Connecticut Next Generation Science Standards Assessment

2020–2021

Volume 3
Setting Performance Standards

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

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 achievement of those standards, the Connecticut State Department of Education 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, AIR conducted the standard-setting workshop to recommend performance standards for Connecticut's Next Generation Science Standards (NGSS) Assessments 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, Connecticut.

Connecticut's NGSS Assessments are designed to measure attainment of the Next Generation Science Standards 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 coverage of the test while limiting increases in testing time and burden on students and schools. Test items were developed by AIR in conjunction with a group of states working to implement three-dimensional Next Generation Science Standards. 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 Next Generation Science Standards.

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 assertions 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.

Panelists reviewed Performance-Level Descriptors (PLDs) describing the degree to which students have achieved Connecticut's Next Generation Science Standards. Range PLDs were reviewed and revised by CSDE prior to the standard-setting workshop. 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 ELA and math for elementary and middle school grades, and SAT evidence-based reading and writing and math college ready indicators for grade 11 for each assertion.

Forty-two Connecticut science educators served as science standard-setting panelists, with 15 participants each for the grade 5 and 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 Next Generation Science Standards 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 math performance standards for elementary and middle school grades, and SAT evidence-based reading and writing and math 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.

1.1.2 Results of the Standard-Setting Workshop

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

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.

[1] 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 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

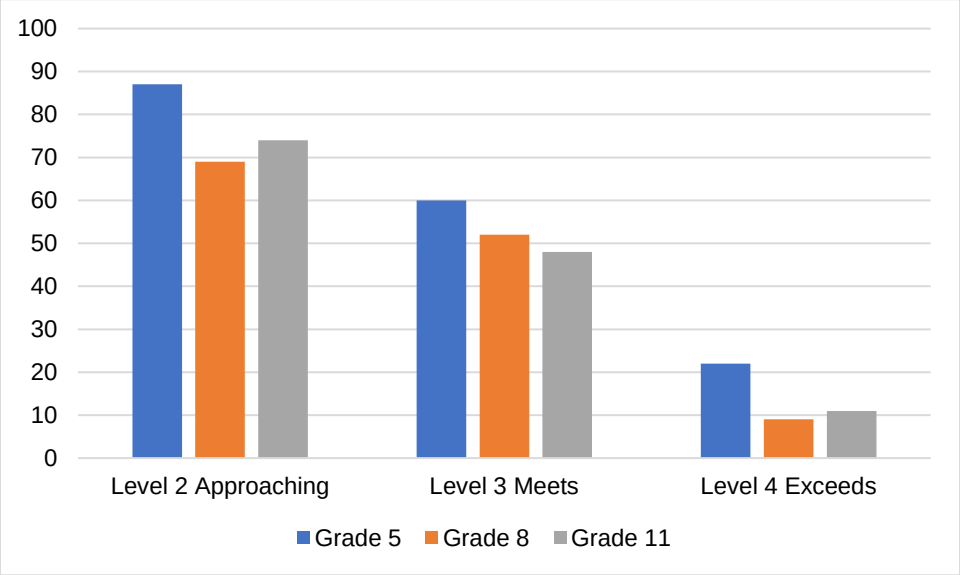
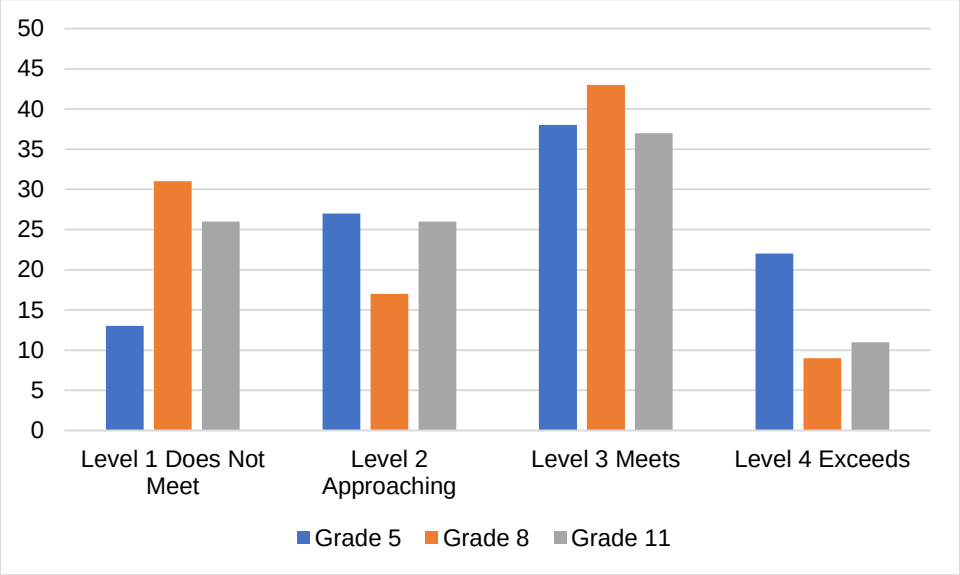


