NORTH DAKOTA DEPARTMENT OF
PUBLIC INSTRUCTION

North Dakota State Assessment for English Language Arts/Literacy and Mathematics

2020–2021

Volume 3
Setting Achievement Standards

Table of Contents

List of Appendices

LIST OF TABLES

LIST OF FIGURES

1. Executive Summary

The American Institutes for Research (AIR) conducted standard-setting workshops for North Dakota in English language arts (ELA) and mathematics (grades 3–8, & 10) from July 17–19, 2018, at the Ramkota Hotel and Conference Center, 800 South Third Street, Bismarck, North Dakota.

AIR used a process of standard setting called the Bookmark method. The Bookmark method is the most common procedure used throughout the country and has been used successfully to set Achievement Level cut scores for previous assessments in North Dakota.

In the Bookmark method, panelists review a booklet of items ordered in difficulty from easy to hard. The panelists also review the North Dakota College- and Career-Readiness Standards, the Achievement Level Descriptors (ALDs), and student achievement distribution, and then they bookmark the page they feel represents the state's achievement standards – Partially Proficient, Proficient, and Advanced. The panelists do this across two rounds of standard setting in a three-day workshop. These cut scores determine the percentage of students in each of the state's four Achievement Levels – Novice, Partially Proficient, Proficient, and Advanced.

The ELA and mathematics assessments were based on the Independent College and Career Readiness (ICCR) item pools developed by the American Institutes for Research for use in statewide assessments. From this pool, North Dakota selected items that meet the unique needs and requirements of the state and meet the blueprints aligned to the North Dakota College and Career Readiness Standards.

On July 16, 2018, all panelists participated in a stakeholders' meeting where they reviewed, modified, and approved the Achievement Level Descriptors used in the standard-setting workshop.

Seventy-three educators participated as workshop panelists (35 for ELA and 38 for mathematics). The panelists represented a group of experienced teachers, specialists, education administrators, instructional coaches, and other stakeholders. Panel composition ensured that a diverse range of perspectives contributed to the standard-setting process and that the selected panelists were representative in terms of gender, race/ethnicity, special education, EL, and region of the state.

2. Standard-Setting Workshops

2.1 Overall Structure of the Standard-Setting Workshops

Table 1 shows the general structure of the standard-setting workshops.

Table 1: Standard-Setting Workshop Structure

Panel	Room	Subject	Grade	Panelists who are Table Leaders	Panelists	Facilitator	Facilitator Assistant	Other AIR Staff
ELA	1	ELA	3, 4	2	8	Patty Hildreth	Terry Hill	Gary Phillips
		ELA	5, 6	2	7			Nik Kalich
		ELA	7, 8	2	6	Julie Benson	Alex Linville	Ahmet Turhan
		ELA	10	2	6			Doug Rogers
Math	2	Math	3, 4	2	7	Paul Maxon	Eileen Hennegan	Michael Dao (IT)
		Math	5, 6	2	7			Samba Ndiaye (IT)
		Math	7, 8	2	8	Kevin Dwyer	Chris Kincheloe	Curtis Mitchell
		Math	10	2	8			Tiffany Chiu
Totals	2			16	57	4	4	8

The key features of the workshops included the following:

- The standard-setting process produced three cut scores (Partially Proficient, Proficient, and Advanced) per grade.
- There were two rounds of standard setting per grade for all cut scores.
- Panelists considered impact data (percentage of students reaching each cut score) in the second round.
- The standard-setting workshops utilized AIR's online standard-setting tool.

2.2 Results of the Standard-Setting Workshops

Figure 1 and Figure 2 display the Achievement Standards recommended by the standard-setting panelists.

Figure 1: Sample State Summary Performance Report

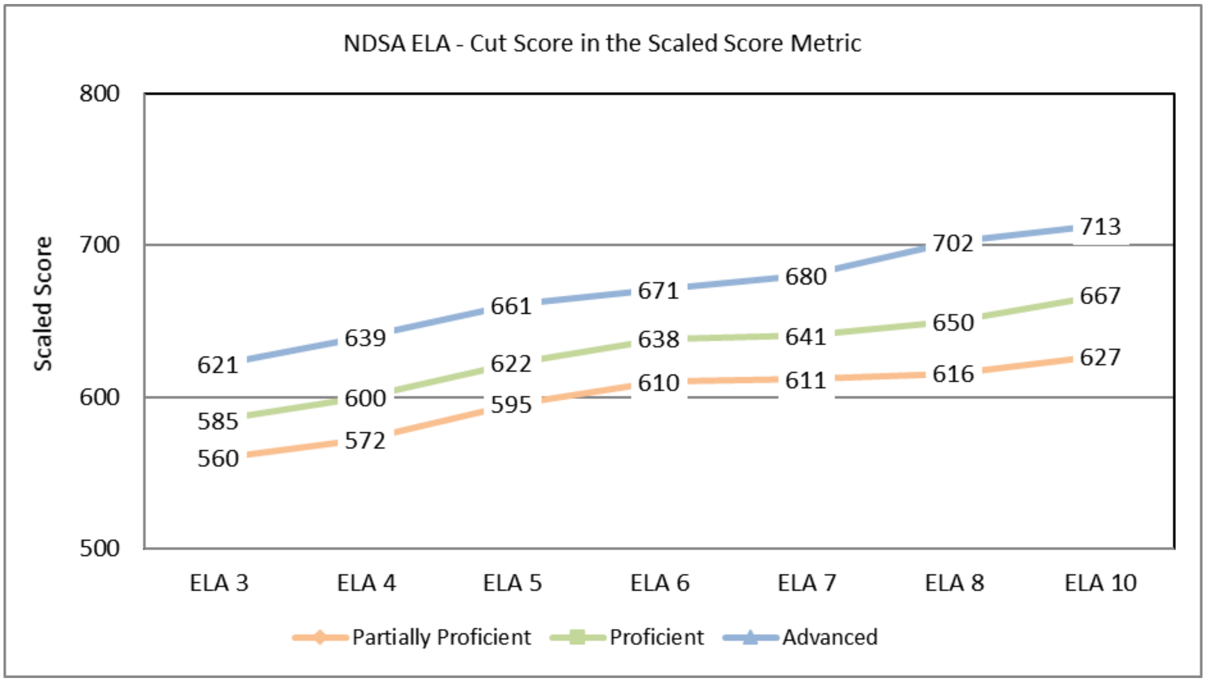


Figure 2. Achievement Standards Recommended for Mathematics

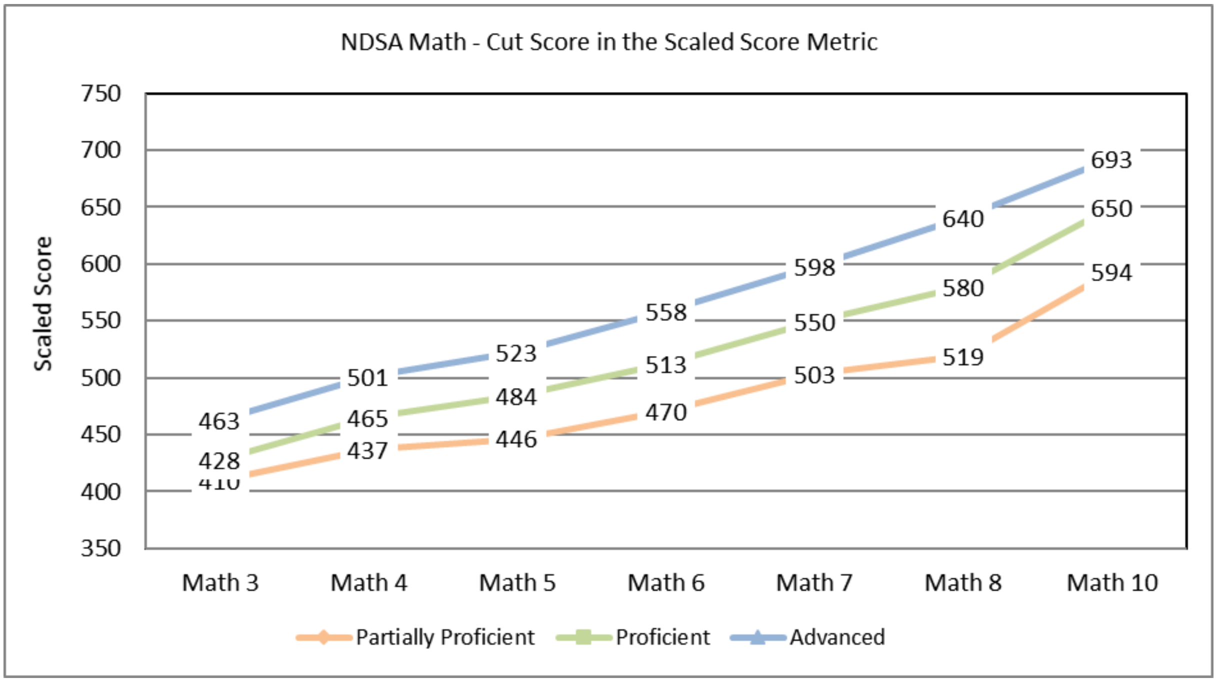
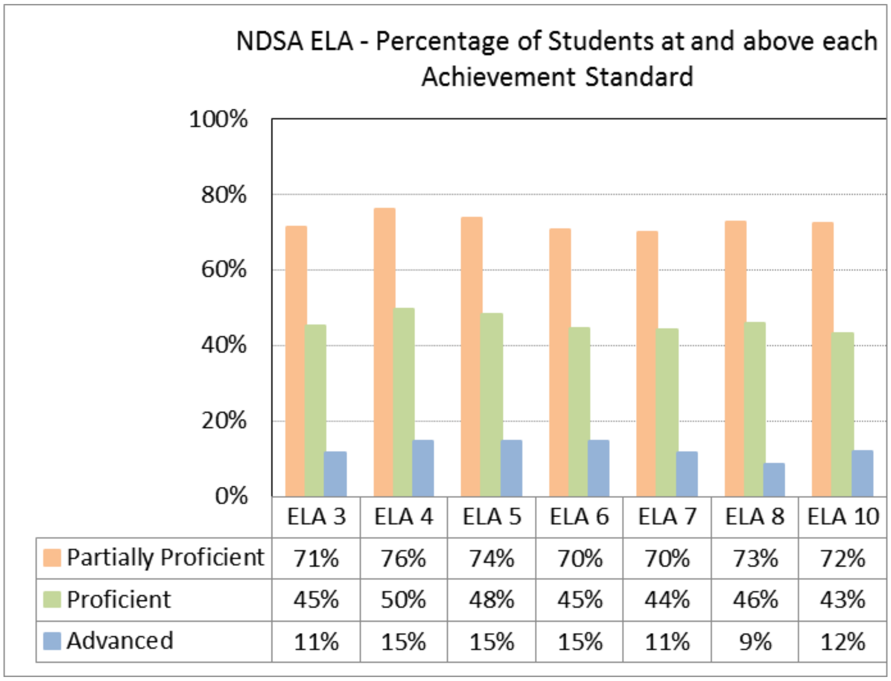


