West Virginia General Summative Assessment

2017-2018

Volume 3 Part 2
Setting Achievement Standards for Science

West Virginia DEPARTMENT OF
EDUCATION

TABLE OF CONTENTS

LIST OF APPENDICES

LIST OF TABLES

LIST OF FIGURES

1. EXECUTIVE SUMMARY

AIR conducted a standard-setting workshop to recommend performance standards for West Virginia's General Summative Assessments (WVGSA) in science at grades 5 and 8. The workshop was conducted July 30 – August 1, 2018, at the Charleston Marriott Town Center in Charleston, WV.

West Virginia's General Summative Assessments in science are designed to measure West Virginia's Next Generation Content Standards and Objectives for Science. Test items were developed by the American Institutes for Research (AIR) in conjunction with a group of states working to implement science standards influenced by the three-dimensional Next Generation Science Standards (NGSS). Test items were developed to ensure that each student is administered a test meeting all elements of WVGSA's science test blueprint, which was constructed to align to the West Virginia's Next Generation Content Standards and Objectives for Science.

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

Panelists reviewed achievement-level descriptors (ALDs) describing the degree to which students have achieved the West Virginia Next Generation Content Standards and Objectives for Science. ALDs were reviewed and revised in a separate workshop conducted prior to the standard-setting workshop. Working through the ordered assertions for each cluster, panelists mapped each assertion into one of the four achievement levels describing proficiency, including: Does Not Meet Standard, Partially Meets Standard, Meets Standard, and Exceeds Standard. The panelists performed the assertion mapping in two rounds of standard setting during the two-day workshop. Panelists' mapping of the scoring assertions was used to identify the location of the three achievement standards used to classify student achievement – Partially Meets Standard, Meets Standard, and Exceeds Standard. Following Round 2, panelists engaged in a moderation session to review and modify recommended achievement standards to facilitate the adoption of an articulated set of achievement standards across grades and subject areas.

Twenty-nine West Virginia science educators served as science standard-setting panelists. They 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.

1.1.STANDARD-SETTING WORKSHOPS

1.1.1. Overall Structure of the Workshops

The key features of the workshops included the following:

- The standard-setting procedure produced three achievement standards (Partially Meets Standard, Meets Standard, and Exceeds Standard) that will be used to classify student science performance on the West Virginia General Summative Assessment (WVGSA).
- Panelists recommended achievement standards in two rounds.
- Impact data (percentage of students reaching each achievement standard) were provided to panelists following the first round of recommending achievement standards.
- The standard-setting workshops were conducted online using AIR's online standard-setting tool. A laptop computer was provided to each panelist at the workshops.
- Following Round 2, panelists engaged in a moderation session to review and modify recommended achievement standards to achieve an articulated system of standards across grades and subject areas.

1.1.2. Results of the Standard-Setting Workshops

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

Table 1. Achievement Standards Recommended for Science

	Achievement Standard		
Grade	Partially Meets Standard	Meets Standard	Exceeds Standard
5	537	555	568
8	837	855	867

Table 2 indicates the percentage of students that we estimate will reach each of the achievement standards in 2018. Figure 1 represents those values graphically.

Table 2. Estimated Percentage of Students Reaching or Exceeding Each Science Achievement Standard in 2018

	Percentage Meeting or Exceeding		
Grade	Partially Meets Standard	Meets Standard	Exceeds Standard
5	83%	38%	11%
8	83%	39%	12%

Figure 1. Estimated Percentage of Students Reaching or Exceeding Each Science Achievement Standard in 2018

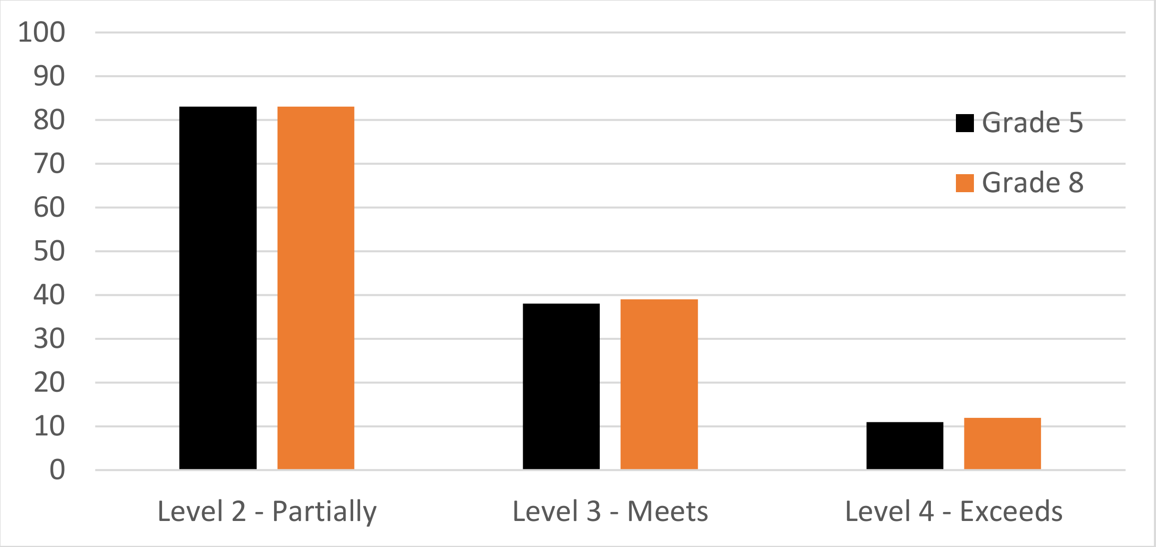
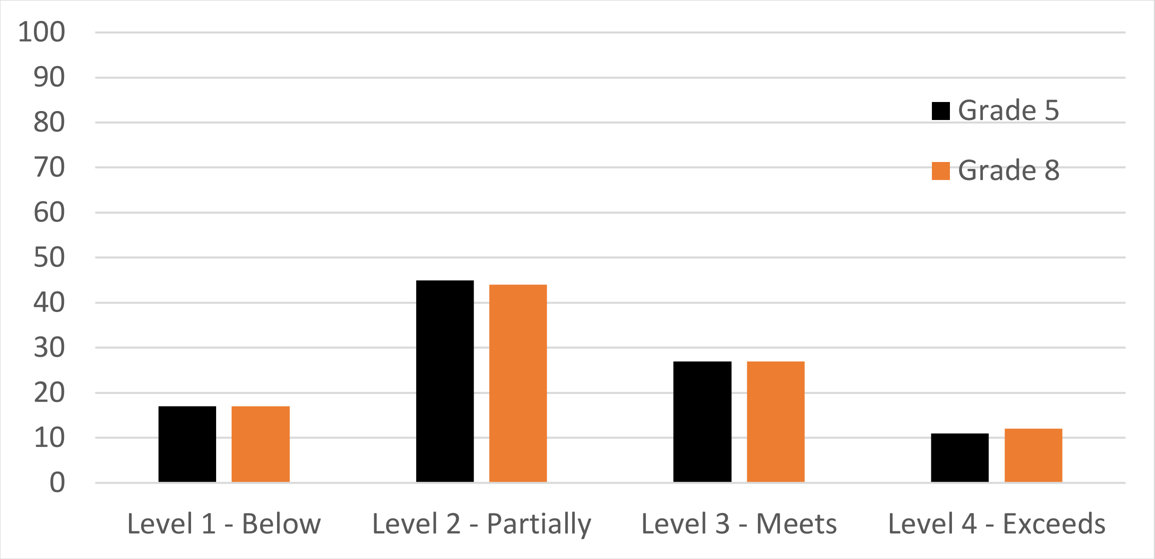