Table 3 indicates the percentage of students classified within each of the performance levels in 2019. The values are displayed graphically in Figure 2.

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) 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 science assessment (NGSS) measures 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 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 report is to document the standard-setting process for NGSS science and resulting performance standard recommendations.

3. THE NEXT GENERATION SCIENCE STANDARDS

The NGSS tests assess the learning objectives described by the Next Generation Science Standards, 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 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 refers to these performed knowledge and skills as *performance expectations* (PEs). Second, while unidimensionality is a typical goal of standards (and the assessments that measure them), 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 connects to other subjects like the Common Core math and 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 *A Framework for K-12 Science Education*:

Science and Engineering Practices	Disciplinary Core Ideas	Crosscutting Concepts
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.	**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.	**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 ASSESSMENTS

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

Item clusters include a stimulus and a series of questions that generally take students about 6–12 minutes to complete. They consist of a phenomenon, an observable fact or design problem that an engaged student explains, models, investigates, or designs using the knowledge and skill described by the performance expectation (PE) to complete a series of activities (comprised of multiple interactions). For example, in Figure 3, proficiency in this single performance expectation requires activities that demonstrate the ability to analyze and evaluate data, knowledge of properties and purposes of different forms of matter, and the application of experimental cause and effect. The stimulus in an item cluster explicitly states a task or goal (for example, "In the questions that follow, you will analyze what happens to the train when the brakes are applied") and subsequent interactions build upon or relate to the task or response to previous questions. The interactions within an item cluster all address the same phenomenon.

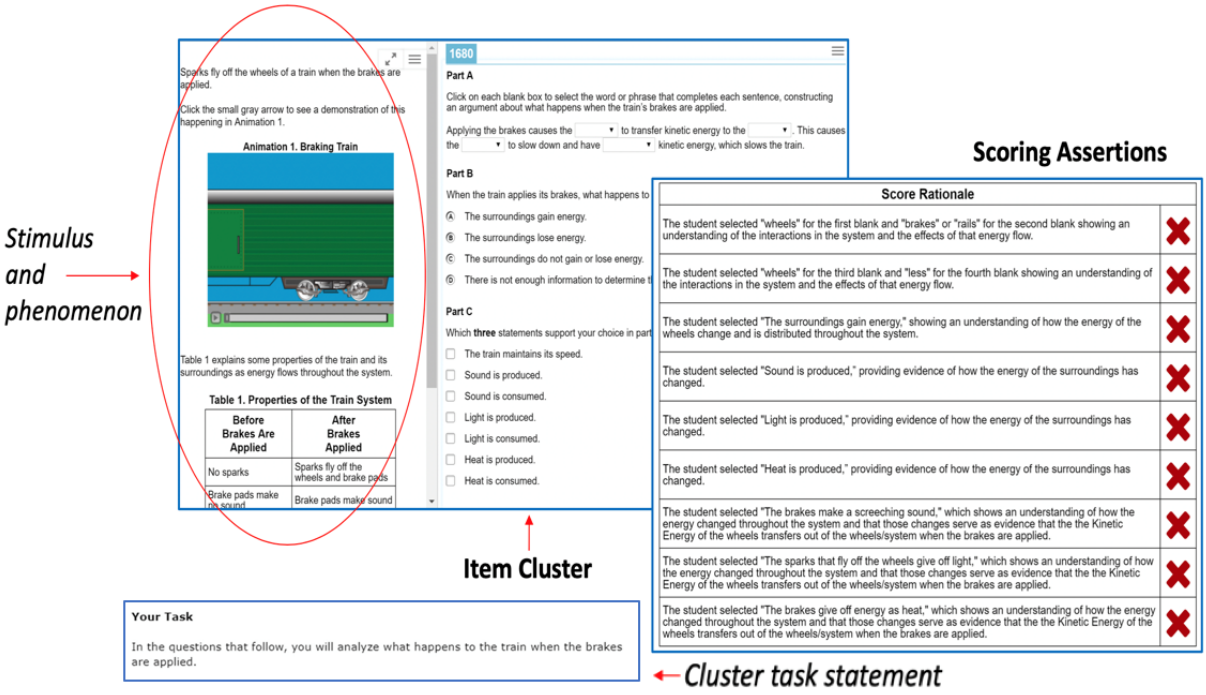
Some added stand-alone items increase the coverage of the test without also increasing testing time or testing burden. Stand-alone items are shorter, unrelated to other items, and generally take students 1–3 minutes to complete. Within each item cluster, there are a variety of interaction types including selected response, multi-select, table match, edit in-line choice, and simulations of science investigations. Stand-alone items can also be these types.

