test with the appropriate supports and reporting any potential data errors to SCs and DCs, as appropriate
- Administering the Connecticut NGSS Assessment
- Reporting all potential test security incidents to the SCs and DCs in a manner consistent with CSDE and district policies

1.2.2 Online Test Administration

Within Connecticut’s testing window, schools can set testing schedules, allowing students to test in intervals (e.g., multiple sessions) rather than in one long test period, which minimizes the interruption of classroom instruction and efficiently utilizes its facility. With online testing, schools do not need to handle test booklets and address the storage and security problems inherent in sending large shipments of materials to a school site.

SCs oversee all aspects of testing at their schools and serve as the main point of contact, while TEs and TAs administer the online assessments only. TEs and TAs are trained in the online testing requirements and the mechanics of starting, pausing, and ending a test session. Training materials for the test administration are provided online.

To start a test session, the TE or TA must first enter the TA Interface of the online testing system using their own computer. A session ID is generated when the test session is created. Students who are taking the assessment with the TE or TA must enter their State Student Identification Number (SSID), first name, and session ID into the Student Interface using computers provided by the school. The TE or TA then verifies that the students are taking the appropriate assessments with the appropriate accessibility feature(s); see Appendix 5-C for a list of accommodations. Students can begin testing only after the TA or TE has confirmed the settings. The TA or TE then reads aloud the *Directions for Administration* in the NGSS TAM to the students and guides them through the login process.

Once an assessment has started, the student must answer the test question presented on a page before proceeding to the next page. Skipping questions is not permitted. For the online computer test, students are allowed to scroll back to review and edit previously answered items, as long as these items are in the same test session and this session has not been paused for more than 20 minutes. Students may review and edit responses they have previously provided before submitting the assessment. During an active online computer test session, if a student reviews and changes the response to a previously answered item, then all items that follow to which the student already responded remain the same. If a student changes the answers, no new items are assigned. For example, a student pauses for 10 minutes after completing item 10. After the pause, the student goes back to item 5 and changes the answer. If the response change in item 5 changes the item score from wrong to right, the student’s overall score will improve; however, there will be no change in items 6–10.

For the summative test, an assessment can be started on one day and completed on another. For the online computer test, the assessment must be completed within 45 calendar days of the start date, or the assessment opportunity will expire.

During a test session, TEs or TAs may pause the test for a student or group of students to take a break. It is up to the TEs or TAs to determine an appropriate stopping point; however, to ensure the integrity of test scores or testing, the online computer test cannot be paused for more than 30 minutes. If that happens, the student must begin a new test session starting where the student left off. Previous responses are no longer available for viewing or editing.

The TAs or TEs must remain in the room at all times during a test session in order to monitor student testing. Once the test session ends, the TAs or TEs must ensure that each student has successfully logged out of the system. Then the TAs or TEs must collect any handouts or scratch paper that students used during the assessment and send them for secure shredding.

1.2.3 Paper-Pencil Test Administration

The paper-pencil versions of the Connecticut NGSS Assessments are provided as an accommodation for students for whom the online test is not accessible. For Connecticut students, paper-pencil tests were offered only in braille and large print as documented in their IEP or Section 504 plan.

The DA must order the accommodated test materials on behalf of the students who need to take the paper-pencil test via the Test Information Distribution Engine (TIDE). Based on the paper-pencil orders submitted in TIDE, the testing contractor ships the appropriate test booklets and the paper-pencil TAM to the district.

After the student has completed the assessment, the TEs and TAs enter the student responses into the Data Entry Interface (DEI) and return the test booklets to the testing vendor. The tests submitted via the DEI are then scored.

1.2.4 Braille Test Administration

The Science fixed-form braille test was available with the same test blueprint.

The braille interface is described as follows:

- The braille interface included a text-to-speech component for mathematics consistent with the read-aloud assessment accommodation. The Job Access with Speech (JAWS) screen-reading software provided by Freedom Scientific was an essential component that students used with the braille interface.
- Science items were presented to students in UEB Contracted with Nemeth Braille code.

Before administering the online summative assessments using the braille interface, TEs or TAs ensured that the technical requirements were met. These requirements applied to the student's computer, the TE's or TA's computer, and any supporting braille technologies used in conjunction with the braille interface.

2. TRAINING AND INFORMATION FOR TEST COORDINATORS AND ADMINISTRATORS

All DAs, DCs, and SCs oversaw all aspects of testing at their schools and served as the main points of contact, and TEs and TAs administered the online assessments. The online CAI TA Certification Course, webinars, user guides, manuals, and training sites were used to train TEs and TAs in the online testing requirements and the mechanics of starting, pausing, and ending a test session. Training materials for test administration were provided online.

2.1 ONLINE TRAINING

Multiple online training opportunities were offered to key staff.

2.1.1 TA Certification Course

CAI's online TA Certification Course was available as an optional course to any user in TIDE. This web-based course was about 30–45 minutes long and covered information on testing policies and steps for administering a test session in the online system. This interactive course required participants to start test sessions under different scenarios. Throughout the training and at the end of the course, participants were required to answer multiple-choice questions about the information provided.

2.1.2 Office Hour Webinars

During the testing window, the CSDE and CAI held office hours every Thursday from 3:00 p.m. to 4:00 p.m. During office hours, the CSDE and CAI staff provided brief, weekly assessment updates and were available for phone support to answer questions from districts. All office-hour sessions were recorded, and the recordings were posted to the portal.

2.1.3 Practice Site

In October 2018, a practice site was opened for TEs/TAs and students. TEs and TAs could practice administering assessments and starting and ending test sessions on the TA Training Site, and students could practice taking a short online assessment on the Student Practice and Training Site. The Connecticut NGSS Assessment practice tests contained the same item types (stand-alone and clusters) students would encounter on the Connecticut NGSS Summative Assessment. The practice tests were designed to provide students and teachers with opportunities to quickly familiarize themselves with the software and navigational tools they would use for the Connecticut NGSS Summative Assessment. Practice tests were organized by grade bands (grades 5, 8, and 11). The practice test was refreshed in August 2023.

A student could log in directly to the practice and training test site as a guest without a TA-generated test session ID, or the student could log in through a practice test session created by the TE or TA.

2.1.4 Manuals and User Guides

The following manuals and user guides were available on the Connecticut portal, <https://ct.portal.cambiumast.com/>.

The *Test Coordinator Manual* (Appendix 5-A) provides information for DCs and SCs regarding policies and procedures for the 2023 NGSS assessments.

The *NGSS Test Administration Manual* (Appendix 5-B) provides information for TEs and TAs administering the NGSS online summative assessments. It includes screen captures and step-by-step instructions on how to administer the online tests.

The *Assistive Technology Manual* provides an overview of the embedded and non-embedded assistive technology tools that can be used to help students with specific accessibility needs complete online tests in the Test Delivery System (TDS). It includes lists of supported devices and applications for each type of assistive technology that students may need, as well as setup instructions for the assistive technologies that require additional configuration in order to work with the TDS.

The technology resource manuals contain technology requirements and instructions that will assist TCs in preparing computers and devices for online testing. A guide is created for each of the approved operating systems (Windows, Mac, iPad, Linux, ChromeOS).

The *Test Information Distribution Engine User Guide* was designed to help users navigate TIDE. It provides information on managing user account information, student account information, student test settings and accommodations, appeals, and voice packs.

The *Centralized Reporting System User Guide* provides information about the Centralized Reporting System (CRS), including instructions for viewing score reports, accessing test management resources, creating and editing rosters, and searching for students for both interim and summative assessments.

The *Test Administrator User Guide* was designed to help users navigate the TDS, including the Student Interface and the TA Interface, and to help TEs/TAs manage and administer online testing for students.

The *Assessment Viewing Application User Guide* provides an overview of how to access and use the Assessment Viewing Application (AVA). AVA allows teachers to view items on the NGSS interim assessments.

All manuals and user guides pertaining to the 2023–2024 online assessments were available on the portal, and DAs, DCs, and SCs used the manuals and user guides to train TAs and TEs in test administration policies and procedures.

2.1.5 Brochures and Quick Guides

The following brochures and quick guides were available on the CT portal, <https://ct.portal.cambiumast.com/>.

Accessing Participation Reports: This brochure provided instructions on how to extract participation reports for the NGSS assessments.

Accessing TIDE: This brochure provided a brief overview of user management in the Test Information Distribution Engine (TIDE) and instructions on how to log in to the system. School

personnel needed to use TIDE account credentials to access all secure online systems used to administer Connecticut Comprehensive Assessment Program online assessments.

Embedded and Non-Embedded Designated Supports for English Learners: This brochure provided recommendations for students who were English learners (ELs) on what supports they might benefit from when participating on the Connecticut statewide assessments. These designated supports were intended as a language support for students with limited English language skills, whether they were identified in the Public-School Information Systems (PSIS) as EL or EL with a disability. The use of these supports may have resulted in the students' needing additional overall time to complete the assessment.

