North Dakota State Assessment for Science

2021–2022

Volume 3: Setting Achievement Standards

NORTH DAKOTA DEPARTMENT OF
PUBLIC INSTRUCTION

TABLE OF CONTENTS

List of Tables

LIST OF FIGURES

LIST OF APPENDICES

1. EXECUTIVE SUMMARY

In February 2019, the North Dakota Department of Public Instruction (NDDPI) adopted the new North Dakota Science Content Standards. The new standards employ a three-dimensional conceptualization of science understanding, including science and engineering practices, crosscutting concepts, and disciplinary core ideas. With the adoption of the new Standards, and the development of new statewide assessments to measure achievement of those standards, the NDDPI convened a standard-setting workshop to recommend a system of achievement standards for determining whether students have met the learning goals defined by the 2019 North Dakota Science Content Standards.

Under contract to NDDPI, Cambium Assessment, Inc. (CAI) conducted the standard-setting workshop to recommend achievement standards for the North Dakota State Assessment (NDSA) for Science in grades 4, 8, and 10. The workshop was conducted remotely on June 16– 17, 2021.

North Dakota's science assessments are designed to measure the attainment of the 2019 North Dakota Science Content Standards adopted by the NDDPI. The assessments are made up of item clusters and stand-alone items. Item clusters represent a series of interrelated student interactions directed toward describing, explaining, and predicting scientific phenomena. Stand-alone items are added to increase the coverage of the test while limiting increases in testing time and burden on students and schools. Test items were developed by CAI, in conjunction with a group of states working to implement three-dimensional science standards. Test items were developed to ensure that each student is administered a test meeting all elements of the NDSA for Science blueprints, which were constructed to align with the 2019 North Dakota Science Content Standards.

North Dakota science educators, serving as standard-setting panelists, followed a rigorous standardized 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 & Lewis, 2012). Consistent with ordered-item procedures in general (e.g., Mitzel, Lewis, Patz, & Green, 2001), workshop panelists reviewed and recommended achievement standards using an ordered set of scoring assertions derived from student interactions within items. Because the new science items—specifically the item clusters—represent multiple, interdependent interactions through which students engage in scientific phenomena, scoring assertions cannot be meaningfully evaluated independently of the item interactions from which they are derived. Thus, panelists were presented with ordered scoring assertions for each item separately rather than for the test overall. Panelists mapped each scoring assertion to the most apt achievement-level descriptor (ALD).

Panelists reviewed ALDs describing the degree to which students have achieved the 2019 North Dakota Science Content Standards. The NDDPI reviewed and revised range ALDs before the standard-setting workshop. After reviewing the range ALDs, standard-setting panelists worked to identify the knowledge and skills characteristic of students just qualifying for entry into each achievement level.

Working through the ordered scoring assertions for each item, panelists mapped each assertion into one of the four achievement levels—Level 1: Novice, Level 2: Partially Proficient, Level 3: Proficient, and Level 4: Advanced. The mapping of scoring assertions was based on the consideration of test content. Panelists were provided additional contextual information, including the percentage of students who performed at or above the achievement level associated with each assertion (impact data), as well as the projected National Assessment of Educational Progress (NAEP) science achievement level corresponding to each assertion. The panelists performed the assertion mapping in two rounds of standard setting. Panelists' mapping of the scoring assertions was used to identify the location of the three achievement standards used to classify student achievement— Partially Proficient, Proficient, and Advanced. Following Round 1, panelists were provided with feedback about the mappings of their fellow panelists and discussed their mappings as a group. 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 assessment systems. A modification to the Partially Proficient achievement standard was recommended for grade 8 during the moderation session.

Thirty-one North Dakota science educators were selected to serve as science standard-setting panelists, with 10 participants serving on the grade 4 panel, 10 participants on the grade 8 panel, and 11 participants on the grade 10 panel. The panelists represented experienced classroom teachers and curriculum specialists as well as district administrators and other stakeholders. The composition of the panel ensured that a diverse range of perspectives and deep experience with the three-dimensional 2019 North Dakota Science Content Standards contributed to the 2021 standard-setting process.

1.1 STANDARD-SETTING WORKSHOP

1.1.1 Overall Workshop Structure

The workshop included the following key features:

- The standard-setting procedure produced three recommended achievement standards (Partially Proficient, Proficient, and Advanced) that will be used to classify student achievement on the NDSA for Science.
- Panelists recommended achievement standards in two rounds.
- Contextual information—including the percentage of students who performed at or above the achievement level associated with each individual assertion (impact data)—was provided to panelists during Round 1 of standard setting.
- Benchmark information—including the projected NAEP science achievement level corresponding to each assertion—was provided to panelists as part of the Round 1 results.
- The standard-setting workshop was conducted using CAI's online standard-setting tool. Because the workshop was conducted remotely, each panelist accessed the tool using their personal computer.
- Following Round 2 of standard setting, panelists engaged in a moderation session, during which they reviewed and modified the recommended achievement standards to achieve an articulated system of standards across grades and assessment systems. A modification to

the Partially Proficient achievement standard was recommended for grade 8 during the moderation session.

1.1.2 Standard-Setting Workshop Results

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

Table 1. Achievement Standards Recommended for Science

Grade	Level 2 Partially Proficient	Level 3 Proficient	Level 4 Advanced
4	380	407	431
8	775	802	835
10	973	1000	1035

Table 2 indicates the percentage of students that will reach or exceed each achievement standard in 2021. Figure 1 represents those values graphically.

Table 2. Percentage of Students Reaching or Exceeding Each Recommended Science Achievement Standard in 2021

Grade	Level 2 Partially Proficient	Level 3 Proficient	Level 4 Advanced
4	76	41	14
8	80	51	10
10	82	50	10

Figure 1. Percentage of Students Reaching or Exceeding Each Recommended Science Achievement Standard in 2021

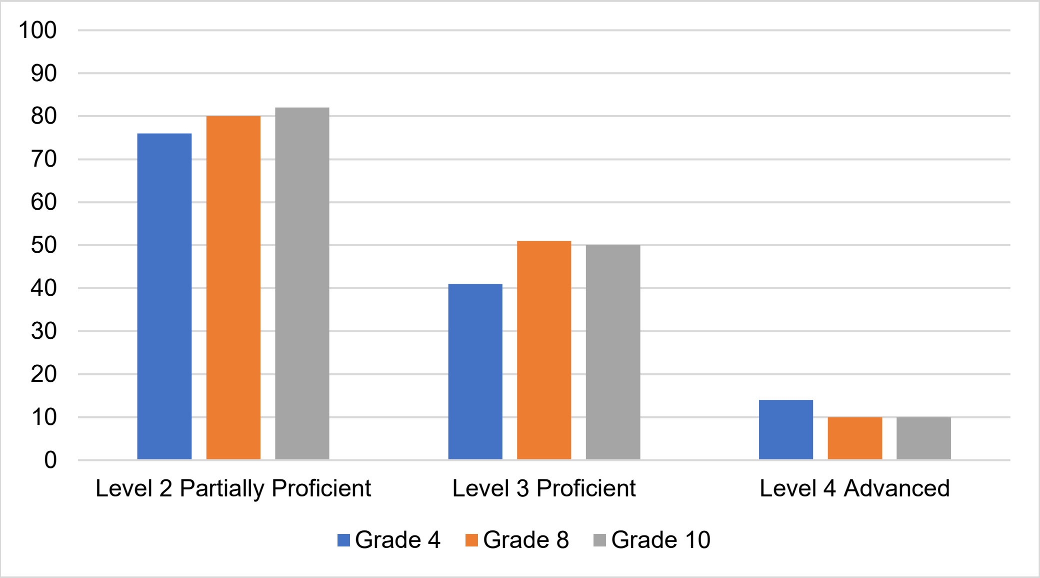
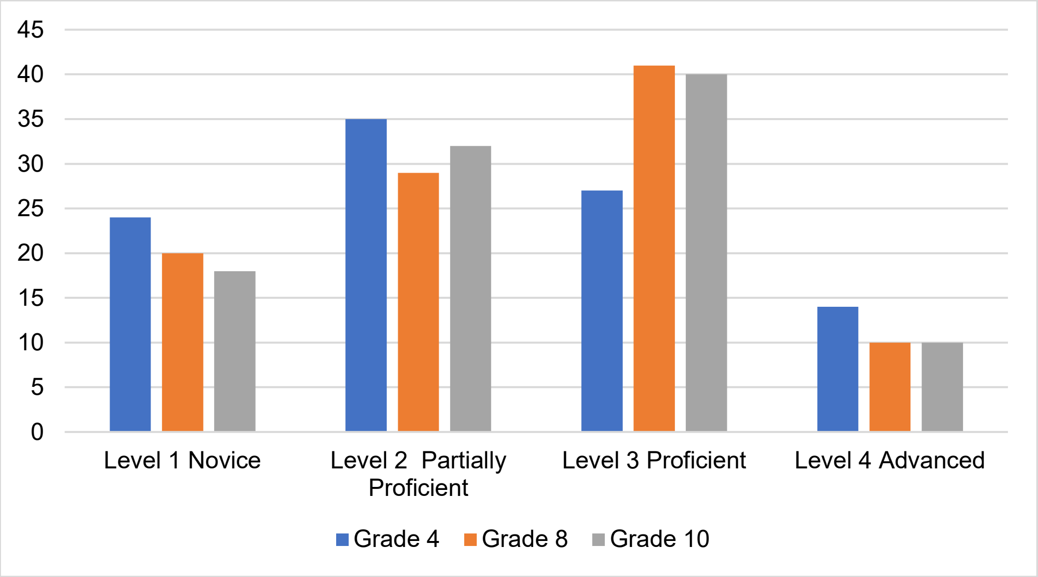


Table 3 indicates the percentage of students classified as having attained each of the achievement levels in 2021. The values are displayed graphically in Figure 2.

Table 3. Percentage of Students Classified Within Each Science Achievement Level in 2021

Grade	Level 1 Novice	Level 2 Partially Proficient	Level 3 Proficient	Level 4 Advanced
4	24	35	27	14
8	20	29	41	10
10	18	32	40	10

Figure 2. Percentage of Students Classified Within Each Science Achievement Level in 2021



2. Introduction

North Dakota adopted three-dimensional science standards as the new 2019 North Dakota Science Content Standards in February of 2019. The North Dakota Department of Public Instruction (NDDPI) and its assessment vendor, Cambium Assessment, Inc. (CAI), developed and administered a new assessment to measure the new standards. In spring 2021, NDDPI and CAI administered to all grade 4, 8, and 10 students in North Dakota new assessments aligned to the three-dimensional science standards.

North Dakota provides information about the science assessments on the NDDPI website at https://www.nd.gov/dpi/districtsschools/assessment/ndsa.

New tests require new achievement standards to link achievement on the test to the content standards. NDDPI contracted with CAI to establish cut scores for the new tests. To fulfill this responsibility, CAI implemented an innovative, defensible, valid, and technically sound method of standard setting; provided training on standard setting to all workshop participants; oversaw the workshop and the standard-setting process; computed real-time feedback data to inform the process; and produced a technical report documenting the method, approach, process, and outcomes. Achievement standards were recommended for grades 4, 8, and 10 in June 2021.

The purpose of this documentation is to detail the standard-setting process for the North Dakota State Assessment (NDSA) for Science and the resulting achievement-standard recommendations.