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.airast.org/.

Figure 4. Example NGSS Item Cluster and Scoring Assertions

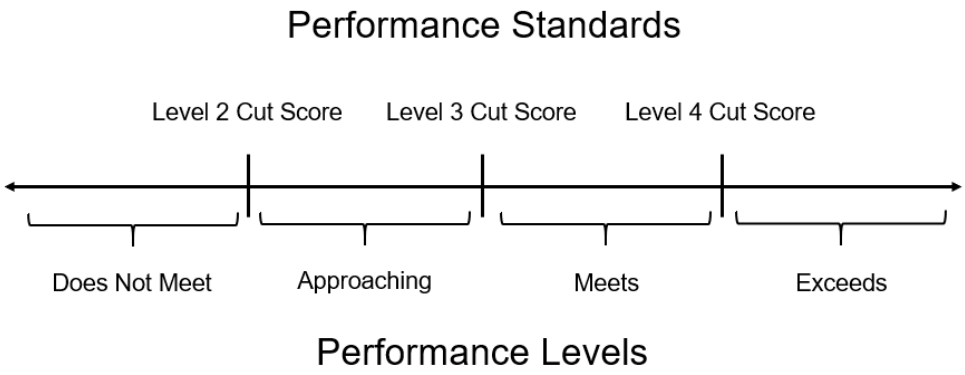


5. STANDARD SETTING

Forty-two educators from Connecticut convened at the Red Lion, Cromwell, Connecticut, from July 31 through August 1, 2019, to complete two rounds of standard setting to recommend three performance standards for the new NGSS science tests.

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 new approach 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 coverage of the test 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 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 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 also 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.

To address these challenges, AIR 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 Bookmarking procedure used frequently for other accountability tests (Rijmen et al., 2018).

The main distinction between AMP and existing ordered-item procedures (e.g., Mitzel, Lewis, Patz, & Green, 2001) 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 ordered-item procedures. 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 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.[2]

[2] AIR 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 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.

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

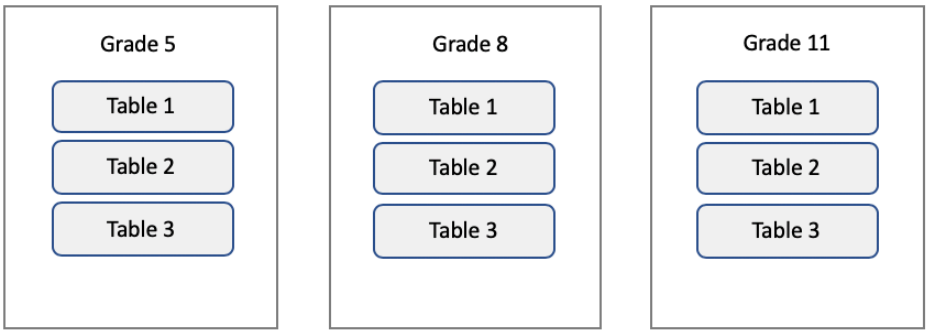


Table 4 summarizes the composition of the tables and the number of facilitators and panelists assigned to each. The 42 standard-setting participants included table leaders and panelists who taught in the content area and grade level for which standards were being set.