How to Access the Data Entry Interface (DEI): This brochure described how to access the Data Entry Interface (DEI) in order to submit the NGSS paper-pencil tests.

Next Generation Science Standards Interim Assessment Quick Guide: This document provided a step-by-step guide on how to start a test session for the NGSS interim assessments. It included a complete list of all interim test labels as they appeared in the TA Interface.

Managing Student Test Settings Brochure: This brochure provided a brief overview on how to manage student test settings in TIDE. Students' embedded accommodations, non-embedded accommodations, and designated supports were set in TIDE prior to test administration so that these settings could be reflected in the TDS.

Monitoring Test Progress: Test Status Code Report and Test Completion Rates: This brochure contained instructions for generating Test Status Code Reports and Test Completion Rates in TIDE. These are excellent tools that should be used to track test completion for students at both the district and the school level.

User Role Permissions for Online Systems Brochure: This brochure outlined the user roles and permissions for each secure online testing system used to administer the online assessments for the Connecticut Comprehensive Assessment Program. These systems included the Test Information Distribution Engine (TIDE), Centralized Reporting System (CRS), Test Administration (TA) Interface, Data Entry Interface (DEI), and Assessment Viewing Application (AVA).

Understanding and Creating Rosters: Rosters are groups of students associated with a teacher in a particular school. Rosters typically represent entire classrooms in lower grades, or individual classroom periods in upper grades. This document provided instructions on how to create, view, and modify rosters in TIDE and in the CRS.

2.2 DISTRICT TEST COORDINATOR TRAINING WORKSHOPS

District Test Coordinator (DC) training workshops were held January 17 - 19, 2024. Training was provided for the administration of the Connecticut NGSS Assessment. During the training, DCs were provided with information to support training of the SCs, TEs, and TAs.

3. TEST SECURITY

All test items, test materials, and student-level testing information are considered secure materials for all assessments. The importance of maintaining test security and the integrity of test items is stressed throughout the webinar trainings and in the user guides, modules, and manuals. Features in the testing system also protect test security. This section describes system security, student confidentiality, and policies on testing improprieties.

3.1 STUDENT-LEVEL TESTING CONFIDENTIALITY

All secure websites and software systems enforce role-based security models that protect individual privacy and confidentiality in a manner consistent with the Family Educational Rights and Privacy Act (FERPA) and other federal laws. Secure transmission and password-protected access are basic features of the current system and ensure authorized data access. All aspects of the system—including item development and review, test delivery, and score reporting—are secured by password-protected logins. Our systems use role-based security models to ensure that users may access only the data to which they are entitled and may edit data only in accordance with their user rights.

There are three dimensions related to identifying that students are accessing appropriate test content:

1. *Test eligibility* refers to the assignment of a test to a particular student.
2. *Test accommodation* refers to the assignment of a test setting to specific students based on need.
3. *Test session* refers to the authentication process of a TE/TA creating and managing a test session, the TE/TA reviewing and approving a test (and its settings) for every student, and the student signing on to take the test.

FERPA prohibits public disclosure of student information or test results. The following are examples of prohibited practices:

- Providing login information (username and password) to other authorized TIDE users or to unauthorized individuals
- Sending a student's name and SSID number together in an email message (If information must be sent via email or fax, include only the SSID number, not the student's name.)
- Having students log in and test under another student's SSID number

Test materials and score reports should not be exposed to identify student names with test scores except by authorized individuals with an appropriate need to know.

All students, including homeschooled students, had to be enrolled or registered at their testing schools in order to take the online, paper-pencil, or braille assessments. Student enrollment information, including demographic data, was generated using a CSDE file and uploaded nightly via a secured file transfer site to the CRS during the testing period.

Students logged in to the online assessment using their legal first name, SSID number, and a test session ID. Only students could log in to an online test session. TEs/TAs, proctors, and other personnel were not permitted to log in to the system on behalf of students, although they were permitted to assist students who needed help logging in. For the paper-pencil versions of the assessments, TEs and TAs were required to affix the student label to the student's answer document.

After a test session, only staff with the administrative roles of DA, DC, SC, or TE were permitted to view their students' scores. TAs did not have access to student scores.

3.2 SYSTEM SECURITY

The objective of system security is to ensure that all data are protected and accessed appropriately by the designated user groups. It is about protecting data and maintaining data and system integrity as intended, including ensuring that all personal information is secured, that transferred data (whether sent or received) is not altered in any way, that the data source is known, and that any service can be performed only by a specific, designated user.

A Hierarchy of Control: As described in Section 1.2.1, Administrative Roles, all DAs, DCs, SCs, TAs, and TEs have defined roles and levels of access to the testing system. When the TIDE testing window opens, the CSDE provides a verified list of DAs to the testing contractor, who uploads the information into TIDE. DAs are then responsible for selecting and entering the DTs' and SCs' information into TIDE, and the SC is responsible for entering TA and TE information into TIDE. Throughout the year, the DA, DC, and SC are also expected to delete information in TIDE for any staff members who have transferred to other schools, have resigned, or no longer serve as TAs or TEs.

Password Protection: All access points by different roles—at the state, district, school principal, and school staff levels—require a password to log in to the system. Newly added SCs, TAs, and TEs receive separate passwords through their personal email addresses assigned by the school.

Secure Browser: A key role of the Technology Coordinator (TC) is to ensure that the secure browser is properly installed on the computers used for the administration of the online assessments. Developed by the testing contractor, the secure browser prevents students from accessing other computers or Internet applications and from copying test information. The secure browser suppresses access to commonly used browsers, such as Internet Explorer and Firefox, and prevents students from searching for answers on the Internet or communicating with other students. The assessments can be accessed only through the secure browser and not by other Internet browsers.

3.3 SECURITY OF THE TESTING ENVIRONMENT

3.3.1 Duties of Testing Personnel

The SCs, TEs, and TAs work together to determine appropriate testing schedules based on the number of computers available, the number of students in each tested grade, and the average amount of time needed to complete each assessment.

Testing personnel are reminded in the online training and user manuals that assessments should be administered in testing rooms that do not crowd students. Good lighting, ventilation, and freedom from noise and interruption are important factors to consider when selecting testing rooms.

TEs and TAs must establish procedures to maintain a quiet environment during each test session, recognizing that some students may finish more quickly than others. If students are allowed to leave the testing room when they finish, TEs or TAs are required to explain the procedures for leaving and where students are expected to report once they leave without disrupting others. If students are expected to remain in the testing room until the end of the session, TEs or TAs are encouraged to prepare some quiet work for these students to do after they finish the assessment.

If a student needs to leave the room for a brief time during testing, the TAs or TEs are required to pause the student's assessment. For the online computer test, if the pause lasts longer than 20 minutes, the student can continue with the rest of the assessment in a new test session, but the system will not allow the student to return to the items answered before the pause. This measure is implemented to prevent students from using the time outside of the testing room to look up answers.

3.3.2 Room Preparation

The room should be prepared before the test session. Any information displayed on bulletin boards, chalkboards, or charts that students might use to help answer test questions should be removed or covered. This rule applies to rubrics, vocabulary charts, student work, posters, graphs, content area strategies charts, and other materials. The cell phones of both testing personnel and students must be turned off and stored in the testing room out of sight. TAs are encouraged to minimize access to the testing rooms by posting signs in halls and entrances in order to promote optimum testing conditions; they should also post “TESTING—DO NOT DISTURB” signs on the doors of testing rooms.

3.3.3 Seating Arrangements

TEs and TAs should provide adequate space between students' seats. Students should be seated so that they will not be tempted to look at others' answers. Because the online computer test is linear on the fly, it is unlikely that students will see the same test questions as other students. However, through appropriate seating arrangements, students should be discouraged from communication with each other.

3.3.4 After the Test

At the end of the test session, TEs or TAs must walk through the classroom to pick up any scratch paper that students used and any papers that display students' SSID numbers and names together. These materials should be securely shredded or stored in a locked area immediately.

For the paper-pencil versions, specific instructions on how to package and secure the test booklets to be returned to the testing contractor's office are provided in the *Paper and Pencil Test Administration Manual*.

3.4 TEST SECURITY VIOLATIONS

Everyone who administers or proctors the assessments is responsible for understanding the security procedures to administer them. Prohibited practices as detailed in the NGSS TAM are categorized into three groups:

1. *Impropriety*: This is a test security incident that has a low impact on the individual or group of students who are testing and has a low risk of potentially affecting student performance on the test, test security, or test validity (for example: student[s] leaving the testing room without authorization).
2. *Irregularity*: This is a test security incident that impacts an individual or group of students who are testing and may potentially affect student performance on the test, test security, or test validity. These circumstances can be contained at the local level (for example: disruption during the test session, such as a fire drill).
3. *Breach*: This is a test security incident that poses a threat to the validity of the test. Breaches require immediate attention and escalation to the CSDE. Examples may include such situations as exposure of secure materials or a repeatable security/system risk. These circumstances have external implications (for example: administrators modifying student answers, or students sharing test items through social media).

