Rhode Island and Vermont Multi-State Science Assessment

2021–2022

Volume 3: Setting Achievement Standards

RIDE Rhode Island Department of Education

VERMONT® AGENCY OF EDUCATION

TABLE OF CONTENTS

List of Tables

LIST OF FIGURES

LIST OF APPENDICES

1. Executive Summary

In 2013, the Rhode Island Department of Education (RIDE) and Vermont Agency of Education (VT AOE) 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, RIDE and VT AOE convened a standard-setting workshop to recommend a system of achievement standards to determine whether students have met the learning goals defined by the NGSS.

Under contract to RIDE and VT AOE, the American Institutes for Research (AIR; currently Cambium Assessment, Inc. [CAI]) conducted the standard-setting workshop to recommend achievement standards for the Rhode Island Next Generation Science Assessment (RI NGSA) and the Vermont Science Assessments (VTSA) at grades 5, 8, and 11. The workshop was conducted August 5–6 2019, at the Grappone Conference Center, 70 Constitution Avenue, Concord, NH.

The RI NGSA and the VTSA are designed to measure attainment of the Next Generation Science Standards. 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 any burdens on students and schools. Test items were developed by AIR in conjunction with a group of states working to implement the three-dimensional NGSS. Test items were developed to ensure that each student is administered a test meeting all elements of the Rhode Island and Vermont Science Assessment blueprints, which were constructed to align to the NGSS.

Rhode Island and Vermont science educators, serving as standard-setting panelists, followed a standardized and rigorous procedure to recommend achievement standards demarcating each achievement level. To recommend achievement standards for the new science assessments, panelists participated in the Assertion Mapping Procedure (AMP), an adaptation of the Item-Descriptor (ID) Matching procedure (Ferrara & Lewis, 2012). Consistent with ordered-item procedures in general (Mitzel, Lewis, Patz, & Green, 2001), workshop panelists reviewed and recommended achievement 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 achievement-level descriptor.

Panelists reviewed achievement-level descriptors (ALDs) describing the degree to which students have performed on the NGSS. Range ALDs were reviewed and revised by educator panels prior to the standard-setting workshop. After reviewing the range ALDs, standard-setting panelists worked to identify knowledge and skills characteristics of students just qualifying for entry into each achievement level.

Working through the ordered scoring assertions for each item, panelists mapped each assertion into one of the four achievement levels—Beginning to Meet Expectations, Approaching Expectations, Meeting Expectations, and Exceeding Expectations. 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 achievement standards used to classify student achievement—Approaching Expectations, Meeting Expectations, and Exceeding Expectations. Mapping of scoring assertions in Round 1 was based on consideration of test content only. 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 contextual information about the percentage of students who would meet or exceed each of the achievement standards recommended in Round 1.

Twenty-six Rhode Island and Vermont science educators were selected to serve as science standard-setting panelists, with nine participants serving on the elementary and middle school panels, and eight participants serving on the high school 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 states.

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 achievement standards (Approaching Expectations, Meeting Expectations, and Exceeding Expectations) that will be used to classify student science achievement on the Rhode Island and Vermont NGSS Assessments.
- Panelists recommended achievement standards in two rounds.
- Context data, including the percentage of students who performed at or above the achievement level associated with each individual assertion, were provided to panelists following the first round of recommending achievement standards.
- The standard-setting workshops were conducted online using AIR's online standard-setting tool. A laptop computer was provided to each panelist at the workshop.

1.1.2 Results of the Standard-Setting Workshop

The science scores are expressed on an integer-valued scale ranging from 1 to 120. Table 1 displays the achievement standards recommended by the standard-setting panelists. Note that the scale for each grade will be re-centered around the Level 3 standard after final approval of the standards. The scale values of the standards will shift accordingly, but the shift will not affect the percentages at or above each of the achievement standards.

Table 1. Achievement Standards Recommended for Science

Grade	Level 2 Approaching	Level 3 Meeting	Level 4 Exceeding
5	45	68	75
8	41	63	77
11	39	63	74

Table 2 indicates the percentage of students who will reach or exceed each of the achievement standards in 2019.

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

Grade	State	Level 2 Approaching	Level 3 Meeting	Level 4 Exceeding
5	**Combined**	**74**	**24**	**12**
	Rhode Island	72	23	12
	Vermont	78	26	13
8	**Combined**	**80**	**35**	**10**
	Rhode Island	78	32	9
	Vermont	84	39	12
11	**Combined**	**90**	**35**	**16**
	Rhode Island	89	31	14
	Vermont	92	42	21

Figure 1 through Figure 3 represent those values graphically.

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

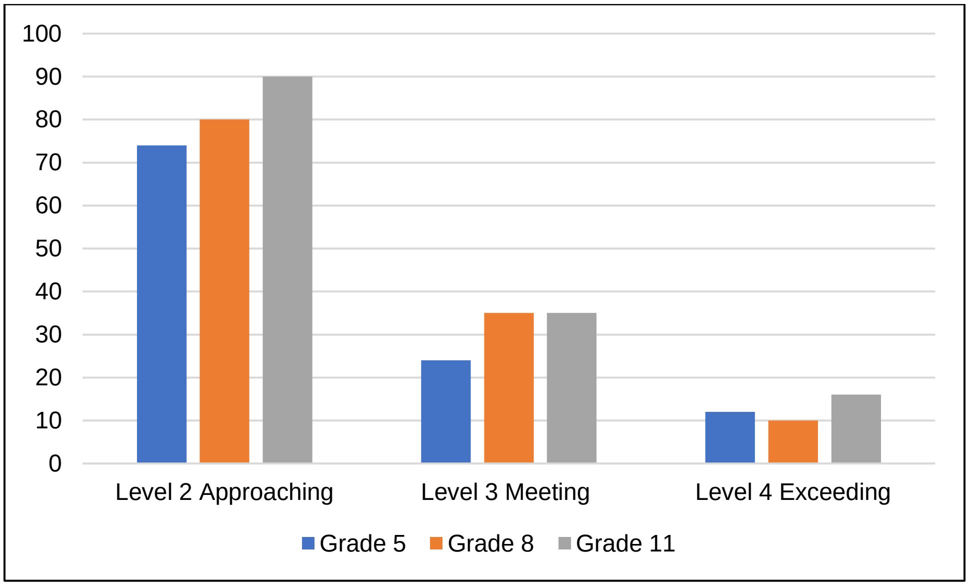


Figure 2. Percentage of Rhode Island Students Reaching or Exceeding Each Recommended Science Achievement Standard in 2019

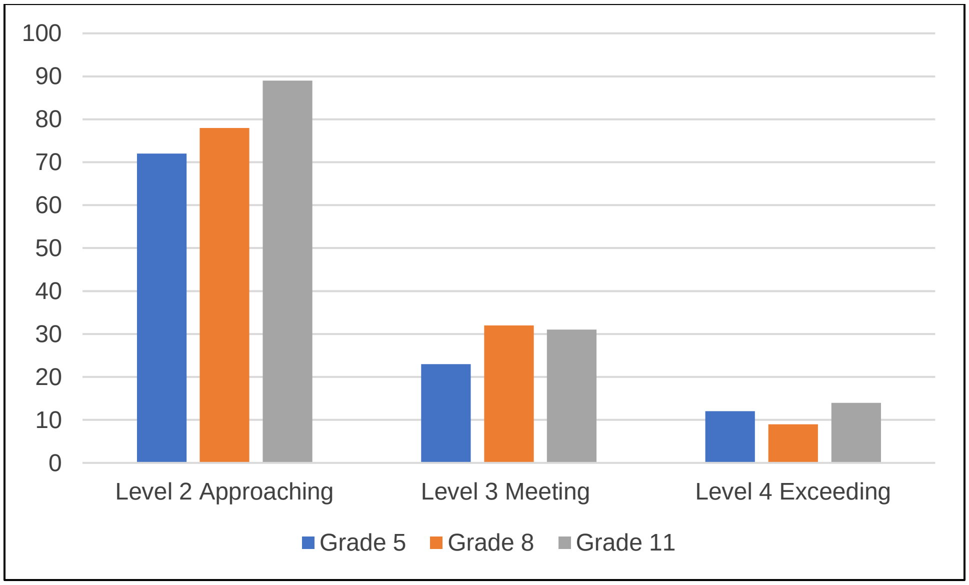


Figure 3. Percentage of Vermont Students Reaching or Exceeding Each Recommended Science Achievement Standard in 2019

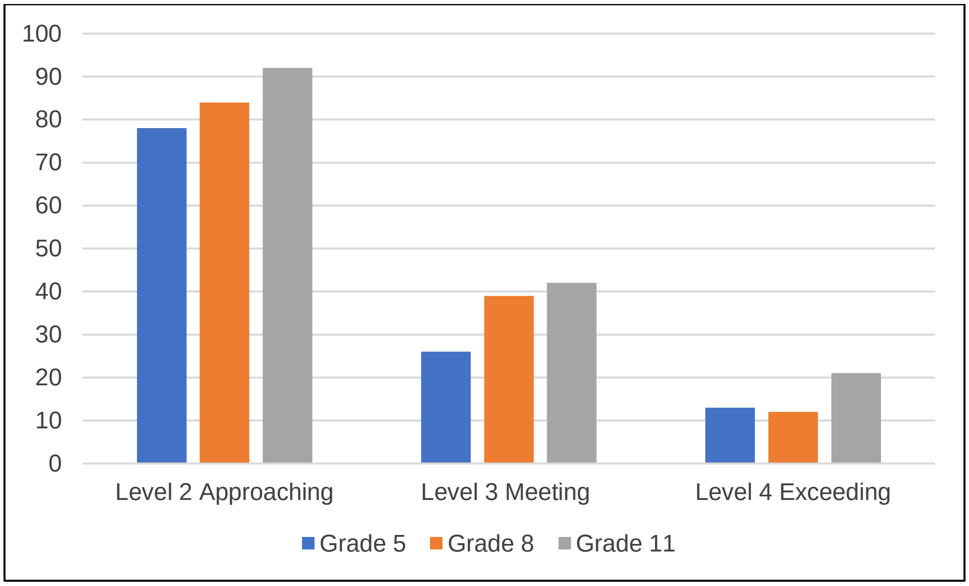


Table 3 indicates the percentage of students classified within each of the achievement levels in 2019. The values are displayed graphically in Figure 4 through Figure 6.

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

Grade	State	Level 1 Beginning to Meet	Level 2 Approaching	Level 3 Meets	Level 4 Exceeds
5	**Combined**	**26**	**50**	**12**	**12**
	Rhode Island	28	49	11	12
	Vermont	22	52	13	13
8	**Combined**	**20**	**45**	**25**	**10**
	Rhode Island	22	46	23	9
	Vermont	16	45	27	12
11	**Combined**	**10**	**55**	**19**	**16**
	Rhode Island	11	58	17	14
	Vermont	8	50	21	21

Figure 4. Percentage of Combined Students Classified Within Each Science Achievement Level in 2019

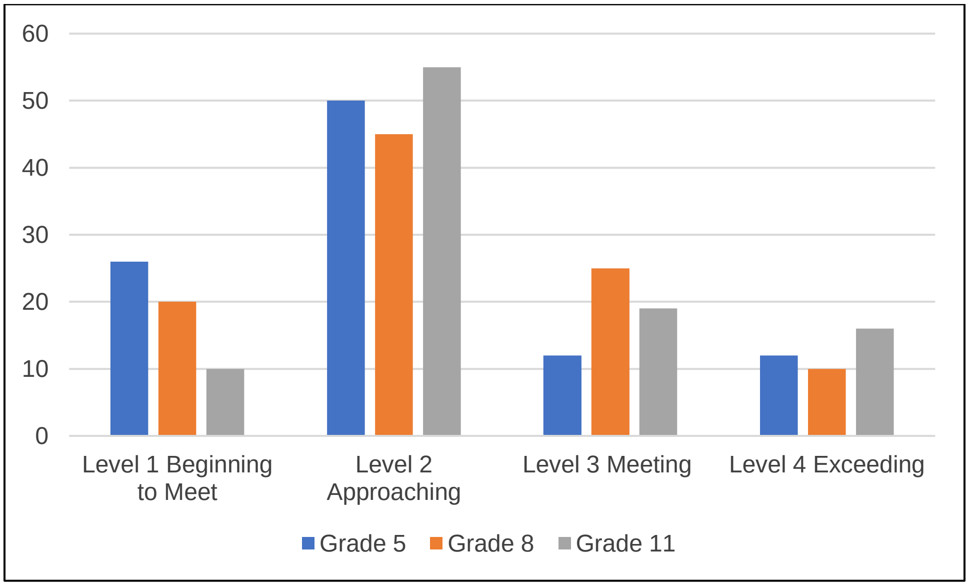


Figure 5. Percentage of Rhode Island Students Classified Within Each Science Achievement Level in 2019