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. 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
- 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 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:

- 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 troubleshot 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. Facilitators were:

- 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 percent male and 29 percent non-white. They included teachers, administrators, and other stakeholder groups who worked in schools (48 percent), districts (29 percent), both schools and districts (17 percent) and elsewhere (7 percent). Panelists represented suburban districts (45 percent), urban districts (33 percent) and rural districts (19 percent) that were most often medium in size (45 percent, followed by small (31 percent) and large (21 percent.)) Ninety-percent taught science and a third (33 percent) 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				
School	33%	50%	60%	48%
District	33%	42%	13%	29%

	Percentage of Panelists by Panel			
	Grade 5	*Grade 8*	*Grade 11*	*Overall*
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. Other stakeholder groups included 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 over 20 years. Over half (52 percent) had taught in their assigned panel's grade and subject for 1-10 years, while 19 percent had taught it for more than 20 years. Most (67 percent) had professional experience outside the classroom and over half between 55 and 60 percent) 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 IEP	47%	42%	87%	60%

Note. Other professional experience in education included positions such as science coordinator, department chair or dean, committee member, coach or specialist. Abbreviation Key: Individualized Education Plan (IEP).

Appendix 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 Next Generation Science Standards. 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 Performance-Level Descriptors (PLDs) from several states' assessments based on three-dimensional science standards. States selected the range PLDs based on standards drafted by the Washington State 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 authors of the Next Generation Science Standards (NGSS) reviewed and applied recommendations to the PLDs. They ensured consistency, coherence and articulation across grades and levels. Appendix 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.

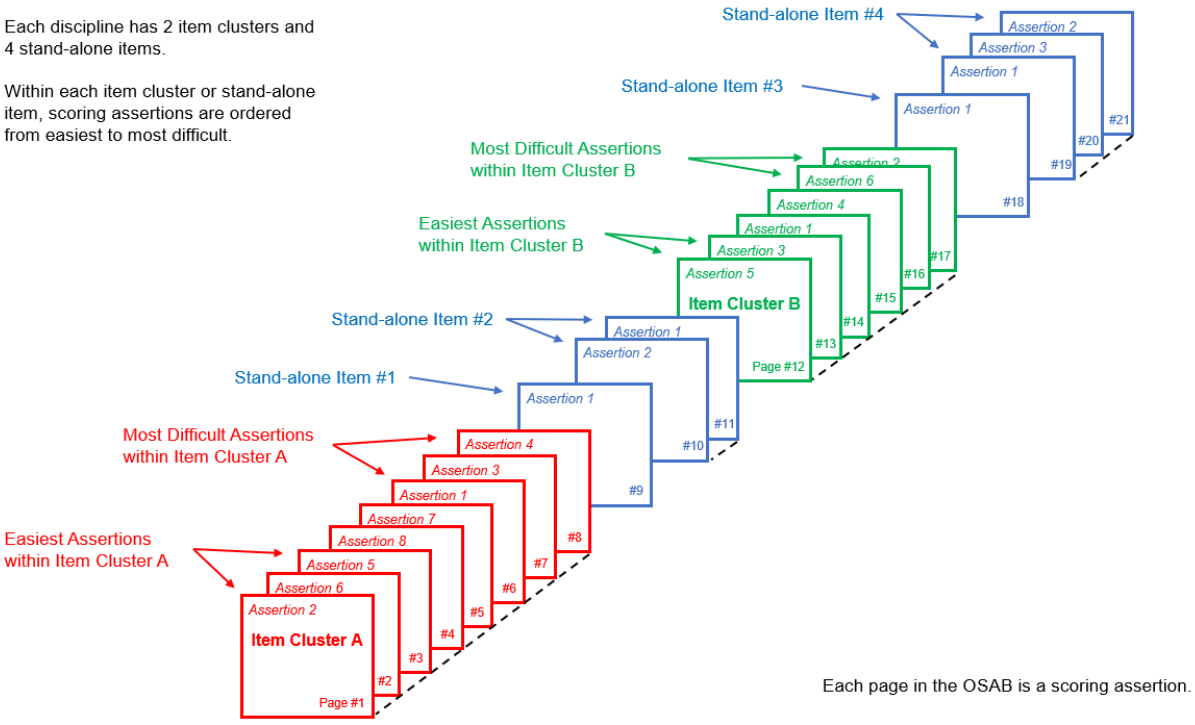
5.4.2 Ordered Scoring Assertion Booklets

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 three percent 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.