District and school personnel are required to document all test security incidents in the test security incident log. The log serves as the document of record for all test security incidents and should be maintained at the district level and submitted to the CSDE at the end of testing.

4. STUDENT PARTICIPATION

4.1 ELIGIBILITY

All students (including retained students) enrolled in grades 5, 8, and 11 at public schools in Connecticut were required to participate in the Connecticut NGSS Assessment. Students had to be tested in the enrolled grade assessment; out-of-grade-level testing was not allowed for the administration of the Connecticut NGSS Summative Assessment.

4.2 HOMESCHOOLED STUDENTS

Home-schooled students are not public-school students and are not eligible to be administered state assessments.

4.3 EXEMPT STUDENTS

The following students were exempt from participating in the Connecticut NGSS Assessment:

- Students who are hospitalized or homebound due to illness should be tested unless there are medical constraints that prevent them from testing and the students have received approved medical exemptions.

5. ONLINE TESTING FEATURES AND TESTING ACCOMMODATIONS

The *CSDE Assessment Guidelines* are intended for school-level personnel and decision-making teams, including Individualized Education Program (IEP) and Section 504 Plan teams, as they prepare for and implement the Connecticut NGSS Assessment. The *Guidelines* provide information for classroom teachers, English language development educators, special education teachers, and instructional assistants to use in selecting and administering universal tools, designated supports, and accommodations for those students who need them. The *Guidelines* are also intended for assessment staff and administrators who oversee the decisions that are made in instruction and assessment.

The *Connecticut Assessment Guidelines* apply to all students. They emphasize an individualized approach to the implementation of assessment practices for those students who have diverse needs and participate in large-scale content assessments. They focus on universal tools, designated supports, and accommodations for the NGSS assessments. At the same time, the *Guidelines* support important instructional decisions about accessibility and accommodations for students who participate in the Connecticut NGSS Assessment.

The summative assessments contain universal tools, designated supports, and accommodations in both embedded and non-embedded versions. Embedded resources are part of the computer administration system, whereas non-embedded resources are provided outside of that system.

State-level users, DCs, and SCs can set embedded and non-embedded designated supports and accommodations based on their specific user role. Designated supports and accommodations must be set in TIDE before starting a test session.

All embedded and non-embedded universal tools will be activated for use by all students during a test session. One or more of the pre-selected universal tools can be deactivated by a TE/TA in the TA Interface of the testing system for a student who may be distracted by the ability to access a specific tool during a test session.

For additional information about the availability of designated supports and accommodations, refer to Appendix 5-C, *Assessment Guidelines*.

Connecticut Next Generation Science Standards Assessment

2023–2024

Volume 6: Score Interpretation Guide



CONNECTICUT STATE
DEPARTMENT OF EDUCATION

TABLE OF CONTENTS

1.	CONNECTICUT SCORE REPORTS	1
1.1	Overview of Connecticut’s Score Reports.....	1
1.2	Overall Scores and Discipline-Level Scores.....	1
1.3	Centralized Reporting System	3
1.4	Available Reports on the Connecticut Centralized Reporting System	3
1.4.1	Reporting by Subgroup	4
1.4.2	Summary Performance Report.....	5
1.4.3	Aggregate-Level Subject Report	6
1.4.4	Aggregate-Level Discipline-Level Report.....	8
1.4.5	Aggregate-Level Disciplinary Core Ideas (DCI) and Science and Engineering Practices (SEP) Report.....	9
1.4.6	Student-Level Subject Report.....	10
1.4.7	Student-Level Discipline-Level Report	11
1.4.8	Individual Student Report	12
1.4.9	Data File	14
1.5	Test information Distribution Engine	15
1.6	Paper Individual Student Reports for Families	15
2.	INTERPRETATION OF REPORTED SCORES.....	16
2.1	Scale Score.....	16
2.2	Standard Error of Measurement.....	16
2.3	Performance Level	17
2.4	Performance Category for Discipline Levels.....	17
2.5	Cut Scores	17
2.6	Aggregated Scores	18
2.7	Relative Strengths and Weaknesses for Disciplinary Core Ideas and Science and Engineering Practices	18
2.8	Appropriate Uses for Scores and Reports	19
3.	SUMMARY	20

LIST OF TABLES

Table 1. Disciplines and Discipline-Level Claims for Science	3
Table 2. Connecticut Score Reports Summary	4
Table 3. Connecticut List of Subgroups	5
Table 4. Connecticut NGSS Assessment Science Performance-Level Cut Scores	18

LIST OF FIGURES

Figure 1. District-Level Summary Performance Distribution Report	6
Figure 2. District Aggregate-Level Subject Report for Grade 11 Science	6
Figure 3. District Aggregate-Level Subject Report for Grade 11 Science by Gender	7
Figure 4. District Aggregate-Level Discipline-Level Report for Grade 11 Science	8
Figure 5. District Aggregate-Level Disciplinary Core Idea (DCI) and Science Engineering Practices (SEP) Report for Grade 11 Science	10
Figure 6. Student Roster Subject Report for Grade 11 Science	11
Figure 7. Student Roster Discipline Report for Grade 11 Science	12
Figure 8. Individual Student Report for Grade 11 Science	13
Figure 9. Data File	15

LIST OF APPENDICES

Appendix 6-A. <i>Centralized Reporting System User Guide</i>
Appendix 6-B. Sample Printed Individual Student Report

1. CONNECTICUT SCORE REPORTS

In spring 2024, the Connecticut Next Generation Science Standards (NGSS) Assessment was administered to Connecticut students in grades 5, 8, and 11. The purpose of the *Score Interpretation Guide* is to document the features of the Connecticut Centralized Reporting System (CRS), which is designed to assist stakeholders in reviewing and downloading the test results and understanding and appropriately using the results of the state assessments. Additionally, this volume describes the score types reported for the spring 2024 assessments, the appropriate uses and inferences that can be drawn from these score types, and features of the score report.

1.1 OVERVIEW OF CONNECTICUT’S SCORE REPORTS

The Connecticut NGSS Assessment was first administered operationally statewide in spring 2019. Due to the COVID-19 pandemic, testing in spring 2020 was cancelled. Testing resumed in spring 2021. Test scores from the spring 2024 assessment were provided to districts and schools through the CRS on May 17, 2024. The CRS provided information on student performance and aggregated summaries at several levels—district, school, and roster.

The CRS (<https://ct.portal.cambiumast.com/>) is a web-based application that provides Connecticut NGSS Assessment results at various levels. Test results are available to users based on their roles and the privileges they receive based on the authentication granted to them. There are four basic levels of user roles, including the state, district, school, and teacher levels. Each user is granted drill-down access to reports in the system based on their assigned role. This means that teachers can access data for their roster(s) of students only, schools can access data for the students in their school only, and districts can access data for all schools and students in their district only.

The following users have access to the system:

- **State Users** have access to all data at the state, district, school, teacher, and student levels.
- **District Administrator (DA), District Test Coordinator (DC), District Reporting (DR)** have access to all data for their district and the schools and students in their district.
- **School Test Coordinator (SC) and School Administrator (SA)** have access to all data for their school and the students in their school.
- **Teacher Examiner (TE)** have access to all aggregate data for their rosters and the students within their rosters.

Access to reports is password protected, and users can access data at and below their assigned level. For example, an SC user can access the school report of students for their school but not for another school.

1.2 OVERALL SCORES AND DISCIPLINE-LEVEL SCORES

Each student receives a single scale score for each subject tested if there is a valid score to report. A student’s score is based only on the operational items on the assessment. A scale score is used to describe how well a student performed on a test and can be interpreted as an estimate of the

student’s knowledge and skills measured. The scale score is transformed from a theta score, which is estimated based on mathematical models. Low scale scores can be interpreted as an indication that the student does not possess sufficient knowledge and skills measured by the test. Conversely, high scale scores can be interpreted as an indication that the student has proficient knowledge and skills measured by the test. Interpretation of scale scores is more meaningful when the scale scores are used along with performance levels and performance-level descriptors (PLDs).

Based on the scale score, students will receive an overall performance level. Performance levels are proficiency categories on a test, which students fall into based on their scale scores. For the Connecticut NGSS Assessment, scale scores are mapped into four performance levels:

1. Does Not Meet Standard
2. Approaching Standard
3. Meets Standard
4. Exceeds Standard

For details on the standard-setting process, refer to Volume 3, Setting Performance Standards, of this technical report.

PLDs are a description of content area, knowledge, and skills that students at each performance level are expected to possess. Thus, performance levels can be interpreted based on PLDs. Generally, students performing on the Connecticut NGSS Assessment at Levels 3 and 4 are considered on track to demonstrate progress toward mastery of the knowledge and skills necessary for college and career readiness.