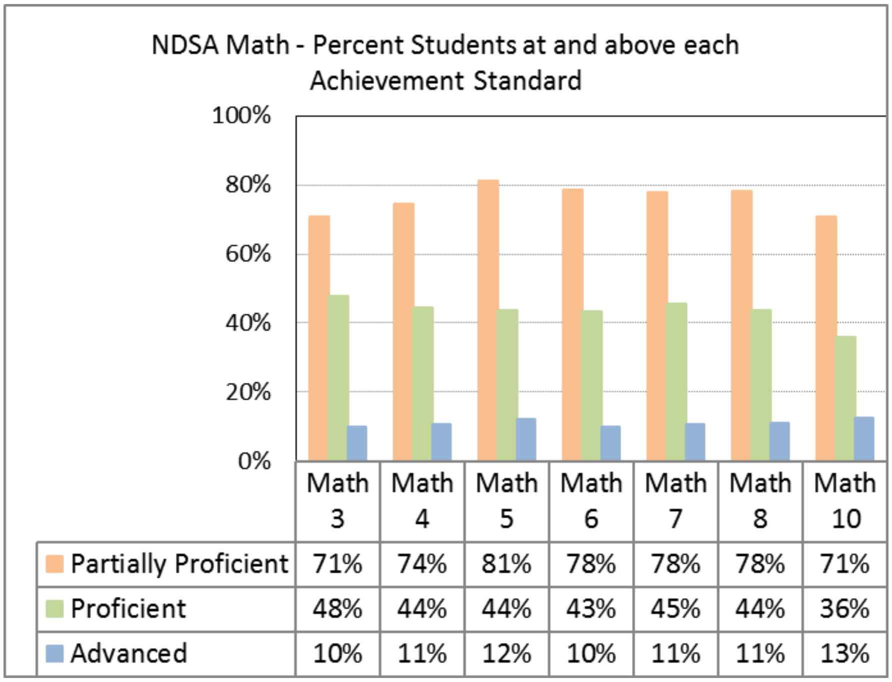
Figure 3 and Figure 4 indicate the percentage of students that we estimate will reach each of the Achievement Standards in 2018.

Figure 3. Percentage of Students at and Above Each Achievement Standard in 2018 ELA



	ELA 3	ELA 4	ELA 5	ELA 6	ELA 7	ELA 8	ELA 10
🟧 Partially Proficient	71%	76%	74%	70%	70%	73%	72%
🟩 Proficient	45%	50%	48%	45%	44%	46%	43%
🟦 Advanced	11%	15%	15%	15%	11%	9%	12%

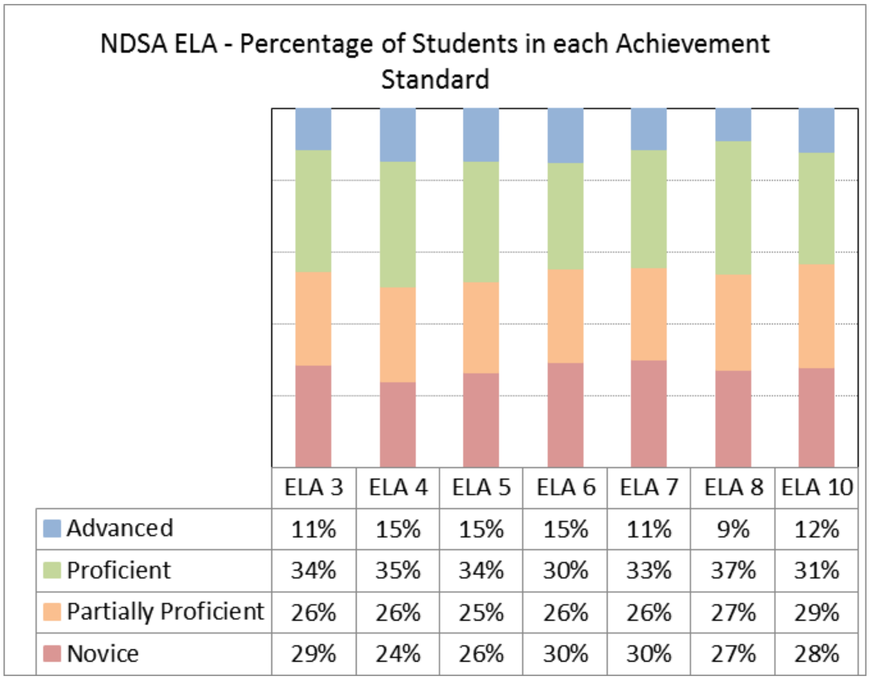
Figure 4. Percentage of Students at and Above Each Achievement Standard in 2018 Mathematics



	Math 3	Math 4	Math 5	Math 6	Math 7	Math 8	Math 10
🟧 Partially Proficient	71%	74%	81%	78%	78%	78%	71%
🟩 Proficient	48%	44%	44%	43%	45%	44%	36%
🟦 Advanced	10%	11%	12%	10%	11%	11%	13%

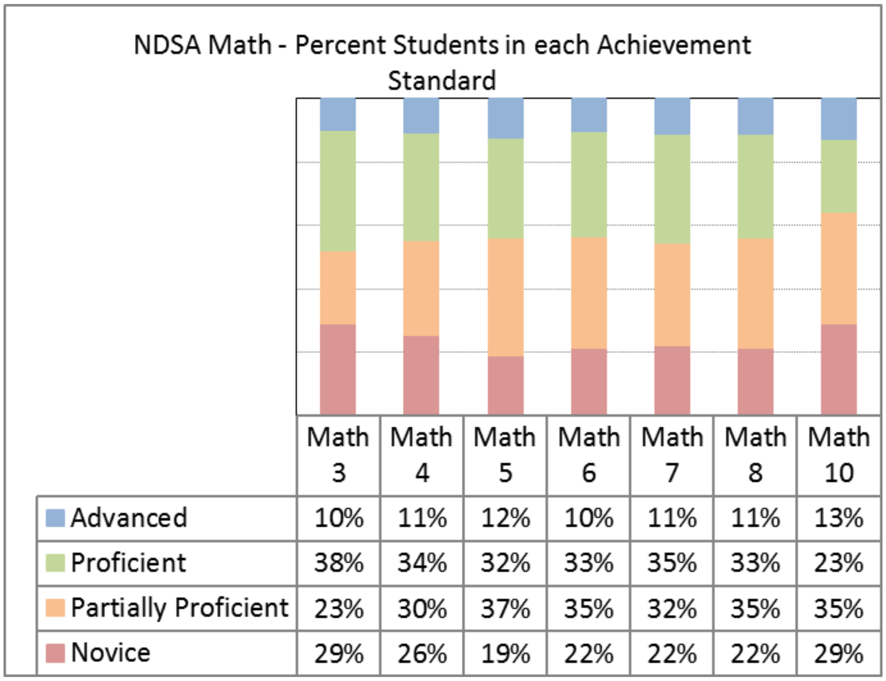
Figure 5 and Figure 6 indicate the percentage of students that we estimate are within each of the Achievement Standards in 2018.

Figure 5. Percentage of Students Within Each Achievement Standard in 2018 ELA



	ELA 3	ELA 4	ELA 5	ELA 6	ELA 7	ELA 8	ELA 10
🟦 Advanced	11%	15%	15%	15%	11%	9%	12%
🟩 Proficient	34%	35%	34%	30%	33%	37%	31%
🟧 Partially Proficient	26%	26%	25%	26%	26%	27%	29%
🟥 Novice	29%	24%	26%	30%	30%	27%	28%

Figure 6. Percentage of Students Within Each Achievement Standard in 2018 Mathematics



	Math 3	Math 4	Math 5	Math 6	Math 7	Math 8	Math 10
🟦 Advanced	10%	11%	12%	10%	11%	11%	13%
🟩 Proficient	38%	34%	32%	33%	35%	33%	23%
🟧 Partially Proficient	23%	30%	37%	35%	32%	35%	35%
🟥 Novice	29%	26%	19%	22%	22%	22%	29%

3. Introduction

The North Dakota Statewide Assessments (NDSA) measure student performance against the state's challenging content and achievement standards in select academic subjects and grades. In spring 2018, the North Dakota Department of Public Instruction (NDDPI) switched from the Smarter Balanced assessment and administered a new state assessment system aligned to the English language arts (ELA) and mathematics standards adopted in April 2017. The online NDSA consists of general, standards-referenced assessments for all students in grades 3–8 and 10 in English language arts/literacy and mathematics, and in grades 4, 8, and 11 in science. In spring 2018, the NDDPI administered new, fixed-form ELA and mathematics assessments developed by the American Institutes for Research (AIR).

The ELA and mathematics assessments were based on the Independent College- and Career-Readiness (ICCR) item pools developed by the American Institutes for Research for use in statewide assessments. North Dakota selected items from the item pool that meet the unique needs and requirements of the state and meet the blueprints aligned to the North Dakota College- and Career-Readiness Standards.

New tests require new achievement standards to link achievement on the test to the content standards. The NDDPI contracted with the American Institutes for Research (AIR) to establish cut scores for the grades 3–8 and 10 ELA and mathematics tests.

To fulfill this responsibility, AIR implemented a defensible, valid, and technically sound method; provided training on standard setting to all participants; oversaw the process; computed real-time feedback data to inform the process; and produced a technical report documenting the method, approach, process, and outcomes.

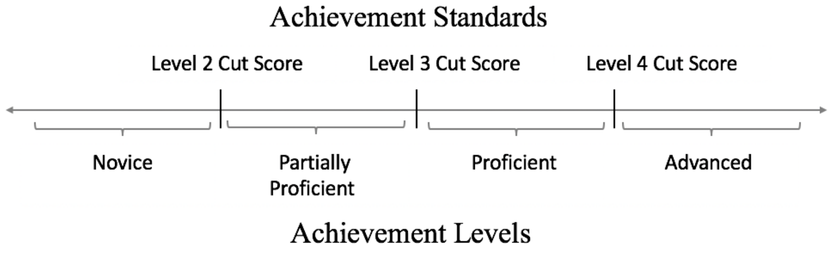
The purpose of this report is to document the standard-setting process and resulting achievement standard recommendations.

