Wisconsin Forward Exam

Science Standard Setting

2019

Final Technical Report

Prepared for the
Wisconsin Department of Public Instruction

Data Recognition Corporation
Maple Grove, MN 55311

Table of Contents

Appendices

I

Summary of Cut Score Recommendations

Summary of Cut Score Recommendations

On May 29–30, 2019, a committee of 27 Wisconsin educators participated in a standard setting for the Wisconsin Forward Exam of grades 4 and 8 science. At the workshop, participants recommended cut scores to divide students into four performance levels: *Below Basic, Basic, Proficient,* and *Advanced.*

The committee comprised 13 educators for grade 4 and 14 educators for grade 8. Participants engaged in the Bookmark Standard Setting Procedure; this well-documented procedure has been used across the country to establish performance standards for large-scale assessments and was used most recently in 2016 on the Wisconsin Forward Exam tests of English language arts and mathematics. The standard setting was sponsored by the Wisconsin Department of Public Instruction (DPI) and was facilitated by Data Recognition Corporation (DRC).

Table 1 shows the recommended cut scores and associated impact data from the workshop. Impact data are the percent of students that would be classified in each performance level on the spring 2019 administration of the assessments if the cut scores were implemented.

Table 1. Recommended Cut Scores and Associated Impact Data for Grades 4 and 8 Science

	Recommended Cut Scores			Percent of Students in Each Performance Level Based on Recommended Cut Scores			
Grade	***Basic***	***Proficient***	***Advanced***	***Below Basic***	***Basic***	***Proficient***	***Advanced***
4	447	496	543	15.0%	32.2%	33.3%	19.5%
8	653	695	737	17.7%	28.3%	31.5%	22.5%

Participants engaged in three rounds of discussions and judgments to make their cut score recommendations. Specifically, the committee performed the following tasks:

1. Participants discussed the state content standards for science and the draft performance level descriptors (PLDs) for their test. The PLDs described the content-based expectations for students in each performance level. Participants refined the PLDs based on their study of the content standards.
2. Participants each examined an ordered item booklet (OIB) which presented test items in order of difficulty. Difficulty was calculated from Wisconsin students' performance.
3. For each item, participants considered whether a student just entering each performance level (e.g., a just *Proficient* student) would have command of the knowledge and skills measured by the item, defined as having at least a 50% chance of answering the item correctly. Participants indicated the set of items in the OIB that measured the content expected of students entering each performance level; they represented these judgments with bookmarks.
4. Participants discussed their bookmarks in three rounds of discussions and decisions. After each round, participants worked individually to revise their bookmark placements.

5. After the second and third rounds, participants examined the impact data for both grades. After the second round, participants reviewed the impact data associated with their recommended cut scores, as well as the impact data for the 2015 NAEP science.
6. After the third round, participants reviewed the PLDs. Participants refined them to reflect the content-based expectations for students in each performance level.
7. Participants' cut score recommendations were recorded in terms of scale score. Each group's recommendation was the median of participants' recommendations.

After the Round 3 of the Bookmark Procedure, participants reviewed their recommendations and associated impact data, as shown in Table 2. Educators expressed satisfaction in their content-based judgments they made during the process. However, participants also voiced an expectation that the percentages of students classified in each performance level would be more consistent across grades 4 and 8.

Table 2. Cut Scores and Associated Impact Data for Wisconsin Science from the Bookmark Procedure

	Cut Scores Before Adjustments			**Impact Data Based on Cut Scores Before Adjustments**			
Grade	***Basic***	***Proficient***	***Advanced***	***Below Basic***	***Basic***	***Proficient***	***Advanced***
4	447	496	543	15.0%	32.2%	33.3%	19.5%
8	653	682	723	17.7%	18.2%	32.2%	31.8%

To promote consistency in the performance standards across grades and testing programs, the Round 3 cut scores for grade 8 *Proficient* and *Advanced* were adjusted using the conditional standard error of measurement (CSEM). The CSEM quantifies the amount of statistical error associated with any point on the test scale. These adjustments promoted consistency among the performance standards across grades.

The CSEM-adjusted cut scores and associated impact data are shown in Table 1. Participants examined these cut scores and considered their reasonableness. Participants indicated the CSEM-adjusted cut scores were consistent with their content-based expectations from the Bookmark Procedure as well as their expectations for the impact data across grades. The committee made the CSEM-adjusted cut scores (shown in Table 1) their final recommendations for the Wisconsin science assessments of grades 4 and 8 science.

II

Standard Setting Methodology

Standard Setting Methodology

On May 29–30, 2019, the Wisconsin Department of Public Instruction (DPI) partnered with Data Recognition Corporation (DRC) to conduct a standard setting for the Wisconsin Forward Exam assessments of grades 4 and 8 science. The purpose of the standard setting was to develop performance standards for the two assessments, including the development of *cut scores* which divide students into four performance levels: *Below Basic, Basic, Proficient*, and *Advanced.*

A total of 27 Wisconsin educators and stakeholders worked individually and in committees to recommend performance standards for the tests. The performance standards were approved by the Superintendent of Public Instruction on June 5, 2019.

About this Section

This section details the planning of the standard setting, the implementation of the workshop, the analysis of Wisconsin educators' recommendations, and the approval of the cut scores. A summary of this work can be found in Chapter I of this report. Further details about the workshop, such as workshop agenda and detailed presentations of participants' recommendations, can be found in appendices of this report.

Background of Testing in Wisconsin

Wisconsin's statewide science tests have recently gone through several changes. The state content standards, published in November 2017, were adapted from the Next Generation Science Standards (Wisconsin Department of Public Instruction, 2017). These three-dimensional standards describe the knowledge and skills that students should be taught in terms of disciplinary core ideas (content), science and engineering practices, and crosscutting concepts. The tests of grades 4 and 8 science measured these new standards for the first time in school year 2018–19.

At the same time, new item types were introduced onto the assessments. These item types, including technology-enhanced items such as drag-and-drop and multi-select, are designed to help students demonstrate their knowledge in authentic ways. Moreover, many of the science items were designed around shared stimuli or *scenarios*: on the test, students read scenarios and then answered several questions about the scenario. These new item types also debuted in school year 2018–19 for science.

The statewide tests of science most recently underwent standard setting just after school year 2015–16 (Data Recognition Corporation, 2016). At that standard setting, performance standards were established for all four content areas of the newly-implemented Wisconsin Forward Exam—English language arts, mathematics, science, and social studies—as the older Wisconsin Knowledge and Concepts Examination (WKCE) program was retired. However, the cut scores established for science assessments in 2016 were based on the older science content standards. Because the science assessments began assessing the new science content standards in school year 2018–19, a new standard setting was required.

At the 2019 standard setting for science, DPI sought to establish cut scores (also known as *passing scores*) for the assessments which reflect the state's expectations for student performance throughout the state. During this standard setting, DPI developed cut scores on the Forward Exam that reflected these content-based expectations on the tests, as informed by test data from well-respected measures of student achievement.

Selecting the Standard Setting Methodology

The Bookmark Standard Setting Procedure (BSSP; Lewis, Mitzel, & Green, 1996; Lewis, Mitzel, Mercado, & Schulz, 2012) was implemented to recommend cut scores for the Wisconsin science tests. This method has been used on assessments in Wisconsin and across the nation (Karantonis & Sireci, 2006), including for the previous version of the Wisconsin science assessments (Data Recognition Corporation, 2016) as well as for the current Wisconsin Forward Exams of English language arts and mathematics.

The BSSP has been well documented in the standard setting literature. Developed in 1996, the BSSP has been implemented in over half of the states in the U.S. and abroad by DRC and by other major testing firms, making it the most widely used standard setting procedure in K–12 education (Karantonis & Sireci, 2006).

As an item-mapping process, the Bookmark Procedure is particularly useful for large-scale assessments that include both traditional multiple-choice and technology-enhanced items, like the Wisconsin Forward Exam. Specifically, the science assessments include multiple-choice items modeled using the three-parameter logistic (3PL) model; and single-point, autoscored technology-enhanced items modeled using the two-parameter partial-credit (2PPC) model. Additional information about the modeling, scaling, and equating of the test forms can be found in the program technical report.

Bookmark Procedure allows these different item types to be ordered together in *ordered item booklets*. In addition, because of its history of use in Wisconsin and across the nation, DPI selected the Bookmark Procedure for the 2019 science standard setting.

The performance level setting also incorporated elements of the *evidence-based standard setting* framework (McClarty, Way, Porter, Beimers, & Miles, 2013). In particular, focused attention was paid before the standard setting to the types of performance standards that DPI would consider reasonable for assessments. Selected policy information was provided to standard setting participants in the form of benchmarks (see Phillips, 2012), allowing educators to consider this policy information in an actionable way as they made their content-based judgments during the Bookmark Procedure.

Performance Level Descriptors (PLDs)

Performance level descriptors (PLDs) summarize the knowledge, skills, and abilities expected of students in each performance level. Specifically, there are four types of PLDs (Egan, Schneider, and Ferrara, 2012), each with a different focus.

1) **Policy PLDs** set out the Department's vision for each performance level. Policy PLDs are not specific to any given test; rather, they represent a policy vision for each performance level. The

policy PLDs for the science assessments are presented in Table 3. In the table, emphasis is added to reveal the differences between the descriptors.

2) **Range PLDs** specify the knowledge, skills, and abilities expected of students in each performance level on a given test. For example, a range PLD may list the expectations of students who are in *Basic* in grade 4 science. These expectations include those for students who are just in the *Basic* level, those who are well within the *Basic* level, and those who are nearly (but not quite) at the *Proficient* level. Range PLDs are often shared with teachers and schools to help them understand the level of construct mastery expected of students in each performance level on each test.

3) **Threshold PLDs** specify the knowledge, skills, and abilities expected of students who are at the point-of-entry in each performance level on a given test. For example, a threshold PLD may list the expectations of students who have just enough skill to be considered *Proficient* in grade 8 science. Whereas the range PLD specifies the expectations for all *Proficient* on the test, the threshold PLD seeks to specify the expectations for a student who has just entered the *Proficient* level. These descriptors are typically used by participants at performance level setting workshops to help inform decisions they make about cut points.

4) **Reporting PLDs**, like range PLDs, specify the knowledge, skills, and abilities expected of students in each performance level on a given test; however, they are designed to communicate this information to stakeholders and educators in the field through score reporting. Reporting PLDs typically comprise a version of the policy or range PLDs, and the language in the reporting PLDs is adjusted to be accessible to a wide audience that may not have in-depth content knowledge.

Table 3. Policy performance level descriptors (PLDs) for Wisconsin Forward Exam

Level	Policy Performance Level Descriptor
Below Basic	Student demonstrates **minimal understanding** of and ability to apply the knowledge and skills for their grade level that are associated with college content-readiness.
Basic	Student demonstrates **partial understanding** of and ability to apply the knowledge and skills for their grade level that are associated with college content-readiness.
Proficient	Student demonstrates **adequate understanding** of and ability to apply the knowledge and skills for their grade level that are associated with college content-readiness.
Advanced	Student demonstrates **thorough understanding** of and ability to apply the knowledge and skills for their grade level that are associated with college content-readiness.

Well in advance of the standard setting, DPI verified that the same policy PLDs used for the other Wisconsin Forward Exams would also be used for the newly-revised science assessments. That is, the same policy PLDs are used for all four content areas in the Wisconsin Forward Exam program.

In the months leading up to the standard setting, DPI worked with its internal content experts and with DRC to develop draft range PLDs for the science assessments. The goal of this team was to develop a set of range PLDs which would (a) help standard setting participants understand the types of knowledge, skills, and abilities expected of students in each performance level; and (b) eventually help Wisconsin educators understand the types of content-based expectations for students in each performance level.

The team quickly realized that developing a set of PLDs that described the expectations for students in each performance level at each and every combination of disciplinary core ideas, science and engineering practices, and crosscutting concepts would create an unwieldy document that would be difficult to use in the real world. Instead, the team considered the expectations for students in each of the four performance levels on the three meta-standards shown here (with emphasis added to reveal the differences between the meta-standards).

- Students use science and engineering practices, crosscutting concepts, and an understanding of **life science** disciplinary core ideas, to make sense of phenomena and solve problems.
- Students use science and engineering practices, crosscutting concepts, and an understanding of **physical science** disciplinary core ideas, to make sense of phenomena and solve problems.
- Students use science and engineering practices, crosscutting concepts, and an understanding of **earth and space science** disciplinary core ideas to make sense of phenomena and solve problems.

Each meta-standard is based on a single *content area* from the disciplinary core standards, and the various science and engineering practices and crosscutting concepts are considered alongside each. For each meta-standard, the team considered the range of knowledge, skills, and abilities that students in each performance level would be expected to have. The draft range PLDs were edited by DPI and DRC, and DPI approved them for use at the standard setting. The draft range PLDs are presented in Appendix G of this report.

At the standard setting, participants studied the draft range PLDs to understand the expectations for students in each performance level. In small groups, the participants then used these expectations to develop informal threshold PLDs for the tests. These threshold PLDs described the types of knowledge and skills expected of students at the point-of-entry of the *Basic, Proficient*, and *Advanced* performance levels. These descriptors were developed by standard setting participants for their use during discussions at the standard setting and are not included in this report. The development of these threshold PLDs is summarized later in this section.

As the standard setting concluded, participants reflected on their recommendations and on what they had learned during the workshop. Participants then recommended several refinements to the draft range PLDs to make them even clearer and more useful to educators in the field. These refined range PLDs are part of the committees' recommendations for the performance standards, and they are presented in Appendix G of this report.

Reporting PLDs were not part of the scope of this standard setting. However, DPI may choose to use the policy PLDs or elements of the range PLDs as part of its overall reporting strategy for the Wisconsin Forward Exam program.

Benchmarks

Benchmarks comprised an important component of the standard setting process. Benchmarks refer to any external content- or policy-based information that is presented to participants to help them make their cut score recommendations. The use of benchmarks at performance level setting is well established (Phillips, 2012; McClarty et al., 2013), especially in the Bookmark Procedure (Lewis, Mitzel, Mercado, & Schulz, 2012). Many states have used benchmarks to provide actionable, policy-based information to performance level setting participants. Participants can then bring their content-based expertise to bear, joining it with the benchmarks. Thoughtful use of benchmarks can bring policy- and content-based information together in a meaningful way.

An exact alignment between the performance of Wisconsin students on the science assessments of the Wisconsin Forward Exam and of NAEP is not expected, for several reasons. For example, the most recent data available for NAEP came from 2015, before the latest major push by the state to focus on the three-dimensional science standards. In addition, the NAEP science assessment is qualitatively different than the Wisconsin Forward Exam (e.g., NAEP includes a hands-on experiential section), and the NAEP achievement standards were developed by a different group for a different purpose. However, DPI acknowledged that some stakeholders would compare student performance on both tests, and that the state had an interest in making sure that tests given in Wisconsin sent consistent signals of student performance to schools and students.

At the same time, DPI noted that it wanted to make sure (a) standard setting participants would make content-based recommendations that linked the cut scores to the Wisconsin state content standards; and (b) standard setting participants were not unduly influenced by the benchmarks. Accordingly, DPI chose to present NAEP-based benchmarks after Round 2 of the Bookmark Procedure. The process used to present the benchmarks is shown later in this report.

Ordered Item Booklets (OIBs)

The ordered item booklet (OIB) is a key component of the BSSP. An OIB contains the items from a test, ordered by difficulty. A separate OIB was prepared for each grade of science at the standard setting.

Within each OIB, items are ordered by their difficulty on the test scale. Easier items appear earlier in the OIB, and harder items appear later. The ordering of the items is based on each item’s scale location, which is based on observed student performance.

Selecting Items for the OIBs

To create the OIBs, DRC selected items for three purposes, shown here.

1) ***Approximately 60 items***. For each test, the operational items were pooled together. However, using all these items would yield an OIB with over 100 pages: a long OIB is often unwieldy at a standard setting. To create an OIB that would focus the time and attention of standard setting participants, approximately 60 items were selected for each OIB.

2) ***Mirror the test blueprint***. The selected items for the OIB proportionally mirrored the reporting categories used for the operational test. In this way, the collection of items presented to participants at the standard setting was representative of future operational forms of the tests.
3) ***Variety of difficulty levels***. A variety of easy, medium, and difficult items were selected. The difficulty of each item was defined as its RP-adjusted location on the test scale.

Response Probability for the OIBs

Items are ordered in the OIB using a response probability (RP) criterion. An RP criterion specifies the probability with which a student with a given ability would be able to correctly answer an item of the same difficulty. For example, if the RP criterion is 0.50 (RP50), students with ability just at the cut score would have a 50% chance of correctly answering items with difficulty at the cut score.

In the BSSP, items are most often ordered using an RP criterion of 0.67 with an adjustment for guessing (RP67GA; Lewis, Green, Mitzel, Baum, & Patz, 1998). However, other RP criteria are sometimes used, including RP50 (Cizek & Bunch, 2007, p. 162; Mitzel, et al., 2001). At the 2016 Wisconsin standard setting, the OIBs were created with a response probability of 0.67 (RP67): the RP-adjusted scale location for each item in the OIB was associated with the scale score needed to have a 67% chance of answering the item correctly. In preparing for that workshop, DPI acknowledged that it had a history of using RP67 at its standard settings.

While creating the OIBs for the 2019 standard setting, however, DRC discovered there were relatively few “easy” items when RP67 was used. DRC created hypothetical OIBs using RP50 and RP67 to order the items. Table 4 presents the bookmark needed to classify 20% of students below the cut score, 30% below, and so on for the various OIBs.

Table 4. Bookmarks needed to classify 20–60% of students below the cut score in hypothetical OIBs using RP50 and RP67 and using selected items (60–64 total items per OIB)

	Grade 4		Grade 8	
Bookmark needed to classify	**RP50**	**RP67**	**RP50**	**RP67**
20% below the cut score	13	4	12	2
30% below the cut score	19	8	18	5
40% below the cut score	23	13	24	11
50% below the cut score	33	18	35	16
60% below the cut score	38	27	41	23

As shown in the table, if RP67 were to be used at the standard setting, then a bookmark before page 10 of the OIB would classify more than 30% of students below the cut score. Using RP67, it would be difficult for participants to use the OIB to recommend cut scores where fewer than 20–30% of students would be classified as *Below Basic*. Given that this is approximately the percentage of Wisconsin students classified as *Below Basic* on NAEP, the potential use of RP67 to create the OIBs caused concern.

In May 2019, DPI consulted with a member of its technical advisory committee (TAC) to consider which RP criterion to use at the upcoming standard setting. After consultation, DPI decided to use RP50 for the standard setting. The decision to use RP50 was made for several reasons. First, DPI noted that it would be difficult for standard setting participants to use OIBs created with RP67 to recommend cut scores that would yield impact data similar to Wisconsin's performance on NAEP, especially for the *Basic* cut score. Although the state does not necessarily expect the results of the test to be similar to NAEP, it is not unreasonable to expect that standard setting participants may want to recommend a performance standard that classifies 20–30% of students as *Below Basic*: such a standard would be difficult to recommend using RP67. Second, DPI acknowledged that it has a history of using RP67 at its standard settings, but that this precedent is not as important as the first point: it is more important to have a useful OIB that can be used by participants to recommend reasonable performance standards than it is important to comport with historically-implemented RP criteria. Third, DPI noted that RP50 has been used frequently by other states and agencies, and the training protocol at the Bookmark Procedure can be adjusted to help participants understand how to interpret an OIB created using RP50. Such an adjustment, focusing on helping participants develop an intuitive understanding of RP50 in the OIB.

Accordingly, RP50 was used to create the OIBs for the standard setting. Additional information about the selection of RP50 for the workshop can be found in Appendix F of this report. The actual OIBs used at the standard setting contain secure test information and are not included in this report. However, the OIB used during the participant training session contains released items and passages; this OIB is included in Appendix B of this report.

Data and Other Workshop Materials

All of the materials used at the standard setting workshop were based on test items and results from the Spring 2019 administration of the Wisconsin Forward Exam for science. Participants used the following materials to help make their cut score recommendations during the workshop.

- **Wisconsin Content Standards for Science**

The state content standards formed the basis for all decisions at the standard setting. These content standards, as adopted by the State Board of Education, detail the knowledge, skills, and abilities that students should be taught in each grade and subject. Copies of the content standards were distributed to workshop participants.

- **Performance Level Descriptors (PLDs)**
- **Ordered Item Books (OIBs)**
- **Item Maps**

The item map summarizes information about the items in an OIB. For each item, the item map indicates: the order of difficulty, the correct answer, and the title of the item's passage.

The operational item maps incorporate secure test information and are not included in this report. However, Appendix B shows the item map that was used during the participant training sessions and does not include secure test information.

- **Benchmarks**

During the BSSP, benchmarks took the form of benchmark-linked bookmarks, termed at the standard setting simply as *benchmarks.* More information about the benchmarks used at the standard setting can be found earlier in this report.

To calculate the OIB benchmarks, the Wisconsin Forward Exam cut scores that most closely yielded the benchmarked impact data were first identified. The OIB positions associated with these benchmarked cut scores were then determined. The benchmarked impact data values and associated OIB benchmarks are presented in Table 5.

As part of the training presentations, participants were instructed that they would see benchmarked bookmarks after Round 2 of the BSSP, and that they should consider the OIB benchmarks. Participants were told about the existence of the benchmarks in the opening session and training presentations, along with a brief description of how they would be used. In the secondary training session, participants were given further instructions on how to interpret the benchmarks.

Participants were asked to consider the knowledge, skills, and abilities measured by the items before each OIB benchmark and to compare them with their bookmarks that they placed in Round 2. If there was good correspondence between the content measured by the items before each bookmark and the content-based expectations for each threshold student, then participants were encouraged to retain their bookmarks. However, if there was not good correspondence, participants were encouraged to move their bookmarks until there was better correspondence, and to use the benchmarked OIB pages as a guide.

Table 5. Benchmarked impact data from 2015 NAEP and associated OIB benchmarks

		OIB Benchmarks			Benchmarked Impact Data			
Content	**Grade**	***Basic***	***Prof.***	***Adv.***	***B.B.***	***Basic***	***Prof.***	***Adv.***
Science	4	15	39	61	21.0%	38.0%	40.0%	1.0%
	8	15	42	63	25.0%	35.0%	38.0%	2.0%

Standard Setting Staff and Participants

Staff members from DPI and DRC collaborated to conduct the standard setting workshop. These staff members worked in facilitative roles and did not contribute to the cut score recommendations during the workshop.

DPI Staff

DPI staff members attended the workshop to monitor the process, answer assessment and curriculum questions, and address DPI policy questions. DPI also monitored participants' cut score recommendations throughout the workshop.

DPI was represented at the workshop by Viji Somasundaram, Director of the Office of Student Assessment (OSA); Philip Cranley, OSA Education Consultant; Duane Dorn, OSA Education Consultant; Alison O'Hara, OSA Education Consultant; and Jennifer Teasdale, OSA Education Program Specialist. These DPI staff members were assisted by additional DPI members who worked in concert to monitor the standard setting.

DRC Staff

The DRC Standard Setting Team was composed of Ricardo Mercado, Research Director; Michelle Boyer, Research Scientist; Sara Kendallen, Sr. Research Analyst; and Julie Pointer Korts, Research Analyst. Prior to the standard setting, this team prepared the materials for the workshop. During the workshop, they were responsible for facilitating the workshop, training participants, entering participant results into a database, performing data analyses, and tracking secure materials. Following the workshop, the team prepared this report.

Content experts from DRC Test Development worked with each group at the workshop to provide content-based support. These content experts were Dave Durette, Sr. Director, Test Development; and Jeannie Hemsworth, Assessment Editor.

Participants

All participants for the workshop committee were recruited, selected, and invited to the workshop by DPI. The recruitment process strived to empanel a sample of participants for the standard setting with diverse demographics (e.g., ethnicity, gender) and diverse points-of-view (e.g., geographic location).

The committee comprised a purposeful mix of educators with a variety of backgrounds. Special care was taken to promote geographic diversity among participants, with representation from across the state. Participants were asked to self-report their demographic characteristics (e.g., ethnicity, number of years in the profession) as part of the participant survey.

Configuration of the Committee

The workshop committee was composed of a total of 27 educators. Two groups were convened for the standard setting, as listed here.

- Grade 4 Science (3 tables, 13 participants)
- Grade 8 Science (3 tables, 14 participants)

Participants of the two groups were divided into three tables of four or five participants per table. One participant at each table served as the table leader. Table leaders moderated discussions at their tables and helped the workshop staff distribute and collect the secure workshop materials. The table leaders were not members of the workshop staff, and they contributed to their committees' recommendations.

Standard Setting

The standard setting workshop took place over a two-day period. All participants began the workshop on the first day and completed the workshop on the second day. The workshop agenda is included in Appendix A.

Opening Session

The workshop began with an opening session by the DPI. During this session, Ms. Somasundaram welcomed the participants to the workshop and described the purpose of the workshop.

Participant Training

Following the opening session, Mr. Mercado from DRC introduced the standard setting methodology. Participants were introduced to the materials that would be used during the workshop. The training presentations are included in Appendix B of this report.

Following the training session, participants were divided into their pre-assigned groups and tables. Each grade convened in a separate breakout room.

Discussion of the PLDs and the Threshold Students

The group leaders instructed participants to read the content standards and policy PLDs, and to consider the knowledge, skills, and abilities (KSAs) that students were expected to demonstrate in each performance level. Specifically, participants were asked to use the policy PLDs, range PLDs, and content standards to develop draft versions of threshold PLDs. These documents are summarized here.

- *Policy PLDs* – Participants examined the policy PLDs, as shown in Table 3, to gain an understanding of the high-level expectations for students in each performance level.
- *Range PLDs* – Participants examined range PLDs which summarize the content-based expectations for students across the range of performance within each performance level. For example, the range PLD for *Proficient* summarized the expectations for students who were at the low end, in the middle, and at the high end of the *Proficient* level.
- *Threshold PLDs* – Participants developed threshold PLDs to summarize the expectations for students who had just enough knowledge, skills, and abilities to be considered in each performance level. Participants were encouraged to imagine a hypothetical *threshold student* to represent each threshold.

Participants engaged in structured discussions about the KSAs they expected to be demonstrated by each of the three threshold students. The three threshold students were just barely *Basic,* just barely *Proficient,* and just barely *Advanced*. To engage in these discussions, participants referred to the PLDs, the content standards, and their knowledge of students.

As a group, participants discussed the PLD for each performance level and the differences between them. During this discussion, participants considered the overall level of rigor implied by each policy PLD. To focus participants on the lines of demarcation between the performance levels, participants were asked to discuss the KSAs that separated students in one performance level from those in another. For example, participants were asked to discuss the KSAs that separated the highest performing *Basic* from the lowest performing *Proficient*. All participants were instructed to refer to the content standards and range PLDs during this discussion.

Participants recorded their expectations for students in each performance level (and at the thresholds of each performance level) on large pieces of paper which were hung conspicuously in the meeting room. Participants were instructed that they could review and revise the PLDs throughout the workshop, but that the expectations for students in each performance level must be based on the content standards and the policy PLDs. DRC informed participants that they would have an opportunity to refine the PLDs before the end of the workshop.

By the end of this discussion, participants had thoroughly considered the PLDs, content standards, and threshold students, and they reached an understanding of the types of skills that the threshold student, for each performance level, should have. Participants' incorporated edits are presented in the policy and range PLDs presented in Appendix G of this report.