In addition to an overall score, students will receive discipline-level scores. For the Connecticut NGSS Assessment, student performance on each discipline level is reported on three performance categories:

1. Below Standard
2. Approaching Standard
3. Above Standard

Unlike the performance levels for the overall test, student performance on each of the discipline levels is evaluated with respect to the *Meets Standard* performance standard. Student performance at either *Below Standard* or *Above Standard* can be interpreted as student performance clearly below or above the *Meets Standard* cut score for a specific discipline. Student performance at *Approaching Standard* can be interpreted as student performance that does not provide enough information to tell whether students reached the *Meets Standard* mark for the specific discipline.

Table 1 displays the disciplines and discipline level claims for science, by grade.

Table 1. Disciplines and Discipline-Level Claims for Science

Grade	Discipline	Claim
5, 8, 11	Practices and Concepts in Life Sciences	The student is able to use the science and engineering practices to demonstrate understanding of the disciplinary core ideas and crosscutting concepts in Life Sciences.
	Practices and Concepts in Physical Sciences	The student is able to use the science and engineering practices to demonstrate understanding of the disciplinary core ideas and crosscutting concepts in Physical Sciences.
	Practices and Concepts in Earth and Space Sciences	The student is able to use the science and engineering practices to demonstrate understanding of the disciplinary core ideas and crosscutting concepts in Earth and Space Sciences.

1.3 CENTRALIZED REPORTING SYSTEM

The Centralized Reporting System (CRS) generates a set of online score reports that describe student performance for students, families, educators, and other stakeholders. The online score reports are produced after the tests are submitted by the students, hand-scored and machine-scored, and finally processed into the CRS. In addition to each individual student’s score report, the CRS produces aggregate score reports for teachers, schools, districts, and states.

Furthermore, to facilitate comparisons, each aggregate report contains the summary results for the selected aggregate unit, as well as all aggregate units above the selected aggregate. For example, if a school is selected, the summary results of the district to which the school belongs are provided so that the school performance can be compared with the district performance. If a teacher is selected, the summary results for the school and the district above the teacher are also provided for comparison purposes.

1.4 AVAILABLE REPORTS ON THE CONNECTICUT CENTRALIZED REPORTING SYSTEM

The Connecticut CRS is hierarchically structured. An authorized user is able to view reports at their own aggregated unit and any lower level of aggregation. For example, a school user can view only the reports and data at the school and student levels of their school. DA users can view the reports and data for their districts and also the student-level results for all of their schools.

Table 2 summarizes the types of score reports that are available in the CRS and the levels at which the reports can be viewed. A description of each report is also provided. Data files are also accessible for districts to download.

For detailed information on available reports and available features, educators can refer to the CRS user guide. The 2023–2024 *Centralized Reporting System User Guide* is included in Appendix 6-A, *Centralized Reporting System User Guide*.

Table 2. Connecticut Score Reports Summary

Report	Description	Unit of Aggregation				
		State	District	School	Roster	Student
Summary Performance	Summary of performance (to date) across grades and subjects or courses for the current administration	✓	✓	✓	✓	
Aggregate-Level Subject Report	Summary of overall performance for a subject and a grade for all students in the defined level of aggregation	✓	✓	✓	✓	
Aggregate-Level Discipline-Level Score Report	Summary of overall performance on each discipline level for each grade across all students within the selected level of aggregation	✓	✓	✓	✓	
Aggregate-Level Disciplinary Core Ideas (DCI) and Science and Engineering Practice (SEP) Report	Summary of overall performance on each disciplinary core idea and each science and engineering practice for a given subject and grade across all students within the selected level of aggregation	✓	✓	✓	✓	
Student-Level Subject Report	List of all students who belong to a school, teacher, or roster with their associated subject or course scores for the current administration			✓	✓	✓
Student-Level Discipline-Level Score Report	List of all students who belong to a school, teacher, or roster with their associated discipline-level performance for the current administration			✓	✓	✓
Individual Student Report (ISR)	Detailed information about a selected student's performance in a specified subject or course; includes overall subject and discipline level results					✓
Data Files	Text/CSV file containing overall and discipline-level scale scores and performance levels along with demographic information		✓	✓	✓	✓

1.4.1 Reporting by Subgroup

The aggregate score reports provide overall student results by default but can at any time be analyzed by subgroups based on demographic data. When used on aggregate-level reports, an additional level of analysis will be provided by aggregating students based on subgroup. For example, when the Gender subgroup is selected, the CRS will display aggregate results for all students, male and female. When used on student-level reports, subgroups can instead be used to

filter individual results. For example, a user will have the option to select Male or Female after the Gender subgroup is selected.

Users can see student assessment results by any subgroup at any time by selecting the desired subgroup from the *Breakdown By* drop-down menu available. Table 3 presents the types of subgroups and subgroup categories provided in the CRS.

Table 3. Connecticut List of Subgroups

Breakdown by Category	Displayed Category
Ethnicity	Two or More Races
	American Indian or Alaskan Native
	Asian
	Hispanic or Latino
	Black or African American
	White
	Pacific Islander
Gender	Male
	Female
IDEA Indicator	Yes
	No
Limited English Proficiency Status	Yes
	No
Enrolled Grade	Grade 5
	Grade 8
	Grade 11

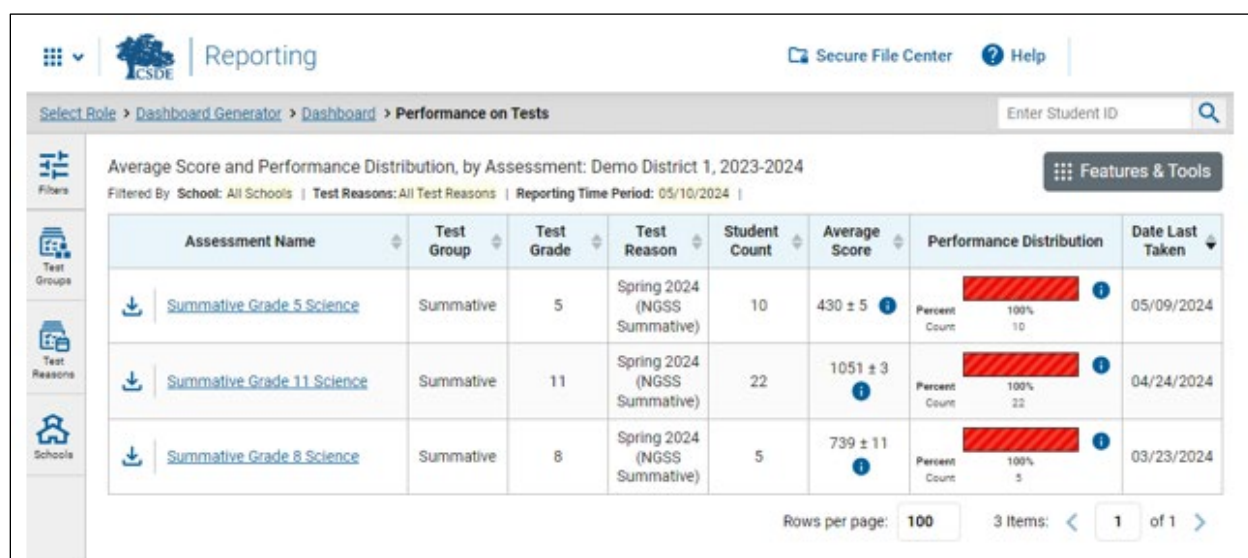
1.4.2 Summary Performance Report

Homepage-authorized users can log on to the CRS dashboard to view summaries of students' performance across grades and subjects. State personnel and district personnel can access district summaries, school personnel can access school summaries, and teachers can access summaries of their students through the dashboard. The Summary Performance Distribution Report can:

- Display summary data separated by grade and subject
- Present level of aggregation based on a user's role
- Report number of students tested and percentage meeting standard

Figure 1 presents a sample Summary Performance Distribution Report at the district level.

Figure 1. Spring 2024 District-Level Summary Performance Distribution Report



1.4.3 Aggregate-Level Subject Report

Detailed summaries of student performance within a grade and subject area are made available in the Aggregate-Level Subject Report. The Aggregate-Level Subject Report presents results for the aggregate unit, as well as results for any higher-level aggregate units. For example, a school's Aggregate-Level Subject Report will contain the summary results of the school's district so that school performance can be compared with district performance.

The Aggregate-Level Subject Report provides the aggregate summaries on a specific subject area, including

- number of students;
- average scale score and standard error of the average scale score;
- percentage of students meeting standard; and
- percentage of students in each performance level.

The summaries are also presented for overall students and by subgroups. Figure 2 presents an example of Aggregate-Level Subject Reports for grade 11 science at the district level without subgroups. Figure 3 highlights grade 11 science at the district level when a user selects a subgroup of gender.