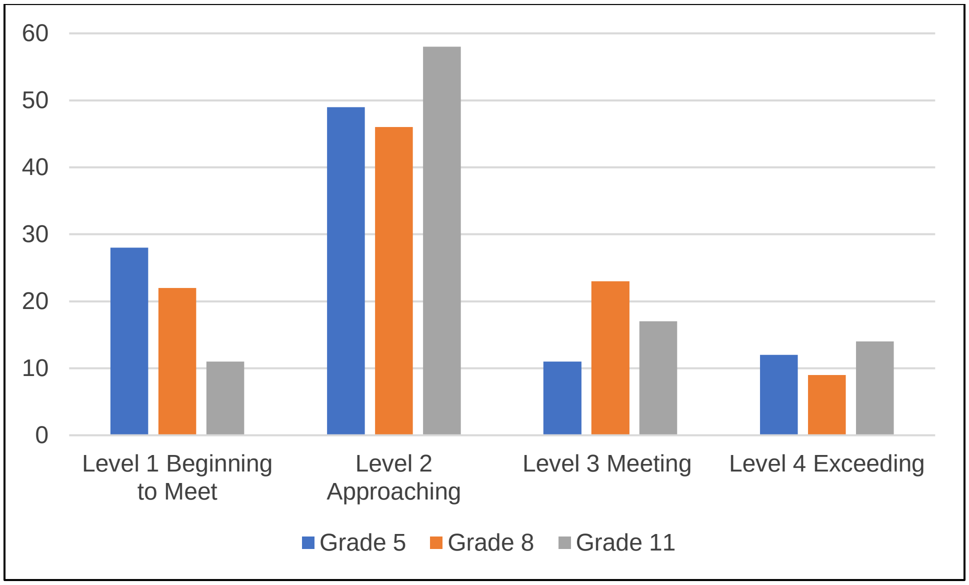
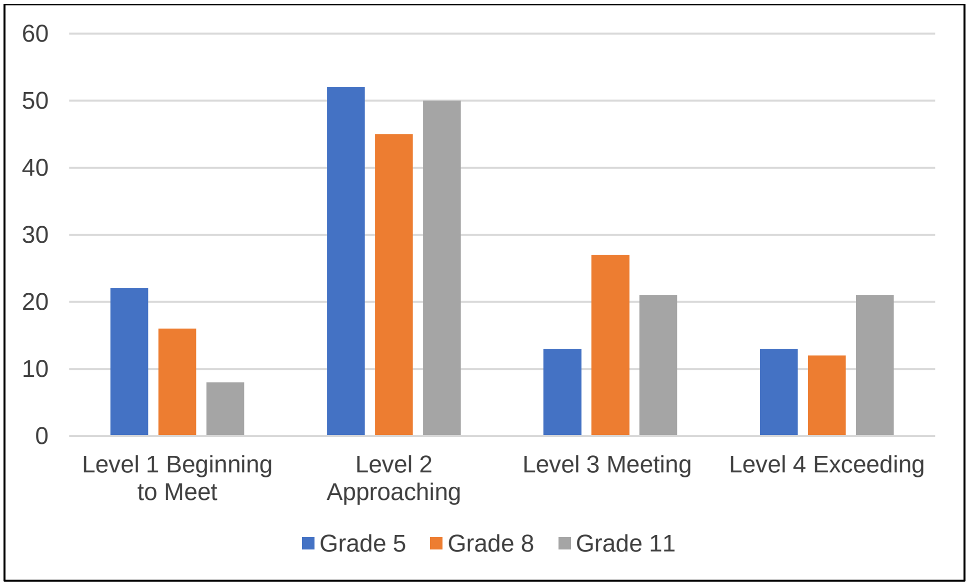


Figure 6. Percentage of Vermont Students Classified Within Each Science Achievement Level in 2019



2. Introduction

Rhode Island and Vermont adopted the Next Generation Science Standards (NGSS) in 2013. The Rhode Island Department of Education (RIDE) and the Vermont Agency of Education (VT AOE) 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. In spring 2019, they administered new assessments aligned to the NGSS to all grade 5, 8, and 11 students in Rhode Island and Vermont. These new assessments, the Rhode Island Next Generation Science Assessment (RI NGSA) and the Vermont Science Assessment (VTSA), were developed jointly by both states and measure the science knowledge and skills of Rhode Island and Vermont students in grades 5, 8, and 11.

Rhode Island provides information about its assessment on its website at https://www.ride.ri.gov/InstructionAssessment/Assessment/NGSAAssessment.aspx and at https://ri.portal.cambiumast.com/resources.

Vermont provides similar information on its website at https://education.vermont.gov/student-learning/assessments/state-and-local-assessments/science and also at https://vt.portal.cambiumast.com/resources#assessment_sm=VTSA .

New tests require new achievement standards to link achievement on the test to the content standards. RIDE and VT AOE contracted 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.

Achievement standards were recommended for grades 5, 8, and 11 science in August 2019. The purpose of this documentation is to detail the standard-setting process for the RI NGSA and the VTSA and resulting achievement standard recommendations.

3. The Next Generation Science Standards

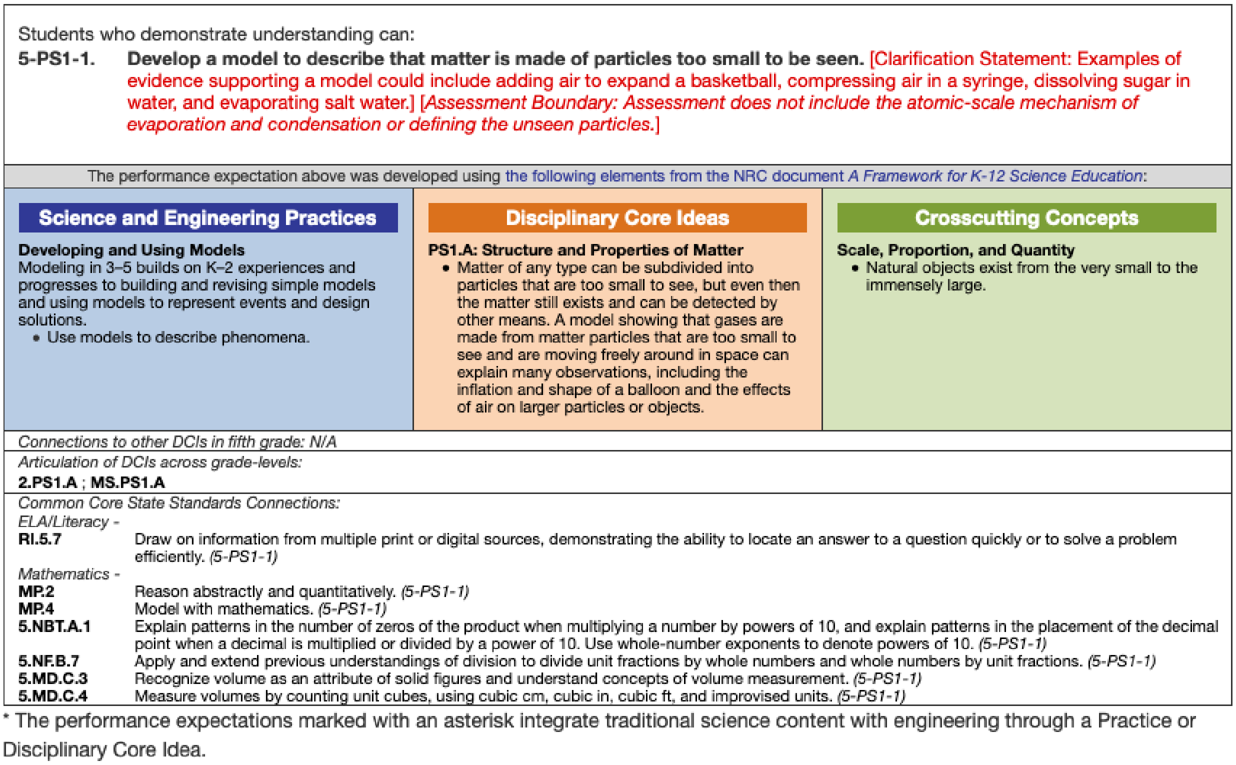
The Next Generation Science Standards (NGSS) tests assess the learning objectives described by the NGSS, adopted in 2013. Information about the NGSS is available at: www.nextgenscience.org.

These Standards reflect the latest research and advances in modern science and differ from previous science standards in multiple ways. First, rather than describe general knowledge and skills that students should know and be able to do, they describe specific performances that demonstrate what students know and can do. The NGSS refers to these performed knowledge and skills as *performance expectations*. Second, while unidimensionality is a typical goal of standards (and the assessments that measure them), the NGSS are intentionally multi-dimensional.

Each performance expectation (PE) 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 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 7 shows the structure of the NGSS for a single grade 5 performance expectation, 5-PS1-1.

Figure 7. 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. Rhode Island and Vermont's NGSS Science Assessment

Due to the unique features of the Next Generation Science Standards (NGSS), items and tests based on the NGSS, such as Rhode Island and Vermont's science assessments, 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*, which represent 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*, which is an observable fact or design problem that an engaged student explains, models, investigates, or designs, to complete a series of activities (comprised of multiple interactions) using the knowledge and skills described by the performance expectation (PE). For example, in Figure 7, 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. The stimulus in an item cluster explicitly states a task or goal (for example, "In the questions that follow, you will develop a model that will allow you to identify moons of Jupiter.") 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.

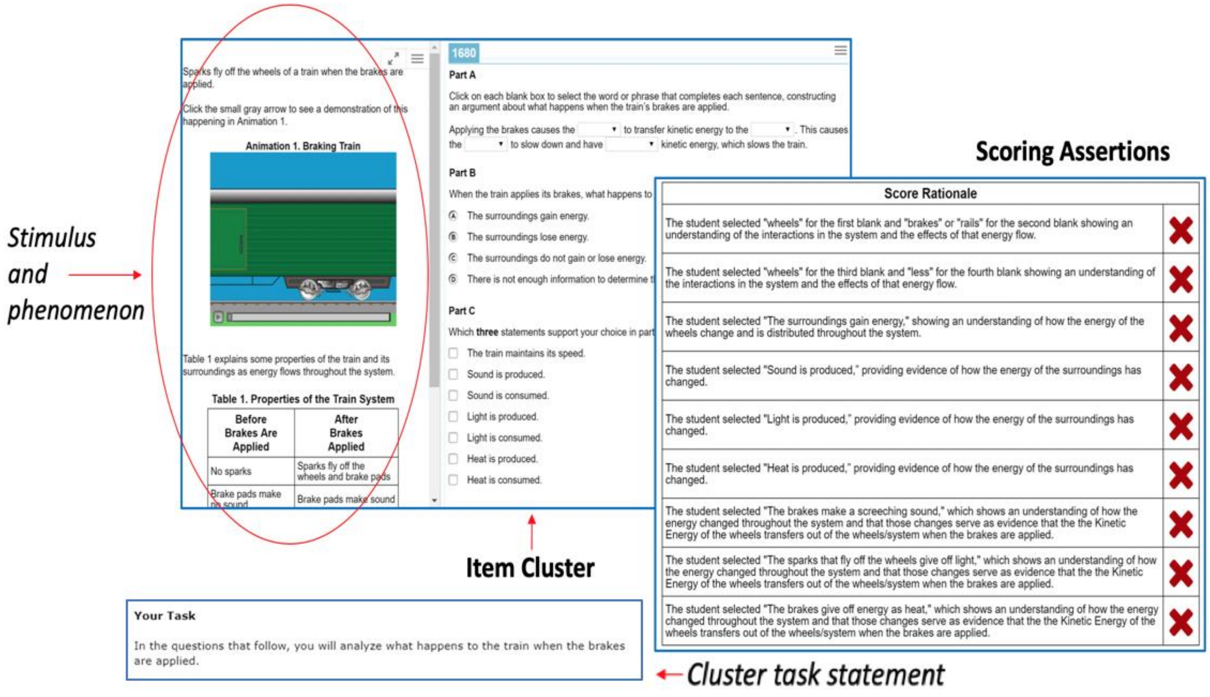
Some added stand-alone items increase the coverage of the test without also increasing testing time or testing burden. Stand-alone items are shorter, unrelated to other items, and generally take students 1–3 minutes to complete. Within each item cluster, there are a variety of interaction types including selected response, multi-select, table match, edit in-line choice, and simulations of science investigations. Stand-alone items can also be the aforementioned 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 8 illustrates an item cluster and associated scoring assertions.