Table 3 indicates the estimated percentage of students within each of the achievement levels in 2018. The values are displayed graphically in Figure 2.

Table 3. Estimated Percentage of Students Within Each Science Achievement Level in 2018

Grade	Percentage Classified Within Achievement Level			
	Does Not Meet Standard	Partially Meets Standard	Meets Standard	Exceeds Standard
5	17%	45%	27%	11%
8	17%	44%	27%	12%

Figure 2. Estimated Percentage of Students Classified Within Each Science Achievement Level in 2018



2. INTRODUCTION

The West Virginia General Summative Assessment (WVGSA) for students in grades 3–8 is a new online summative criterion-referenced test given toward the end of the school year to measure student performance on the state's content standards in mathematics and English language arts (ELA) in grades 3–8 and science in grades 5 and 8.

New tests require new achievement standards to link performance on the test to the content standards. The West Virginia Department of Education (WVDE) contracted with the American Institutes for Research (AIR) to establish cut scores for the new tests. To fulfill this responsibility, AIR implemented an innovative, defensible, valid, and technically-sound method; provided training on standard setting to all participants; oversaw the process; computed real-time feedback data to inform the process; and produced a technical report documenting the method, approach, process, and outcomes. Achievement standards were set for grades 3–8 mathematics and ELA in June 2018 and for grades 5 and 8 science in July 2018.

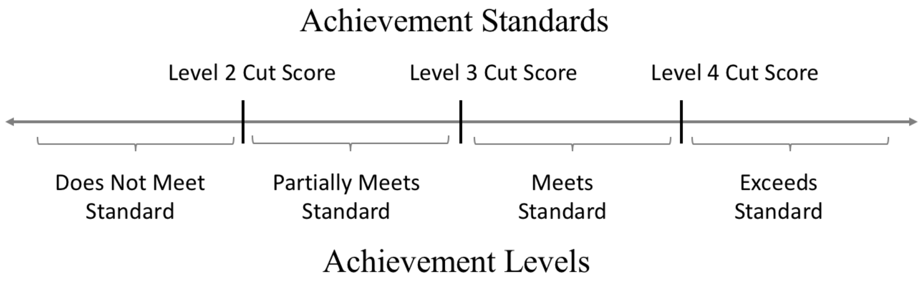
The purpose of this report is to document the standard-setting process for WVGSA science and resulting achievement standard recommendations.

3. STANDARD SETTING

Thirty educators from West Virginia (15 for each grade-level test) convened at the Charleston Marriott Town Center in Charleston, WV, from July 31 through August 1, 2018, to complete two rounds of standard setting to recommend three achievement standards for the WVGSA science tests.

Standard setting is the process used to define achievement on the WVGSA. Achievement levels are defined by achievement standards, or cut scores, that specify how much of the content standards students must know and be able to do in order to meet the minimum for each achievement level. As shown in Figure 3, three achievement standards are sufficient to define four achievement levels.

Figure 3. Three Achievement Standards Defining West Virginia's Four Achievement Levels



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

3.1.The Assertion Mapping Procedure

A new approach to setting achievement standards is necessary for tests based on the Next Generation Science Standards (NGSS) due to the structure of the content standards, and subsequently, the structure of test items assessing the standards. Tests based on the NGSS, such as the WVGSA, adopt a three-dimensional conceptualization of science understanding. Each item aligns to a science practice, one or more crosscutting concepts, and one or more disciplinary core ideas. Accordingly, the new science assessments are comprised mostly of item clusters representing a series of interrelated student interactions directed toward describing, explaining, and predicting scientific phenomena. Some stand-alone items are added to increase the coverage of the test without also increasing testing time or testing burden.

Within each item or 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, thereby not supporting the assertion that the student can interpret relationships expressed graphically.

While some other assessments, especially ELA, comprise items probing a common stimulus, the degree of interdependence among such items is limited and student achievement on such items can be evaluated independently of student achievement on other items within the stimulus set. This is not the case with the new science 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 between the interactions and resulting assertions of an item cluster are too substantial to ignore because those item interactions and assertions are more intrinsically related to each other. The interdependence of student interactions within items has consequences both for scoring and recommending achievement standards.

To account for the cluster-specific variation of related item clusters, additional dimensions can be added to the IRT model. Typically, these are nuisance dimensions unrelated to student ability. Examples of IRT models that follow this approach are the bi-factor model (Gibbons & Hedeker,

1992) and the testlet model (Bradlow, Wainer, & Wang, 1999). The testlet model is a special case of the bi-factor model (Rijmen, 2010).