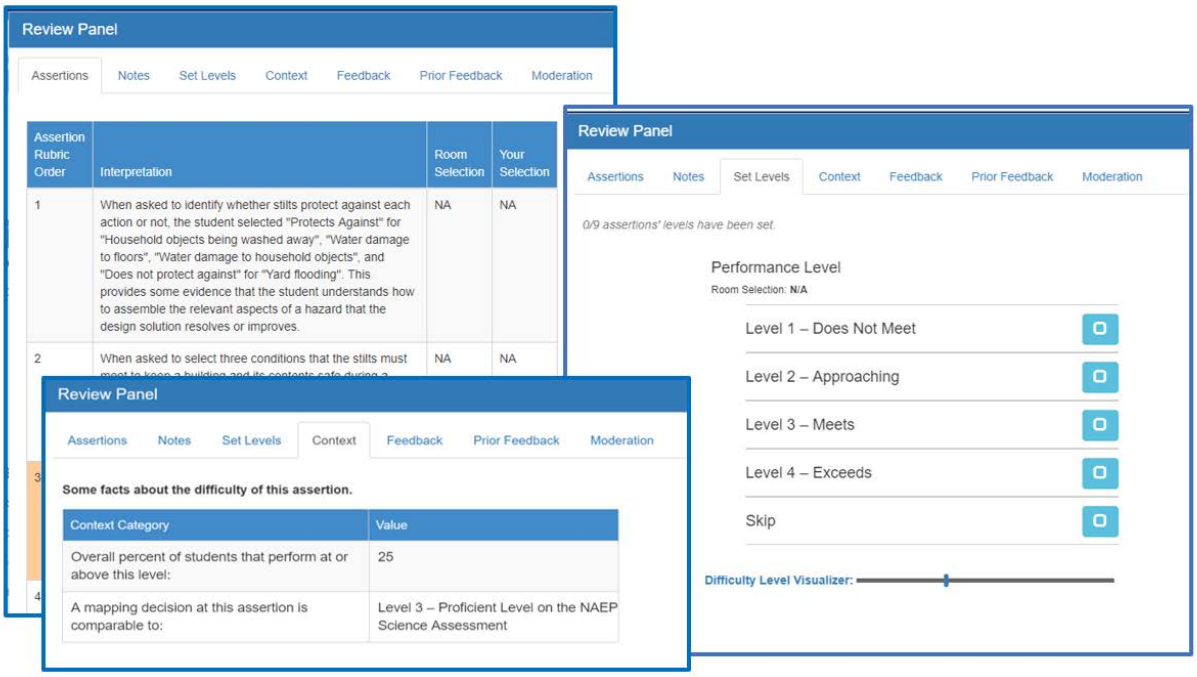
Each OSAB contains three disciplines and 18 items (item clusters and stand-alone items). The grade 5 OSAB contained 70 assertions, the grade 8 OSAB contained 76 assertions, and the grade 11 OSAB contained 80 assertions. Each comprised of 6 item clusters and 12 stand-alone items.

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 IT 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 C, Standard-Setting Workshop Agenda, provides the full workshop agenda.

Table 7. Standard-Setting Agenda Summary

Day 1: Wednesday, July 31, 2019
• Table leader orientation • Large-group introductory training • Take the test • PLD review • Create threshold PLDs • OSAB review
Day 2: Thursday, August 1, 2019
• 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 were to provide the following throughout the workshop; they:

- 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 serve as panelists and set individual cut scores.

Appendix 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 people 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 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 table, drafted descriptions for "Meets," "Exceeds," and then "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), their notes from the OSAB review, and their professional judgments. Figure 9 graphically describes the assertion-mapping process.