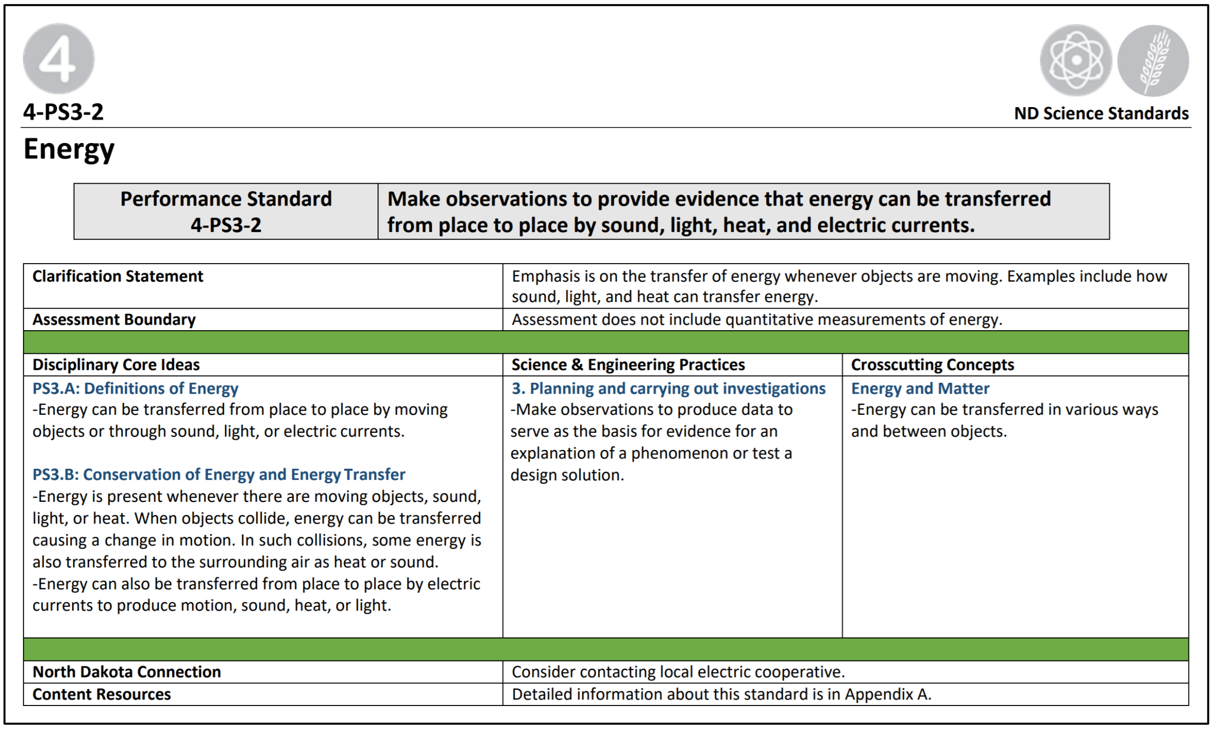
3. THE 2019 NORTH DAKOTA SCIENCE CONTENT STANDARDS

The North Dakota State Assessment (NDSA) for Science assesses the learning objectives described by the North Dakota Science Content Standards, adopted by North Dakota in 2019. Information about the 2019 North Dakota Science Content Standards is available online at https://www.nd.gov/dpi/districtsschools/k-12-education-content-standards.

The three-dimensional science standards, which are based on *A Framework for K–12 Science Education* (National Research Council, 2012), reflect the latest research and advances in modern science education and differ from previous science standards in multiple ways. First, rather than describe general knowledge and skills that students should know and be able to do, they describe specific performances that demonstrate what students know and can do. The North Dakota Science Content Standards refer to these performed knowledge and skills as *performance standards*. Second, while unidimensionality is a typical goal of standards (and the items that measure them), the 2019 North Dakota Science Content Standards are intentionally multi-dimensional. Each performance standard incorporates all three dimensions from *A Framework for K–12 Science Education* (National Research Council, 2012): a science or engineering practice, a disciplinary core idea, and a crosscutting concept. Another unique feature of the North Dakota Science Content Standards is the assumption that students should learn all science disciplines, rather than select a few, as is traditionally the case in many high schools, where students may elect, for example, to take biology and chemistry but not physics or astronomy.

Figure 3 shows the structure of the 2019 North Dakota Science Content Standards for a single grade 4 performance standard, 4-PS3-2.

Figure 3. Structure of the 2019 North Dakota Science Content Standard for One Performance Standard



4

4-PS3-2

ND Science Standards

Energy

Performance Standard 4-PS3-2	Make observations to provide evidence that energy can be transferred from place to place by sound, light, heat, and electric currents.

Clarification Statement	Emphasis is on the transfer of energy whenever objects are moving. Examples include how sound, light, and heat can transfer energy.
Assessment Boundary	Assessment does not include quantitative measurements of energy.

Disciplinary Core Ideas	Science & Engineering Practices	Crosscutting Concepts
PS3.A: Definitions of Energy -Energy can be transferred from place to place by moving objects or through sound, light, or electric currents. **PS3.B: Conservation of Energy and Energy Transfer** -Energy is present whenever there are moving objects, sound, light, or heat. When objects collide, energy can be transferred causing a change in motion. In such collisions, some energy is also transferred to the surrounding air as heat or sound. -Energy can also be transferred from place to place by electric currents to produce motion, sound, heat, or light.	**3. Planning and carrying out investigations** -Make observations to produce data to serve as the basis for evidence for an explanation of a phenomenon or test a design solution.	**Energy and Matter** -Energy can be transferred in various ways and between objects.

North Dakota Connection	Consider contacting local electric cooperative.
Content Resources	Detailed information about this standard is in Appendix A.

Source.
https://www.nd.gov/dpi/sites/www/files/documents/Academic%20Support/FINAL%20ND%20Science%20Content%20Standards_rev2.12.10.19.pdf.

4. The North Dakota State Assessment for Science

Due to the unique features of the three-dimensional 2019 North Dakota Science Content Standards, items and tests based on these standards, such as the North Dakota State Assessment (NDSA) for Science, must also incorporate similarly unique features. The most impactful of these changes is that new science tests are multi-dimensional and are thus made up mostly of *item clusters* representing a series of interrelated student interactions directed toward describing, explaining, and predicting scientific phenomena.

4.1 Item Clusters and Stand-Alone Items

Item clusters include a stimulus and a series of questions that generally take students approximately 6–12 minutes to complete. They consist of a phenomenon—an observable fact or design problem—that an engaged student explains, models, investigates, or designs using the knowledge and skill described by the performance standard to complete a series of activities (made up of multiple interactions). For example, in Figure 3, proficiency in this single performance standard requires activities that demonstrate the ability to make observations to provide evidence that energy can be transferred from place to place by sound, light, heat, and electric currents. The stimulus in an item cluster explicitly states a task or goal (for example, “In the questions that follow, you will analyze what happens to the train when the brakes are applied”), and subsequent interactions build on or relate to the task or to the responses to previous questions. The interactions within an item cluster all address the same phenomenon.

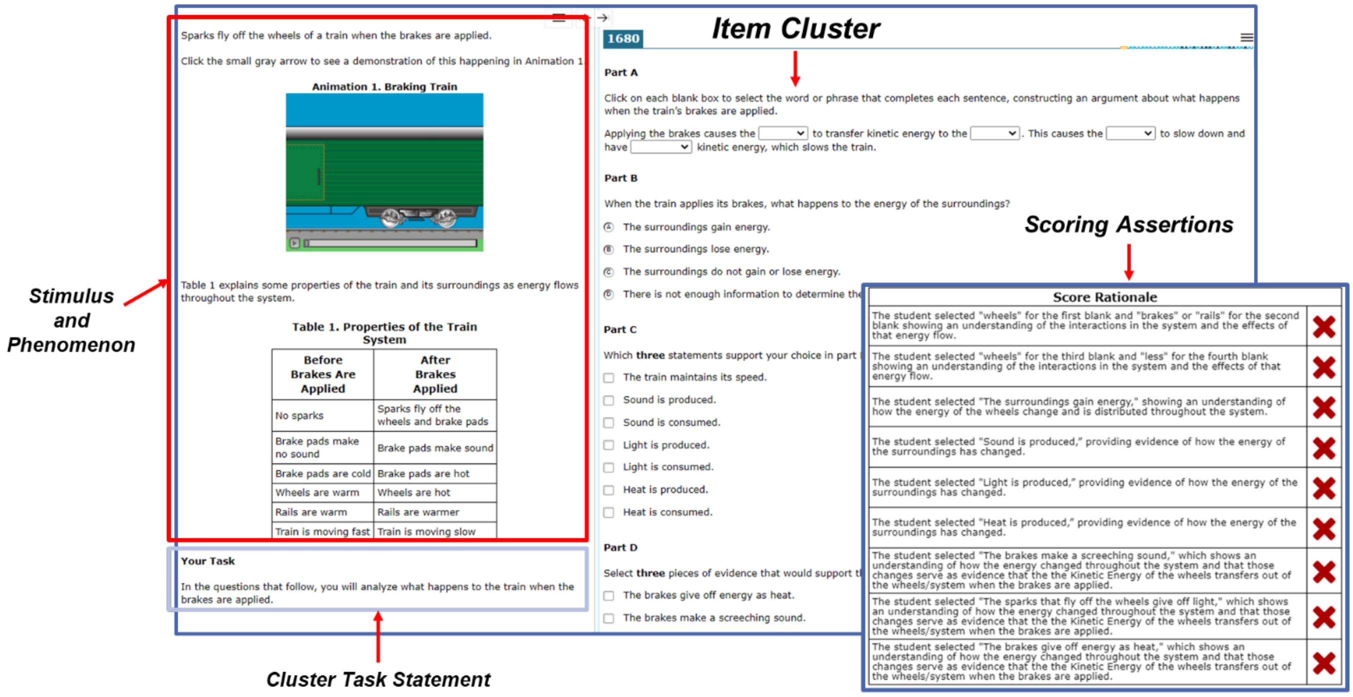
Some added stand-alone items increase the coverage of the test without also increasing testing time or testing burden. Stand-alone items are shorter, are unrelated to other items, and generally take students one to three minutes to complete. Within each item cluster, there is a variety of interaction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

4.2 Scoring Assertions

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

Figure 4 illustrates an item cluster and associated scoring assertions.

Figure 4. Example of Three-Dimensional Science Item Cluster and Scoring Assertions



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

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

Animation 1. Braking Train

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

Table 1. Properties of the Train System

Before Brakes Are Applied	After Brakes Applied
No sparks	Sparks fly off the wheels and brake pads
Brake pads make no sound	Brake pads make sound
Brake pads are cold	Brake pads are hot
Wheels are warm	Wheels are hot
Rails are warm	Rails are warmer
Train is moving fast	Train is moving slow

Your Task

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

1680

Part A

Click on each blank box to select the word or phrase that completes each sentence, constructing an argument about what happens when the train's brakes are applied.

Applying the brakes causes the [▾] to transfer kinetic energy to the [▾]. This causes the [▾] to slow down and have [▾] kinetic energy, which slows the train.

Part B

When the train applies its brakes, what happens to the energy of the surroundings?

Ⓐ The surroundings gain energy.

Ⓑ The surroundings lose energy.

Ⓒ The surroundings do not gain or lose energy.

Ⓓ There is not enough information to determine the

Part C

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

□ The train maintains its speed.

□ Sound is produced.

□ Sound is consumed.

□ Light is produced.

□ Light is consumed.

□ Heat is produced.

□ Heat is consumed.

Part D

Select **three** pieces of evidence that would support t

□ The brakes give off energy as heat.

□ The brakes make a screeching sound.

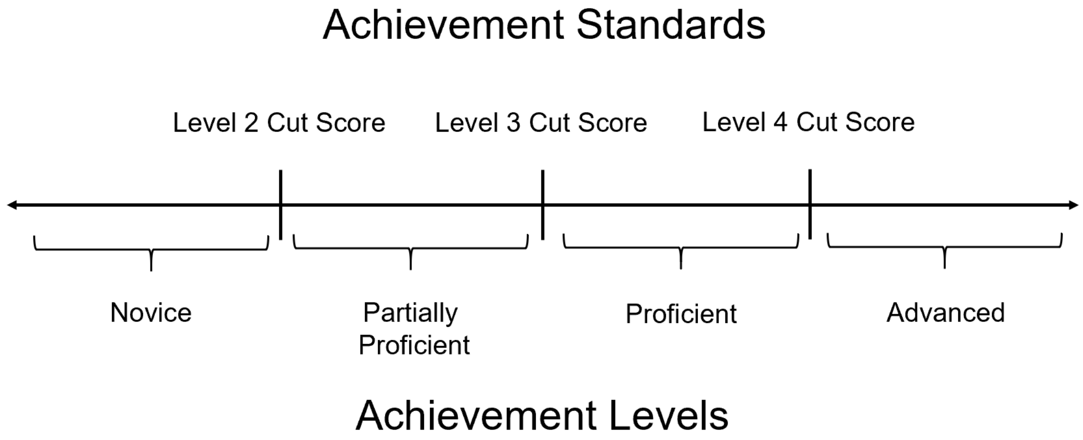
Score Rationale	
The student selected "wheels" for the first blank and "brakes" or "rails" for the second blank showing an understanding of the interactions in the system and the effects of that energy flow.	✗
The student selected "wheels" for the third blank and "less" for the fourth blank showing an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

5. Standard Setting

Thirty-one educators from North Dakota convened remotely on June 16–17, 2021, to complete two rounds of standard setting to recommend three achievement standards for the North Dakota State Assessment (NDSA) for Science.