4. The Standard-Setting Process

Standard setting is the process used to define achievement on the NDSA. AIR implemented a process of standard setting called the Bookmark method. The Bookmark method is the most common procedure used throughout the country and has been used successfully to set achievement-level cut scores for previous assessments in North Dakota. In the Bookmark method, panelists review a booklet of items ordered in difficulty from easy to hard. The panelists also review the North Dakota College- and Career-Readiness Standards, the Achievement Level Descriptors (ALDs), and student achievement distribution, and then they bookmark the page they feel represents the state's achievement standards – Partially Proficient, Proficient, and Advanced. The panelists do this across two rounds of standard setting in a three-day workshop. These cut scores determine the percentage of students in each of the state's four Achievement Levels.

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 each Achievement Level. As shown in Figure 7, three achievement standards are sufficient to define four Achievement Levels.

Figure 7. 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.

4.1 The Bookmark Method

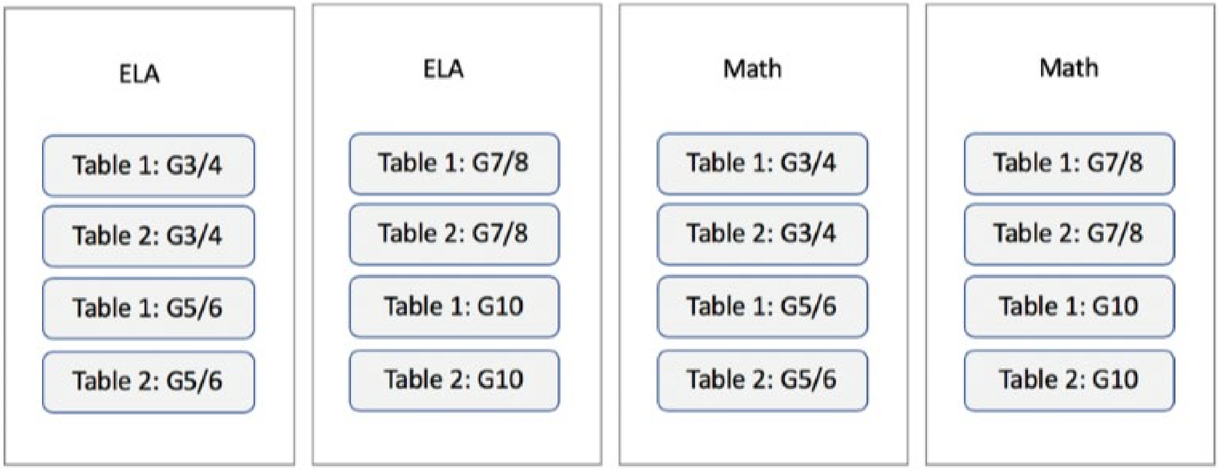
The Bookmark method of standard setting is well suited to support the establishment of cut scores on high stakes tests. It is appropriate for tests, like the NDSA, that are scored using item response theory (IRT) and that use mixed-type items (e.g., multiple choice with one key and selective response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

The Bookmark method derives its name from the primary task required of panelists—the placement of a bookmark in an ordered-item booklet (OIB) to represent a cut score recommendation. Over the course of two rounds of judgments, panelists recommended content-based cut scores using information from the policy descriptors, target student descriptors, test content viewed in the OIBs, panelist discussions, and impact data.[1]

4.2 Workshop Structure

One large meeting room served as the all-participant training room. Four breakout rooms served as workspaces for the subject and grade-level panels. As shown in Table 2, each room contained four tables, allowing for two tables per subject and grade. Each table set standards for two grades: an anchor grade (the even grades) and an adjacent grade (the odd grades).

Table 2. Room Structure



ELA	ELA	Math	Math
Table 1: G3/4	Table 1: G7/8	Table 1: G3/4	Table 1: G7/8
Table 2: G3/4	Table 2: G7/8	Table 2: G3/4	Table 2: G7/8
Table 1: G5/6	Table 1: G10	Table 1: G5/6	Table 1: G10
Table 2: G5/6	Table 2: G10	Table 2: G5/6	Table 2: G10

Table 3 summarizes the setup of the tables and the number of facilitators and panelists assigned to each table. The 73 total standard-setting participants included table leaders and panelists. Each table included four special education teachers and one general education teacher who taught in the content area and grade level for which standards were being set.

Table 3. Table Assignments

Panel	Room	Grade(s)	Table Leaders	Panelists	Facilitator	Facilitator Assistant
ELA	1	3/4	2	8	Patty Hildreth	Terry Hill

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

		5/6	2	7		
	2	7/8	2	6	Julie Benson	Alex Linville
		10	2	6		
Math	3	3/4	2	7	Paul Maxon	Eileen Heneghan
		5/6	2	7		
	4	7/8	2	8	Kevin Dwyer	Chris Kincheloe
		10	2	8		
Totals	4		16	57	4	4

4.3 Participants and Roles

4.3.1 North Dakota Department of Public Instruction Staff

NDDPI staff were present throughout the process, providing overall policy context and answering any policy questions that arose. They included the following staff members:

- Rob Bauer, NDSA Testing Coordinator & Director, Office of Student Assessment
- Bonnie Weisz, Assistant Director, Office of Student Assessment
- Gwyn Marback, NAEP Coordinator, Office of Student Assessment
- Patricia Laubach, Program Administrator, Office of Student Assessment

4.3.2 AIR Staff

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

- Dr. Gary Phillips, Vice President and AIR Institute Fellow
    - Dr. Phillips facilitated and oversaw the workshop. He provided training to all participants, including the facilitators, table leaders, and panelists; he also supervised the psychometric analyses conducted during and after the workshop.
- Doug Rogers, Senior Program Manager
    - Managed logistics throughout the meeting.
- Dr. Ahmet Turhan, Lead Psychometrician
    - Provided psychometric analysis.
- Nik Kalich, Psychometric Support Manager
    - Oversaw the setup of analytics technology and psychometrics.
- Curtis Mitchell and Tiffany Chiu, Psychometric Support Assistants
    - Floated among tables and provided support as needed.
- Michael Dao and Samba Ndiaye, System Support Agents
    - Set up, tested, and performed troubleshooting for technology during the workshop.

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

- Patty Hildreth, Manager of Test Development, and Julie Benson, Senior Manager of Test Development, served as ELA room facilitators.
- Terry Hill and Alex Linville, Test Developers, served as ELA room facilitator assistants.
- Paul Maxon, Director of Test Development, and Kevin Dwyer, Senior Director of Test Development, served as mathematics room facilitators.
- Eileen Heneghan and Chris Kincheloe, Test Developers, served as mathematics room facilitator assistants.

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

4.3.3 Educator Table Leaders

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

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

Table leaders trained as a group early in the morning of the first day. Dr. Gary Phillips from AIR led the training. Training consisted of an overview of their responsibilities and some process guidance.

Table leaders provided the following throughout the workshop:

- Helping panelists see “the big picture”
- Leading table discussions
- Supporting panelists with tasks
- Monitoring security of materials
- Reporting issues or misunderstandings to room facilitators
- Maintaining a supportive atmosphere of professionalism and respect

4.3.4 Educator Participants

Seventy-three educators (35 for ELA and 38 for mathematics) from North Dakota convened to complete two rounds of standard setting to recommend three achievement standards for the NDSA tests in ELA and mathematics.

Recruitment

To set the bookmarks, NDDPI recruited a diverse set of participants from across the state. In recruiting panelists, NDDPI targeted the recruitment of participants to be representative of the gender, geographic, special education, and EL representation of the teacher population found in North Dakota.

Characteristics

The panelists represented a group of experienced educators—including classroom teachers, specialists, education administrators, instructional coaches, four higher education faculty, and other stakeholders—to ensure that a diverse range of perspectives contributed to the standard-setting process and product. They represented both schools and districts and represented the state in terms of district size and urbanicity. Table 4 summarizes characteristics of the panels.

Table 4. Panelist Characteristics

	Percentage of Panelists	
	ELA	**Math**
Male	3%	21%
Non-White	0%	3%
Position		
Educator	69%	89%
Administrator	0%	5%
Specialist	17%	8%
Coach	9%	3%
Other	11%	8%
Location of Current Position		
School	89%	92%
District	17%	24%
Other	6%	3%
District Size		
Large	26%	47%
Medium	40%	26%
Small	34%	24%
District Location		
Urban	20%	50%
Suburban	14%	11%
Rural	57%	37%

Note. Values are rounded to the nearest whole percentage. Totals for position and location of current position do not equal 100% due to panelists being able to check off more than one box when providing information. "Other" stakeholders included a school psychologist, higher education faculty, and educational strategists.

Qualifications

For results of the Bookmark method to be valid, the judgments must be made by individuals who are qualified to make them. Participants in the North Dakota standard-setting workshop were highly qualified. They brought a variety of expertise in instruction, curriculum, assessment, and student populations. Most had taught for 12 years or more, with half of those teaching in their assigned grade for 6 or more years. Many had professional experience in addition to teaching, and most had a master's degree or higher. Table 5 summarizes the qualifications of the panels.

Table 5. Panelist Qualifications

	Percentage of Panelists	
	ELA	**Math**
Years of Teaching Experience		
5 Years or Less	14%	16%
6 to 10 Years	20%	21%
11 Years or More	66%	63%
Years of Teaching in Assigned Grade/Subject		
5 Years or Less	46%	42%
6 to 10 Years	11%	16%
11 Years or More	43%	42%
Highest Degree Earned		
Bachelor's	43%	18%
Master's	51%	76%
Doctorate	6%	5%
Other	0%	0%
Percent with Professional Experience (in addition to teaching)	43%	36%

Note. Values are rounded to the nearest whole percentage.