Study of the OIBs and Item Maps

Participants at each table examined the items in the OIB in terms of what each item measured and why it was more difficult than the items preceding it. Participants were instructed to take notes on the item maps about the knowledge, skills, and abilities required to answer the items correctly.

Secondary Training on Placing Bookmarks

Mr. Mercado provided the participants with training for placing bookmarks. Participants were told how cut score recommendations could be represented by bookmarks. Participants were instructed that all items preceding the bookmark contain the knowledge, skills, and abilities that a student who is just barely in the *Basic* level, for example, is expected to know. The training presentation is included in Appendix B. The training materials used during this session are also included in Appendix B.

Participants were also informed that they should have a content-based rationale for each of their bookmarks, and that these rationales should refer to the alignment between the knowledge, skills, and abilities in the PLDs and those in the items before the bookmark. Participants were instructed that they would share these rationales verbally with their tables after Round 1 was complete.

Following training, participants were tested on their understanding of bookmark placement with a short quiz, termed a *mid-process evaluation.* Participants completed the mid-process evaluation. Afterwards, participants were provided the correct answers for the mid-process evaluation, as well as explanations of those answers. The mid-process evaluation and results are presented under the heading "Committee Training," as well as in Appendix B.

Round 1 Bookmarks

Participants then made their Round 1 bookmark judgments. Participants were informed that bookmark placement is an individual activity. They referred to their OIBs, item maps, PLDs, and content standards.

Participants recorded their bookmark placements on a form, along with a few words about their content-based rationale for doing so. Participants were instructed that they should have a content-based rationale for each bookmark placement that linked the content measured by the items before their bookmark and the content-based expectations for the threshold student. These content-based rationales were solely for participants' reference during their table's discussion before Round 2. Participants then completed Round 1 by recording their bookmark placements on a secure web-based survey platform.

Presentation of Round 1 Recommendations

Following Round 1 bookmark placements, DRC calculated the bookmark recommendations for each group. Participants were presented with a summary of their Round 1 recommendations. Specifically, participants were shown the range of bookmark placements in the room, median bookmark placements for each table, as well as the overall median bookmark for the group. Table 7, under heading "Results" presents participant's Round 1 recommendations and associated impact data. Appendices C and D present the Detailed Reports of Participants' Judgments and the Graphical Representations of Participants' Judgments, respectively.

Round 2 Bookmarks

For each performance level, participants discussed the rationales behind their Round 1 bookmark placements. Participants were instructed to engage in a content-based discussion by focusing on the items in the OIB between the lowest and highest bookmarks for Round 1. Participants were also informed that they could discuss items outside the range of their bookmarks. These content-based discussions took place at each table. Participants referred to their OIBs, item maps, PLDs, and the content standards throughout the discussions.

Following this discussion, participants placed their Round 2 bookmarks. Participants were reminded that bookmark placement is an individual activity. Participants were also reminded that they would be free to retain their bookmarks from Round 1 or to change one or more of them; however, in either case, participants would need to have content-based rationales for their decisions.

Presentation of Round 2 Recommendations

Following Round 2 bookmark placements, DRC calculated the bookmark recommendations for each group. Participants were presented with a summary of their Round 2 recommendations and associated impact data. Table 8, under heading "Results" presents participant's Round 2 recommendations and associated impact data. Appendices C and D present the Detailed Reports of Participants' Judgments and the Graphical Representations of Participants' Judgments, respectively.

Presentation of Benchmarks

In each breakout room, the benchmarks were shown to participants in terms of OIB position. The underlying benchmarked impact data used to calculate the OIB benchmarks were shown to participants to provide contextual information for consideration. Participants were given instructions on how to use the OIB benchmarks as points-of-reference as they considered their Round 3 judgments, and they were asked to consider how similar or different their Round 2 bookmarks were from the OIB benchmarks.

Round 3 Bookmarks

For each performance level, participants discussed the rationales behind their Round 2 bookmark placements. Participants were instructed to engage in a content-based discussion by focusing on the items in the OIB between the lowest and highest bookmarks for Round 2. Participants were also informed that they could discuss items outside the range of their bookmarks. These content-based discussions took place as a group. Participants referred to their OIBs, item maps, benchmarks, PLDs, and the content standards throughout the discussions.

Following this discussion, participants placed their Round 3 bookmarks. Participants were reminded that bookmark placement is an individual activity. Participants were also reminded that they would be free to retain their bookmarks from Round 2 or to change one or more of them; however, in either case, participants would need to have content-based rationales for their decisions.

Presentation of Round 3 Recommendations

In their groups, participants were shown their Round 3 recommendations and associated impact data. Then participants were convened as a single committee, and DRC presented the impact data associated with the cut score recommendations from both groups, as taken from their Round 3 median bookmark placements.

Participants were encouraged to consider these questions:

1. How would you describe the pattern in the impact data across grades?
2. Are there any cut scores from your group that you have questions about? If so, what kind of latitude do we have around our final bookmarks that are still consistent with the content?
3. Are there any cut scores from the other group that you have questions about?

Table 9, under heading “Results” presents participant’s Round 3 recommendations and associated impact data. Appendices C and D present the Detailed Reports of Participants’ Judgments and the Graphical Representations of Participants’ Judgments, respectively.

Across-Grade Discussion

Participants then discussed the recommendations as a committee. Participants shared their views of their recommended cut scores, including their reactions to the three questions posted above.

Participants were informed that they could recommend adjustments to the cut scores, if needed, to promote better articulation across grades. However, participants were cautioned against suggesting adjustments which were inconsistent with the content: any adjusted bookmarks should still link the PLDs, tested content, and content standards.

Wisconsin 2018 impact data was presented to participants at this time to aide their discussion. The Spring 2018 Wisconsin Science impact data is presented here in Table 6.

Table 6. Wisconsin Science Spring 2018 impact data

Content	Grade	Impact Data			
		Below Basic	***Basic***	***Proficient***	***Advanced***
Science	4	15.2%	34.1%	34.4%	16.3%
	8	17.2%	34.0%	34.2%	14.7%

The committee worked by consensus. As described later in this section, the committee had an in-depth conversation about the articulation between the two grades, and they recommended changes to two cut scores to promote better articulation across the grades. Participants viewed their final recommended cut scores as they left the across-grade discussion.

PLD Refinement

Back in their breakout rooms, participants then worked to refine the PLDs. A facilitator in each room asked participants to consider their learnings from the standard setting about the content-based expectations for students in each performance level.

DRC asked participants to suggest refinements to make the PLDs more clear, concise, and ultimately more useful to educators in the field. However, participants were cautioned against suggesting refinements which adjusted the overall level of rigor associated with each performance level.

Participants worked in each group to suggest refinements to the PLDs, and the participants recorded these suggestions.

Workshop Evaluation

To conclude the workshop, participants were asked to complete a written evaluation. Selected results are presented later in this chapter. The complete results of the evaluations are included in Appendix H.

Workshop Security

Throughout the workshop, security was of paramount importance. Secure test materials used during the workshop were numbered and assembled into packets. Each participant signed out a specific packet and signed his or her name on each of the materials in the packet. At all times, DRC staff monitored the meeting rooms to prevent the removal of secure materials. At the end of each day, each participant's materials were collected and inventoried against a master list. The secure materials were stored overnight in a secure room. At the conclusion of the workshop, the secure materials were collected and inventoried against the sign-out lists for a final time.

In addition, participants were required to sign non-disclosure agreements to participate in the workshop. These agreements were signed by participants and were collected by the DRC staff at the beginning of the workshop.

Results

The standard setting was conducted according to the plans created by DPI and DRC prior to the workshop. The results of the workshop are presented in this section.

Participants' Round 1 Recommendations

Table 7 shows participants' recommendations from Round 1 of the BSSP. (All of the impact data shown in the table and in this section are based on Wisconsin students' performance in Spring 2019.) In Round 1, participants made their recommendations independently and without discussion.

Table 7. Cut score recommendations and associated impact data from Round 1 of the standard setting

		Round 1 Cut Scores			Associated Impact Data			
Content	**Grade**	***Basic***	***Prof.***	***Adv.***	***B.B.***	***Basic***	***Prof.***	***Adv.***
Science	4	447	496	543	15.0%	32.2%	33.3%	19.5%
	8	654	690	719	18.2%	23.9%	23.2%	34.8%

Participants' Round 2 Recommendations

Table 8 shows participants' recommendations from Round 2 of the BSSP. Before placing their Round 2 bookmarks, participants discussed their Round 1 bookmark placements with their colleagues at their table. After this discussion, participants placed their bookmarks independently.

Table 8. Cut score recommendations and associated impact data from Round 2 of the standard setting

		Round 2 Cut Scores			Associated Impact Data			
Content	**Grade**	***Basic***	***Prof.***	***Adv.***	***B.B.***	***Basic***	***Prof.***	***Adv.***
Science	4	447	496	538	15.0%	32.2%	30.4%	22.4%
	8	653	673	718	17.7%	12.0%	34.8%	35.5%

Participants' Round 3 Recommendations

Table 9 shows participants' recommendations from Round 3 of the BSSP. Before placing their Round 3 bookmarks, participants were shown benchmarks based on Wisconsin students' performance on the 2015 NAEP science assessments, and participants discussed their Round 2 bookmarks with their colleagues in their group. After this discussion, participants placed their bookmarks individually. Participants' individual recommendations from all rounds may be found in Appendix C of this report. Graphical representations of participant's individual recommendations may be found in Appendix D of this report.

Table 9. Cut score recommendations and associated impact data from Round 3 of the standard setting

Content	Grade	Round 3 Cut Scores			Associated Impact Data			
		Basic	***Prof.***	***Adv.***	***B.B.***	***Basic***	***Prof.***	***Adv.***
Science	4	447	496	543	15.0%	32.2%	33.3%	19.5%
	8	653	682	723	17.7%	18.2%	32.2%	31.8%

Recommendations from the Articulation Discussion

Throughout the standard setting process, participants were informed they would have an opportunity at the end of the workshop to consider the across-grade articulation of the performance standards. Participants were told that performance standards were well-articulated when the impact data associated with a set of cut scores formed a reasonable, explainable pattern across grades.

The committee inspected the impact data associated with their recommendations. Participants noted that the content standards were different in each grade, and that the grades were not consecutive. Educators also acknowledged that their impact data reflected the current level of performance of Wisconsin students in relationship to the Wisconsin Standards for Science.

However, participants also expressed an expectation that there would be a correspondence between grades 4 and 8 impact data. As shown in Table 9, the percentage of students classified as *Basic* and as *Advanced* was different across the two grades: most participants reported that they had expected these percentages to be more similar. DRC asked participants to elaborate on the pattern they expected to see in the impact data across both grades: participants reported that they had expected as many or fewer students to be classified as *Proficient* and above in grade 8 as in grade 4.

DPI heard the concern expressed by participants, and it considered the performance of Wisconsin students in grade 4 and grade 8. That is, DPI understood that participants were reasonably confident that their bookmarks were well-linked to the knowledge and skills expected of students in each performance level; however, participants were also concerned that there was not better alignment of the impact data across the two grades. In addition, DPI noted that the impact data associated with grade 4 were more similar to that of Wisconsin NAEP than that of grade 8.

To promote consistency in the performance standards across grades and testing programs, DPI considered adjustments to the cut scores using the conditional standard error of measurement (CSEM). The CSEM quantifies the amount of statistical error associated with any point on the test scale. If a student were to be tested and retested with the Wisconsin Forward Exam (or any other test), one would expect his or her scores to be similar but not exactly the same each time. The CSEM quantifies this "statistical noise" around any test score: one would expect the hypothetical student's test scores to fall within a range of ±1 CSEM about two-thirds of the time. If the difference between two test scores is less than 1 CSEM, it is often considered difficult to describe the difference as significant.

CSEM values are frequently used to adjust cut scores after standard settings, and DPI reserved the right to adjust participants' cut scores after the standard setting, should it be needed. However, DPI also wished to be highly transparent with participants and to share how CSEM values could be used to adjust cut scores.

As part of the across-grade discussion, participants were presented with hypothetical adjustments to their cut scores, for the consideration of the committee. The CSEM-based adjustments are summarized here.

- Grade 8 *Proficient*, +1 CSEM
- Grade 8 *Advanced*, +1 CSEM

These adjusted cut scores, along with their associated impact data, are shown in Table 10. Participants were told about CSEM and how they are sometimes used after standard settings. Participants were also told that these hypothetical adjustments were suggested by their feedback from earlier in the across-grade discussion: participants had expressed confidence in their content-based bookmarks, but dissatisfaction in the articulation of the impact data across both grades. Participants were told that if CSEM values were used to make adjustments to the cut scores, the underlying cut scores would still reflect the content-based expectations of the committee, and they would be tempered by the real-world implementation of the tests.

DPI adjusted the grade 8 *Proficient* and *Advanced* cut scores upward by 1 CSEM. The adjustments were made to promote consistency among the performance standards across grades. No adjustments were made to the grade 4 cut scores.

Participants viewed the hypothetical adjustments to the cut scores, as shown in Table 10. Participants generally reported that they found this set of impact data to better match their expectations for student performance in both grades. DRC asked if the committee would recommend making these adjustments, and accordingly make the cut scores shown in Table 10 their final recommendations. The committee gave its unanimous assent to these cut scores. Accordingly, the cut scores (and associated impact data) presented in Table 10 are considered to be the standard setting committee's final recommendations. Additional information concerning CSEM may be found in Appendix E of this report.

Table 10. Cut score recommendations and associated impact data from the standard setting, as taken from the articulation discussion

		Cut Scores from Articulation			Associated Impact Data			
Content	**Grade**	***Basic***	***Prof.***	***Adv.***	***B.B.***	***Basic***	***Prof.***	***Adv.***
Science	4	447	496	543	15.0%	32.2%	33.3%	19.5%
	8	653	695	737	17.7%	28.3%	31.5%	22.5%

Approved Cut Scores

The Superintendent of Public Instruction approved the cut scores for the Wisconsin Forward Exam assessments on June 5, 2019. The Superintendent approved the cut scores as taken from the across-grade discussion. These cut scores and associated impact data are shown in Table 11.

Table 11. Final, adjusted cut score recommendations and associated impact data

Content	Grade	Recommended Cut Scores			Associated Impact Data			
		Basic	***Prof.***	***Adv.***	***B.B.***	***Basic***	***Prof.***	***Adv.***
Science	4	447	496	543	15.0%	32.2%	33.3%	19.5%
	8	653	695	737	17.7%	28.3%	31.5%	22.5%

Evidence of Procedural Validity

The standard setting was conducted using a diverse, well-trained committee, and was perceived as valid by participants. This section supports these claims.

Committee Diversity

As part of the workshop evaluation, participants were asked about their backgrounds. The self-reported demographic characteristics of the participants are documented in this section. Of the 27 participants in the standard setting committee, all 27 responded to a request on the first day of the workshop to share background and demographic information. Later, 26 participants responded to the workshop evaluations administered on the last afternoon of the workshop.

Participants were asked to report their gender, race, and ethnicity. As shown in Table 12, most of the participants were female and white, and non-Hispanic.

Participants were asked to report their highest level of education, their profession, and the number of years in the profession. As shown in Table 13, approximately 89% of participants had master's or doctoral degrees, and more than 50% of the participants were teachers. Approximately two-thirds of participants indicated they had taught for 11 years or longer, and 40% reported they had worked for 16 years or longer in their profession.

In addition, participants responded whether they had experience with students in special education, English language learners (ELLs), alternate education, and vocational education. As shown in Table 14, a large majority of the committee had experience teaching special education students, ELLs, or both.

In Table 12 through 14, the percentages may not sum to 100% due to rounding and due to individual participants omitting their responses to certain questions, especially questions about race and ethnicity. When asked about their experience teaching different groups, participants were allowed to select multiple responses. The full results of the participant evaluations, including participants' self-reported demographic and background information, may be found in Appendix H.

Table 12. Participants' self-reported gender, race, and ethnicity

	Gender		Race/Ethnicity			
N	Female	Male	White	Black	Asian/Pacific Islander	Hispanic/ Latino
27	81%	19%	89%	0%	4%	7%

Table 13. Participants' self-reported level of education and profession

	Education				Profession				
N	High School	Bach-elor's	Mast-er's	Doct-rate	Teacher	Educator, Non-Teacher	Higher Education	Other	No Response
27	0%	11%	85%	4%	52%	22%	7%	15%	4%

Table 14. Participants' self-reported experience teaching special populations

	Experience with Special Populations				
N	Spec Ed, Mainstream	Spec Ed, Self-Contained	ELL	Alternative Education	Vocational Education
27	78%	11%	63%	15%	7%

Committee Training

During the standard setting workshop, it was clear to the facilitators that participants understood how to make judgments as part of the standard setting methodology (e.g., setting bookmarks). To confirm participants' knowledge of the methodology, they were given a short quiz, termed a *mid-process evaluation*, after training. The mid-process evaluation and detailed results are shown in Appendix B. Of the standard setting committee participants, all 27 submitted completed mid-process evaluations.

Participants answered items 1–5 on the mid-process evaluation correctly most of the time. This indicates that, on the whole, participants were well prepared to make judgments and that the training was effective. Results of the mid-process evaluation are shown in Table 15. All questions on the mid-process evaluation were scored dichotomously.

Table 15. Participants answering each item correctly on the training mid-process evaluation

N	#1	#2	#3	#4	#5
27	96%	89%	56%	100%	89%

As shown in Table 15, participants tended to do well on the mid-process evaluation quiz, but they had more difficulty with item #3 (which introduced a scenario where a student does not quite meet a given performance level based on his or her demonstrated skills). The answer for this item was discussed with the committee, and participants asked questions about the item before the process continued. After the mid-process evaluation, facilitators checked-in with participants to make sure they understood how cut scores could be represented in OIBs with bookmarks.

Participants' Perceived Validity of the Workshop

Participants indicated their perceived validity of the workshop and their recommendations as part of the workshop evaluation. Hambleton (2001) noted that evaluations are important evidence for establishing the validity of performance levels.

Satisfaction with Workshop and Recommendations

Generally, participants were satisfied with their recommendations and with the workshop as a whole.

Table 16 shows participants' level of satisfaction with their recommendations. Particularly, participants understood the connection between the benchmarks and their cut score recommendations, and participants generally agreed that the final recommendations reflected the work of the standard setting committee.

The full workshop evaluation, along with participants' responses, are presented in Appendix H of this report.

Table 16. Participants' agreement with various statements on the workshop evaluation regarding their satisfaction with the process and the final recommendations

	Response					Agree + Strongly Agree
Statement	**Strongly Disagree**	**Disagree**	**Neutral**	**Agree**	**Strongly Agree**	
The opening session provided a clear overview of the standard setting process.	3%	0%	0%	35%	62%	97%
I considered the standards when I placed my bookmarks.	0%	0%	0%	27%	73%	100%
I considered the threshold students when making my bookmarks.	0%	0%	0%	27%	73%	100%
Discussing the threshold students helped me make my bookmarks.	0%	0%	0%	31%	69%	100%
My group's work was reflected in the presentation of recommendations across grades.	0%	0%	0%	27%	73%	100%
Overall, I believe my opinions were considered and valued by my group.	0%	0%	0%	15%	85%	100%

References

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

Data Recognition Corporation. (2016). Wisconsin standard setting 2016 final technical report. Madison, WI: Wisconsin Department of Public Instruction.

Egan, K.L., Schneider, M.C., & Ferrara, S. (2012). Performance level descriptors: History, practice, and a proposed framework. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 79–106). New York, NY: Routledge.

Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 89–116). Mahwah, NJ: Lawrence Erlbaum.

Karantonis, A., & Sireci, S. G. (2006). The Bookmark standard setting method: A literature review. *Educational Measurement: Issues and Practice, 25*, 4–12.

Lewis, D. M., Green, D. R., Mitzel, H. C., Baum, K., & Patz, R. J. (1998, April). The bookmark standard setting procedure: Methodology and recent implementations. Paper presented at the annual meeting of the National Council for Measurement in Education, San Diego, CA.

Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996, June). Standard setting: A bookmark approach. In D. R. Green (Chair), *IRT-based standard-setting procedures utilizing behavioral anchoring.* Symposium conducted at the Council of Chief State School Officers National Conference on Large-scale Assessment, Phoenix, AZ.

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 ed., pp. 225–253). New York, NY: Routledge.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, 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.

Phillips, G.W. (2012). The benchmark method of standard setting. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 232–346). New York, NY: Routledge.

Wisconsin Department of Public Instruction. (2017). *Wisconsin standards for science.* Retrieved from https://dpi.wi.gov/sites/default/files/imce/science/WI-Standards-for-Science-2017.pdf

Appendix A

Agenda

Workshop Agenda

Wisconsin Forward Exam
Grades 4 and 8 Science

Standard Setting Workshop
Madison, Wisconsin
May 29–30, 2019

Welcome to the standard setting workshop for the Wisconsin Forward Exam! As part of your work, you will focus on Grade 4 Science or Grade 8 Science.

The Wisconsin Department of Public Instruction (DPI) and Data Recognition Corporation (DRC) would like to thank you for your time and expertise during this important process. Please use this agenda to orient yourself during the workshop. If you have any questions or concerns, please do not hesitate to contact a facilitator.

Wednesday, May 29

Welcome!

7:30 AM **Breakfast and Participant Registration**
Participants register at the reception table to sign the confidentiality agreement, receive a nametag, and collect any other necessary information.

8:30 AM **Opening Session**
DPI welcomes participants, overviews the testing program, discusses the reasons for the standard setting, and describes the desired outcome of the workshop.

8:45 AM **Participant Training**
DRC introduces participants to the Bookmark Standard Setting Procedure and explains how a cut score can be represented in the ordered item book (OIB) as a bookmark.

10:00 AM **Break and Adjournment to Breakout Rooms**

10:15 AM **Distribution of Secure Materials**
After brief introductions, DRC distributes the secure materials.

- Be sure to write your name on each of the secure materials.
- Your folder number is shown on the cover of your folder. Write your name and folder number on the materials sign-out list.
- Please remember that the secure materials, including the OIBs and item maps, must remain in your breakout room; and that your discussions of the secure materials must remain confidential.

Discuss Threshold Students

10:30 AM **Discuss the Performance Level Descriptors (PLDs) and the Threshold Students**

As a table, appoint a scribe to take notes during these discussions.

- Look over the content standards to refresh your understanding of them.
- Then review and discuss the PLDs at your table.
- Consider the knowledge, skills, and abilities that students are expected to demonstrate if they are *Proficient.* Then do the same for *Advanced* and *Basic*.
- Participants then engage in discussions about the knowledge and skills they expect to be demonstrated by each threshold student, that is, a student who is just barely entering a performance level.
- There are three threshold students to consider: just *Basic*, just *Proficient*, and just *Advanced*.
- For each threshold student, create a brief, bulleted list that describes the skills expected of the threshold student.
- To engage in these discussions, participants refer to the PLDs, the content standards, and their knowledge of students.

11:30 AM **Discuss the Threshold Students Across Tables**

Using the PLDs and content standards, participants discuss the threshold students across tables.

- A spokesperson from your table should be prepared to report some of the highlights from your table's discussion of the threshold students.
- During the discussion, refer to the PLDs and the content standards.
- Take notes during the discussion and update your bulleted lists of the skills expected of each of the three threshold students.

12:15 PM **Lunch**

The group breaks for 45 minutes.

1:00 PM **Examine the Test Items**

Participants examine the test items from the student's perspective.

- Study and answer the items to get a sense of what is measured by the test and how it is measured.
- Although some discussion about individual test items is normal, focus should be toward examining the test and away from prolonged debate.
- If necessary, use the provided index cards to record comments about test items.

Study the Ordered Item Booklet and Round 1

1:30 PM **Break**

1:45 PM **Discussion of the Ordered Item Booklet (OIB)**

Group leaders introduce this task by instructing participants to find the item map in their secure materials. The group leader leads the group in a review of each column on the item map. Participants at each table examine the items in the OIB.

- Participants engage in a discussion with everyone at their table about each of the items in the OIB. Starting with the first item, participants briefly discuss each item in turn, focusing on what each item measures and what makes it harder than the previous items. Each participant records these details on his or her item map.
- Group leaders remind participants to use the index cards, as necessary, to record comments about items.
- Group leaders monitor the tables to check that each participant has a chance to speak.

3:45 PM **Bookmark Placement Training**

DRC introduces bookmark placement, explaining and illustrating how bookmarks are placed and what bookmarks mean.

- DRC explains how participants make cut score recommendations by placing bookmarks in the OIB.
- After the training, a short mid-process evaluation is administered and discussed.

4:15 PM **Round 1 Bookmark Placement**

Group leaders direct all participants to place their Round 1 bookmarks.

- Remember that bookmark placement is an individual activity.
- Participants mark their bookmark placements on their bookmark worksheets.
- Then participants transfer their bookmark placements into the *kiosk* system.

4:45 PM **Secure Materials Collection**

Group leaders facilitate collection of the secure materials from all participants. All participants return their secure materials to the facilitator for safekeeping.

5:00 PM **Dismissal**

Rounds 2 and 3

7:30 AM **Breakfast and Participant Sign-in**
Please be sure to sign in for the day.

8:30 AM **Discuss Round 1 Results as a Table**
The group leader presents results from Round 1, as well as *benchmarks* from other tests of science. When instructed by the group leader, participants begin a discussion of their Round 1 judgments at their tables.

- Participants discuss the items between the lowest and highest bookmarks, explaining the rationales behind their judgments.
- Remember that the benchmarks are for reference and should be considered alongside all the other information presented at the workshop.

10:00 AM **Round 2 Bookmark Placement**
Group leaders direct all participants to place their Round 2 bookmarks.

- Remember that bookmark placement is an individual activity.
- Participants mark their bookmark placements on their bookmark worksheets.
- Then participants transfer their bookmark placements into the *kiosk* system.

10:30 AM **Break**

10:45 AM **Discuss Round 2 Results as a Group**
DRC presents a summary of the Round 2 judgments to the entire group. Afterwards, the group leaders begin a discussion of each bookmark with all the tables, similar to the table-level discussions of Round 2.

12:15 PM **Lunch**
The group breaks for 45 minutes.

1:00 PM **Round 3 Bookmark Placement**
Group leaders direct all participants to place their Round 3 bookmarks.

- Remember that bookmark placement is an individual activity.
- Participants mark their bookmark placements on their bookmark worksheets.
- Then participants transfer their bookmark placements into the *kiosk* system.

1:30 PM **Break**

1:45 PM **Presentation of Recommendations**

A summary of the recommendations are shown to the participants. The group leader shows the recommendations from grades 4 and 8 to each team.

- Participants should consider the consistency of the recommendations across the two grades. If needed, the group leader will record recommendations on adjustments to the cut scores to improve this across-grade articulation.

2:15 PM **PLD Refinements**

The group leader introduces the activity by inviting participants to recommend refinements to the PLDs that make the document clearer and more useful to educators in the field. Then, the group leader elicits these recommendations.

- Participants should use what they have learned during the standard setting to suggest refinements to the PLDs.
- PLD refinements should *not* adjust the overall level of rigor associated with each descriptor. Instead, the refinements should make the descriptors clearer.
- As needed, the group will break into tables to focus on different aspects of the PLDs, and will then reconvene to discuss their all the proposed refinements.

4:45 PM **Evaluation**

Each participant completes an evaluation of the standard setting.

4:50 PM **Secure Materials Collection**

Group leaders facilitate collection of the secure materials from all participants. All participants return their secure materials to the facilitator for safekeeping.