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

Figure 5. Three Achievement Standards Defining North Dakota's Four Achievement Levels



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

5.1 The Assertion-Mapping Procedure

A new approach to standard setting was necessary for the North Dakota State Assessment (NDSA) for Science due to the structure of the content standards and, subsequently, the structure of the test items assessing the standards. The 2019 North Dakota Science Content Standards adopted a three-dimensional conceptualization of science understanding, including science practices, crosscutting concepts, and disciplinary core ideas. Accordingly, the new NDSA for Science tests were comprised mostly of item clusters representing a series of interrelated student interactions directed toward describing, explaining, and predicting scientific phenomena. Some stand-alone items were added to increase the coverage of the test without also increasing testing time or testing burden.

Within each item, a series of explicit assertions were made regarding the knowledge and skills that a student demonstrated based on specific features of the student's responses across multiple interactions. For example, students may have correctly graphed data points, indicating that they could construct a graph showing the relationship between two variables, but may have made an incorrect inference regarding the relationship between the two variables, thereby failing to support the assertion that they could interpret relationships expressed graphically.

While some other assessments, especially English language arts (ELA), comprise items probing a common stimulus, the degree of interdependence among such items is limited; and student performance on such items can be evaluated independently of student achievement on other items within the stimulus set. This is not the case with the new science items, which could, for example, involve multiple steps in which students interact with the products of previous steps. However, unlike 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 could be added to the item response theory (IRT) model. Typically, these are nuisance dimensions unrelated to student ability. Examples of IRT models that follow this approach are the bi-factor model (Gibbons & Hedeker, 1992) and the testlet model (Bradlow, Wainer, & Wang, 1999). The testlet model is a special case of the bi-factor model (Rijmen, 2010).

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

To address these challenges, CAI psychometricians designed a new method for setting achievement standards on cluster-based assessments. CAI implemented this method for the New Hampshire, Utah, and West Virginia statewide assessments in 2018; for the Connecticut, Oregon, and the joint Multi-State Science Assessment (MSSA) for Rhode Island and Vermont in 2019; and for the South Dakota, Hawaii, and Utah statewide assessments in 2021.

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

The main distinction between AMP and existing ordered-item procedures (e.g., Mitzel, Lewis, Patz, & Green, 2001) is that the panelists evaluate scoring assertions rather than individual items. Scoring assertions are not test items but inferences that are supported (or not supported) by students' responses in one or more interactions within an item cluster or stand-alone item. Because item clusters represent multiple, interdependent interactions through which students engage in scientific phenomena, scoring assertions cannot be meaningfully evaluated independently of the item from which they are derived. Therefore, the scoring assertions from the same item cluster or stand-alone item are always presented together. Within each item cluster or stand-alone item, scoring assertions are ordered by difficulty (i.e., the IRT difficulty parameter) consistent with ordered-item procedures. One can think of the resulting booklet as consisting of different chapters, where each chapter represents an item cluster or stand-alone item. Within each chapter, the (ordered) pages represent scoring assertions. As in ID matching, panelists are asked to map each scoring assertion to the most apt achievement-level descriptor (ALD) during two rounds of standard setting. As with the Bookmark method, assertion mappings are made independently with the goal of convergence rather than consensus over two rounds of rating.[1]

[1] CAI historically implements two rounds of standard setting as best practice in the Bookmark method and extends this practice to the AMP method. In addition to lessening the panelists' burden of needing to repeat a cognitively

5.2 Workshop Structure

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

Figure 6. Workshop Panels, Per Room

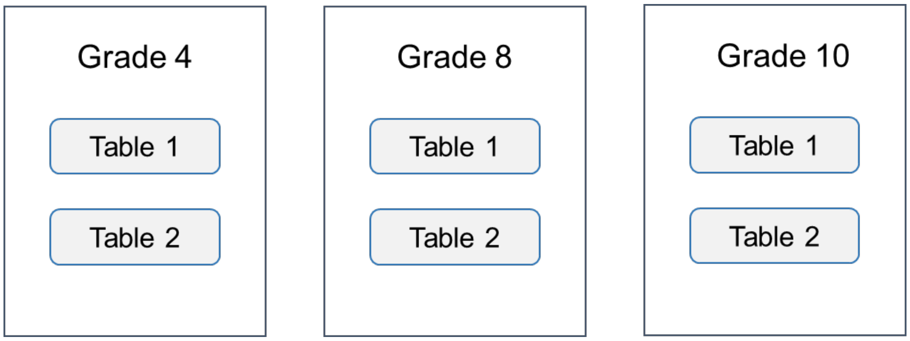


Table 4 summarizes the composition of the tables and the number of facilitators and panelists assigned to each. The 31 standard-setting participants included table leaders and panelists from North Dakota who taught in the content area and grade for which standards were being set.

Table 4. Table Assignments

Room	Grade	Tables and Table Leaders (One per Table)	Panelists (per Table)	Facilitator	Facilitator Assistant
1	4	2	5 / 5	Meg McMahon Heather MacRae	Azza Hussein Ethan Yosebashvili
2	8	2	5 / 5	Kevin Dwyer Vanessa Johnson	Jennifer Chou Maggie Lindsay
3	10	2	6 / 5	Matthew Davis Kam Mangis de Mark	Tracy Graham Rehan Mustafa

5.3 Participants and Roles

5.3.1 North Dakota Department of Public Instruction Staff

Staff from the North Dakota Department of Public Instruction (NDDPI) were present throughout the process and provided overall policy context and answered any policy questions that arose.

From NDDPI, attendees included:

demanding task for a third time, using two rounds introduces significant cost efficiency by reducing the number of days needed for standard setting. Panels typically converge in Round 2, and panelists completing two rounds report levels of confidence in the outcomes that are similar to the confidence expressed by panelists participating in three rounds. Psychometric evaluation of the reliability and variability in results from two and three rounds are generally consistent. CAI has used two rounds in standard setting in more than 17 states and 38 assessments, beginning in 2001 with the enactment of the No Child Left Behind (NCLB) Act.

- Stanley Schauer, Director of Assessment Office
- Bonnie Weisz, Assistant Director of Assessment Office
- Karla Egan, North Dakota Technical Advisory Committee

5.3.2 Cambium Assessment, Inc. Staff

CAI facilitated the workshop and were available in each of the content-area rooms, provided psychometric and statistical support, and oversaw technical set-up and logistics. CAI team members were highly qualified to lead the workshop and conduct analyses, and included the following personnel:

- Dr. Stephan Ahadi, Managing Director of Psychometrics, facilitated and oversaw all AMP processes and tasks and provided training to participants.
- Dr. Frank Rijmen, Senior Director of Psychometrics, supervised all psychometric analyses conducted during and after the workshop.
- Dr. Widad Abdalla, Psychometrician, provided psychometric analyses.
- Alesha Ballman, Psychometric Project Coordinator, oversaw analytics technology and psychometrics.
- Azza Hussein and Ethan Yosebashvili, Psychometric Support Assistants, provided support as needed.
- Maggie Lindsay, Jennifer Chou, Caroline Lempres, Marie Musumeci, and Rehan Mustafa, Program Management Team, managed the process and logistics throughout the meeting.
- Floyd Helm, Nicholas Brennan, and Jesse Justiniano, System Support Agents, troubleshot technology during the workshop.

5.3.3 Room Facilitators

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

- Meg McMahon and Heather MacRae facilitated the grade 4 panel.
- Kevin Dwyer and Vanessa Johnson facilitated the grade 8 panel.
- Matthew Davis and Kam Mangis de Mark facilitated the grade 10 panel.

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

5.3.4 Educator Participants

To establish achievement standards, the NDDPI recruited participants from across the state. Panelists included science teachers, administrators, and representatives from other stakeholder groups (e.g., coaches, college faculty) to ensure that a range of perspectives contributed to the standard-setting process and products. In recruiting panelists, the NDDPI targeted participants who would be representative of the gender and geographic demographics of North Dakota's teacher population. All participants also had to be familiar with the 2019 North Dakota Science Content Standards and test.

The NDDPI selected classroom teachers from the potential panelist pool and invited them to participate in the workshop. Overall, the standard-setting workshop panelists were 23% male and 10% non-white; they represented stakeholder groups that included special education teachers, coaches, curriculum coordinators, administrators, and higher education faculty or administrators, with general education teachers comprising 94% of the panels overall. The majority of panelists taught in the grades to which they were assigned to set standards. Overall, 26% of panelists taught elementary school, 23% taught middle school, and 16% taught high school; the remainder taught some combination of grades. Most panelists worked in schools (90%), although some worked in districts (10%). Districts included rural (52%), suburban (23%), and urban (26%), and were small (42%), medium (19%), and large (35%). Table 5 summarizes the characteristics of the panels.

Table 5. Panelist Characteristics

	Percentage of Panelists, by Panel			
	Science Grade 4	*Science Grade 8*	*Science Grade 10*	*Overall*
Characteristics				
Male	0%	30%	36%	23%
Non-White	0%	20%	9%	10%
Stakeholder Groups[a]				
General Education Teacher	90%	100%	91%	94%
Special Education Teacher	10%	0%	0%	3%
Coach	10%	0%	0%	3%
Curriculum Coordinator	0%	0%	9%	3%
Administrator	0%	0%	9%	3%
Higher Education Teacher	10%	0%	0%	3%
Other	10%	0%	0%	3%
Current Position[b]				
School	90%	100%	82%	90%
District	0%	0%	27%	10%
Other[c]	10%	0%	0%	3%
District Size				
Large	40%	40%	27%	35%

	Percentage of Panelists, by Panel			
	Science Grade 4	*Science Grade 8*	*Science Grade 10*	*Overall*
Medium	10%	30%	18%	19%
Small	40%	30%	55%	42%
Not applicable	10%	0%	0%	3%
District Urbanicity				
Urban	30%	30%	18%	26%
Suburban	30%	30%	9%	23%
Rural	40%	40%	73%	52%
Primary Grades Taught				
ES (grades 1–5)	80%	0%	0%	26%
MS (grades 6–8)	0%	70%	0%	23%
HS (grades 9–12)	0%	0%	45%	16%
ES and MS (grades 1–8)	0%	0%	0%	0%
MS and HS (grades 6–12)	0%	20%	45%	23%
ES, MS, and HS (all grades)	0%	10%	0%	3%
MS and College	10%	0%	0%	3%
College	10%	0%	9%	6%

[a] The total sums to over 100% for "Stakeholder Groups" as many participants held multiple roles.
[b] The total sums to over 100% for "Current Position" as one panelist reported the location of their current position as both school and district.
[c] Other Current Position includes College.

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 NDSA for Science standard-setting workshop were highly qualified. They brought a variety of experience and expertise. Overall, 61% of panelists had earned a master's degree or higher. Many had taught for more than 10 years, and just over 25% had professional experience outside the classroom. Ninety-seven percent of panelists taught science, and many taught other subjects, too. The average time teaching the 2019 North Dakota Science Content Standards was nearly two years. Over 60% of each panel had experience teaching special populations, such as those eligible to receive free or reduced-price lunch (87% overall), English learners (61% overall), and students on Individualized Education Plans (90% overall). Table 6 summarizes the qualifications of the panels.

Table 6. Panelist Qualifications

Qualifications	Percentage of Panelists, by Panel			
	Science Grade 4	*Science Grade 8*	*Science Grade 10*	*Overall*
Highest Degree				
Bachelor	60%	20%	36%	39%

Qualifications	Percentage of Panelists, by Panel			
	Science Grade 4	*Science Grade 8*	*Science Grade 10*	*Overall*
Master	30%	60%	45%	45%
Doctoral	10%	20%	18%	16%
Years Teaching Experience				
None	0%	0%	0%	0%
Less than 1 year	0%	0%	0%	0%
1–5 years	20%	10%	9%	13%
6–10 years	20%	40%	18%	26%
11–15 years	40%	20%	18%	26%
16–20 years	10%	10%	27%	16%
More than 20 years	10%	20%	27%	19%
Years Teaching Experience in Assigned Grade				
None	10%	0%	0%	3%
Less than 1 year	20%	0%	9%	10%
1–5 years	20%	20%	18%	19%
6–10 years	0%	40%	9%	16%
11–15 years	40%	10%	9%	19%
16–20 years	10%	20%	27%	19%
More than 20 years	0%	10%	27%	13%
Subject Areas Currently Teaching[a]				
English Language Arts (ELA)	80%	0%	0%	26%
Mathematics	80%	10%	0%	29%
Social Studies	80%	0%	0%	26%
Science	90%	100%	100%	97%
Other[b]	10%	10%	0%	6%
Other professional experience in education	30%	20%	27%	26%
Years Professional Experience in Education				
None	70%	80%	73%	74%
Less than 1 year	0%	0%	0%	0%
1–5 years	10%	20%	9%	13%
6–10 years	10%	0%	0%	3%
11–15 years	0%	0%	9%	3%
16–20 years	10%	0%	9%	6%
More than 20 years	0%	0%	0%	0%
Experience Teaching Special Student Populations				
Students eligible to receive free/reduced price lunch	70%	100%	91%	87%

Qualifications	Percentage of Panelists, by Panel			
	Science Grade 4	*Science Grade 8*	*Science Grade 10*	*Overall*
English learners (ELs)	50%	80%	55%	61%
Students on an Individualized Education Plan (IEP)	80%	100%	91%	90%
Average years teaching the 2019 North Dakota Science Content Standards	1	3	2	2

[a] The total sums to more than 100% for Subject Areas Currently Teaching, as many participants taught multiple subjects.