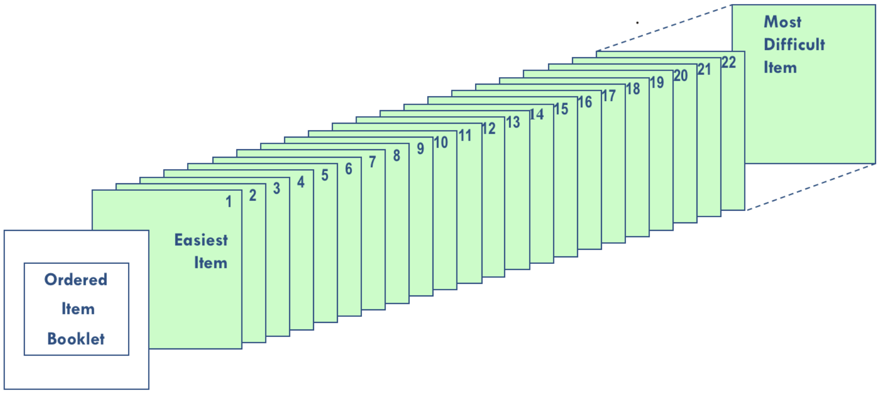
Appendix A identifies the individual panelists.

4.4 Materials

4.4.1 Ordered-Item Booklets

The Bookmark method utilizes ordered-item booklets (OIBs) as the key tool for setting standards. OIBs contain sets of test items ordered by difficulty by grade and subject. Each page of the online OIB presents a single item, with the easier items located in the front of the OIB and the more difficult items in the back of the OIB. Item difficulty, and thus item ordering, is determined by analyses of actual student achievement on the items. Panelists use the OIBs to place bookmarks that identify sets of items students meeting each standard should be able to answer correctly. Each page of the OIB corresponds to a cut score; thus, when panelists place their "bookmarks" for each Achievement Level, they are in fact selecting the achievement standard, indicated by the RP50 value of the item, for that level.

Figure 8. Ordered-Item Booklet (OIB)



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

For North Dakota, the OIBs ranged from 85 to 110 pages.

4.4.2 Content Standards

The North Dakota College- and Career-Readiness Standards provide a rigorous and content appropriate framework for instruction to increase student achievement. They describe expectations for what students should know and be able to do at each grade in each subject area. North Dakota's ELA and mathematics content standards were written by groups of North Dakota mathematics and English teachers.

NDDPI provides the content standards at the following link: https://www.nd.gov/dpi/SchoolStaff/Standards

4.4.3 Achievement Level Descriptors

A prerequisite to standard setting is to determine the nature of the categories into which students are classified. These categories, or Achievement Levels, are associated with Achievement Level Descriptors (ALDs). ALDs link the standards to the achievement standards and describe what a student knows and can do to demonstrate proficiency on a content standard.

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 prior to the workshop, these detailed grade- and content-area-specific descriptions communicate exactly what students performing at each level know and can do.
3. Target ALDs: Typically created during and used for standard setting only, these describe what a student just barely scoring into each Achievement Level knows and can do.
4. Reporting ALDs: These are abbreviated range 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: (1) "Novice," (2) "Partially Proficient," (3) "Proficient," (4) "Advanced."

When developing Achievement Level Descriptors (ALDs) for the North Dakota State Assessments, AIR started with the Common Core State Standards in mathematics and English language arts/literacy (ELA/L). High-level policy ALDs were written first, to clearly define what it means at the achievement level to be college and career ready. Then, AIR adapted these policy ALDs to the other achievement levels, describing what a student's achievement would look like at each point on the continuum.

With policy ALDs in place, range ALDs were written for each assessed standard. AIR started with the language of the standard as the achievement level, adapting it to show how the performance of a student would differ at the "advanced," "partially proficient," and "novice" levels. As AIR moved toward the "advanced" level in ELA, for example, language like "complex inference" was used, rather than simply "inference" to indicate that higher-performing students would be expected to read and draw conclusions from more complex texts. These range ALDs offer observable evidence of student achievement within each standard, and they change and become more (or less) sophisticated across achievement levels.

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

With the ALDs in place, North Dakota state standards were reviewed to identify any standards that differed from the Common Core State Standards. In cases where the standards differed, a unique range ALD was written to represent that standard.

The North Dakota Department of Public Instruction then reviewed the ALDs to ensure that the language accurately represented the goals and policies of their state. AIR worked with them to make revisions where necessary.

On July 16, 2018, all panelists participated in a stakeholder's meeting where they reviewed, modified and approved the Achievement Level Descriptors (ALDs) used in the standard-setting workshop. They provided a final level of stakeholder review by verifying there were no errors in the ALDs and confirming that the ALDs aligned to the standards and progressed in rigor from one level to the next. Panelists approved the draft ALDs, some with minor recommendations for NDDPI consideration.

The ALDs can be found in Appendix E.

4.4.4 Workshop Technology

Panelists used AIR's online standard setting tool. Using this application, panelists placed multiple rounds of bookmarks, reviewed the content alignment and score points for each item, and evaluated the impact that proposed cut scores would have on students. During designated results rounds, panelists saw their own bookmarks, their table's bookmarks, the other tables' bookmarks, and the overall bookmarks for both tables. They were able to add notes and comments on the items as they reviewed them and examine reference and benchmark data onscreen following each round. Benchmark data allowed panelists to compare spring 2018 cut score impact to 2017 results in North Dakota, using the National Assessment of Educational Progress (NAEP).

Each panelist was provided an AIR laptop or Chromebook on which they took the test, reviewed items and ancillary materials, and placed bookmarks.

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

4.5 EVENTS

The standard-setting workshop occurred over a period of three days. Table 6 summarizes each day's events, and this section describes each event listed in greater detail. Appendix B provides the full workshop agenda.

Table 6. Standard Setting Agenda Summary

Day 1: Tuesday, July 17	Day 2: Wednesday, July 18	Day 3: Thursday, July 19
• Table leader training • Orientation and introductions • Large group training • Take the test • Review content standards • Review anchor grade range ALDs • Create anchor grade target student descriptions (ALDs) • Review anchor grade OIBs	• Review RP50 • Practice placing bookmarks • Standard setting readiness evaluation • Place round 1 anchor grade bookmarks • Review round 1 feedback, impact data, and benchmark data and discuss • Place round 2 anchor grade bookmarks • Anchor grade moderation • Review adjacent grade range ALDs • Create adjacent grade target student descriptions (ALDs)	• Review adjacent grade OIBs • Place round 1 adjacent grade bookmarks • Review round 1 feedback, impact data, and benchmark data and discuss • Place round 2 adjacent grade bookmarks • Review round 2 feedback, impact data, and benchmark data and discuss • Standard setting workshop evaluations • Final moderation

4.5.1 Orientation

Robert Bauer, State Testing Director from the North Dakota Department of Public Instruction (NDDPI), and Gary Phillips, Vice President and AIR Institute Fellow from the American Institutes for Research (AIR), welcomed panelists to the workshop.

Dr. Phillips then described the purpose and objectives of the meeting, explained the process that would meet those objectives, and outlined the events that would happen each day. He outlined the responsibilities of the three groups of people at the workshop: panelists, AIR staff, and NDDPI personnel. He explained that panelists were selected because they were experts and described how the process to be implemented over the three days was designed to elicit and apply their expertise to recommend new cut scores. He described how standard setting works and what would happen once the panelists had finalized their recommendations.

4.5.2 Confidentiality and Security

Standard setting uses live test items from the operational NDSA tests, requiring confidentiality to maintain their security. Participants were not allowed to do the following during or after the workshop:

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

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

4.5.3 Take the Test

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

Taking the same test that 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.

4.5.4 Review Content Standards and Draft ALDs

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

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

4.5.5 Write Target ALDs

After reviewing and discussing the ALDs, panelists worked in their table groups to draft target ALDs that described the skills that students just barely in one Achievement Level have that students just below the Achievement Level don't have. Target ALDs describe students who are not typical of students at an Achievement Level; although, at "Just Barely," they do reach the standard.

4.5.6 Place Bookmarks

Bookmark Placement Training

The objective of standard setting is aspirational: to identify what all students *should* know and be able to do, not what they *actually* know and can do. To accomplish this, panelists think about the target ALDs that describe students "just barely" meeting each Achievement Level as they review the OIB.

Panelists applied a 50/50 response probability rule when placing bookmarks. This rule requires panelists to identify the page in the OIB where 50% of students who "just barely" meet the standard (those described by the target ALDs) should be able to get the item on that page correct.[2]

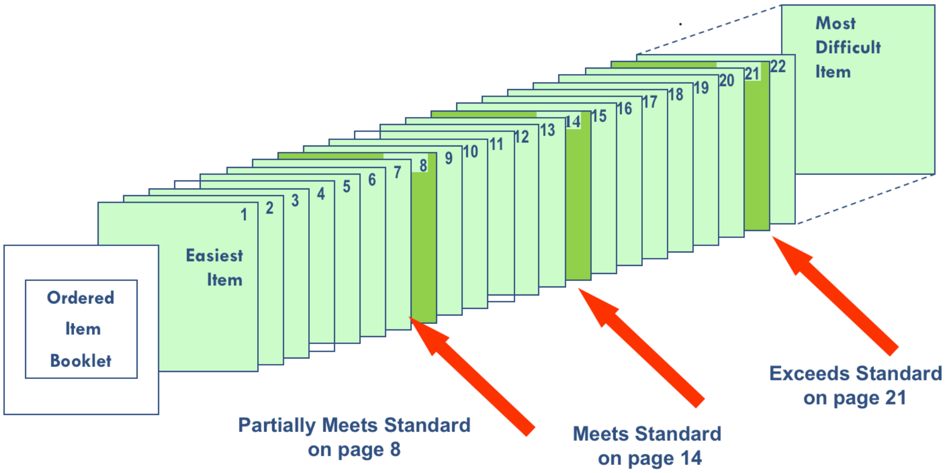
The explanation of this rule provided to panelists was as follows:

> *"Of 100 students who are 'just barely' at the standard, what percentage would get this item correct?"*

These "just barely" students are more likely to be able to correctly answer items at the beginning of the OIB and are less likely to be able to correctly answer items towards the end of the OIB. As panelists work through the OIB, they will come across an item, or small group of items, where they think that about half of the "Just Barely Meets Standard" students (for example) would get the item correct. Items before that point in the OIB are items that more than half of the "Just Barely Meets Standard" students would correctly answer. Items beyond that point in the OIB are items that less than half of the "just barely" students would correctly answer. Panelists place their bookmark on the first page in the OIB where they believe the "Just Barely Meets Standard" student would NOT have at least a 50% chance of answering correctly. Panelists repeated this process for the "Just Barely Emerging" student and the "Just Barely Exceeds Standard" student.