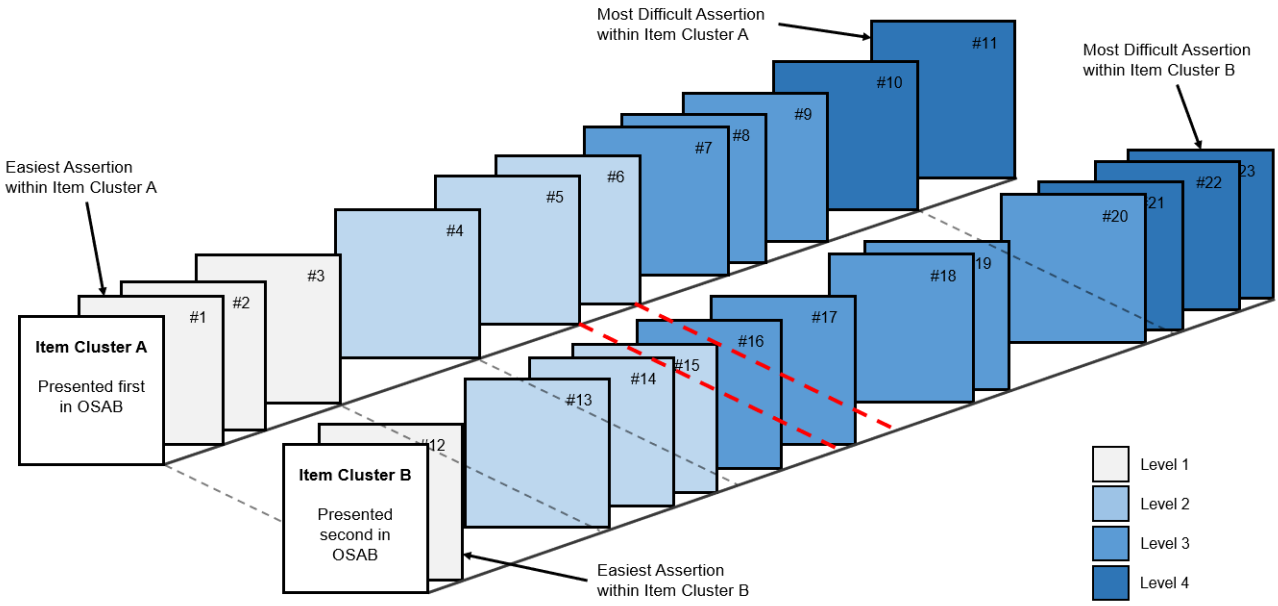
Facilitators provided the following process to guide the mapping of assertions onto PLDs:

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 done 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 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; and,
- 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 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, and
- 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 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.[3] 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 assertions that were mapped to the two adjacent levels were considered (e.g., for the second cut, only the assertions

[3] 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 performance-level descriptor (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).

that were 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 grade 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.[4] 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 math 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		
	A	*M*	*E*	*A*	*M*	*E*
Grade 5	**465**	**493**	**522**	**87**	**60**	**25**
Table 1	478	502	519	76	49	28

[4] The National Assessment of Educational Progress (NAEP) provides state-level benchmark data in science for grade 8; benchmark data for grade 5 is interpolated, and for grade 11 it is extrapolated.

Grade and Table	Cut Scores			Impact Data		
	A	*M*	*E*	*A*	*M*	*E*
Table 2	478	499	522	76	52	25
Table 3	465	484	522	87	70	25
Grade 8	**783**	**798**	**842**	**69**	**52**	**9**
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 abbreviation key: Approaching (A), Meets (M), 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

Grade and Table	Cut Scores			Impact Data		
	A	*M*	*E*	*A*	*M*	*E*
Table 3	1077	1107	1141	75	38	11

Note. The grade-level row summarizes the room data (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 2 cut scores. Performance level abbreviation key: A = Approaching, M = Meets, E = Exceeds.

Given the recommended round 2 cut scores, for all grades, between 48 and 60 percent of students would meet the recommended standard, between 9 and 22 percent would exceed the standard, and between 69 and 87 percent 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