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

To address these challenges, AIR psychometricians designed a new method for setting achievement standards on new tests of the NGSS, including the WVGSA 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 procedures such as the Bookmarking procedure WVDE 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 supported (or not) 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 achievement-level descriptor during two rounds of standard setting. Like the Bookmark method, assertion mappings are made independently with the goal of convergence over two rounds of rating, rather than consensus.[1]

3.2.Workshop Structure

One large meeting room served as an all-participant training room. This room broke into two separate working rooms, one for each set of grade-level panels, after the all-group orientation. As shown in Figure 4, two separate panels set achievement standards for each grade.

[1] AIR historically implements two rounds of standard setting as best practice in the Bookmark method and extends this practice to the AMP method. Panels typically converge in Round 2, and the moderation session provides the opportunity for any necessary articulation. In addition to lessening panelist burden from having to repeat a cognitively-demanding task for a third time, using two rounds introduces significant cost efficiency by reducing the number of days needed for standard setting. Panelists completing two rounds report levels of confidence in the outcomes that are similar to the confidence expressed by panelists participating in three rounds. Psychometric evaluation of the reliability and variability in results from two and three rounds are generally consistent. AIR has used two rounds in standard setting in over 12 states and 30 assessments, beginning in 2001 with the enactment of NCLB.

Figure 4. WVGSA Science Room Structure

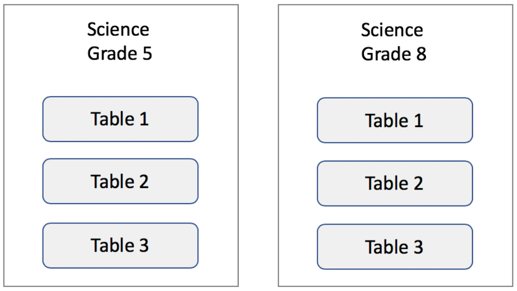


Table 4 summarizes the composition of the tables and the number of facilitators and panelists assigned to each. The 30 standard-setting participants included table leaders and panelists who taught in the content area and grade level for which standards were being set.

Table 4. WVGSA Science Table Assignments

Panel	Room	Table	Table Leaders	Panelists	Facilitator	Facilitator Assistant
Grade 5	1	1	1	5	Kevin Chandler	Danielle Peterford
		2	1	5		
		3	1	5		
Grade 8	2	1	1	5	Josh Smith	Matt Shina
		2	1	5		
		3	1	5		
Totals	2	6	6	30	2	2

3.3.Participants and Roles

3.3.1. West Virginia Department of Education Staff

WVDE staff from the Office of Assessment were present throughout the process and provided overall policy context and answered any policy questions that arose. They included:

- Dr. Timothy Butcher, Coordinator, Office of Assessment
- Sonja Phillips, Coordinator, Office of Assessment
- Sonya White, Assistant Director, Office of Assessment
- Courtney Dexter, Intern, Office of Assessment
- Rob Surface, Coordinator, Office of Assessment
- Carrie Christy, Secretary, Office of Assessment
- Robin Sizemore, Coordinator, Middle and Secondary Learning

3.3.2. 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:

- Dr. Gary Phillips, AIR Vice President and Institute Fellow, and Dr. Stephan Ahadi, Managing Director of Psychometrics

  - Facilitated and oversaw the workshop. Dr. Ahadi provided training to all participants, including the facilitators, the table facilitators, and all participants, supervised the psychometric analyses conducted during and after the workshop, and presented impact and benchmark data to panelists after each round.
- Dr. Frank Rijmen, Lead Psychometrician
  - Provided psychometric services; Alesha Ballman, Psychometric Support Assistant, provided support.
- Patrick Kozak, Psychometric Support Manager,
  - Oversaw analytics technology and psychometrics
- Matthew Shina and Danielle Peterford, Psychometric Support Assistants
  - Provided support as needed.
- Mark Lewis, Senior Program Manager
  - Managed process and logistics throughout the meeting.
- Michael Dao and Samba Ndiaye, System Support Agents
  - Set up, tested, and troubleshot technology during the workshop.

3.3.3. Observers

Given the newness of the Assertion Mapping Procedure, representatives from states planning to employ the process in the near future observed the second day of the workshop. They did not interact with panelists or impact the process in any way.

3.3.4. Room Facilitators

An AIR room facilitator and assistant facilitator guided the process in each room. Facilitators were content experts experienced in leading standard-setting processes, had led AMP processes before, 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. Facilitators were:

- Kevin Chandler served as the grade 5 room facilitator, and Danielle Peterford served as assistant room facilitator.
- Joshua Smith served as the grade 8 room facilitator, and Matt Shina served as assistant room facilitator.

Each facilitator was trained to be extensively knowledgeable of the constructs, processes, and technologies used in standard setting. All facilitators and assistant facilitators participated in a full-day process training and a technology training prior to each workshop.

3.3.5. Table Leaders

WVDE 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.

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

Table leaders provided the following support throughout the workshop:

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

3.3.6. Educator Participants

To establish achievement standards, WVDE recruited a diverse set of participants from across the state. Panelists included science teachers, administrators, and representatives from other stakeholder groups (e.g., parents, business representatives) to ensure that a diverse range of perspectives contributed to the standard-setting process and product. In recruiting panelists, WVDE targeted the recruitment of participants to be representative of the gender and geographic representation of the teacher population found in West Virginia. Table 5 summarizes characteristics of the panels.

Table 5. Panelist Characteristics

	Percentage of Panelists		
	Grade 5	Grade 8	Overall
Male	40%	21%	31%
Non-White	13%	0%	7%
District Size			
Large	33%	43%	38%
Medium	33%	14%	24%
Small	13%	36%	24%
Not Applicable	20%	7%	14%
District Urbanicity			
Urban	7%	0%	3%
Suburban	20%	21%	21%
Rural	53%	57%	55%
Not Applicable	20%	7%	14%
Unknown	0%	14%	7%
Stakeholder Group			
Educator	67%	64%	66%
Administrator	13%	14%	14%
Educator or Other	20%	14%	17%

Note. The “Other” category included a parent and a community/business member.

For results of any judgment-based method to be valid, the judgments must be made by individuals who are qualified to make them. Participants in the West Virginia standard-setting workshop were highly qualified. They brought a variety of experience and expertise. Many had taught for 12 years or more, and most had professional experience outside the classroom. They also represented a range of stakeholders, such as educators, administrators, parents, and business leaders. Table 6 summarizes the qualifications of the panels.