[b] Other categories in Subject Areas Currently Teaching include health and special education.

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

5.3.5 Table Leaders

Volunteers from the participant pool served as table leaders. In addition to serving as panelists and mapping assertions, table leaders participated in the moderation session.

5.4 Materials

5.4.1 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, the NDDPI must adopt a similar system of achievement, or 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 content 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. **Threshold 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. They may also be called Target ALDs or Just Barely ALDs.
4. **Reporting ALDs.** These are much-abbreviated ALDs (typically 350 or fewer characters) created following state approval of the achievement standards used to describe student achievement on score reports.

North Dakota uses four achievement levels to describe student achievement: Novice, Partially Proficient, Proficient, and Advanced. At the policy level, these achievement levels are defined as follows:

- **Novice.** Students who achieve at this level demonstrate initial understanding of knowledge and skills specific to the North Dakota Science Content Standards. The student generally performs significantly below the standard for the grade level, is likely able to partially access grade-level content, and engages with higher-order thinking skills with extensive support.
- **Partially Proficient.** Students who achieve at this level demonstrate minimal understanding of the knowledge and skills specific to the North Dakota Science Content Standards. The student generally performs slightly below the standard for the grade level, is likely able to access grade-level content, and engages in higher-order thinking skills with some independence and support.
- **Proficient.** Students who achieve at this level demonstrate satisfactory understanding of knowledge and skills specific to the North Dakota Science Content Standards. The student generally performs at the standard for grade-level content, is likely able to access above grade-level content, and engages in higher-order thinking skills with some independence and support.
- **Advanced.** Students who achieve at this level demonstrate advanced understanding of knowledge and skills specific to the North Dakota Science Content Standards. The student generally performs above the standard for grade-level content, can access above grade-level content, and engages in higher-order thinking skills independently.

Development of Science Range Achievement-Level Descriptor

The NDDPI drafted range ALDs that describe observable evidence for what student performance looks like in science at each achievement level and grade. The NDDPI and CAI reviewed the draft range ALDs to ensure that the language accurately represented the goals and policies of the state. CAI worked with NDDPI to make revisions where necessary.

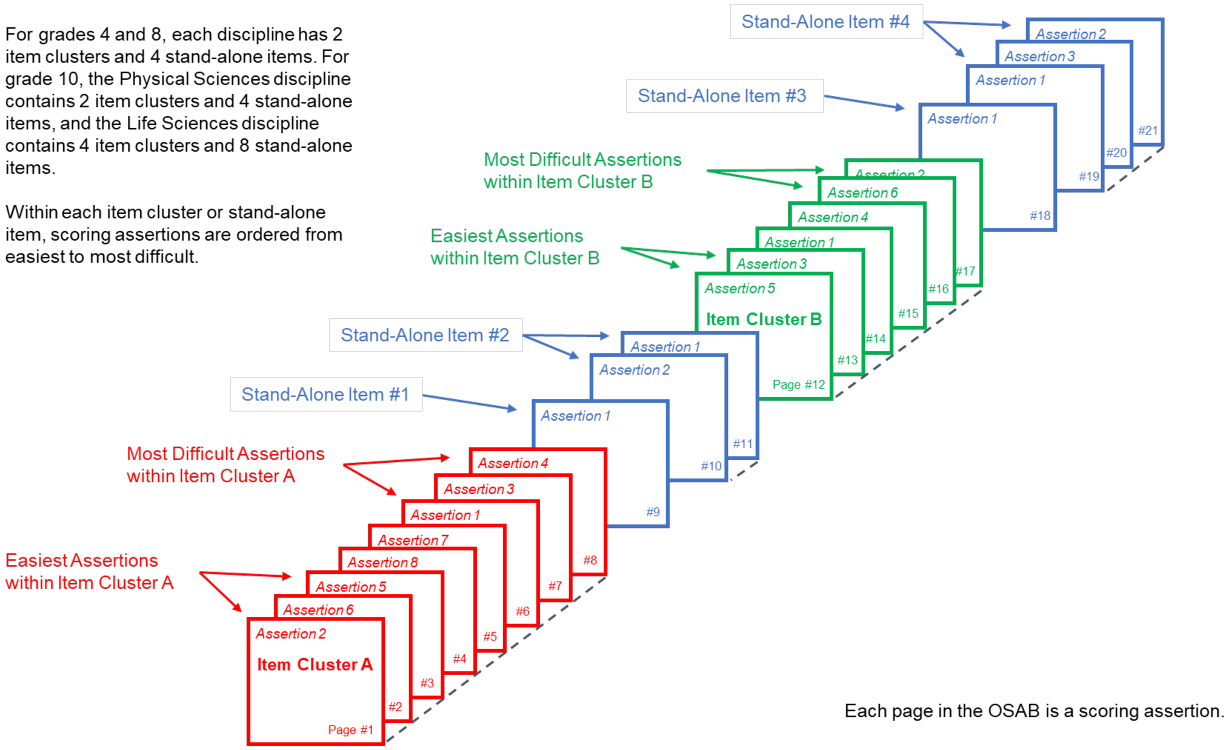
Prior to the standard-setting workshop, a group of experienced North Dakota educators, familiar with students and the subject matter, convened on May 12, 2021, to review, revise, and approve the range ALDs. Appendix B, NDSA for Science Range Achievement-Level Descriptors, provides the final range ALDs for the NDSA for Science.

5.4.2 Ordered Scoring Assertion Booklets

Like the Bookmark method used to establish achievement standards for traditional science tests, the AMP method uses booklets of ordered test materials to set 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 state item pool.

The OSABs were assembled using a mixed-integer programming approach. The objective function that was minimized was the number of gaps between the impact values of the assertions across the entire OSAB. A gap was defined as a difference of 3% or more between the impact values of two consecutive assertions ordered by difficulty. The linear constraints of the mixed-integer problem represented the constraints implied by the blueprint. In addition, the total number of assertions was not allowed to exceed 85. A set of feasible solutions was further evaluated based on the distribution of the impact values of assertions across the OSAB. The candidate solution was then reviewed internally by content experts and by the NDDPI and approved without any changes for all three grades. Figure 7 describes the structure of the OSAB.

Figure 7. Ordered Scoring Assertion Booklet (OSAB)



For grades 4 and 8, each discipline has 2 item clusters and 4 stand-alone items. For grade 10, the Physical Sciences discipline contains 2 item clusters and 4 stand-alone items, and the Life Sciences discipline contains 4 item clusters and 8 stand-alone items.

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

Each page in the OSAB is a scoring assertion.

Since the operational test was adaptive, the order of the items was different across students. The items in the OSABs were grouped by science discipline so that panelists worked through all items associated with one discipline before moving to the next, allowing panelists to focus on the knowledge and skill requirements for one discipline at a time.

- For the grade 4 OSAB, the Earth and Space Science discipline items were presented first, then Life Sciences items, and then Physical Sciences items.
- For the grade 8 OSAB, the Physical Sciences discipline items were presented first, then Earth and Space Sciences items, and then Life Sciences items.
- For the grade 10 OSAB, Life Sciences discipline items were presented first, followed by the Physical Sciences items.
- For grades 4 and 8, two item clusters and four stand-alone items represent each discipline.

- For grade 10, four item clusters and eight stand-alone items represent the Life Sciences discipline, and two item clusters and four stand-alone items represent the Physical Sciences discipline.

Within a discipline, the item clusters were presented first, followed by the stand-alone items. The item clusters and stand-alone items were further ordered by mean difficulty of the assertions within the item. This approach may help to reduce some of the cognitive demands on panelists by making clear that some items, and their associated interactions, are easier for students to access, even though the assertions they support are similar in content.

Within each item cluster or stand-alone item, scoring assertions were ordered by difficulty. Easier assertions are those that 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 in the OSAB was typically more difficult than the easiest assertion of the next item in the OSAB. That is, the order of assertions in Figure 7 represents the order of presentation to the panelists, but assertions were not ordered by overall difficulty across all items. (See Figure 8 for a depiction of the overlapping difficulty of assertions in the complete OSAB.)

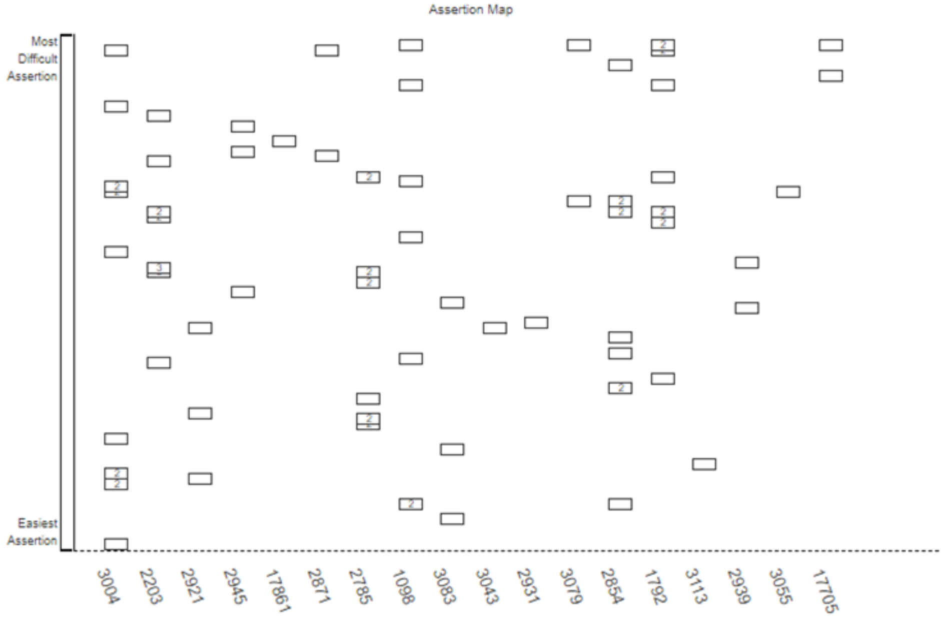
Not all items have assertions that will map onto all achievement levels. For example, an item cluster may have assertions that map onto Novice, Partially Proficient, and Proficient, but not Advanced.

The grade 4 and 8 OSABs contain three disciplines and 18 items (item clusters and stand-alone items), and the grade 10 OSAB contains two disciplines and 18 items. The grade 4 OSAB contained 68 assertions, the grade 8 OSAB contained 74 assertions, and the grade 10 OSAB contained 73 assertions. Each comprised of six item clusters and 12 stand-alone items.

5.4.3 Assertion Maps

Assertion maps were provided to panelists to help reduce the cognitive load of the AMP. The assertion maps were displayed in CAI’s online standard-setting tool and listed all scoring assertions in each OSAB by item ID and assertion, and plotted all assertions by difficulty. The assertion maps provided panelists with context about student performance on the assertions in the OSAB, describing the difficulty of each assertion in the underlying OSAB. This was to help panelists easily identify more- or less-difficult assertions and compare the difficulty of assertions across items. The assertion maps were provided during the OSAB review. After Round 1 and Round 2, the assertion maps were updated to also display the tentative standards (refer to Section 5.7.2.2, Feedback Data, for more details). Figure 8 presents the assertion map for grade 4. The assertions maps for all three grades are presented in Appendix C, NDSA for Science Standard-Setting Assertion Maps.

