New Hampshire Statewide Assessment System

2018–2019

Volume 3

Setting Performance Standards

New Hampshire
Department of Education
Serving New Hampshire's Education Community

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

1. Executive Summary of English Language Arts and Mathematics

American Institutes for Research (AIR) conducted standard-setting workshops for the New Hampshire Department of Education (NHDOE) in English language arts (ELA) and mathematics grades 3–8 from July 25–27, 2018, at the Courtyard Marriott Grappone Conference Center in Concord, New Hampshire.

The ELA and mathematics assessments are based on the Independent College and Career Readiness (ICCR) item pools developed by AIR for use in statewide assessments. New Hampshire selected items from the item pool that meet the unique needs and requirements of the state and match the NHDOE blueprints aligned to the New Hampshire College and Career Readiness Standards (NH CCRS).

On July 24, 2018, all table leaders who would be at the standard-setting workshop participated in an additional stakeholder's meeting during which they reviewed, modified, and approved the Performance-Level Descriptors (PLDs) used in the standard-setting workshop.

AIR used a process of standard setting called the Bookmark method for ELA and mathematics. The Bookmark method is the nation's most commonly used standard-setting procedure, and it has been used successfully to set performance-level cut scores for previous assessments in New Hampshire.

In the Bookmark method, panelists are given a booklet of items ordered by difficulty from easy to hard. After reviewing the New Hampshire College and Career Readiness Standards (NH CCRS), the PLDs, and the student performance distribution, panelists then bookmarked the page they feel represents the cutoff between each of the state's three performance levels: *Approaching Proficient*, *Proficient*, and *Above Proficient*. The panelists repeated this process across two rounds of standard setting during a three-day workshop. These cuts ultimately determined the percentage of students in each of the state's four performance levels: *Below Proficient*, *Approaching Proficient*, *Proficient*, and *Above Proficient*.

Fifty-eight panelists (29 for ELA and 29 for mathematics) were selected to participate in the workshop. The panelists represented experienced classroom teachers, curriculum specialists, education administrators, and other stakeholders. The composition of the panel ensured that a diverse range of perspectives contributed to the standard-setting process. The panel was also representative in terms of gender, race/ethnicity, and region of the state.

1.1 Overall Structure of the ELA and Mathematics Standard-Setting Workshops

The key features of the workshops included the following:

- The standard-setting process produced three cut scores per grade, corresponding to *Approaching Proficient*, *Proficient*, and *Above Proficient* performance levels.

- There were two rounds per grade for all cut scores.

- Impact data (percentage of students reaching each cut score in the spring 2018 test administration) was available in the second round.
- A computer was available for each panelist at the workshops. The standard-setting workshops were conducted online using AIR's online standard-setting tool.

1.2 Results of the ELA and Mathematics Standard-Setting Workshops

Table 1 and Figure 1 describe the performance standards recommended for ELA.

Table 1: Performance Standards Recommended for ELA

Grade	**Performance Level**		
	Approaching Proficient	***Proficient***	***Above Proficient***
3	557	587	616
4	580	605	635
5	594	621	664
6	605	642	688
7	608	644	697
8	625	661	711

Figure 1. Performance Standards Recommended for ELA

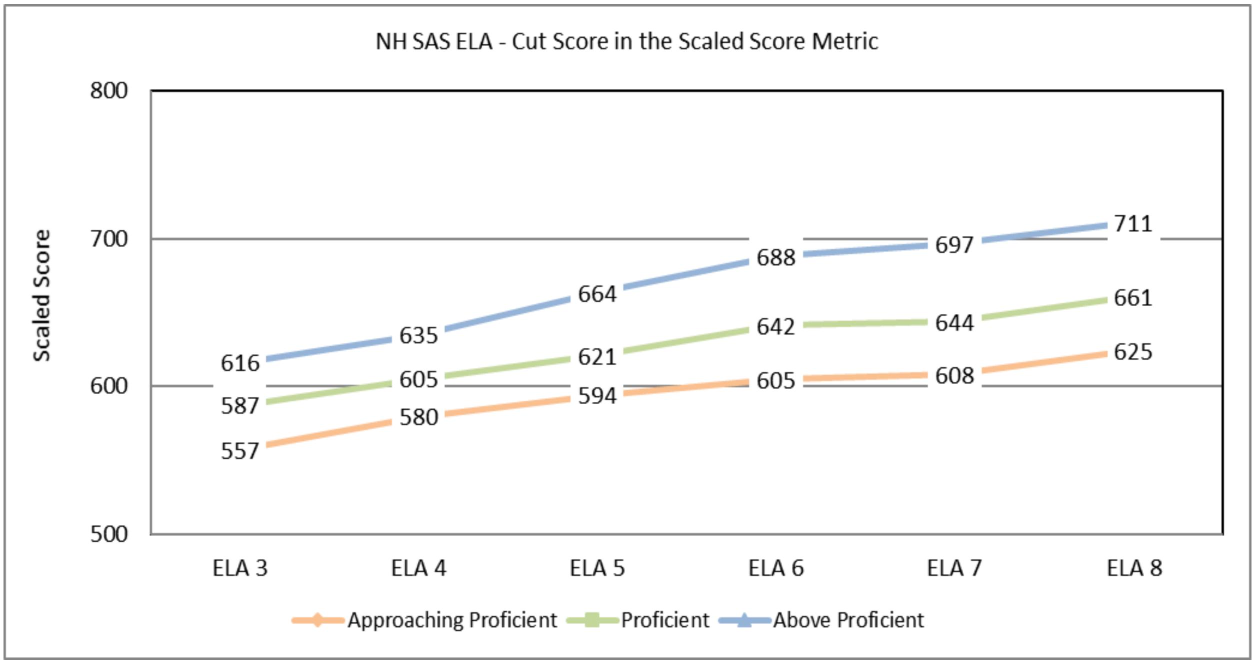


Table 2 and Figure 2 describe the performance standards recommended for mathematics.

Table 2: Performance Standards Recommended for Mathematics

Grade	Performance Level		
	Approaching Proficient	***Proficient***	***Above Proficient***
3	410	431	455
4	431	460	492
5	460	495	522
6	479	518	556
7	507	552	587
8	539	591	625

Figure 2. Performance Standards Recommended for Mathematics

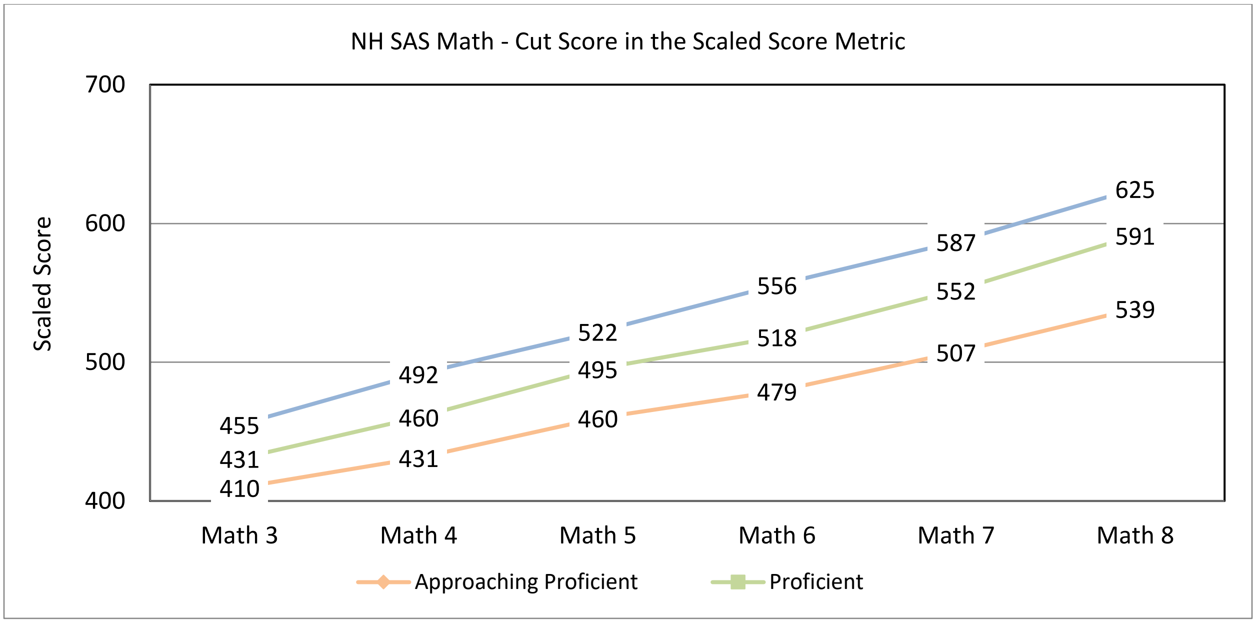
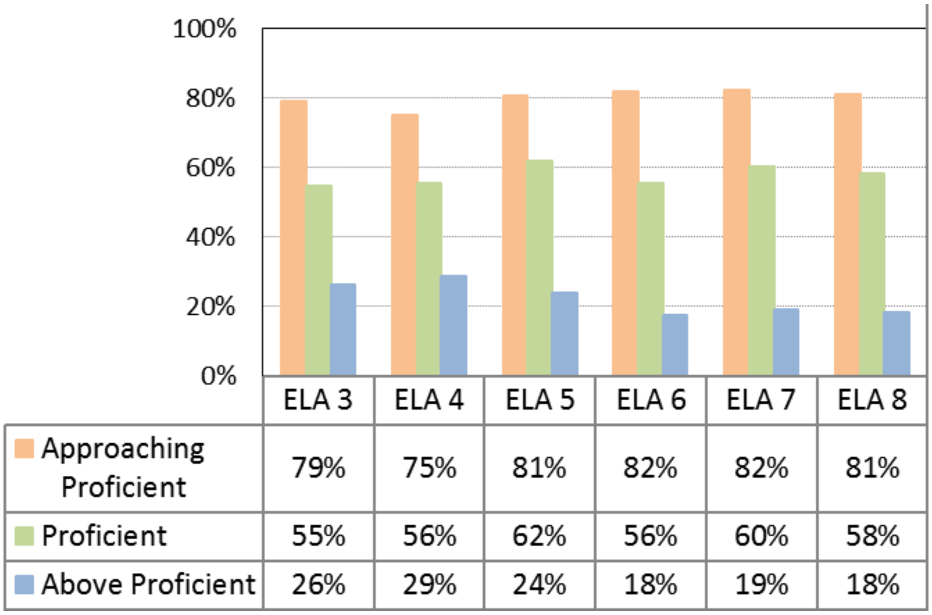


Table 3 indicates the percentage of students that we estimate reached each of the performance levels in ELA in 2018. Figure 3 represents those values graphically.

Table 3: Students Reaching or Exceeding Each Performance Level, Spring 2018 ELA

Grade	**Percentage at or Above**		
	Approaching Proficient	***Proficient***	***Above Proficient***
3	79	55	26
4	75	56	29
5	81	62	24
6	82	56	18
7	82	60	19
8	81	58	18

Figure 3. Students Reaching or Exceeding Each Performance Level, Spring 2018 ELA



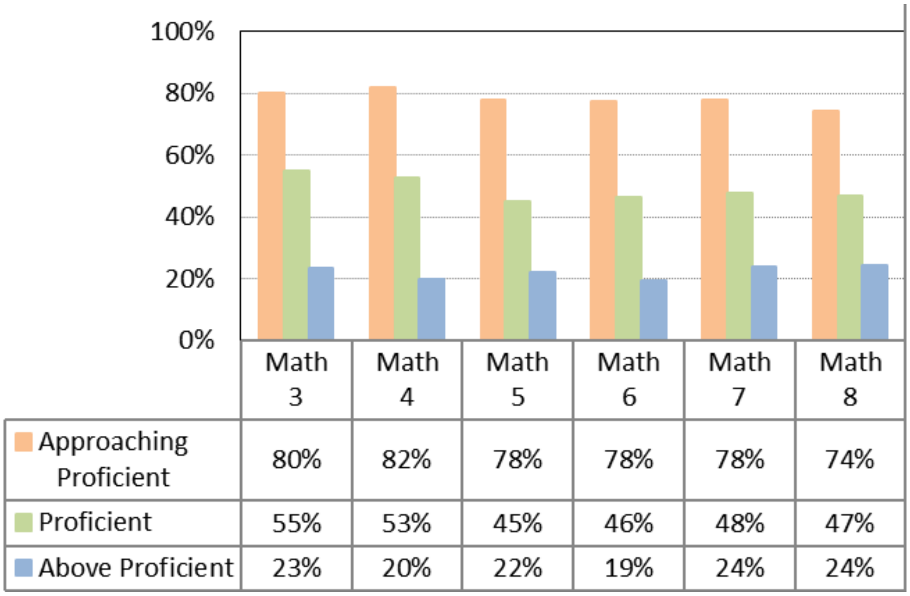
	ELA 3	ELA 4	ELA 5	ELA 6	ELA 7	ELA 8
🟧 Approaching Proficient	79%	75%	81%	82%	82%	81%
🟩 Proficient	55%	56%	62%	56%	60%	58%
🟦 Above Proficient	26%	29%	24%	18%	19%	18%

Table 4 indicates the percentage of students that we estimate reached each of the performance levels in mathematics in 2018. Figure 4 represents those values graphically.

Table 4: Students Reaching or Exceeding Each Performance Level, Spring 2018 Mathematics

Grade	**Percentage at or Above**		
	Approaching Proficient	***Proficient***	***Above Proficient***
3	80	55	23
4	82	53	20
5	78	45	22
6	78	46	19
7	78	48	24
8	74	47	24

Figure 4. Students Reaching or Exceeding Each Performance Level, Spring 2018 Mathematics



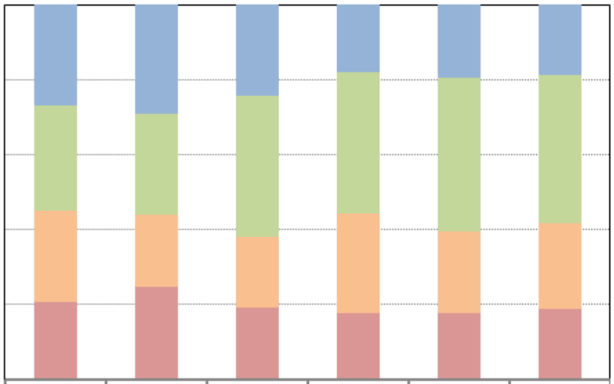
	Math 3	Math 4	Math 5	Math 6	Math 7	Math 8
🟧 Approaching Proficient	80%	82%	78%	78%	78%	74%
🟩 Proficient	55%	53%	45%	46%	48%	47%
🟦 Above Proficient	23%	20%	22%	19%	24%	24%

Table 5 and Figure 5 describe the percentage of students within each of the performance levels for ELA.