5:00 PM **Dismissal**

DPI and DRC thank you for your participation in the Wisconsin Forward Exam Standard Setting!

Agenda at a Glance

Wisconsin Science Standard Setting

Wednesday, May 29

7:30 AM	Breakfast and Participant Registration
8:30 AM	Opening Session
8:45 AM	Participant Training
10:00 AM	Break and Adjournment to Breakout Rooms
10:15 AM	Distribution of Secure Materials
10:30 AM	Discuss the Performance Level Descriptors (PLDs) and the Threshold Students
11:30 AM	Discuss the Threshold Students Across Tables
12:15 PM	Lunch
1:00 PM	Examine the Test Items
1:30 PM	Break
1:45 PM	Discussion of the Ordered Item Booklet (OIB)
3:45 PM	Bookmark Placement Training
4:15 PM	Round 1 Bookmark Placement
4:45 PM	Secure Materials Collection
5:00 PM	Dismissal

Thursday, May 30

7:30 AM	Breakfast and Participant Sign-in
8:30 AM	Discuss Round 1 Results as a Table
10:00 AM	Round 2 Bookmark Placement
10:30 AM	Break
10:45 AM	Discuss Round 2 Results as a Group
12:15 PM	Lunch
1:00 PM	Round 3 Bookmark Placement
1:30 PM	Break
1:45 PM	Presentation of Recommendations
2:15 PM	PLD Refinements
4:45 PM	Evaluation
4:50 PM	Secure Materials Collection
5:00 PM	Dismissal

Appendix B

Training Presentation and Materials

Wisconsin Science Standard Setting

Grades 4 and 8 Science

Opening Session & Training
May 29, 2019

Opening Session

Viji Somasundaram
Director, Office of Student Assessment
Wisconsin Department of Public Instruction

Training Session

Rick Mercado
Director, Research
Data Recognition Corporation

DRC Staff Members

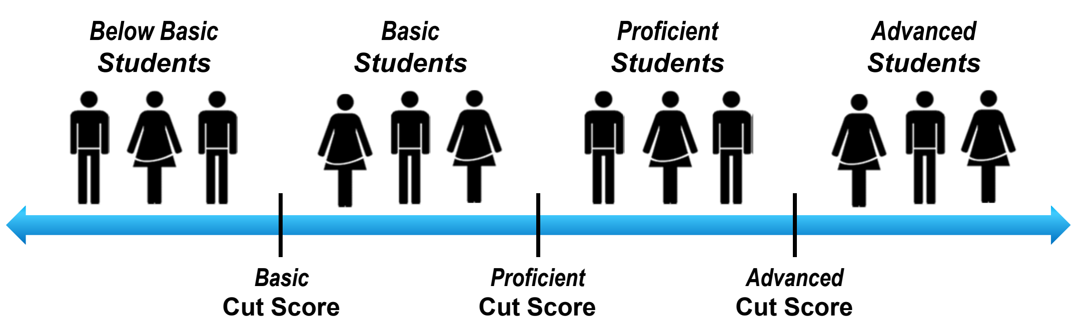
- **Rick Mercado**, Group Leader
- **Michelle Boyer**, Group Leader
- **Dave Durette**, Science Content Expert
- **Jeannie Hemsworth**, Science Content Expert
- **Mary Basch**, Science Content Expert
- **Sara Kendallen**, Data Analyst
- **Julie Pointner Korts**, Data Analyst
- **Courtney Johnson**, Project Manager
- **Natalie Flynn**, Project Manager

Workshop Goal

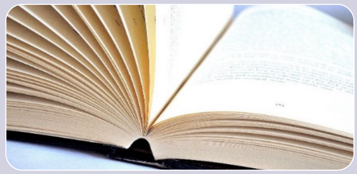
- To recommend cut scores that categorize students into one of four performance levels:
  - *Below Basic*
  - *Basic*
  - *Proficient*
  - *Advanced*

Cut Scores & Performance Levels

- Three cut scores classify students into four performance levels.



Bookmark Procedure



Item-centered method

Content-based decisions

Iterative process

Process Overview

Today

- Review the content standards
- Study the performance level descriptors (PLDs)
- Discuss the threshold students
- Review the test format
- Study the ordered item booklet
- **Round 1**: Make cut score recommendations on your own

Tomorrow

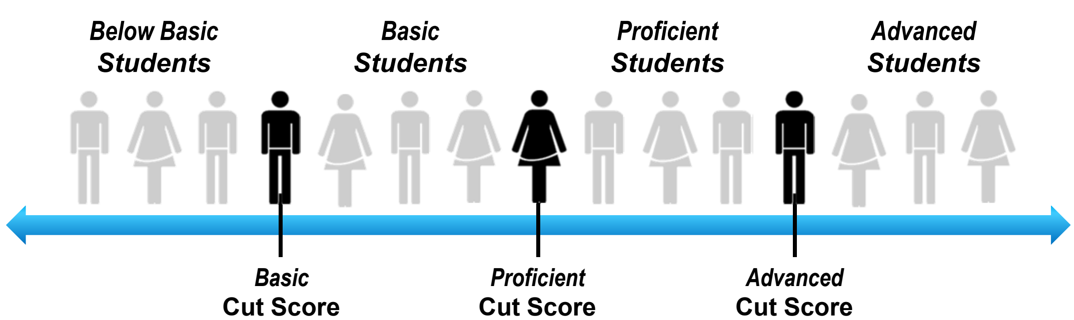
- Discuss recommendations with your table
- **Round 2**: Make cut score recommendations on your own
- Discuss your recommendations with your group
- **Round 3**: Make cut score recommendations on your own
- Review the group's recommendations
- Refine the PLDs
- Evaluate the workshop

Performance Level Descriptors (PLDs)

- PLDs describe the knowledge, skills, and abilities expected of students in each performance level.
  - They are linked to the content standards.
  - PLDs describe students in the middle of each level, not on the *thresholds*.

PLDs and Performance Levels

- PLDs describe the student in the middle of each performance level.



Three Threshold Students

- Threshold students are those just barely leaving one level and entering the next level.
  - The PLDs do *not* describe these students directly.
  - There are three threshold students.

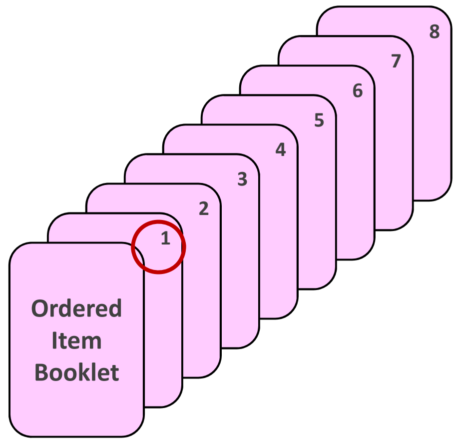
Threshold Below Basic/**Basic** ***Student***	Threshold Basic/**Proficient** ***Student***	Threshold Proficient/**Advanced** ***Student***

Take the Test

- By taking the online tools training (OTT), you will better understand students' testing experience on test day.
  - The OTT shows how many items are grouped in scenarios.
  - The OTT also shows the various item types.
- After taking the OTT, you will see the live items on paper, ordered by difficulty, in the ordered item booklet (OIB).

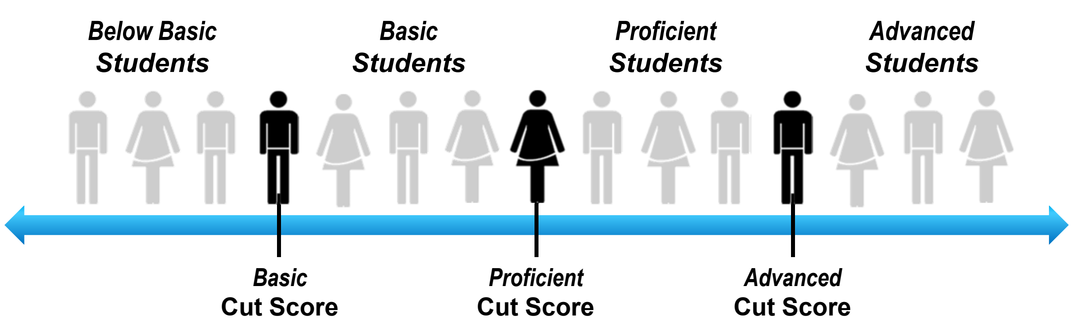
Ordered Item Booklet (OIB)

- The OIB comprises items from the spring test.
  - One item per page
  - Easiest item first
  - Hardest item last
  - Items ascend in difficulty as based on student performance



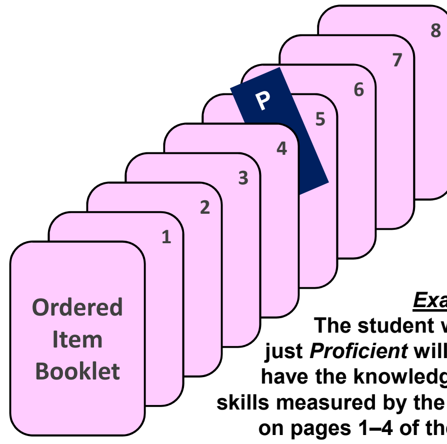
Three Threshold Students

- Bookmark judgments and cut scores are linked to the student *just* in each level.



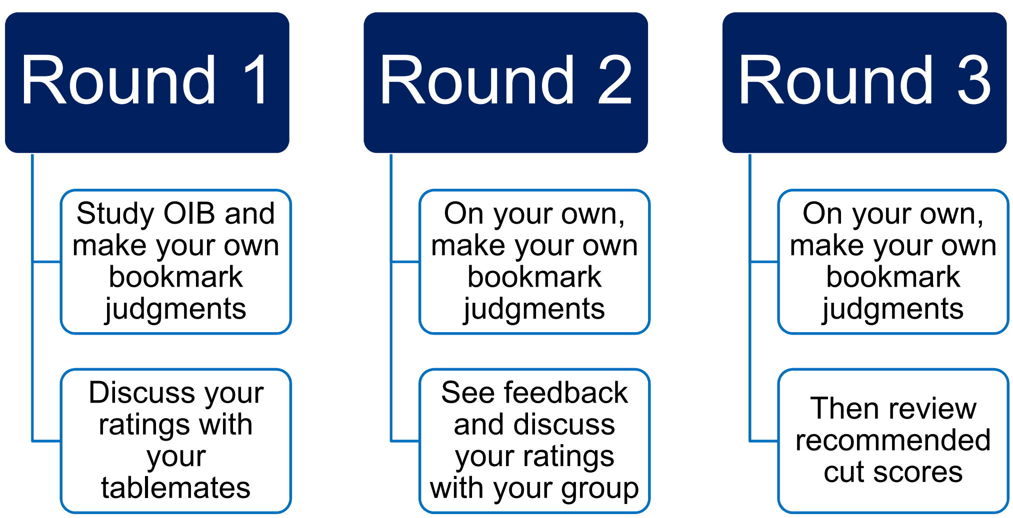
Threshold Students and the OIB

- You will consider the three threshold students.
- You will make statements in the OIB using *bookmarks.*
- These bookmarks are linked to cut score recommendations.



Example:
The student who is just *Proficient* will likely have the knowledge and skills measured by the items on pages 1–4 of the OIB.

Three Rounds



Roles and Responsibilities

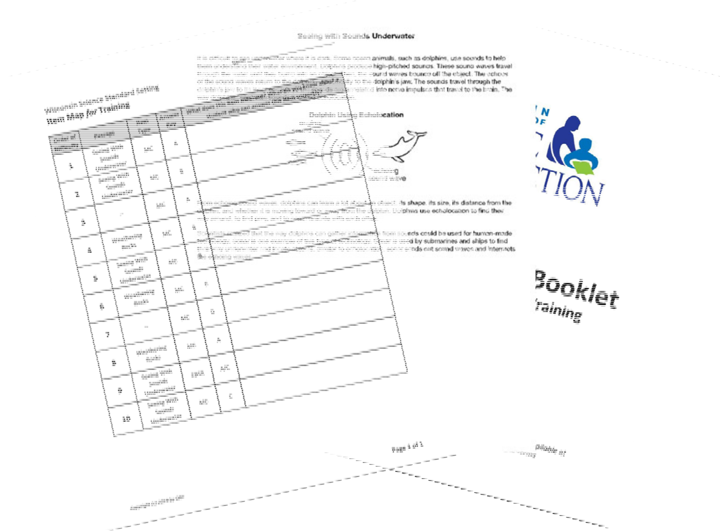
- You will recommend performance standards to DPI.
- During the workshop, remember to:
  - Contribute to discussions at your table
  - Participate in group-wide discussions
  - Place your bookmarks independently
  - Ask a member of staff any questions
  - Use workshop materials only in meeting rooms
  - Keep workshop conversations confidential

Workshop Security

- Your facilitators will collect your materials each afternoon in a structured way.
- Always leave the workshop materials in the meeting rooms. Do not discuss the contents of the materials outside your meeting room.
- You are welcome to use phones, tablets, and laptops in the lunchroom and hallways, but never in the meeting rooms.

Training Materials

- Item map
- Training scenarios
- Training ordered item booklet (OIB)



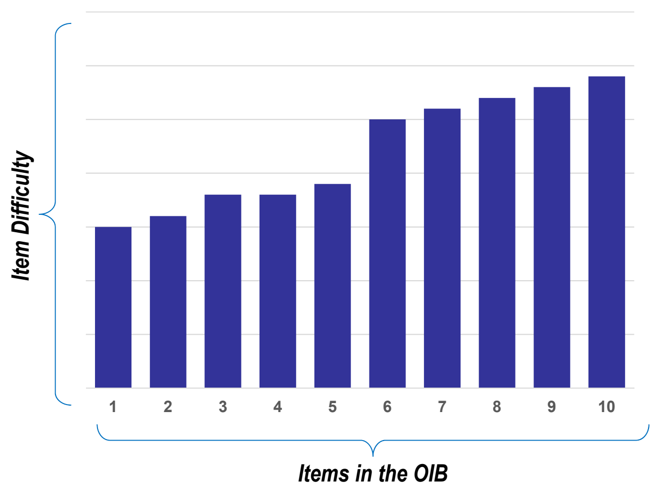
Item Map

Name: ___________________________

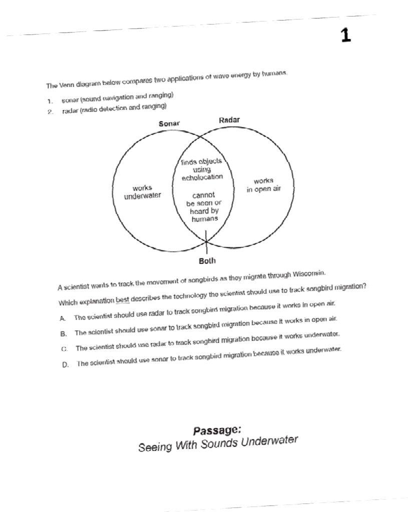
Wisconsin Science Standard Setting
Item Map *for* Training

Order of Difficulty	Passage	Item Type	Answer Key	What does this item measure? What do you know about a student who can answer this item correctly?
1	Seeing With Sounds Underwater	MC	A	
2	Seeing With Sounds Underwater	MC	B	
3	--	MC	A	
4	Weathering Rocks	MC	B	
5	Seeing With Sounds Underwater	MC	B	
6	Weathering Rocks	MC	C	

Example: Item Separation Chart



Examining an Item



1

The Venn diagram below compares two applications of wave energy by humans.

1. sonar (sound navigation and ranging)
2. radar (radio detection and ranging)

A scientist wants to track the movement of songbirds as they migrate through Wisconsin.

Which explanation best describes the technology the scientist should use to track songbird migration?

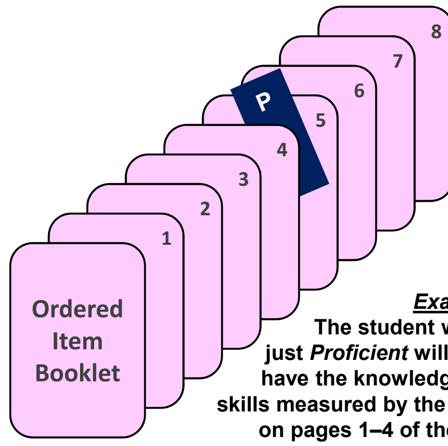
A. The scientist should use radar to track songbird migration because it works in open air.
B. The scientist should use sonar to track songbird migration because it works in open air.
C. The scientist should use radar to track songbird migration because it works underwater.
D. The scientist should use sonar to track songbird migration because it works underwater.

Passage:
Seeing With Sounds Underwater

- Make a brief note to yourself about what the item measures.
  - What knowledge and skills does a student need to have in order to answer the item correctly?
  - If a student answers the item correctly, what do you know about the student?

Finding a Possible Bookmark Range

- You will consider the three threshold students.
- You will make statements in the OIB using *bookmarks.*
- These bookmarks are linked to cut score recommendations.



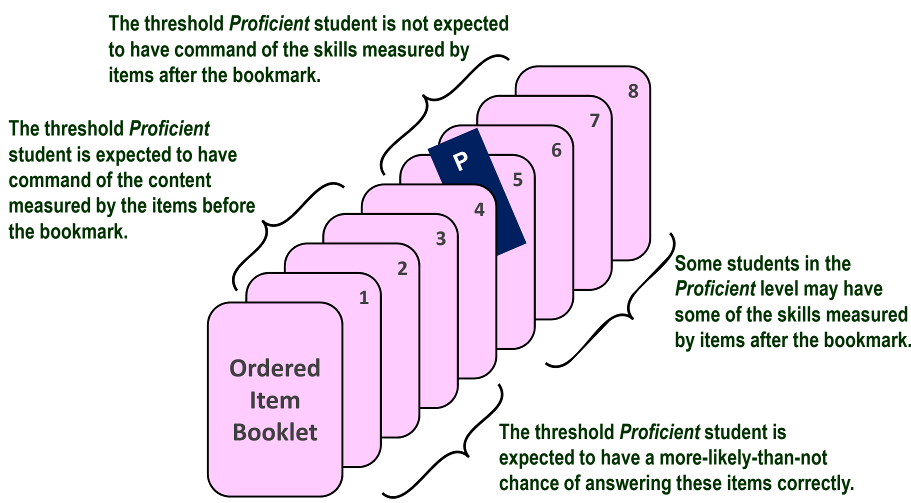
Example:
The student who is just *Proficient* will likely have the knowledge and skills measured by the items on pages 1–4 of the OIB.

Possible Bookmark Range

- You will find a range of items where you could set your bookmark.
  - The possible bookmark range may be a couple of items wide, or may be more than that.
  - Do not get stuck on a single item.

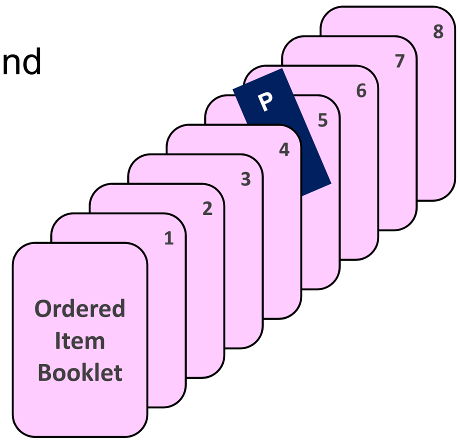
Finding the Possible Bookmark Range

- Progress through the OIB until you reach an item that the threshold student would not have a more-likely-than-not chance of answering correctly.
  - This is the start of your possible bookmark range.
- Keep going until you have reached the last item that a student would have a more-likely-than-not chance of answering correctly.
  - The possible bookmark range ends after that page.



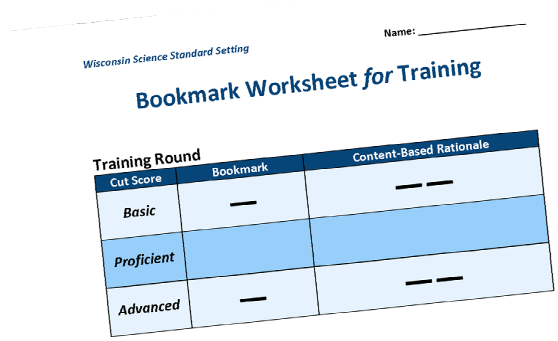
Recording Your Bookmark

- Place your bookmark within your possible bookmark range.
  - Use the PLDs, the benchmarks, and your professional judgment as guides.
- Record the page number **after** your bookmark.



Bookmark Worksheet

- Write your bookmarks on the *Bookmark Worksheet*.
  - You will place three bookmarks.
  - Write a few words to help you remember why you placed your bookmarks where you did.



Name: ________

Wisconsin Science Standard Setting

Bookmark Worksheet *for* Training

Training Round

Cut Score	Bookmark	Content-Based Rationale
Basic	—	——
Proficient		
Advanced	—	——

Recording Your Bookmarks

- In the actual workshop, you will then record your bookmarks in an online system.
  - You will record your bookmarks online, *not* your rationales.

Pacing

- Some people will take longer than others to study the test items and place their bookmarks.
  - Today, Round 1 bookmark placement is the last activity for the day. Please be considerate of others as you leave the workshop.
  - Tomorrow, during conversations before Rounds 2 and 3, please be considerate of others at your table and in the room.

Practice Exercise

Grades 4 and 8 Science

Opening Session & Training
May 29, 2019

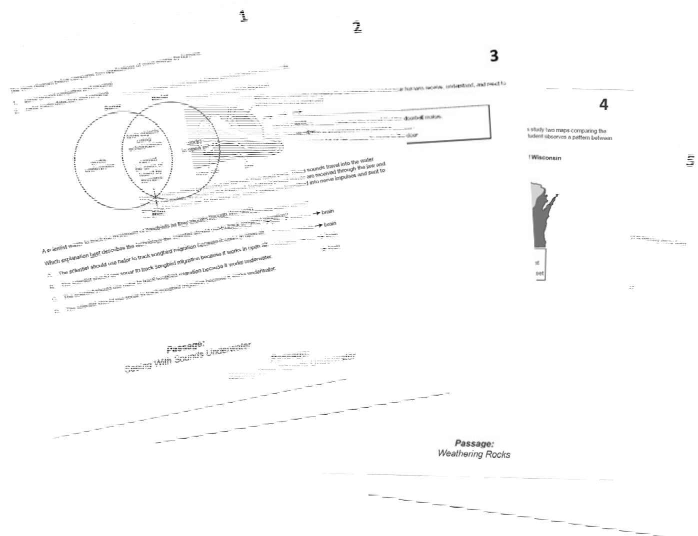
Consider the Threshold Student

- Review these sample PLDs for *Basic* and *Proficient*.
  - Consider the student who is just barely *Proficient*.
  - What knowledge, skills, and abilities would you expect of this threshold student?

	Basic	Proficient
	Student demonstrates partial understanding of and ability to apply the knowledge and skills for their grade level that are associated with college content-readiness.	Student demonstrates adequate understanding of and ability to apply the knowledge and skills for their grade level that are associated with college content-readiness.
Standard: **Students use science and engineering practices, crosscutting concepts, and an understanding of life science disciplinary core ideas to make sense of phenomena and solve problems.**	can analyze evidence to determine if it supports a claim about the role of external structures of plants and animals in supporting survival and reproduction.	can provide feedback and ask questions about a claim and its supporting evidence about the role of internal and external structures of plants and animals in supporting survival, growth, behavior, and reproductive success.
	can give evidence of the sequence of events resulting in a given animal behavior (i.e., sensory input, sense receptor, brain processing, behavioral output).	can develop a model of an animal behavior (phenomenon) showing various components (i.e., sensory input, sense receptor, the brain, behavioral output) working together as a system.
	can describe how data shows a cause and effect relationship between an environmental stimuli and an animal's behavior.	can develop a model of sensory systems showing how animals' memories can impact future behavior, survival, and reproduction.

Study the Test Items

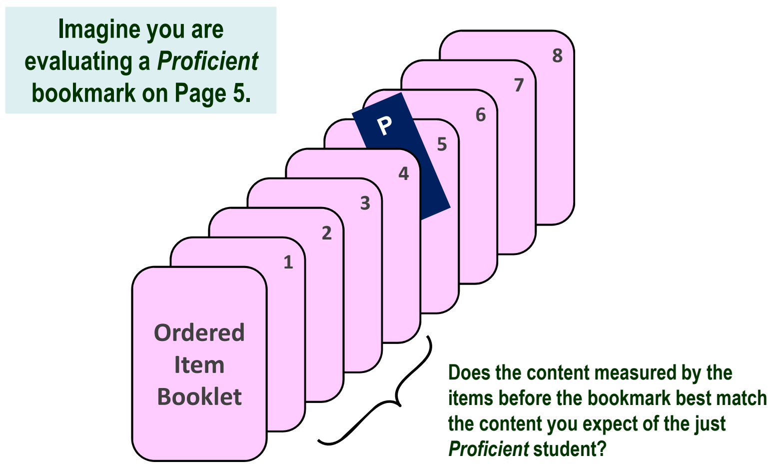
- For each question, ask yourself:
  - what does the item measure?
  - if a student can answer the item correctly, what do we know he or she can do?



Place Your Bookmark

- Consider the *Proficient* threshold student.
- The student is expected to have at least a more-likely-than-not chance of answering items correctly before the bookmark.
- The probability after the bookmark is less than half, but not zero.

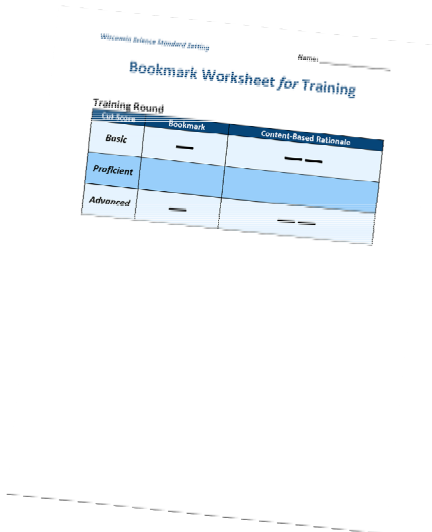
Evaluating a Bookmark Holistically



Write a Rationale

- Good rationales link the content of the items in front of the bookmark to the content-based expectations for the threshold student.
  - For example, “Students must describe important parts of the cell in a real-world context, as expected of the threshold student.”
  - Or, “Students have to begin to make simple inferences from empirical data, not just report those data, as listed in PLDs.”
- Not-so-good rationales don’t make reference to the content of the items.
  - For example, “The first evidence-based selected response item is just after the bookmark.”

Make Your Bookmark Placements


Wisconsin Science Standard Setting

Name: ______

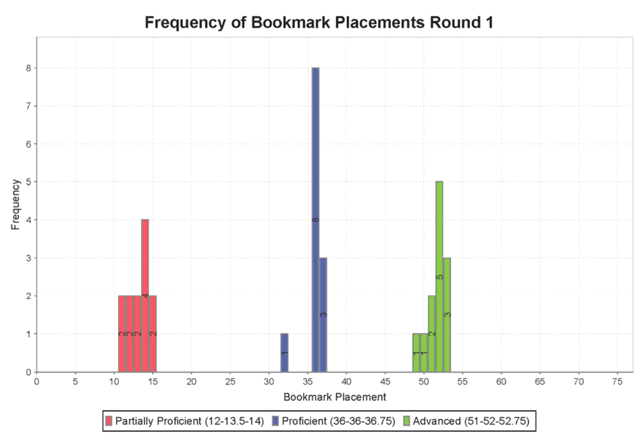
Bookmark Worksheet *for* Training

Training Round

Cut Score	Bookmark	Content-Based Rationale
Basic	___	___
Proficient		
Advanced	___	___

- Write your bookmark placement on your training Bookmark Worksheet.
- Turn your Worksheet over when you're done.