Figure 2. Spring 2024 District Aggregate-Level Subject Report for Grade 11 Science

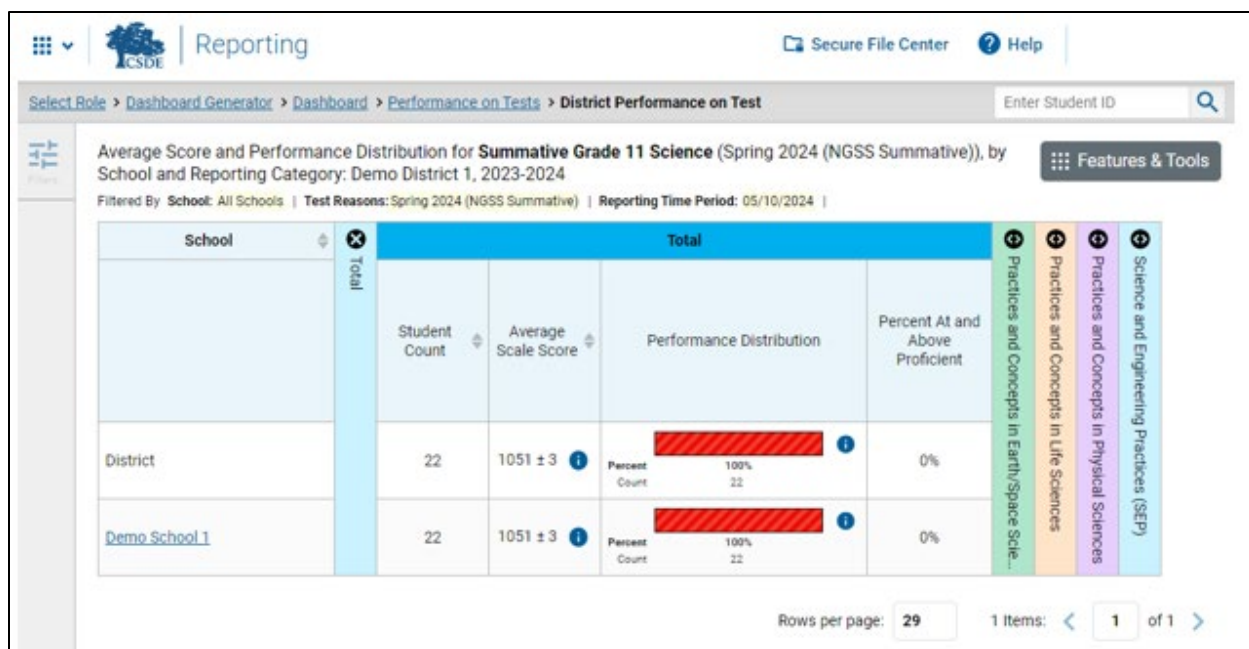
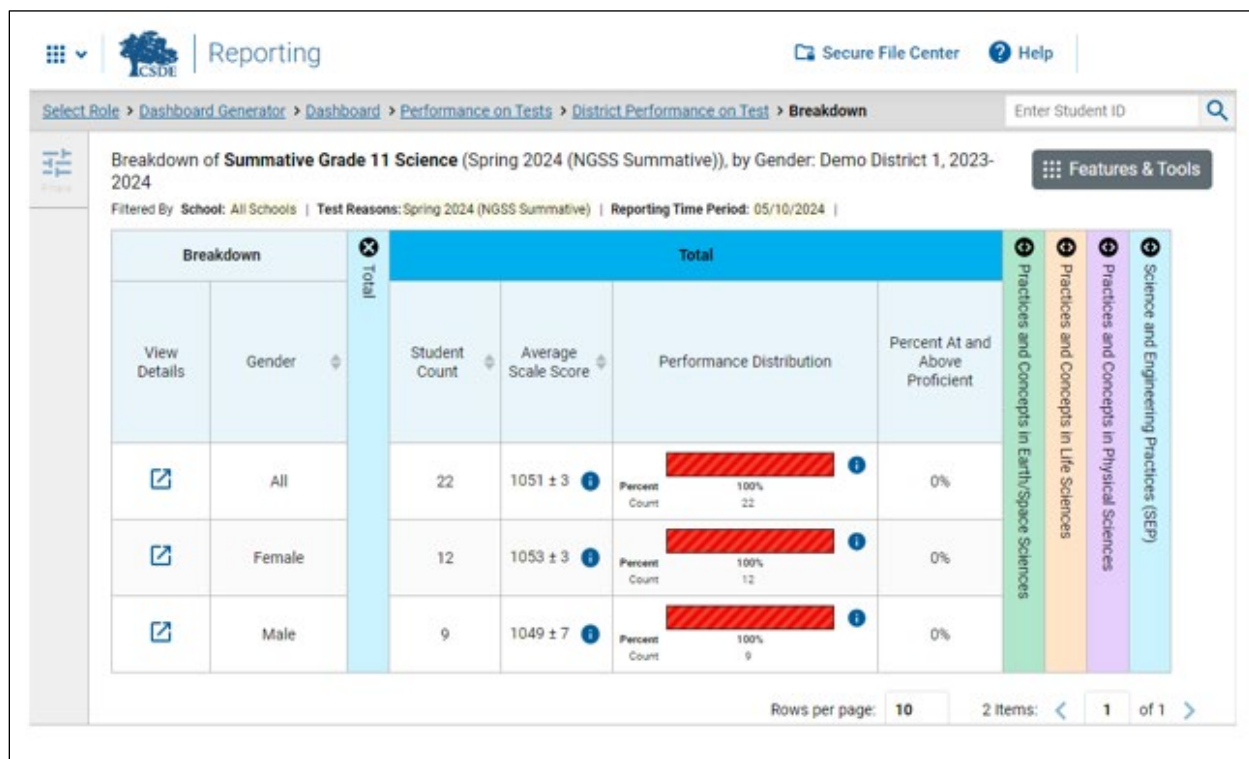


Figure 3. Spring 2024 District Aggregate-Level Subject Report for Grade 11 Science by Gender



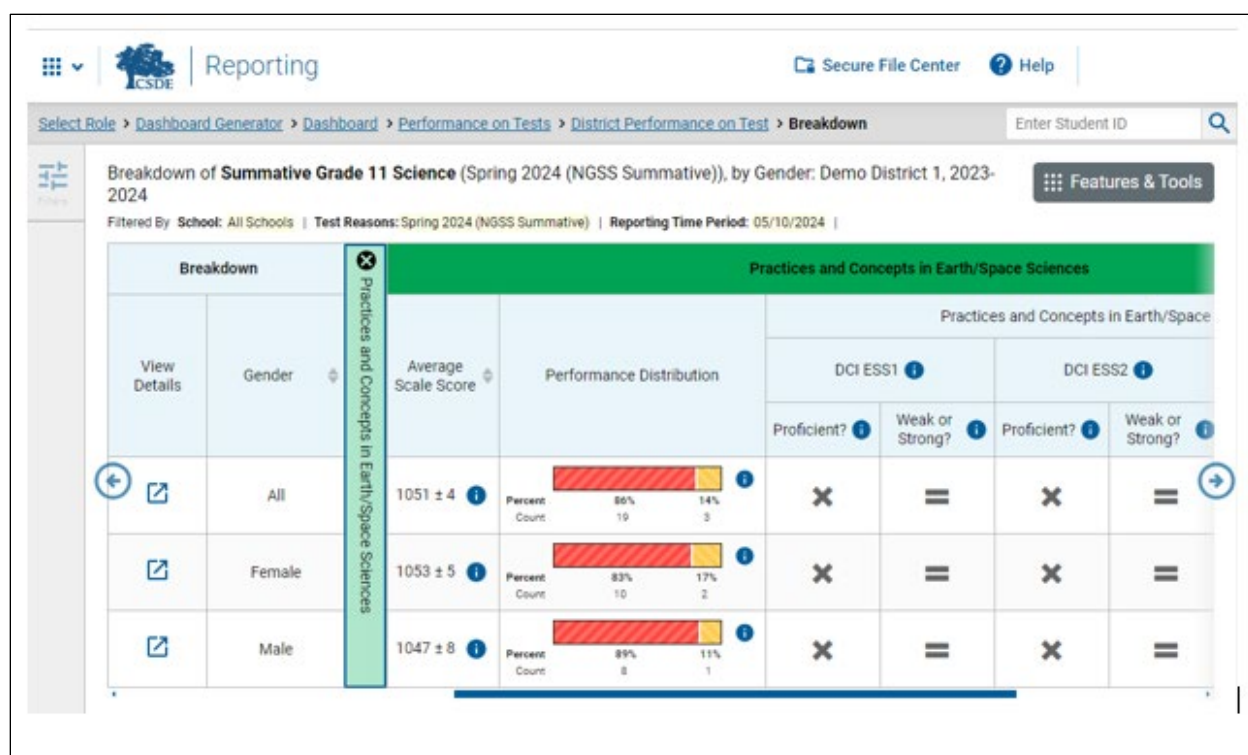
1.4.4 Aggregate-Level Discipline-Level Report

The Aggregate-Level Discipline-Level Report provides the aggregate summaries on student performance in each discipline level for each grade. The summaries on the Aggregate-Level Discipline-Level Report include

- number of students;
- average scale score and standard error of the average scale score;
- percentage of students meeting standard; and
- percentage of students in each performance category for each of the disciplines.