Table 5: Students Within Each Performance Level, Spring 2018 ELA

Grade	Percentage Within Each Level			
	Below Proficient	***Approaching Proficient***	***Proficient***	***Above Proficient***
3	21	24	28	26
4	25	19	27	29
5	19	19	38	24
6	18	26	38	18
7	18	22	41	19
8	19	23	40	18

Figure 5. Students Within Each Performance Level, Spring 2018 ELA



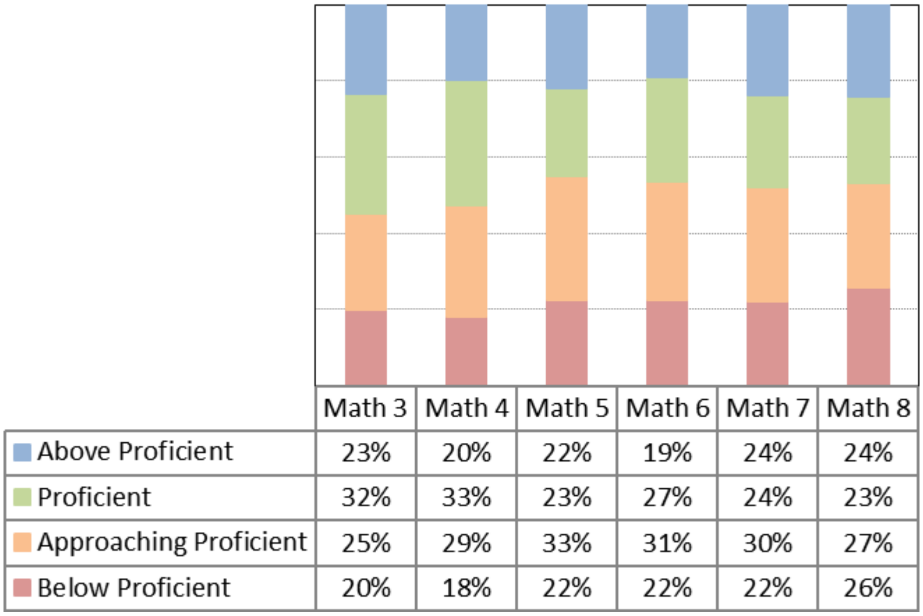
	ELA 3	ELA 4	ELA 5	ELA 6	ELA 7	ELA 8
🟦 Above Proficient	26%	29%	24%	18%	19%	18%
🟩 Proficient	28%	27%	38%	38%	41%	40%
🟧 Approaching Proficient	24%	19%	19%	26%	22%	23%
🟥 Below Proficient	21%	25%	19%	18%	18%	19%

Table 6 and Figure 6 describe the percentage of students within each of the performance levels for mathematics.

Table 6: Students Within Each Performance Level, Spring 2018 Mathematics

Grade	Percentage Within Each Level			
	Below Proficient	***Approaching Proficient***	***Proficient***	***Above Proficient***
3	20	25	32	23
4	18	29	33	20
5	22	33	23	22
6	22	31	27	19
7	22	30	24	24
8	26	27	23	24

Figure 6. Students Within Each Performance Level, Spring 2018 Mathematics



	Math 3	Math 4	Math 5	Math 6	Math 7	Math 8
🟦 Above Proficient	23%	20%	22%	19%	24%	24%
🟩 Proficient	32%	33%	23%	27%	24%	23%
🟧 Approaching Proficient	25%	29%	33%	31%	30%	27%
🟥 Below Proficient	20%	18%	22%	22%	22%	26%

2. Executive Summary of Science

AIR conducted a standard-setting workshop to recommend performance standards for New Hampshire’s Statewide Assessment System in Science at grades 5, 8, and 11. The workshop was conducted on July 25–26, 2018, at the Grappone Conference Center, 70 Constitution Ave., Concord, New Hampshire.

New Hampshire’s Statewide Assessments in Science are designed to measure the New Hampshire College and Career Ready Science Standards. Test items were developed by AIR in conjunction with a consortium of states working to implement three-dimensional Next Generation Science Standards (NGSS). Test items were developed to ensure that each student is administered a test matching all elements of New Hampshire's Science assessment blueprint, which was constructed to align to the New Hampshire College and Career Ready Science Standards.

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

Panelists reviewed Performance-Level Descriptors (PLD) describing the degree to which students have achieved the New Hampshire College and Career Ready Science Standards. PLDs were reviewed and revised in a separate workshop conducted prior to the standard-setting workshop. All standard-setting workshop table leaders participated in the PLD review workshop. Working through the ordered assertions for each cluster, panelists mapped each assertion into one of the four performance levels: *Below Proficient, Approaching Proficient, Proficient*, and *Above Proficient*. The panelists performed the assertion mapping in two rounds of standard setting during the two-day workshop. Panelists' mapping of the scoring assertions was used to identify the location of the three performance standards used to classify student achievement: *Approaching Proficient*, *Proficient*, and *Above Proficient*. Following Round 2, panelists engaged in a moderation session to review and modify recommended performance standards to facilitate the adoption of an articulated set of performance standards across grades and subject areas.

Thirty New Hampshire science educators were selected to serve as science standard-setting panelists. The panelists represented a group of experienced teachers and curriculum specialists, as well as school administrators and other stakeholders. The composition of the panel ensured that a diverse range of perspectives contributed to the standard-setting process. The panel was also representative in terms of gender, race/ethnicity, and region of the state.

2.1 Overall Structure of the Science Workshops

The key features of the workshops included the following:

- The standard-setting procedure produced three performance standards (*Approaching Proficient, Proficient,* and *Above Proficient*) that will be used to classify student science performance on the New Hampshire Statewide Assessment System.

- Panelists recommended performance standards in two rounds.

- Impact data (percentage of students reaching each performance standard) was provided to panelists following the first round of recommending performance standards.
- The standard-setting workshops were conducted online using AIR's online standard-setting tool. A laptop computer was provided to each panelist at the workshops.
- Following Round 2, panelists engaged in a moderation session to review and modify recommended performance standards to achieve an articulated system of standards across grades and subject areas.

2.2 Results of the Science Standard-Setting Workshops

Table 7 displays the performance level cut scores recommended by the standard-setting panelists.

Table 7: Performance Level Cut Scores Recommended for Science

Grade	*Approaching Proficient*	*Proficient*	*Above Proficient*
5	544	554	566
8	845	854	870
11	1146	1153	1176

Table 8 indicates the percentage of students that we estimate will reach each of the performance levels in 2018. Figure 7 represents those values graphically.

Table 8: Students Reaching or Exceeding Each Performance Level, Spring 2018 Science

Grade	Percentage at or Above		
	Approaching Proficient	*Proficient*	*Above Proficient*
5	70	44	14
8	68	44	9
11	60	40	5

Figure 7. Students Reaching or Exceeding Each Performance Level, Spring 2018 Science

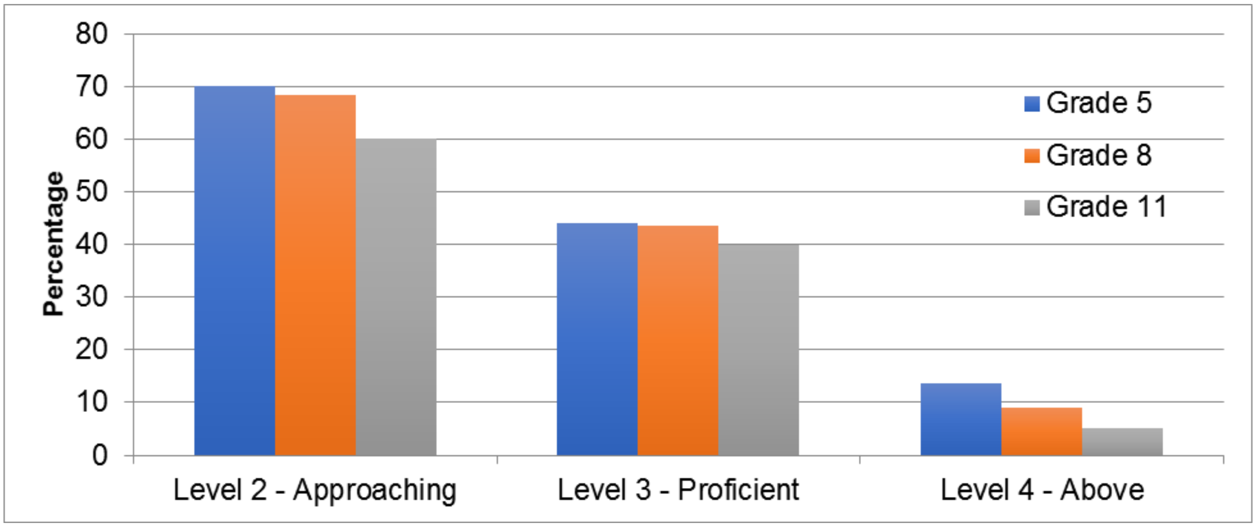
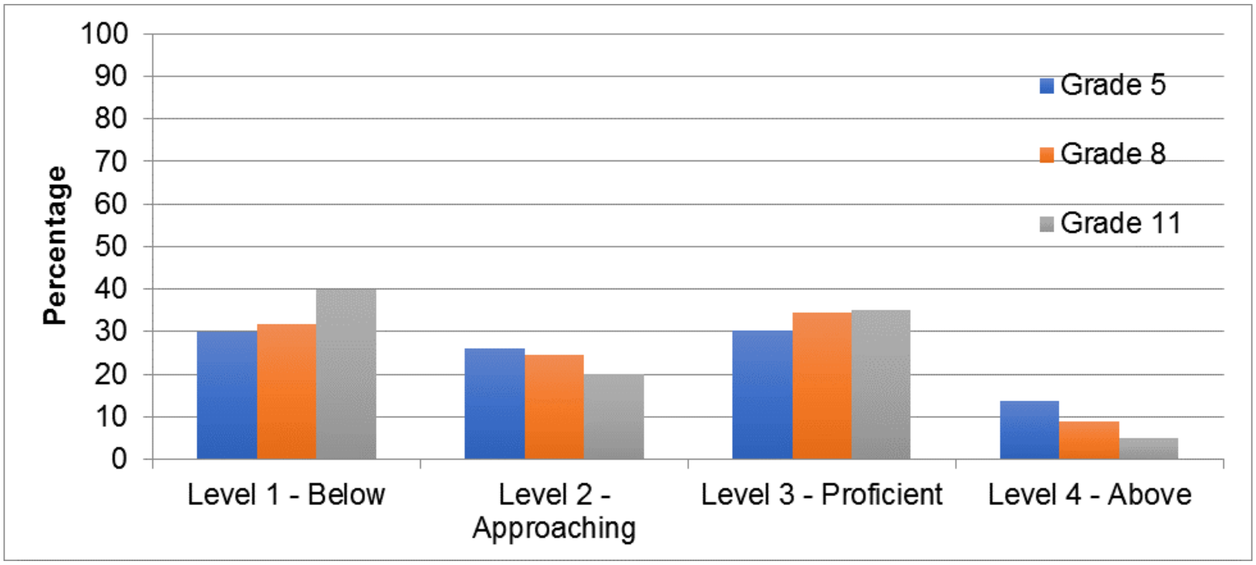


Table 9 indicates the percentage of students classified within each of the performance levels in 2018. The values are displayed graphically in Figure 8.

Table 9: Students Within Each Performance Level, Spring 2018 Science

Grade	Percentage Within Each Level			
	Below Proficient	***Approaching Proficient***	***Proficient***	***Above Proficient***
5	30	26	30	14
8	32	24	35	9
11	40	20	35	5

Figure 8. Students Within Each Performance Level, Spring 2018 Science



3. Introduction

In 2017–2018, New Hampshire transitioned from the Smarter Balanced Assessment Consortium (SBAC) in ELA and mathematics and the New England Common Assessment Program (NECAP) in science to the NH SAS. The NH SAS was designed to measure the New Hampshire College and Career Ready Standards (NH CCRS) for ELA, mathematics, and science. The tests measure academic progress for students in grades 3–8 for mathematics and ELA and in grades 5, 8, and 11 for science. The New Hampshire Department of Education (NHDOE) provides a summary of the new tests at https://www.education.nh.gov/instruction/assessment/index.htm.

New tests require new performance standards to link performance on the test to the content standards. The NHDOE contracted with the American Institutes for Research (AIR) to establish cut scores for the grades 3–8 ELA and mathematics tests, and grades 5, 8, and 11 science tests.

To fulfill this responsibility, AIR implemented a 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.

The purpose of this report is to document the standard-setting process and resulting performance standard recommendations.

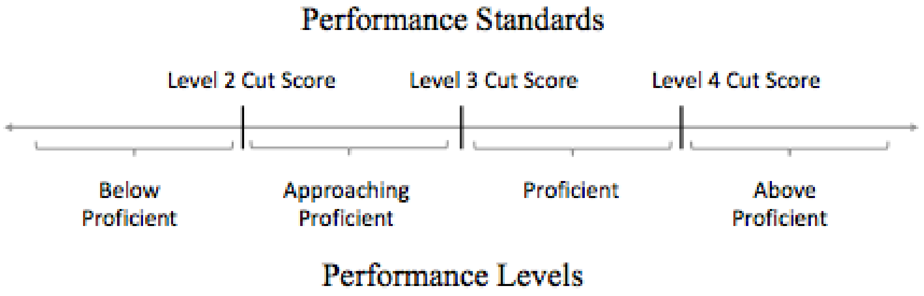
4. Standard Setting

Eighty-six educators from New Hampshire (8-10 for each grade-level test) convened in Concord, New Hampshire, from July 25–27, 2018, to complete two rounds of standard setting to recommend three performance standards for the NH SAS tests in ELA, mathematics, and science.[1]

Standard setting is the process used to define achievement on the NH SAS. Performance levels are defined by performance standards that specify how much of the content standards students must know and be able to do in order to meet each performance level. As shown in Figure 9, three performance standards are sufficient to define four performance levels.

[1] AIR has implemented two rounds of standard setting as best practice for more than 15 years. The approach has been approved by state Technical Advisory Committees and federal accountability peer reviewers. Panels typically converge in Round 2 with only modest improvements in Round 3, and the moderation session provides the opportunity for any necessary articulation that has not occurred after Round 2. In addition to lessening panelists' burden by not having them repeat a cognitively demanding task for a third time, using two rounds also introduces significant cost efficiency by reducing the number of days needed for standard setting. Panelists who complete two rounds report levels of confidence in the outcomes which are similar to the confidence expressed by panelists who participate in three rounds. Psychometric evaluation of the reliability and variability in results from two and three rounds are generally consistent. AIR has used two rounds for standard setting in more than 12 states and 20 NCLB-approved assessments.

Figure 9. Three Performance Standards Defining New Hampshire's Four Performance Levels



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

4.1 Methods

4.1.1 The Bookmark Method for ELA and Mathematics

The student-centered Bookmark method of standard setting is well suited to support the establishment of cut scores on high-stakes tests. It is appropriate for tests, like the NH SAS ELA and mathematics tests, that are scored using item response theory (IRT) and that use mixed-type items. This approach is appropriate for these types of tests and simplifies the decision process for panelists by allowing them to make the same judgment task for all items, regardless of item type. Because the Bookmark method directly relies on judgments made by experts, panelists and stakeholders report high confidence in the outcomes. It has proven to be technically sound in litigation, and more than 30 states have selected and implemented it, making it the most frequently used method of setting performance standards on high-stakes state accountability assessments (Karantonis & Sireci, 2006; Mitzel, Lewis, Patz, & Green, 2001; Perie, 2005). For these reasons, the NHDOE chose to apply the Bookmark standard setting method to establish new performance standards.

The Bookmark method derives its name from the primary task required of panelists—the placement of a bookmark in an ordered-item booklet (OIB) to represent a cut score recommendation. Over the course of two rounds of judgments, panelists recommended content-based cut scores using information from the policy descriptors, target student descriptors, test content viewed in the OIBs, panelist discussions, and impact and benchmark data.

Standard-setting judgements are made independently with the goal of convergence over two rounds of rating, rather than consensus.

4.1.2 The Assertion Mapping Procedure for Science

A different approach is necessary for tests based on the Next Generation Science Standards (NGSS), due to the structure of the content standards, and subsequently, the structure of test items assessing the standards. Tests based on the NGSS, such as the NH SAS science test, adopt a three-dimensional conceptualization of science understanding, including science and engineering practices, crosscutting concepts, and disciplinary core ideas. Accordingly, the new science assessments are composed 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 and item cluster, 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, therefore not supporting the assertion that the student can interpret relationships expressed graphically.