Example: Round 1 Feedback



Discussion of Round 1 Ratings

- In the actual workshop, you will discuss your Round 1 bookmarks at your table.
- Feel free to discuss:
  - Your bookmarks
  - Your possible bookmark ranges (and any overlaps)
- After discussion, you will have a second opportunity to make bookmark judgments.
  - You can change any, all, or none of your bookmarks.
  - Bookmark placement is always an individual activity.

Suggestions for Discussions

- Practice active listening.
- Be open to changing your mind.
- Work to understand your colleagues' rationales for their bookmark placements.
- In a respectful manner, feel free to ask questions of your colleagues.
- Do not discuss your bookmarks until everyone at the table has placed theirs.
- Keep the contents of your discussions private.

After Round 2

- After Round 2, you will see:
  - the medians from the group's Round 2 bookmarks
  - *impact data*, the percent of students that would be classified in each performance level if the Round 2 cut scores were implemented
  - *benchmarks*, the percent of Wisconsin students who were classified in each performance level on the NAEP Science test
- The impact data and benchmarks are provided as contextual information for you to consider.

Round 3

- After Round 2, you will discuss your bookmark placements *across tables*.
  - Again, you will share where you placed your bookmarks and why you placed them there.
- Then you will place your Round 3 bookmarks.
  - Bookmark placement is always an individual activity.

After Round 3

- After Round 3, your facilitator will show you a presentation of the Round 3 recommendations from both grades.
  – You will be asked to look at the articulation of the performance standards across grades.
  – You may wish to consider adjustments to your recommendations to improve the articulation across grades.

PLD Refinement

- Participants will suggest refinements to the PLDs.
  – You will be asked to recommend refinements to the PLDs to make them even clearer and more useful to educators.
  – These refinements should not adjust the overall level of rigor associated with each performance level.

After the Workshop

- Your recommendations will be considered by DPI.
  – The recommendations from both groups will be considered by DPI and its advisors.
  – The Superintendent has the ultimate responsibility to implement the cut scores.

Workshop Structure

- Study PLDs
- Study OIB and make Round 1 ratings
- Discuss Round 1 at tables
- Make Round 2 ratings
- Discuss Round 2 as a group
- Make Round 3 ratings
- Review recommendations
- Refine PLDs

Questions

- Do you have any questions?
  - If questions come up later, ask your facilitator, or write them on an index card.

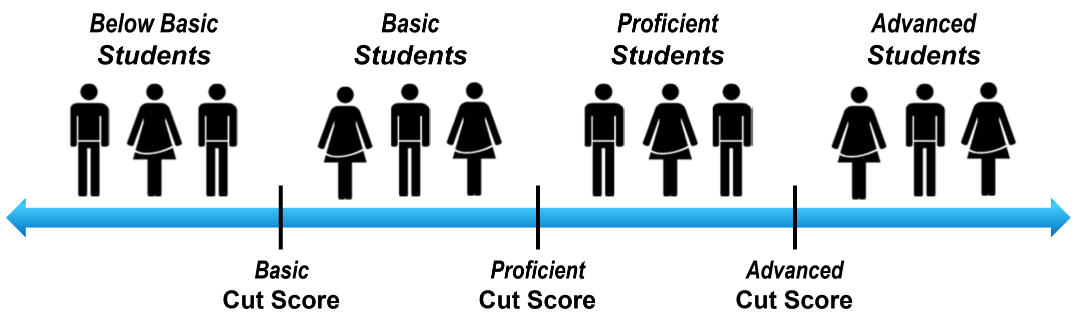
Bookmark Refresher Training

Grades 4 and 8 Science

Opening Session & Training
May 29, 2019

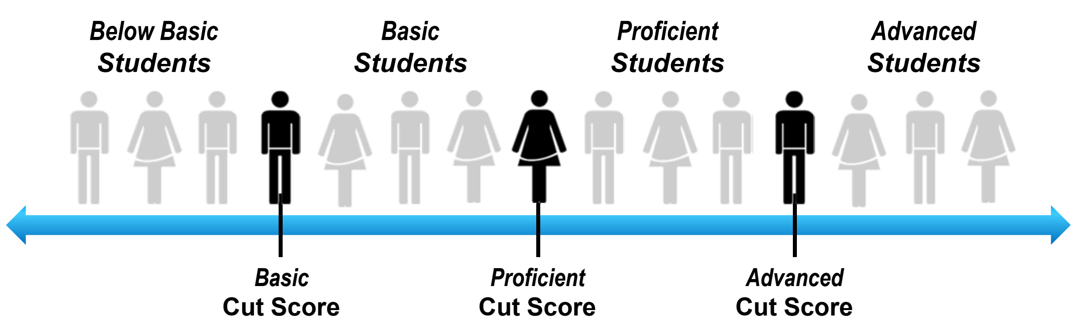
Cut Scores & Performance Levels

- Three cut scores classify students into four performance levels.



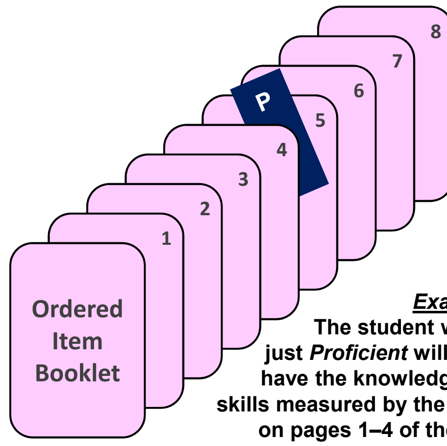
Three Threshold Students

- Bookmark judgments and cut scores are linked to the student *just* in each level.



Threshold Students and the OIB

- You will consider the three threshold students.
- You will make statements in the OIB using *bookmarks.*
- These bookmarks are linked to cut score recommendations.



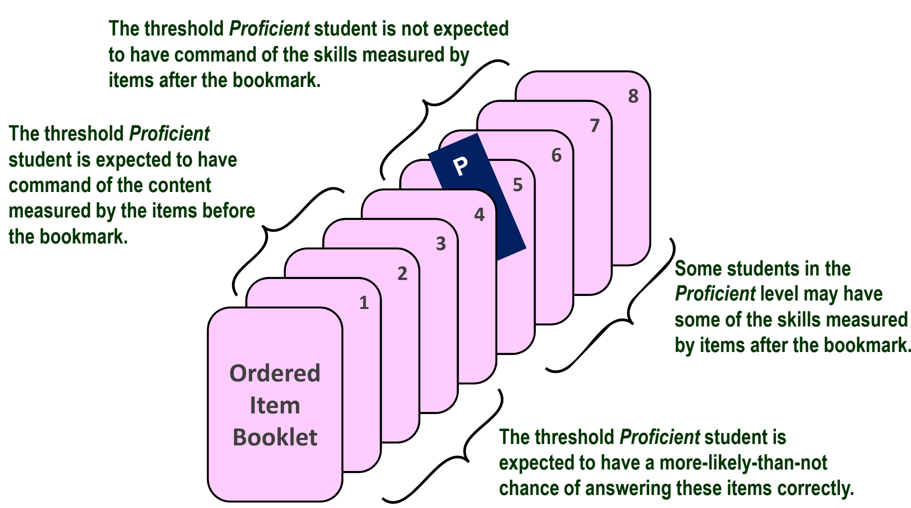
Example:
The student who is just *Proficient* will likely have the knowledge and skills measured by the items on pages 1–4 of the OIB.

Possible Bookmark Range

- You will find a range of items where you could set your bookmark.
  - The possible bookmark range may be a couple of items wide, or may be more than that.
  - Do not get stuck on a single item.

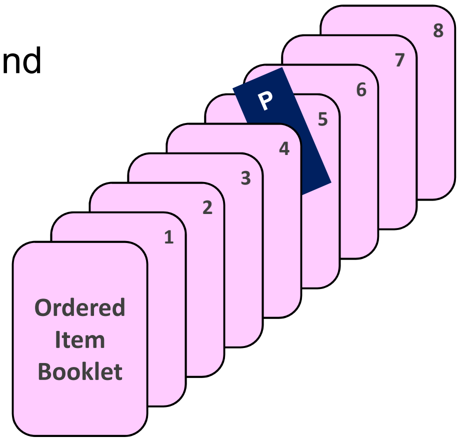
Finding the Possible Bookmark Range

- Progress through the OIB until you reach an item that the threshold student would not have a more-likely-than-not chance of answering correctly.
  - This is the start of your possible bookmark range.
- Keep going until you have reached the last item that a student would have a more-likely-than-not chance of answering correctly.
  - The possible bookmark range ends after that page.



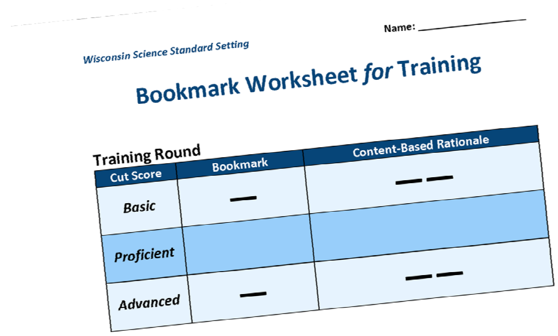
Recording Your Bookmark

- Place your bookmark within your possible bookmark range.
  - Use the PLDs, the benchmarks, and your professional judgment as guides.
- Record the page number **after** your bookmark.



Bookmark Worksheet

- Write your bookmarks on the *Bookmark Worksheet*.
  - You will place three bookmarks.
  - Write a few words to help you remember why you placed your bookmarks where you did.

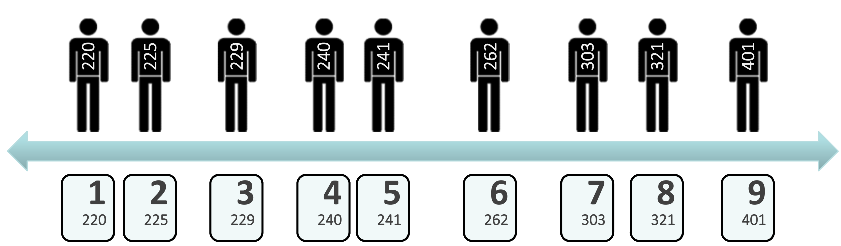


Not "Number Correct"

- Your bookmark placement does ***not*** correspond directly with the number of points a student needs to earn to be classified in a performance level.
  - For example, if you place your Proficient bookmark on Page 10, this does *not* mean a student needs to get 10 points on the test to be Proficient.
- Instead, your cut score recommendations are made on the test scale.

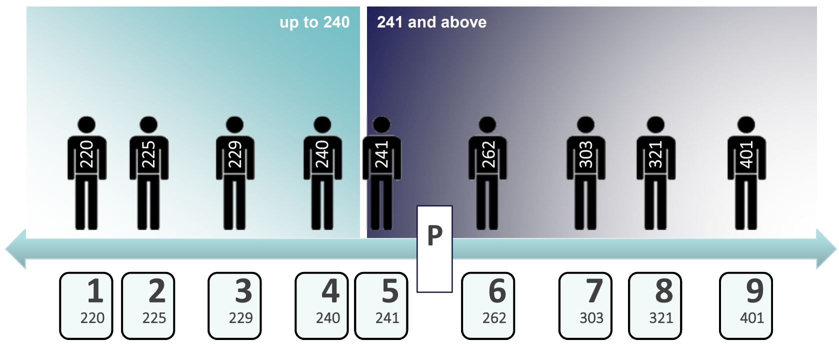
Test Scale

- Items are ordered by difficulty, easy to hard.
- Students are ordered by performance, low to high.



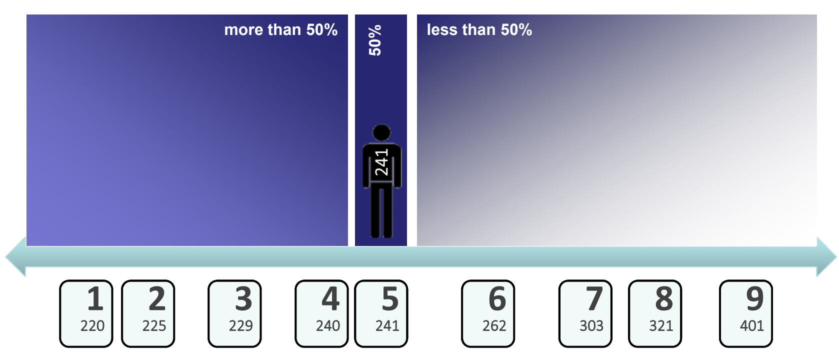
Cut Score

- The bookmark separates items.
- The cut score separates students.



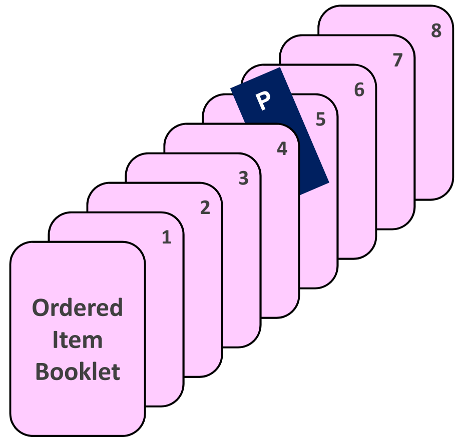
Threshold Student

- The threshold student has a 50% chance of answering the item just before the bookmark.



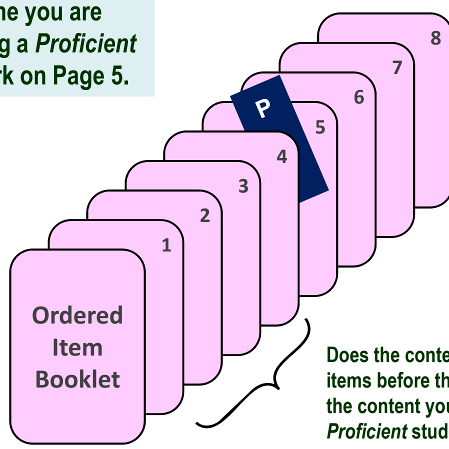
Place Your Bookmark

- Consider the *Proficient* threshold student.
- The student is expected to have at least a more-likely-than-not chance of answering items correctly before the bookmark.
- The probability after the bookmark is less than half, but not zero.



Evaluating a Bookmark Holistically

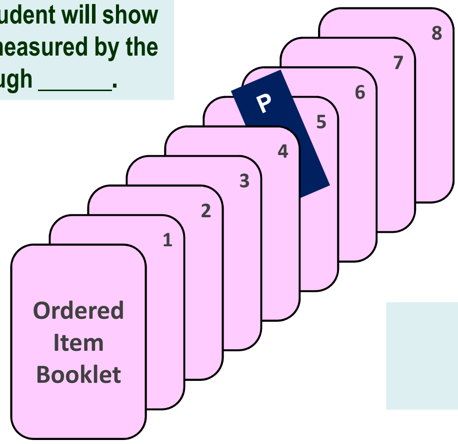
Imagine you are evaluating a *Proficient* bookmark on Page 5.



Does the content measured by the items before the bookmark best match the content you expect of the just *Proficient* student?

Bookmark on Page 5

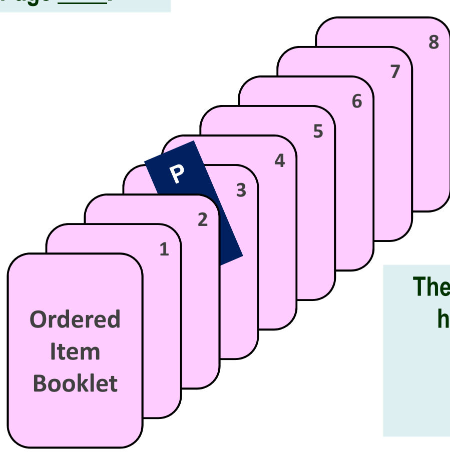
If the *Proficient* bookmark is on Page 5, the threshold *Proficient* student will show command of the content measured by the items on Pages 1 through ______.



Item 2 is ______ than Item 7.
a) Easier
b) Harder

Bookmark on Page ___

This is a bookmark on Page ____.



The threshold *Proficient* student would have a _____ chance of answering Items 1–2 correctly.
a) More likely than not
b) Nearly 100% chance

Write a Rationale

- Good rationales link the content of the items in front of the bookmark to the content-based expectations for the threshold student.
  - For example, “Students must describe important parts of the cell in a real-world context, as expected of the threshold student.”
  - Or, “Students have to begin to make simple inferences from empirical data, not just report those data, as listed in PLDs.”
- Not-so-good rationales don’t make reference to the content of the items.
  - For example, “The first evidence-based selected response item is just after the bookmark.”

Rounds

- **Round 1**: Place bookmarks on your own
- **Round 2**: See feedback, discuss with your tablemates, place bookmarks on your own
- **Round 3**: See feedback and impact, discuss with the group, place bookmarks on your own

Mid-Process Evaluation

- Before we continue, let's complete the mid-process evaluation.

Name: ________________________________

Wisconsin Science Standard Setting

Item Map *for* Training

Order of Difficulty	Passage	Item Type	Answer Key	What does this item measure? What do you know about a student who can answer this item correctly?
1	Seeing With Sounds Underwater	MC	A	
2	Seeing With Sounds Underwater	MC	B	
3	--	MC	A	
4	Weathering Rocks	MC	B	
5	Seeing With Sounds Underwater	MC	B	
6	Weathering Rocks	MC	C	
7	--	MC	D	
8	Weathering Rocks	MC	A	
9	Seeing With Sounds Underwater	EBSR	A/C	
10	Seeing With Sounds Underwater	MC	C	

Sample

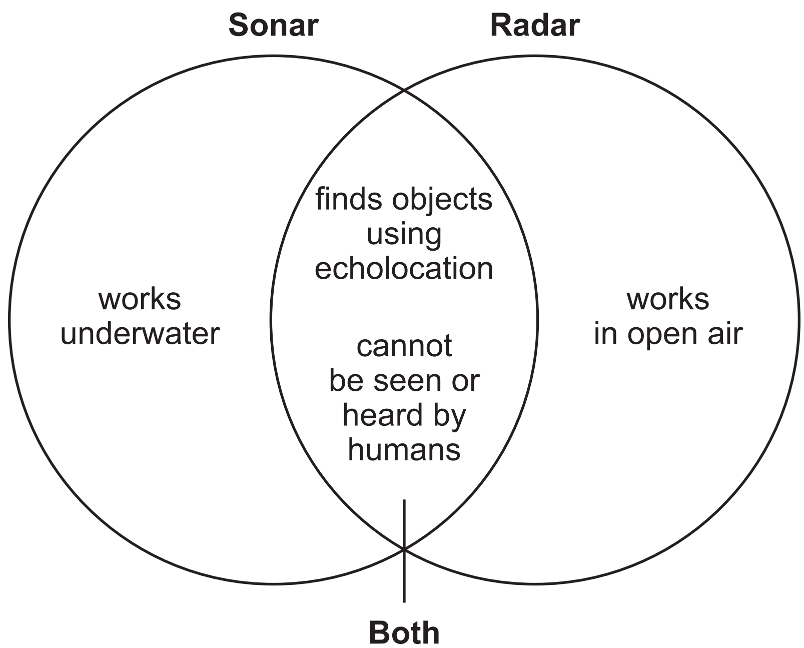
Ordered Item Booklet

for Standard Setting Training

These released items are adapted from the
Wisconsin Forward Exam item sampler for grade 4 science, available at
https://dpi.wi.gov/assessment/forward/sample-items

The Venn diagram below compares two applications of wave energy by humans.

1. sonar (sound navigation and ranging)
2. radar (radio detection and ranging)



A scientist wants to track the movement of songbirds as they migrate through Wisconsin.

Which explanation best describes the technology the scientist should use to track songbird migration?

A. The scientist should use radar to track songbird migration because it works in open air.

B. The scientist should use sonar to track songbird migration because it works in open air.

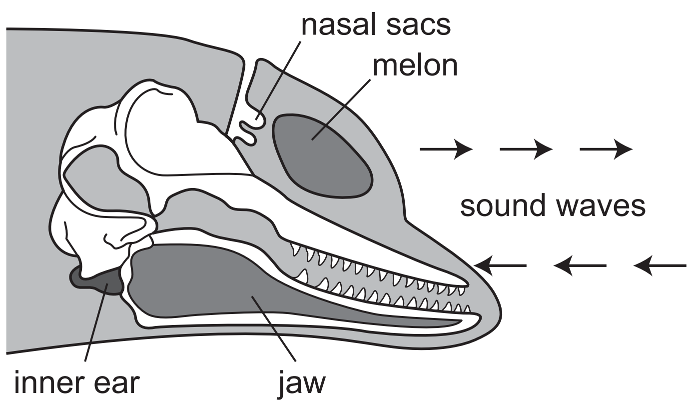
C. The scientist should use radar to track songbird migration because it works underwater.

D. The scientist should use sonar to track songbird migration because it works underwater.

Passage:
Seeing With Sounds Underwater

The diagram below shows how dolphins produce and receive sounds.

Dolphin Echolocation System



Dolphins make sounds by blowing air through their nasal sacs. These sounds travel into the water through the melon, an organ in the forehead. Returning sound waves are received through the jaw and then sent to the inner ear. In the inner ear, sound waves are translated into nerve impulses and sent to the brain.

Which model best shows how dolphins use incoming sound waves?

A. incoming sound waves → nerve impulses → jaw → inner ear → brain

B. incoming sound waves → jaw → inner ear → nerve impulses → brain

C. incoming sound waves → inner ear → nerve impulses → jaw → brain

D. incoming sound waves → nerve impulses → inner ear → jaw → brain

Passage:
Seeing With Sounds Underwater

Which words correctly complete the sentences to model how humans receive, understand, and react to information from a doorbell?

1. The person's _______ detect(s) the sound that the doorbell makes.
2. The information is interpreted in the person's _______ .
3. The person uses his or her _______ to move to the door.

A. 1. ears
2. brain
3. body

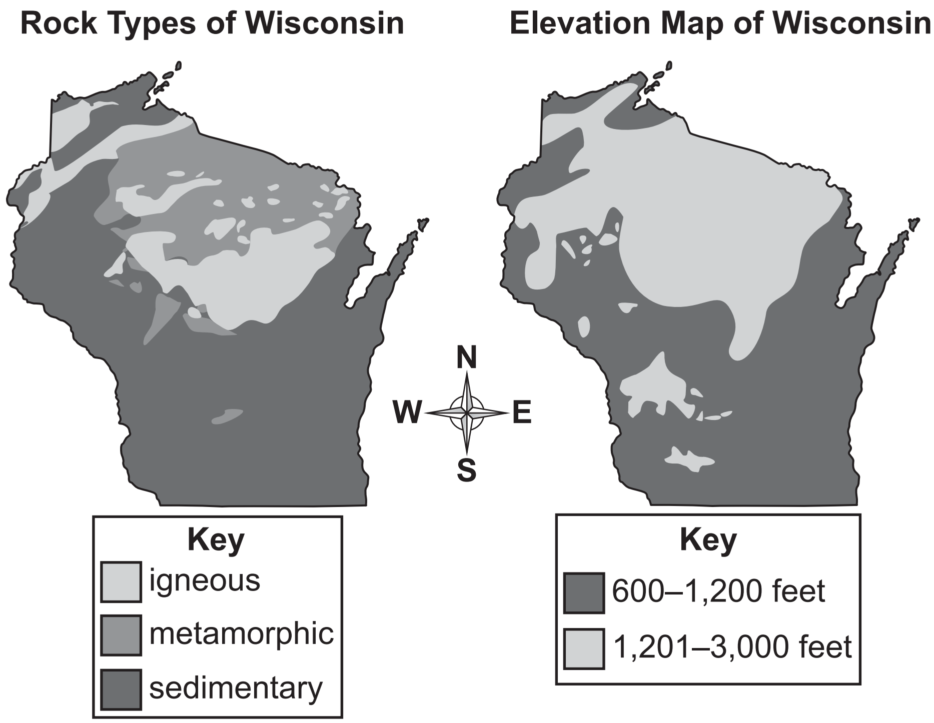
B. 1. ears
2. eyes
3. body

C. 1. body
2. brain
3. eyes

D. 1. brain
2. ears
3. eyes

4

Students are studying characteristic rocks in Wisconsin. The students study two maps comparing the types of rocks in Wisconsin and the elevation across the state. One student observes a pattern between the rock type and elevation.



Which chart shows the pattern the student observed?

A.

Elevation (feet)	Rock Types
600–1,200	metamorphic and sedimentary
1,201–3,000	igneous

B.

Elevation (feet)	Rock Types
600–1,200	sedimentary
1,201–3,000	igneous and metamorphic

C.

Elevation (feet)	Rock Types
600–1,200	igneous
1,201–3,000	metamorphic and sedimentary

D.

Elevation (feet)	Rock Types
600–1,200	igneous and metamorphic
1,201–3,000	sedimentary

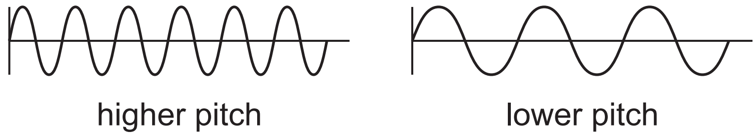
Passage:

Weathering Rocks

5

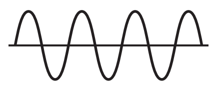
A student studies models of waves with different pitches.

Sounds Produced by a Dolphin

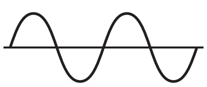


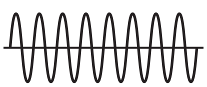
Echoing waves with a lower pitch than the original sound wave suggest the object is moving away from the source. A dolphin produces the sound wave shown below.

Dolphin Sound Wave

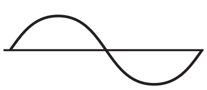


Which model best shows an echoing wave for an object moving toward the dolphin?

A. 

B. 

C. 

D. 

Passage:

Seeing With Sounds Underwater

The students decide to leave the rocks in the rock tumbler for a fourth day. One student claims that the mass of the sandstone sample on day 4 can be predicted since weathering conditions remained the same throughout the investigation.

What is the most likely mass of the sandstone sample on day 4 of the investigation?

A. 120 grams

B. 60 grams

C. 30 grams

D. 10 grams

Passage:

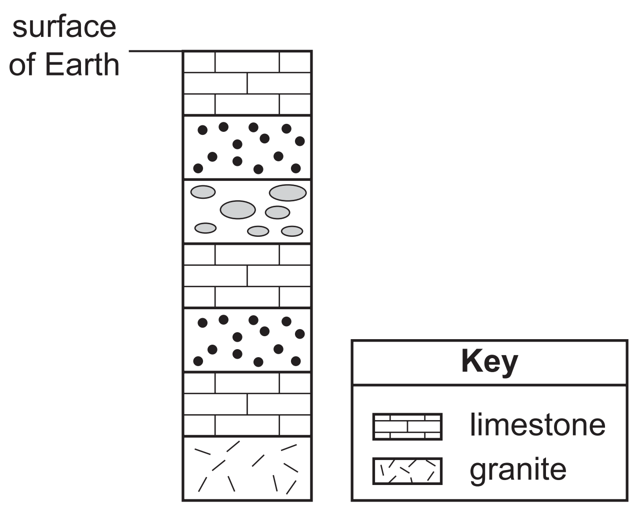
Weathering Rocks

A student is looking at the diagram below, which uses symbols to represent layers of rock in an area. Each type of rock has a unique symbol.