Similar to the Aggregate-Level Subject Report, the Aggregate-Level Discipline-Level Report presents the summary results for the selected aggregate unit, as well as the summary results for the aggregate unit above the selected aggregate. In addition, summaries can be presented for all students within an aggregate and by students within a defined subgroup. Figure 4 presents an example of the District Aggregate-Level Discipline-Level Report for grade 11 science. Reports by subgroups are also available for the Aggregate-Level Discipline-Level Report, similar to Figure 3.

Figure 4. Spring 2024 District Aggregate-Level Discipline-Level Report for Grade 11 Science



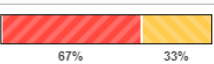


Aggregate-Level Disciplinary Core Ideas (DCI) and Science and Engineering Practices (SEP) Report

The Aggregate-Level Disciplinary Core Ideas (DCI) and Science and Engineering Practices (SEP) Report lists data on the performance of student groups on each standard of a subject for the current window and reports Areas Where Performance Indicates Proficiency and Areas of Strongest and Weakest Performance. For Areas Where Performance Indicates Proficiency, a performance indicator produces information on how a group of students in a class, school, or district performed on the standard compared to the proficiency cuts. It shows whether performance on this standard for this group was above, no different than, or below what is expected of students at the proficient level. This indicator shows strengths and weaknesses for a group of students and is provided only at an aggregate level since it is unstable at the individual level. For Areas of Strongest and Weakest Performance, the expected performance is determined based on the students' overall performance on the entire test.

Figure 5 demonstrates examples of the Aggregate-Level DCI and SEP Report for grade 11 science.

Figure 5. Spring 2024 District Aggregate-Level Disciplinary Core Idea (DCI) and Science Engineering Practices (SEP) Report for Grade 11 Science

Practices and Concepts in Earth/Space Sciences							
Average Scale Score	Performance Distribution	EarthAndSpaceScience					
		DCI ESS1		DCI ESS2		DCI ESS3	
		Proficient?	Weak or Strong?	Proficient?	Weak or Strong?	Proficient?	Weak or Strong?
1063 ± 16	 Percent Count: 57% 4, 43% 3	*	+	*	–	*	=
1063 ± 25	 Percent Count: 50% 2, 50% 2	*	=	*	–	*	=
1063 ± 26	 Percent Count: 67% 2, 33% 1	*	=	*	–	*	=

Note: The figure is an example of a DCI report, but a SEP report would be similar.




1.4.5 Student-Level Subject Report

The Student-Level Subject Report lists all students who belong to the selected aggregate level, such as a school, and reports the following measures for each student:

- Scale score
- Overall subject performance level

Figure 6 demonstrates an example of the Student-Level Subject Report for grade 11 science.

Figure 6. Spring 2024 Student Roster Subject Report for Grade 11 Science

Student	Student ID	Total	Total		Practices and Concepts in Earth/Space Sciences	Practices and Concepts in Life Sciences	Practices and Concepts in Physical Sciences
			Scale Score	Performance			
District			1050 ± 4	 Percent: 100% Count: 9			
School			1046	 Percent: 100% Count: 1			
My Students			1046	 Percent: 100% Count: 1			
Smith, Thomas	11111		1046 ± 11	Level 1			

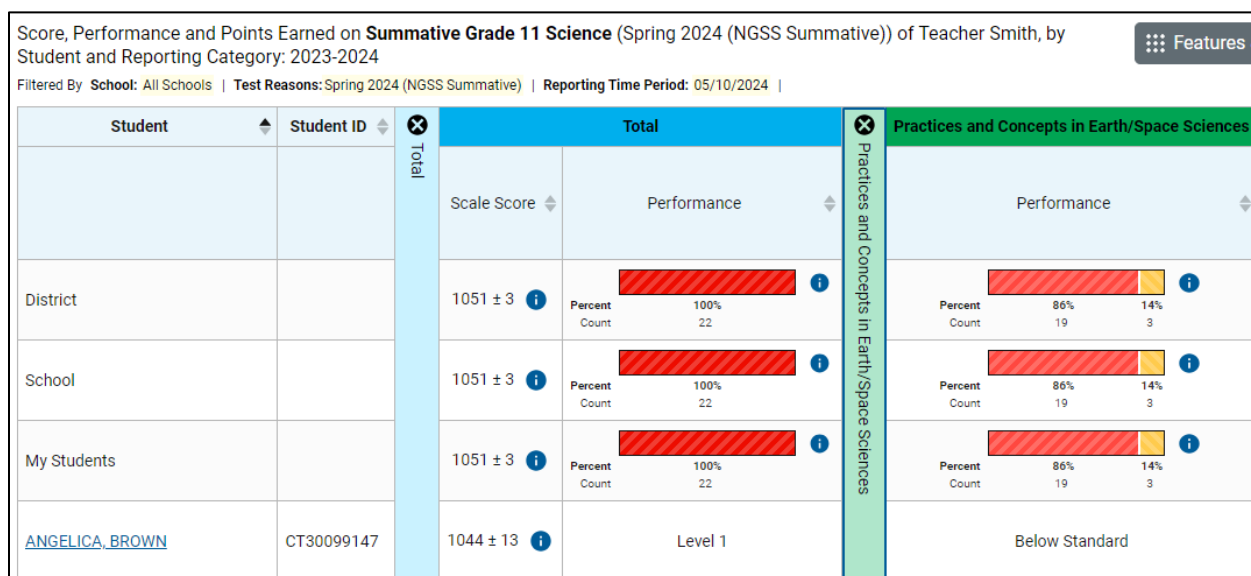
1.4.6 Student-Level Discipline-Level Report

The Student-Level Discipline-Level Report lists all students who belong to the selected aggregate level, such as a school, and reports the following measures for each student:

- Scale score
- Overall subject performance level
- Discipline performance category (i.e., Earth and Space Sciences, Life Sciences, and Physical Sciences)

Figure 7 presents an example of the Student-Level Discipline Report for grade 11 science.

Figure 7. Spring 2024 Student Roster Discipline Report for Grade 11 Science



1.4.7 Individual Student Report

When a student receives a valid test score, an Individual student report (ISR) can be generated in the CRS. The ISR contains the following measures:

- Scale score and standard error of measurement (SEM)
- Overall subject performance level
- Average scale scores for student's district and school
- Performance category in each discipline (science)

At the top of the report, information includes:

- Student's name
- Scale score with SEM
- Performance level

In the middle section of the report, information includes:

- Barrel chart with student's scale score and SEM (using a sign of "±")
- PLDs with cut scores at each performance level
- Average scale scores and standard errors for district and school aggregation levels


- Note: the “±” next to the student’s scale score is the standard error of measurement of the scale score, whereas the “±” next to the average scale scores for aggregate levels represents the standard error of the average scale scores.

At the bottom of the report, information includes:

- Detailed information on student performance on each discipline level
 - Note: Bar charts in the Discipline table show how students performed on each discipline (black bar), relative to the discipline-level performance standard (dashed white line). Green boxes show the score range the student would likely fall within if they took the test multiple times.

Figure 8 presents an example ISR for grade 11 science. An example of the printed ISRs is displayed in Appendix 6-B, Sample Printed Individual Student Report, of this technical report.

Figure 8. Spring 2024 Individual Student Report for Grade 11 Science



CONNECTICUT STATE DEPARTMENT OF EDUCATION
COMPREHENSIVE ASSESSMENT PROGRAM

Reporting

Individual Student Report

ANGELICA, BROWN

Student ID: CT30099147 | Student DOB: 3/9/2001 | Enrolled Grade: 11

Date Taken: 2/3/2024

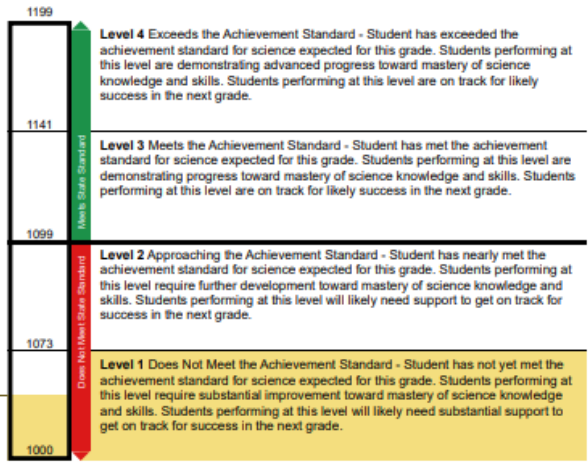
Scale Score: 1044±13 Performance: Level 1

Summative Grade 11 Science 2023-2024

Demo District 1

Demo School 1

How Did Your Child Do on the Test?