Figure 8. Example NGSS Item Cluster and Scoring Assertions



Sparks fly off the wheels of a train when the brakes are applied.

Click the small gray arrow to see a demonstration of this happening in Animation 1.

Animation 1. Braking Train

Table 1 explains some properties of the train and its surroundings as energy flows throughout the system.

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

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 t

Part C

Which **three** statements support your choice in part

□ The train maintains its speed.

□ Sound is produced.

□ Sound is consumed.

□ Light is produced.

□ Light is consumed.

□ Heat is produced.

□ Heat is consumed.

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 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 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 the Kinetic Energy of the wheels transfers out of the wheels/system when the brakes are applied.	✗

Your Task

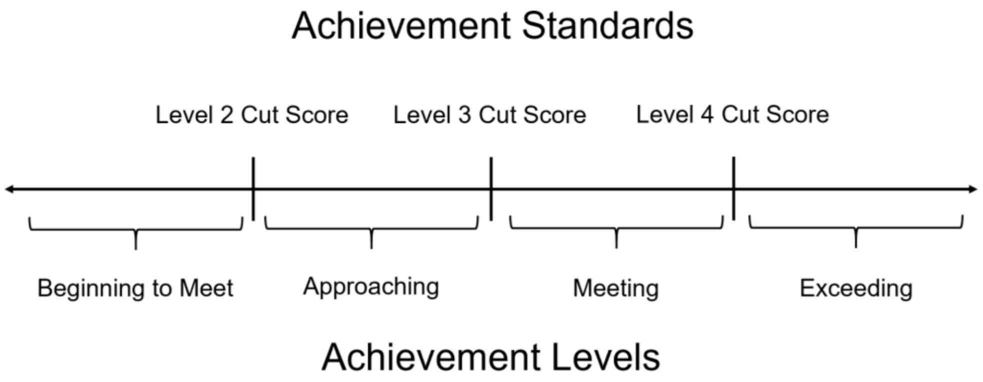
In the questions that follow, you will analyze what happens to the train when the brakes are applied.

5. Standard Setting

Twenty-six educators from Rhode Island and Vermont convened at the Grappone Conference Center in Concord, NH, from August 5–6, 2019, to complete two rounds of standard setting to recommend three achievement standards for the Rhode Island Next Generation Science Assessment (RI NGSA) and the Vermont Science Assessments (VTSA).

Standard setting is the process used to define achievement on the test. Achievement levels are defined by achievement 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 achievement level. As shown in Figure 9, three achievement standards are sufficient to define Rhode Island and Vermont's four achievement levels.

Figure 9. Three Achievement Standards Defining Rhode Island and Vermont's Four Achievement Levels



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

5.1 The Assertion-Mapping Procedure

A new approach to setting achievement standards is necessary for tests based on the Next Generation Science Standards (NGSS) due to the structure of the PEs and, subsequently, the structure of test items assessing the PEs. While traditional tests and measurement models assume unidimensionality, tests based on the Next Generation Science Standards (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 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 achievement 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 achievement 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 could also be appropriate for recommending achievement 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 achievement standards on new tests of the NGSS. AIR implemented this method for three 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.

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) 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 empirical difficulty 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. Similar to ID matching, panelists are asked to map each scoring assertion to the most apt achievement-level descriptor (ALD) 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.[1]

5.2 Workshop Structure

During the workshop, 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 10, three separate panels set achievement standards for each grade.

Figure 10. Workshop Panels Per Room

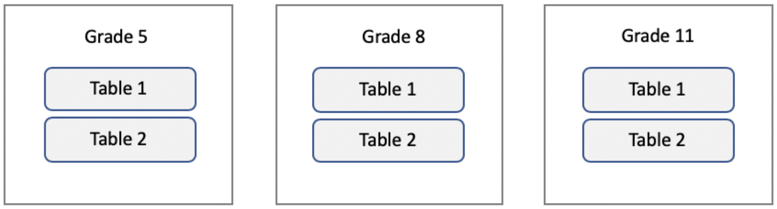


Table 4 summarizes the composition of the tables and the number of facilitators and panelists assigned to each. The 26 standard-setting participants included table leaders and panelists from Rhode Island and Vermont who taught in the content area and grade level for the standards being set.

Table 4. Table Assignments

Room	Grade	Tables (Table Leaders)	Panelists (Per Table)	Number of Panelists		Facilitator	Facilitator Assistant
				Rhode Island	Vermont		
1	5	Overall (2)	9	4	5	Jim McCann	Matt Davis
		Table 1 (1)	5	2	3		
		Table 2 (1)	4	2	2		
2	8	Overall (2)	9	7	2	Kevin Dwyer	Hibbah Haddam
		Table 1 (1)	4	3	1		
		Table 2 (1)	5	4	1		
3	11	Overall (2)	8	6	2	Meg McMahon	Kam Mangis de Mark
		Table 1 (1)	4	3	1		
		Table 2 (1)	4	3	1		

[1] AIR historically implemented two rounds of standard setting as best practice in the Bookmark method and extended this practice to the AMP method. In addition to lessening the panelists' burden of having to repeat a cognitively demanding task for a third time, using two rounds introduced significant cost efficiency by reducing the number of days needed for standard setting. Panels typically converged in round 2, and panelists completing two rounds reported 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 were generally consistent. AIR has used two rounds in standard setting in more than 16 states and 34 assessments, beginning in 2001 with the enactment of the No Child Left Behind Act (NCLB).

5.3 Participants and Roles

5.3.1 Departments of Education Staff

Staff from the Rhode Island Department of Education (RIDE) and the Vermont Agency of Education (VT AOE) were present throughout the process, provided overall policy context, and answered any policy questions that arose.

From RIDE, they included:

- Phyllis Lynch, Director, State Assessment
- Erin Escher, Science Specialist
- Kate Schulz, Instructional Improvement/Science Specialist
- Kamlyn Keith, Assessment Specialist
- Ana Karantonis, Assessment Specialist

From VT AOE, attendees included:

- Margaret Carrera-Bly, Science Specialist
- Gabriel McGann, Statewide Assessment Coordinator

5.3.2 AIR Staff

AIR (now Cambium Assessment, Inc.) facilitated the workshop and the sessions in 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 Assertion-Mapping Procedure (AMP) processes and tasks. He provided training to participants, including the facilitators and table leaders.
- Dr. Frank Rijmen, Director of Psychometrics, supervised all psychometric analyses conducted during and after the workshop.
- Dr. Mengyao Cui, Psychometrician, provided psychometric analyses.
- Alesha Ballman, Psychometric Project Coordinator, oversaw analytics technology and psychometrics.
- Azza Hussein and Matthew Andersen, Psychometric Support Assistants, provided support as needed.
- Elizabeth Mortimer, SooYun Chung, and Hannah Binder, members of the Program Management Team, managed process and logistics throughout the meeting.
- Drew Azar, System Support Agent, set up, tested, and troubleshot technology during the workshop.

5.3.3 Observers

Barbara Plake, a member of the Technical Advisory Committee (TAC) for Rhode Island and Vermont, attended the workshop. As an observer, she did not interact with panelists or impact the process in any way.

5.3.4 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 included the individuals below.

- Jim McCann served as the grade 5 room facilitator, and Matt Davis served as assistant room facilitator.
- Kevin Dwyer served as the grade 8 room facilitator, and Hibbah Haddam served as assistant room facilitator.
- Meg McMahon served as the grade 11 room facilitator, and Kam Mangis de Mark served as assistant room facilitator.

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

5.3.5 Educator Participants

To establish achievement standards, the RIDE and the VT AOE recruited a diverse variety of participants from across Rhode Island and Vermont. 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, RIDE and VT AOE targeted participants who were representative of the gender and geographic representation of the teacher population found in both states and the diversity of the students they serve. All participants also had to be familiar with NGSS content and tests.

Overall, panelists were 23% male and 8% non-white. Ninety-two percent were teachers (all of whom taught science), and 8% were either coaches or administrators. Most worked in schools (81%), and exactly half represented large districts. Panelists came from rural (38%), suburban (38%), and urban (23%) districts. Table 5 summarizes the characteristics of the panels.

Table 5. Panelist Characteristics

	Percentage of Panelists by Panel			
	Grade 5	**Grade 8**	**Grade 11**	**Overall**
Characteristics				
Male	11%	0%	63%	23%
Non-White	0%	11%	13%	8%
Stakeholder Group				
Administrator	0%	11%	0%	4%
Coach	11%	0%	0%	4%
Teacher	78%	56%	100%	77%
Teacher, Coach	11%	0%	0%	4%
Teacher, Other	0%	11%	0%	4%
Teacher, Specialist	0%	11%	0%	4%
Teacher, Specialist, Coach	0%	11%	0%	4%
Current Position				
District	0%	22%	0%	8%
School	89%	67%	88%	81%
School, District	11%	0%	13%	8%
School, District, Other	0%	11%	0%	4%
District Size				
Large	33%	56%	63%	50%
Medium	22%	22%	25%	23%
Small	44%	22%	13%	27%
District Urbanicity				
Urban	0%	44%	25%	23%
Suburban	22%	33%	63%	38%
Rural	78%	22%	13%	38%
Primary Grades Taught				
Elementary School (grades K–5)	67%	0%	0%	23%
Middle School (grades 6–8)	0%	78%	0%	27%
High School (grades 9–12)	0%	0%	100%	31%
Elementary School and Middle School (grades 1–8)	33%	22%	0%	19%
Middle School and High School (grades 6–12)	0%	0%	0%	0%
Elementary School, Middle School, and High School (all grades)	0%	0%	0%	0%
N/A (Non-educators)	0%	0%	0%	0%
Subjects Taught				

	Percentage of Panelists by Panel			
	Grade 5	Grade 8	Grade 11	Overall
Science	100%	100%	100%	100%
Other (including N/A)	0%	0%	0%	0%

For results of any judgment-based method to be valid, the judgments must be made by qualified individuals. Participants in the Rhode Island and Vermont standard-setting workshop were highly qualified and brought a variety of experience and expertise. Many had taught for more than 11 years, over a third had taught for more than 20 years, and 42% also had additional professional experience outside the classroom. Many had experience teaching special populations. In addition, 92% taught students receiving free/reduced price lunch, 69% taught English language learners (ELLs), and 96% taught students on an Individualized Educational Program (IEP). The participants represented a range of stakeholders, such as educators, administrators, parents, and business leaders. Table 6 summarizes the qualifications of the panelists.

Table 6. Panelist Qualifications

	Percentage of Panelists by Grade			
	Grade 5	Grade 8	Grade 11	Overall
Highest Degree				
Bachelors	44%	22%	13%	27%
Masters	56%	78%	88%	73%
Doctorate	0%	0%	0%	0%
Other	0%	0%	0%	0%
Years Teaching Experience				
0 years	0%	0%	0%	0%
1–5 years	22%	0%	13%	12%
6–10 years	0%	22%	13%	12%
11–15 years	22%	22%	25%	23%
16–20 years	22%	22%	13%	19%
21+ years	33%	33%	38%	35%
Years Teaching Experience in Assigned Grade/Subject				
0 years	0%	0%	0%	0%
1–5 years	56%	11%	13%	27%
6–10 years	11%	22%	13%	15%
11–15 years	22%	11%	25%	19%
16–20 years	0%	11%	13%	8%
21+ years	11%	44%	38%	31%
Other professional experience in education	33%	56%	38%	42%

	Percentage of Panelists by Grade			
	Grade 5	Grade 8	Grade 11	Overall
Years Professional Experience in Education				
0 years	67%	44%	63%	58%
1–5 years	11%	44%	25%	27%
6–10 years	11%	0%	0%	4%
11–15 years	11%	0%	0%	4%
16–20 years	0%	0%	0%	0%
21+ years	0%	11%	13%	8%
Experience Teaching Special Student Populations				
Students receiving free/reduced price lunch	89%	100%	88%	92%
English Language Learners	44%	89%	75%	69%
Students on an IEP	100%	100%	88%	96%

Note. Percentages in table describe all participants, not just educator participants. Abbreviation Key: IEP = Individualized Educational Program.