Figure 8. Standard-Setting Assertion Map, Science Grade 4



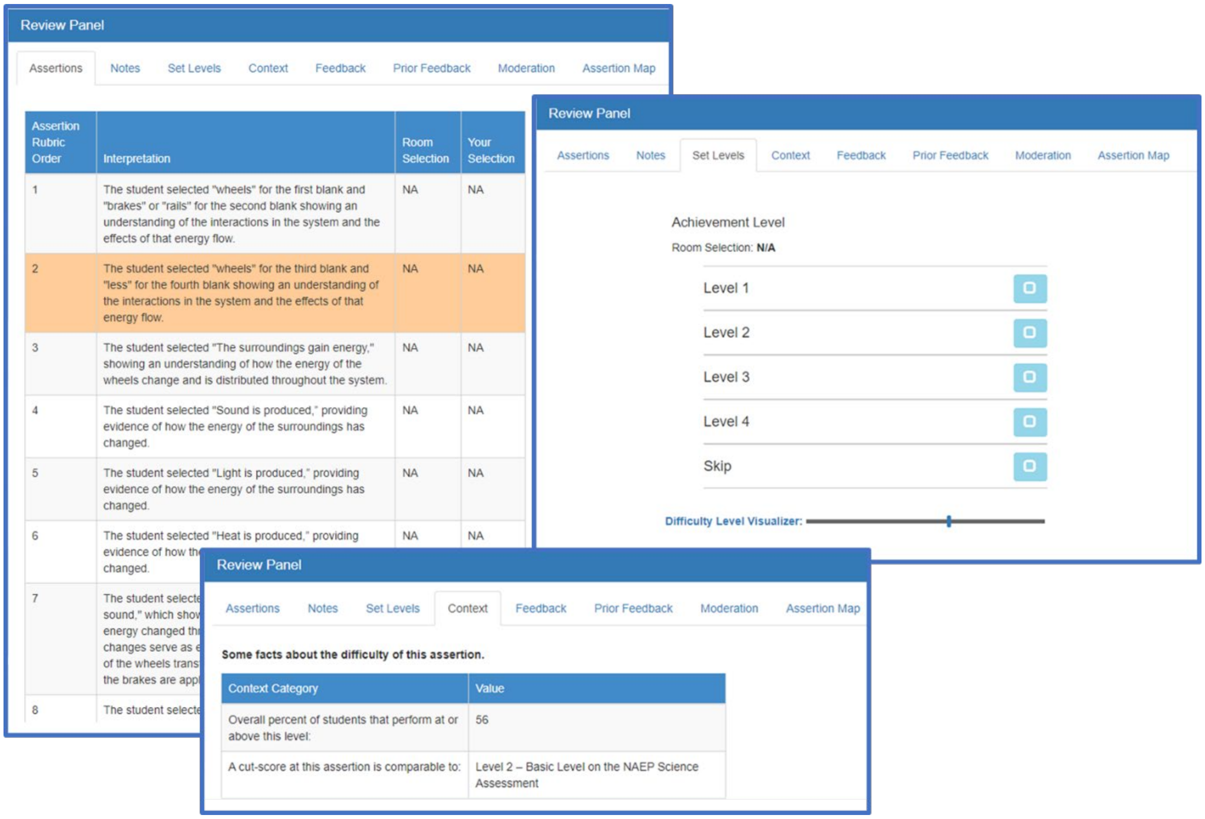
5.5 Workshop Technology

The standard-setting panelists used CAI's online application for standard setting. Each panelist used their own computer on which they took the test, reviewed item clusters and stand-alone items and ancillary materials, and mapped assertions to achievement levels.

Using tabs in the review panel of the tool (see Figure 9), panelists could review the items and scoring assertions, determine the relative difficulty of assertions to other assertions in the same item, examine the content alignment of each item (via the alignment of the assertions within an item, which all align to the same performance standard), assign assertions to achievement levels, add notes and comments on the assertions as they reviewed them, and review contextual information and feedback data. Additionally, they had access to a difficulty level visualizer, a graphic representation of the difficulty of each assertion relative to all other assertions in the OSAB (not just within the item).[2] Panelists also reviewed their assertion placement, their table's placement, the other tables' placement, and the overall placement for both tables.

[2] The difficulty level visualizer represented the percentage of students who would fall at or above the difficulty level of that assertion.

Figure 9. Example Features in CAI's Standard-Setting Tool



Assertion Rubric Order	Interpretation	Room Selection	Your Selection
1	The student selected "wheels" for the first blank and "brakes" or "rails" for the second blank showing an understanding of the interactions in the system and the effects of that energy flow.	NA	NA
2	The student selected "wheels" for the third blank and "less" for the fourth blank showing an understanding of the interactions in the system and the effects of that energy flow.	NA	NA
3	The student selected "The surroundings gain energy," showing an understanding of how the energy of the wheels change and is distributed throughout the system.	NA	NA
4	The student selected "Sound is produced," providing evidence of how the energy of the surroundings has changed.	NA	NA
5	The student selected "Light is produced," providing evidence of how the energy of the surroundings has changed.	NA	NA
6	The student selected "Heat is produced," providing evidence of how th changed.	NA	NA
7	The student selecte sound," which show energy changed thr changes serve as e of the wheels trans the brakes are appl		
8	The student selecte		

Achievement Level

Room Selection: **N/A**

- Level 1 □
- Level 2 □
- Level 3 □
- Level 4 □
- Skip □

Difficulty Level Visualizer:

Some facts about the difficulty of this assertion.

Context Category	Value
Overall percent of students that perform at or above this level:	56
A cut-score at this assertion is comparable to:	Level 2 – Basic Level on the NAEP Science Assessment

Full-time CAI information technology specialists answered questions and ensured that technological processes ran smoothly and without interruption throughout the remote workshop.

5.6 Events

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

Table 7. Standard-Setting Agenda Summary

Day 1: Wednesday, June 16, 2021
• Large-Group Orientation • Review and Take the Operational Test • Review Range ALDs • Discuss Threshold ALDs • OSAB Review
Day 2: Thursday, June 17, 2021
• Continue OSAB Review • Assertion-Mapping Training • Round 1 Assertion Mapping • Round 1 Feedback and Impact Data Review and Discussion

- Round 2 Assertion Mapping
- Round 2 Feedback and Impact Data Review
- Standard-Setting Workshop Evaluations
- Across-Grade Moderation and Articulation

5.6.1 Participant Login

Panelists were required to attend a technical check prior to the standard-setting workshop to ensure they had access to the required sites needed for the workshop. They also received and signed affidavits of non-disclosure at this time, affirming that they would not reveal any secure information they would have access to during the workshop. Panelists arrived at the workshop, virtually, on the first day, and followed the instructions given for joining the workshop via Microsoft Teams.

5.6.2 Large-Group Orientation

Stanley Schauer, NDDPI Director of Assessment, welcomed panelists to the workshop and provided context and background for the NDSA for Science. The NDDPI outlined the roles and responsibilities of the participants at the workshop: panelists, CAI staff, and NDDPI personnel. 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 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 finalized their recommendations. Appendix E, NDSA for Science Standard-Setting Training Slides, provides the slides used during the large-group training.

5.6.3 Confidentiality and Security

Workshop leaders and room facilitators addressed confidentiality and security during orientation and again in each room. Standard setting uses live science test items from the operational NDSA for Science, requiring confidentiality to maintain their security. Participants were forbidden to do the following either during, or after, the workshop:

- Discuss test items outside of the meeting
- Discuss judgments or cut scores (their own or others') with anyone outside of the meeting
- Discuss secure materials with non-participants
- Create any form of electronic copy of test content (e.g., screenshots, electronic notes)
- Create any hand-written notes of test content
- Use a personal computer during the course of the meeting for any purpose other than participating in the standard-setting workshop and item review (e.g., email, web browsing, social media)
- Save notes about item or passage content to a computer

While participants could engage in general conversation regarding the process and days' events, workshop leaders warned them against discussing details, particularly those involving test items, cut scores, or any other confidential information.

5.6.4 Take the Operational Test

Following the large-group orientation, panelists broke out into their separate grade-level virtual meeting rooms. As their introduction to the standard-setting process, panelists took a form of the test that students took in 2021, in the grade for which they would be setting achievement standards. Panelists took the tests online via the same tool used to deliver operational tests to students; and the testing environment closely matched that of students when they took the test.

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

5.6.5 Range Achievement-Level Descriptor Review

After taking the operational test, panelists completed a thorough review of the range ALDs for their assigned grade. Panelists were provided with an overview of the ALDs and their importance to standard setting. The ALDs were used as a reference for evaluating student performance, so it was important for panelists to understand the critical role of ALDs in the standard-setting process.

Panelists began their review of the range ALDs that define what students in each achievement level know and are able to do with respect to the 2019 North Dakota Science Content Standards. Workshop facilitators provided panelists with draft range ALDs, test blueprints, and the 2019 North Dakota Science Content Standards. The facilitators lead panelists through a thorough review of the range ALDs for their assigned grade using the materials as references and drawing on the expertise of the panelists.

Panelists identified key words describing the skills necessary for achievement at each level and discussed the skills and knowledge that differentiate achievement in each of the four levels.

Reviewing the range ALDs ensured that participants understood what students in North Dakota should know and be able to do and how much knowledge and skill students are expected to demonstrate at each level of achievement.

5.6.6 Discuss Threshold Achievement-Level Descriptors

After reviewing and discussing the range ALDs, panelists worked in their grade-level groups to develop a shared understanding of the threshold ALDs that describe the skills that students just barely able to score in one achievement level have but that students scoring just below the achievement level do not have. Facilitators encouraged panelists to consider the characteristics of students who just barely qualify for entry into the achievement level from those just below. Looking at each ALD, panelists identified the skills needed to just barely perform at that level. The following two questions guide the process:

1. What skills and knowledge must the student demonstrate to qualify for entrance into this achievement level?

2. How does this differ from the upper range of the adjacent (lower) achievement level?

These discussions yielded common descriptions of students just barely characterized by each ALD within each room.

The AMP employed the range ALDs since panelists were mapping items across the full range of the ALD. The purpose of the threshold ALD discussion was to enhance the panelists' understanding of the differences between ALD levels by paying special attention to the transition areas between achievement levels.

5.6.7 Ordered Scoring Assertion Booklet Review

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

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

5.6.8 Assertion-Mapping Training

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

Panelists were to match each assertion to the achievement level best supported by the assertion using the ALDs, the difficulty level visualizer (described in Section 5.5, Workshop Technology), the assertion map (described in Section 5.4.3, Assertion Maps), their notes from the OSAB review, and their professional judgments. Figure 10 graphically describes the assertion-mapping process.

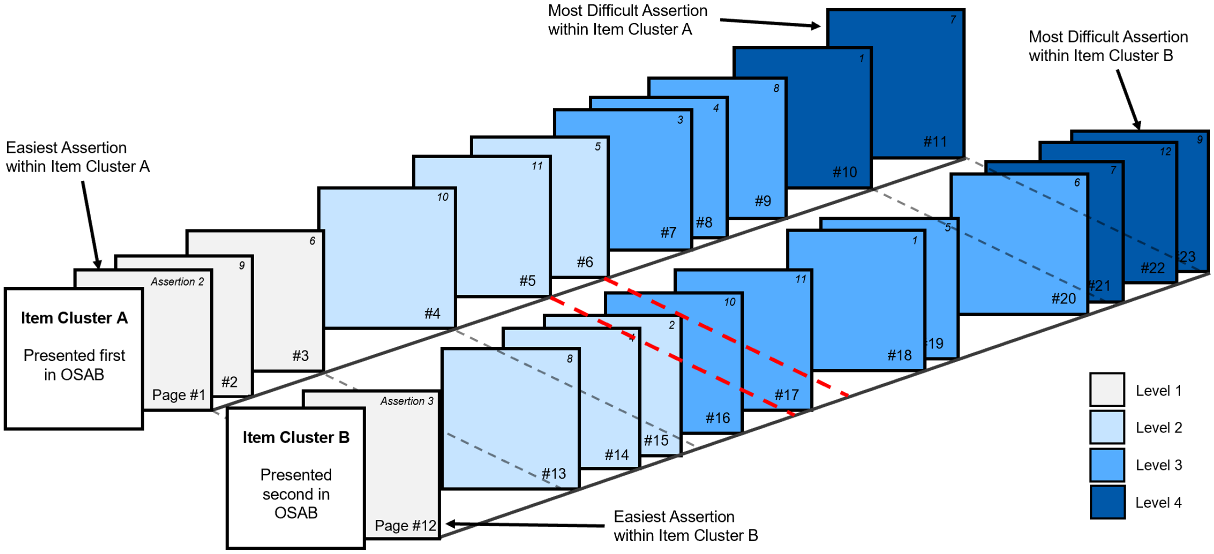
Facilitators provided the following process to guide the mapping of assertions onto ALDs:

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 (within the item)?
3. Which ALD most ably describes this assertion and the underlying interactions?