Limestone, a type of rock that is usually deposited when a surface is covered by an ocean, is shown by a brick-like pattern.

Granite, a rock that is commonly found in areas that have had volcanic eruptions, is shown by a pattern that looks like very short lines arranged in different directions.

Rock Layer Symbols



Which statements provide the best description of this area?

A. The deepest and oldest layer shows limestone, indicating that the land was formed by a volcano. The area was repeatedly affected by an ocean, as shown by the layers of granite.

B. The deepest and oldest layer shows limestone, indicating that the land was formed by an ocean. The area was repeatedly affected by a volcano, as shown by the layers of limestone.

C. The deepest and oldest layer shows granite, indicating that the land was formed by an ocean. The area was repeatedly affected by a volcano, as shown by the layers of granite.

D. The deepest and oldest layer shows granite, indicating that the land was formed by a volcano. The area was repeatedly affected by an ocean, as shown by the layers of limestone.

Engineers are exploring locations to construct a new building. They study a chart showing factors that affect rates of weathering.

Factors That Affect Rates of Weathering

Factor	weathering rate fast ⟶ slow		
precipitation	high	medium	low
thickness of soil layer	thin	medium	thick
hills	steep	medium	gentle

Next, the engineers study a chart showing characteristics of four locations in Wisconsin.

Location	**Average Yearly Precipitation (inches)**	**Thickness of Soil Layer**	**Hills**
1	31–32	thick	gentle
2	37–38	thin	gentle
3	32–33	medium	steep
4	34–35	thin	medium

Which location most likely has the slowest rate of rock weathering?

A. location 1

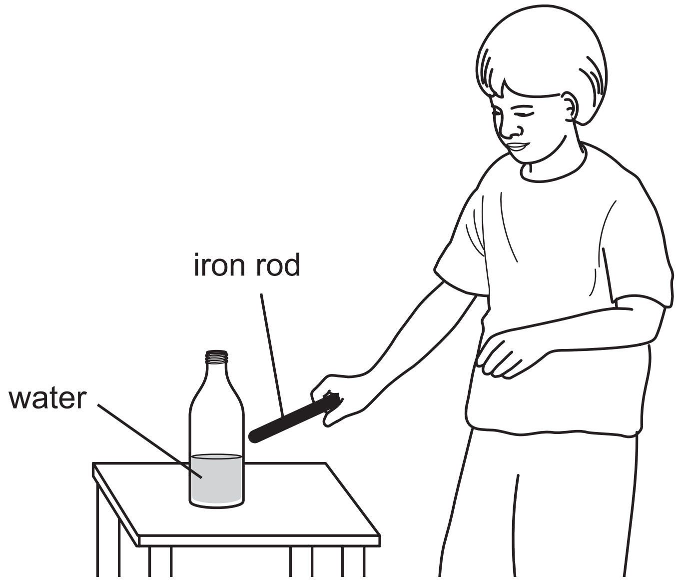
B. location 2

C. location 3

D. location 4

Passage:
Weathering Rocks

A student pours water into a glass bottle. Next, the student gently taps the outside of the bottle with an iron rod.



Part A

Which idea is most likely being investigated by the student?

A. energy transfer through materials

B. heat conduction through materials

C. magnetic properties of materials

D. reflective properties of materials

Part B

Which observations best support the answer to Part A?

A. The temperature of the water remains the same after the student taps the bottle with the iron rod.
The iron rod and the glass bottle are made of different materials.

B. A sound is produced when the student taps the bottle with the iron rod.
The temperature of the water remains the same after the student taps the bottle with the iron rod.

C. A sound is produced when the student taps the bottle with the iron rod.
Waves are produced in the water, showing vibration.

D. Waves are produced in the water, showing vibration.
The iron rod and the glass bottle are made of different materials.

Passage:

Seeing With Sounds Underwater

Dolphins cannot detect fishing nets using echolocation. Sometimes dolphins get caught in these nets. A student listed two possible solutions to improve the design of the nets.

Possible Design Solutions
Solution 1: increase the size of the openings in the net so dolphins can swim out
Solution 2: attach a device to the net that reflects echolocation sounds from dolphins

The goals for the new nets are listed below.

Goals for the Nets

Goal A: prevent dolphins from getting trapped

Goal B: help dolphins to locate fishing nets

Which table best identifies the goal(s) that each solution meets?

A.

	Goal A	Goal B
Solution 1		×
Solution 2	×	

B.

	Goal A	Goal B
Solution 1	×	×
Solution 2		×

C.

	Goal A	Goal B
Solution 1	×	
Solution 2	×	×

D.

	Goal A	Goal B
Solution 1	×	
Solution 2		×

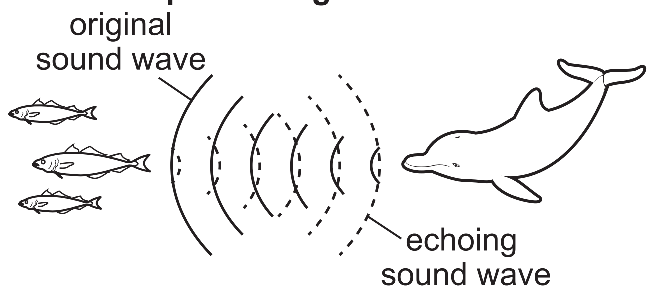
Passage:

Seeing With Sounds Underwater

Seeing with Sounds Underwater

It is difficult to see underwater where it is dark. Some ocean animals, such as dolphins, use sounds to help them understand their water environment. Dolphins produce high-pitched sounds. These sound waves travel through the water until they bump into an object. Then, the sound waves bounce off the object. The echoes of the sound waves return to the dolphin—specifically to the dolphin's jaw. The sounds travel through the dolphin's jaw to its inner ear, where the sounds are translated into nerve impulses that travel to the brain. The way dolphins "see" with sounds is called echolocation.

Dolphin Using Echolocation



From echoing sound waves, dolphins can learn a lot about an object: its shape, its size, its distance from the dolphin, and whether it is moving toward or away from the dolphin. Dolphins use echolocation to find their way around, to find prey, and to communicate with each other.

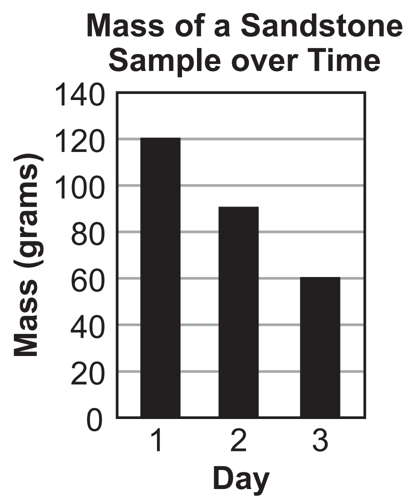
Scientists realized that the way dolphins can gather information from sounds could be used for human-made technology. Sonar is one example of this type of technology. Sonar is used by submarines and ships to find their way underwater and locate objects. Similar to echolocation, sonar sends out sound waves and interprets the echoing waves.

Weathering Rocks

Students learn that igneous, sedimentary, and metamorphic rocks can be observed in Wisconsin. These rocks weather at different rates.

The students study data from an experiment that used a rock tumbler, which is a machine used to weather rocks. A rock tumbler spins and tumbles rocks similar to how a washing machine spins and tumbles clothes. A rock tumbler is filled with sand and water to help weather the rocks inside. The data from the experiment suggest that certain igneous rocks weather at a slower rate than some sedimentary and metamorphic rocks exposed to the same conditions.

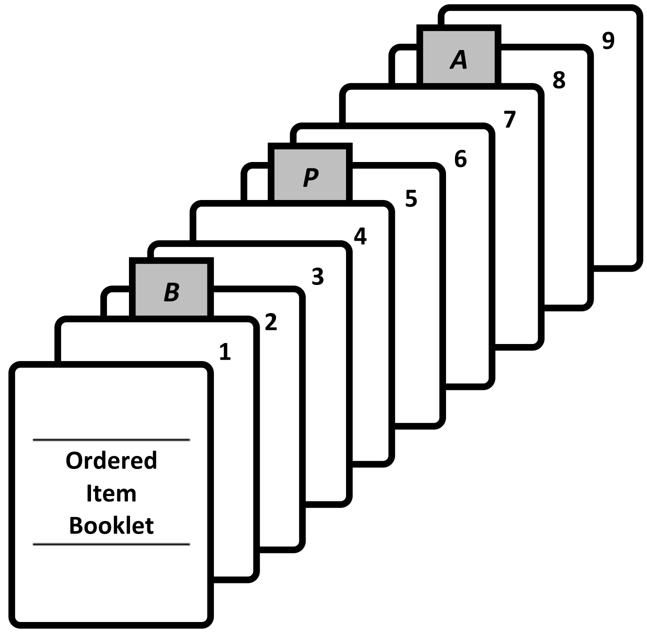
The students decided to conduct a similar experiment with one type of rock in a rock tumbler half-filled with sand and water. The students selected a rock sample of sandstone, which is a sedimentary rock. Over three days, the students measured the mass of the sandstone. The students recorded the data in the graph below.



Mid-Process Evaluation

Suppose the bookmarks were placed in this sample ordered item booklet (OIB) as follows:

Cut Score	Bookmark on Page
Basic	2
Proficient	5
Advanced	8



1. Of which items does a student need to have command to just make it into the *Proficient* performance level?

   ○ 1 to 4 ○ 1 to 5 ○ 1 to 7

2. If a student has command of the content in ***only*** Items 1 through 3 (and nothing else), in which performance level would this student be?

   ○ *Basic* ○ *Proficient* ○ *Advanced*

3. Suppose a student has command of the content in Items 1 through 7. In which performance level is this student?

   ○ *Basic* ○ *Proficient* ○ *Advanced*

4. For the *Proficient* threshold student, will the items ***before*** the *Proficient* bookmark be easier, about the same, or harder to answer correctly than the items ***after*** the bookmark?

   ○ Easier to answer ○ About the same ○ Harder to answer

5. What does an *Advanced* bookmark placed on Page 8 represent?

   ○ Students must have command of the content measured by the items on Pages 1–7 to be in the *Advanced* level.

   ○ Students must answer ***all*** of the items before Page 8 correctly to be in the *Advanced* level.

   ○ Students must have command of the content measured by the items on Pages 8–9 to be in the *Advanced* level.

6. Are you ready to proceed?

   ○ Yes, I am ready. ○ Not yet; I have questions that I have written on the back of this form.

6: If you are not ready to proceed, please write your questions below.

Legend: Correct: Incorrect: Distractors Chosen More than Correct Answer: Changed Answer:

1. Of which items does a student need to have command to just make it into the Proficient performance level?

Response	Frequency	Percent
***1 to 4**	26	96.30
1 to 5	1	3.70
1 to 7	0	0.00

2. If a student has command of the content in only Items 1 through 3 (and nothing else), in which performance level would this student be?

Response	Frequency	Percent
***Basic**	24	88.89
Proficient	3	11.11
Advanced	0	0.00

3. Suppose a student has command of the content in Items 1 through 7. In which performance level is this student?

Response	Frequency	Percent
Basic	0	0.00
Proficient	12	44.44
***Advanced**	15	55.56

4. For the Proficient threshold student, will the items before the Proficient bookmark be easier, about the same, or harder to answer correctly than the items after the bookmark?

Response	Frequency	Percent
***Easier to answer**	27	100.00
About the same	0	0.00
Harder to answer	0	0.00

5. What does an Advanced bookmark placed on Page 8 represent?

Response	Frequency	Percent
***Students must have command of the content measured by the items on Pages 1-7 to be in the Advanced level.**	24	88.89
Students must answer all of the items before Page 8 correctly to be in the Advanced level.	0	0.00
Students must have command of the content measured by the items on Pages 8-9 to be in the Advanced level.	3	11.11

6. Are you ready to proceed?

Response	Frequency	Percent
***Yes I am ready**	26	96.30
Not yet	0	0.00
No Response	1	3.70

Appendix C

Detailed Reports of Participants' Judgments

Wisconsin Grade 4 Science
Round 1 Bookmark Placements

Table	Participant	Basic	Proficient	Advanced
1	1	12	32	44
1	2	5	35	44
1	3	5	20	37
1	4	9	31	51
1	5	11	32	59
2	6	8	35	50
2	7	8	31	51
2	8	25	38	51
2	9	7	31	51
3	10	24	44	51
3	11	10	32	51
3	12	17	27	53
3	13	8	29	38

Overall	Median	9	32	51
	25th %ile	7.5	30	44
	75th %ile	14.5	35	51
	Minimum	5	20	37
	Maximum	25	44	59

Wisconsin Grade 4 Science
Round 1 Cut Scores

Table	Participant	Basic	Proficient	Advanced
1	1	450	496	520
1	2	436	501	520
1	3	436	475	506
1	4	447	495	543
1	5	448	496	570
2	6	447	501	542
2	7	447	495	543
2	8	488	509	543
2	9	439	495	543
3	10	487	520	543
3	11	448	496	543
3	12	466	492	548
3	13	447	494	509

Overall	Median	447	496	543
	25th %ile	443	495	520
	75th %ile	458	501	543
	Minimum	436	475	506
	Maximum	488	520	570

Wisconsin Grade 4 Science
Round 1 Summary of Bookmark Placements

Statistic	Table	Basic	Proficient	Advanced
Median	1	9	32	44
Median	2	8	33	51
Median	3	13.5	30.5	51
Median	Overall	9	32	51
25th %ile	1	5	25.5	40.5
25th %ile	2	7.25	31	50.25
25th %ile	3	8.5	27.5	41.25
25th %ile	Overall	7.5	30	44
75th %ile	1	11.5	33.5	55
75th %ile	2	20.75	37.25	51
75th %ile	3	22.25	41	52.5
75th %ile	Overall	14.5	35	51
Minimum	1	5	20	37
Minimum	2	7	31	50
Minimum	3	8	27	38
Minimum	Overall	5	20	37
Maximum	1	12	35	59
Maximum	2	25	38	51
Maximum	3	24	44	53
Maximum	Overall	25	44	59

Overall	Median	9	32	51
	25th %ile	7.5	30	44
	75th %ile	14.5	35	51
	Minimum	5	20	37
	Maximum	25	44	59

Wisconsin Grade 4 Science
Round 1 Summary of Cut Scores

Statistic	Table	Basic	Proficient	Advanced
Median	1	447	496	520
Median	2	447	498	543
Median	3	457	495	543
Median	Overall	447	496	543
25th %ile	1	436	485	513
25th %ile	2	441	495	542
25th %ile	3	447	493	518
25th %ile	Overall	443	495	520
75th %ile	1	449	498	556
75th %ile	2	478	507	543
75th %ile	3	482	514	547
75th %ile	Overall	458	501	543
Minimum	1	436	475	506
Minimum	2	439	495	542
Minimum	3	447	492	509
Minimum	Overall	436	475	506
Maximum	1	450	501	570
Maximum	2	488	509	543
Maximum	3	487	520	548
Maximum	Overall	488	520	570

Overall	Median	447	496	543
	25th %ile	443	495	520
	75th %ile	458	501	543
	Minimum	436	475	506
	Maximum	488	520	570

Wisconsin Grade 4 Science
Round 1 Median Bookmark Summary

Table	Basic	Proficient	Advanced
1	9	32	44
2	8	33	51
3	13.5	30.5	51
Overall	9	32	51

Impact Data

	Below Basic	Basic	Proficient	Advanced
Overall	15.0	32.2	33.3	19.5

Wisconsin Grade 4 Science
Round 2 Bookmark Placements

Table	Participant	Basic	Proficient	Advanced
1	1	10	32	44
1	2	9	32	45
1	3	9	32	52
1	4	9	31	44
1	5	10	32	49
2	6	8	31	50
2	7	7	31	51
2	8	9	33	51
2	9	7	31	51
3	10	8	20	44
3	11	9	32	51
3	12	9	27	47
3	13	9	32	41

Overall	Median	9	32	49
	25th %ile	8	31	44
	75th %ile	9	32	51
	Minimum	7	20	41
	Maximum	10	33	52

Wisconsin Grade 4 Science
Round 2 Cut Scores

Table	Participant	Basic	Proficient	Advanced
1	1	448	496	520
1	2	447	496	520
1	3	447	496	544
1	4	447	495	520
1	5	448	496	538
2	6	447	495	542
2	7	439	495	543
2	8	447	497	543
2	9	439	495	543
3	10	447	475	520
3	11	447	496	543
3	12	447	492	525
3	13	447	496	515

Overall	Median	447	496	538
	25th %ile	447	495	520
	75th %ile	447	496	543
	Minimum	439	475	515
	Maximum	448	497	544

Wisconsin Grade 4 Science
Round 2 Summary of Bookmark Placements

Statistic	Table	Basic	Proficient	Advanced
Median	1	9	32	45
Median	2	7.5	31	51
Median	3	9	29.5	45.5
Median	Overall	9	32	49
25th %ile	1	9	31.5	44
25th %ile	2	7	31	50.25
25th %ile	3	8.25	21.75	41.75
25th %ile	Overall	8	31	44
75th %ile	1	10	32	50.5
75th %ile	2	8.75	32.5	51
75th %ile	3	9	32	50
75th %ile	Overall	9	32	51
Minimum	1	9	31	44
Minimum	2	7	31	50
Minimum	3	8	20	41
Minimum	Overall	7	20	41
Maximum	1	10	32	52
Maximum	2	9	33	51
Maximum	3	9	32	51
Maximum	Overall	10	33	52

Overall	Median	9	32	49
	25th %ile	8	31	44
	75th %ile	9	32	51
	Minimum	7	20	41
	Maximum	10	33	52

Wisconsin Grade 4 Science
Round 2 Summary of Cut Scores

Statistic	Table	Basic	Proficient	Advanced
Median	1	447	496	520
Median	2	443	495	543
Median	3	447	494	523
Median	Overall	447	496	538
25th %ile	1	447	495	520
25th %ile	2	439	495	542
25th %ile	3	447	479	516
25th %ile	Overall	447	495	520
75th %ile	1	448	496	541
75th %ile	2	447	497	543
75th %ile	3	447	496	538
75th %ile	Overall	447	496	543
Minimum	1	447	495	520
Minimum	2	439	495	542
Minimum	3	447	475	515
Minimum	Overall	439	475	515
Maximum	1	448	496	544
Maximum	2	447	497	543
Maximum	3	447	496	543
Maximum	Overall	448	497	544

Overall	Median	447	496	538
	25th %ile	447	495	520
	75th %ile	447	496	543
	Minimum	439	475	515
	Maximum	448	497	544

Wisconsin Grade 4 Science
Round 2 Median Bookmark Summary

Table	Basic	Proficient	Advanced
1	9	32	45
2	7.5	31	51
3	9	29.5	45.5
Overall	9	32	49

Impact Data

	Below Basic	Basic	Proficient	Advanced
Overall	15.0	32.2	30.4	22.4

Wisconsin Grade 4 Science
Round 3 Bookmark Placements

Table	Participant	Basic	Proficient	Advanced
1	1	10	32	47
1	2	9	32	50
1	3	10	34	52
1	4	9	32	51
1	5	10	32	56
2	6	9	31	47
2	7	7	31	51
2	8	9	32	51
2	9	9	31	51
3	10	9	32	51
3	11	9	32	53
3	12	9	27	47
3	13	10	32	45

Overall	Median	9	32	51
	25th %ile	9	31	47
	75th %ile	10	32	51.5
	Minimum	7	27	45
	Maximum	10	34	56

Wisconsin Grade 4 Science
Round 3 Cut Scores

Table	Participant	Basic	Proficient	Advanced
1	1	448	496	525
1	2	447	496	542
1	3	448	499	544
1	4	447	496	543
1	5	448	496	553
2	6	447	495	525
2	7	439	495	543
2	8	447	496	543
2	9	447	495	543
3	10	447	496	543
3	11	447	496	548
3	12	447	492	525
3	13	448	496	520

Overall	Median	447	496	543
	25th %ile	447	495	525
	75th %ile	448	496	544
	Minimum	439	492	520
	Maximum	448	499	553

Wisconsin Grade 4 Science
Round 3 Summary of Bookmark Placements

Statistic	Table	Basic	Proficient	Advanced
Median	1	10	32	51
Median	2	9	31	51
Median	3	9	32	49
Median	Overall	9	32	51
25th %ile	1	9	32	48.5
25th %ile	2	7.5	31	48
25th %ile	3	9	28.25	45.5
25th %ile	Overall	9	31	47
75th %ile	1	10	33	54
75th %ile	2	9	31.75	51
75th %ile	3	9.75	32	52.5
75th %ile	Overall	10	32	51.5
Minimum	1	9	32	47
Minimum	2	7	31	47
Minimum	3	9	27	45
Minimum	Overall	7	27	45
Maximum	1	10	34	56
Maximum	2	9	32	51
Maximum	3	10	32	53
Maximum	Overall	10	34	56

Overall	Median	9	32	51
	25th %ile	9	31	47
	75th %ile	10	32	51.5
	Minimum	7	27	45
	Maximum	10	34	56

Wisconsin Grade 4 Science
Round 3 Summary of Cut Scores

Statistic	Table	Basic	Proficient	Advanced
Median	1	448	496	543
Median	2	447	495	543
Median	3	447	496	534
Median	Overall	447	496	543
25th %ile	1	447	496	534
25th %ile	2	441	495	530
25th %ile	3	447	493	521
25th %ile	Overall	447	495	525
75th %ile	1	448	497	549
75th %ile	2	447	495	543
75th %ile	3	448	496	547
75th %ile	Overall	448	496	544
Minimum	1	447	496	525
Minimum	2	439	495	525
Minimum	3	447	492	520
Minimum	Overall	439	492	520
Maximum	1	448	499	553
Maximum	2	447	496	543
Maximum	3	448	496	548
Maximum	Overall	448	499	553

Overall	Median	447	496	543
	25th %ile	447	495	525
	75th %ile	448	496	544
	Minimum	439	492	520
	Maximum	448	499	553

Wisconsin Grade 4 Science
Round 3 Median Bookmark Summary

Table	Basic	Proficient	Advanced
1	10	32	51
2	9	31	51
3	9	32	49
Overall	9	32	51

Impact Data

	Below Basic	Basic	Proficient	Advanced
Overall	15.0	32.2	33.3	19.5

Wisconsin Grade 8 Science
Round 1 Bookmark Placements

Table	Participant	Basic	Proficient	Advanced
1	1	12	23	40
1	2	11	23	39
1	3	22	32	54
1	4	9	26	44
2	5	13	36	52
2	6	19	36	58
2	7	14	43	56
2	8	17	43	50
2	9	16	42	57
3	10	19	37	46
3	11	9	19	47
3	12	10	24	47
3	13	9	20	47
3	14	2	22	45

Overall	Median	12.5	29	47
	25th %ile	9	22.75	44.75
	75th %ile	17.5	38.25	54.5
	Minimum	2	19	39
	Maximum	22	43	58

Wisconsin Grade 8 Science
Round 1 Cut Scores

Table	Participant	Basic	Proficient	Advanced
1	1	654	683	707
1	2	652	683	706
1	3	682	693	731
1	4	650	688	716
2	5	655	701	727
2	6	673	701	747
2	7	662	716	740
2	8	671	716	723
2	9	669	714	746
3	10	673	702	719
3	11	650	673	719
3	12	651	684	719
3	13	650	681	719
3	14	620	682	718

Overall	Median	654	690	719
	25th %ile	650	682	717
	75th %ile	671	705	733
	Minimum	620	673	706
	Maximum	682	716	747

Wisconsin Grade 8 Science
Round 1 Summary of Bookmark Placements

Statistic	Table	Basic	Proficient	Advanced
Median	1	11.5	24.5	42
Median	2	16	42	56
Median	3	9	22	47
Median	Overall	12.5	29	47
25th %ile	1	9.5	23	39.25
25th %ile	2	13.5	36	51
25th %ile	3	5.5	19.5	45.5
25th %ile	Overall	9	22.75	44.75
75th %ile	1	19.5	30.5	51.5
75th %ile	2	18	43	57.5
75th %ile	3	14.5	30.5	47
75th %ile	Overall	17.5	38.25	54.5
Minimum	1	9	23	39
Minimum	2	13	36	50
Minimum	3	2	19	45
Minimum	Overall	2	19	39
Maximum	1	22	32	54
Maximum	2	19	43	58
Maximum	3	19	37	47
Maximum	Overall	22	43	58

Overall	Median	12.5	29	47
	25th %ile	9	22.75	44.75
	75th %ile	17.5	38.25	54.5
	Minimum	2	19	39
	Maximum	22	43	58

Wisconsin Grade 8 Science
Round 1 Summary of Cut Scores

Statistic	Table	Basic	Proficient	Advanced
Median	1	653	685	712
Median	2	669	714	740
Median	3	650	682	719
Median	Overall	654	690	719
25th %ile	1	650	683	706
25th %ile	2	658	701	725
25th %ile	3	635	677	718
25th %ile	Overall	650	682	717
75th %ile	1	675	691	727
75th %ile	2	672	716	747
75th %ile	3	662	693	719
75th %ile	Overall	671	705	733
Minimum	1	650	683	706
Minimum	2	655	701	723
Minimum	3	620	673	718
Minimum	Overall	620	673	706
Maximum	1	682	693	731
Maximum	2	673	716	747
Maximum	3	673	702	719
Maximum	Overall	682	716	747

Overall	Median	654	690	719
	25th %ile	650	682	717
	75th %ile	671	705	733
	Minimum	620	673	706
	Maximum	682	716	747

Wisconsin Grade 8 Science
Round 1 Median Bookmark Summary

Table	Basic	Proficient	Advanced
1	11.5	24.5	42
2	16	42	56
3	9	22	47
Overall	12.5	29	47

Impact Data

	Below Basic	Basic	Proficient	Advanced
Overall	18.2	23.9	23.2	34.8

Wisconsin Grade 8 Science
Round 2 Bookmark Placements

Table	Participant	Basic	Proficient	Advanced
1	1	12	22	44
1	2	10	22	42
1	3	12	22	47
1	4	10	22	48
2	5	12	18	39
2	6	12	19	41
2	7	12	19	41
2	8	12	19	40
2	9	12	19	42
3	10	9	19	47
3	11	9	19	47
3	12	10	20	47
3	13	9	20	47
3	14	9	19	47

Overall	Median	11	19	45.5
	25th %ile	9	19	41
	75th %ile	12	22	47
	Minimum	9	18	39
	Maximum	12	22	48

Wisconsin Grade 8 Science
Round 2 Cut Scores

Table	Participant	Basic	Proficient	Advanced
1	1	654	682	716
1	2	651	682	714
1	3	654	682	719
1	4	651	682	721
2	5	654	672	706
2	6	654	673	710
2	7	654	673	710
2	8	654	673	707
2	9	654	673	714
3	10	650	673	719
3	11	650	673	719
3	12	651	681	719
3	13	650	681	719
3	14	650	673	719

Overall	Median	653	673	718
	25th %ile	650	673	710
	75th %ile	654	682	719
	Minimum	650	672	706
	Maximum	654	682	721