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

Figure 9. Example Bookmark Placement



Workshop leaders from AIR and NDDPI advised panelists that while some items may seem out of order, the item order is determined by item difficulty, which is computed from actual student achievement on the items and not by content or cognitive process. The ordering of items in the OIB does not follow the sequence of instruction or the order of item presentation on the test.

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

OIB Review

After completing the target ALDs, panelists independently reviewed all items in the online OIB. For each item, they could take notes on the items that would help them as they placed the round 1 and round 2 bookmarks. Suggested review steps included noting what students need to know and be able to do to correctly answer each item and identifying what made each item more difficult than the one before. Following OIB review, panelists practiced placing bookmarks in the OIB.

Bookmark Placement Practice Round

The purpose of the practice round was to ensure that panelists were comfortable with the technology and item types prior to setting any actual bookmarks. On the afternoon of the second day, panelists practiced placing a series of cut scores. They used an abbreviated OIB designed to give them an understanding of the bookmarking process and how to set cut scores using AIR's online tool. Panelists asked questions, and the room facilitators provided clarifications and further instructions until everyone had completed the practice round.

Bookmark Placement Readiness Assessment

Prior to placing bookmarks, panelists completed a practice quiz and readiness assessment. The quiz assessed panelists' understanding in multiple ways. They must

- indicate on a diagram of how achievement standards and levels work together, where students "just barely" meeting each of the standards fall;
- answer questions about relative item difficulty in a hypothetical OIB; and
- demonstrate understanding by correctly applying the 50/50 rule to a hypothetical bookmark placement.

After completing the quiz, panelists affirm that they understand how to set bookmarks. Every panelist must affirm readiness on the readiness assessment before placing bookmarks in both rounds of the workshop. Any panelist who is unable to affirm understanding is not allowed to place bookmarks.

All panelists demonstrated understanding of the task necessary to pass the readiness assessment.

Round 1 Bookmark Placement

In round 1, panelists set the bookmarks based on the difficulty and content of the items for "Proficient," then for "Partially Proficient," and lastly, for "Advanced." Each panelist independently placed his or her own bookmark. The median of the individual bookmarks across each table or grade level became the round 1 cut score.

Table 7 presents the bookmarks and associated impact data for round 1.

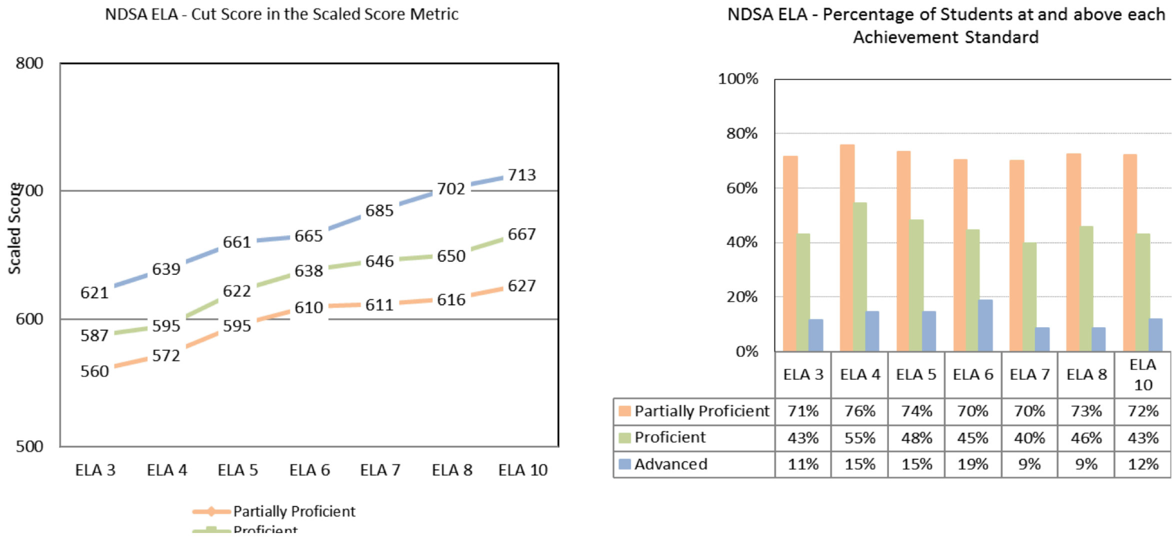
Table 7. Round 1 Results

Table	Median Round 1 Bookmark (Page #)			Impact Data (Percent at or Above)		
	PP	P	A	PP	P	A
ELA						
Grade 3	13	27	50	71.5	43.0	11.5
Grade 4	16	33	60	81.9	57.6	9.5
Grade 5	15	30	48	73.5	48.1	14.5
Grade 6	17	34	48	70.5	40.6	19.6
Grade 7	13	28	49	70.0	39.6	8.7
Grade 8	15	30	51	72.6	45.8	8.2
Grade 10	18	32	53	72.2	43.1	11.9
Mathematics						
Grade 3	15	34	53	70.8	36.4	9.9
Grade 4	19	39	53	59.6	28.8	10.5
Grade 5	9	24	49	74.6	43.8	12.2
Grade 6	14	33	57	77.4	30.5	4.5
Grade 7	14	38	56	76.9	45.5	10.8
Grade 8	15	35	62	80.7	44.9	11.0
Grade 10	22	37	55	41.5	19.5	4.6

Note. Each grade-level row summarizes the data across both tables.
Achievement Level abbreviation key: Partially Proficient (PP), Proficient (P), Advanced (A).

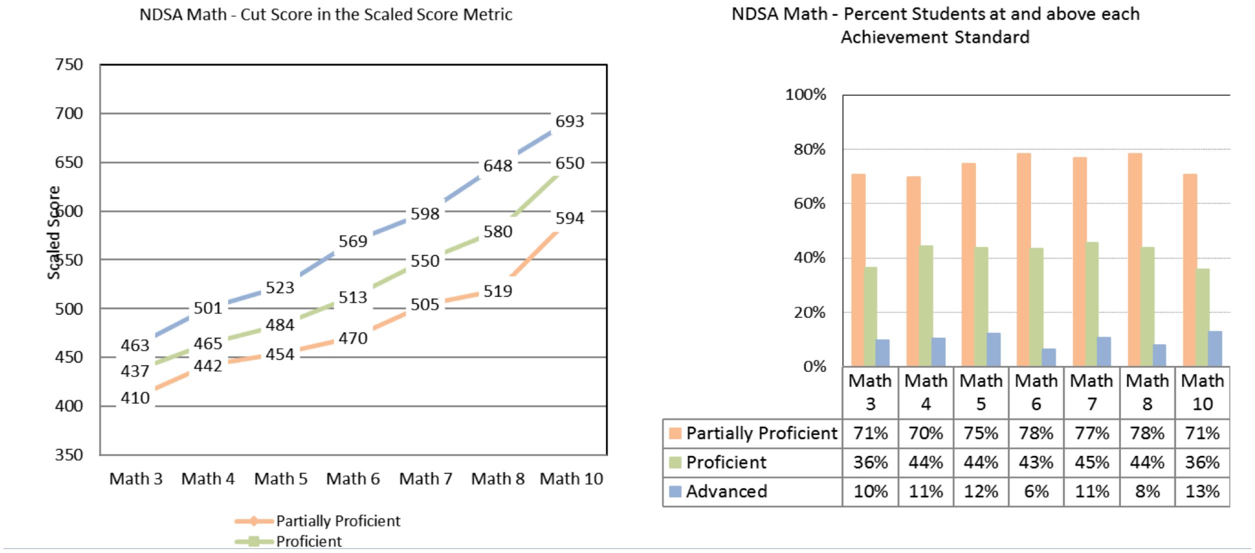
Figure 10 and Figure 11 present the round 1 bookmarks as scaled scores and graphically depict impact data.

Figure 10. Round 1 ELA Cut Scores and Impact Data



	ELA 3	ELA 4	ELA 5	ELA 6	ELA 7	ELA 8	ELA 10
🟧 Partially Proficient	71%	76%	74%	70%	70%	73%	72%
🟩 Proficient	43%	55%	48%	45%	40%	46%	43%
🟦 Advanced	11%	15%	15%	19%	9%	9%	12%

Figure 11. Round 1 Mathematics Cut Scores and Impact Data



	Math 3	Math 4	Math 5	Math 6	Math 7	Math 8	Math 10
🟧 Partially Proficient	71%	70%	75%	78%	77%	78%	71%
🟩 Proficient	36%	44%	44%	43%	45%	44%	36%
🟦 Advanced	10%	11%	12%	6%	11%	8%	13%

After placing the round 1 bookmarks, workshop facilitators provided panelists with additional instruction for placing the round 2 bookmarks. First, they described the goal of round 2 as one of convergence, not consensus, on a common achievement standard. A second goal was articulation of bookmarks across grade levels. Panelists reviewed and discussed example sets of cut scores across grades 3–8, showing multiple ways of disarticulation, until they understood why articulation was important and should be a consideration in placing the round 2 bookmarks.

Workshop leaders discussed the impact of cut scores from a policy perspective and presented impact data showing the percentage of students who would score at or above each Achievement Level, given the bookmarks set in round 1. They also provided feedback data (the cut scores placed by other members of their own table), the median bookmark from the other grade-level/subject table, and the median bookmark overall, across both tables.

This information was provided to inform, but not determine, their round 2 decisions.

Round 2 Bookmark Placement

Panelists discussed the feedback data, impact data, and articulation associated with the median round 1 bookmark. Panelists then independently placed their round 2 bookmarks.

Round 2 results were also well articulated. Table 8 presents the bookmarks and associated impact data for round 2.