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

5.3.6 Table Leaders

The RIDE and the VT AOE pre-selected table leaders from the participant pool for their specialized knowledge or experience with the assessment, items, or NGSS. In addition to serving as panelists, table leaders had the additional responsibility of ensuring that table activities remain focused, ensuring that panelists understood their assignment and alerting workshop leaders to any issues encountered by panelists.

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:

- Led table discussions
- Helped panelists see the “big picture”
- Monitored materials security
- 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 Achievement-Level Descriptors

With the adoption of the new standards in science, and the development of new statewide tests to assess achievement of those standards, Rhode Island and Vermont adopted a similar system of achievement, or achievement standards, to determine whether students have met the learning goals defined by the new science standards.

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

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

1. *Policy ALDs*. These are brief descriptions of each achievement level that do not vary across grade or content area.
2. *Range ALDs*. 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 ALDs*. Typically created during standard setting and used for standard setting only, these describe what a student Just Barely scoring into each achievement level knows and can do. They may also be called Target ALDs or Just Barely ALDs.
4. *Reporting ALDs*: These are much-abbreviated ALDs (typically 350 or fewer characters) created following state approval of the achievement standards used to describe student achievement on score reports.

Rhode Island and Vermont use four achievement levels to describe student achievement: “Beginning to Meet Expectations,” “Approaching Expectations,” “Meeting Expectations,” and “Exceeding Expectations.” At the policy-level, these achievement levels are described as follows:

- **Beginning to Meet Expectations.** Students who achieve at this level demonstrate initial understanding of knowledge and skills needed to apply three dimensions of science to question, evaluate, and explain science phenomena. Student performance based on assessment results begins to meet grade-level expectations.
- **Approaching Expectations.** Students who achieve at this level demonstrate minimal understanding of knowledge and skills needed to apply three dimensions of science to question, evaluate, and explain science phenomena. Student performance based on assessment results partially meets grade-level expectations.
- **Meeting Expectations.** Students who achieve at this level demonstrate satisfactory understanding of knowledge and skills needed to apply three dimensions of science to question, evaluate, and explain science phenomena. Student performance based on assessment results meets grade-level expectations.

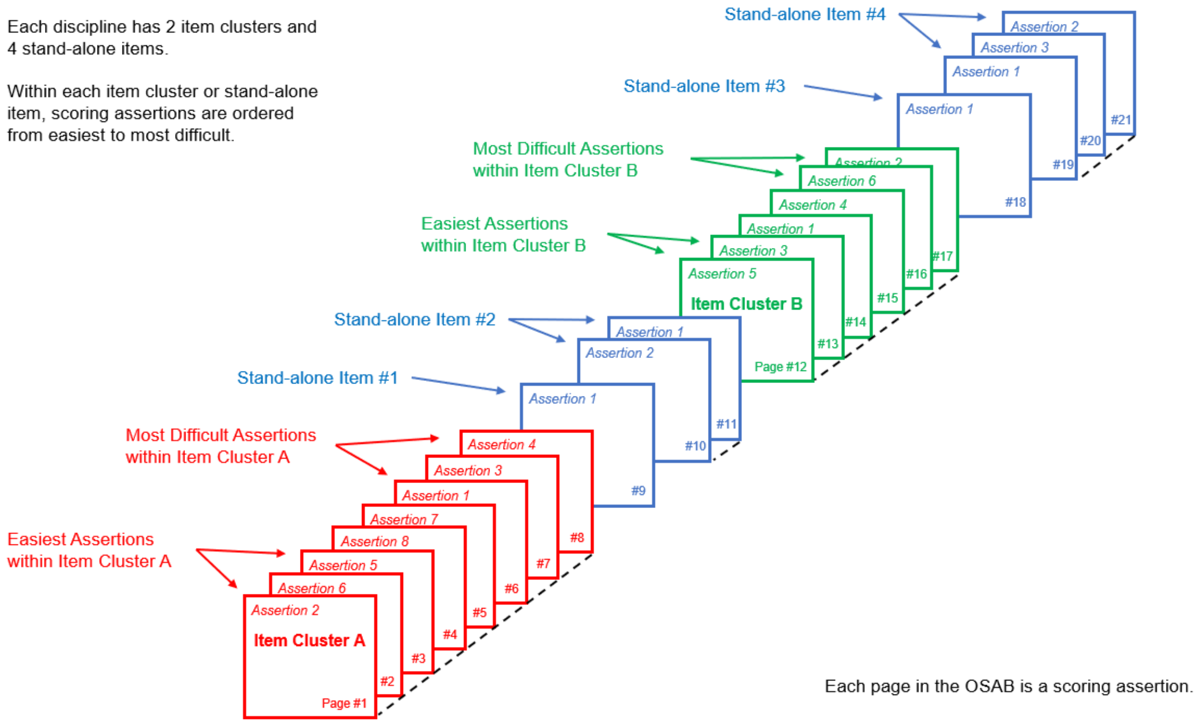
- **Exceeding Expectations.** Students who achieve at this level demonstrate advanced understanding of knowledge and skills needed to apply three dimensions of science to question, evaluate, and explain science phenomena. Student performance based on assessment results exceeds grade-level expectations.

Appendix B, Achievement-Level Descriptors, provides the final ALDs for the RI NGSA and the VTSA.

5.4.2 Ordered Scoring Assertion Booklets

Like the Bookmark method used for establishing achievement standards for traditional science tests, 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 bank. Figure 11 describes the structure of the OSAB.

Figure 11. Ordered Scoring Assertion Booklet (OSAB)



Each discipline has 2 item clusters and 4 stand-alone items.

Within each item cluster or stand-alone item, scoring assertions are ordered from easiest to most difficult.

Each page in the OSAB is a scoring assertion.

For the OSABs, the item clusters and stand-alone items are presented by discipline; 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 intermixed, just like item clusters and stand-alone items would be selected at random by the algorithm that was used to assemble operational tests linearly on the fly.

Within each item cluster or stand-alone item, scoring assertions are 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. Note that assertions were ordered by difficulty within items only. 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 11 represents the order of presentation to the panelists, but assertions were not ordered by overall difficulty across all items.

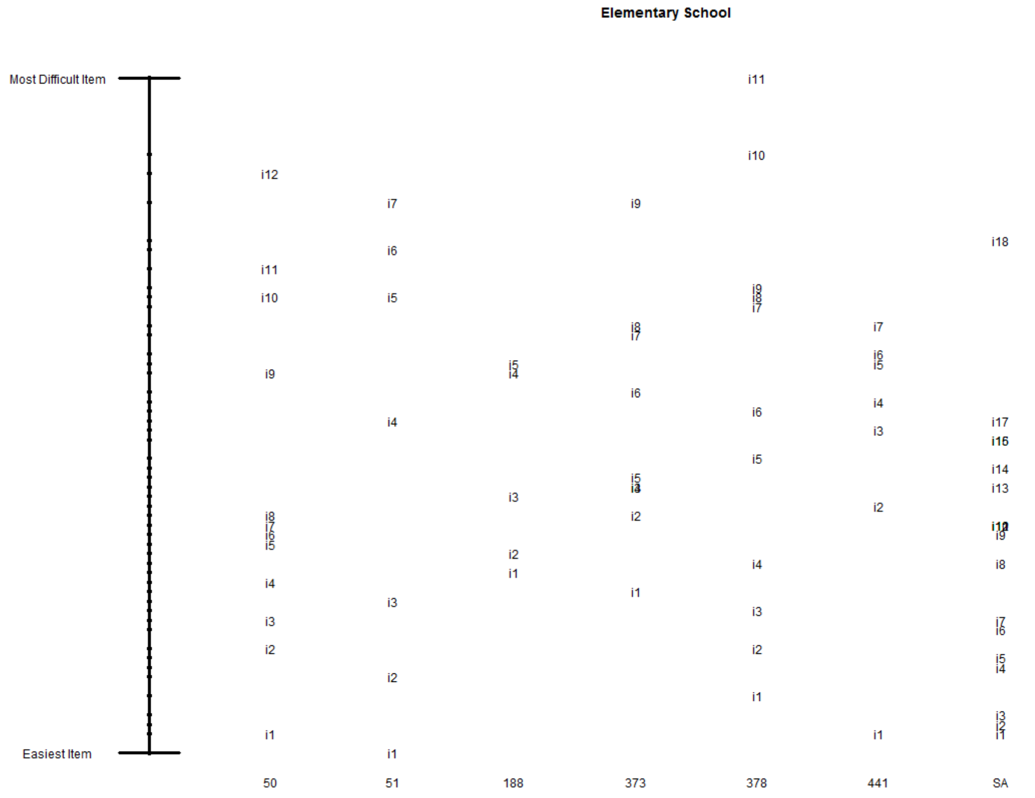
Not all items have assertions that will map onto all achievement levels. For example, an item cluster may have assertions that map onto “Beginning to Meet Expectations,” “Approaching Expectations,” and “Meeting Expectations,” but not “Exceeding Expectations.”

Each OSAB contains three disciplines and 18 items (item clusters and stand-alone items). The grade 5 OSAB contained 69 assertions, the grade 8 OSAB contained 78 assertions, and the grade 11 OSAB contained 78 assertions. Each was comprised of six item clusters and 12 stand-alone items.

5.4.3 Assertion Maps

Assertion maps listed all scoring assertions in the OSAB by page number, item ID, and item type (i.e., part of an item cluster or stand-alone item) and plotted all assertions by difficulty. The maps provided panelists with context about student performance on the assertions in the OSAB, describing the difficulty of each assertion in the underlying OSAB. This was to help panelists easily identify more- or less-difficult assertions and compare the difficulty of assertions across items. The assertion maps were provided during the OSAB review. After Round 1, the assertion maps were updated to also display the tentative standards. Figure 12 presents the assertion map for grade 5. The assertions maps for grades 8 and 11 are presented in Appendix C, Standard-Setting Assertion Maps.

Figure 12. Elementary School Assertion Map

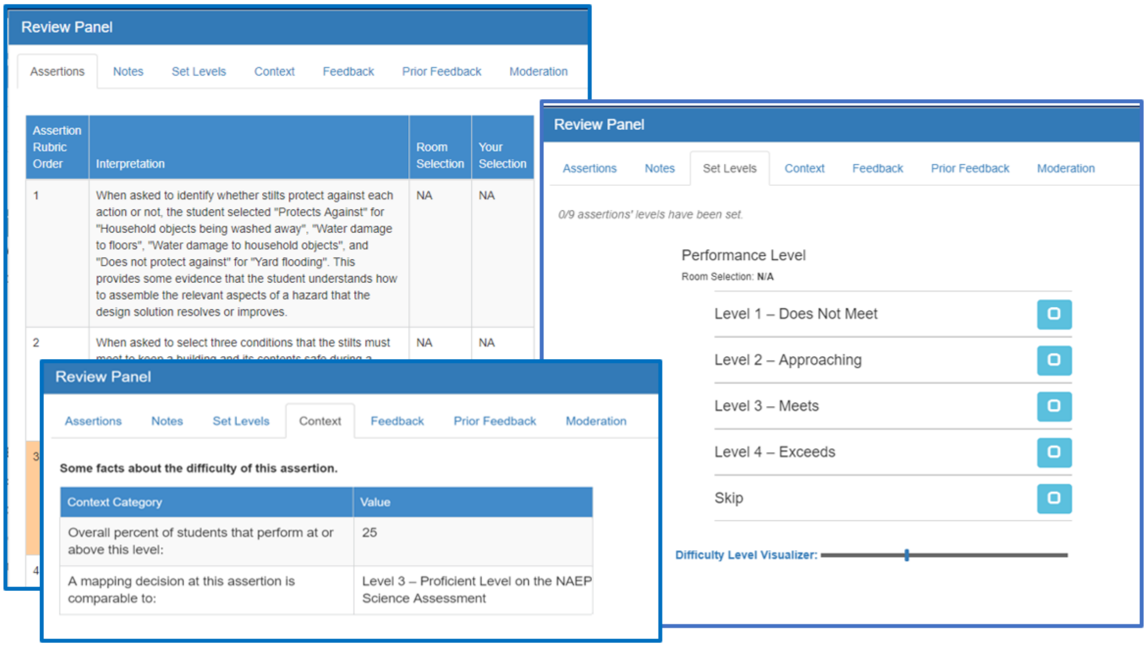


5.5 Workshop Technology

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, stand-alone items, and ancillary materials, and mapped assertions to achievement levels.