Wisconsin Grade 8 Science
Round 2 Summary of Bookmark Placements

Statistic	Table	Basic	Proficient	Advanced
Median	1	11	22	45.5
Median	2	12	19	41
Median	3	9	19	47
Median	Overall	11	19	45.5
25th %ile	1	10	22	42.5
25th %ile	2	12	18.5	39.5
25th %ile	3	9	19	47
25th %ile	Overall	9	19	41
75th %ile	1	12	22	47.75
75th %ile	2	12	19	41.5
75th %ile	3	9.5	20	47
75th %ile	Overall	12	22	47
Minimum	1	10	22	42
Minimum	2	12	18	39
Minimum	3	9	19	47
Minimum	Overall	9	18	39
Maximum	1	12	22	48
Maximum	2	12	19	42
Maximum	3	10	20	47
Maximum	Overall	12	22	48

Overall	Median	11	19	45.5
	25th %ile	9	19	41
	75th %ile	12	22	47
	Minimum	9	18	39
	Maximum	12	22	48

Wisconsin Grade 8 Science
Round 2 Summary of Cut Scores

Statistic	Table	Basic	Proficient	Advanced
Median	1	653	682	718
Median	2	654	673	710
Median	3	650	673	719
Median	Overall	653	673	718
25th %ile	1	651	682	715
25th %ile	2	654	673	707
25th %ile	3	650	673	719
25th %ile	Overall	650	673	710
75th %ile	1	654	682	720
75th %ile	2	654	673	712
75th %ile	3	650	681	719
75th %ile	Overall	654	682	719
Minimum	1	651	682	714
Minimum	2	654	672	706
Minimum	3	650	673	719
Minimum	Overall	650	672	706
Maximum	1	654	682	721
Maximum	2	654	673	714
Maximum	3	651	681	719
Maximum	Overall	654	682	721

Overall	Median	653	673	718
	25th %ile	650	673	710
	75th %ile	654	682	719
	Minimum	650	672	706
	Maximum	654	682	721

Wisconsin Grade 8 Science
Round 2 Median Bookmark Summary

Table	Basic	Proficient	Advanced
1	11	22	45.5
2	12	19	41
3	9	19	47
Overall	11	19	45.5

Impact Data

	Below Basic	Basic	Proficient	Advanced
Overall	17.7	12.0	34.8	35.5

Wisconsin Grade 8 Science
Round 3 Bookmark Placements

Table	Participant	Basic	Proficient	Advanced
1	1	12	26	50
1	2	10	22	47
1	3	12	22	47
1	4	10	26	54
2	5	12	19	52
2	6	12	22	48
2	7	12	22	39
2	8	12	35	50
2	9	12	27	50
3	10	9	19	47
3	11	9	22	54
3	12	10	22	50
3	13	9	22	50
3	14	9	22	50

Overall	Median	11	22	50
	25th %ile	9	22	47
	75th %ile	12	26	50.5
	Minimum	9	19	39
	Maximum	12	35	54

Wisconsin Grade 8 Science
Round 3 Cut Scores

Table	Participant	Basic	Proficient	Advanced
1	1	654	688	723
1	2	651	682	719
1	3	654	682	719
1	4	651	688	731
2	5	654	673	727
2	6	654	682	721
2	7	654	682	706
2	8	654	698	723
2	9	654	688	723
3	10	650	673	719
3	11	650	682	731
3	12	651	682	723
3	13	650	682	723
3	14	650	682	723

Overall	Median	653	682	723
	25th %ile	650	682	719
	75th %ile	654	688	724
	Minimum	650	673	706
	Maximum	654	698	731

Wisconsin Grade 8 Science
Round 3 Summary of Bookmark Placements

Statistic	Table	Basic	Proficient	Advanced
Median	1	11	24	48.5
Median	2	12	22	50
Median	3	9	22	50
Median	Overall	11	22	50
25th %ile	1	10	22	47
25th %ile	2	12	20.5	43.5
25th %ile	3	9	20.5	48.5
25th %ile	Overall	9	22	47
75th %ile	1	12	26	53
75th %ile	2	12	31	51
75th %ile	3	9.5	22	52
75th %ile	Overall	12	26	50.5
Minimum	1	10	22	47
Minimum	2	12	19	39
Minimum	3	9	19	47
Minimum	Overall	9	19	39
Maximum	1	12	26	54
Maximum	2	12	35	52
Maximum	3	10	22	54
Maximum	Overall	12	35	54

Overall	Median	11	22	50
	25th %ile	9	22	47
	75th %ile	12	26	50.5
	Minimum	9	19	39
	Maximum	12	35	54

Wisconsin Grade 8 Science
Round 3 Summary of Cut Scores

Statistic	Table	Basic	Proficient	Advanced
Median	1	653	685	721
Median	2	654	682	723
Median	3	650	682	723
Median	Overall	653	682	723
25th %ile	1	651	682	719
25th %ile	2	654	677	713
25th %ile	3	650	677	721
25th %ile	Overall	650	682	719
75th %ile	1	654	688	729
75th %ile	2	654	693	725
75th %ile	3	650	682	727
75th %ile	Overall	654	688	724
Minimum	1	651	682	719
Minimum	2	654	673	706
Minimum	3	650	673	719
Minimum	Overall	650	673	706
Maximum	1	654	688	731
Maximum	2	654	698	727
Maximum	3	651	682	731
Maximum	Overall	654	698	731

Overall	Median	653	682	723
	25th %ile	650	682	719
	75th %ile	654	688	724
	Minimum	650	673	706
	Maximum	654	698	731

Wisconsin Grade 8 Science
Round 3 Median Bookmark Summary

Table	Basic	Proficient	Advanced
1	11	24	48.5
2	12	22	50
3	9	22	50
Overall	11	22	50

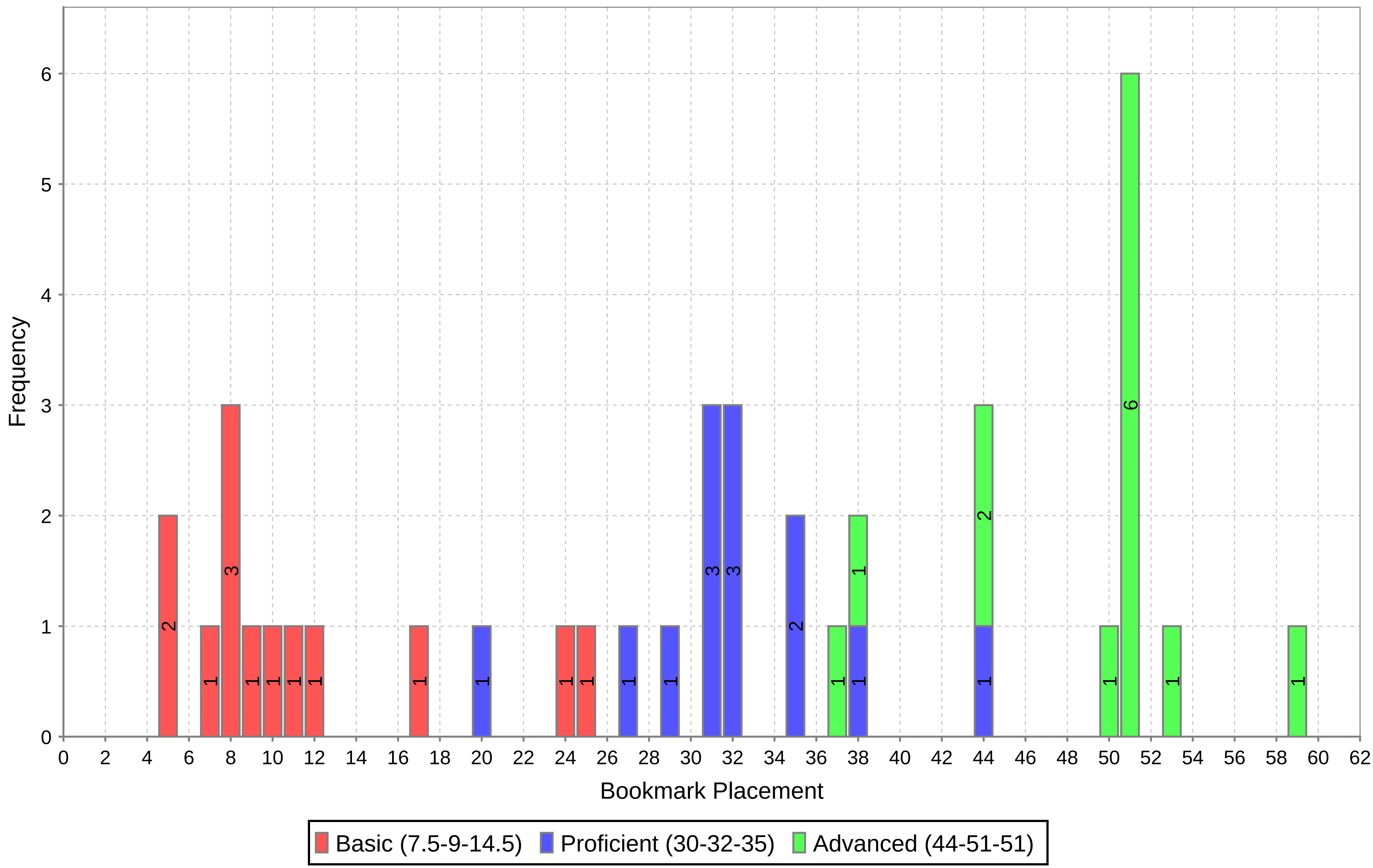
Impact Data

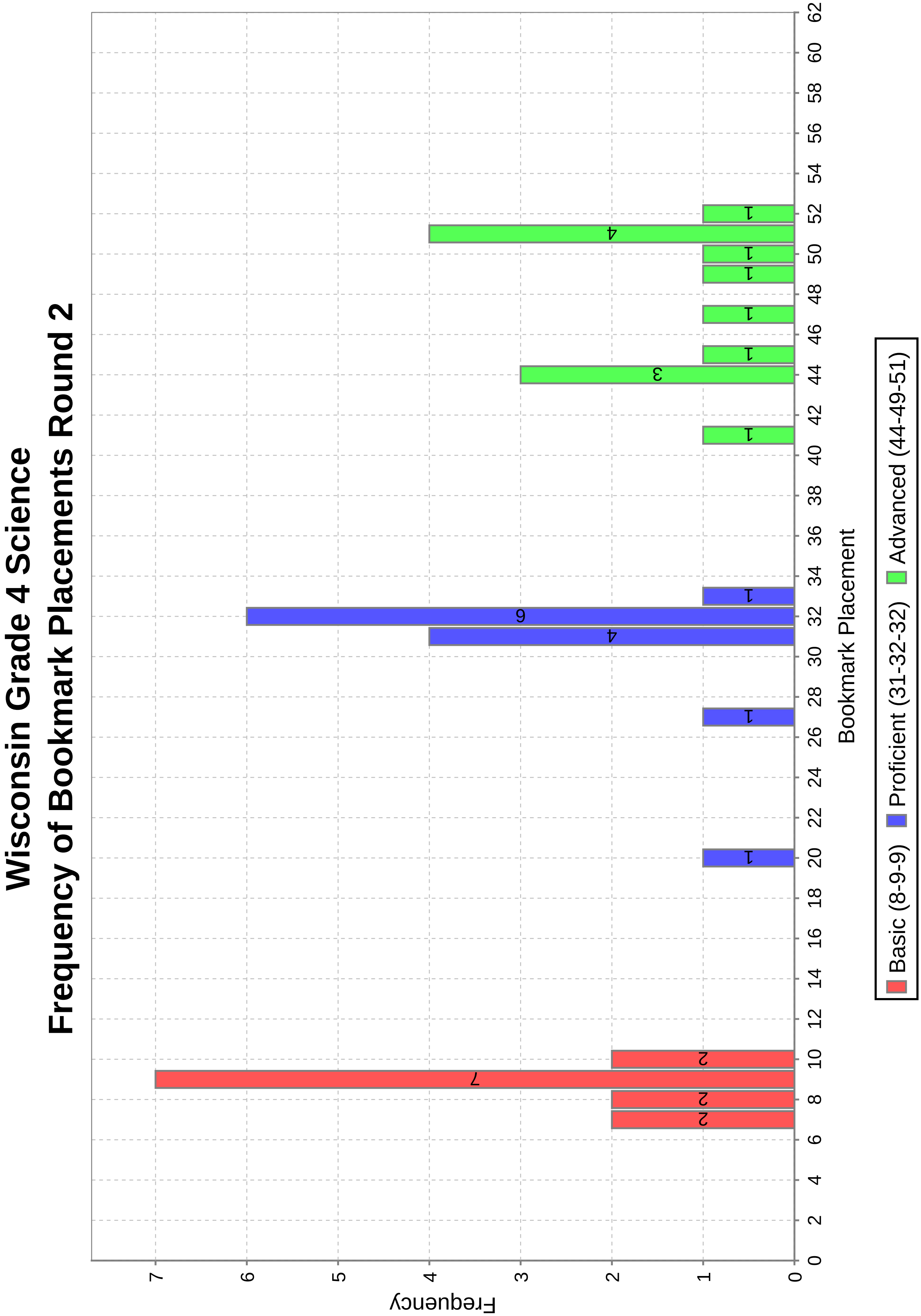
	Below Basic	Basic	Proficient	Advanced
Overall	17.7	18.2	32.2	31.8

Appendix D

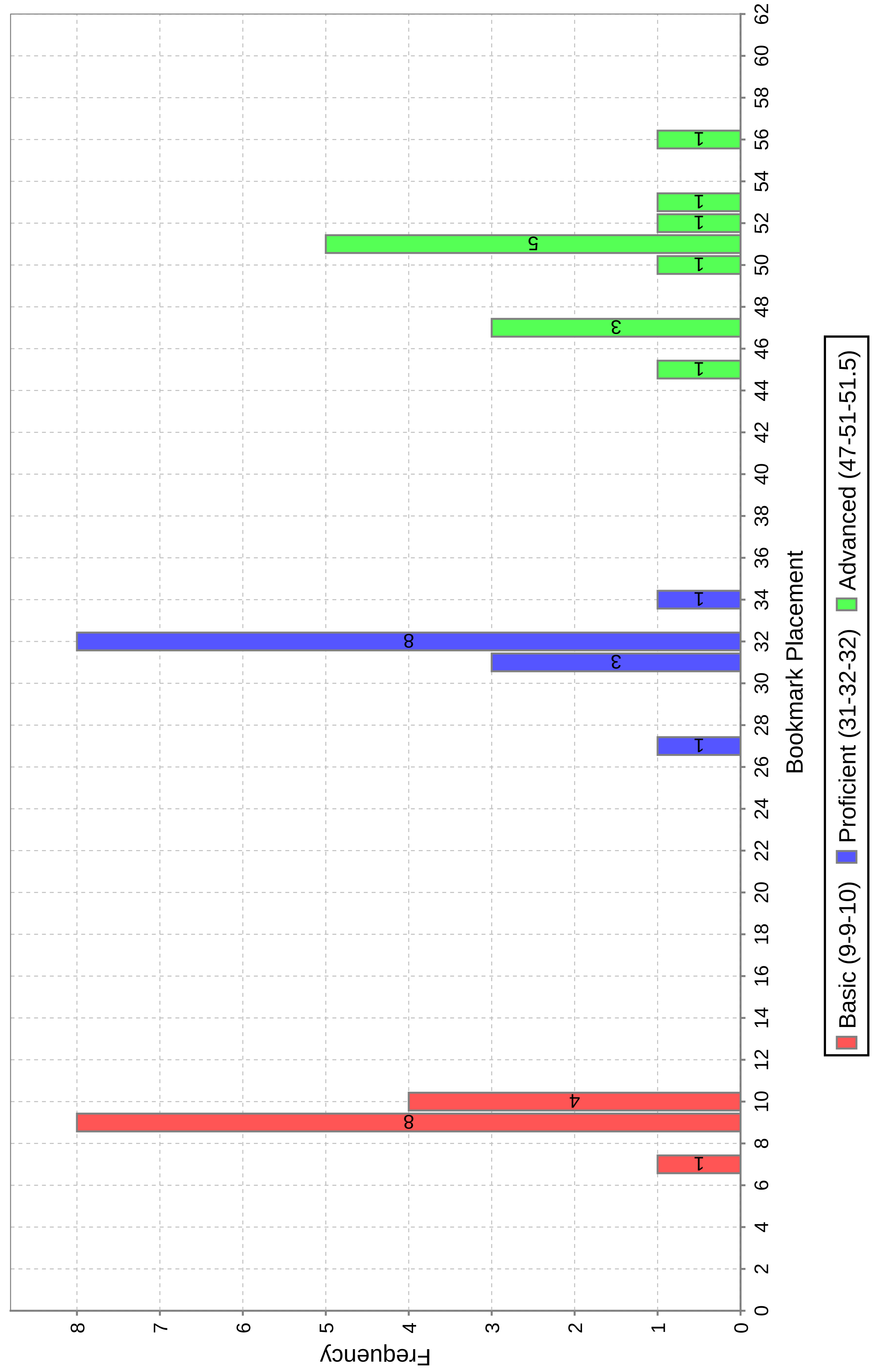
Graphical Representation of Participants' Judgments

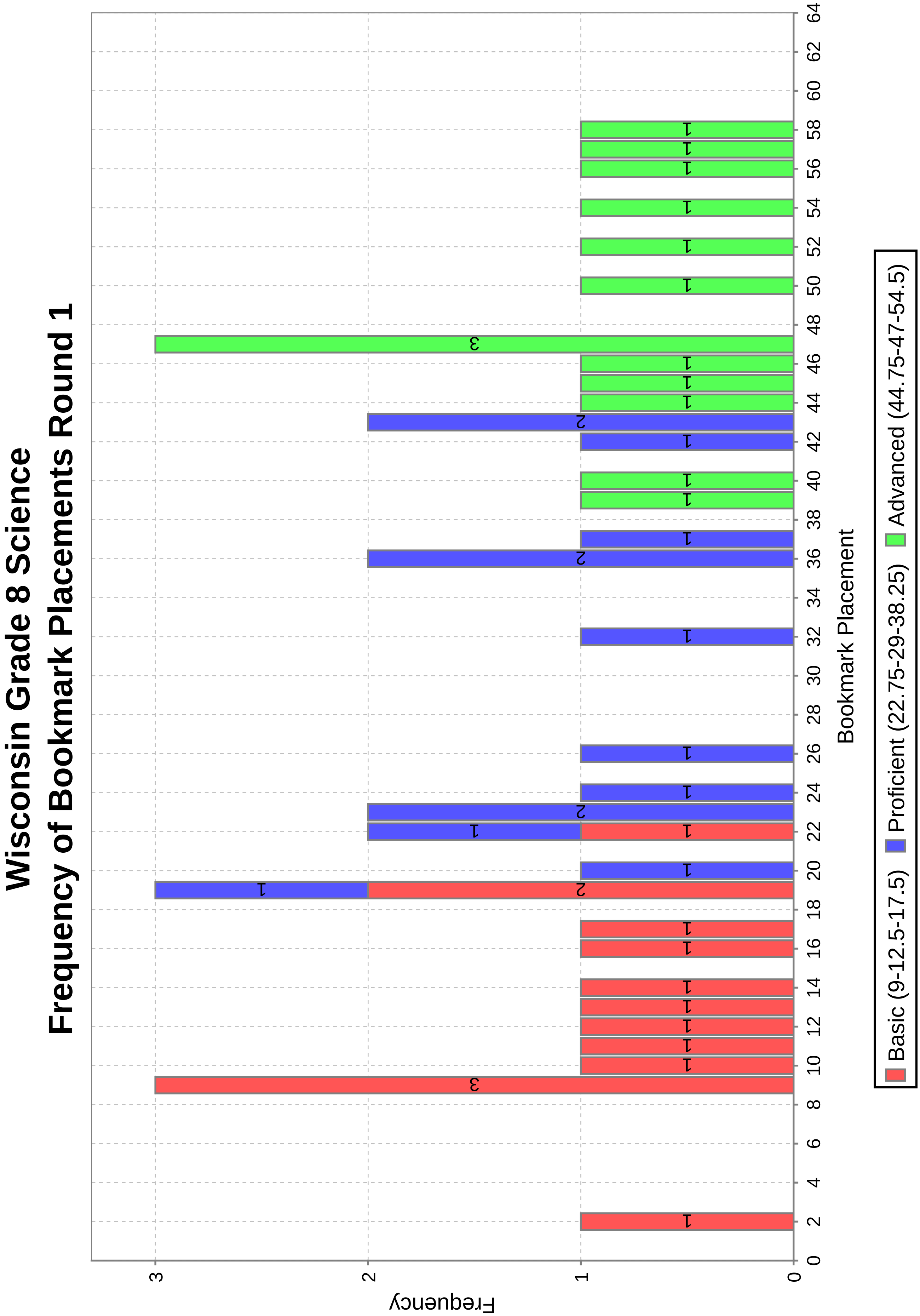
Wisconsin Grade 4 Science
Frequency of Bookmark Placements Round 1



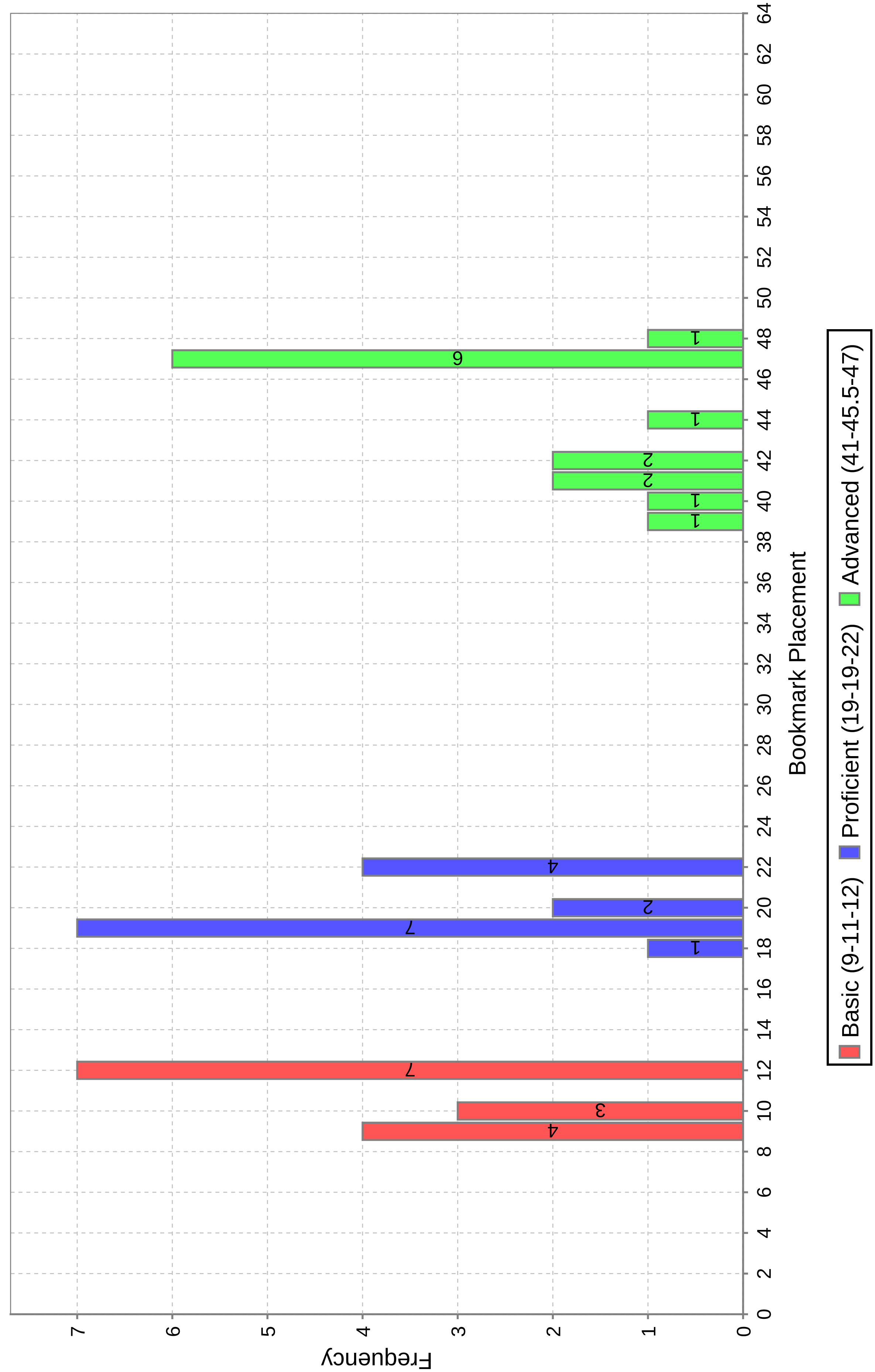

Wisconsin Grade 4 Science
Frequency of Bookmark Placements Round 2
Frequency
Bookmark Placement
Basic (8-9-9)
Proficient (31-32-32)
Advanced (44-49-51)

Wisconsin Grade 4 Science
Frequency of Bookmark Placements Round 3

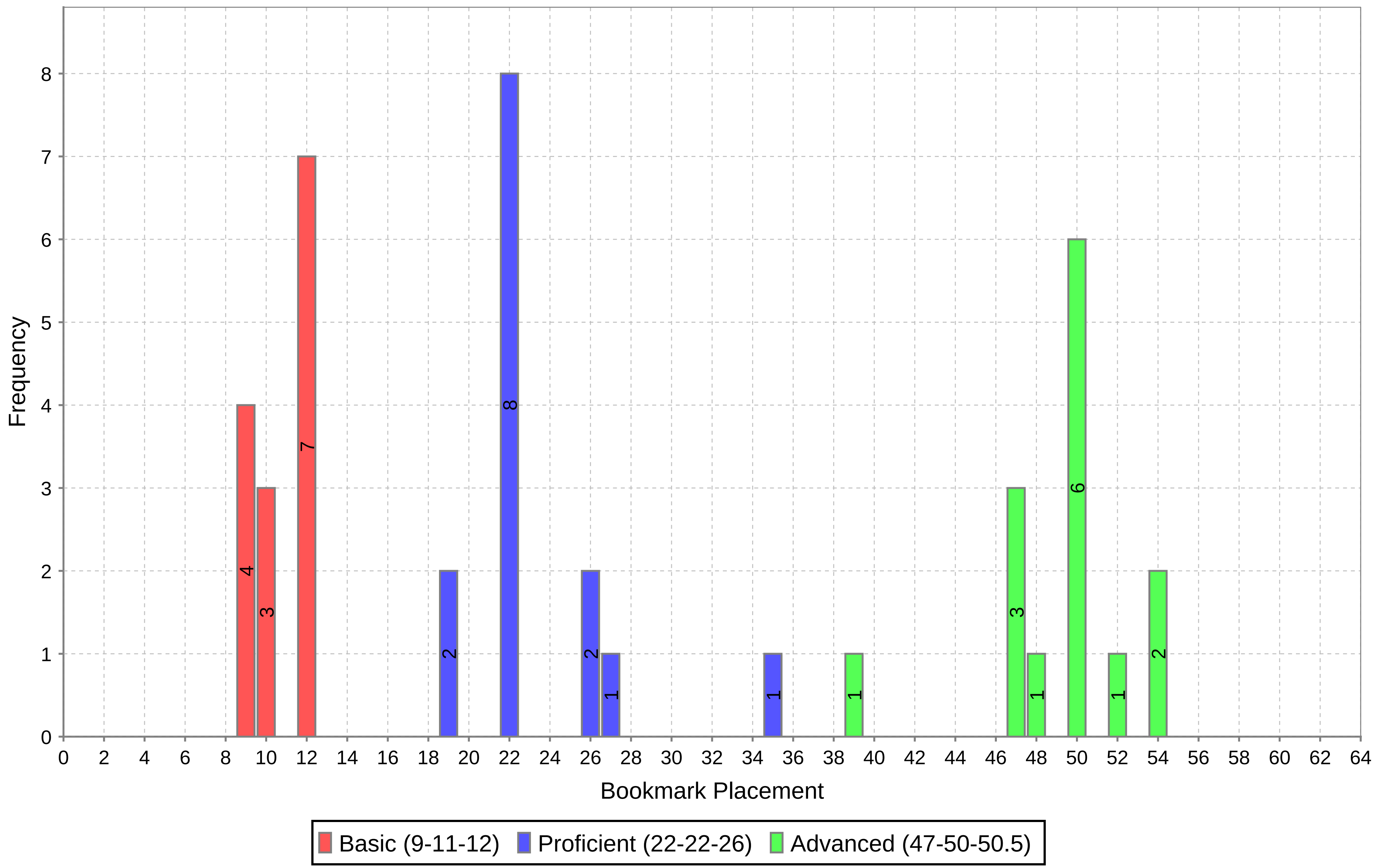



Wisconsin Grade 8 Science
Frequency of Bookmark Placements Round 1
Frequency
Bookmark Placement
Basic (9-12.5-17.5)
Proficient (22.75-29-38.25)
Advanced (44.75-47-54.5)

Wisconsin Grade 8 Science
Frequency of Bookmark Placements Round 2



Wisconsin Grade 8 Science
Frequency of Bookmark Placements Round 3



Appendix E

Standard Errors Associated with Cut Scores

Calculating a Meaningful Standard Error for the Bookmark Cut Score

In the Bookmark Standard Setting Procedure for a given grade and content area, participants are assigned to roughly equivalent small groups that work independently through Round 2. Thus, the set of Round 2 cut scores provide some information about the stability of consensus in Bookmark cut scores across independent small group replications. To quantify this degree of consensus, we calculate the cluster sample standard error (Cochran, 1963, p. 210) of the Round 2 mean cut score. Cluster sample standard errors are appropriate when, as may be reasonably assumed here, data are collected from groups and independence can be assumed between groups but not within groups.