Although other assessments, especially ELA, comprise items probing a common stimulus, the degree of interdependence among such items is limited and student performance on such items can be evaluated independently of student performance on other items within the stimulus set. This is not the case with the new science item types, 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 among 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 one another. The interdependence of student interactions within items has consequences both for scoring and recommending performance standards.

The effects of item clusters can be accounted for by including additional dimensions in the IRT model to account for cluster specific variation. These dimensions are considered to be 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 used to recommend performance standards. However, the BoW method is manageable only with small numbers of performance tasks and quickly becomes onerous when the number of clusters approaches 10 or more.

To address these challenges, AIR psychometricians designed a new method for setting standards on new tests of the NGSS, including the NH SAS Science test. New Hampshire is breaking new ground by being the first to apply this method to set performance standards on the science test.

The test-centered Assertion Mapping Procedure (AMP) is an adaptation of the Item-Descriptor (ID) Matching procedure (Ferrara and Lewis, 2012) that preserves the integrity of the item clusters while also taking advantage of ordered-item-based procedures such as the Bookmarking procedure used for the ELA and mathematics tests.

The main distinction between AMP and existing ordered-item procedures (e.g., Mitzel, Lewis, Patz, and Green, 2001) is that the panelists evaluate scoring assertions rather than individual items. Scoring assertions are not test items, but inferences that are (or are not) supported by students' responses in one or more interactions within an item cluster. Because item clusters represent multiple, interdependent interactions through which students engage in scientific phenomena, scoring assertions cannot be meaningfully evaluated independently of the cluster from which they are derived. Therefore, the scoring assertions from the same item or item cluster are always presented together. Within each item or item cluster, 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 or item cluster. Within each chapter, the (ordered) pages represent scoring assertions. Like in ID matching, panelists are asked to map each scoring assertion to the most apt performance-level descriptor (PLD) during two rounds of standard setting. Like the Bookmark method, assertion mappings are made independently with the goal of convergence over two rounds of rating, rather than consensus.

4.2 Workshop Structure

Two large meeting rooms served as training rooms for all participants, with one room for each standard setting method. Five breakout rooms served as workspaces for the subject- and grade-level panels. As shown in Figure 10, the ELA and mathematics rooms contained six tables each, and the three science rooms contained two tables each, allowing for two tables per test and grade or grade band.

Figure 10. Room Structure

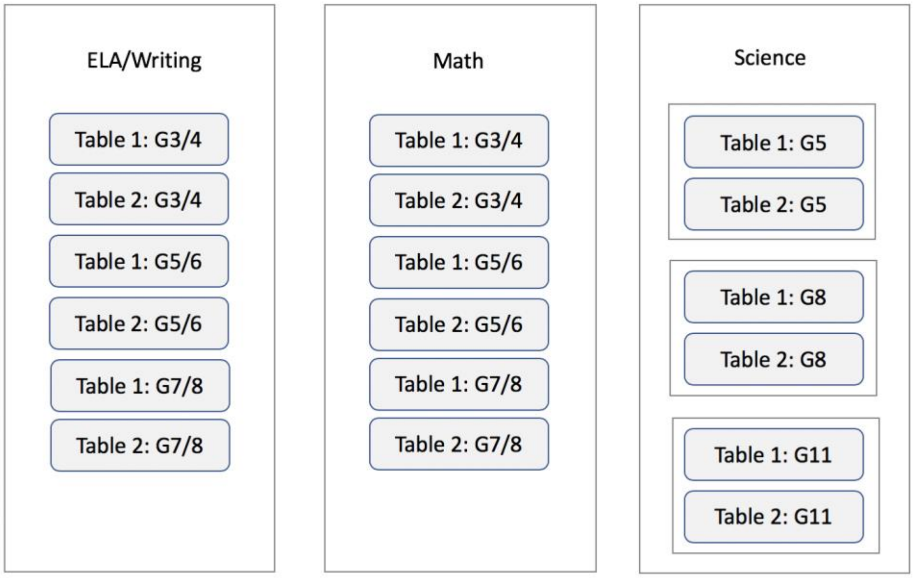


Table 10 summarizes the composition of the tables and the number of facilitators and panelists assigned to each. The 86 standard-setting participants included table leaders and panelists who taught in the content area for which standards were being set.

Table 10: Table Assignments

Panel	Room	Grade	Tables	Table Leaders	Panelists	Facilitator	Facilitator Assistant(s)
ELA	1	3 and 4	2	2	8	Brett Craycraft	Ann Harshbarger Allison Hahn
		5 and 6	2	2	7		
		7 and 8	2	2	8		
Mathematics	2	3 and 4	2	2	7	Jim McCann	Richard Yang Alysa Giustino
		5 and 6	2	2	8		
		7 and 8	2	2	8		
Science	3	5	2	2	8	Margaret McMahon	Ashley Gilliam
	4	8	2	2	6	Kevin Chandler	S. Alexa McDorman
	5	11	2	2	8	Joshua Smith	Matthew Andersen
Total Participants	N/A	N/A	N/A	18	68	N/A	N/A

4.3 Participants and Roles

4.3.1 New Hampshire Department of Education Staff

Julie Couch, Assessment Administrator for NHDOE, oversaw the workshop, provided overall policy context, and answered any policy questions that arose.

4.3.2 AIR Staff

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

- Tom Glorfield, Senior Program Manager; Evelyn Chester, Senior Project Coordinator
  - managed the process throughout the meeting.
- Dr. Gary Phillips, AIR Vice President and Institute Fellow
  - facilitated and oversaw the Bookmark method process and tasks. He provided training to all ELA and mathematics participants, including the facilitators, table leaders, and all participants, and he supervised the psychometric analyses conducted during and after the workshop.
- Dr. Stephan Ahadi, Managing Director of Psychometrics
  - facilitated and oversaw the AMP process and tasks. He provided training to all science participants, including the facilitators, table leaders, and all participants, and he supervised the psychometric analyses conducted during and after the workshop.
- Dr. Ahmet Turhan, Lead Psychometrician
  - provided psychometric analysis for ELA and mathematics.
- Dr. Frank Rijmen, Director of Psychometrics
  - provided psychometric analysis for AMP (science).

- Nik Kalich and Patrick Kozak, Psychometric Support Managers; Alesha Ballman, Psychometric Support Assistant
    - oversaw analytics technology.
- Drew Azar and Jim Unger, System Support Agents
    - set up, tested, and troubleshot technology during the workshop.

AIR provided a facilitator and an assistant facilitator to guide the process in each room. Facilitators were content experts experienced in leading standard-setting processes and could answer any questions about the process or about the items or what the items are intended to measure. They also monitored time and motivated panelists to complete tasks within the scheduled time. They included:

- Brett Craycraft, Senior Test Developer, serving as the ELA room facilitator;
- Ann Harshbarger, Senior Test Developer, and Allison Hahn, Test Developer, serving as ELA facilitator assistants;
- Jim McCann, Senior Test Developer, serving as the mathematics room facilitator;
- Richard Yang, Item Writer, and Alysa Giustino, Test Developer, serving as mathematics facilitator assistants;
- Margaret McMahon, Vice President of Content & Test Development, Kevin Chandler, Test Development Manager, and Joshua Smith, Director of Test Development, serving as the science room facilitators; and
- Ashley Gilliam, Test Developer, and S. Alexa McDorman and Matthew Andersen, Psychometric Support Assistants, serving as science facilitator assistants.

Prior to the workshop, it was necessary to ensure that each facilitator was extensively knowledgeable of the constructs, processes, and technologies used in standard setting. Thorough training is essential to standardize the training and procedures across the grade/subject committees. All facilitators and assistant facilitators participated in a full-day process training and a technology training prior to each workshop.

4.3.3 Table Leaders

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

As with room facilitators, it was necessary to ensure that each table leader was knowledgeable of the constructs, processes, and technologies used in standard setting and able to adhere to a standardized process across the grade/subject committees.

Table leaders trained as a group early in the morning of the first day. Training consisted of an overview of their responsibilities and some process guidance. Table leaders provided the following throughout the workshop:

- Lead table discussions;
- Helped panelists see the ‘big picture’;
- Monitored the security of materials;

- Monitored panelists' understanding and reported issues or misunderstandings to room facilitators; and
- Maintained a supportive atmosphere of professionalism and respect.

4.3.4 Educator Participants

To set the bookmarks, NHDOE recruited a diverse set of participants from across the state, ensuring that a range of perspectives contributed to the standard-setting process and product. In recruiting panelists, NHDOE targeted the recruitment of participants to be representative of the gender and geographic representation of the teacher population found in New Hampshire. Table 11 summarizes characteristics of the panels.

Table 11: Panelist Characteristics

Characteristic	Panelists by Subject Area (%)		
	ELA	*Mathematics*	*Science*
Male	4	38	50
Non-White	4	7	7
District Size			
Large	28	24	14
Medium	24	38	46
Small	48	24	36
Unknown	0	14	4
District Urbanicity			
Urban	14	14	11
Suburban	34	38	43
Rural	52	34	39
Unknown	0	14	7
Stakeholder Group			
Educator	72	72	71
Administrator	14	7	18
Coach	0	3	0
Specialist	17	14	0
Other	10	4	11

For the results of any judgment-based method to be valid, the judgments must be made by individuals who are qualified to make them. Participants in the New Hampshire standard-setting workshop were highly qualified. They brought a variety of expertise in instruction, curriculum, assessment, and special student populations. Most had professional experience in addition to teaching, and most had taught for 11 years or more. Table 12 summarizes the qualifications of the panels.

Table 12: Panelist Qualifications

Qualification	Panelists by Subject Area (%)		
	ELA	*Mathematics*	*Science*
Years Teaching Experience			
5 Years or Less	3%	20%	18%
6–10 Years	21%	24%	29%
11 Years or More	72%	55%	54%
Unknown	3%	0%	0%

Qualification	Panelists by Subject Area (%)		
	ELA	*Mathematics*	*Science*
Years Professional Experience			
5 Years or Less	66%	62%	71%
6–10 Years	24%	14%	4%
11 Years or More	10%	24%	21%
Unknown	0%	0%	4%
Highest Degree Earned			
Bachelor's	17%	14%	18%
Master's	62%	59%	68%
Doctorate	10%	7%	4%
Other	10%	21%	11%
Experience with ELs	34%	45%	68%
Experience with SWDs	66%	76%	89%
Experience with Low-SES Students	59%	69%	79%

Notes. Abbreviation Key: English Learners (ELs), Students with disabilities (SWDs), Socio-economic Status (SES).

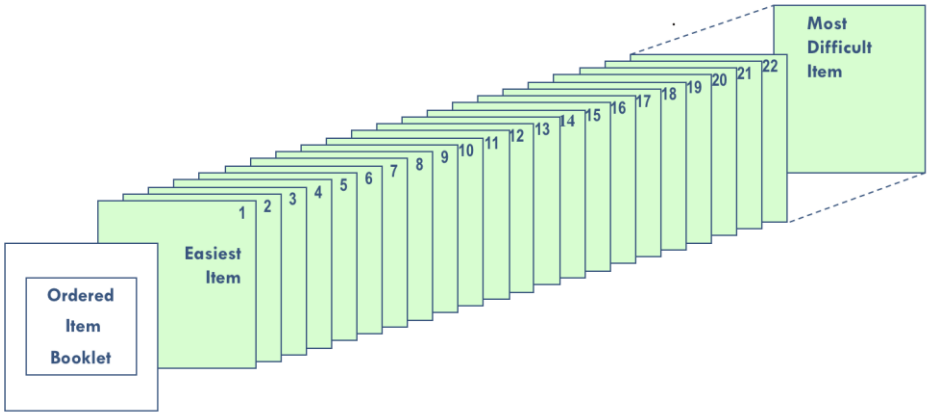
Appendix A describes the characteristics of individual panelists.

4.4 MATERIALS

4.4.1 Ordered-Item Booklets for ELA and Mathematics

The Bookmark method utilizes OIBs as the key tool for setting standards. OIBs contain sets of test items ordered by difficulty by grade and subject. Each page of the online OIB presents a single item, with the easier items located in the front of the OIB and the more difficult items in the back of the OIB. Item difficulty, and thus item ordering, is determined by analyses of actual student performance on the items. Mathematics and ELA panelists use the OIBs to place bookmarks that identify sets of items students meeting each standard should be able to answer correctly. Each page of the OIB corresponds to a cut score; thus, when panelists place their "bookmarks" for each performance level, they are, in fact, selecting the performance standard for that performance level.

Figure 11. Ordered-Item Booklet



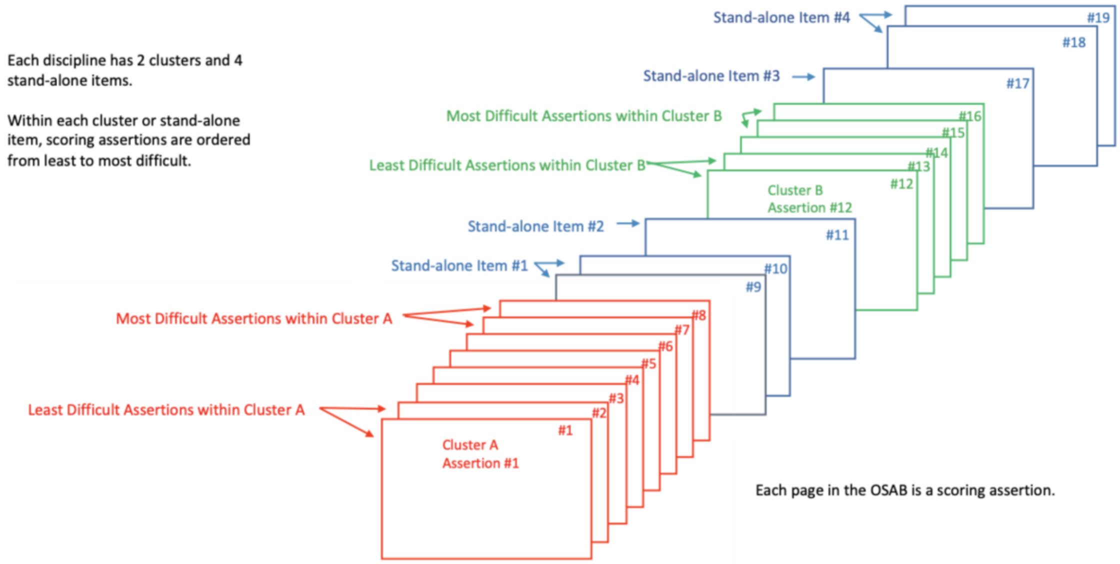
Some multi-select items provide multiple score points on a test; these items are presented on multiple pages in the OIB, one page for each possible score point. As such, the number of pages in the OIB is equal to the number of score points in the OIB, not the number of items.

For New Hampshire, the OIBs ranged from 85 to 110 pages.

4.4.2 Ordered Scoring Assertion Booklets for Science

Like the Bookmark method, the AMP 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 12 describes the structure of the OSAB.

Figure 12. Ordered Scoring Assertion Booklet


Each discipline has 2 clusters and 4 stand-alone items.
Within each cluster or stand-alone item, scoring assertions are ordered from least to most difficult.
Stand-alone Item #4
#19
#18
Stand-alone Item #3
#17
Most Difficult Assertions within Cluster B
#16
#15
#14
#13
Least Difficult Assertions within Cluster B
#12
Cluster B Assertion #12
Stand-alone Item #2
#11
Stand-alone Item #1
#10
#9
Most Difficult Assertions within Cluster A
#8
#7
#6
#5
#4
#3
#2
Least Difficult Assertions within Cluster A
#1
Cluster A Assertion #1
Each page in the OSAB is a scoring assertion.