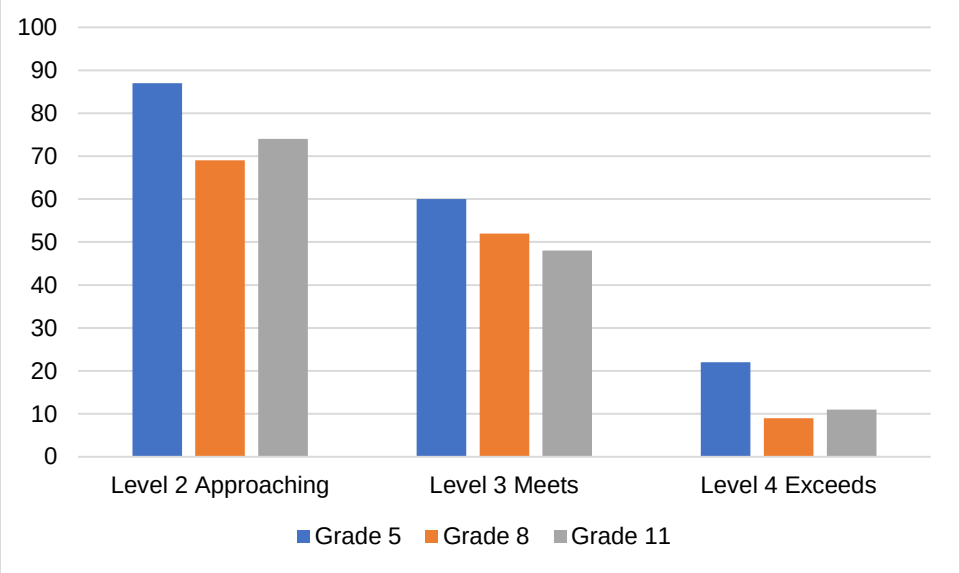
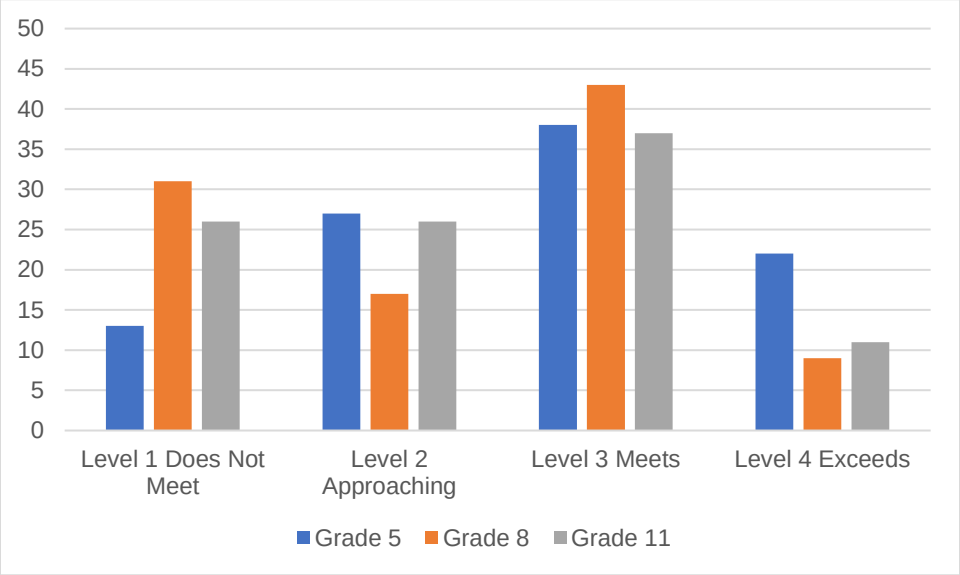


Table 10 indicates the percentage of students classified within each of the performance levels in 2019. The values are displayed graphically in Figure 11.

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 abbreviation key: A = Approaching, M = Meets, E = Exceeds.