Table 6. Panelist Qualifications

	Percentage of Panelists		
	Grade 5	Grade 8	Overall
Years Teaching Experience			
5 Years or Less	20%	21%	21%
6 to 10 Years	20%	36%	28%
11 years or More	53%	43%	48%
Years Professional Experience			
5 Years or Less	53%	57%	55%
6 to 10 Years	13%	7%	10%
11 years or More	27%	36%	31%
Highest Degree Earned			
Bachelor’s	47%	21%	34%
Master’s	40%	57%	48%
Doctorate	7%	14%	10%
Other	7%	7%	7%
Experience with ELLs	40%	43%	41%
Experience with SWDs	80%	79%	79%
Experience with Low-SES Students	73%	79%	76%

Notes. Percentages in table describe all participants, not just educator participants. Percentages may not sum to 100% because not all participants were educators and therefore did not have valid responses for years of teaching experience or experience with different student populations. Abbreviation Key: English Language Learners (ELLs), Students with Disabilities (SWDs), Socio-economic Status (SES).

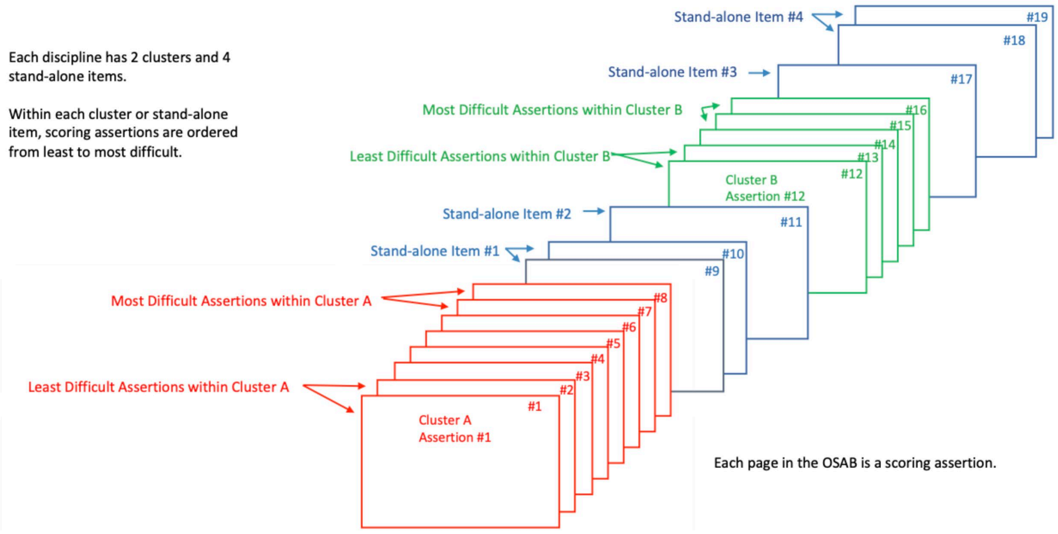
Panelist expertise informed table composition. Someone at each table had served as an item writer, item reviewer, or test scorer and was familiar with the scoring assertions. Appendix A: Standard Setting Panelists provides additional information about the individuals participating in the standard-setting workshop.

3.4.Materials

3.4.1. Ordered Scoring Assertion Booklets

Like the Bookmark method used for establishing achievement standards for the WVGSA mathematics and ELA tests, the Assertion Mapping Procedure (AMP) method uses booklets of ordered test materials for setting standards. Instead of test items, the AMP uses scoring assertions presented in grade-specific booklets called ordered scoring assertion booklets (OSABs). Each OSAB represents one possible testing instance resulting from applying the test blueprints to the item bank. Figure 5 describes the structure of the OSAB.

Figure 5. Ordered Scoring Assertion Booklet (OSAB)


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, just like clusters and stand-alone items would be selected at random by the algorithm that was used to assemble operational tests linearly on the fly. Within each item 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 5 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 all achievement levels. For example, a cluster may have assertions that map onto “Does Not Meet Standard,” “Partially Meets Standard,” and “Meets Standard,” but not “Exceeds Standard.” Clusters may have as few as four assertions or as many as 20 assertions. Each assertion is worth one score-point.

Each OSAB contains three disciplines and 18 tasks (clusters and items). The grade 5 and 8 OSABs contained 69 assertions (in grade 5) and 68 assertions (in grade 8), each comprised of 6 item clusters and 12 stand-alone items.

3.4.2. West Virginia’s Next Generation Content Standards and Science Objectives

The WVGSA assess the learning objectives described by West Virginia’s Next Generation Content Standards and Objectives for Science, adopted in 2016.

The Next Generation Content Standards and Objectives for Science are available at: https://wvde.state.wv.us/instruction/NxGen.html.

3.4.3. Achievement-Level Descriptors

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

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

ALDs link the standards to the achievement standards. There are four types of ALDs:

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

West Virginia uses four achievement levels to describe student performance: “Does Not Meet Standard,” “Partially Meets Standard,” “Meets Standard,” and “Exceeds Standard.”

The Washington State Office of Superintendent of Public Instruction (OSPI) drafted initial range PLDs based on the Next Generation Science Standards (NGSS). AIR, West Virginia Department of Education (WVDE) 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 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.

The WVDE then reviewed the policy and range PLDs to ensure that the language accurately represented the goals and policies of their state. AIR worked with them to make revisions where necessary. Prior to the standard-setting workshop, WVDE hosted an ALD review meeting with stakeholders to review, revise, and approve the policy and range PLDs.

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

3.5.Workshop Technology

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

Using the tool, panelists could review the item clusters and scoring assertions, they could determine the relative difficulty of assertions to other assertions in the same cluster, examine the content alignment of each assertion, assign assertions to achievement 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 all other assertions in the OSAB (not just within the cluster). Panelists also reviewed their own assertion placement, their table's placement, the other tables' placement, and the overall placement for both tables.

Panelists were able to add notes and comments on the items, item clusters, or assertions as they reviewed them and examine impact 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.