The items and item clusters are presented by discipline. For the operational test, the order of the disciplines was randomized over students. For the OSABs, Earth and Space Sciences items were presented first, then Life Sciences items, and then Physical Sciences items. Two item clusters and four stand-alone items represent each discipline. Within a discipline, clusters and stand-alone items were presented intermixed, in the same way that clusters and stand-alone items would be selected at random by the algorithm that was used to linearly assemble operational tests. Within each item or item cluster, 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 items in Figure 12 represents the order of presentation to the panelists, but items were not ordered by overall item difficulty.

Not all clusters have assertions that will map onto to all performance levels. For example, a cluster may have assertions that map onto *Below Proficient*, *Approaching Proficient*, and *Proficient*, but not *Above Proficient*. Clusters may have as few as four or as many as 20 assertions. Each assertion is worth one score-point.

The grade 5 OSAB contained 67 assertions, the grade 8 OSAB contained 77 assertions, and the grade 11 OSAB contained 81 assertions. Each OSAB was composed of 6 item clusters and 12 stand-alone items.

4.4.3 New Hampshire College and Career Readiness Standards

The NH SAS assesses the learning objectives described by the NH CCRS. The ELA and mathematics standards are based on the Common Core State Standards (CCSS, adopted in 2010) and science is based upon the Next Generation Science Standards (NGSS, adopted in 2016).

The College and Career Readiness Standards are available at https://www.education.nh.gov/instruction/curriculum/index.htm.

4.4.4 Performance-Level Descriptors

With the adoption of the new standards and the development of new statewide assessments to assess achievement of those standards, NHDOE must adopt a similar system of performance standards to determine if students have met the learning goals defined by the new standards in ELA, mathematics, and science.

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

PLDs link the standards to the performance standards. There are four types of PLDs:

1. Policy PLDs: These are brief descriptions of each performance level that do not vary across grade or content area.
2. Range PLDs: Provided to panelists to review and endorse during the workshop, these detailed grade- and content area-specific descriptions communicate exactly what students performing at each level know and can do.
3. Target PLDs: Typically created during and used for standard setting only, these describe what a student “just barely” scoring into each performance level knows and can do.
4. Reporting PLDs: These are much abbreviated PLDs (typically 350 or fewer characters) created following state approval of the performance standards used to describe student performance on score reports.

New Hampshire uses four performance levels to describe student performance: *Below Proficient, Approaching Proficient*, *Proficient*, and *Above Proficient*.

Mathematics and ELA PLD Development

When developing PLDs, AIR started with the Common Core State Standards in mathematics and English language arts/literacy (ELA/L), high-level policy PLDs were written first to clearly define what it means at the achievement level to be college and career ready. Then, AIR adapted these policy PLDs to the other performance levels, describing what a student's achievement would look like at each point on the continuum. With policy PLDs in place, range PLDs were written for each assessed standard. AIR started with the language of the standard as the performance level, adapting it to show how the performance of a student would differ at the *Above Proficient*, *Approaching Proficient,* and *Below Proficient* levels. As AIR moved toward the *Above Proficient* level in ELA, for example, language like "complex inference" was used, rather than simply "inference," to indicate that higher-performing students would be expected to read and draw conclusions from more complex texts. These range PLDs offer observable evidence of student achievement within each standard, and they change and become more (or less) sophisticated across performance levels.

Once the range PLDs for a single grade were completed, the process of writing PLDs for the other grades began, keeping in mind the ways in which language was adapted for the grades above and below. When all the PLDs were drafted, senior reviewers at AIR reviewed the PLDs across all grades to ensure a clear vertical articulation from grade to grade.

With the PLDs in place, New Hampshire state standards were reviewed to identify any standards that differed from the Common Core State Standards. In cases where the standards differed, a unique range PLD was written to represent that standard.

Science PLD Development

The Washington State Office of Superintendent of Public Instruction (OSPI) drafted initial range PLDs based on the Next Generation Science Standards (NGSS). AIR, state Department of Education staff, and educators from the 10 states using AIR's science assessment convened in May of 2018 to review and refine the draft PLDs.[2] The panels created policy-level descriptors and reviewed and identified refinements to the range PLDs to describe observable evidence for what student achievement looks like in in science at each performance level and grade. AIR and one of the authors of the NGSS reviewed and applied the recommendations to the PLDs. They ensured consistency, coherence and articulation across grades and levels.

Panelist PLD Review

The NHDOE then reviewed the PLDs to ensure that the language accurately represented the goals and policies of the state. AIR worked with them to make revisions where necessary.

The day before the standard setting workshop, the group of New Hampshire educators selected to be table leaders, who were intimately familiar with students and the subject matter, convened to review, revise, and approve the policy and range PLDs. More information on this is available in Sections 4.6.4 and 4.6.5.

[2] These states included Oregon, Hawaii, New Hampshire, Oregon, Rhode Island, Utah, Vermont, West Virginia and Wyoming.

4.5 Workshop Technology

Panelists used AIR's online application for standard setting. Each panelist used an AIR laptop or Chromebook on which they took the test, reviewed items and ancillary materials, and placed bookmarks.

The Bookmarking panelists in ELA and mathematics could review each item, examine the content alignment and score points for each item, and evaluate the impact that proposed cuts will have on students. Panelists also saw their own bookmarks, their table's bookmarks, the other tables' bookmarks, and the overall bookmarks for both tables.

AMP panelists in science could review the items, item clusters and scoring assertions, examine the content alignment of each assertion, assign assertions to performance levels, and review impact and benchmark data. Additionally, they had access to a difficulty visualizer, a graphic representation of the difficulty of each assertion relative to the other assertions in the OSAB. Panelists also reviewed their own assertion placement, their table's placement, the other tables' placement, and the overall placement for both tables.

All panelists were able to add notes and comments on the items, item clusters, or assertions as they reviewed them and examine reference and benchmark data onscreen following each round.

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

4.6 Events

The standard-setting workshop occurred over three days for ELA and mathematics and two days for science. Because each ELA and mathematics table set standards for two grades, an anchor grade (the even numbered grades, 4, 6, and 8) and an adjacent grade (the odd numbered grades 3, 5, and 7), they needed more time than did science participants, who set standards for a single grade. Table 13 summarizes each day's events, and this section describes each event listed in greater detail.

Appendix E provides the full agendas for ELA, mathematics, and science.

Table 13: Standard-Setting Agenda Summary

ELA/Mathematics	Science
Day 1: Wednesday, July 25	
• Table leader training • Orientation and introductions • Take the test • Content standards review • PLD review • Write "just barely" target PLDs • OIB review (anchor grade)	• Table leader training • Orientation and introductions • Take the test • Content standards review • PLD review • Item cluster review • OSAB review

Day 2: Thursday, July 26	
• Bookmark placement training • Bookmark placement practice • Standard setting readiness evaluation • Round 1 anchor-grade bookmark placement • Round 1 feedback, impact data, and benchmark data review and discussion • Round 2 anchor-grade bookmark placement • Round 2 feedback, impact data, and benchmark data review and discuss	• Assertion mapping training • Assertion mapping practice • Standard setting readiness evaluation • Round 1 assertion mapping • Round 1 feedback, impact data, and benchmark data review and discussion • Round 2 assertion mapping • Round 2 feedback, impact data, and benchmark data review and discussion • Standard-setting workshop evaluations • Final moderation • Dismissal
Day 3: Friday, July 27	
• OIB review (adjacent grade) • Round 1 adjacent-grade bookmark placement • Round 1 feedback, impact data, and benchmark data review and discussion • Round 2 adjacent-grade bookmark placement • Round 2 feedback, impact data, and benchmark data review and discussion • Standard-setting workshop evaluations • Final moderation	

4.6.1 Orientation

Julie Couch, the Assessment Administrator for NHDOE, along with Frank Edelblut from the Commission of Education and Gary Phillips from AIR, welcomed panelists to the workshop.

The ELA and mathematics panelists then split from the science panelists for separate, large-group orientations. Dr. Phillips led the ELA and mathematics orientation, and Dr. Ahadi led the science orientation. Both described the purpose and objectives of the meeting, explained the process to be implemented to meet those objectives, and outlined the events that would happen each day. They outlined the responsibilities of the three groups at the workshop (panelists, AIR staff, and NHDOE personnel) and explained that the panelists were selected because they were experts, as well as how the process to be implemented was designed to elicit and apply their expertise to recommend new cut scores. Finally, they described how standard setting works and what would happen once the panelists had finalized their recommendations.

4.6.2 Confidentiality and Security

Confidentiality and security were addressed once during orientation and again by the facilitators in each room. Standard setting uses live test items from the operational NH SAS tests, so confidentiality is required to maintain their security. Participants were NOT allowed to do the following during and after the workshop:

- Discuss the test items outside of the meeting
- Remove any secure materials from the room on breaks or at the end of the day
- Discuss judgments or cut scores (theirs 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 to the workshop

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

4.6.3 Take the Test

Following the large-group training, panelists broke out into their assigned rooms, where they took a sample of the test that students took in 2018, in the subject area and grade for which they would be setting performance standards. They took the tests online via the same test engine used to deliver operational tests to students, and the testing environment closely matched that of students when they took the test. While testing, panelists could not discuss the items, hold any conversations, or access their phones.

Taking the same test as students take provides the opportunity to interact and become familiar with the test items and the look and feel of the student testing experience.

4.6.4 Review Content Standards and PLDs

After completing the test, panelists completed a thorough review of the standards and PLDs for their grade and subject area. They identified key words describing the skills necessary for performance at each level and discussed the skills and knowledge that differentiated performance in each of the four levels.

Reviewing the content standards ensured that participants understood what students in New Hampshire are expected to know and be able to do, and reviewing the performance standards ensured that they understood how much knowledge and skill students are expected to demonstrate at each level of achievement.

4.6.5 Write Target PLDs

After reviewing and discussing the PLDs, the ELA and mathematics panelists worked in their table groups to draft target PLDs that described the skills that students "just barely" in one performance level have that students just below the performance level don't. Target PLDs describe students who are not typical of students at a performance level, though, at "just barely," they do reach the standard. One subject/grade-level table drafted the target PLDs for the *Approaching Proficient* level, the other table drafted the target PLDs for the *Proficient level*, and then the two tables discussed and created the *Above Proficient* target PLDs together.

Because science performance standards are set by matching student performances to performance levels rather than by identifying the thresholds that demarcate each performance level, science panelists did not need or write target PLDs.

4.6.6 Ordered-Item Booklet/Ordered Scoring Assertion Booklet Review

After completing the target PLDs, ELA and mathematics panelists independently reviewed all items in the online OIB. For each item, they could take notes on the items that would help them as they placed the Round 1 and Round 2 bookmarks. Suggested review steps included noting what students need to know and be able to do to answer each item correctly, identifying what made each item more difficult than the one before and documenting how the item related to the performance levels.

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

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

4.6.7 Training

The objective of both methods of standard setting is aspirational: to identify what all students should know and be able to do, not what a student or group of students actually know and can do. The sections below describe how each method addressed this objective.

Bookmark Training for ELA and Mathematics

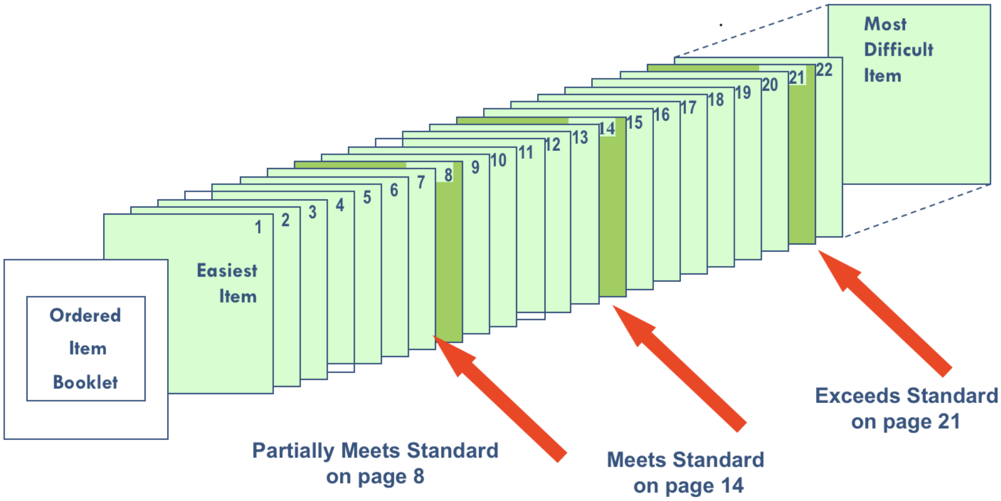
The ELA and mathematics panelists considered the target PLDs that describe students “just barely” meeting each performance level as they reviewed the OIB and applied a two-thirds response probability rule (RP67) when placing bookmarks. This rule required panelists to identify the page in the OIB at which two-thirds of students who “just barely” meet the standard (those described by the target PLDs) should be able to get the item on that page correct.

The explanation of this rule provided to panelists was as follows:

“Of 100 students who are ‘just barely’ at the standard, what percentage would get this item correct?”

These “just barely” students are more likely to be able to correctly answer items at the beginning of the OIB and less likely to be able to correctly answer items towards the end of the OIB. As panelists work through the OIB, they will come across an item, or small group of items, for which they think about 67% of the “Just Barely Proficient” students (for example) would get the item correct. Items before that point in the OIB are items that more than 67% of the “Just Barely Proficient” students would answer correctly. Items beyond that point in the OIB are items that less than 67% of the “just barely” students would correctly answer. Panelists place their bookmark on the first page of the OIB for which they believe the “Just Barely Proficient” student would NOT have at least a 67% chance of answering correctly. Panelists repeated this process for the "Just Barely Approaching Proficient" student and the "Just Barely Above Proficient" student.

Figure 13. Example Bookmark Placement


Most Difficult Item
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Easiest Item
Ordered Item Booklet
Partially Meets Standard on page 8
Meets Standard on page 14
Exceeds Standard on page 21

The workshop leaders from AIR and NHDOE advised panelists that, though some items may seem out of order, the item order is determined by item difficulty calculated from actual student performance on the items and not determined by content or cognitive processes. The order of items in the OIB does not follow the sequence of instruction or the order of item presentation on the test.

To keep panelists focused on the standard-setting task, and not on item critique, panelists could refer item-related questions or comments to workshop facilitators and NHDOE staff for investigation. Bookmarks were not to be placed on any item that panelists disagreed with or felt might be incorrect or unfair. Finally, panelists were not to set standards for individual students they knew, or for students in their classrooms, but to set performance standards for all students across the state.

Assertion Mapping Training for Science

Because science panelists considered scoring assertions rather than items, they did not consider "just barely" students. Instead, they considered the interactions required by each assertion and the PLDs. Facilitators provided the following process to guide the mapping of assertions onto PLDs:

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

Like the items in the OIB, the scoring assertion order within each item was determined by actual student performance.