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

Because the assertion mapping was conducted separately for each item, there might have been 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 (i.e., mapping inversions of assertions could occur across items, but mapping inversions of assertions were not allowed within an item). For example, in Figure 10, the difficulty of the assertion on page 6 of item cluster A ("Level 2") has a higher difficulty than the assertion on page 17 of item cluster B ("Level 3"). However, it was expected for the higher achievement levels to be assigned more frequently with increasing assertion difficulty across items. Appendix E, NDSA for Science Standard-Setting Training Slides, provides the training slides used during the breakout room training.

Figure 10. Example of Assertion Mapping



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

5.6.9 Practice Quiz

Panelists completed a practice quiz before beginning a practice round. The quiz assessed panelists' understanding in multiple ways. They must be able to perform the following:

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

Room facilitators reviewed the quizzes with the panelists and provided additional training for incorrect responses on the quiz. Appendix F, NDSA for Science Standard-Setting Practice Quiz, provides the quiz that panelists completed before mapping any assertions.

5.6.10 Practice Round

Following the practice quiz, panelists practiced mapping assertions to ALDs in a short practice OSAB consisting of one item cluster and one stand-alone item. The purpose of the practice round was to ensure that panelists were comfortable with the technology, items, item interactions, and scoring assertions before mapping any assertions in the OSAB. Panelists discussed their practice mappings and asked questions, and the room facilitators provided clarifications and further instructions until everyone had completed the practice round.

5.6.11 Readiness Assertion

After completing the practice round, and before mapping assertions to achievement levels 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 determine and map assertions to ALDs in the standard-setting tool
- Understanding how to use the assertion map when reviewing the OSAB and considering assertion mapping decisions
- Understanding the contextual information (student impact data and benchmarking data) when considering assertion mapping decisions
- Readiness to begin the Round 1 task

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

- Understanding of the feedback data and impact data
- Understanding of the Round 2 task
- Readiness to complete the Round 2 task

Room facilitators reviewed the readiness forms and provided additional training to panelists not asserting understanding or readiness. However, every panelist affirmed readiness before mapping assertions in both rounds of the workshop. Appendix G, NDSA for Science Standard-Setting Readiness Forms, contains the forms that panelists completed prior to each round of standard setting.

5.7 Assertion Mapping

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

5.7.1 Calculating Cut Scores from the Assertion Mapping

Cut scores were calculated by treating every possible scale value as a hypothetical cut score and evaluating the number of discrepancies between the assertion mappings of the panelists and the achievement levels of the assertions implied by hypothetical cut score. The implied achievement level of an assertion was determined by comparing the response probability of an assertion to the hypothetical cut.[3] Each cut score was defined as the score point that minimized the weighted number of discrepancies. 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 "Partially Proficient" and "Proficient" were used). Specifically, let n_k be the number of assertions put at achievement level k, t_k be the cut to be estimated, d_i be the assigned achievement level, and θ_i be the RP value of the ith assertion. For each assertion placed at levels k and $k+1$, the misclassification indicator is defined as

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

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

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

Unlike the Bookmark method, the cut scores for a table or room were not the median value of the cut scores of the individual panelists. Instead, cut scores at the table and room (grade) level were computed using the same method but considering the assigned levels of all the raters at the table and in the room, respectively. Applying these cut scores to the 2021 operational test data created data describing the percentage of students falling into each achievement level. This algorithm calculated cut scores from the assertion mappings by panelist, by table, and for the room.

[3] Typically, the response probability used in standard setting is 0.67 ("RP67" [Huynh, 1994]). RP67 is the assertion difficulty point where 67% of the students would earn the score point. The reason to adopt RP50 for grades 4, 8, and 10 for North Dakota was because the difficulty of most items exceeded students' abilities. RP50 better aligned with the ALD and therefore led to more appropriate achievement cut scores. Using RP50 prevented panelists from mapping the first cut score onto the lowest-difficulty assertions on the test. This approach has been adopted for other high-stakes tests, such as the Smarter Balanced Assessments (see Cizek & Koons, 2014).

5.7.2 Contextual Information and Feedback Data

To be adoptable, achievement standards for a statewide system must be coherent across grades and subjects. 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.

The standard-setting tool developed by CAI provides feedback data and allows for displaying contextual information to ensure standard-setting recommendations are well articulated.

5.7.2.1 Contextual Information

Panelists were also provided with additional contextual information to help inform their primary content-driven achievement standard recommendations. The standard-setting tool developed by CAI allows for displaying both impact and benchmark data to ensure standard-setting recommendations are well articulated. The contextual information provided included impact data and benchmark data for each of the assertions of the OSAB, as described in the following sections.

Impact Data

The impact data for an assertion was defined as the percentage of students who performed at or above the specified RP value associated with the assertion. Panelists were asked to consider the impact data when making their content-based assertion mappings.

Benchmark Data

The 2015 National Assessment of Educational Progress (NAEP) science scores provided benchmark data, another source of contextual 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, panelists could evaluate the reasonableness of the proposed achievement standards. NAEP provides state-level data in science for grades 4 and 8; benchmark data for grade 10 is extrapolated. For each ordered scoring assertion, panelists were provided with the associated achievement level for the NAEP science. An example of the benchmark information provided for each assertion in the review panel of the standard-setting tool is shown in Figure 9. This provided external evidence of student achievement for panelists to consider when mapping assertions to achievement levels in Round 2.

5.7.2.2 Feedback Data

The online standard-setting tool created feedback data and cut scores corresponding to the assertion mappings for each panelist, for each table, and for the room overall (across both tables).

In addition, panelists were shown impact data based on the cut scores resulting from their assertion mappings. Impact data were defined for panelists as the percentages of students who would reach or exceed each of the achievement standards given the assertion mappings. Percentages were calculated using the student data from the 2021 administration of the NDSA for Science. This information allowed panelists to compare their mappings to other panelist's mappings to evaluate the impact of their current mappings.

The standard-setting tool also generated variance monitor data and the assertion maps in the tool were updated to display the tentative standards for panelists to evaluate before Round 2 (the variance data and assertion maps are described in more detail below). 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 students before mapping assertions in Round 2.

After reviewing the feedback data, the 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. The second goal was to encourage articulation across grade levels. Each panel spent time reviewing and discussing assertion mappings and articulation, beginning with table-level feedback and discussion, and progressing to the room-level discussion. After completing these discussions, panelists again worked through mapping all OSAB assertions to achievement levels for Round 2.

Variance Monitor Data

Feedback included a review of a variance monitor, part of CAI'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. Figure 11 illustrates the types of information available in the variance monitor. Room facilitators and panelists reviewed and discussed the assertions with the most variable mappings.

Figure 11. Variance Monitor in CAI's Standard-Setting Tool

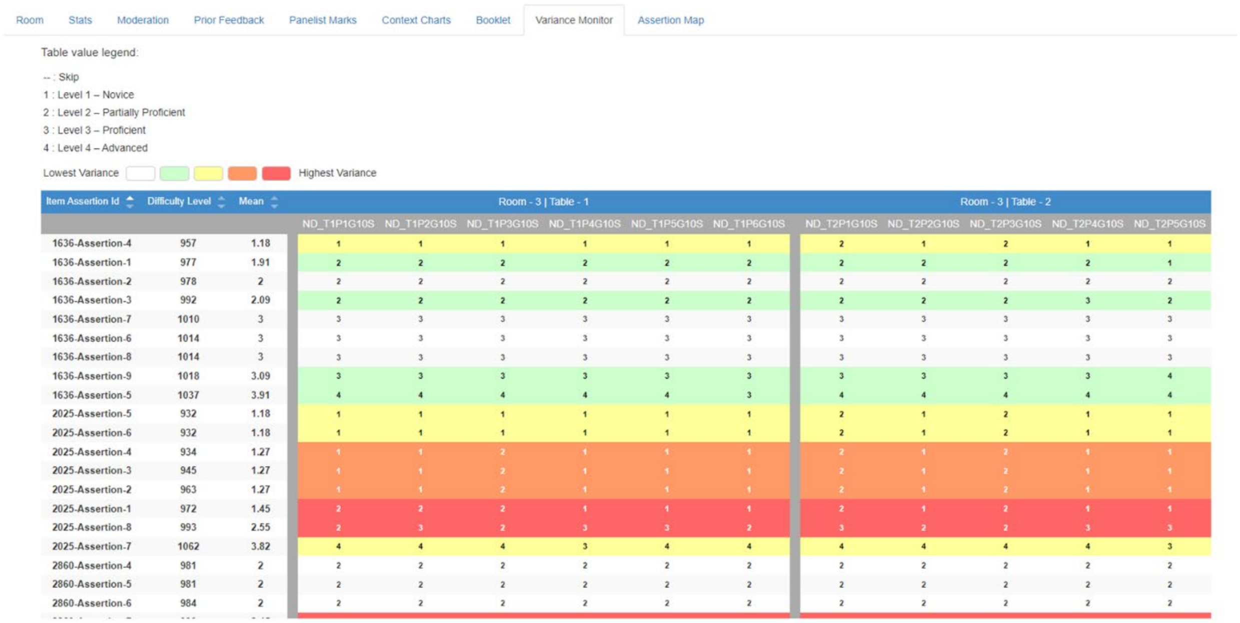


Table value legend:

-- : Skip
1 : Level 1 – Novice
2 : Level 2 – Partially Proficient
3 : Level 3 – Proficient
4 : Level 4 – Advanced

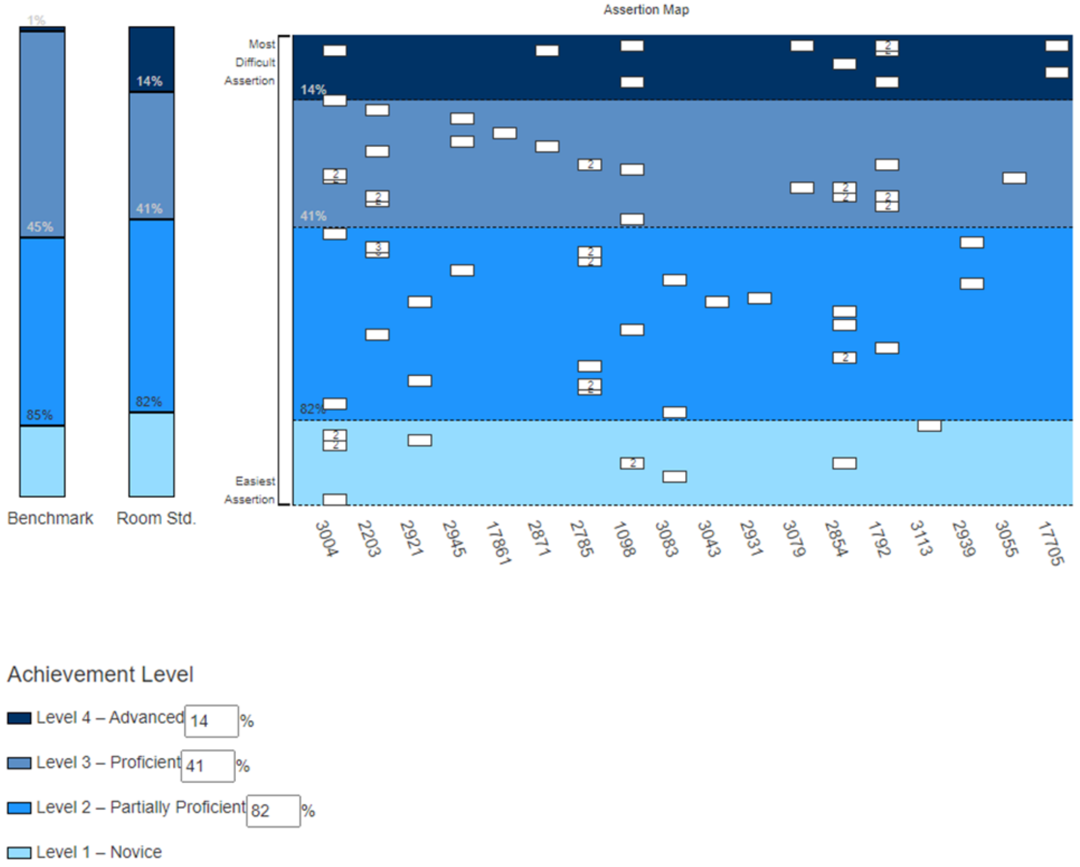
Lowest Variance ⬜ 🟩 🟨 🟧 🟥 Highest Variance