Using tabs in the review panel of the toolbar (see Figure 13), panelists could review the items and scoring assertions, 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 achievement levels, add notes and comments on the assertions as they reviewed them, and review context 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 13. Example Features in Standard-Setting Tool



Assertion Rubric Order	Interpretation	Room Selection	Your Selection
1	When asked to identify whether stilts protect against each action or not, the student selected "Protects Against" for "Household objects being washed away", "Water damage to floors", "Water damage to household objects", and "Does not protect against" for "Yard flooding". This provides some evidence that the student understands how to assemble the relevant aspects of a hazard that the design solution resolves or improves.	NA	NA
2	When asked to select three conditions that the stilts must	NA	NA

0/9 assertions' levels have been set.

Performance Level

Room Selection: **N/A**

Level 1 – Does Not Meet □

Level 2 – Approaching □

Level 3 – Meets □

Level 4 – Exceeds □

Skip □

Difficulty Level Visualizer:

Some facts about the difficulty of this assertion.

Context Category	Value
Overall percent of students that perform at or above this level:	25
A mapping decision at this assertion is comparable to:	Level 3 – Proficient Level on the NAEP Science Assessment

A 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 D, Standard-Setting Workshop Agenda, provides the full workshop agenda.

Table 7. Standard-Setting Agenda Summary

Day 1: Monday, August 5, 2019
• Table leader orientation • Registration • Large-group introductory training • Take the test • ALD review • OSAB review
Day 2: Tuesday, August 6, 2019

- OSAB review (continued)
- Assertion-mapping training
- Round 1—assertion mapping
- Round 1—feedback and context data review and discussion
- Round 2—assertion mapping
- Round 2—feedback and context data review
- 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 a 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 materials’ security
- Monitor panelist understanding and report 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 E, Standard-Setting Training Slides, provides the slides used during the table leader orientation.

5.6.2 Registration

As panelists arrived at the workshop, they received packets of materials to refer to during the workshop and signed affidavits of non-disclosure, affirming that they would not reveal any secure information they would have access to during the workshop.

5.6.3 Large-Group Introductory Training

Phyllis Lynch from RIDE and Gabriel McGann from VT AOE welcomed panelists to the workshop and provided context and background for the Rhode Island and Vermont NGSS Assessments. AIR’s Dr. Stephan 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 reviewed the responsibilities of the three groups of participants at the workshop, including panelists, AIR staff, and RIDE and VT AOE personnel. 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 E, Standard-Setting Training Slides, provides the slides used during the large-group training.

5.6.4 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 instructed 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 during 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.

5.6.5 Take the Test

Following the large-group introductory training, participants broke out into their separate grade-level rooms. As an 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 achievement 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 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.

5.6.6 Achievement-Level Descriptor Review

After taking the test, panelists completed a thorough review of the ALDs for their assigned grade. They identified key words describing the skills necessary for achievement at each level and discussed the skills and knowledge that differentiated achievement in each of the four levels.

Facilitators encouraged panelists to pay special attention to the transition areas between achievement levels and consider the characteristics of students who Just Barely qualify for entry into the achievement level from those just below. These students are not typical of students in the achievement level; they are poor examples of the achievement level, but they do Just Barely meet the expectation.

Reviewing the ALDs ensured that participants understood what students are expected to know and be able to do, how much knowledge and skills students are expected to demonstrate at each level of achievement, and how to differentiate performance at each level of achievement.

5.6.7 Ordered Scoring Assertion Booklet Review

After reviewing the ALDs, 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 ALDs.

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 achievement 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 ALDs, 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 instructed to match each assertion to the achievement level best supported by the assertion using the ALDs, the difficulty visualizer (described in Section 5.5, Workshop Technology), their notes from the OSAB review, and their professional judgments. Figure 14 graphically describes the assertion-mapping process.

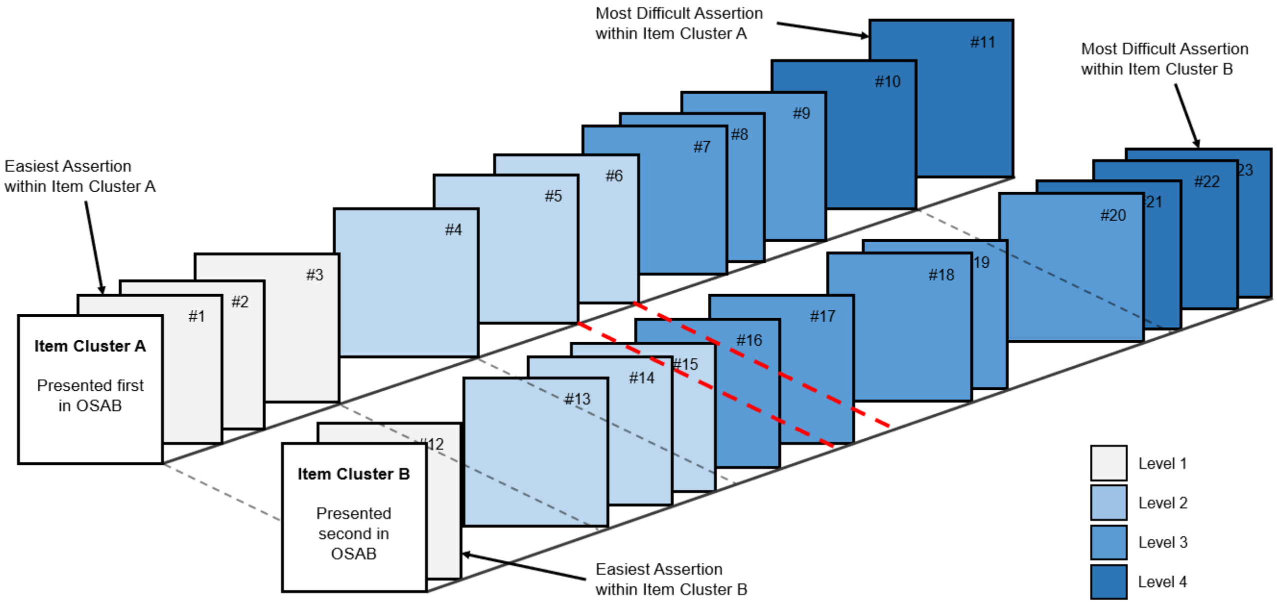
Facilitators provided the following three-part process to guide the mapping of assertions onto ALDs:

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 ALD best describes this assertion?

It was emphasized that assertions within an item were ordered by difficulty, and therefore, the assigned achievement levels should be ordered, as well. Within each item, panelists were not allowed to place an assertion into a lower achievement 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 achievement level to) the assertion. However, this was to be used as a last resort.

Because the assertion mapping was performed 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 this frequently occurred) that an assertion of one item had a higher difficulty but lower assigned achievement level than another assertion from a different item. For example, in Figure 14, 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 achievement levels to be assigned more frequently with increasing assertion difficulty across items. Appendix E, Standard-Setting Training Slides, provides the training slides used during the breakout room training.

Figure 14. Example of Assertion Mapping


Most Difficult Assertion within Item Cluster A
Most Difficult Assertion within Item Cluster B
Easiest Assertion within Item Cluster A
Item Cluster A
Presented first in OSAB
#1
#2
#3
#4
#5
#6
#7
#8
#9
#10
#11
Item Cluster B
Presented second in OSAB
12
#13
#14
#15
#16
#17
#18
19
#20
21
#22
23
Easiest Assertion within Item Cluster B
Level 1
Level 2
Level 3
Level 4

Note. Figure 14 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 an achievement scale;
- indicate on a diagram how achievement standards define achievement 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 F, Standard-Setting Practice Quiz, provides the quiz that panelists completed prior to mapping any assertions.

5.6.10 Practice Round

Following the practice quiz, panelists practiced mapping assertions to ALDs in a short practice OSAB consisting of one item cluster. 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 discussed their practice mappings and asked questions, and room facilitators provided clarifications and further instructions until everyone had successfully completed the practice round.

5.6.11 Readiness Form

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 ALDs
- The structure, use, and importance of the OSAB
- The process to determine and map assertions to ALDs in the standard-setting tool
- The readiness to begin the Round 1 task

The readiness form for Round 2 focused on affirming understanding of the context 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 G, Standard-Setting Readiness Forms, provides the form that panelists completed prior to each round of standard setting. Notwithstanding the readiness forms and additional training, the room facilitator for grade 11 flagged one panelist for not fully understanding the task of mapping assertions to ALDs. After a discussion with AIR psychometricians and RIDE and VT AOE staff, it was decided to let the panelists proceed to Round 1 but monitor the actual ratings.

5.7 Assertion Mapping

Panelists mapped assertions independently, using the ALDs, their notes from reviewing each assertion, and the difficulty visualizer to place each of the assertions into one of the four achievement 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.[2] 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 that were mapped onto the levels "Approaching" and "Meeting" were used). Specifically, let n_k be the number of assertions put at achievement 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 ith assertion. For each assertion placed at levels k and $k+1$, define the misclassification indicator as:

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

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

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

Cut scores at the table and grade level were computed using the same method while taking into account the assigned levels of all the raters at the table and grade, respectively. Applying these cut scores to the 2019 test data created data describing the percentage of students falling into each achievement level. This algorithm calculated cut scores from the assertion maps by panelist, by table, and for the room.

5.7.2 Feedback Data and Impact Data

Feedback included the cut scores corresponding to the assertion mappings for each panelist, each table, and for the room overall (across both 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 achievement standards given the assertion mappings. Percentages were calculated using real student data from

[2] 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 ALD 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).

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 achievement level to which each panelist assigned the assertion. The tool highlights assertions that panelists have assigned to different achievement 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 proficiency level associated with each individual assertion. Percentages were calculated using real student data from the Rhode Island and Vermont 2019 NGSS administration.

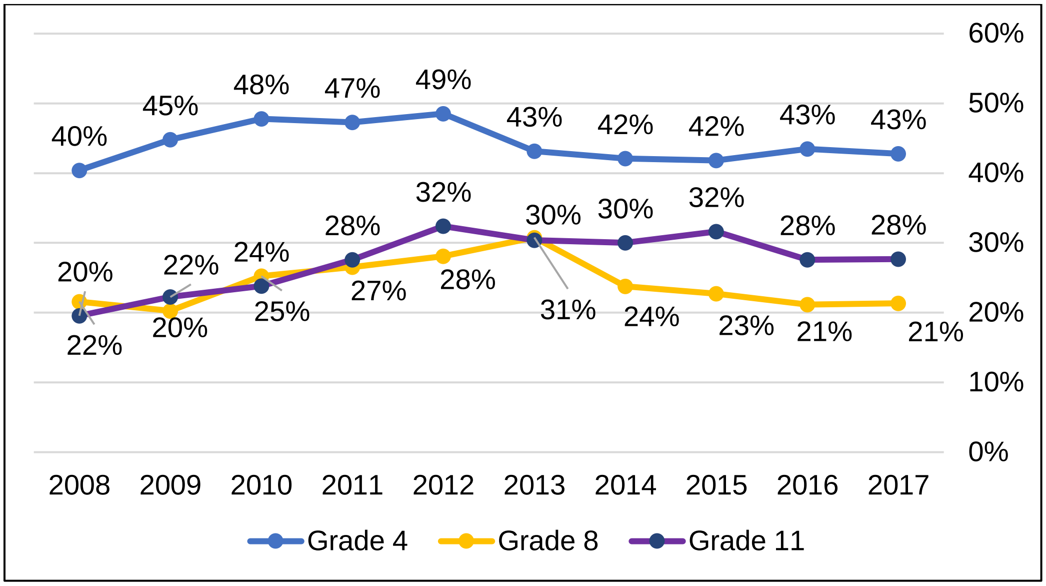
5.7.4 Articulation

To be adoptable, achievement 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. Workshop leaders described the following characteristics of well-articulated standards and asked panelists to consider articulation in Round 2:

- The cut scores for each achievement level should increase smoothly with each increasing grade.
- The cut scores should result in a reasonable percentage of students at each achievement level; reasonableness can be determined by the percentage of students in the achievement 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.

To support panelists as they considered articulation, they were provided with the percentage of students proficient on the previous science assessment (see Figure 15).