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

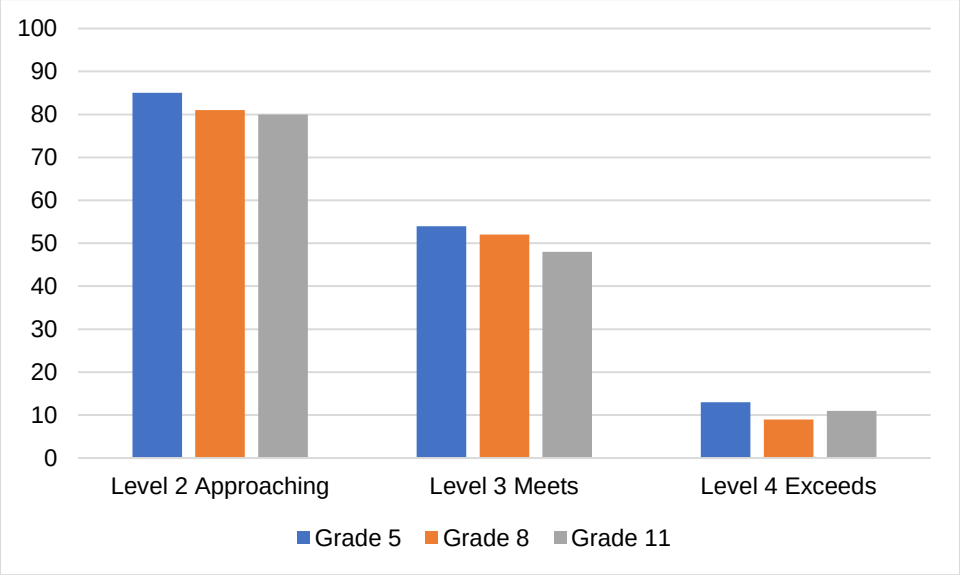
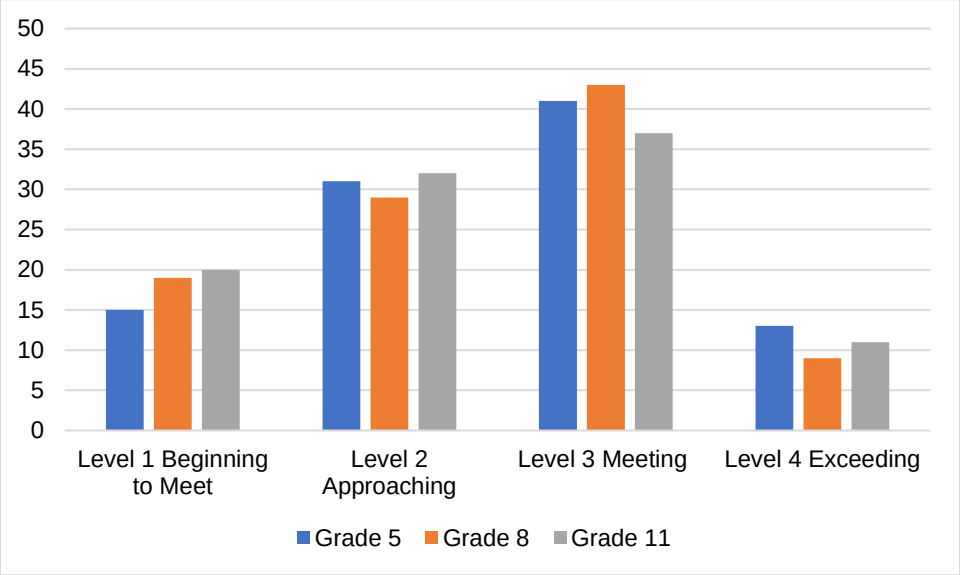


Table 12 indicates the percentage of students classified within each of the performance levels in 2019 proceeding from CSDE refinements to the recommended performance standards. The values are displayed graphically in Figure 13.

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. Table 13, Table 14, Table 15, Table 16, and Table 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 is 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, 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, 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 indicted 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, 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 his or her opinions throughout the workshop.	100%	100%	100%	100%

Note. Number of responses = 39 (Grade 5 responses = 14, Grade 8 responses = 12, 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, 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 twenty-one 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; Mehrens, 1995). The process also adhered to the following professional standards recommended in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) related to standard setting:

> *Standard 5.21:* When proposed score interpretation involves 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 and distinct substantive interpretations should be informed by sound empirical data concerning the relation of test performance to the relevant criteria.

The sections of this report documenting the rationale and procedures used in the standard-setting workshop address Standard 5.21. The 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 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 United States 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 1.2:** Substantive involvement and input of educators and subject-matter experts

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 achievement standards and academic achievement standards to ensure they are valid and reliable.

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 5.6.8, Assertion-Mapping Training; and Section 5.6.10, Practice Round) and confidence in the workshop process and outcomes (Section 5.10.1, Workshop Participant Feedback) 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 the 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 A, Standard-Setting Panelist Characteristics).

7. REFERENCES

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

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.

Mehrens, W. (1995). *Licensure Testing: Purposes, Procedures, and Practices*, ed. James C. Impara (Lincoln, NE: Buros Institute of Mental Measurements, University of Nebraska–Lincoln, 1995).

Mitzel, H.C., Lewis, D.M., Patz, R.J., & Greene, D.R. (2001). The Bookmark procedure: Psychological perspectives. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Erlbaum.

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 28). Scoring and reporting for assessments developed for the new science standards [Symposium]. National Conference on Student Assessment, San Diego, CA, United States.

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.