3.6.Events

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

Table 7. Standard-Setting Agenda Summary

Day 1: Tuesday, July 31	• Table leader training • Orientation and introductions • Content standards review • Take the test • ALD review • Create "Does Not Meet Standard" ALDs • Item and item cluster review • OSAB review
Day 2: Wednesday, August 1	• 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

3.6.1. Orientation

Dr. Timothy Butcher from the WVDE Office of Assessment welcomed panelists to the workshop and provided context and background. Dr. Stephan Ahadi then oriented participants to the workshop by describing the purpose and objectives of the meeting, explaining the process to be implemented to meet those objectives, and outlining the events that would happen each day. He reviewed the responsibilities of the three groups of people at the workshop: panelists, AIR staff, and WVDE personnel, and explained that panelists were selected because they were experts, and how the process to be implemented over the two days was designed to elicit and apply their expertise to recommend new cut scores. Finally, he described how standard setting works and what would happen once the panelists had finalized their recommendations.

3.6.2. Confidentiality and Security

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

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

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

3.6.3. Review ALDs

Panelists completed a thorough review of the ALDs for their assigned grade. They identified key words describing the skills necessary for achievement at each level and discussed the skills and knowledge that differentiated achievement in each of the four levels. Reviewing the ALDs ensured that participants understood what students in West Virginia are expected to know and be able to do and how much knowledge and skill students are expected to demonstrate at each level of achievement.

3.6.4. Take the Test

Following ALD review, panelists took a form of the test that students took in 2018, in the grade level to 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. Taking the same test as students take provides the opportunity to interact with and become familiar with the test items and the look and feel of the student experience while testing.

3.6.5. OSAB Review

After reviewing the ALDs, 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 described why an assertion might be more or less difficult than a previous assertion. They also noted how each assertion related to the ALDs.

After reviewing the item interactions and scoring assertions individually, panelists engaged in discussion with table members about the skills required and relationships among the reviewed test materials and achievement levels. This process ensured that panelists built a solid understanding of how the scoring assertions relate to the item interactions and how the items relate to the ALDs, and also helped to facilitate a common understanding among workshop panelists.[3]

3.6.6. Assertion Mapping Training

The objective of standard setting is aspirational; to identify what all students should know and be able to do, not what any particular group of students actually knows and can do. Facilitators provided the following process to guide the mapping of assertions onto ALDs:

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

Like the items in the ordered-item booklet (OIB), scoring assertion order within each item was determined by actual student performance.

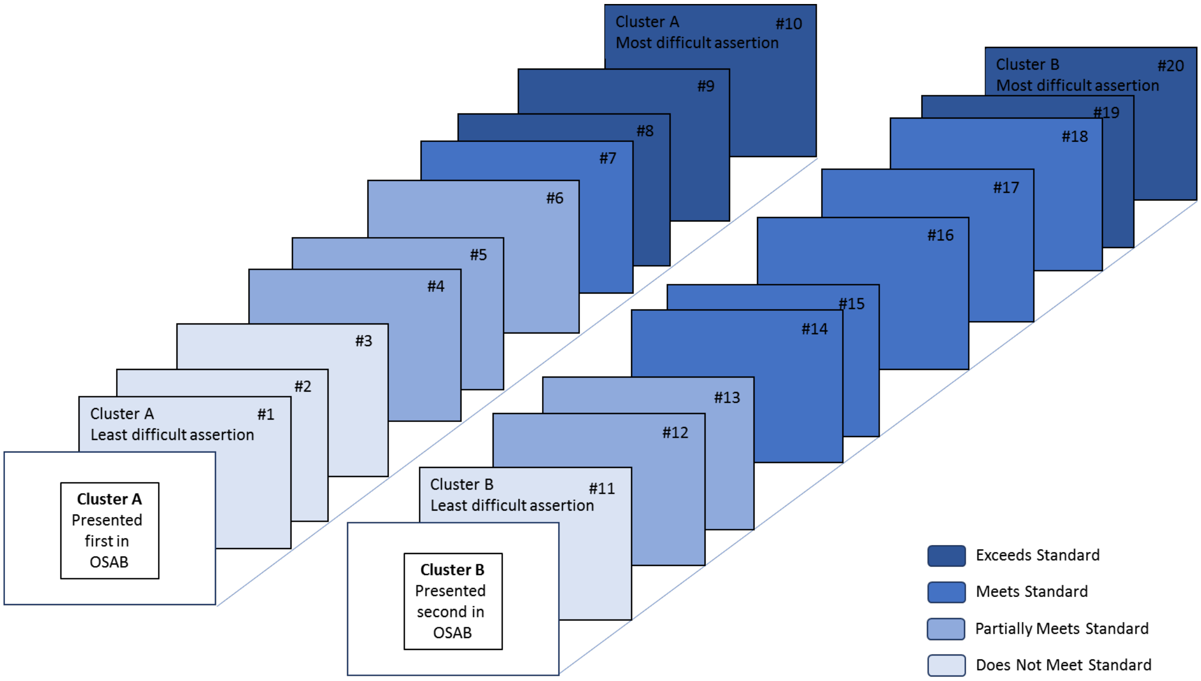
Panelists were to match each assertion to the achievement level best supported by the assertion using the ALDs, an online difficulty visualizer (described in Section 4.5), their notes from the OSAB review, and their professional judgments. Figure 6 graphically describes the assertion mapping process.

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

Because the assertion mapping was done separately for each item, it was possible that there was no perfect ordering of the assigned levels of the assertions across all items as a function of assertion difficulty. It was allowed (and it occurred frequently) that an assertion of one item had a higher difficulty but lower assigned achievement level than another assertion from a different item. For example, in Figure 6, the difficulty of Assertion 3 of cluster A ("Does Not Meet Standard") has a higher difficulty than Assertion 13 of cluster B ("Partially Meets Standard"). However, it was expected for the higher achievement levels to be assigned more frequently with increasing assertion difficulty across items.

[3] One careful panelist was unable to review enough items in the OSAB to map assertions onto ALDs. The data from this participant was excluded from Round 1 and Round 2.

Figure 6. Example Assertion Mapping



Note. Figure 6 describes scoring assertion mapping across two clusters, where assertions 1, 2, 3, and 11 are mapped onto the "Does Not Meet Standard" level, assertions 4, 5, 6, 12, and 13 are mapped onto the "Partially Meets Standard" level, assertions 7, 14, 15, 16, 17, and 18 are mapped onto the "Meets Standard" level, and assertions 8, 9, 10, 19, and 20 are mapped onto the "Exceeds Standard" level.