Figure 15. Rhode Island and Vermont Proficiency on New England Common Assessment Program (NECAP) Science Assessment



They were also provided with the percentage proficient on the previous National Assessment of Educational Progress (NAEP) science assessment (see Table 8).

Table 8. Achievement on NAEP Science Assessment

	Average Scale Score Grade 4	Percentage at or Above Proficient Grade 4	Average Scale Score Grade 8	Percentage at or Above Proficient Grade 8
Rhode Island	152	36	151	32
Vermont	163	48	163	44
National Public	153	37	153	33

Each table spent time reviewing and discussing the assertion mappings and context data, beginning with table-level feedback and discussion, and progressing to room-level discussion. After completing these discussions, panelists again worked through the OSAB, independently mapping assertions to achievement levels for Round 2.

5.8 Workshop Results

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

5.8.1 Round 1

Table 9 presents the achievement standards and associated context data from Round 1. Based on the Round 1 results, and depending on grade, between 61% and 95% of students fell at or above

Approaching Expectations, between 24% and 45% fell at or above Meeting Expectations, and between 1% and 11% fell at Exceeding Expectations.

Table 9. Round 1 Results

Grade and Table	Cut Scores			Context Data		
	AE	ME	EE	AE	ME	EE
Grade 5	**47**	**68**	**100**	**70**	**24**	**1**
Table 1	47	68	100	70	24	1
Table 2	53	67	78	57	26	9
Grade 8	**51**	**63**	**77**	**61**	**35**	**10**
Table 1	51	63	82	61	35	5
Table 2	41	66	77	80	28	10
Grade 11	**34**	**58**	**79**	**95**	**45**	**11**
Table 1	62	65	79	37	31	11
Table 2	34	58	72	95	45	19

Note. The grade-level row summarizes the room data (across both tables). Context data describes the percentage of students falling at or above each of the achievement standards based on the recommended Round 1 cut scores. Achievement standard: AE = Approaching Expectations, ME = Meeting Expectations, and EE = Exceeding Expectations.

After reviewing the feedback 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 achievement 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.

As discussed in Section 5.6.11, the room facilitator for grade 11 flagged one panelist before Round 1 started for having difficulties with the mapping task. The results of Round 1 confirmed this observation. The standards computed for this rater showed an aberrant pattern with a value for the “Meeting Expectations” standard lower than the value for the “Approaching Expectations” standard.

5.8.2 Round 2

Table 10 presents the recommended achievement standards and associated context data for Round 2. The panelist of grade 11 that was flagged for not understanding the mapping task again assigned mappings that resulted in the same aberrant pattern of computed achievement standards as observed after Round 1, when computing cuts based on the ALD assignments of this rater only. Therefore, the panelist was excluded from computation of the achievement standards for Round 2.

Table 10. Round 2 Results

Grade and Table	**Cut Scores**			**Context Data**		
	AE	**ME**	**EE**	**AE**	**ME**	**EE**
Grade 5	**45**	**68**	**75**	**74**	**24**	**12**
Table 1	45	68	75	74	24	12
Table 2	45	67	78	74	26	9
Grade 8	**41**	**63**	**77**	**80**	**35**	**10**
Table 1	41	63	83	80	35	5
Table 2	41	63	77	80	35	10
Grade 11	**39**	**63**	**74**	**90**	**35**	**16**
Table 1	39	66	83	90	29	8
Table 2	34	63	74	95	35	16

Note. The grade-level row summarizes the room data (across both tables). Context data describes the percentage of students falling at or above each of the achievement standards based on the recommended Round 2 cut scores. Achievement standard: AE = Approaching Expectations, ME = Meeting Expectations, and EE = Exceeding Expectations.

Based on the Round 2 results, and depending on grade, between 74% and 90% of students would fall at or above Approaching Expectations, between 24% and 35% would fall at or above Meeting Expectations, and between 10% and 16% would fall at Exceeding Expectations.–Figure 16 represents those values graphically.

Figure 16. Percentage of Students Reaching or Exceeding Each Recommended Science Achievement Standard in 2019

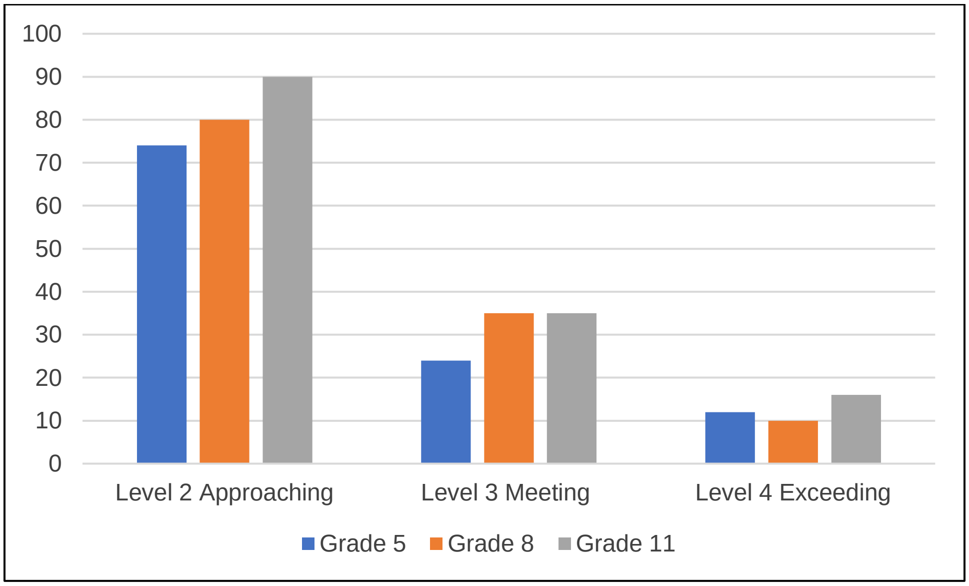


Table 11 indicates the percentage of students classified within each of the achievement levels in 2019. The values are displayed graphically in Figure 17 through Figure 19.

Table 11. Percentage of Students Classified Within Each Recommended Science Achievement Level in 2019

Grade	State	Level 1 Beginning to Meet	Level 2 Approaching	Level 3 Meets	Level 4 Exceeds
5	**Combined**	**26**	**50**	**12**	**12**
	Rhode Island	28	49	11	12
	Vermont	22	52	13	13
8	**Combined**	**20**	**45**	**25**	**10**
	Rhode Island	22	46	23	9
	Vermont	16	45	27	12
11	**Combined**	**10**	**55**	**19**	**16**
	Rhode Island	11	58	17	14
	Vermont	8	50	21	21

Figure 17. Percentage of Combined Students Classified Within Each Recommended Science Achievement Level in 2019

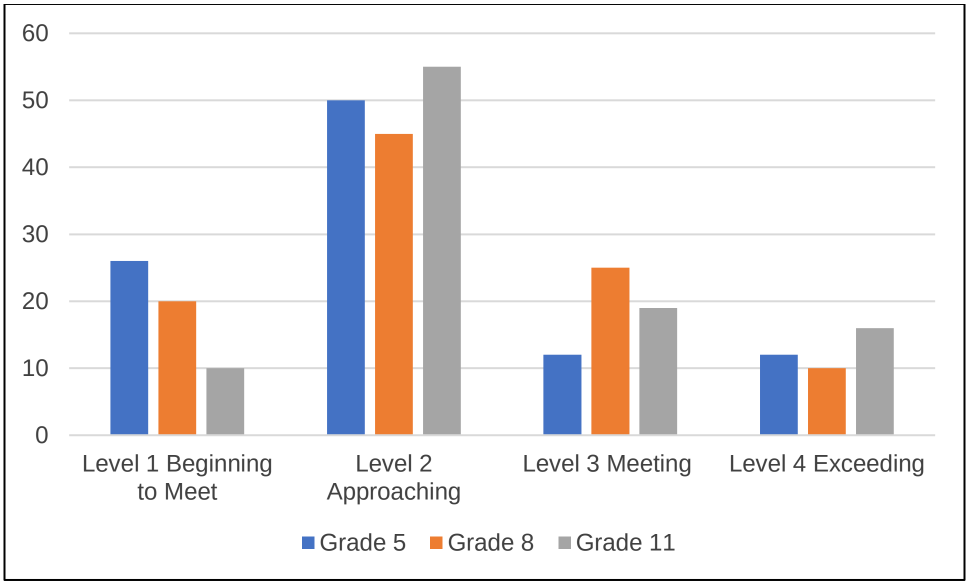


Figure 18. Percentage of Rhode Island Students Classified Within Each Recommended Science Achievement Level in 2019

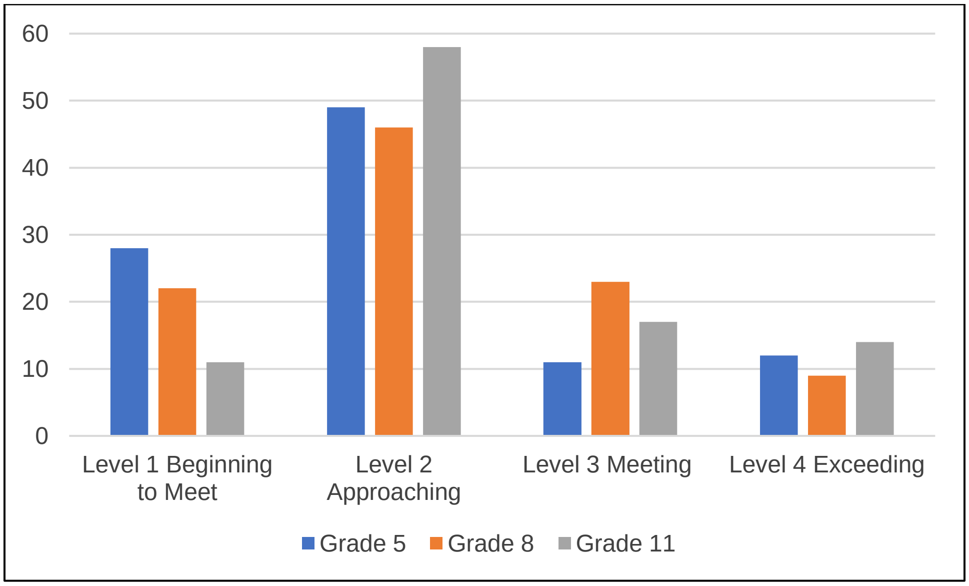
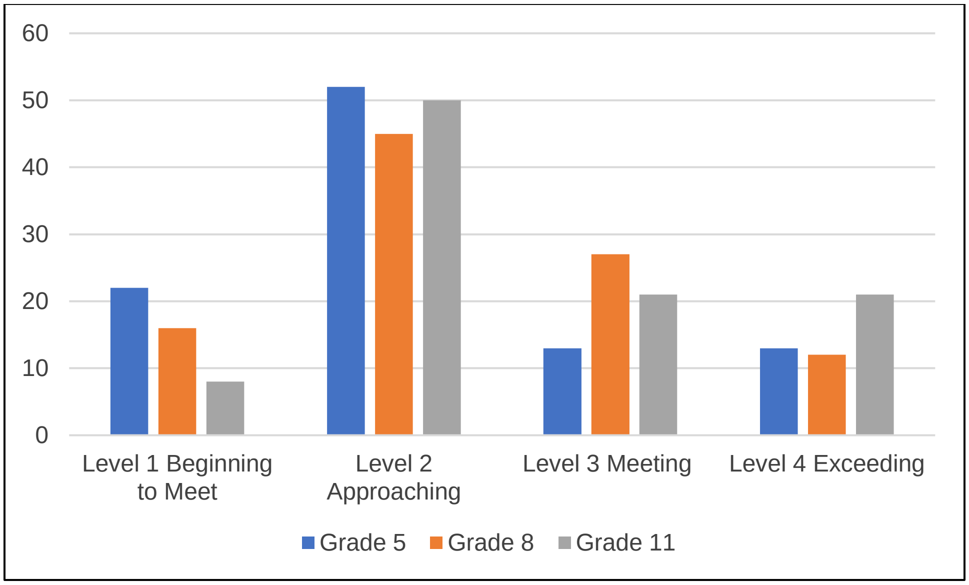


Figure 19. Percentage of Vermont Students Classified Within Each Recommended Science Achievement Level in 2019



5.8.3 Post-Workshop Refinements

Following the workshop, the RIDE and the VT AOE made some refinements to the workshop recommendations by lowering some cut scores. Table 12 presents the final achievement standards for state adoption. Figure 20 through Figure 22 represent those values graphically.