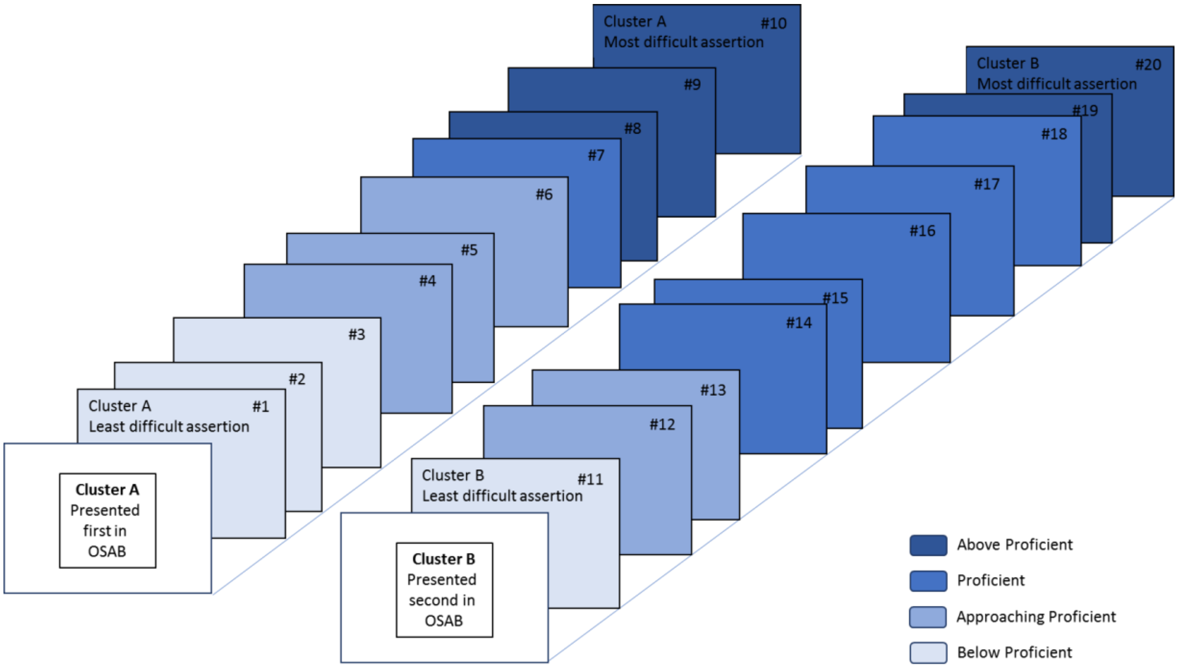
Panelists were to match each assertion to the performance level best supported by the assertion using the PLDs, an online difficulty visualizer (described in section 4.5), their notes from the

OSAB review, and their professional judgment. Figure 14 graphically describes the assertion mapping process.

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

Because the assertion mapping was done separately for each item, it was possible that there was no perfect ordering of the assigned levels of the assertions across all items as a function of assertion difficulty. It was allowed (and it occurred frequently) that an assertion of one item had a higher difficulty but lower assigned performance level than another assertion from a different item. For example, in Figure 14, the difficulty of Assertion #3 of cluster A (*Below Proficient*) has a higher difficulty than Assertion #13 of cluster B (*Approaching Proficient*). However, it was expected for the higher performance levels to be assigned more frequently with increasing assertion difficulty across items.

Figure 14. Example Assertion Mapping



Note. Figure describes scoring assertion mapping across two clusters, where assertions 1, 2, 3, and 11 are mapped onto the Below Proficient *level, assertions 4, 5, 6, 12, and 13 are mapped onto the* Approaching Proficient *level, assertions 7, 14, 15, 16, 17, and 18 are mapped onto the* Proficient *level, and assertions 8, 9, 10, 19, and 20 are mapped onto the* Above Proficient *level.*

4.6.8 Readiness Assessment

The quiz assesses panelists' understanding in multiple ways. ELA and mathematics panelists must be able to:

- indicate where students “just barely” meeting each of the standards fall on a diagram of how performance standards and levels work together;
- answer questions about relative item difficulty in a hypothetical OIB; and
- demonstrate understanding by correctly applying the RP67 rule to a hypothetical bookmark placement.

Science panelists must be able to:

- answer questions about the assertion mapping process;
- identify the most and least difficult assertions using the difficulty visualizer; and
- indicate on a diagram how performance standards differentiate proficiency levels.

Room facilitators review the quizzes and provide additional training for incorrect responses on the quiz. However, all the panelists answered all items correctly.

4.6.9 Practice Round

Following the readiness assessment, panelists practiced placing bookmarks in the OIB or mapping assertions in the OSAB. The purpose of the practice round was to ensure that panelists were comfortable with the technology and item types or assertions prior to setting any actual bookmarks or mapping any assertions. Panelists asked questions, and the room facilitators provided clarifications and further instructions until everyone had completed the practice round.

4.6.10 Readiness Assertion

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

- The knowledge and skills described by the PLDs and the skills and interactions that differentiate levels
- The structure, use, and importance of the OIB or OSAB
- The process to set cut scores or to map assertions from the OSAB onto the PLDs

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

- Understanding of the impact and feedback data
- Understanding of the Round 2 task
- Readiness to complete 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 placing bookmarks or mapping assertions in both rounds of the workshop.

4.6.11 Round 1

In Round 1, ELA and mathematics panelists set the bookmarks based on the difficulty and content of the items for *Proficient*, then for *Approaching Proficient*, and last for *Above Proficient*. Each panelist independently placed his or her own bookmarks. The median of the individual bookmarks across each table became the Round 1 cut score for that grade level.

Science panelists mapped assertions independently, using the PLDs, their notes from reviewing each assertion, and the difficulty visualizer to place each of the assertions into one of the four performance levels. AIR psychometricians then created cut scores from these science assertion mappings, one for each participant, table, and grade overall, and then also generated feedback, impact data, and reference data for the panelists to evaluate before Round 2. A proprietary algorithm utilized RP67 to minimize misclassifications to calculate cut scores based on the assertion mappings.

For science, 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 Proficient* and *Proficient* were used). Cut scores at the table and grade level were computed using the same method, but taking into account the assigned levels of all the raters at the table and grade, respectively. Applying these cut scores to the 2018 test data created impact 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.

Table 14, Table 15, and Table 16 present the bookmarks and associated impact and benchmark data for ELA, mathematics, and science, respectively. Panelists discussed the benchmark data, impact data, and articulation associated with their Round 1 recommendations. This information and resulting discussion informed their Round 2 judgments.

Table 14: Round 1 Results, ELA

Table	Median Round 1 Bookmark (Page #)			Impact Data (Percentage At or Above)			Benchmark Data (Comparable Performance Level Using 2017 Test)		
	App P	***P***	***Abv P***	***App P***	***P***	***Abv P***	***App P***	***P***	***Abv P***
G3	13	25	40	69	46	25	App P	P	Abv P
1	12	22	40	73	50	25	App P	P	Abv P
2	13	26	39	69	44	26	App P	P	Abv P
G4	16	33	52	70	45	13	App P	P	Abv P
1	15	28	48	71	53	19	App P	P	Abv P
2	16	33	54	70	45	10	App P	P	Abv P
G5	9	24	44	81	62	27	App P	App P	P
1	10	24	43	79	62	28	App P	App P	P
2	9	24	44	81	62	27	App P	App P	P

Table	Median Round 1 Bookmark (Page #)			Impact Data (Percentage At or Above)			Benchmark Data (Comparable Performance Level Using 2017 Test)		
	App P	***P***	***Abv P***	***App P***	***P***	***Abv P***	***App P***	***P***	***Abv P***
G6	12	24	45	73	50	15	App P	P	Abv P
1	15	26	44	69	47	15	App P	P	Abv P
2	10	23	46	79	54	14	App P	P	Abv P
G7	10	21	47	79	60	19	App P	P	Abv P
1	9	19	46	79	64	20	App P	App P	Abv P
2	12	23	48	77	57	17	App P	App P	Abv P
G8	12	30	52	81	55	18	App P	P	P
1	10	24	52	84	60	18	B P	App P	P
2	18	34	52	69	49	18	App P	App P	P

Note. The grade-level row summarizes the room data (the median across both tables). Benchmark data describes the percentage at or above each performance level using data from the 2017 Smarter Balanced test. Performance level abbreviation key: Below Proficient *(B P),* Approaching Proficient *(App P),* Proficient *(P),* Above Proficient *(Abv P).*

Table 15: Round 1 Results, Mathematics

Table	Median Round 1 Bookmark (Page #)			Impact Data (Percentage At or Above)			Benchmark Data (Comparable Performance Level Using 2017 Test)		
	App P	***P***	***Abv P***	***App P***	***P***	***Abv P***	***App P***	***P***	***Abv P***
G3	18	38	61	80	55	23	App P	P	Abv P
1	17	38	61	82	55	23	App P	P	Abv P
2	22	39	60	76	54	24	App P	P	P
G4	17	40	68	81	53	20	App P	App P	P
1	21	47	72	78	44	16	App P	P	Abv P
2	16	38	64	82	56	25	App P	App P	P
G5	17	42	62	78	47	22	B P	App P	Abv P
1	17	42	60	78	47	25	B P	App P	P
2	17	37	63	78	52	21	B P	App P	Abv P
G6	17	39	65	70	39	18	App P	P	Abv P
1	17	38	65	70	40	18	App P	P	Abv P
2	15	45	68	73	34	16	App P	P	Abv P
G7	16	33	58	68	46	24	App P	P	Abv P
1	15	34	57	70	45	25	App P	P	P
2	16	32	58	68	47	24	App P	P	Abv P

Table	Median Round 1 Bookmark (Page #)			Impact Data (Percentage At or Above)			Benchmark Data (Comparable Performance Level Using 2017 Test)		
	App P	***P***	***Abv P***	***App P***	***P***	***Abv P***	***App P***	***P***	***Abv P***
G8	16	38	60	74	55	24	B P	App P	P
1	25	43	62	64	44	23	App P	P	Abv P
2	15	27	56	74	63	27	B P	App P	P

Note. The grade-level row summarizes the room data (the median across both tables). Benchmark data describes the percentage at or above each performance level using data from the 2017 Smarter Balanced test. Performance level abbreviation key: Below Proficient *(B P),* Approaching Proficient *(App P),* Proficient *(P),* Above Proficient *(Abv P).*

Table 16: Round 1 Results, Science

Table	Cut Score (Scaled Score)			Impact Data (Percentage At or Above)			Benchmark Data (NAEP)		
	App P	***P***	***Abv P***	***App P***	***P***	***Abv P***	***App P***	***P***	***Abv P***
G5	540	554	566	79	44	14	87	50	1
1	543	554	573	73	44	6	87	50	1
2	537	554	566	83	44	14	87	50	1
G8	854	854	879	44	44	3	81	46	2
1	-	854	879	-	44	3	81	46	2
2	854	854	880	44	44	2	81	46	2
G11	1150	1173	1188	51	7	1	75	42	3
1	1150	-	1188	51	-	1	75	42	3
2	1131	1173	1188	95	7	1	75	42	3

Note. The grade-level row summarizes the room data (across both tables). Benchmark data describes the percentage at or above each performance level using data from the grade 8 National Assessment of Educational Progress (NAEP); grade 5 benchmark data is interpolated from grade 8 NAEP, and grade 11 is extrapolated from grade 8 NAEP. Performance level abbreviation key: Approaching Proficient *(App P),* Proficient *(P),* Above Proficient *(Abv P). Blank cells are the result of mappings that provided insufficient information to calculate a specific cut score (for example, panelists mapping no assertions into the lowest level would not provide sufficient information to calculate an* Approaching Proficient *cut score).*

4.6.12 Round 2

After Round 1, workshop facilitators provided panelists with additional instruction for completing Round 2. First, they described the goal of Round 2 as one of convergence, but not consensus, on a common performance standard. A second goal was articulation across grade levels. Panelists reviewed and discussed example sets of performance standards across grade, showing multiple ways of disarticulation until they understood why articulation was important and should be a consideration in making their Round 2 judgements. After explaining articulation, AIR psychometricians presented example scenarios that would maximize articulation and showed panelists what articulation looked like. Articulation was just another piece of information for panelists to consider, and like all feedback provided, they could consider it in their Round 2 judgments or not.

Workshop facilitators also provided panelists with additional information to inform the Round 2 judgments. This information included the judgments made by the other members of their table, the judgment from the other grade-level/subject table (for ELA and math), and the judgment overall, across both tables. For ELA, Round 1 judgments were summarized using the median cut score for each table and the room. For science, the Round 1 mappings were summarized across all judgments.

For science, workshop facilitators provided panelists with feedback data to inform the Round 2 judgments. Feedback included the cut scores corresponding to the assertion mappings for each panelist, each table, and for the room overall (across both tables). Feedback also included review of a variance monitor, part of AIR's online standard setting tool that color codes the variance of assertion classifications. For all assertions, the variance monitor shows the performance level that each panelist assigned the assertion to. The tool highlights assertions that panelists have assigned to different performance levels by the panelists. Room facilitators and panelists reviewed and discussed the assertions with the most variable mappings.

Panelists received impact data showing the percentage of students who would score at or above each performance level given the Round 1 judgments and benchmark data describing student performance on a measure other than the one on which they were setting performance standards. For ELA and mathematics, this was the percentage of students from the 2017 Smarter Balanced assessment scoring in each performance level; for science, this was the percentage of students scoring into each performance level on the 2015 NAEP. This provided external evidence of student performance for panelists to consider when placing Round 2 bookmarks or mapping the Round 2 assertions.

This information was to inform, but not to determine, their Round 2 decisions. Panelists discussed this information and the impact that the Round 1 cut scores may have on New Hampshire students before making Round 2 judgments. Table 17, Table 18, and Table 19 show the Round 2 recommendations and associated impact and benchmark data for ELA, mathematics, and science. Appendix F presents box-and-whisker plots for ELA/mathematics for both rounds.