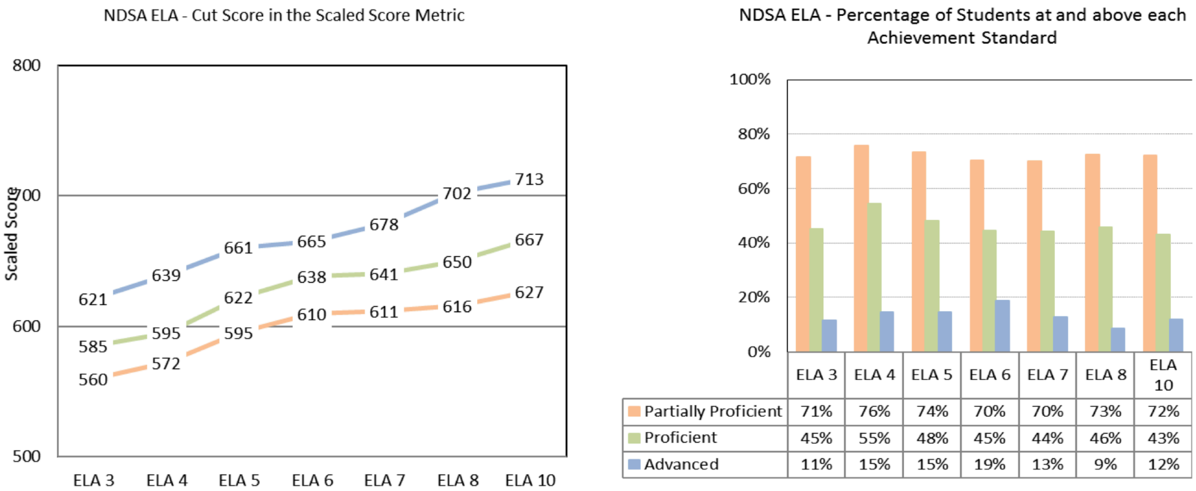
Table 8. Round 2 Results

Table	Median Round 2 Bookmark (Page #)			Impact Data (Percent at or Above)		
	PP	P	A	PP	P	A
ELA						
Grade 3	13	25	50	71.5	45.3	11.5
Grade 4	18	35	56	75.9	54.5	14.6
Grade 5	15	30	48	73.5	48.1	14.5
Grade 6	17	32	50	70.5	44.6	18.7
Grade 7	13	26	45	70.0	44.2	12.8
Grade 8	15	30	50	72.6	45.8	8.7
Grade 10	18	32	53	72.2	43.1	11.9
Mathematics						
Grade 3	15	28	53	70.8	47.7	9.9
Grade 4	12	27	53	69.8	44.5	10.5
Grade 5	8	24	49	81.1	43.8	12.2
Grade 6	13	29	54	78.5	43.3	6.3
Grade 7	12	38	58	81.6	45.5	5.6
Grade 8	17	36	65	78.4	43.8	7.9
Grade 10	11	24	43	70.8	35.8	12.7

Note. Each grade-level row summarizes the data across both tables.
Achievement Level abbreviation key: Partially Proficient (PP), Proficient (P), Advanced (A).

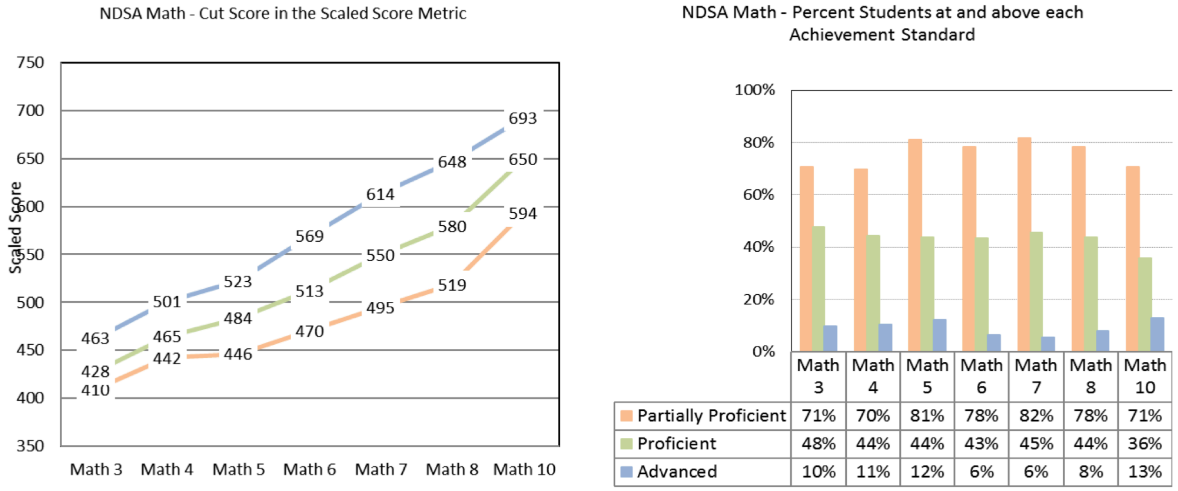
Figure 12 and Figure 13 present the round 2 bookmarks as scaled scores and graphically depict impact data.

Figure 12. Round 2 ELA Cut Scores and Impact Data



	ELA 3	ELA 4	ELA 5	ELA 6	ELA 7	ELA 8	ELA 10
Partially Proficient	71%	76%	74%	70%	70%	73%	72%
Proficient	45%	55%	48%	45%	44%	46%	43%
Advanced	11%	15%	15%	19%	13%	9%	12%

Figure 13. Round 2 Mathematics Cut Scores and Impact Data



	Math 3	Math 4	Math 5	Math 6	Math 7	Math 8	Math 10
Partially Proficient	71%	70%	81%	78%	82%	78%	71%
Proficient	48%	44%	44%	43%	45%	44%	36%
Advanced	10%	11%	12%	6%	6%	8%	13%

Appendix C and Appendix D provide box and whisker plots for the results of each round.

4.5.7 Moderation

Achievement standards for a statewide system must be coherent across grades and subjects. There should be no irregular peaks and valleys, and they should be orderly across subjects with no dramatic differences in expectation. On the last day of the workshop, table leaders and panelists met to discuss and resolve any issues or needs related to cross-grade articulation.

Table leaders and panelists who wanted to attend moderation viewed the round 2 bookmarks for all subjects and grades with the associated impact and benchmark data.

To better articulate across grades and to even the expectations across Achievement Levels, ELA panelists considered placing the grade 6 and 7 bookmark for Exceeds later in the OIB (raising the standard), and the grade 4 bookmark for Proficient earlier in the OIB (lowering the standard). Mathematics panelists considered placing the grade 6, 7, and 8 bookmarks for Advanced and the grade 4 bookmark for Partially Proficient earlier in the OIB (lowering the standard) and the grade 7 bookmark for Partially Proficient later in the OIB (raising the standard). Table 9 provides their final recommendations reflecting these considerations.

Table 9. Moderated Results

Table	Final Bookmarks (Page #)			Impact Data (Percent at or Above)		
	PP	P	A	PP	P	A
ELA						
Grade 3	13	25	50	71.5	45.3	11.5
Grade 4	18	39*	56	75.9	49.5*	14.6
Grade 5	15	30	48	73.5	48.1	14.5
Grade 6	17	32	52*	70.5	44.6	14.8*
Grade 7	13	26	46*	70.0	44.2	11.5*
Grade 8	15	30	50	72.6	45.8	8.7
Grade 10	18	32	53	72.2	43.1	11.9
Math						
Grade 3	15	28	53	70.8	47.7	9.9
Grade 4	9*	27	53	74.4*	44.5	10.5
Grade 5	8	24	49	81.1	43.8	12.2
Grade 6	13	29	50*	78.5	43.3	10.1*
Grade 7	13*	38	56*	77.9*	45.5	10.8*
Grade 8	17	36	62*	78.4	43.8	11.0*
Grade 10	11	24	43	70.8	35.8	12.7

Note. Each grade-level row summarizes the data across both tables.
Achievement Level abbreviation key: Partially Proficient (PP), Proficient (P), Advanced (A).
**Bookmark changed during moderation.*

Figure 14 and Figure 15 display the achievement standards recommended by the standard-setting panelists.