3.6.7. Readiness Assessment

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

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

Room facilitators reviewed the quizzes with the panelists and provided additional training for incorrect responses on the quiz.

3.6.8. Practice Round

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

3.6.9. Readiness Assertion

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

- The knowledge and skills described by the ALDs, and the skills and interactions that differentiate levels
- The structure, use, and importance of the OSAB
- The process to map assertions from the OSAB onto the ALDs

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 mapping assertions in both rounds of the workshop.

3.6.10. Round 1

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

AIR psychometricians then created cut scores from these 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.

3.6.10.1. Calculating Cut Scores from the Assertion Mapping

A proprietary algorithm utilized RP67 (for grade 5) and RP50 (for grade 8) to minimize misclassifications to calculate cut scores based on the assertion mappings.[4] 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 "Partially Meets Standard" and "Meets Standard" 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

[4] Typically, the probability used in standard setting is .67 ("RP67", Huynh, 1994). RP67 is the item difficulty point where 67% of the students would earn the score point. The reason to adopt RP50 for grade 8 was because most of the items were more difficult than students' abilities. RP50 better aligned with the performance-level descriptor (PLD), and therefore, led to more appropriate achievement cut scores. Using RP50 prevented panelists from setting the first cut score on the lowest-difficulty items on the test. This approach has been taken by other high-stakes tests, such as the Smarter Balanced Assessment Consortium (see Cicek & Koons, 2014)).

students falling into each achievement level. This algorithm calculated cut scores from the assertion maps by panelist, table, and for the room.

3.6.10.2. Feedback Data

Feedback included the cut scores corresponding to the assertion mappings for each panelist, each table, and for the room overall (across all three 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 achievement level to which each panelist assigned the assertion. The tool highlights assertions that panelists have assigned to different achievement levels by the panelists. Room facilitators and panelists reviewed and discussed the assertions with the most variable mappings.

3.6.10.3. Impact Data

Applying the Round 1 cut scores to student data from the 2018 administration of the science WVGSA provided impact data. This showed panelists the projected percentage of students who would fall into each of the achievement levels given the current mapping of assertions.

3.6.10.4. Benchmark Data

The 2015 National Assessment of Educational Progress (NAEP) science scores provided benchmark data, another source of information that panelists could use to evaluate and adjust their assertion mapping. By comparing the results of each round against the percentage proficient on NAEP, it is possible to judge the reasonableness of the proposed achievement standards. NAEP provides state-level data in science for grade 8; benchmark data for grade 5 is interpolated. This provided external evidence of student achievement for panelists to consider when mapping the Round 2 assertions.

Finally, AIR psychometricians described the need for articulated achievement standards and presented cut scores that would maximize articulation for panelist consideration. These were another piece of information for panelists to deliberate.

All feedback and information served to inform, but not determine, their Round 2 decisions. Panelists discussed this information and the impact that the Round 1 cut scores may have on West Virginia students before mapping the Round 2 assertions.

Table 8 presents the achievement standards and associated impact and benchmark data.

Table 8. Round 1 Results

Table	Cut Scores			Impact Data (Percentage At or Above Each Cut Score)			Benchmark Data (2015 NAEP)		
	PM	M	E	PM	M	E	Basic	Proficient	Advanced
Grade 5	535	547	558	88	61	31	72	30	1
1	535	547	558	88	61	31	72	30	1
2	536	552	571	86	48	8	72	30	1
3	555	559	576	38	29	4	72	30	1
Grade 8	832	855	870	92	39	9	63	27	1
1	832	854	868	92	42	12	63	27	1
2	852	860	870	46	27	9	63	27	1
3	836	858	870	86	32	9	63	27	1

Note. The grade-level row summarizes the room data (across the three tables). Impact data describes the projected percentage of students falling at or above each of the achievement levels based on the Round 1 cut scores. Benchmark data describes the percentage at or above each achievement level using data from the 2015 NAEP (interpolated for grade 5). Achievement level abbreviation key: Partially Meets Standard (PM), Meets Standard (M), Exceeds Standard (E).

3.6.11. Round 2

Round 2 began with a discussion of the feedback data from Round 1, beginning with table-level feedback and discussion, and progressing to room-level discussion.

After reviewing the feedback data, workshop facilitators provided panelists with additional instructions for completing Round 2. First, they described the goal of Round 2 as one of convergence, but not consensus, on a common achievement standard. A second goal was to encourage articulation across grade levels.

Each table spent time reviewing and discussing assertion mappings and articulation. After completing these discussions, panelists again worked through the OSAB, placing their Round 2 cut scores for all three achievement levels.

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

Table 9. Round 2 Results

Table	Cut Scores			Impact Data (Percentage At or Above Each Cut Score)			Benchmark Data (2015 NAEP)		
	PM	M	E	PM	M	E	Basic	Proficient	Advanced
Grade 5	537	555	571	83	38	8	72	30	1
1	535	553	571	88	43	8	72	30	1
2	539	551	571	79	51	8	72	30	1
3	534	558	571	89	31	8	72	30	1
Grade 8	837	855	870	83	39	9	63	27	1
1	842	852	868	72	46	12	63	27	1
2	837	860	870	83	27	9	63	27	1
3	836	858	871	86	32	8	63	27	1

Note. The grade-level row summarizes the room data (across the three tables). Impact data describes the projected percentage of students falling at or above each of the achievement levels based on the Round 2 cut scores. Benchmark data describes the percentage at or above each achievement level using data from the 2015 NAEP (interpolated for grade 5). Achievement level abbreviation key: Partially Meets Standard (PM), Meets Standard (M), Exceeds Standard (E).

3.6.12. Moderation

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

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

Panelists receive the information necessary for articulation prior to Round 2. Often, panelists intuitively create well-articulated sets of achievement standards, but sometimes minor changes might significantly improve articulation. Calculated based on panelist recommendations and approved by WVDE, minor changes were offered for consideration to a subset of the panelists after Round 2. After discussion, the moderation panel recommended the achievement standards described in Table 10. Results are displayed graphically in Figure 7 and Figure 8.