Table 17: Round 2 Results, ELA

Table	**Median Round 2 Bookmark (Page #)**			**Impact Data (Percentage At or Above)**			**Benchmark Data (Comparable Performance Level Using 2017 Test)**		
	App P	***P***	***Abv P***	***App P***	***P***	***Abv P***	***App P***	***P***	***Abv P***
G3	11	22	39	74	50	26	App P	P	Abv P
1	12	22	39	73	50	26	App P	P	Abv P
2	10	22	38	76	50	28	App P	P	Abv P
G4	12	26	47	75	56	20	App P	P	Abv P
1	11	25	44	77	56	26	App P	P	Abv P
2	12	28	48	75	53	19	App P	P	Abv P

Table	**Median Round 2 Bookmark (Page #)**			**Impact Data (Percentage At or Above)**			**Benchmark Data (Comparable Performance Level Using 2017 Test)**		
	App P	***P***	***Abv P***	***App P***	***P***	***Abv P***	***App P***	***P***	***Abv P***
G5	9	24	45	81	62	24	App P	App P	Abv P
1	10	24	45	79	62	24	App P	App P	Abv P
2	9	24	45	81	62	24	App P	App P	Abv P
G6	9	22	43	81	56	18	App P	P	Abv P
1	9	23	43	81	54	18	App P	P	Abv P
2	7	22	43	84	56	18	App P	P	Abv P
G7	8	21	47	82	60	19	App P	P	Abv P
1	8	19	46	82	64	20	App P	App P	Abv P
2	9	21	47	79	60	19	App P	P	Abv P
G8	12	27	52	81	58	18	App P	P	P
1	10	24	52	84	60	18	B P	App P	P
2	12	27	52	81	58	18	App P	P	P

Note. The grade-level row summarizes the room data (the median across both tables). Benchmark data describes the percentage at or above each performance level using data from the 2017 Smarter Balanced test. Performance level abbreviation key: Below Proficient *(B P),* Approaching Proficient *(App P),* Proficient *(P),* Above Proficient *(Abv P).*

Table 18: Round 2 Results, Mathematics

Table	**Median Round 2 Bookmark (Page #)**			**Impact Data (Percentage At or Above)**			**Benchmark Data (Comparable Performance Level Using 2017 Test)**		
	App P	***P***	***Abv P***	***App P***	***P***	***Abv P***	***App P***	***P***	***Abv P***
G3	18	38	61	80	55	23	App P	P	Abv P
1	18	38	61	80	55	23	App P	P	Abv P
2	19	37	59	79	56	25	App P	App P	P
G4	16	40	68	82	53	20	App P	App P	P
1	16	41	68	82	51	20	App P	P	P
2	16	40	68	82	53	20	App P	App P	P
G5	17	43	62	78	45	22	B P	P	Abv P
1	18	42	60	76	47	25	App P	App P	P
2	16	48	62	78	36	22	B P	P	Abv P
G6	14	34	63	74	46	19	App P	P	Abv P
1	14	34	63	74	46	19	App P	P	Abv P
2	15	34	63	73	46	19	App P	P	Abv P

Table	Median Round 2 Bookmark (Page #)			Impact Data (Percentage At or Above)			Benchmark Data (Comparable Performance Level Using 2017 Test)		
	App P	***P***	***Abv P***	***App P***	***P***	***Abv P***	***App P***	***P***	***Abv P***
G7	10	31	58	75	48	24	App P	P	Abv P
1	10	30	58	75	50	24	App P	P	Abv P
2	9	33	58	76	46	24	App P	P	Abv P
G8	16	39	60	74	53	24	B P	App P	P
1	18	43	62	73	44	23	App P	P	Abv P
2	15	29	56	74	62	27	B P	App P	P

Note. The grade-level row summarizes the room data (the median across both tables). Benchmark data describes the percentage at or above each performance level using data from the 2017 Smarter Balanced test. Performance level abbreviation key: Below Proficient *(B P),* Approaching Proficient *(App P),* Proficient *(P),* Above Proficient *(Abv P).*

Table 19: Round 2 Results, Science

Table	Cut Score (Scaled Score)			Impact Data (Percentage At or Above)			Benchmark Data (NAEP)		
	App P	***P***	***Abv P***	***App P***	***P***	***Abv P***	***Basic***	***Proficient***	***Advanced***
G5	540	554	566	79	44	14	87	50	1
1	540	554	568	79	44	11	87	50	1
2	540	554	566	79	44	14	87	50	1
G8	845	854	879	68	44	3	81	46	2
1	845	854	877	68	44	4	81	46	2
2	845	854	879	68	44	3	81	46	2
G11	1150	1162	1188	51	21	1	75	42	3
1	1148	1162	1180	55	21	3	75	42	3
2	1150	1172	1188	51	8	1	75	42	3

Note. The grade-level row summarizes the room data (across both tables). Benchmark data describes the percentage at or above each performance level using data from the grade 8 NAEP; grade 5 benchmark data is interpolated from grade 8 NAEP and grade 11 is extrapolated from grade 8 NAEP. Performance level abbreviation key: Approaching Proficient *(App P),* Proficient *(P),* Above Proficient *(Abv P).*

4.6.13 Moderation and Results

To be adoptable, performance standards for a statewide system must be coherent across grades and subjects. There should be no irregular peaks and valleys, and they should be orderly across subjects with no dramatic differences in expectation. The following are characteristics of well-articulated standards:

- The cut scores for each performance level increase smoothly with each increasing grade.
- The cut scores should result in a reasonable percentage of students at each performance level; reasonableness can be determined by the percentage of students in the performance levels on historical tests or contemporaneous tests measuring the same or similar content.

- Barring significant content standard changes (e.g., major changes in rigor), the percentage *Proficient* might on new tests should not be radically different from the percentage proficient on historical tests, which is shown through benchmark data from the 2017 SBAC testing administration.

Panelists receive the information necessary for articulation prior to Round 2. Often, panelists intuitively create well-articulated sets of performance standards, but sometimes minor changes might improve articulation greatly. Calculated based on panelist recommendations and approved by NHDOE, these cuts are offered to a subset of panelists after Round 2 for consideration in a step referred to a moderation.

On the last day of the workshop, table leaders and panelists met to discuss and resolve any issues or needs related to cross-grade articulation.

Workshop leaders reminded panelists that content is one of multiple considerations in setting performance standards—perhaps the most important, but not the only consideration; panelists also considered impact and policy in Round 2. Table 20, Table 24, and Table 28 show the moderated recommendations and associated impact and benchmark data for ELA, mathematics, and science.

ELA Moderation and Final Results

ELA panelists adjusted the grade 3 bookmarks to better articulate across the grade.

Table 20: Moderated Results, ELA

Table	Median Moderation Bookmark (Page #)			Impact Data (Percentage At or Above)			Benchmark Data (Comparable Performance Level Using 2017 Test)		
	App P	***P***	***Abv P***	***App P***	***P***	***Abv P***	***App P***	***P***	***Abv P***
G3	9*	21*	39	79	55	26	App P	P	Abv P
G4	12	26	42*	75	56	29	App P	P	P
G5	9	24	45	81	62	24	App P	App P	Abv P
G6	9	22	43	81	56	18	App P	P	Abv P
G7	8	21	47	82	60	19	App P	P	Abv P
G8	12	27	52	81	58	18	App P	P	P

Note. The grade-level row summarizes the room data (the median across both tables). Benchmark data describes the percentage at or above each performance level using data from the 2017 Smarter Balanced test. Performance level abbreviation key: Approaching Proficient *(App. P),* Proficient *(P),* Above Proficient *(Abv. P).*

Table 21 and Figure 15 describe the performance level cut scores recommended for ELA.

Table 21: Performance Standards Recommended for ELA

Grade	Performance Standard		
	Approaching Proficient	***Proficient***	***Above Proficient***
3	557	587	616
4	580	605	635
5	594	621	664
6	605	642	688
7	608	644	697
8	625	661	711

Figure 15. Performance Standards Recommended for ELA

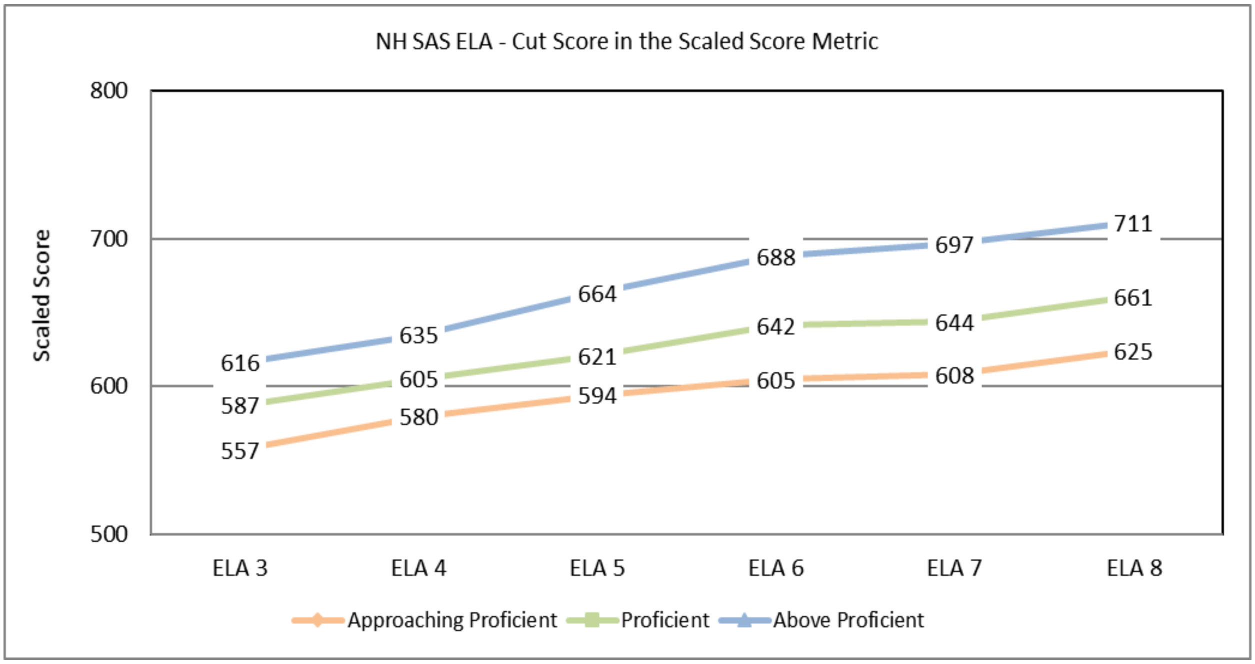
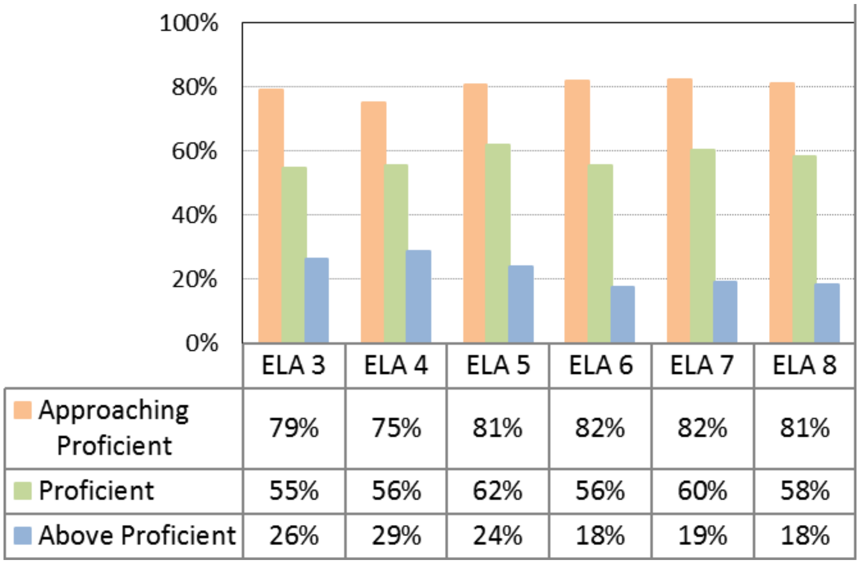


Table 22 indicates the percentage of students that we estimate reached each of the performance levels in ELA in 2018. Figure 16 represents those values graphically.

Table 22: Students Reaching Each Performance Standard, Spring 2018 ELA

Grade	**Percentage at or Above**		
	Approaching Proficient	***Proficient***	***Above Proficient***
3	79	55	26
4	75	56	29
5	81	62	24
6	82	56	18
7	82	60	19
8	81	58	18

Figure 16. Students Reaching Each Performance Standard, Spring 2018 ELA



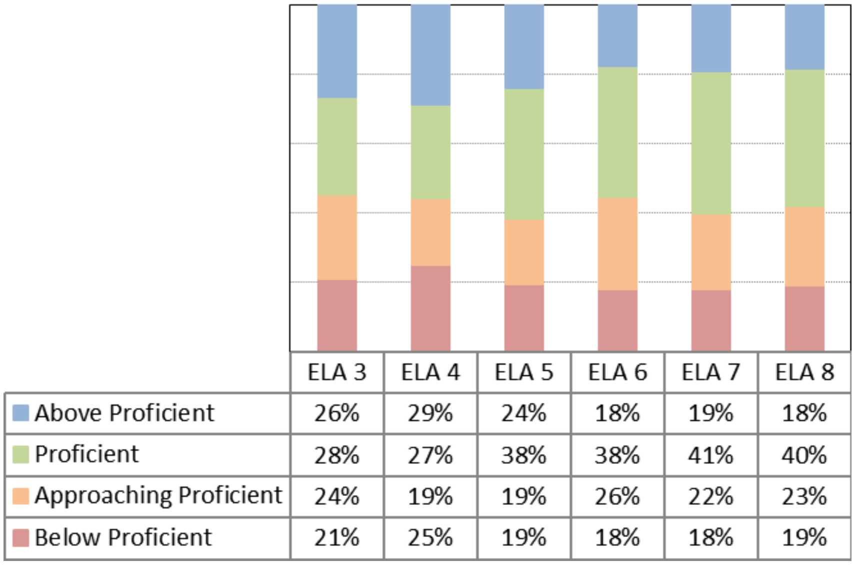
	ELA 3	ELA 4	ELA 5	ELA 6	ELA 7	ELA 8
🟧 Approaching Proficient	79%	75%	81%	82%	82%	81%
🟩 Proficient	55%	56%	62%	56%	60%	58%
🟦 Above Proficient	26%	29%	24%	18%	19%	18%

Table 23 and Figure 17 describe the percentage of students within each of the performance standards for ELA.

Table 23: Students Within Each Performance Standard, Spring 2018 ELA

Grade	Percentage Within Each Level			
	Below Proficient	***Approaching Proficient***	***Proficient***	***Above Proficient***
3	21%	24%	28%	26%
4	25%	19%	27%	29%
5	19%	19%	38%	24%
6	18%	26%	38%	18%
7	18%	22%	41%	19%
8	19%	23%	40%	18%

Figure 17. Students Within Each Performance Standard, Spring 2018 ELA



	ELA 3	ELA 4	ELA 5	ELA 6	ELA 7	ELA 8
🟦 Above Proficient	26%	29%	24%	18%	19%	18%
🟩 Proficient	28%	27%	38%	38%	41%	40%
🟧 Approaching Proficient	24%	19%	19%	26%	22%	23%
🟥 Below Proficient	21%	25%	19%	18%	18%	19%

Mathematics Moderation and Final Results

The mathematics moderators made minor adjustments to the *Approaching Proficient* bookmark in grades 6 and 7 and the *Above Proficient* bookmark in grade 8 (Table 24).

Table 24: Moderated Results: Mathematics

Grade	Median Round 2 Bookmark (Page #)			Impact Data (Percentage At or Above)			Benchmark Data (Comparable Performance Level Using 2017 Test)		
	App P	***P***	***Abv P***	***App P***	***P***	***Abv P***	***App P***	***P***	***Abv P***
3	18	38	61	80	55	23	App P	P	Abv P
4	16	40	68	82	53	20	App P	App P	P
5	17	43	62	78	45	22	B P	P	Abv P
6	13*	34	63	74	46	19	App P	P	Abv P
7	8*	31	58	75	48	24	App P	P	Abv P
8	16	39	42*	74	53	24	B P	App P	P

Note. The grade-level row summarizes the room data (the median across both tables). Benchmark data describes the percentage at or above each performance level using data from the 2017 Smarter Balanced test. Performance level abbreviation key: Below Proficient *(B P),* Approaching Proficient *(App P),* Proficient *(P),* Above Proficient *(Abv P).*

Table 25 and Figure 18 describe the performance standards recommended for mathematics.