Score
1044 ±13

How Does Your Child's Score Compare?

Name	Average Scale Score
Demo District 1	1051±3
Demo School 1	1051±3




Information on Standard Error of Measurement

A student's score is best interpreted when recognizing that the student's knowledge and skills fall within a score range and not just a precise number. For example, 2300 (±10) indicates a score range between 2290 and 2310.

How Did Your Child Perform on Different Areas of the Test?

The table and the graph below indicate student performance on individual reporting categories. The black dot indicates the student's score on each reporting category. The lines to the left and right of the dot show the range of likely scores your student would receive if he or she took the test multiple times.

⚠ Below Standard ⚠ Approaching Standard ✅ Above Standard

Category	Performance	Performance Level	Performance level Description
Practices and Concepts in Earth/Space Sciences		⚠	In Earth/space sciences, student performance includes: using evidence to explain the history of the universe, explain how stars produce matter and energy, and predict orbital motion; using models and analyzing data to describe the processes that cycle matter and energy and affect Earth's surface, atmosphere, and inhabitants; using data as evidence to explain interactions between humans and Earth, including changes in climate; and designing solutions to problems resulting from the increasing use of Earth's resources.
Practices and Concepts in Life Sciences		⚠	In life sciences, student performance includes: using models and evidence from investigations of cell processes and living systems; analyzing factors affecting the stability of populations and ecosystems; making and defending claims for DNA coding of traits and to explain genetic variations; using multiple lines of evidence to support evolutionary relationships; and designing and evaluating solutions that minimize the effects of humans on biodiversity.
Practices and Concepts in Physical Sciences		⚠	In physical sciences, student performance includes: using periodic table patterns to describe and model changes in energy and matter; using data and mathematical representations to explain changes in motion due to forces and to investigate electromagnetic interactions; developing and using models to show energy transfers in systems; evaluating claims and evidence relating waves and electromagnetic radiation; and designing devices to optimize forces, results of chemical reactions, or energy conversion.

1.4.8 Data File

CRS users have the option to quickly generate a comprehensive data file of their students' scores. Data files (See Figure 9) can be downloaded in Microsoft Excel or CSV format and contain a wide variety of data, which includes scale scores, reporting discipline scores, demographic data, and performance levels. Data files can be useful as a resource for further analysis. Data files can be generated at the district, school, teacher, or roster level.

Figure 9. Spring 2024 Data File

H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
Enrolled	Enrolled	Enrolled	Enrolled	Enrolled	Science Sc	Standard	Science A	Practices	Practices	Standard	Practices	Practices	Standard	Practices	Practices	Standard	Error
8	Demo Sch	999999999	Demo Dis	999999999	752	12	Level 1	Below Sta	743	24	Below Sta	765	17	Below Sta	740	24	
8	Demo Sch	999999999	Demo Dis	999999999	700	17	Level 1	Below Sta	700	44	Below Sta	733	21	Below Sta	700	35	
8	Demo Sch	999999999	Demo Dis	999999999	751	12	Level 1	Below Sta	766	20	Below Sta	749	19	Below Sta	739	22	
8	Demo Sch	999999999	Demo Dis	999999999	751	10	Level 1	Below Sta	742	17	Below Sta	772	16	Below Sta	727	25	
8	Demo Sch	999999999	Demo Dis	999999999	751	11	Level 1	Below Sta	754	17	Below Sta	750	19	Below Sta	748	21	
8	Demo Sch	999999999	Demo Dis	999999999	747	12	Level 1	Below Sta	734	22	Below Sta	759	19	Below Sta	745	21	
8	Demo Sch	999999999	Demo Dis	999999999	717	15	Level 1	Below Sta	754	18	Below Sta	700	44	Below Sta	700	38	
8	Demo Sch	999999999	Demo Dis	999999999	700	19	Level 1	Below Sta	700	37	Below Sta	714	24	Below Sta	700	44	
8	Demo Sch	999999999	Demo Dis	999999999	727	12	Level 1	Below Sta	726	20	Below Sta	755	17	Below Sta	700	33	
8	Demo Sch	999999999	Demo Dis	999999999	760	10	Level 1	Below Sta	751	20	Below Sta	755	17	Approach	774	18	
8	Demo Sch	999999999	Demo Dis	999999999	736	13	Level 1	Below Sta	741	23	Below Sta	741	20	Below Sta	720	27	
8	Demo Sch	999999999	Demo Dis	999999999	762	11	Level 1	Below Sta	730	25	Approach	798	17	Below Sta	738	23	
8	Demo Sch	999999999	Demo Dis	999999999	747	12	Level 1	Below Sta	741	21	Below Sta	740	23	Below Sta	759	19	
8	Demo Sch	999999999	Demo Dis	999999999	700	26	Level 1	Below Sta	700	44	Below Sta	700	39	Below Sta	700	38	
8	Demo Sch	999999999	Demo Dis	999999999	709	17	Level 1	Below Sta	700	44	Below Sta	723	23	Below Sta	717	26	
8	Demo Sch	999999999	Demo Dis	999999999	705	15	Level 1	Below Sta	700	44	Below Sta	700	44	Below Sta	745	18	
8	Demo Sch	999999999	Demo Dis	999999999	741	12	Level 1	Below Sta	762	18	Below Sta	709	27	Below Sta	739	22	
8	Demo Sch	999999999	Demo Dis	999999999	746	11	Level 1	Below Sta	761	17	Below Sta	737	20	Below Sta	731	25	
8	Demo Sch	999999999	Demo Dis	999999999	707	18	Level 1	Below Sta	700	44	Below Sta	732	23	Below Sta	700	37	

1.5 TEST INFORMATION DISTRIBUTION ENGINE

Test Completion Rate Reports were available on the Test Information Distribution Engine (TIDE) website (<https://ct.tide.cambiumast.com>). These reports indicate the students who completed or need to complete computer-based testing and allow users to view participation summary statistics (counts and percentages) of students who have tested.

Once a user logs in, they are directed to the homepage, which allows them to access the Test Completion Rate Reports.

The Test Completion Rate Report allows teachers, principals, and district staff to see which students have not yet completed their tests. Users can select from a series of options to customize the group of students whose participation status is to be reviewed for a particular grade and subject, such as those who started but have not completed their test or those who have not yet begun their test. Users can export the list into a Microsoft Excel file and download the file.

1.6 PAPER INDIVIDUAL STUDENT REPORTS FOR FAMILIES

ISRs were delivered as printed materials to the districts where students were enrolled as of June 2, 2023, at 11:59:59 p.m. The primary purpose of the ISR was to provide a document that enabled families to understand their child's performance in the subject in which the child tested. The ISR also presented information that indicated how a student's performance compared to that of other students who took the same test. The report is organized as follows:

- **Top of Report.** The student's name, student ID, test grade, test date, school, and district are identified. Here, science reports include a frequently asked questions section.
- **Connecticut NGSS Assessment Scores.** The student's scale score and corresponding performance level are displayed graphically and explained in accompanying text. A range of scores that is \pm SEM is given with explanatory text.

- **Student Performance Compared.** Included with the Connecticut NGSS Assessment scores graphic, this section provides a comparison between the student’s scale score and that of the student’s school and district.
- **Discipline Level Scores.** Discipline level tables show how students performed on each discipline level. This section includes graphical displays of the Performance Category (*Below Standard, Approaching Standard, or Above Standard*), or relative strength/weakness, for each of the assessed discipline levels. These results are explained in greater detail next to the graphics, including “Next Steps,” that families and students may take to improve student performance.

An example of the printed ISRs is displayed in Appendix 6-B, Sample Printed Individual Student Report, of this technical report.

2. INTERPRETATION OF REPORTED SCORES

A student’s performance on a test is reported as a scale score and a performance level for the overall test and as a performance level for each discipline level. Students’ scores and performance levels are summarized at the aggregate levels. This section describes how to interpret these scores.

2.1 SCALE SCORE

A scale score is used to describe how well a student performed on a test and can be interpreted as an estimate of a student’s knowledge and skills as measured by their performance on the test. A scale score is the student’s overall numeric score. These scores fall on a continuous scale. The Connecticut NGSS Assessment scale scores are not expressed on a vertical scale, which means that scores from different grades cannot be compared.