Item Assertion Id	Difficulty Level	Mean	Room - 3 | Table - 1						Room - 3 | Table - 2				
			ND_T1P1G10S	ND_T1P2G10S	ND_T1P3G10S	ND_T1P4G10S	ND_T1P5G10S	ND_T1P6G10S	ND_T2P1G10S	ND_T2P2G10S	ND_T2P3G10S	ND_T2P4G10S	ND_T2P5G10S
1636-Assertion-4	957	1.18	1	1	1	1	1	1	2	1	2	1	1
1636-Assertion-1	977	1.91	2	2	2	2	2	2	2	2	2	2	1
1636-Assertion-2	978	2	2	2	2	2	2	2	2	2	2	2	2
1636-Assertion-3	992	2.09	2	2	2	2	2	2	2	2	2	3	2
1636-Assertion-7	1010	3	3	3	3	3	3	3	3	3	3	3	3
1636-Assertion-6	1014	3	3	3	3	3	3	3	3	3	3	3	3
1636-Assertion-8	1014	3	3	3	3	3	3	3	3	3	3	3	3
1636-Assertion-9	1018	3.09	3	3	3	3	3	3	3	3	3	3	4
1636-Assertion-5	1037	3.91	4	4	4	4	4	3	4	4	4	4	4
2025-Assertion-5	932	1.18	1	1	1	1	1	1	2	1	2	1	1
2025-Assertion-6	932	1.18	1	1	1	1	1	1	2	1	2	1	1
2025-Assertion-4	934	1.27	1	1	2	1	1	1	2	1	2	1	1
2025-Assertion-3	945	1.27	1	1	2	1	1	1	2	1	2	1	1
2025-Assertion-2	963	1.27	1	1	2	1	1	1	2	1	2	1	1
2025-Assertion-1	972	1.45	2	2	2	1	1	1	2	1	2	1	1
2025-Assertion-8	993	2.55	2	3	2	3	3	2	3	2	2	3	3
2025-Assertion-7	1062	3.82	4	4	4	3	4	4	4	4	4	4	3
2860-Assertion-4	981	2	2	2	2	2	2	2	2	2	2	2	2
2860-Assertion-5	981	2	2	2	2	2	2	2	2	2	2	2	2
2860-Assertion-6	984	2	2	2	2	2	2	2	2	2	2	2	2

Assertion Maps

In addition to providing the numerical value of the cut scores and impact data, the feedback was shown on the assertion maps. After each round of assertion mapping, the assertion maps displayed in CAI's online standard-setting tool were updated with the overall room cut scores and the individual panelist cut scores for Round 1 and Round 2. Figure 12 presents the assertion map for grade 4 with the overall room cut scores for Round 1. The Round 1 and Round 2 assertion maps with overall room cut scores for all three grades are presented in Appendix H, NDSA for Science Round 1 and Round 2 Standard-Setting Assertion Maps.

Figure 12. Round 1 Standard-Setting Assertion Map, Grade 4



Panelists were instructed to consider their assertion mappings to compare the room cut score and assertions to their cut scores and assertion mappings. They were again reminded to evaluate the relative location of the assertions on the assertion maps.

5.8 Assertion Mapping Results

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

5.8.1 Round 1 Results

Table 8 presents the achievement standards and associated impact data (percentage of students falling at or above each of the achievement standards based on the recommended Round 1 cut scores) from Round 1.

Table 8. Round 1 Results

Grade and Table	Cut Score			Impact Data		
	Level 2 Partially Proficient	Level 3 Proficient	Level 4 Advanced	Level 2 Partially Proficient (%)	Level 3 Proficient (%)	Level 4 Advanced (%)
Grade 4	**375**	**407**	**431**	**82**	**41**	**14**
Table 1	371	407	431	85	41	14
Table 2	375	409	425	82	38	20
Grade 8	**775**	**802**	**827**	**80**	**51**	**16**
Table 1	783	802	827	73	51	16
Table 2	768	804	827	86	48	16
Grade 10	**973**	**1000**	**1035**	**82**	**50**	**10**
Table 1	973	1000	1035	82	50	10
Table 2	973	1000	1035	82	50	10

Note. The grade row summarizes the room data (across both tables). Impact data describes the percentage of students falling at or above each of the achievement standards based on the recommended Round 1 cut scores.

Review of the Round 1 results began with a discussion of the feedback data from Round 1, beginning with table-level feedback and progressing to the room-level discussion. After reviewing the feedback (i.e., individual cuts, cuts by a table, cuts by a room) and impact data, workshop facilitators provided panelists with additional instructions for completing Round 2. They described the goal of Round 2 as one of convergence but not consensus on a common achievement standard. Panelists then spent time reviewing and discussing assertion mappings. After completing these discussions, panelists again worked through the OSAB, mapping assertions for Round 2.

5.8.2 Round 2 Results

Table 9 presents the recommended achievement standards and associated impact data (percentage of students falling at or above each of the achievement standards based on the recommended Round 2 cut scores) from Round 2.

Table 9. Round 2 Results

Grade and Table	Cut Score			Impact Data		
	Level 2 Partially Proficient	Level 3 Proficient	Level 4 Advanced	Level 2 Partially Proficient (%)	Level 3 Proficient (%)	Level 4 Advanced (%)
Grade 4	**380**	**407**	**431**	**76**	**41**	**14**
Table 1	371	407	431	85	41	14
Table 2	380	407	433	76	41	13
Grade 8	**762**	**802**	**835**	**90**	**51**	**10**
Table 1	775	802	827	80	51	16

Grade and Table	Cut Score			Impact Data		
	Level 2 Partially Proficient	*Level 3 Proficient*	*Level 4 Advanced*	*Level 2 Partially Proficient (%)*	*Level 3 Proficient (%)*	*Level 4 Advanced (%)*
Table 2	762	809	835	90	40	10
Grade 10	**973**	**1000**	**1035**	**82**	**50**	**10**
Table 1	973	1000	1035	82	50	10
Table 2	973	998	1035	82	53	10

Note. The grade row summarizes the room data (across both tables). Impact data describes the percentage of students falling at or above each of the achievement standards based on the recommended Round 2 cut scores.

5.8.3 Convergence Across Rounds

While consensus is not an objective of standard setting, convergence is. Indicators of panelist convergence over rounds are the interquartile range (IQR) and standard deviations (SD) of the standards computed for individual panelists based on their mappings. The interquartile range and standard deviations for each grade and after each round are presented in Table 10. For the Level 2 (except in grade 10) and Level 3 standards, the indicators consistently show that there is a convergence in individual standards. For the Level 4 standards, the pattern is not consistent across grades.

Table 10. Interquartile Range and Standard Deviation of Panelist Recommended Achievement Standards

Grade	Statistic	Level 2 Partially Proficient		Level 3 Proficient		Level 4 Advanced	
		Round 1	*Round 2*	*Round 1*	*Round 2*	*Round 1*	*Round 2*
4	IQR	13.75	7.50	8.75	1.75	11.25	9.00
	SD	8.34	6.26	6.42	4.09	9.17	8.21
8	IQR	16.75	2.5	16.00	7.00	8.00	17.50
	SD	17.65	6.02	10.11	4.85	8.21	8.87
10	IQR	1.00	5.00	4.50	2.00	0	4.00
	SD	4.00	5.55	4.76	2.11	8.44	8.32

5.8.4 Moderation and Results

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 NDDPI, these cuts are offered to a subset of panelists after Round 2 for consideration in a step referred to as moderation.

On the last day of the workshop, table leaders met to discuss and resolve any issues or needs related to cross-grade articulation, resulting in the final recommendations provided in Table 11. Workshop leaders reminded panelists that content is one of multiple considerations in setting achievement standards—perhaps the most important, but not the only consideration; panelists also considered impact and policy in Round 2. After discussion, the moderation panel made a minor adjustment to the grade 8 Partially Proficient cut for better articulation across the grades.

Table 11 displays the moderated achievement standards recommended by the standard-setting panelists.

Table 11. Moderated Results for Science

Grade	Cut Score			Impact Data		
	Level 2 Partially Proficient	*Level 3 Proficient*	*Level 4 Advanced*	*Level 2 Partially Proficient (%)*	*Level 3 Proficient (%)*	*Level 4 Advanced (%)*
4	380	407	431	76	41	14
8	775*	802	835	80*	51	10
10	973	1000	1035	82	50	10

Note. *Minor adjustment made during the moderation session.

Figure 13 displays the percentage of students that will reach or exceed each of the recommended science achievement standards in 2021.

Figure 13. Percentage of Students Reaching or Exceeding Each Recommended Science Achievement Standard in 2021

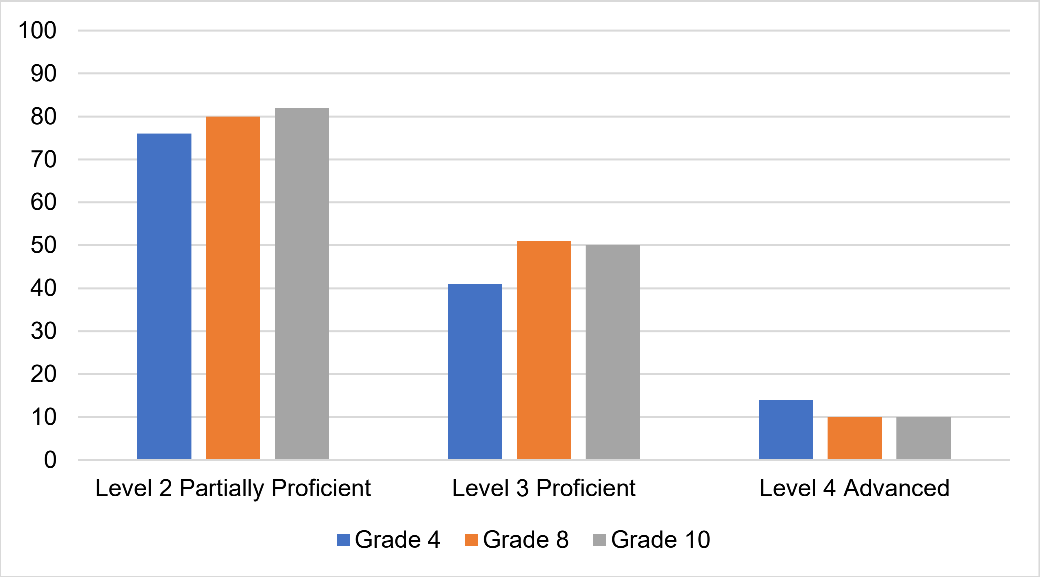
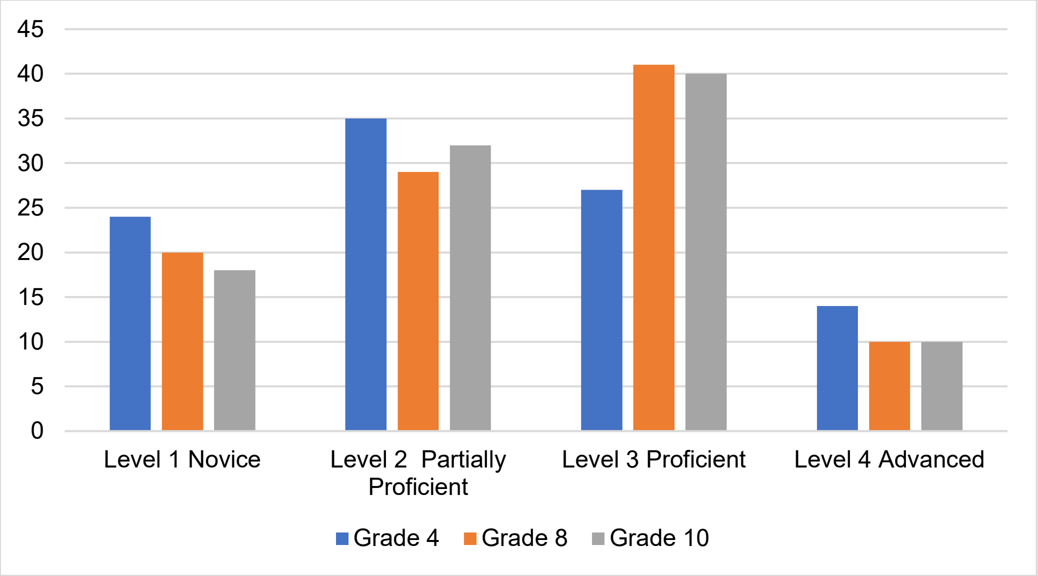


Table 12 indicates the percentage of students classified within each of the achievement levels in 2021. These values are displayed graphically in Figure 14.