Table 25: Performance Standards Recommended for Mathematics

Grade	Performance Level		
	Approaching Proficient	***Proficient***	***Above Proficient***
3	410	431	455
4	431	460	492
5	460	495	522
6	479	518	556
7	507	552	587
8	539	591	625

Figure 18. Performance Standards Recommended for Mathematics

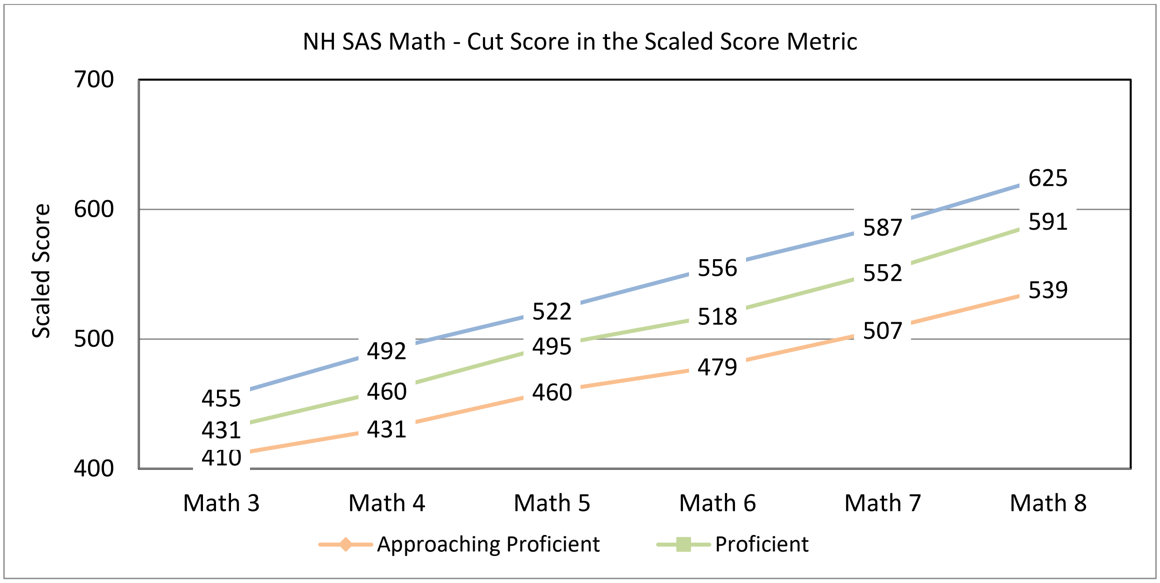
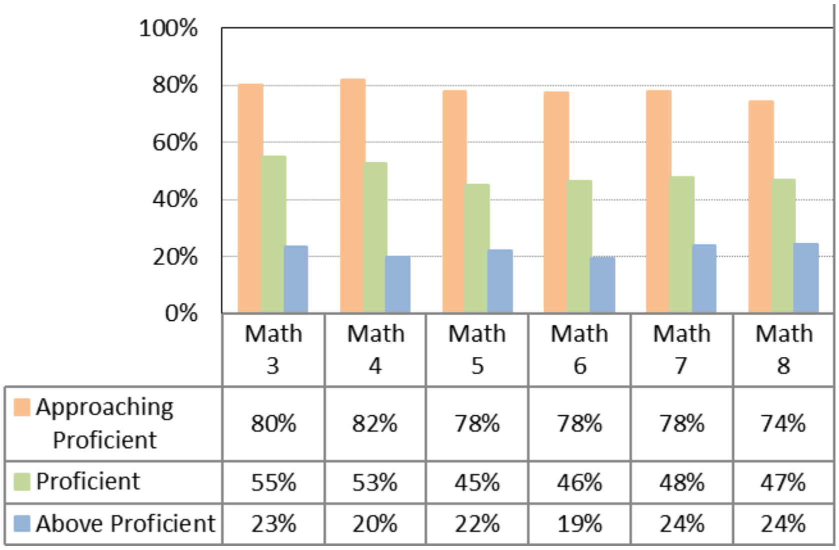


Table 26 indicates the percentage of students that we estimate reached each of the performance standards in mathematics in 2018. Figure 19 represents those values graphically.

Table 26: Students Reaching Each Performance Standard, Spring 2018 Mathematics

Grade	**Percentage at or Above**		
	Approaching Proficient	***Proficient***	***Above Proficient***
3	80	55	23
4	82	53	20
5	78	45	22
6	78	46	19
7	78	48	24
8	74	47	24

Figure 19. Students Reaching Each Performance Standard, Spring 2018 Mathematics



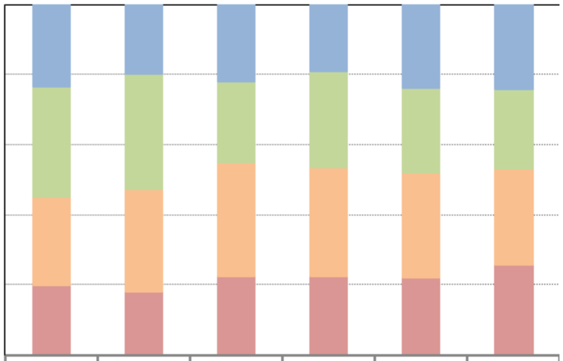
	Math 3	Math 4	Math 5	Math 6	Math 7	Math 8
🟧 Approaching Proficient	80%	82%	78%	78%	78%	74%
🟩 Proficient	55%	53%	45%	46%	48%	47%
🟦 Above Proficient	23%	20%	22%	19%	24%	24%

Table 27 and Figure 20 show the percentage of students within each of the performance standards for mathematics.

Table 27: Students Within Each Performance Standard, Spring 2018 Mathematics

Grade	**Percentage Within Each Level**			
	Below Proficient	***Approaching Proficient***	***Proficient***	***Above Proficient***
3	20	25	32	23
4	18	29	33	20
5	22	33	23	22
6	22	31	27	19
7	22	30	24	24
8	26	27	23	24

Figure 20. Students Within Each Performance Standard, Spring 2018 Mathematics



	Math 3	Math 4	Math 5	Math 6	Math 7	Math 8
🟦 Above Proficient	23%	20%	22%	19%	24%	24%
🟩 Proficient	32%	33%	23%	27%	24%	23%
🟧 Approaching Proficient	25%	29%	33%	31%	30%	27%
🟥 Below Proficient	20%	18%	22%	22%	22%	26%

Science Moderation and Final Results

From a policy perspective, the percentage *Proficient* on the new science test might be similar to benchmark data such as the percentage *Proficient* on the old science test (approximately 40%), performance on NAEP, or performance on the ACT/SAT. The grade 5 and 8 recommendations were close to benchmark data, and as a result, they made only minor changes based on Round 2 feedback. However, grade 11 recommendations raised expectations so high that only 20% of students would be *Proficient* and less than one percentage would be *Above Proficient*.

Panelists were not required to change their standards, but during moderation they were asked to consider if their recommended performance standards would be adoptable by the Board of Education. The grade 11 science educators discussed and made the following points in support of the Round 2 recommendations:

- The NGSS, and therefore the New Hampshire College and Career Readiness standards, are more rigorous and higher in cognitive demand than the previous science standards; therefore, a drop in the percentage of students able to demonstrate proficiency is not unreasonable.

- Because science courses merge the disciplines, and students must apply writing and mathematics skills in addition to science knowledge, they may present more challenges than mathematics or ELA courses, and therefore have lower test scores.

- The NGSS in high school are more rigorous and demanding than are the NGSS in grades 5 and 8, and therefore, fewer students in high school may be expected to demonstrate proficiency than students in elementary or middle school.

- Required science courses in New Hampshire tend to include biology and physical science, which may not sufficiently prepare students for proficient performance on the comprehensive new grade 11 test, which includes physics and chemistry.

If few students met the standards, students may be encouraged to take more advanced science courses, and the state may allocate additional resources and support for science education. Standard setting is aspirational by design, and they felt that the standards they set were high but reflected what students should know and be able to do, even if few students are currently able to meet those expectations.

About half the panelists wanted to stand by the content-only-based standards, and the other half agreed to consider the impact data and political context and were willing to revise their expectations to make them adoptable. Recognizing that the standards they really wanted to recommend were not implementable, the group revised them so that 40% of students could reach *Proficient* and ultimately recommended the performance standards described in Table 28.

Table 28: Moderated Results: Science

Table	Cut Score (Scaled Score)			Impact Data (Percentage At or Above)			Benchmark Data (NAEP)		
	App P	***P***	***Abv P***	***App P***	***P***	***Abv P***	***Basic***	***Proficient***	***Advanced***
G5	544	554	566	70	44	14	87	50	1
G8	845	854	870	68	44	9	81	46	2
G11	1146	1153	1176	60	40	5	75	42	3

Note. The grade-level row summarizes the room data (across both tables Benchmark data describes the percentage at or above each performance level using data from the grade 8 NAEP; grade 5 benchmark data is interpolated from grade 8 NAEP and grade 11 is extrapolated from grade 8 NAEP. Performance level abbreviation key Approaching Proficient *(App P),* Proficient *(P),* Above Proficient *(Abv P).*

Figure 21 describes the percentage of students reaching or exceeding each of the final recommended performance standards for science, and Figure 22 describes the percentage of students falling into each performance level.

Figure 21. Students Reaching or Exceeding Each Performance Standard, Spring 2018 Science

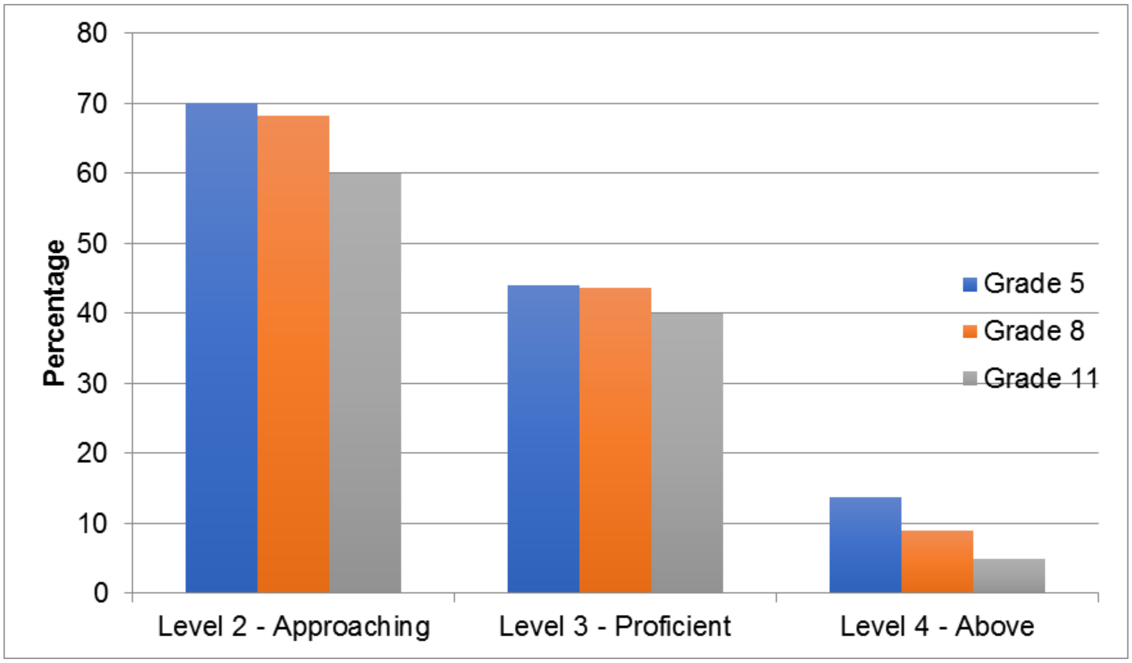
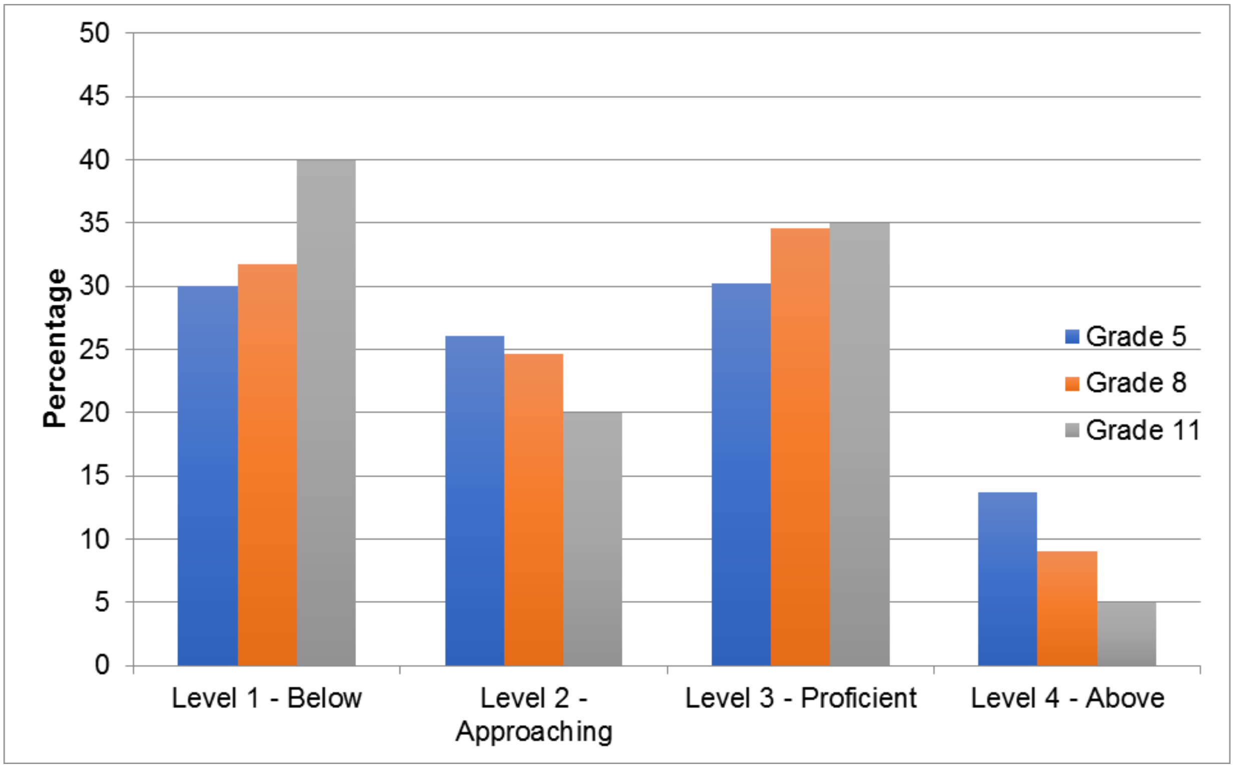


Figure 22. Students Within Each Performance Level, Spring 2018 Science



4.7 Workshop Evaluations

After finishing all activities, panelists completed online meeting evaluations independently, in which they described and evaluated their experience taking part in the standard setting. Table 29, Table 30, Table 31, Table 32, and Table 33 summarize the results of the Bookmark method evaluations and Table 34, Table 35, Table 36, Table 37, and Table 38 summarize the results of the AMP evaluations. Evaluation items not endorsed by at least 90% of panelists are discussed in text.

4.7.1 Bookmark Procedure Evaluations: ELA and Mathematics

Workshop participants overwhelmingly indicated clarity in the instructions, materials, data, and process (Table 29).

Table 29: Bookmark Evaluation (ELA and Mathematics), Clarity of Materials and Process

Please rate the clarity of the following components of the workshop.	"Somewhat Clear" or "Very Clear" (%)	
	ELA	Mathematics
Instructions provided by the Workshop Leader	96	100
Performance-Level Descriptors (PLDs)	96	100
Ordered-item booklet (OIB)	96	100
Panelist agreement data	100	100
Impact data (percentage of students that would achieve at the level indicated by the OIB page)	96	97

Note. Abbreviation Key: Number of responses = 57. Evaluation options included "Very Clear," "Somewhat Clear," "Somewhat Unclear," and "Very Unclear."