Figure 14. Final Achievement Standards Recommended for ELA

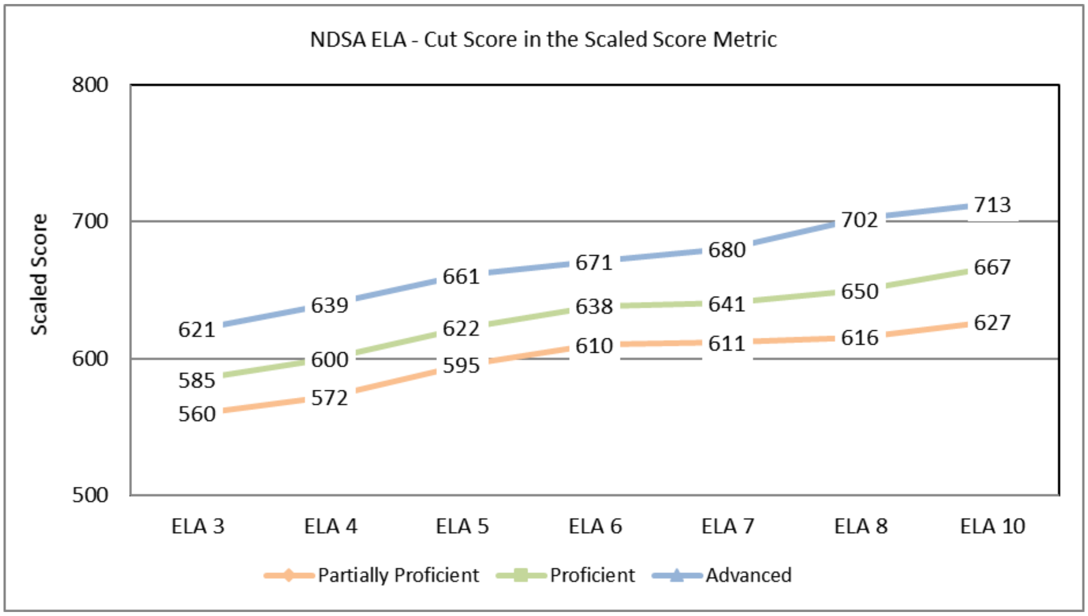


Figure 15. Final Achievement Standards Recommended for Mathematics

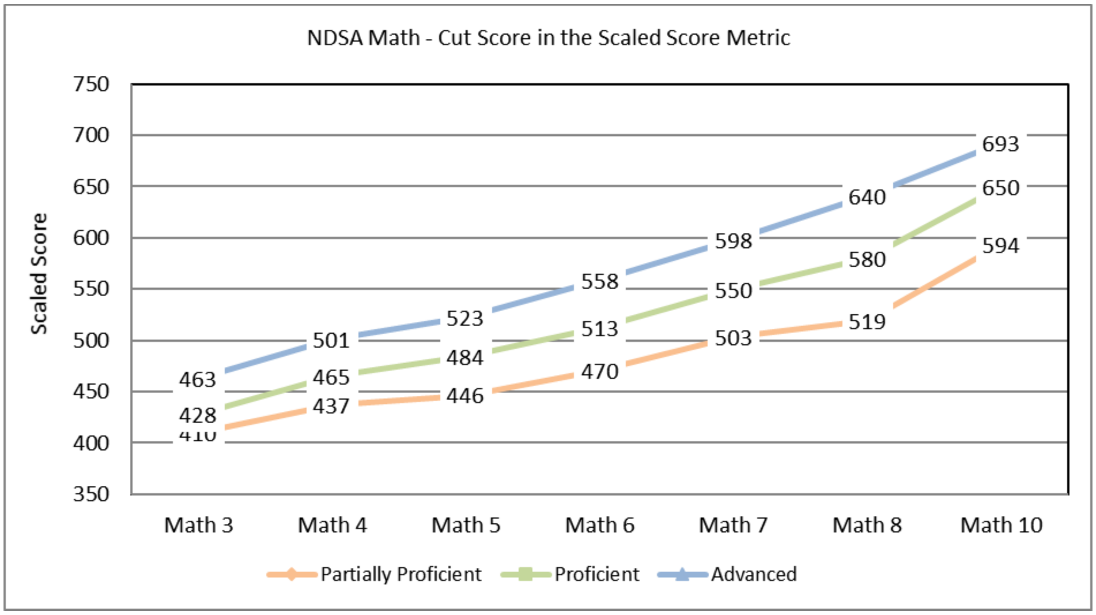
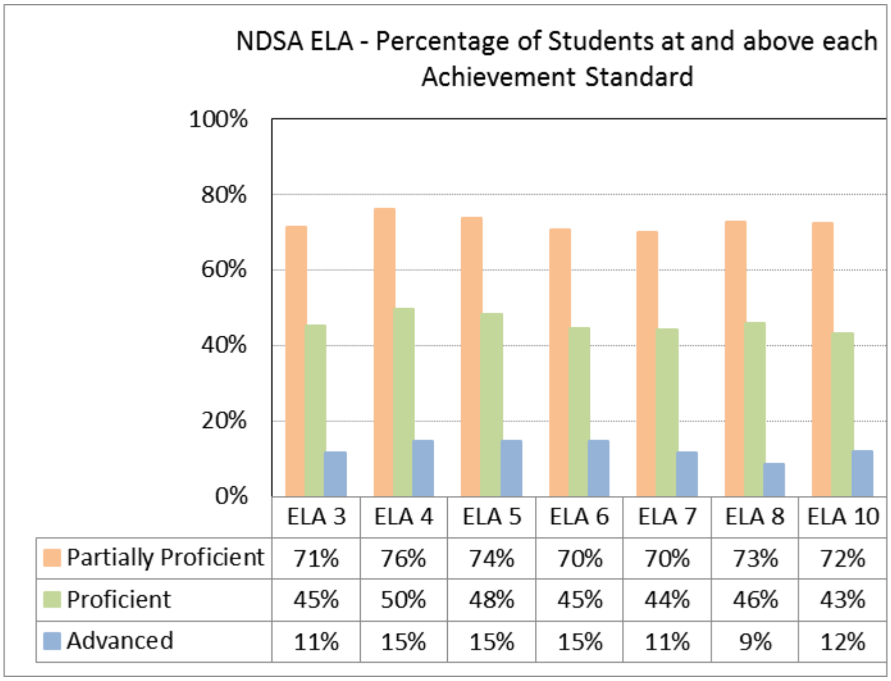


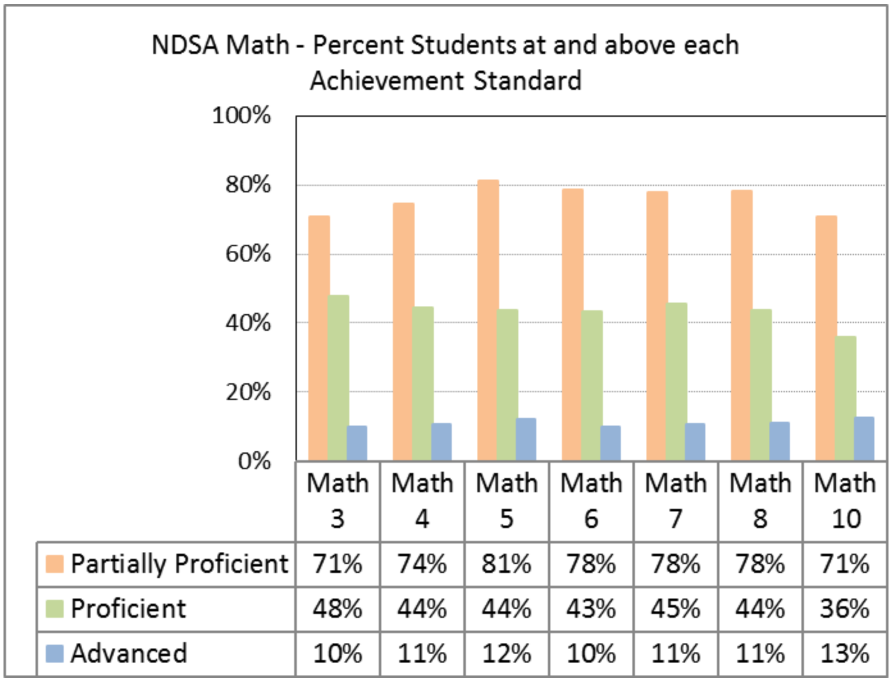
Figure 16 and Figure 17 provide the percentage of students estimated to reach each of the achievement standards in 2018.

Figure 16. Percentage of Students at and Above Each Achievement Standard in 2018 ELA



	ELA 3	ELA 4	ELA 5	ELA 6	ELA 7	ELA 8	ELA 10
🟧 Partially Proficient	71%	76%	74%	70%	70%	73%	72%
🟩 Proficient	45%	50%	48%	45%	44%	46%	43%
🟦 Advanced	11%	15%	15%	15%	11%	9%	12%

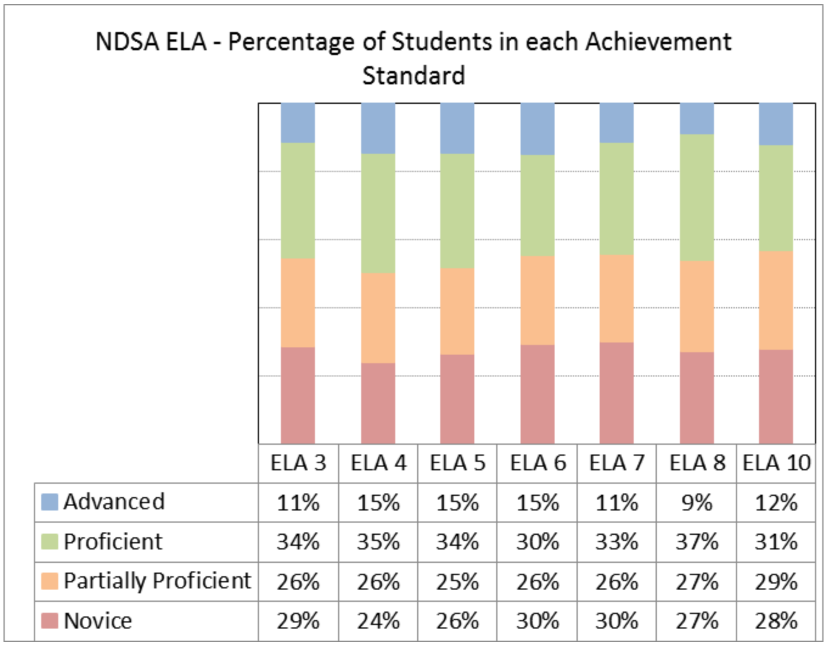
Figure 17. Percentage of Students at and Above Each Achievement Standard in 2018 Mathematics



	Math 3	Math 4	Math 5	Math 6	Math 7	Math 8	Math 10
🟧 Partially Proficient	71%	74%	81%	78%	78%	78%	71%
🟩 Proficient	48%	44%	44%	43%	45%	44%	36%
🟦 Advanced	10%	11%	12%	10%	11%	11%	13%

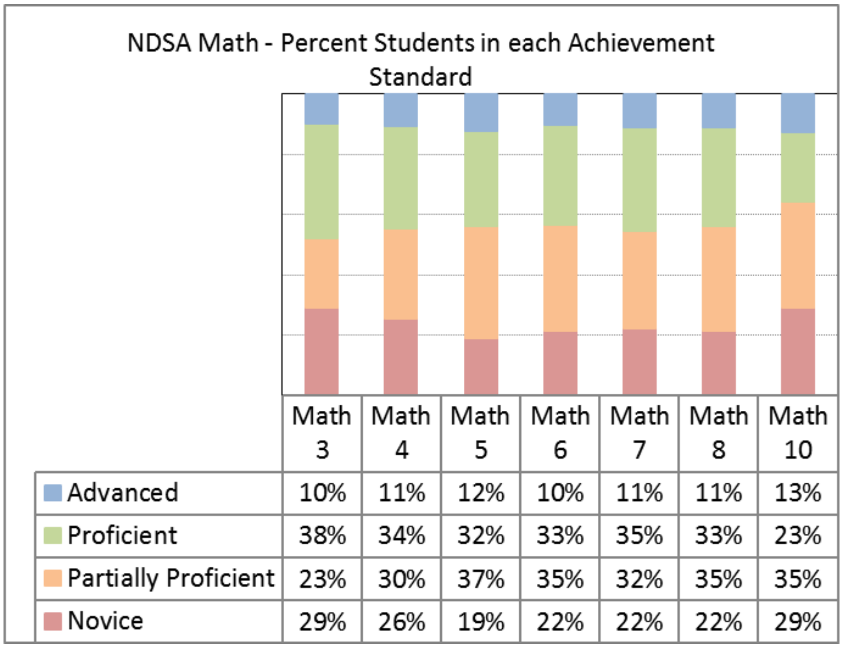
Figure 18 and Figure 19 describe the percentage of students estimated within each of the Achievement Standards in 2018.

Figure 18. Percentage of Students Within Each Achievement Standard in 2018 ELA



	ELA 3	ELA 4	ELA 5	ELA 6	ELA 7	ELA 8	ELA 10
🟦 Advanced	11%	15%	15%	15%	11%	9%	12%
🟩 Proficient	34%	35%	34%	30%	33%	37%	31%
🟧 Partially Proficient	26%	26%	25%	26%	26%	27%	29%
🟥 Novice	29%	24%	26%	30%	30%	27%	28%

Figure 19. Percentage of Students Within Each Achievement Standard in 2018 Mathematics



	Math 3	Math 4	Math 5	Math 6	Math 7	Math 8	Math 10
🟦 Advanced	10%	11%	12%	10%	11%	11%	13%
🟩 Proficient	38%	34%	32%	33%	35%	33%	23%
🟧 Partially Proficient	23%	30%	37%	35%	32%	35%	35%
🟥 Novice	29%	26%	19%	22%	22%	22%	29%

4.5.8 Workshop Evaluations

After finishing all activities, panelists completed online meeting evaluations independently, in which they described and assessed their experience taking part in the standard setting using the Bookmark method.