Table 12. Percentage of Students Classified Within Each Science Achievement Level, 2021

Grade	Level 1 Novice	Level 2 Partially Proficient	Level 3 Proficient	Level 4 Advanced
4	24	35	27	14
8	20	29	41	10
10	18	32	40	10

Figure 14. Percentage of Students Classified Within Each Science Achievement Level, 2021



5.9 Workshop Evaluations

After finishing all activities, panelists completed online workshop evaluations independently, in which they described and evaluated their experience taking part in the standard setting. Tables 13 through 17 summarize the results of the evaluations. Evaluation items endorsed by fewer than 90% of panelists are discussed in the text, and the least endorsed items are discussed in terms of the number and type of response.

Generally, workshop participants indicated clarity in the instructions, materials, data, and process (see Table 13). Three panelists reported the least clarity with impact data.

Table 13. Evaluation Results: Clarity of Materials and Process

Please rate the clarity of the following components of the workshop.	Percentage Indicating “Somewhat Clear” or “Very Clear”			
	Science Grade 4	Science Grade 8	Science Grade 10	Overall
Instructions provided by the workshop leader	100	90	100	97
Achievement-level descriptors (ALDs)	90	100	100	97
Ordered Scoring Assertion Booklet (OSAB)	100	100	100	100
Assertion Map	90	100	100	97
Impact data (percentage of students that would achieve at the level indicated by the assertion difficulty)	80	100	91	90
Panelist agreement data	90	100	91	94

Note. Number of responses = 31 (grade 4 responses = 10, grade 8 responses = 10, grade 10 responses = 11). Evaluation response options included “Very Unclear,” “Somewhat Unclear,” “Somewhat Clear,” and “Very Clear.”

Participants felt they had sufficient time to complete all activities. Some indicated having too much time to complete some tasks (see Table 14). Eleven panelists indicated the large-group orientation was too long. Seven panelists reported the time for experiencing the operational test was too short, while one panelist indicated that it was too long. Four panelists reported the time for ALD review was too long. Four panelists reported having too little time to discuss the “just barely” students, while six others reported having too much time to discuss. One panelist indicated spending too little time on reviewing the OSAB, while three others reported spending too much time on this task. Four panelists felt too much time was spent mapping scoring assertions, while two panelists felt not enough time was spent. Finally, four panelists indicated that not enough time was allowed for Round 1 results discussion, while one panelist reported the discussion took too much time.

Table 14. Evaluation Results: Appropriateness of Process

How appropriate was the amount of time you were given to complete the following components of the standard-setting process?	Percentage Indicating “About Right”			
	Science Grade 4	Science Grade 8	Science Grade 10	Overall
Large-group orientation	50	80	64	65
Experiencing the online assessment	100	80	45	74
Reviewing the achievement-level descriptors (ALDs)	80	90	91	87
Discussion of the skills demonstrated by students who are “just barely” described by each ALD	60	70	73	68
Reviewing the Ordered Scoring Assertion Booklet (OSAB)	100	80	82	87
Mapping your scoring assertions to achievement levels in each round	80	80	82	81
Round 1 results discussion	90	80	82	84

Note. Number of responses = 31 (grade 4 responses = 10, grade 8 responses = 10, and grade 10 responses = 11). Evaluation response options included “Too Little,” “Too Much,” and “About Right.”

Participants appreciated the importance of the multiple factors contributing to assertion mapping, with nearly all participants rating each factor as important or very important (see Table 15). Two grade 4 panelists indicated the "Just Barely" ALDs, their perception of the difficulty of the scoring assertions and items, and impact data were not important.

Table 15. Evaluation Results: Importance of Materials

How important were each of the following factors in your mapping of scoring assertions to achievement levels?	Percentage Indicating "Somewhat Important" or "Very Important"			
	Science Grade 4	*Science Grade 8*	*Science Grade 10*	*Overall*
Achievement-level descriptors (ALDs)	100	100	100	100
"Just Barely" ALDs	80	100	100	94
Your perception of the difficulty of the scoring assertions and items in general	80	100	100	94
Your experience with students	100	90	100	97
Discussions with other panelists	90	100	100	97
Assertion map	90	100	100	97
External benchmark data	90	100	100	97
Impact data (percentage of students that would achieve at the level indicated by the assertion difficulty)	80	100	100	94
Room agreement data (room, table, and individual standards)	90	100	91	94

Note. Number of responses = 31 (grade 4 responses = 10, grade 8 responses = 10, and grade 10 responses = 11). Evaluation response options included "Not Important," "Somewhat Important," and "Very Important."

Participant understanding of the workshop processes and tasks was high (see Table 16). The least agreed with statements are related to the just barely ALDs and impact data. A total of four panelists disagreed with the just barely ALD statement, and a total of four panelist disagreed with the impact data statement.

Table 16. Evaluation Results: Understanding Processes and Tasks

At the end of the workshop, please rate your agreement with the following statements.	Percentage Indicating "Agree" or "Strongly Agree"			
	Science Grade 4	*Science Grade 8*	*Science Grade 10*	*Overall*
I understood the purpose of this standard-setting workshop.	80	100	100	94
The procedures used to recommend achievement standards were fair and unbiased.	90	90	91	90
The training provided me with the information I needed to recommend achievement standards.	100	100	100	100
Taking the online assessment helped me to better understand what students need to know and be able to do to answer each question.	100	100	100	100
The achievement-level descriptors (descriptions of what students within each achievement level	70	100	100	90

At the end of the workshop, please rate your agreement with the following statements.	**Percentage Indicating "Agree" or "Strongly Agree"**			
	Science Grade 4	*Science Grade 8*	*Science Grade 10*	*Overall*
are expected to know and be able to do) provided a clear picture of expectations for student achievement at each level.				
I was able to develop an understanding of the knowledge and skills demonstrated by students who are "just barely" described by the ALDs.	80	100	82	87
I understood how to review each assertion in the Ordered Scoring Assertion Booklet (OSAB) to determine what students must know and be able to do to answer each assertion correctly.	90	100	100	97
I understood how to map assertions to the most apt achievement level.	90	100	100	97
I found the assertion map helpful in my decisions about the assertions I mapped to achievement levels.	90	100	100	97
I found the benchmark data and discussions helpful in my decisions about the assertions I mapped to achievement levels.	80	100	100	94
I found the impact data (percentage of students that would achieve at the level indicated by the assertion difficulty) and discussions helpful when mapping assertions to achievement levels.	80	100	82	87
I found the panelist agreement data (room, table, and individual cuts) and discussion helpful in my decisions about assertions I mapped to achievement levels.	80	100	100	94
I felt comfortable expressing my opinions throughout the workshop.	90	100	100	97
Everyone was given the opportunity to express his or her opinions throughout the workshop.	100	100	100	100

Note. Number of responses = 31 (grade 4 responses = 10, grade 8 responses = 10, and grade 10 responses = 11). Evaluation response options included "Strongly Disagree," "Disagree," "Agree," and "Strongly Agree."

Finally, nearly all participants agreed that the standards set during the workshop reflected the intended grade-level expectations (see Table 17).

Table 17. Evaluation Results: Student Expectations

Please read the following statement carefully and indicate your response.	**Percentage Indicating "Agree" or "Strongly Agree"**			
	Science Grade 4	*Science Grade 8*	*Science Grade 10*	*Overall*
"A student performing at Partially Proficient nearly meets proficiency for the grade."	90	90	100	94
"A student performing at Proficient meets proficiency for the grade."	100	100	100	100

Please read the following statement carefully and indicate your response.	**Percentage Indicating "Agree" or "Strongly Agree"**			
	Science Grade 4	*Science Grade 8*	*Science Grade 10*	*Overall*
"A student performing at Advanced exceeds proficiency for the grade."	90	100	100	97

Note. Number of responses = 31 (grade 4 responses = 10, grade 8 responses = 10, and grade 10 responses = 11). Evaluation response options included "Strongly Disagree," "Disagree," "Agree," and "Strongly Agree."

5.9.1 Workshop Participant Feedback

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

Twenty-eight panelists responded to the first question, and 21 responded to the second. Most responses indicated the training was effective and the process was clear. Participants provided minor suggestions, such as shortening or lengthening the time allocated for some tasks and having the workshop in person. Many appreciated the organization, well-prepared materials, and technology, and many panelists complimented the professionalism and expertise of the facilitators.

Additional participant comments included:

> *"I really valued this opportunity as an educator and appreciated the professionalism and patience from the leaders."*
>
> *"Thank you! This was a great experience and I got to 'see' teachers from around that state that I haven't seen in a number of years."*
>
> *"This was very informational and worth my time. I enjoyed getting to see how questions are presented to students. I previously taught in MN and we were able to do practice questions as a class prior to them taking the state exams. This allowed me to show students how the basic functions of test questions work. For example, how to highlight when reading a passage, where the calculator is located, moving of question items. This is something that is very helpful for students."*

6. Validity Evidence

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

6.1 Evidence of Adherence to Professional Standards and Best Practices

The North Dakota State Assessment (NDSA) for 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). The process also adhered to the following professional standards recommended in the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014) related to standard setting:

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

The sections of this documentation discussing the rationale and procedures used in the standard-setting workshop address Standard 5.21. The Assertion-Mapping Procedure (AMP) for standard setting is appropriate for tests of this type—with interrelated sets of three-dimensional item clusters and scaled using item response theory (IRT). Section 5.1, The Assertion-Mapping Procedure, provides the justification for and the additional benefits of selecting the AMP method to establish the cut scores; Section 5.6, Events, through Section 5.8.1, Round 1 Results, document the process followed to implement the method.

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

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

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

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

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

The provision of benchmark, context, and articulation data to panelists after Round 1 addresses Standard 5.23 (see Section 5.7.2, Contextual Information and Feedback Data). This set of empirical data provides necessary and additional context describing student performance given the recommended standards.

6.2 EVIDENCE IN TERMS OF PEER-REVIEW CRITICAL ELEMENTS

The U.S. Department of Education (USDOE) guides 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; U.S. Department of Education, 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

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

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

> **Critical Element 6.2**: Achievement standards setting. The state used a technically sound method and process that involved panelists with appropriate experience and expertise for setting its academic achievement standards and academic achievement standards to ensure they are valid and reliable.

Evidence to support this critical element includes:

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

2) Documentation that the method used for setting cut scores allowed panelists to apply their knowledge and experience reasonably and supported the establishment of reasonable and defensible cut scores (Section 5.6, Events; Section 5.6.2, Large-Group Orientation; Section 5.8, Assertion Mapping Results; and Section 6.1, Evidence of Adherence to Professional Standards and Best Practices).

3) Panelists self-reported readiness to undertake the task (Section 5.6.9, Practice Quiz; and Section 5.6.11, Readiness Assertion) and confidence in the workshop process and outcomes supporting the validity of the process (Section 5.8, Assertion Mapping Results; and Section 5.8.1, Round 1 Results).

4) The standard-setting panels consisted of panelists with appropriate experience and expertise, including content experts with experience teaching North Dakota's science content standards, and individuals with experience and expertise teaching special population and general education students in North Dakota (Section 5.3.4, Educator Participants; and Appendix A, NDSA for Science Standard-Setting Panelist Characteristics).

7. References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: Author.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*(2), 153–168.

Cizek, G. J., & Koons, H. (2014). Observation and Report on Smarter Balanced Standard Setting: October 12–20, 2014. Accessed from https://portal.smarterbalanced.org/library/en/standard-setting-observation-and-report.pdf.

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

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

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

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

Huynh, H. (1994, Oct.). Some technical aspects in standard setting. In *Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessment Programs* (co-sponsored by National Assessment Governing Board and National Center for Education Statistics), Washington, DC, October 5–7, 1994, pp. 75–91.

Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. J. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods, and Perspectives* (pp. 53–88). Mahwah, NJ: Lawrence Erlbaum.

Kingston, N. M., Kahl, S. R., Sweeney, K. P., & Bay, L. (2001). Setting performance standards using the body of work method. In G. J. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods, and Perspectives* (pp. 219– 248). Mahwah, NJ: Lawrence Erlbaum.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Greene, D. R. (2001). The Bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting Performance Standards: Concepts, methods, and perspectives* (pp. 249–281). Mahwah, NJ: Lawrence Erlbaum Associates.

National Research Council. (2012). *A framework for K–12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: The National Academies Press.

Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement*, *47*(3), 361−372.

Rijmen, F., Cohen, J., Butcher, T., & Farley, D. (2018, June). Scoring and reporting for assessments developed for the new science standards [Symposium]. National Conference on Student Assessment, San Diego, CA.

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, DC. Accessed from https://www2.ed.gov/policy/elsec/guid/assessguid15.pdf.