Participants felt that they had sufficient time to complete all activities. In fact, some indicated having too much time to complete some tasks (see Table 30). Participants indicated that the discussion of target PLDs, review of OIBs, and orientation session could have been shorter. For example

- Twenty-four panelists felt that the orientation was too long.
- Five panelists felt that the time allowed for taking the test was too long, and three indicated that it was too short.
- Four panelists indicated having too much time to review PLDs, and one reported that more time was needed.
- Eight panelists reported that there was too much time allocated for discussion of the PLDs, and two wanted more time to discuss the PLDs.
- Eleven panelists reported too much time to review the OIB, and three reported not enough time to review the OIB.
- Twenty panelists indicated that the time to place bookmarks was too long and one indicated that it was too short.

Table 30: Bookmark Evaluation (ELA and Mathematics), Appropriateness of Process

How appropriate was the amount of time you were given to complete the following components of the standard setting process?	**“About Right” (%)**	
	ELA	***Mathematics***
Large group orientation	46	69
Experiencing the online assessment	86	86
Review of the Performance-Level Descriptors (PLDs)	82	100
Discussion of the skills demonstrated by students who are "just barely" described by each PLD	82	83
Review of the ordered-item booklet (OIB)	75	76
Placement of your bookmarks in each round	60	66
Round 1 discussion	93	97

Note. Number of responses = 57. Evaluation options included “About Right,” “Too Much,” and “Too Little.”

Participants appreciated the value of the multiple factors contributing to bookmark placement, with participants rating each factor as important or very important (Table 31).

Table 31: Bookmark Evaluation (ELA and Mathematics), Importance of Materials

How important was each of the following factors in your placement of the bookmarks?	**“Somewhat Important” or “Very Important” (%)**	
	ELA	***Mathematics***
Performance-Level Descriptors (PLDs)	93	97
Your perception of the difficulty of the items	100	100
Your experience with students	100	97
Discussions with other panelists	100	100
External benchmark data	100	90
Room agreement data (room medians and individual bookmark placements)	100	100
Impact data (percentage of students that would achieve at the level indicated by the OIB page)	96	93

Note. Number of responses = 57. Evaluation options included “Not Important,” “Somewhat Important,” and “Very Important.”

Participants’ understanding of the workshop processes and tasks was consistently high (see Table 32). Four mathematics panelists and one ELA panelist found the impact data to be less helpful than other factors in placing their bookmarks.

Table 32: Bookmark Evaluation (ELA and Mathematics), Understanding Processes and Tasks

At the end of the workshop, please rate your agreement with the following statements.	"Agree" or "Strongly Agree" (%)	
	ELA	***Mathematics***
I understood the purpose of this standard setting workshop.	100	100
The procedures used to recommend performance standards were fair and unbiased.	96	90
The training provided me with the information I needed to recommend performance standards.	100	97
Taking the online assessment helped me to better understand what students need to know and be able to do to answer each question.	93	93
The Performance-Level Descriptors (PLDs) described what students within each performance level are expected to know and be able to do; they provided a clear picture of expectations for student performance at each level.	100	100
I was able to develop an understanding of the knowledge and skills demonstrated by students who are "just barely" described by the PLDs.	100	100
I understood how to review each page in the ordered-item booklet (OIB) to determine what students must know and be able to do to answer each item correctly.	96	100
I was able to interpret having an approximate 50% chance of answering an item correctly as indicating mastery.	100	100
I understood how to place my bookmarks.	100	100
I found the benchmark data and discussions helpful in my decisions about where to place my bookmarks.	96	97
I found the panelist agreement data (room medians and individual bookmark placements) and discussion helpful in my decision about where to place my bookmarks.	96	100
I found the impact data (percentage of students that would achieve at the level indicated by the OIB page) and discussions helpful in my decisions about where to place my bookmarks.	96	86
I felt comfortable expressing my opinions throughout the workshop.	100	100
Everyone was given the opportunity to express his or her opinions throughout the workshop.	96	100

Note. Number of responses = 57. Evaluation options included "Strongly Agree" "Agree," "Disagree," and "Strongly Disagree."

Participants agreed that the standards set during the workshop reflected the intended grade-level expectations (Table 33).

Table 33: Bookmark Evaluation (ELA and Mathematics), Student Expectations

Please read the following statement carefully and indicate your response.	**“Agree” or “Strongly Agree” (%)**	
	ELA	***Mathematics***
A student performing at Level 3 meets expectations for the grade level.	100	100
A student performing at Level 2 is below expectations for the grade level.	93	90
A student performing at Level 4 exceeds expectations for the grade level.	100	100

Note. Number of responses = 57. Evaluation options included “Strongly Agree,” “Agree,” “Disagree,” and “Strongly Disagree.”

4.7.2 AMP Evaluations: Science

Workshop participants generally indicated clarity in the instructions, materials, data, and process (see Table 34). Five panelists, three from grade 11 and one each from grades 5 and 8, indicated that the PLDs were somewhat unclear, three panelists (two from grade 11) reported that the OSABs were somewhat unclear, and two grade 11 panelists indicated that the impact data was somewhat unclear.

Table 34: AMP Evaluations (Science), Clarity of Materials and Processes

Please rate the clarity of the following components of the workshop.	**“Somewhat Clear” or “Very Clear” (%)**
Instructions provided by the Workshop Leader	96
Performance-Level Descriptors (PLDs)	81
Ordered Scoring Booklet (OSAB)	88
Panelist agreement data	100
Impact data (percentage of students that would achieve at the level indicated by the OSAB page)	88

Note. Abbreviation Key: Number of responses = 26. Evaluation options included “Very Clear,” “Somewhat Clear,” “Somewhat Unclear,” and “Very Unclear.”

Panelists felt that the time allocated to various workshop tasks may be adjusted (Table 35). Of the panelists who did not indicate that the time allocation for a task was “About Right”

- eleven indicated that the large-group orientation was too long;
- eleven (four from grade 5 and seven from grade 11) indicated not having enough time for taking the test;
- one reported having too much time to review the PLDs, and two reported having too much time for review;

- nine panelists indicated having too little time to review the OSAB (one from grade 5, four each from grade 8 and 11), and one panelist (grade 11) reported having too much time for review;
- three panelists indicated having too much time to place scoring assertions, and one panelist indicated having too much time to place them; and
- five panelists (one from grade 5 and four from grade 11) reported having too little time for discussing Round 1, and two grade 11 panelists indicated having too much time for discussion.

Table 35: AMP Evaluations (Science), Appropriateness of Process

How appropriate was the amount of time you were given to complete the following components of the standard setting process?	**“About Right” (%)**
Large group orientation	58
Experiencing the online assessment	58
Review of the Performance-Level Descriptors (PLDs)	88
Review of the Ordered Scoring Assertion Booklet (OSAB)	62
Placement of your scoring assertion mapping decisions in each round	85
Round 1 discussion	73

Note. Number of responses = 26. Evaluation options included “About Right,” “Too Much,” and “Too Little.”

Panelists felt that the materials used throughout the process were important to identifying the cut scores. Impact data and external benchmark data was perceived as less important than other factors (Table 36).

Table 36: AMP Evaluations (Science), Importance of Materials

How important were each of the following factors in your placement of the scoring assertion mapping decisions?	**“Somewhat Important” or “Very Important” (%)**
Performance-Level Descriptors (PLDs)	100
Your perception of the difficulty of the items	100
Your experience with students	100
Discussions with other panelists	96
External benchmark data	88
Room agreement data (room and individual assertion mapping placements)	96
Impact data (percentage of students that would achieve at the level indicated by the OSAB)	85

Note. Number of responses = 26. Evaluation options included “Not Important,” “Somewhat Important,” and “Very Important.”

With a few exceptions, panelists indicated a high level of agreement to statements assessing understanding of the processes and materials used throughout the workshop (Table 37).

Five panelists from grades 8 and 11 disagreed with the statement that the process was fair and unbiased, six panelists (two from each grade) indicated that the PLDs provided a clear picture of expectations for student achievement, and four grade 11 panelists felt that the impact data was not helpful in placing cut scores.

Table 37: AMP Evaluations (Science), Understanding Processes and Tasks

At the end of the workshop, please rate your agreement with the following statements.	**"Agree" or "Strongly Agree" (%)**
I understood the purpose of this standard setting workshop.	100
The procedures used to recommend performance standards were fair and unbiased.	81
The training provided me with the information I needed to recommend performance standards.	100
Taking the online assessment helped me to better understand what students need to know and be able to do to answer each question.	96
The Performance-Level Descriptors (description of what students within each performance level are expected to know and be able to do) provided a clear picture of expectations for student achievement at each level.	77
I understood how to review each assertion in the Ordered Scoring Assertion Booklet (OSAB) to determine what students must know and be able to do to answer each item correctly.	100
I understood how to place my scoring assertion mapping decisions.	100
I found the benchmark data and discussions helpful in my decisions about where to place my scoring assertion mapping decisions.	100
I found the panelist agreement data (room and individual assertion placements) and discussion helpful in my decisions about where to place my scoring assertion mapping decisions.	100
I found the impact data (percentage of students that would achieve at the level indicated by the OSAB) and discussions helpful in my decisions about where to place my scoring assertion mapping decisions.	85
I felt comfortable expressing my opinions throughout the workshop.	100
Everyone was given the opportunity to express his or her opinions throughout the workshop.	100

Note. Number of responses = 26. Evaluation options included "Strongly Agree" "Agree," "Disagree," and "Strongly Disagree."

The majority of panelists agreed that the performance standards described grade-level expectations. However, a few grade 11 panelists (and one grade 8 panelist) disagreed. Five disagreed with the level 3 statement, and four disagreed with the level 4 and level 2 statements (see Table 38).

Table 38: AMP Evaluations (Science), Student Expectations

Please read the following statement carefully and indicate your response.	"Agree" or "Strongly Agree" (%)
A student performing at Level 3 meets expectations for the grade level.	81
A student performing at Level 2 is below expectations for the grade level.	85
A student performing at Level 4 exceeds expectations for the grade level.	85

Note. Number of responses = 26. Evaluation options included "Strongly Agree," "Agree," "Disagree," and "Strongly Disagree."

4.7.3 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." Forty-five ELA and mathematics participants and 25 science participants responded to the first question and 38 ELA participants and 23 science participants responded to the second question.

Although most participants indicated the process was clear and did not need to be improved, some suggested that there be less downtime between tasks, that the process be demonstrated while it is being introduced and explained, that the large-group orientations be shorter, that instructions be printed, and that more visual examples be provided. Participants expressed gratitude for being involved in setting performance standards and for interacting with so many educators from across the state. They appreciated the organization, well-prepared materials, and technology, and many panelists complemented the professionalism and expertise of the facilitators.

Additional participant comments included:

> *"I thought this workshop was a great use of time, and I enjoyed being a part of the panel."*
>
> *"I appreciate the process that we took to arrive at the benchmarks we selected."*
>
> *"This was a great way to better understand how our students are assessed."*
>
> *"The process was interesting and enlightening. I am glad I participated."*

5. Validity Evidence

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

5.1 Evidence of Adherence to Professional Standards and Best Practices

The NH SAS standard-setting workshop was designed and executed to be consistent with established practices and best practice principles (Hambleton & Pitoniak, 2006; Hambleton, Pitoniak, & Copella, 2012; Kane, 2001; Mehrens, 1995). The process also adhered to the following professional standards recommended by the AERA/APA/NCME Standards for Educational and Psychological Testing (2014) related to standard setting:

> Standard 5.21: When proposed score interpretation involves one or more cut scores, the rationale and procedures used for establishing cut scores should be documented clearly.
>
> Standard 5.22: When cut scores defining pass-fail or proficiency levels are based on direct judgments about the adequacy of item or test performances, the judgmental process should be designed so that the participants providing the judgments can bring their knowledge and experience to bear in a reasonable way.
>
> Standard 5.23: When feasible and appropriate, cut scores defining categories and distinct substantive interpretations should be informed by sound empirical data concerning the relation of test performance to the relevant criteria.

The sections of this report documenting the rationale and procedures used in the standard-setting workshop address Standard 5.21. The standard-setting procedures applied are appropriate for tests like the NH SAS. Section 4.1 provides the justification for and the additional benefits of selecting the methods to establish the cut scores, and Sections 4.6 and 4.7 document the process followed to implement the methods.

The design and implementation of the Bookmark and AMP procedures address Standard 5.22. Both methods directly leverage the subject matter expertise of the panelists and incorporate 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, including:

- understanding the test and test items (from an educator and student perspective);
- describing the content measured by the test as described by the content standards;
- identifying the skills associated with each test item;
- describing the skills associated with “just barely” students for each performance level;
- selecting which test items students in each performance level should be able to answer correctly;

- describing the skills necessary for an assertion to be scored as correct;
- evaluating and applying feedback and reference data to their Round 2 bookmarks; and
- considering the impact of the recommended cut scores on students.

Additionally, panelists' readiness evaluations provided evidence of a successful orientation to the process and understanding of the standard setting procedures, and 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 and knowledgeable about the subject area and students' developmental levels. Section 4.3.4 summarizes details about the panel demographics and qualifications.

The provision of benchmark and impact data to panelists after Round 1 addresses Standard 5.23. This empirical data provides necessary and additional context describing student performance given the recommended standards.

5.2 Evidence in Terms of Peer Review Critical Elements

The United States Department of Education (ED) 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; ED, 2015). The following critical elements are relevant to standard setting; evidence supporting each element immediately follows.

> Critical Element 1.2: Substantive involvement and input of educators and subject-matter experts

New Hampshire educators played a critical role in establishing performance levels for the NH SAS. They reviewed and revised the PLDs, drafted and applied target PLDs to delineate performance at each performance level, considered benchmark data and the impact of their recommendations, and formally recommended performance standards.

Many subject-matter experts contributed to the development of New Hampshire's performance standards. Contributing educators were subject-matter experts in their content area, the content standards and curriculum that they teach, and the developmental and cognitive capabilities of their students. AIR's facilitators were subject-matter experts in the subjects tested, alternate assessments, and 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. Finally, Dr. Phillips is a nationally known subject-matter expert in assessment and measurement, including multiple methods of standard setting.

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

Evidence to support this critical element includes:

1) The rationale for and technical sufficiency of the Bookmark method selected to establish performance standards (Section 4.1)
2) The rationale for selecting and applying the AMP (Section 4.1.2)
3) 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 4.6 and 5.1)
4) Panelists self-reported readiness to undertake the tasks (Section 4.6.8) and confidence in the workshop process and outcomes (Section 4.7) supporting the validity of the process
5) The standard-setting panels consisted of panelists with appropriate experience and expertise, including content experts with experience teaching the New Hampshire's academic content standards and prioritized standards in the tested grades and subjects, and individuals with experience and expertise teaching special and general education students in New Hampshire (Section 4.3.4).

REFERENCES

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

Bradlow, E.T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*, 153-168.

Cizek, G. J., & Bunch, M. B. (2007). Standard setting: A guide to establishing and evaluating performance standards on tests. Thousand Oaks, CA: Sage.

Cizek, G. J., and 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.

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, NY: Routledge.

Huynh, H. (2006), A Clarification on the Response Probability Criterion RP67 for Standard Settings Based on Bookmark and Item Mapping. *Educational Measurement: Issues and Practice*, 25: 19–20.

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.

Karantonis, A. & Sireci, S. (2006). The Bookmark Standard-Setting Method: A Literature Review. *Educational Measurement: Issues and Practice*. 25. 4–12.

Lewis, D. M., Mitzel, H. C., Mercado, R. L., & Schulz, E. M. (2012). The bookmark standard setting procedure. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations (2nd Edition)* (pp. 225–253). New York, NY: Routledge.

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

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

Perie, M. (2005, April). Angoff and Bookmark methods. Workshop presented at the annual Meeting of the National Council on Measurement in Education, Montreal, Canada.

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