For the Bookmark Procedure, the standard error of the Bookmark cut score (SE_{cut}) is based on the cluster sample standard error of the Round 2 mean cut score. Because the final Bookmark cut scores are based on the *median* of the group instead of the mean, this cluster sample standard error (SE_{cut}) is adjusted by $\sqrt{\frac{\pi}{2}}$ (Huynh, 2003). The standard error of the Bookmark cut score is:

$$SE_{cut} = \left(\sqrt{\frac{\pi}{2}}\right)\left(\sqrt{\frac{S^2}{N}\left[1+\left(\frac{N}{n}-1\right)r\right]}\right),$$

where S^2 is the sample variance of individual Round 2 cut scores, r is the Round 2 intraclass correlation, N is the number of participants, and n is the number of groups. To be precise, if Y_{ik} is the cut score from the i^{th} participant in the k^{th} group, $\bar{Y}_k$ is the average cut score for group k, and $\bar{\bar{Y}}$ is the average of all Round 2 cut scores, then

$$r = \frac{Var(\bar{Y}_k)}{Var(\bar{Y}_k) + Var(Y_{ik} - \bar{Y}_k)} \quad \text{and} \quad S^2 = \frac{1}{N-1}\sum_{n,k}\left(Y_{nk} - \bar{\bar{Y}}\right)^2$$

If we have only two groups ($n=2$) and perfect dependence (agreement) within groups ($r=1$), then the cluster sample standard error simplifies to $SE_{cut} = \left(\sqrt{\frac{\pi}{2}}\right)\left(\frac{|Y_1 - Y_2|}{2}\right)$, which is the standard error formula employed by NAEP for two independent replications of a modified Angoff procedure (ACT, 1983, pp. 4-8). If, on the other hand, individual participants acted independently of their groups ($r=0$), then the cluster sample standard error simplifies to the traditional standard error of the mean for independent observations, $SE_{cut} = \left(\sqrt{\pi/2}\right)\left(\sqrt{S^2/N}\right)$. In this manner, SE_{cut} provides a simple, flexible, and general way to quantify the amount of uncertainty associated with final Bookmark cut scores.

It is appropriate (if statistically imprecise) to say that repeated replications of this very standard setting procedure with different judges sampled from the same population of potential judges would result in a range of cut scores, most of which would fall in a band of width 4* SE_{cut}. In the graphical displays of participant data, we depict such an interval centered at the median of the Round 3 cut score. The purpose of calculating statistics like SE_{cut} and producing graphs of the types displayed here is to effectively communicate the complex information that is gathered during a Bookmark Standard Setting Procedure.

References

ACT (1993). Setting achievement levels on the 1992 National Assessment of Educational Progress in Mathematics, Reading, and Writing: A technical report on reliability and validity.

Cochran, W. G. (1963). *Sampling techniques*. New York: John Wiley & Sons.

Huynh, H. (2003, August). Technical Memorandum for Computing Standard Error in Bookmark Standard Setting. (The South Carolina PACT 2003 Standard Setting Support Project). Columbia: University of South Carolina.

Wisconsin Grade 4 Science

Recommended Cut Points* Plus/Minus Selected Standard Errors (SEs) of the Cut Score

Performance Level	Below Basic	Basic	Proficient	Advanced	
SE (cut score)		1.76	2.98	6.66	
Recommended Cut Point* + 3 SE		452	504	563	+ 3 SE
Percent of Students in Each Level	17.4	35.9	36.5	10.2	
Recommended Cut Point* + 2 SE		451	501	556	+ 2 SE
Percent of Students in Each Level	16.9	34.1	36.0	13.0	
Recommended Cut Point* + 1 SE		449	499	550	+ 1 SE
Percent of Students in Each Level	15.9	33.6	34.7	15.8	
Recommended Cut Point*		447	496	543	Recommended Cut Points*
Percent of Students in Each Level	15.0	32.2	33.3	19.5	
Recommended Cut Point* -1 SE		445	493	536	-1 SE
Percent of Students in Each Level	14.0	30.9	31.6	23.6	
Recommended Cut Point* -2 SE		444	490	530	-2 SE
Percent of Students in Each Level	13.5	29.1	30.0	27.4	
Recommended Cut Point* -3 SE		442	487	523	-3 SE
Percent of Students in Each Level	12.5	27.9	27.3	32.3	

* Participants' Large Group Medians

Wisconsin Grade 4 Science

Recommended Cut Points* Plus/Minus Selected Standard Errors (SEs) of Measurement

Performance Level	Below Basic	Basic	Proficient	Advanced	
Standard Error (SE) measurement		15.00	14.00	16.00	
Recommended Cut Point* + 3 SE		492	538	591	+ 3 SE
Percent of Students in Each Level	44.1	33.5	18.7	3.7	
Recommended Cut Point* + 2 SE		477	524	575	+ 2 SE
Percent of Students in Each Level	33.0	35.3	24.9	6.7	
Recommended Cut Point* + 1 SE		462	510	559	+ 1 SE
Percent of Students in Each Level	23.1	34.9	30.3	11.7	
Recommended Cut Point*		447	496	543	Recommended Cut Points*
Percent of Students in Each Level	15.0	32.2	33.3	19.5	
Recommended Cut Point* -1 SE		432	482	527	-1 SE
Percent of Students in Each Level	8.3	28.3	33.9	29.6	
Recommended Cut Point* -2 SE		417	468	511	-2 SE
Percent of Students in Each Level	3.9	23.0	31.8	41.2	
Recommended Cut Point* -3 SE		402	454	495	-3 SE
Percent of Students in Each Level	1.6	16.9	27.8	53.6	

* Participants' Large Group Medians

Wisconsin Grade 4 Science

Recommended Cut Points* Plus/Minus Selected Standard Errors (SEs) of Measurement and the Cut Score

Performance Level	Below Basic	Basic	Proficient	Advanced	
Standard Error (SE) measurement + cutscore		15.10	14.31	17.33	
Recommended Cut Point* + 3 SE		493	538	595	+ 3 SE
Percent of Students in Each Level	44.8	32.8	19.3	3.1	
Recommended Cut Point* + 2 SE		477	524	578	+ 2 SE
Percent of Students in Each Level	33.0	35.3	25.5	6.1	
Recommended Cut Point* + 1 SE		462	510	560	+ 1 SE
Percent of Students in Each Level	23.1	34.9	30.7	11.3	
Recommended Cut Point*		447	496	543	Recommended Cut Points*
Percent of Students in Each Level	15.0	32.2	33.3	19.5	
Recommended Cut Point* -1 SE		432	481	526	-1 SE
Percent of Students in Each Level	8.3	27.6	33.9	30.2	
Recommended Cut Point* -2 SE		417	467	508	-2 SE
Percent of Students in Each Level	3.9	22.3	30.2	43.5	
Recommended Cut Point* -3 SE		402	453	491	-3 SE
Percent of Students in Each Level	1.6	16.4	25.4	56.6	

* Participants' Large Group Medians

Wisconsin Grade 8 Science

Recommended Cut Points* Plus/Minus Selected Standard Errors (SEs) of the Cut Score

Performance Level	Below Basic	Basic	Proficient	Advanced	
SE (cut score)		1.39	2.93	3.45	
Recommended Cut Point* + 3 SE		657	690	734	+ 3 SE
Percent of Students in Each Level	19.9	22.2	33.6	24.4	
Recommended Cut Point* + 2 SE		655	688	730	+ 2 SE
Percent of Students in Each Level	18.7	21.8	32.6	26.9	
Recommended Cut Point* + 1 SE		654	685	727	+ 1 SE
Percent of Students in Each Level	18.2	20.1	32.7	29.0	
Recommended Cut Point*		653	682	723	Recommended Cut Points*
Percent of Students in Each Level	17.7	18.2	32.2	31.8	
Recommended Cut Point* -1 SE		651	679	720	-1 SE
Percent of Students in Each Level	16.7	17.1	32.2	34.0	
Recommended Cut Point* -2 SE		650	676	716	-2 SE
Percent of Students in Each Level	16.2	15.5	31.3	37.0	
Recommended Cut Point* -3 SE		648	673	713	-3 SE
Percent of Students in Each Level	15.2	14.5	30.9	39.4	

* Participants' Large Group Medians

Wisconsin Grade 8 Science

Recommended Cut Points* Plus/Minus Selected Standard Errors (SEs) of Measurement

Performance Level	Below Basic	Basic	Proficient	Advanced	
Standard Error (SE) measurement		15.00	13.00	14.00	
Recommended Cut Point* + 3 SE		698	721	765	+ 3 SE
Percent of Students in Each Level	48.4	18.3	23.7	9.6	
Recommended Cut Point* + 2 SE		683	708	751	+ 2 SE
Percent of Students in Each Level	36.7	19.9	28.3	15.1	
Recommended Cut Point* + 1 SE		668	695	737	+ 1 SE
Percent of Students in Each Level	26.4	19.5	31.5	22.5	
Recommended Cut Point*		653	682	723	Recommended Cut Points*
Percent of Students in Each Level	17.7	18.2	32.2	31.8	
Recommended Cut Point* -1 SE		638	669	709	-1 SE
Percent of Students in Each Level	10.8	16.3	30.4	42.6	
Recommended Cut Point* -2 SE		623	656	695	-2 SE
Percent of Students in Each Level	5.8	13.5	26.7	54.0	
Recommended Cut Point* -3 SE		608	643	681	-3 SE
Percent of Students in Each Level	2.7	10.3	22.2	64.8	

* Participants' Large Group Medians

Wisconsin Grade 8 Science

Recommended Cut Points* Plus/Minus Selected Standard Errors (SEs) of Measurement and the Cut Score

Performance Level	Below Basic	Basic	Proficient	Advanced	
Standard Error (SE) measurement + cutscore		15.06	13.33	14.42	
Recommended Cut Point* + 3 SE		698	722	766	+ 3 SE
Percent of Students in Each Level	48.4	19.0	23.2	9.3	
Recommended Cut Point* + 2 SE		683	708	752	+ 2 SE
Percent of Students in Each Level	36.7	19.9	28.7	14.7	
Recommended Cut Point* + 1 SE		668	695	738	+ 1 SE
Percent of Students in Each Level	26.4	19.5	32.1	21.9	
Recommended Cut Point*		653	682	723	Recommended Cut Points*
Percent of Students in Each Level	17.7	18.2	32.2	31.8	
Recommended Cut Point* -1 SE		637	668	709	-1 SE
Percent of Students in Each Level	10.4	16.1	31.0	42.6	
Recommended Cut Point* -2 SE		622	655	694	-2 SE
Percent of Students in Each Level	5.6	13.1	26.5	54.8	
Recommended Cut Point* -3 SE		607	642	680	-3 SE
Percent of Students in Each Level	2.5	9.9	22.0	65.5	

* Participants' Large Group Medians

Appendix F

Selecting a Response Probability Criterion

Selecting a Response Probability Criterion

On May 29–30, 2019, the Wisconsin Department of Public Instruction (DPI) will partner with DRC to conduct a standard setting for the Wisconsin Forward Exams in grades 4 and 8 science. The tests administered in spring 2019 reflect challenging new content standards and new item types, and it is expected that students' performance in 2019 will be somewhat lower than that seen in previous years.

The standard setting is designed to follow the same implementation of the Bookmark Standard Setting Procedure as used at the original Wisconsin Forward Exam standard setting in 2016. Accordingly, DRC is preparing an ordered item booklet (OIB) and item map for each test, and participants will engage in three rounds of discussions and judgments to recommend cut scores for the tests.

In advance of the standard setting, two questions arose, both sparked by analysis of the 2019 data and by reflection on the 2016 standard setting methodology. These questions are shown here. The subsequent document elaborates on the questions and presents the decisions made by DPI in response to the questions, as informed by a call with a member of the technical advisory committee (TAC).

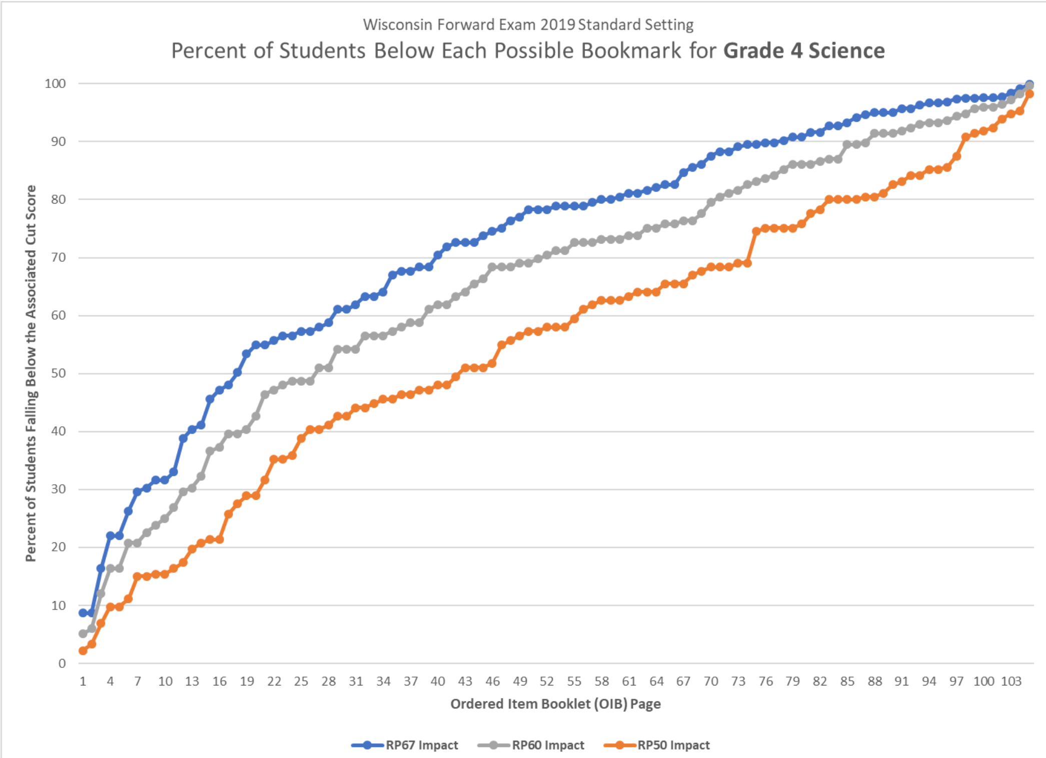
1) What response probability (RP) value should be used to create the OIB, RP67 or RP50?
2) When should impact data be shared with standard setting participants, after Round 1 or 2?

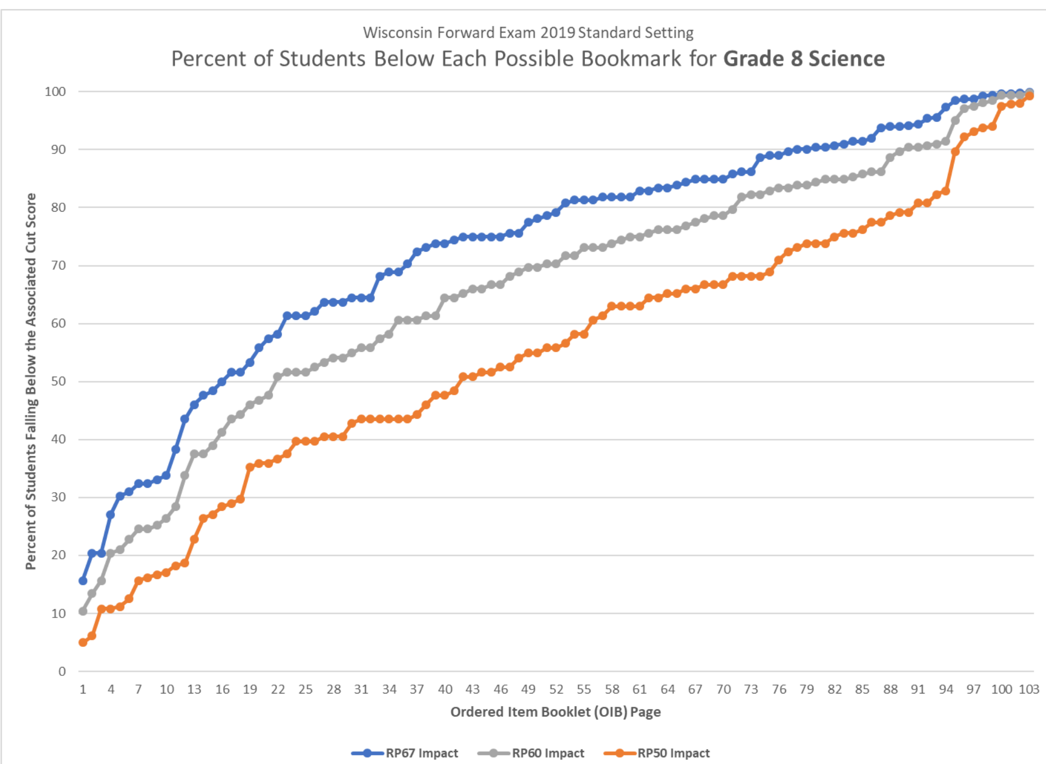
Selecting a response probability (RP) value

In 2016, the OIBs were created with a response probability of 0.67 (RP67): the RP-adjusted scale location for each item in the OIB was associated with the scale score needed to have a 67% chance of answering the item correctly. In preparing for that workshop, DPI acknowledged that it had a long history of using RP67 at its standard settings. RP67, with a correction for guessing (also called RP67GA), is frequently associated with the Bookmark Procedure; however, many state agencies have used other RP criteria for their standard settings, including RP50 (associated with a 50% chance of answering an item correctly).

The classical item analysis shows that the average adjusted *p*-value for the 40 operational items for grade 4 science was .56, and for grade 8 was .54. For each test, field test items were placed on the scale, making them suitable to be included in the OIBs. Flagged field test items were excluded from each grade's original 99 field test items, this left 65 field test items in grade 4 and 63 field test items in grade 8 from the spring 2019 administration analysis to be considered for inclusion in the OIBs with the 40 operational items. Hypothetical OIBs were created using RP67 and all 105 items in grade 4 and 103 items in grade 8. DRC calculated the percent of students who would be classified below the cut score for each possible bookmark at the Bookmark Procedure (i.e., the percent of students that would fall below the cut score associated with OIB page 1, with OIB page 2, and so on). To extend this analysis, DRC also created two additional hypothetical OIBs using RP50 and RP60. (These same values were used during analogous analyses in 2016). The results of these analyses (*impact data* curves) are presented in Figure 1.

Figure 1. Percent of students classified below each possible bookmark in hypothetical ordered item booklets (OIBs) using three response probability (RP) criterion and all 2019 items, by grade





The RP analysis showed that there are relatively few easy items available for selection in the OIBs. If RP67 is used, the percentage of students classified below the cut score may be high, even for a bookmark early in the OIB. For example, in grade 4, even if all 105 available items were used in the OIB, a bookmark after page 18 would classify 50% of students below the cut score, and a bookmark after page 4 would classify 22% of students below the cut score. For grade 8, bookmarks after page 16 and after page 2 would classify 50% and 20% of students below the cut score, respectively. Table 1 shows additional bookmarks associated with different percentages.

Table 1. Bookmarks needed to classify 20–60% of students below the cut score in hypothetical OIBs using three selected response probability (RP) criteria and all 2019 items (105 total items per grade 4 OIB, 103 total items per grade 8 OIB)

	Grade 4			Grade 8		
Bookmark needed to classify	**RP50**	**RP60**	**RP67**	**RP50**	**RP60**	**RP67**
20% below the cut score	13	6	4	12	4	2
30% below the cut score	20	13	8	18	11	5
40% below the cut score	26	18	13	26	15	11
50% below the cut score	42	27	18	42	22	16
60% below the cut score	55	38	29	56	35	23

Again, the hypothetical OIBs represented in Figure 1 and Table 1 use all available operational and field test items: no additional, easier items are available to augment the OIB. However, OIBs of this length are unwieldy at standard settings: having many items around the same scale locations is not typically meaningful to standard setting participants. For the standard setting, DRC intends to create OIBs that:

- have approximately 60 items (or about 20 items per cut score to be recommended)
- reflect the test blueprint (with roughly equal numbers of items from each reporting category)
- include a mix of easy, medium, and hard items along the test scale

DRC selected items for hypothetical OIBs under two of the RP criteria described above: RP67 and RP50. RP67 was selected because it reflects the criterion historically used in Wisconsin. RP50 was selected because has been used frequently by other states and agencies (e.g., Smarter). Moreover, RP50 was selected because Figure 1 shows that when RP50 is applied, more of the items are mapped at scale locations associated with the first two deciles of examinees. Of course, different RP criteria have different meanings at the Bookmark Procedure, as described later in this document.

The *impact data curves* for hypothetical OIBs created using RP50 and RP67 are presented in Figures 2 and 3 for grades 4 and 8, respectively. In each OIB, items are selected purposefully using the three guidelines listed above. The bookmarks associated with having 20–60% of students below the cut score are shown in Table 2.

As shown in Figures 2 and 3, using RP67 to create OIBs for the standard setting creates a situation where bookmarks placed early in the OIB will yield many students below the associated cut score. Although this is precedented in Wisconsin—in 2016, a similar pattern was seen in middle-school mathematics—DPI may wish to consider whether to continue to use RP67 for the upcoming science standard setting.

Figure 2. Percent of students classified below each possible bookmark in hypothetical grade 4 science OIBs using RP50 and RP67 and using selected items

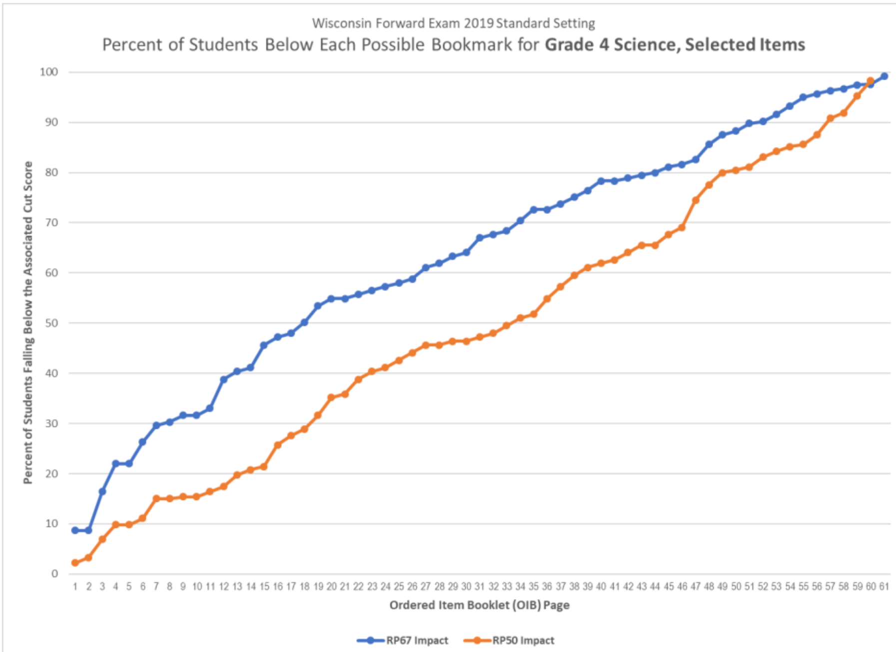


Figure 3. Percent of students classified below each possible bookmark in hypothetical grade 8 science OIBs using RP50 and RP67 and using selected items

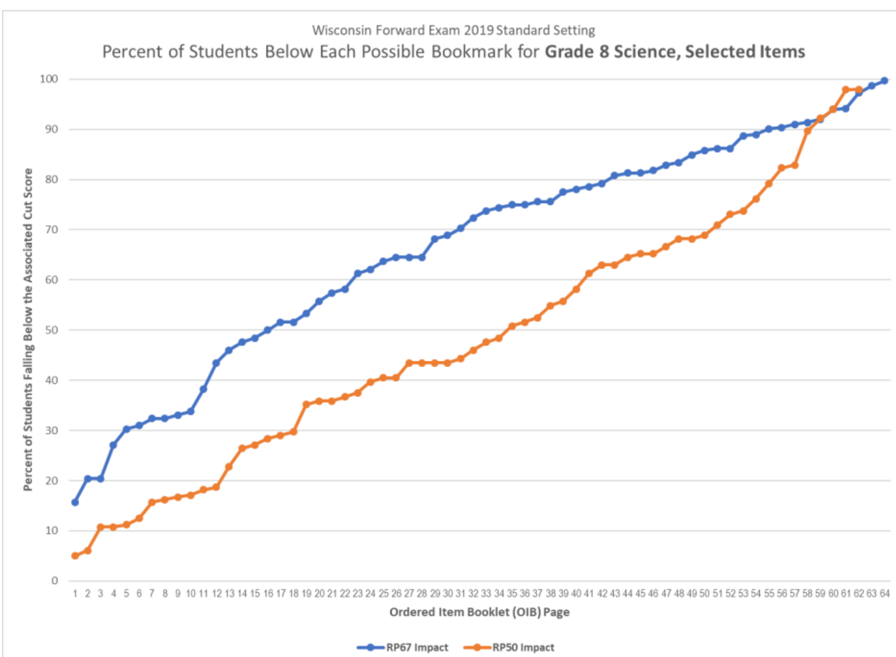


Table 2. Bookmarks needed to classify 20–60% of students below the cut score in hypothetical OIBs using RP50 and RP67 and using selected items (60–64 total items per OIB)

	Grade 4		Grade 8	
Bookmark needed to classify	**RP50**	**RP67**	**RP50**	**RP67**
20% below the cut score	13	4	12	2
30% below the cut score	19	8	18	5
40% below the cut score	23	13	24	11
50% below the cut score	33	18	35	16
60% below the cut score	38	27	41	23

Advantages and challenges of retaining RP67. By using RP67, DPI would continue to use the RP criterion that it used at the 2016 and previous standard settings. At the standard setting, participants would place their bookmarks based on the content that they expect students to have *mastery* of, defined as content measured by items they expect a student at the threshold of a performance level would have at least a 67% chance of answering correctly. However, it would be difficult for participants to recommend cut scores where fewer than 40% of students fall below the cut score—especially for grade 8—because not many easy items exist in the current test pool (when compared to current student performance). Some participants may feel cognitive dissonance during the workshop: they may have a disconnect between the content they expect students to master and the impact data they see during the workshop. Participants may be tempted during later rounds of the Bookmark Procedure to change their bookmarks dramatically in response to impact data presented.