Scale scores can be used to illustrate a student’s current levels of performance. Low scale scores indicate that a student does not possess sufficient knowledge and skills measured by the test. Conversely, high scale scores indicate that a student has proficient knowledge and skills measured by the test. When combined across a student population, scale scores can also describe school- and district-level changes in performance and reveal gaps in performance among different groups of students. In addition, scale scores can be averaged across groups of students, allowing educators to use group comparison. Interpretation of scale scores is more meaningful when the scale scores are used along with performance levels and performance-level descriptors (PLDs). It should be noted that the utility of scale scores is limited when comparing smaller differences among scores (or averaged group scores), particularly when the difference among scores is within the standard error of measurement (SEM). Furthermore, the scale score of individual students should be cautiously interpreted when comparing two scale scores, because small differences in scores may not reflect real differences in performance.

2.2 STANDARD ERROR OF MEASUREMENT

A student’s score is best interpreted when recognizing that the student’s knowledge and skills fall within a score range and are not just precise numbers. A scale score (the observed score on any test) is an estimate of the true score. If a student takes a similar test several times, the resulting

scale scores will vary across administrations; sometimes the scores will be a little higher, a little lower, or the same. The SEM represents the precision of the scale score, or the range in which the student would likely score if a similar test were administered several times. The SEM can be interpreted as the degree of uncertainty of a student’s score based on a statistical analysis of the student’s answers on a test. When interpreting scale scores, it is recommended to always consider the range of scale scores incorporating the SEM of the scale score.

The “ \pm ” next to a student’s scale score provides information about the certainty, or confidence, of the score’s interpretation. The boundaries of the score band are one SEM above and below the student’s observed scale score, representing a range of score values that is likely to contain the true score. For example, “680 \pm 10” indicates that if a student were tested again, it is likely that a student would receive a score between 670 and 690.

2.3 PERFORMANCE LEVEL

Performance levels are proficiency categories on a test, which students fall into based on their scale scores. For the Connecticut NGSS Assessment, scale scores are mapped into four performance levels (*Does Not Meet Standard*, *Approaching Standard*, *Meets Standard*, *Exceeds Standard*) using performance standards (refer to Section 2.5, Cut Scores). PLDs are a description of content-area knowledge and skills that students at each performance level are expected to possess. Thus, performance levels can be interpreted based on PLDs. Students performing on the Connecticut NGSS Assessment at *Meets Standard* and *Exceeds Standard* are considered on track to demonstrate progress toward mastery of the knowledge and skills necessary for college and career readiness.

2.4 PERFORMANCE CATEGORY FOR DISCIPLINE LEVELS

Students’ performance on each reporting discipline is reported for three performance categories: *Below Standard*, *Approaching Standard*, and *Above Standard*. Unlike the performance levels for the overall test, student performance on each of the discipline levels is evaluated with respect to the *Meets Standard* performance standard. Students performing at either *Below Standard* or *Above Standard* can be interpreted as having student performance that is clearly below or above the *Meets Standard* cut score for a specific discipline level. Students performing at *Approaching Standard* can be interpreted as having student performance that does not provide enough information to tell whether students reached the *Meets Standard* mark for the specific discipline level.

2.5 CUT SCORES

For all grades in the Connecticut NGSS Assessment, scale scores are mapped onto four performance levels (*Does Not Meet Standard*, *Approaching Standard*, *Meets Standard*, *Exceeds Standard*). For each performance level, there is a minimum and a maximum scale score that define the range of scale scores students in each performance level have achieved. Collectively, these minimum and maximum scale scores are defined as cut scores and are the cut-off points for each performance level. Table 4 presents the cut scores for science for all grades.

Table 4. Connecticut NGSS Assessment Science Performance-Level Cut Scores

Grade	Does Not Meet Standard	Approaching Standard	Meets Standard	Exceeds Standard
5	400–467	468–497	498–534	535–599
8	700–771	772–797	798–841	842–899
11	1000–1072	1073–1098	1099–1140	1141–1199

2.6 AGGREGATED SCORES

Students’ scale scores are aggregated at the roster, teacher, school, and district levels to represent how a group of students performs on a test. When students’ scale scores are aggregated, the aggregated scale scores can be interpreted as an estimate of knowledge and skills that a group of students possesses. This interpretation makes aggregated scores a powerful tool when comparing student performance across different groups of students, whether it be at a similar level of aggregation (e.g., school to school) or an analysis of a subgroup (e.g., comparing a teacher’s roster to the overall school).

Given that student scale scores are estimates, the aggregated scale scores are also estimates and are subject to measures of uncertainty, as expressed using the calculated SEM for an aggregate average scale score. In addition to the aggregated scale scores, the percentage of students in each performance level is reported at the aggregate level to represent how well a group of students performs overall and by discipline level.

2.7 RELATIVE STRENGTHS AND WEAKNESSES FOR DISCIPLINARY CORE IDEAS AND SCIENCE AND ENGINEERING PRACTICES

For Disciplinary Core Idea (DCI) and Science and Engineering Practices (SEP) performance, relative strengths and weaknesses at each standard are reported for aggregate levels only (e.g., classroom, school, district). Because an individual student responds to too few items within a standard to generate reliable data, the standard performance is produced by aggregating all items within a standard across students at an aggregate level.

The “Areas Where Performance Indicates Proficiency” for a standard shows how a group of students performed in each standard relative to the expected performance for proficiency. For summative tests, this is the expected level of performance necessary to achieve *Meets Standard* performance. This is a standards-based report with the group performance in each standard being compared to the performance standard for that standard. Similar to the performance levels provided for the total test, this is an indication of students’ performance in the standard with respect to the standards.

Since the “Areas Where Performance Indicates Proficiency” data for each standard is a comparison to the standards-based expectations, performance across groups can be compared.

For “Areas of Strongest and Weakest Performance,” the expected performance is determined based on the students’ overall performance on the entire test. It shows how a group of students performed in each standard relative to their performance on the test overall. Rather than comparing across

groups, “Areas of Strongest and Weakest Performance” provides more information regarding the relative strength and weakness on different standards on the test within a group.

2.8 APPROPRIATE USES FOR SCORES AND REPORTS

Assessment results can be used to provide information on individual student performance on the test. Overall, assessment results tell what a student knows and is able to do in certain subject areas and gives further information on whether a student is on track to demonstrate the knowledge and skills necessary for college and career readiness. Additionally, assessment results can be used to identify a student’s relative strengths and weaknesses in certain content areas. For example, performance categories for reporting disciplines can be used to identify an individual student’s relative strengths and weaknesses among reporting categories within a content area.

Assessment results on student performance on the test can be used to help teachers or schools make decisions on how to support students’ learning. Aggregate score reports at the teacher and school level provide information about the strengths and weaknesses of students and can be used to improve teaching and student learning. For example, a group of students may have performed very well overall, but possibly did not perform as well in several standards compared to their overall performance. In this case, teachers or schools can identify strengths and weaknesses of their students through the group performance by standards and promote instruction on specific areas where student performance is below their overall performance. Further, by narrowing down the student performance result by subgroup, teachers and schools can determine what strategies may be needed to improve teaching and student learning, particularly for students from disadvantaged subgroups. For example, teachers might see student assessment results by gender and observe that a particular group of students is struggling with Physical Sciences. Teachers can then provide additional instructions that focus on Physical Sciences for these students.

In addition, assessment results can be used to compare student performance among different students and among different groups. Teachers can evaluate how their students perform compared with other students in schools and districts for overall scores and by discipline level. Although all students are administered different sets of items under the linear-on-the-fly test design, scale scores are comparable across students.

While assessment results provide valuable information to understand student performance, these scores and reports should be used with caution. It is important to note that scale scores are estimates of true scores and hence do not represent the precise measure for student performance. A student’s scale score is associated with measurement error, and thus users need to consider measurement error when using student scores to make decisions about student performance. Moreover, although student scores may be used to help make important decisions about student placement and retention or teachers’ instructional planning and implementation, the assessment results should not be used as the only source of information. Given that assessment results measured by a test provide limited information, other sources on student performance, such as classroom assessment and teacher evaluation, should be considered when making decisions on student learning. Finally, when student performance is compared across groups, users need to take into account the group size. The smaller the group, the larger the measurement error related to these aggregate data, thus requiring a more cautious interpretation.

3. SUMMARY

Connecticut NGSS Assessment results are reported online via the Centralized Reporting System (CRS), as well as through printed individual student reports (ISRs) sent to families. The results were released after the testing window closed and standard setting was completed.

The CRS is interactive. When educators or administrators log in, they see a summary of data about students for whom they are responsible (e.g., a principal will see the students in their school; a teacher will see students in their class). Users can then drill down through various levels of aggregation all the way to individual student reports. The system allows them to tailor the content more precisely, moving from subject area through reporting categories, and even to standards-level reports for aggregates. Aggregate reports are available at every level, and authorized users can print these reports or download them (or the data on which they are based). ISRs can be produced individually or batched as PDF reports.

All authorized users can download files, including data about students for whom they are responsible, at any time. The various reports available may be used to inform stakeholders regarding student performance and instructional strategies.