Table 12. Post-Standard-Setting Workshop: Final Cut Scores (Change from Workshop Recommendation) and Context Data

Grade	State	Cut Scores (Revision)			Context Data		
		AE	ME	EE	AE	ME	EE
5	**Combined**	40 (–5)	63 (–5)	75	**83**	**34**	**12**
	Rhode Island				81	32	12
	Vermont				85	38	13
8	**Combined**	41	63	77	**80**	**35**	**10**
	Rhode Island				78	32	9
	Vermont				84	39	12
11	**Combined**	39	63	74	**90**	**35**	**16**
	Rhode Island				89	31	14
	Vermont				92	42	21

Note. Context data describes the percentage of students falling at or above each of the achievement standards based on the final cut scores. Achievement standard: AE = Approaching Expectations, ME = Meeting Expectations, and EE = Exceeding Expectations.

Figure 20. Post-Standard-Setting Workshop: Percentage of Combined Students Reaching or Exceeding Each Science Achievement Standard in 2019

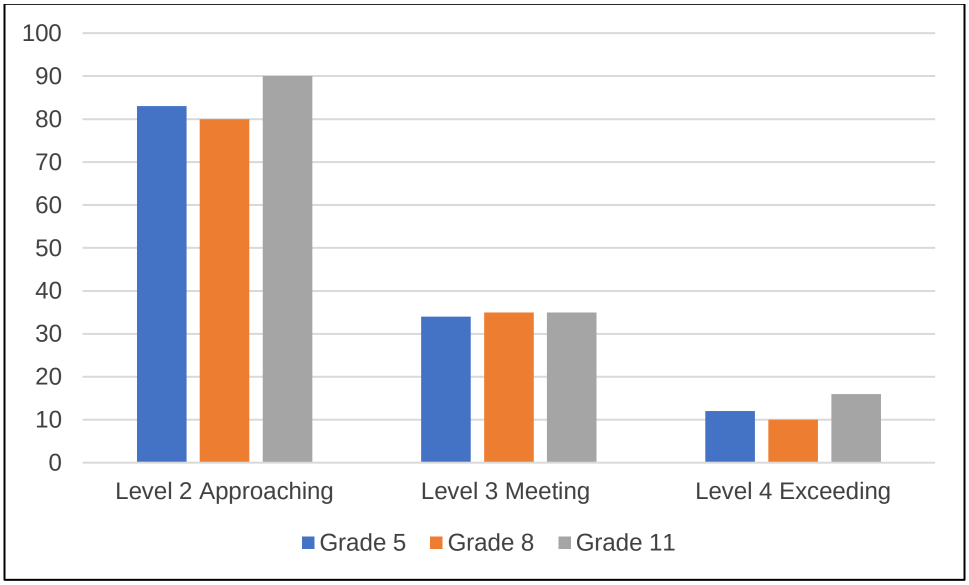


Figure 21. Post-Standard-Setting Workshop: Percentage of Rhode Island Students Reaching or Exceeding Each Science Achievement Standard in 2019

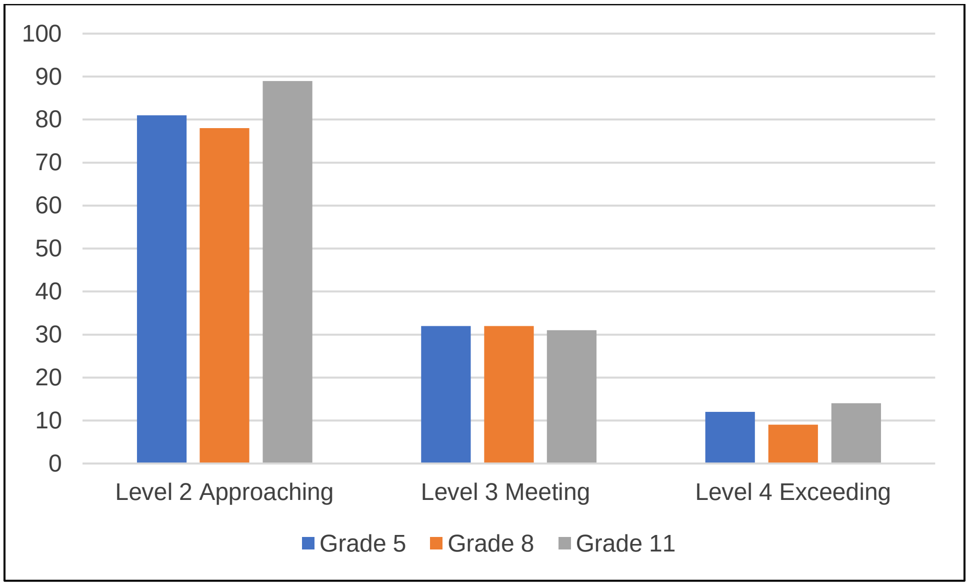


Figure 22. Post-Standard-Setting Workshop: Percentage of Vermont Students Reaching or Exceeding Each Science Achievement Standard in 2019

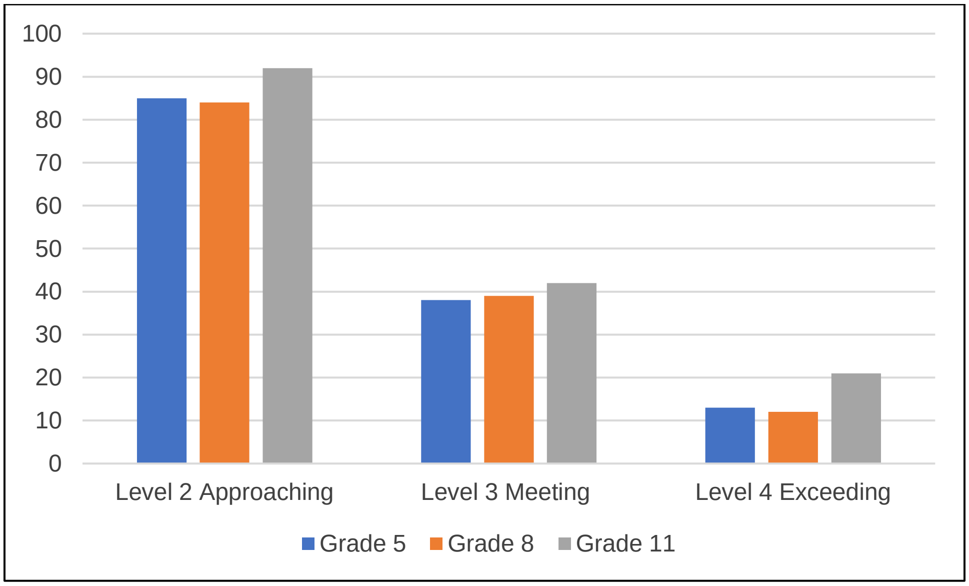


Table 13 indicates the percentage of students classified within each of the achievement levels in 2019 resulting from RIDE and VT AOE refinements to the recommended achievement standards. The values are displayed graphically in Figure 23 through Figure 25.

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

Grade	State	Level 1 Beginning to Meet	Level 2 Approaching	Level 3 Meeting	Level 4 Exceeding
5	**Combined**	**17**	**49**	**22**	**12**
	Rhode Island	19	49	20	12
	Vermont	15	47	25	13
8	**Combined**	**20**	**45**	**25**	**10**
	Rhode Island	22	46	23	9
	Vermont	16	45	27	12
11	**Combined**	**10**	**55**	**19**	**16**
	Rhode Island	11	58	17	14
	Vermont	8	50	21	21

Figure 23. Post-Standard-Setting Workshop: Percentage of Combined Students Classified Within Each Science Achievement Level in 2019

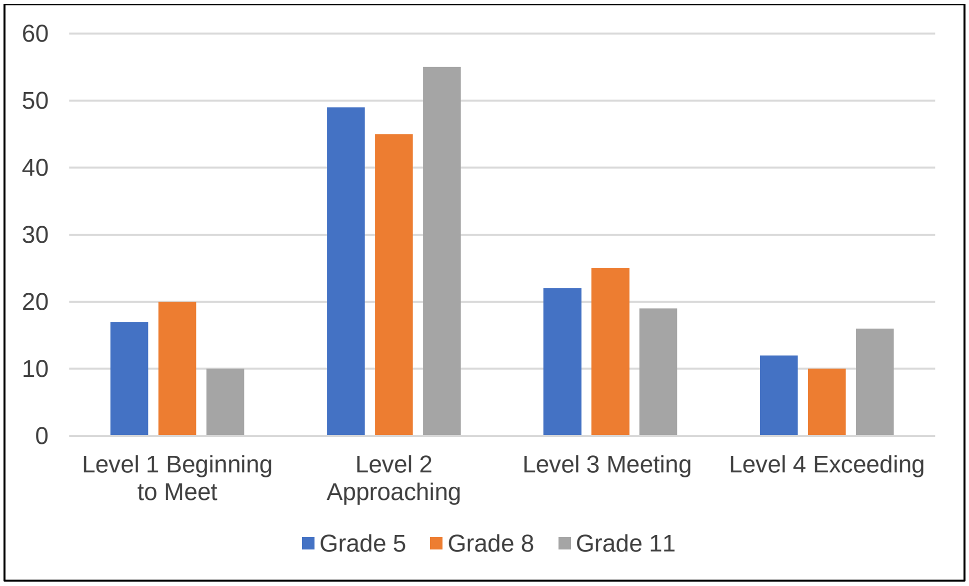


Figure 24. Post-Standard-Setting Workshop: Percentage of Rhode Island Students Classified Within Each Science Achievement Level in 2019

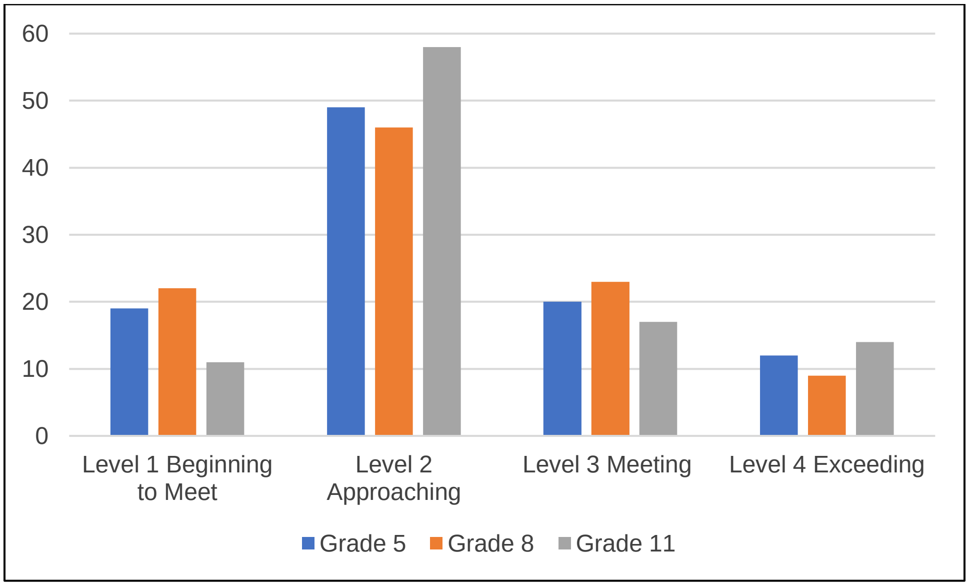
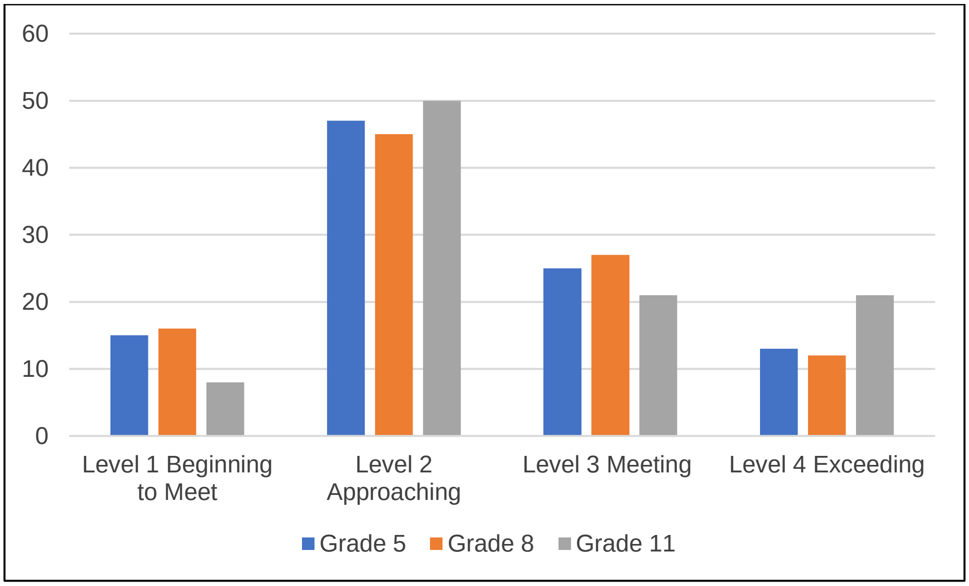


Figure 25. Post-Standard-Setting Workshop: Percentage of Vermont Students Classified Within Each Science Achievement Level in 2019



As mentioned in Section 1.1.2, Results of the Standard-Setting Workshop, the scale for each grade will be re-centered around the Level 3 standard after final approval of the standards. After the post workshop refinements, the Level 3 cut score was set at 63 on the proposed scale for all three grades. In order to center the reporting scale around the Level 3 standard, the scale was translated by minus 3, the difference between tentative and final cut scores expressed on the reporting scale. Table 14 presents the final achievement standards after centering around the Level 3 standard. The percentages at or above each of the achievement standards are not affected.