Table 10. Moderated Results

Table	Cut Score			Impact Data (Percentage At or Above Each Cut Score)		
	PM	M	E	PM	M	E
Grade 5	537	555	568	83	38	11
Grade 8	837	855	867	83	39	12

Note. Impact data describes the projected percentage of students falling at or above each of the achievement levels based on the final cut scores. Achievement level abbreviation key: Partially Meets Standard (PM), Meets Standard (M), Exceeds Standard (E).

Figure 7. Percentage of Students Reaching or Exceeding Each Achievement Standard

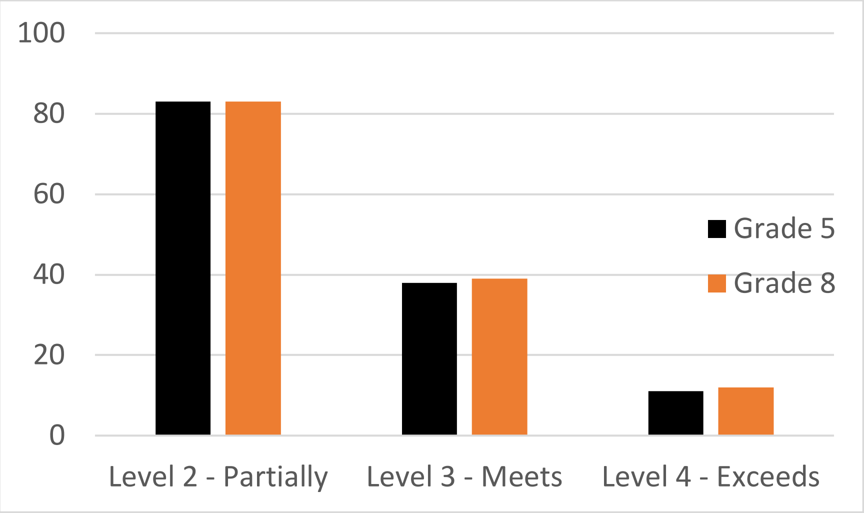
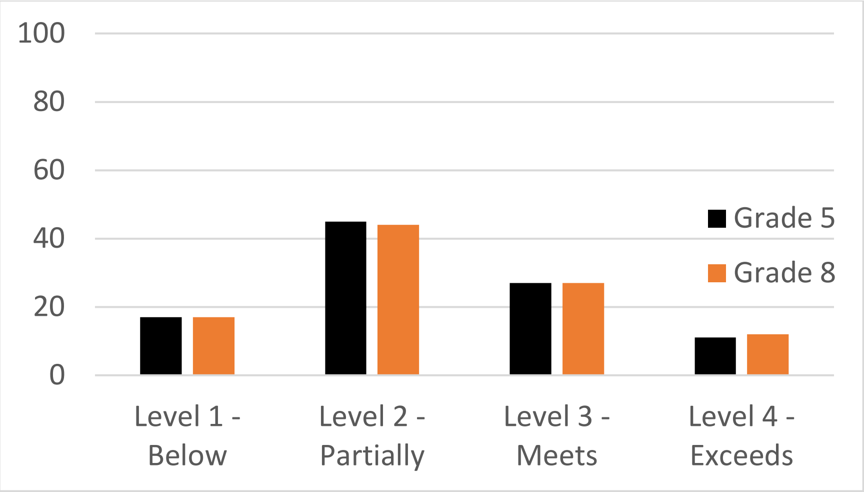


Figure 8 summarizes the percentage of students who would be classified into each achievement level.

Figure 8. Percentage of Students Classified Into Each Achievement Level



3.7. 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 11, Table 12, Table 13, Table 14, and Table 15 summarize the results of the evaluations. Evaluation items endorsed by fewer than 90% of panelists are discussed in text.

Workshop participants overwhelmingly indicated clarity in the instructions, materials, data, and process (see Table 11).

Table 11. Evaluation Results: Clarity of Materials and Process

Please rate the clarity of the following components of the workshop.	Percentage "Somewhat Clear" or "Very Clear"		
	Grade 5	Grade 8	Overall
Instructions provided by the Workshop Leader	93%	100%	97%
Achievement-Level Descriptors (ALDs)	100%	93%	97%
Ordered Scoring Assertion Booklet (OSAB)	100%	100%	100%
Panelist agreement data	100%	100%	100%
Impact data (percentage of students who would reach any standard that you select)	100%	100%	100%

Note. Abbreviation Key: Number of responses = 29. Evaluation options included "Very Clear," "Somewhat Clear," "Somewhat Unclear," and "Very Unclear."

Participants felt they had sufficient time to complete all activities. In fact, some indicated having too much time to complete some tasks (see Table 12). Two panelists indicated having too much and too little time to review ALDs. Three panelists reported having too much time for OSAB review, while one wanted more time, and four panelists indicated having too much time for mapping scoring assertions.

Table 12. Evaluation Results: Appropriateness of Process

How appropriate was the amount of time you were given to complete the following components of the standard-setting process?	Percentage "About Right"		
	Grade 5	Grade 8	Overall
Large group orientation	87%	100%	93%
Experiencing the online assessment	100%	86%	93%
Review of the Achievement-Level Descriptors (ALDs)	87%	86%	86%
Review of the Ordered Scoring Assertion Booklet (OSAB)	87%	86%	86%
Mapping of your scoring assertions in each round	93%	79%	86%
Round 1 discussion	93%	93%	93%

Note. Number of responses = 29. Evaluation options included "About Right," "Too Much," and "Too Little."

Participants appreciated the importance of the multiple factors contributing to assertion mapping, with participants rating each factor as important or very important (Table 13).

Table 13. Evaluation Results: Importance of Materials

How important was each of the following factors in your placement of the scoring assertion mapping decisions?	Percentage "Somewhat Important" or "Very Important"		
	Grade 5	Grade 8	Overall
Achievement-Level Descriptors (ALDs)	100%	100%	100%
Your perception of the difficulty of the items	100%	93%	97%
Your experience with students	93%	100%	97%
Discussions with other panelists	100%	100%	100%
External benchmark data	100%	100%	100%
Room agreement data (room medians and individual mappings of assertions)	93%	100%	97%
Impact data (percentage of students who would reach any standard that you select)	100%	93%	97%

Note. Number of responses = 29. Evaluation options included "Not Important," "Somewhat Important," and "Very Important."