Workshop participants indicated clarity in the instructions, materials, data, and process (see Table 10), although eight mathematics panelists indicated that the ALDs were “somewhat unclear.” They were spread across table and grade—with three from grade 3/4, two from grade 5/6, one from grade 7/8, and two from grade 10—and generally indicated that they wanted the standards and Achievement Level Descriptors to be more state-specific to North Dakota.

Table 10. Evaluations: Clarity of Materials and Process

Please rate the clarity of the following components of the workshop.	Percent “Somewhat Clear” or “Very Clear”	
	ELA	Math
Instructions provided by the Workshop Leader	100%	91%
Achievement Level Descriptors (ALDs)	98%	79%
Ordered-Item Booklet (OIB)	100%	100%
Panelist agreement data	100%	100%
Impact data (percentage of students that would achieve at the level indicated by the OIB page)	100%	100%

Abbreviation key: Number of responses = 66. Evaluation options included “Very Clear,” “Somewhat Clear,” “Somewhat Unclear,” and “Very Unclear.”

Participants felt that they had sufficient time to complete all activities. In fact, some indicated having too much time to complete some tasks (see Table 11). Participants indicated that the discussion of target ALDs, review of OIBs, and orientation session could have been shorter.

For the large group orientation, all participants not selecting “About Right” selected “Too Much,” indicating that time allowed for the orientation may have been too long.

Six ELA panelists reported having too little time for taking the test, five mathematics panelists (four from grade 10) indicated not having enough time to review the ALDs, three mathematics panelists indicated not having enough time to place their bookmarks, and six mathematics and three ELA panelists reported having too much time to do so.

Table 11. Evaluations: Appropriateness of Process

How appropriate was the amount of time you were given to complete the following components of the standard-setting process?	Percent “About Right”	
	ELA	Math
Large group orientation	65%	36%
Experiencing the online assessment	80%	94%
Review of the Achievement Level Descriptors (ALDs)	93%	82%
Discussion of the skills demonstrated by students who are “just barely” described by each ALD	85%	73%
Review of the Ordered-Item Booklet (OIB)	85%	91%
Placement of your bookmarks in each round	73%	67%
Round 1 discussion	90%	85%

Note. Number of responses = 73. Evaluation options included "About Right," "Too Much," and "Too Little."

Participants appreciated the value of the multiple factors contributing to bookmark placement, with participants rating each factor as important or very important (Table 12).

Table 12. Evaluation: Importance of Materials

How important was each of the following factors in your placement of the bookmarks?	Percent "Somewhat Important" or "Very Important"	
	ELA	Math
Achievement Level Descriptors (ALDs)	100%	100%
Your perception of the difficulty of the items	98%	97%
Your experience with students	98%	94%
Discussions with other panelists	100%	97%
External benchmark data	95%	94%
Room agreement data (room medians and individual bookmark placements)	100%	97%
Impact data (percentage of students that would achieve at the level indicated by the OIB page)	100%	100%

Note. Number of responses = 73. Evaluation options included "Not Important," "Somewhat Important," and "Very Important."

Participant understanding of the workshop processes and tasks was consistently high (see Table 13).

Table 13. Evaluation: Understanding Processes and Tasks

At the end of the workshop, please rate your agreement with the following statements:	Percent "Agree" or "Strongly Agree"	
	ELA	Math
I understood the purpose of this standard-setting workshop.	100%	97%
The procedures used to recommend achievement standards were fair and unbiased.	95%	100%
The training provided me with the information I needed to recommend achievement standards.	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%	94%
The Achievement Level Descriptors (description of what students within each Achievement Level are expected to know and be able to do) provided a clear picture of expectations for student achievement at each level.	98%	85%

I was able to develop an understanding of the knowledge and skills demonstrated by students who are "just barely" described by the Achievement Level Descriptors.	100%	100%
I understood how to review each page in the Ordered-Item Booklet (OIB) to determine what students must know and be able to do to answer each item correctly.	100%	100%
I was able to interpret having an approximate 50% chance of answering an item correctly as indicating mastery.	100%	100%
I understood how to place my bookmarks.	100%	100%
I found the benchmark data and discussions helpful in my decisions about where to place my bookmarks.	98%	100%
I found the panelist agreement data (room medians and individual bookmark placements) and discussion helpful in my decision about where to place my bookmarks.	100%	100%
I found the impact data (percentage of students that would achieve at the level indicated by the OIB page) and discussions helpful in my decisions about where to place my bookmarks.	98%	91%
I felt comfortable expressing my opinions throughout the workshop.	98%	94%
Everyone was given the opportunity to express his or her opinions throughout the workshop.	98%	94%

Note. Number of responses = 73. Evaluation options included "Strongly Agree," "Agree," "Disagree," and "Strongly Disagree."

Participants agreed that the standards set during the workshop reflected the intended grade-level expectations (Table 14).

Table 14. Evaluation: Student Expectations

Please read the following statement carefully and indicate your response.	Percent "Agree" or "Strongly Agree"	
	ELA	Math
A student performing at Level 3 meets expectations for the grade level.	100%	94%
A student performing at Level 2 is below expectations for the grade level.	100%	91%
A student performing at Level 4 exceeds expectations for the grade level.	100%	94%

Note. Number of responses = 73. Evaluation options included "Strongly Agree," "Agree," "Disagree," and "Strongly Disagree."

Finally, panelists responded to two open-ended questions: "What suggestions do you have to improve the training or standard-setting process?" and "Do you have any additional comments? Please be specific." Forty-eight participants responded to the first question and 34 participants responded to the second question. While several participants indicated that the process was clear and did not need to be improved, some suggested that there should be less down-time between tasks and less repetition in training and that instructions be printed and include more visual examples that are easier to see from the back of the room.

Additional participant comments included:

> *Thank you AIR and DPI for the opportunity to be a part of this process. I always learn something new that will benefit me as a teacher, my students, and the students in the state of North Dakota.*
>
> *Do training for each day one day at a time instead of all at the beginning*
>
> *It was great to be a part of this process, it gives me a much better understanding of how our students' proficiency levels are judged and will likely help me better interpret student assessment data within my own classroom.*
>
> *Many of the slides and instructions were repeated far too many times. The large group introduction was so long and every need piece was restated in later instructional periods.*
>
> *It was definitely a learning experience and I feel much more confident about teaching these standards than I ever have. Thanks for all you did to make this process run quite smoothly.*

5. VALIDITY EVIDENCE

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

5.1 EVIDENCE OF ADHERENCE TO PROFESSIONAL STANDARDS AND BEST PRACTICES

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

> Standard 5.21: When proposed score interpretation involves one or more cut scores, the rationale and procedures used for establishing cut scores should be documented clearly.
>
> Standard 5.22: When cut scores defining pass-fail or proficiency levels are based on direct judgments about the adequacy of item or test performance, the judgmental process should be designed so that the participants providing the judgments can bring their knowledge and experience to bear in a reasonable way.

> Standard 5.23: When feasible and appropriate, cut scores defining categories and distinct substantive interpretations should be informed by sound empirical data concerning the relation of test achievement to the relevant criteria.

The sections of this report documenting the rationale and procedures used in the standard-setting workshop address Standard 5.21. The Bookmark standard-setting procedure is appropriate for tests of this type: with multiple item formats and scaled using item response theory (IRT). Section 3.1 provides the justification for and the additional benefits of selecting the Bookmark method to establish the cut scores; Sections 3.5.1 through 3.5.8 document the process followed to implement the Bookmark method.

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

- understanding the test and test items (from an educator and student perspective);
- describing the content measured by the test as described by the content standards;
- identifying the skills associated with each test item;
- describing the skills associated with "just barely" students for each Achievement Level;
- selecting which test items that students in each Achievement Level should be able to answer correctly;
- evaluating and applying feedback and reference data to their round 2 bookmarks; and
- considering the impact of the recommended cut scores on students.

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

The recruitment process resulted in panels which were representative of important regional and demographic groups and were knowledgeable about the subject area and students' developmental level. Section 3.3.4 summarizes details about the panel demographics and qualifications.

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

5.2 Evidence in Terms of Peer Review Critical Elements

The United States Department of Education (ED) provides guidance for the peer review of state assessment systems. This guidance is intended to support states in meeting statutory and regulatory requirements under Title I of the Elementary and Secondary Education Act of 1965 (ESEA)" (ED, 2015). The three 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 NDSA. They reviewed and revised the ALDs, drafted and applied target ALDs to delineate performance at each Achievement Level, considered feedback 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, the content standards and curriculum that they teach, and the developmental and cognitive capabilities of their students. AIR's facilitators were subject-matter experts in the subjects tested, alternate assessments, and facilitation of effective standard-setting workshops. The psychometricians performing the analyses and calculations throughout the meeting were subject-matter experts in the measurement and statistics principles required of the standard-setting process. Finally, Dr. Phillips is a nationally known subject-matter expert in assessment and measurement, including multiple methods of standard setting.

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

Evidence to support this critical element includes:

1) The rationale for and technical sufficiency of the Bookmark method selected to establish performance standards (Section 3.1).
2) Documentation that the method used for setting cut scores allowed panelists to apply their knowledge and experience in a reasonable manner and supported the establishment of reasonable and defensible cut scores (Section 3.5 and 4.1).
3) Panelists' self-reported readiness to undertake the task (Section 0) and confidence in the workshop process and outcomes (Section 3.5.8) supporting the validity of the process.
4) The standard-setting panels consisted of panelists with appropriate experience and expertise, including content experts with experience teaching the North Dakota's academic content standards and prioritized standards in the tested grades and subjects, and individuals with experience and expertise teaching special education and general education students in North Dakota (Section 3.3.4).

6. References

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

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

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

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

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

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

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

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

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

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

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

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

U. S. Department of Education, (2015). *Non-Regulatory Guidance for States for Meeting Requirements of the Elementary and Secondary Education Act of 1965, as amended.* Washington, D.C. Accessed from https://www2.ed.gov/policy/elsec/guid/assessguid15.pdf.