Table 14. Final Cut Scores After Re-Centering Around Level 3 Standards

Grade	Cut Scores		
	AE	ME	EE
5	37	60	72
8	38	60	74
11	36	60	71

Note. Achievement standard: AE = Approaching Expectations, ME = Meeting Expectations, and EE = Exceeding Expectations.

5.9 Workshop Evaluations

After finishing all activities, panelists completed online workshop evaluations independently, in which they described and evaluated their experience taking part in the standard setting. Table 15 through Table 19 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.

Generally, workshop participants indicated clarity in the instructions, materials, data, and process (see Table 15). However, 63% of grade 11 panelists indicated the ALDs were clear and 75% of grade 5 panelists indicated the OSABs were clear.

Table 15. 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	88%	100%	88%	92%
Achievement-Level Descriptors (ALDs)	100%	100%	63%	88%
Ordered Scoring Assertion Booklet (OSAB)	75%	100%	100%	92%
Panelist agreement data	100%	100%	100%	100%
Context data (percentage of students who would reach any standard you select)	88%	100%	88%	92%
Assertion map	100%	100%	88%	96%

Note. Number of responses = 25 (grade 5 responses = 8, grade 8 responses = 9, and grade 11 responses = 8). Evaluation options included "Very Unclear," "Somewhat Unclear," "Somewhat Clear," and "Very Clear."

Some panelists indicated having too much time to complete some tasks (see Table 16). Nine panelists indicated the large-group training was too long, six indicated having too little time to review ALDs, and two indicated having too much time to review the ALDs. Five panelists indicated having too much time for mapping scoring assertions, while three reported spending too much time on the Round 1 discussion, and one reported not spending enough time on the Round 1 discussion.

Table 16. 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	63%	78%	50%	64%
Experiencing the online assessment	88%	100%	75%	88%
Reviewing the Achievement-Level Descriptors (ALDs)	50%	100%	50%	68%
Reviewing the Ordered Scoring Assertion Booklet (OSAB)	88%	100%	75%	88%
Mapping your scoring assertions to achievement levels in each round	63%	89%	88%	80%
Round 1 discussion	88%	100%	63%	84%

Note. Number of responses = 25 (grade 5 responses = 8, grade 8 responses = 9, and grade 11 responses = 8). Evaluation options included “Too Little,” “Too Much,” and “About Right.”

Participants appreciated the importance of the multiple factors contributing to assertion mapping, with all but a single panelist in some grades rating each factor as important or very important (see Table 17).

Table 17. Evaluation Results: Importance of Materials

How important were each of the following factors in your mapping of scoring assertions to achievement levels?	Percentage Responding “Somewhat Important” or “Very Important”			
	Grade 5	Grade 8	Grade 11	Overall
Achievement-Level Descriptors (ALDs)	100%	100%	88%	96%
Your perception of the difficulty of the scoring assertions and items in general	88%	100%	88%	92%
Your experience with students	100%	100%	100%	100%
Discussions with other panelists	100%	100%	100%	100%
Room agreement data (room, table, and individual cuts)	100%	100%	88%	96%
Context data (percentage of students who would reach any standard you select)	88%	100%	88%	92%
Assertion map	100%	100%	88%	96%

Note. Number of responses = 25 (grade 5 responses = 8, grade 8 responses = 9, and grade 11 responses = 8). Evaluation options included “Not Important,” “Somewhat Important,” and “Very Important.”

Although participant understanding of the workshop processes and tasks was high (see Table 18), three grade 11 panelists disagreed that the procedures used were fair and unbiased, four panelists disagreed that the ALDs provided clear expectations, and three panelists indicated the context data were not helpful.

Table 18. 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 achievement standards were fair and unbiased.	100%	100%	63%	88%
The training provided me with the information I needed to recommend achievement 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%	89%	100%	96%
The Achievement-Level Descriptors (descriptions of what students within each achievement level are expected to know and be able to do) provided a clear picture of expectations for student achievement at each level.	75%	100%	75%	84%
I understood how to review each assertion in the Ordered Scoring Assertion Booklet 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 achievement level.	100%	100%	100%	100%
I found the assertion map helpful in my decisions about the assertions I mapped to achievement levels.	100%	100%	88%	96%
I found the context data (percentage of students who would achieve at the level indicated by the assertion difficulty) and discussions helpful in my decisions about the assertions I mapped to achievement levels.	88%	100%	75%	88%
I found the panelist agreement data (room, table, and individual cuts) and discussion helpful in my decisions about assertions I mapped to achievement levels.	100%	100%	88%	96%
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 = 25 (grade 5 responses = 8, grade 8 responses = 9, and grade 11 responses = 8). 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 19).

Table 19. 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 expectations for the grade.	100%	100%	100%	100%

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 3 is meeting expectations for the grade.	100%	100%	100%	100%
A student performing at Level 4 is exceeding expectations for the grade.	100%	89%	100%	96%

Note. Number of responses = 25 (grade 5 responses = 8, grade 8 responses = 9, and grade 11 responses = 8). Evaluation options included "Strongly Disagree," "Disagree," "Agree," and "Strongly Agree."

5.9.1 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."

Twenty-three panelists responded to the first question, and nine responded to the second. Most responses indicated the training was effective and the process was clear. Participants provided minor suggestions, such as shortening or lengthening the time allocated for some tasks, making the rooms smaller or the tables larger, and providing less practice time and more task completion time. Many commented on the value of discussions and interactions with other panelists.

Additional participant comments included:

> *"Thank you for the opportunity and the experience. Greatly appreciated."*

> *"I am quite pleased that I was selected to work on this and provide input. While the task was quite intense, it was a valuable learning experience."*

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 necessary 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 *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 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 judgment 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 Sections 5.6 through 5.9.1 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 achievement 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 achievement level;
- identifying which test item scoring assertions students at each achievement 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 5.3.5, 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. This set of empirical data provides necessary and additional context describing student performance given the recommended standards.

Further evidence of the validity of the AMP as a standard-setting process and the adherence to professional standards and best practices is provided by the observations of an independent standard-setting expert. The observations of Dr. Barbara Plake, who was present during the entire standard-setting workshop, are presented in Appendix H. Synopsis of Validity Evidence for the Cutscores. Dr. Plake concluded her report as follows:

> *These steps [of the standard-setting workshop] are consistent with current practice for the conducting a test-centered standard-setting method. For the most part, these steps were successfully implemented, and when minor issues emerged, they were handled immediately and appropriately. There is no evidence to suggest that there is any reason to question the validity of the resultant cut scores produced by these panels.*

The Rhode Island and Vermont Technical Advisory Committee for the Science Assessment also endorsed the standard-setting method and the final standards during their October 2019 meeting.

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 critical elements described in this section 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

Rhode Island and Vermont educators played a critical role in establishing achievement levels for the Multi-State Science Assessment (MSSA). They created the item clusters, reviewed and revised the achievement-level descriptors (ALDs), mapped assertions to achievement levels to delineate performance at each achievement level, considered benchmark data and the impact of their recommendations, and formally recommended achievement standards.

Many subject-matter experts contributed to developing Rhode Island's and Vermont's achievement 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.

Four pieces of evidence to support this critical element include:

1) The rationale for and technical sufficiency of the AMP method selected to establish achievement standards (Section 5.1)

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, 5.9, and 6.1)

3) Panelists self-reported readiness to undertake the task (Sections 5.6.9 and 5.6.11) and confidence in the workshop process and outcomes (Section 5.9) 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 Rhode Island's and Vermont's science content standards, and individuals with experience and expertise teaching special population and general education students in Rhode Island and Vermont (Section 5.3.5, Educator Participants, and Appendix 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.

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* (pp. 249–281). Mahwah, NJ: Lawrence Erlbaum Associates.

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.

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.

Appendix A

Standard-Setting Panelist Characteristics

Standard-Setting Panelist Characteristics

Table A-1. Standard-Setting Panelists, Science Grade 5

State	Position	Gender	Race/ Ethnicity	Level of Education	Years Teaching Experience	Years Professional Experience	Table Leader
Rhode Island	Teacher, Coach	Female	White	Bachelor's degree	21+ years	1–5 years	Yes
Vermont	Teacher	Female	White	Bachelor's degree, +45 hours in graduate classes	16–20 years	0 years	
Rhode Island	Teacher	Female	White	Bachelor's degree, Master's degree, National Board Certified	21+ years	11–15 years	
Vermont	Teacher	Male	White	Master's degree	11–15 years	0 years	
Vermont	Teacher	Female	White	Bachelor's degree, Master's degree	1–5 years	0 years	
Vermont	Teacher	Female	White	Bachelor's degree	1–5 years	0 years	Yes
Vermont	Coach	Female	White	Master's degree	11–15 years	0 years	
Rhode Island	Teacher	Female	White	Bachelor's degree	21+ years	0 years	
Rhode Island	Teacher	Female	White	Master's degree	16–20 years	6–10 years	

Table A-2. Standard-Setting Panelists, Science Grade 8

State	Position	Gender	Race/ Ethnicity	Level of Education	Years Teaching Experience	Years Professional Experience	Table Leader
Rhode Island	Teacher, Department Head K–12	Female	White	Bachelor's degree, Master's degree	16–20 years	1–5 years	Yes
Rhode Island	Administrator	Female	White	Master's degree	11–15 years	21+ years	
Rhode Island	Teacher, Specialist	Female	White	Master's degree	16–20 years	0 years	
Vermont	Teacher, Specialist, Coach	Female	White	Master's degree	6–10 years	1–5 years	
Vermont	Teacher	Female	White	Bachelor's degree, Master's degree	11–15 years	0 years	Yes
Rhode Island	Teacher	Female	White	Master's degree	6–10 years	0 years	
Rhode Island	Teacher	Female	Asian	Bachelor's degree	21+ years	0 years	
Rhode Island	Teacher	Female	White	Bachelor's degree, Master's degree	21+ years	1–5 years	
Rhode Island	Teacher	Female	White	Bachelor's degree	21+ years	1–5 years	

Table A-3. Standard-Setting Panelists, Science Grade 11

State	Position	Gender	Race/ Ethnicity	Level of Education	Years Teaching Experience	Years Professional Experience	Table Leader
Rhode Island	Teacher	Female	White	Master's degree	21+ years	21+ years	Yes
Vermont	Teacher	Male	East Asian & White	Master's degree	11–15 years	0 years	
Rhode Island	Teacher	Male	White	Bachelor's degree, Master's degree	21+ years	0 years	
Rhode Island	Teacher	Female	White	Master's degree	1–5 years	0 years	
Vermont	Teacher	Male	White	Master's degree	16–20 years	0 years	Yes
Rhode Island	Teacher	Female	White	Master's degree	11–15 years	1–5 years	
Rhode Island	Teacher	Male	White	Master's degree	21+ years	1–5 years	
Rhode Island	Teacher	Male	White	Bachelor's degree	6–10 years	0 years	