Participant understanding of the workshop processes and tasks was consistently high (see Table 14).

Table 14. Evaluation Results: Understanding Processes and Tasks

At the end of the workshop, please rate your agreement with the following statements.	Percentage “Agree” or “Strongly Agree”		
	Grade 5	Grade 8	Overall
I understood the purpose of this standard-setting workshop.	100%	100%	100%
The procedures used to recommend achievement standards were fair and unbiased.	100%	100%	100%
The training provided me with the information I needed to recommend achievement standards.	100%	100%	100%
Taking the online assessment helped me to better understand what students need to know and be able to do to answer each question.	93%	100%	97%
The Achievement-Level Descriptors (description of what students within each achievement level are expected to know and be able to do) provided a clear picture of expectations for student achievement at each level.	93%	93%	93%
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%	100%	100%
I understood how to place my scoring assertion mapping decisions.	100%	100%	100%
I found the benchmark data and discussions helpful in my decisions about where to place my scoring assertion mapping decisions.	100%	93%	97%
I found the panelist agreement data (room and individual scoring assertion placements) and discussion helpful in my decisions about where to place my scoring assertion mapping decisions.	93%	100%	97%
I found the impact data (percentage of students who would achieve at the level indicated by the OSAB) and discussions helpful in my decisions about where to place my scoring assertion mapping decisions.	100%	93%	97%
I felt comfortable expressing my opinions throughout the workshop.	100%	100%	100%
Everyone was given the opportunity to express his or her opinions throughout the workshop.	100%	100%	100%

Note. Number of responses = 29. 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 15).

Table 15. Evaluation Results: Student Expectations

Please read the following statement carefully and indicate your response.	Percentage "Agree" or "Strongly Agree"		
	Grade 5	Grade 8	Overall
A student performing at Level 3 meets expectations for the grade level.	93%	100%	97%
A student performing at Level 2 is below expectations for the grade level.	93%	100%	97%
A student performing at Level 4 exceeds expectations for the grade level.	93%	100%	97%

Note. Number of responses = 29. Evaluation options included "Strongly Agree," "Agree," "Disagree," and "Strongly Disagree."

3.7.1. Workshop Participant Feedback

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

Fourteen participants responded to the first and second questions. Most responses indicated the training was effective and the process was clear. Participants provided minor suggestions, such as providing an acronym list or shortening the time allocated for some tasks. Many commented on the value of discussions and interactions with other panelists.

Additional participant comments included:

> *"Overall, this was a great experience. I feel like I had good interactions with my group and have a much better understanding of the processes that go into developing testing. I'm glad I had the opportunity to participate and that my questions were answered, and suggestions were considered."*

> *"This was very difficult because I am not a teacher but I think my non-teacher point of view was important."*

> *"The workshop was very well managed!"*

4. VALIDITY EVIDENCE

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

4.1.Evidence of Adherence to Professional Standards and Best Practices

The WVGSA science standard-setting workshop was designed and executed consistent with established practices and best practice principles (Hambleton & Pitoniak, 2006; Hambleton, Pitoniak, & Copella, 2012; Kane, 2001; Mehrens, 1995). The process also adhered to the following professional standards recommended 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 AMP standard-setting procedure is appropriate for tests of this type—with interrelated sets of three-dimensional item clusters and scaled using item response theory (IRT). Section 4.1 provides the justification for and the additional benefits of selecting the AMP method to establish the cut scores; and Sections 4.6 through 4.7.1 document the process followed to implement the method.

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

- understanding the test and test items (from an educator and student perspective),
- describing the knowledge and skills measured by the test,
- identifying the skills associated with each test item,
- describing the skills associated with student performance in each performance level,
- identifying which test items students in each performance level should be able to answer correctly, and
- evaluating and applying feedback and reference data to the Round 2 recommendations and considering the impact of the recommended cut scores on students.

Additionally, panelists' readiness evaluations provided evidence of a successful orientation to the process and understanding of the process, while their workshop evaluations provide evidence of confidence in the process and resulting recommendations.

The recruitment process resulted in panels that were representative of important regional and demographic groups who were knowledgeable about the subject area and students' developmental level. Section 4.3.6 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.

4.2.Evidence in Terms of Peer Review Critical Elements

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

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

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

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

> Critical Element 6.2: Achievement standards setting. The state used a technically sound method and process that involved panelists with appropriate experience and expertise for setting its academic achievement standards and 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 AMP method selected to establish performance standards (Section 4.1)
2) Documentation that the method used for setting cut scores allowed panelists to apply their knowledge and experience in a reasonable manner and supported the establishment of reasonable and defensible cut scores (Section 4.6, 4.7 and 5.1)
3) Panelists self-reported readiness to undertake the task (Section 4.6.7) and confidence in the workshop process and outcomes (Section 4.7) supporting the validity of the process
4) The standard-setting panels consisted of panelists with appropriate experience and expertise, including content experts with experience teaching the West Virginia's science content standards, and individuals with experience and expertise teaching special population and general education students in West Virginia (Section 4.3.6).

5. 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.

Ferrara, S., & Lewis, D. M. (2012). The item-descriptor (ID) matching method. In G. J. Cizek (Ed.), *Setting performance standards. Foundations, methods, and innovations* (2nd ed., pp. 255-282). New York: Routledge.

Gibbons, R.D., & Hedeker, D.R. (1992). Full-information bi-factor analysis. *Psychometrika*, *57*, 423-436.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433-470). Westport, CT: Praeger.

Hambleton, R. K., Pitoniak, M. J., & Copella, J. M. (2012). Essential steps in setting performance standards on educational tests and strategies for assessing the reliability of results. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 47–76). New York, NY: Routledge.

Huynh, H. (2006). A Clarification on the Response Probability Criterion RP67 for Standard Setting 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.

Appendices A and B are available upon request.

To request these documents, please contact the Office of Assessment:

West Virginia Department of Education
Office of Assessment
1900 Kanawha Blvd
Charleston, WV
(304) 558-2546