Advantages and challenges of using RP50. By using RP50, DPI would acknowledge that the performance of Wisconsin students is not yet well-aligned with the difficulty of the items on the test. This is unsurprising, given the shift in content standards and test format over the last year. At the standard setting, participants would place their bookmarks based on the content that they expect students to have *command* of, defined as the items that measure content measured by items they expect a student at the threshold of the performance level would have at least a 50% chance of answering correctly. Participants would be able to use the items in the OIB to recommend a full range of less- or more-stringent performance standards, as informed by their study of the performance level descriptors (PLDs) and tested content. To date, however, grades 4 and 8 science would be the only tests in the Forward Exam program which use RP50 at the standard setting: the change in RP would need to be well documented, and the standard setting training would need to fully reflect the shift in RP.

Decision to use RP50. On May 20, 2019, DPI consulted with a member of its technical advisory committee (TAC) to consider which RP criterion to use at the upcoming standard setting. After consultation, DPI decided to use RP50 for the standard setting. The decision to use RP50 was made for several reasons. First, DPI noted that it would be difficult for standard setting participants to use OIBs created with RP67 to recommend cut scores that would yield impact data similar to Wisconsin's performance on NAEP (as shown in Table 3), especially for the *Basic* cut score. Although the state does not necessarily expect the results of the test to be similar to NAEP, it is not unreasonable to expect that standard setting participants may want to recommend a performance standard that classifies 20–30% of students as *Below Basic*: such a standard would be difficult to recommend using RP67. Second, DPI acknowledged that it has a history of using RP67 at its standard setting, but that this precedent is not as important as the first point: it is more important to have a useful OIB that can be used by participants to recommend reasonable performance standards than it is important to comport with historically-implemented RP criteria. Third, DPI noted that RP50 has been used frequently by other states and agencies, and the training protocol at the Bookmark Procedure can be adjusted to help participants understand how to interpret an OIB created using RP50. Such an adjustment, focusing on helping participants develop an intuitive understanding of RP50 in the OIB, will be implemented by DRC.

Timing of impact data

The science standard setting will take place over two days. On Day 1, participants will be trained, will study the content standards and PLDs, will study the OIBs, and will place their Round 1 bookmarks. DPI has indicated that Round 1 will be driven entirely by content: participants will not be shown benchmarks or impact data before Round 1, and participants will place their bookmarks individually.

On Day 2, participants will go through two more rounds of bookmark placements. During this time, DPI is interested in showing participants:

- Impact data based on the performance of students on the spring 2019 assessments
- External, benchmark data from Wisconsin students' performance on NAEP Science in 2015

These NAEP benchmarks were also shared with participants during the 2016 standard setting. However, these benchmarks were shared with participants before Round 1, and some observers believed these benchmarks may have unduly swayed participants' judgments during the workshop. Accordingly, DPI is interested in sharing these benchmarks later in the workshop in 2019.

Table 3 shows the external NAEP data, as well as the impact data from the 2018 Wisconsin science assessments. DPI has indicated it is unlikely that students would perform as well on the 2019 science tests as they did in 2018 because of shifts to the content standards. Accordingly, DPI has indicated that these prior-year impact data will be shared with standard setting participants only if necessary.

Table 3. Impact data from the performance of Wisconsin students on NAEP Science (2015) and the Wisconsin Forward Exam science tests (2018)

		Performance Level			
Testing Program	Grade	Below Basic	Basic	Proficient	Advanced
State NAEP Science (2015)	4	21%	38%	40%	1%
	8	25%	35%	38%	2%
Wisconsin Science (2018)	4	15.24%	34.07%	34.43%	16.26%
	8	17.18%	33.96%	34.16%	14.70%

Typically, current-year impact data are shared with participants after Round 2 of the Bookmark Procedure; however, there is a great deal of variation among states in this practice. DPI is interested in making sure the impact data (and external benchmarks) shared with participants are timely, helpful, and not distracting.

Impact data and benchmarks after Round 2. After consulting with a member of its TAC, DPI reaffirmed that it does not have an *a priori* expectation of the impact data it expects from the standard setting. DPI would like Wisconsin educators to use the standards, tested content, PLDs, and their professional judgment as their primary guides at the standard setting, not necessarily external test data. However, DPI would like to provide participants with an external benchmark, based on Wisconsin's performance on NAEP Science, later in the standard setting process. The of the information that to be provided to participants during the Bookmark Procedure is shown here.

- *Before Round 1:* Participants use only the content to guide their decisions
- *After Round 1:* Participants see their group's median bookmarks from Round 1
- *After Round 2:* Participants see their group's median bookmarks from Round 2, plus the 2019 impact data associated with their median cut score recommendations, plus the impact data from the 2015 NAEP Science assessment for reference
- *After Round 3:* Participants see their group's median bookmarks from Round 3, plus the 2019 impact data from both grades 4 and 8

After each Round, DRC will share with DPI the groups' cut score recommendations and associated impact data. Should it be needed, DPI reserves the right to adjust this plan and present additional data (e.g., prior-year performance on the Wisconsin science tests) to help participants through the standard setting.

Appendix G

Performance Level Descriptors (PLDs)

PLDs Prior to Participants’ Refinements

	Performance Level Descriptors (PLDs)			
	Below Basic	**Basic**	**Proficient**	**Advanced**
Policy	Student demonstrates minimal understanding of and ability to apply the knowledge and skills for their grade level that are associated with college content-readiness.	Student demonstrates partial understanding of and ability to apply the knowledge and skills for their grade level that are associated with college content-readiness.	Student demonstrates adequate understanding of and ability to apply the knowledge and skills for their grade level that are associated with college content-readiness.	Student demonstrates thorough understanding of and ability to apply the knowledge and skills for their grade level that are associated with college content-readiness.
Standard: Students use science and engineering practices, crosscutting concepts, and an understanding of life science disciplinary core ideas to make sense of phenomena and solve problems.	may describe primary functions of main structures in everyday plants and animals. may identify important sense receptors within a system that supports basic animal behaviors. may use a model to recognize that a variety of factors in the environment can be sensed by animals (e.g., sound, light, odor, temperature).	can analyze evidence to determine if it supports a claim about the role of external structures of plants and animals in supporting survival and reproduction. can give evidence of the sequence of events resulting in a given animal behavior (i.e., sensory input, sense receptor, brain processing, behavioral output). can describe how data shows a cause and effect relationship between an environmental stimuli and an animal's behavior.	can provide feedback and ask questions about a claim and its supporting evidence about the role of internal and external structures of plants and animals in supporting survival, growth, behavior, and reproductive success. can develop a model of an animal behavior (phenomenon) showing various components (i.e., sensory input, sense receptor, the brain, behavioral output) working together as a system. can develop a model of sensory systems showing how animals' memories can impact future behavior, survival, and reproduction.	can develop a model showing different plant or animal structures working together as part of a system to support survival, growth, behavior, and reproductive success. can create or improve a model of a phenomenon based on evidence to explain how sensory systems and behavioral output function to support animal survival, growth, and reproductive success. can analyze an animal's behavior and describe reasonable, possible initial causes based on given evidence.

Standard: Students use science and engineering practices, crosscutting concepts, and an understanding of physical science disciplinary core ideas to make sense of phenomena and solve problems.	may use given evidence to describe the relative speed of an object (e.g., faster vs. slower).	can describe that the speed of a given object is related to the energy of the object (e.g., the faster an object is moving, the more energy it possesses).	can interpret given quantitative data to support the idea that the speed of a given object is related to the energy of the object (e.g., the faster an object is moving, the more energy it possesses).	can use evidence and reasoning to construct an explanation for how a given phenomenon affects the speed and related energy of an object.
	may identify examples showing a transfer of energy.	can describe the purpose of an investigation of a phenomenon related to energy transfer (e.g., moving objects, sound, light, heat, electric currents).	can plan and conduct an investigation that fairly tests a phenomenon involving the transfer of energy from place to place (e.g., moving objects, sound, light, heat, electric currents).	can obtain and evaluate evidence from multiple sources to design a solution to a problem related to the transfer of energy.
	may identify a device that converts energy from one type to another (e.g., a light bulb to convert electrical energy into light energy).	can identify a possible solution to a given problem involving the conversion of energy from one form to another.	can design an evidence-based improvement to local transportation systems or energy grids to reduce the environmental impact of the conversion of energy from one form to another.	can analyze and interpret evidence gathered from testing a device that converts energy from one form to another and use the results of the test to address problems in the design or improve its functioning.
	may identify a phenomenon in which waves can cause an object to move.	can compare waves in phenomena in terms of amplitude and wavelength.	can develop a model of a phenomenon related to wave behavior that describes wave amplitude, wavelength, or motion of objects (e.g., wave models of loud vs soft sound).	can design a solution to transfer information over a distance, comparing digitized information transfer to other methods in addressing particular criteria and constraints.

Standard: Students use science and engineering practices, crosscutting concepts, and an understanding of earth and space science disciplinary core ideas to make sense of phenomena and solve problems.	may use fossil evidence to infer a basic feature of what an environment used to be like (e.g. marine fossils indicate that in the past a landscape was covered in water).	can ask cause and effect questions about rock layers, fossils, and geological features that could lead to productive investigations about these phenomena.	can use a diagram of rock layers and fossils, as well as other geological features such as canyons, to help explain how an environment has changed over time.	can design a solution, that addresses particular criteria and constraints, to prevent water, ice, or wind from impacting a particular landscape.
	may recognize which type of maps can be used to best locate different land and water features on Earth.	can use evidence from given topographic maps to identify various Earth features (e.g., mountain ranges, ocean trenches, ocean floor structures, fault lines, volcanoes).	can use patterns in a map as evidence to explain where geologic processes are likely to occur (e.g., earthquakes, erosion, volcanoes).	can evaluate a map of a fantasy land to describe where it does or does not show reasonable patterns of geologic features.
	may identify examples of natural resources that humans use for energy.	can use given evidence to identify cause and effect relationships between the use of a natural resource and its likely impact on the environment.	can analyze and interpret patterns in evidence to describe that energy and fuels are derived from natural resources (e.g., fossil fuels, solar, wind, water) and their uses can have various effects on the environment.	can design a solution based on evidence from multiple sources to a problem related to the use of natural resources and their effects on the environment.
	may identify possible negative impacts to humans from a natural Earth process (e.g., an earthquake, volcano, flood, landslide).	can use evidence to describe how one Earth process can have a greater negative impact compared to another Earth process in a given area or region.	can use evidence to design a possible solution to reduce the impacts of natural Earth processes on humans.	can use evidence to generate multiple possible solutions to reduce the impacts of natural Earth processes on humans and then evaluate which one bests addresses criteria and constraints.

Standard: Students use science and engineering practices, crosscutting concepts, and an understanding of engineering, technology and applications of science disciplinary core ideas, to make sense of phenomena and solve problems.	may use scientific understanding to define a simple design problem that includes responding to a human need or want. may use scientific understanding to identify which tools and methods could be used to collect data for a given investigation.	can use scientific understanding to define a problem related to local phenomena that can be solved with the development of a new or improved object, tool, process, or system. can use scientific understanding to make conclusions related to how well a model works or a prototype performs against given criteria and constraints.	can use given scientific information and information about an everyday situation or phenomenon to design a solution to a problem that includes responding to a needs or wants of humans. can conduct fair tests in which variables are controlled and possible failure points are considered when designing a prototype to solve problems related to a particular local phenomenon.	can use scientific understanding to evaluate multiple possible solutions to a problem and describe how well the solution addresses the constraints within which the problem must be solved. can use scientific understanding to evaluate a range of new technologies to determine how they may change how people live and interact with each other.

*Not all 4th grade crosscutting concepts, science and engineering practices, and disciplinary core ideas, or combinations thereof, are contained within this document. These Performance Level Descriptors are intended as examples of the type of the types of skills and understanding students would have at each performance level.

	Performance Level Descriptors (PLDs)			
	Below Basic	**Basic**	**Proficient**	**Advanced**
Policy	Student demonstrates minimal understanding of and ability to apply the knowledge and skills for their grade level that are associated with college content-readiness.	Student demonstrates partial understanding of and ability to apply the knowledge and skills for their grade level that are associated with college content-readiness.	Student demonstrates adequate understanding of and ability to apply the knowledge and skills for their grade level that are associated with college content-readiness.	Student demonstrates thorough understanding of and ability to apply the knowledge and skills for their grade level that are associated with college content-readiness.
Standards: Students use science and engineering practices, crosscutting concepts, and an understanding of life science disciplinary core ideas, to make sense of phenomena and solve problems.	**A student at this level** may make observations to use as evidence that an object is a living or non-living thing. may identify evidence supporting the claim that some animal behaviors can affect the survival of another species. may ask questions that could guide an investigation into differences between how plants and animals obtain food. may engage in the understanding of phenomena associated with sense receptors responding to different inputs (i.e., mechanical, chemical, electromagnetic).	**A student at this level** can collect and analyze data during an investigation to determine whether existence of cells can be a distinguishing characteristic of living things. can use patterns to predict how a certain animal behavior will likely affect the reproductive success of multiple other species. can use a scientific model to describe how food molecules are rearranged through chemical reactions to form new molecules that support growth and/or to release energy as this matter moves through an organism. can construct an explanation, based on cause and effect evidence, of a phenomenon that results in sensory receptors sending signals to the brain.	**A student at this level** can conduct an investigation to provide evidence that tissues and organs are made of cells with specialized functions in the body system. can develop a model showing how genetic and/or environmental factors can affect an organism's growth and reproductive success. can construct a scientific explanation based on evidence for the role of photosynthesis in the cycling of matter and flow of energy into and out of ecosystems. can develop and use a model to show sensory receptors responding to stimuli by sending messages to the brain for immediate behavior or storage as memories.	**A student at this level** can use arguments based on scientific reasoning and evidence from multiple sources to support the idea that a body is a system of interacting subsystems composed of various cells. can analyze the validity and reliability of given evidence to solve problems related to biological and environmental factors affecting organisms. can develop a model to track the changes of energy and matter into, out of, and within the systems of a plant, noting the limitations of the model. can synthesize information that provides evidence of causal relationships between information received by sensory receptors and behavior, at various time scales.

	may describe how a change in the amount or availability of a natural resource can result in changes in a population of organisms.	can analyze and interpret graphical displays of data (e.g., graphs, charts) to provide evidence of the relationships between resource availability and organism abundance.	can make sense of phenomena related to patterns of interactions among organisms to help explain relationships within ecosystems.	can use reasoning and evidence in an explanation of interactions within an ecosystem and predict future interactions based on patterns in that evidence.
	may identify examples of producers, consumers, and/or decomposers within an ecosystem and what effects they have in that system.	can construct a model of how matter and energy are transferred between producers, consumers, and decomposers within an ecosystem.	can construct an evidence-based argument describing the conservation of matter within and flow of energy into and out of an ecosystem.	can develop and use a model to explain the transfer of matter (atoms) and energy between living and nonliving parts of the ecosystem at various levels within the system, and define the boundaries of that system.
	may recognize that structural changes to genes (i.e., mutations) may result in observable changes in organisms.	can use a model (i.e., Punnett square) to show that more genetic variation occurs in organisms that reproduce sexually compared to organisms that reproduce asexually.	can use evidence related to a phenomenon (i.e., pictures of a litter of puppies or diagrams of fern spores and amoeba) to predict the differences in genetic variation resulting from sexual and asexual reproduction.	can use a model of a protein to explain how changes to protein structure can lead to changes in its function that may cause beneficial, neutral, or harmful changes in the structures or functions of organisms.
	may identify a similar anatomical feature shared by organisms that suggests they are likely to be more closely related than other organisms that do not share the similar anatomical feature.	can construct an argument based on evidence that some organisms sharing a pattern of anatomical features are likely to be more closely related than organisms that do not share a pattern of anatomical features.	can apply scientific ideas and evidence to construct an explanation for the anatomical similarities and differences between modern and fossil organisms that describes evolution over time.	can connect multiple sources of evidence comparing modern living animals and fossilized animals to support an argument for past connections of multiple lines of descent of different species.
	may recognize that data shows that some organisms will better survive and reproduce from generation to generation due to traits that are advantageous in a specific environment.	can use patterns as evidence to support claims that some traits have advantages that make it more probable that an organism will be able to survive and reproduce in a specific environment.	can evaluate different explanations about natural selection within a population of organisms to determine which is better supported by evidence.	can analyze data to trace the increase or decrease of particular traits in a population over time and make claims about how those changes were likely the result of particular historical phenomena that changed their environment.

Standards: Students use science and engineering practices, crosscutting concepts, and an understanding of physical science disciplinary core ideas, to make sense of phenomena and solve problems.	may use simple models of different atoms and molecules as support for why different substances have different properties.	can plan experiments to identify different substances based on their characteristic physical properties (e.g., density, melting point).	can analyze data to identify changes in physical and chemical properties of substances before and after an interaction to make a claim about whether a chemical reaction occurred.	can use simple models to provide evidence for an argument that a change in properties of substances can be related to the rearrangement of atoms in a chemical reaction.
	may make sense of a phenomenon where a collision occurs by describing how a force was exerted by the first object on the second object.	can use evidence to model the components within the system that are involved in a collision between two objects.	can address a given problem involving a collision of two objects (e.g. toy people falling out of toy cars that collide) by detailing a process or system that helps solve the problem.	can evaluate data on two designs for solving a problem involving a collision of two objects to determine which design better meets the criteria and constraints of the situation.
	may use information from a model to recognize that an object subjected to balanced forces does not change its motion and an object subjected to unbalanced forces does change its motion.	can use mathematical thinking to explain how changes in an object's motion can be due to the degree of balanced or unbalanced forces acting on the object as well as the mass of the object.	can conduct an investigation involving the change in motion of an object and gather evidence identifying various factors affecting the object's motion.	can plan an investigation to provide evidence that the change in an object's motion depends on specific variables such as the initial motion of the object, the total forces acting on the object, and the mass of the object.
	may identify the relative magnitude and direction of the forces between objects in a given system.	can identify evidence from a given phenomenon to support the idea that gravitational forces are attractive and mass dependent.	can construct an argument based on evidence that gravitational interactions are attractive and dependent on the masses of interacting objects.	can evaluate evidence on how well it supports the idea that gravitational forces are attractive and mass dependent, identifying strengths and weaknesses of the evidence, and including possible alternative interpretations of it.
	may recognize that energy increases if either the mass or the speed of the object increases and that energy decreases if either the mass or the speed of the object decreases.	can interpret graphical displays of data to describe the relationships of kinetic energy to the mass of an object and to the speed of an object.	can construct graphical displays or other models to communicate the idea that the mass and speed of an object affect its kinetic energy.	can design a solution to a problem identified by analyzing data from multiple sources that show the relationships of kinetic energy to the mass and speed of an object.

	may identify that the interactions of two objects at a distance can cause a transfer of energy between the objects. may identify examples of mechanical waves that need a medium through which they are transmitted. may identify examples of waves interacting with materials by being reflected, absorbed, or transmitted.	can analyze and interpret data to provide evidence that as the relative position of two objects changes, the potential energy of the system changes (e.g. an object higher off the ground has more gravitational potential energy). can use mathematical representations to describe a simple model for waves (repeating pattern) that includes how the amplitude of a wave is related to the energy in a wave. can use a given model to make sense of given phenomena involving reflection, absorption, or transmission properties of different materials for light and matter waves.	can develop a model to describe that when the arrangement of objects interacting at a distance changes, different amounts of potential energy are stored in the system. can use a model to describe a simple model for waves (repeating pattern) that includes evidence for how the amplitude of a wave is related to the energy in a wave. can develop and use a model about phenomena involving light and/or matter waves to describe the differences between how light and matter waves interact with different materials.	can plan and conduct an investigation to make sense of a given phenomenon involving two objects interacting at a distance. can develop a model to compare mechanical and electromagnetic waves and explain similarities and differences in how energy is transferred by each type of wave. can design a solution to a problem using an understanding of waves and their applications in technologies for information transfer.
Standards: Students use science and engineering practices, crosscutting concepts, and an understanding of earth and space science disciplinary core ideas to make sense of	may use evidence to describe how the Earth's rotation causes day/night cycle in Wisconsin. may recognize that gravity causes a pattern of smaller/less massive objects orbiting around larger/more massive objects.	can use a model of the Earth-Moon-Sun system to explain patterns in lunar phases. can make sense of a model that shows gravity as an attractive force between solar system and galaxy objects that increases with the mass of the interacting objects increases, or decreases as the distances between objects increases.	can develop and use an Earth-Moon-Sun model to explain solar and lunar eclipses. can ask scientific questions to clarify the role of gravity in the motions within galaxies and the solar system.	can develop and use a model of the Earth-Sun system, including the earth's atmosphere and tilt, to describe the cyclic patterns of seasons. can use computational thinking to develop and refine a qualitative or quantitative model that describes the role of gravity in the motions within galaxies and the solar system.

phenomena and solve problems.	may organize given data on solar system objects (e.g., surface features, object layers) from various Earth- and space-based instruments to allow for analysis and interpretation.	can use quantitative analyses to describe similarities and differences among solar system objects by describing features of those objects at different scales.	can use patterns in given data at varying scales to make conclusions about the identifying characteristics of different categories of solar system objects (e.g., planets, meteors, asteroids, comets) based on their features, composition, and locations within the solar system.	can describe how advances in solar system science have been made possible by improved engineering (e.g., knowledge of the evolution of the solar system from lunar exploration and space probes) and new developments in engineering made possible by advances in science (e.g., space-based telescopes using different wavelengths).
	may describe how newer rock layers sit on top of older rock layers, allowing for a relative ordering in time of the formation of the layers (assuming no disturbance of the layers).	can construct an explanation of how the fossil record can provide relative dates of change over time based on the appearance or disappearance of organisms (e.g., within fossil layers).	can obtain, evaluate, and synthesize information about local geological features to use as evidence to construct an explanation about the relative order of events and relative ages of rock units.	can make sense of a major and event in Earth's history by constructing an argument supported by evidence from specific changes in fossils and geologic features over time (e.g., volcanic eruptions, glaciations, asteroid impacts).
	may identify examples of how humans can positively and negatively impact the environment.	can identify patterns through mapping the history of natural hazards in a region and understanding related geological forces.	can construct an explanation based on evidence for how the availability of natural resources, occurrence of natural hazards, and changes in climate have influenced human activity.	can analyze and interpret data on natural hazards to forecast future catastrophic events and inform the development of technologies to mitigate their effects.

Standards: Students use science and engineering practices, crosscutting concepts, and an understanding of engineering, technology and applications of science disciplinary core ideas, to make sense of phenomena and solve problems.	may identify examples of how humans can positively and negatively impact the environment.	can define criteria and constraints for a solution to a local community problem that can be addressed through engineering.	can analyze data from tests to determine similarities and differences among several design solutions to identify the best characteristics of each that can be combined into a new, optimized solution to better meet the criteria for success.	can analyze and interpret data from multiple sources to determine patterns among several design solutions to identify the best approach to a new solution to better meet the criteria for success within the given the constraints.
	may identify limitations and possible negative consequences of the use of technologies.	can design a technology-based solution to monitor or minimize negative effects that human activities have on the environment.	can develop a model (e.g., prototype) of a proposed object, tool, or process that can generate data through iterative testing so that an optimal design can be achieved.	can refine a model to generate data for iterative testing and modification of a proposed object, tool, or process such that an optimal design can be achieved.
	may identify problems that could be solved with the help of technology.	can ask questions to clarify evidence of the factors that have caused the rise in global temperatures over the past century.	can use a model of relevant systems to define a problem and show how possible solutions may help solve a problem.	can develop and refine models of local and global systems to define a problem and show how possible solutions may help solve a local and/or global problem.
	may identify individual or societal needs and wants that can have environmental impacts.	can use scientific principles to identify relationships between human activity and likely negative environmental impacts and determine whether a particular design will mitigate those impacts.	can use evidence and scientific reasoning to design a solution that addresses the negative environmental impacts of a particular human activity.	can plan and conduct investigations to evaluate how well designs monitor and minimize negative effects that human activities have on the environment.

*Guiding statement - Not all crosscutting concepts, science and engineering practices, and disciplinary core ideas, or combinations thereof, are contained within this document. These Performance Level Descriptors are intended as examples of the type of the types of skills and understanding students would have at each performance level.

Final PLDs After Participants' Refinements

These Performance Level Descriptors are intended as examples of the type of the types of skills and understanding students would have at each performance level. Not all 4th grade crosscutting concepts, science and engineering practices, and disciplinary core ideas, or combinations thereof, are contained within this document.

	Performance Level Descriptors (PLDs)			
	Below Basic	**Basic**	**Proficient**	**Advanced**
Policy	Student demonstrates minimal understanding of and ability to apply the knowledge and skills for their grade level that are associated with college content-readiness.	Student demonstrates partial understanding of and ability to apply the knowledge and skills for their grade level that are associated with college content-readiness.	Student demonstrates understanding of and ability to apply the knowledge and skills for their grade level that are associated with college content-readiness.	Student demonstrates exemplary understanding of and ability to apply the knowledge and skills for their grade level that are associated with college content-readiness.
Standards: Students use science and engineering practices, crosscutting concepts, and an understanding of **life science** disciplinary core ideas to make sense of phenomena and solve problems.	**A student at this level** Sometimes describes primary functions of main structures in everyday plants and animals. Sometimes identifies important sense receptors within a system that supports basic animal behaviors. Sometimes uses a model to recognize that a variety of factors in the environment Can be sensed by animals (e.g., sound, light, odor, temperature).	**A student at this level** Can analyze evidence to determine if it supports a claim about the role of external structures of plants and animals in supporting survival and reproduction. Can give evidence of the sequence of events resulting in a given animal behavior (i.e., sensory input, sense receptor, brain processing, behavioral output). Can describe how data shows a cause and effect relationship between an environmental stimuli and an animal's behavior.	**A student at this level** Can provide feedback and ask questions about a claim and its supporting evidence about the role of internal and external structures of plants and animals in supporting survival, growth, behavior, and reproductive success. Can develop a model of an animal behavior (phenomenon) showing various components (i.e., sensory input, sense receptor, the brain, behavioral output) working together as a system. Can analyze an animal's behavior and describe reasonable, possible initial causes for it based on given evidence.	**A student at this level** Can create, improve, or analyze a model showing different plant or animal structures working together as parts of a system to support survival, growth, behavior, and reproductive success. Can create or improve a model of a phenomenon based on evidence to explain how sensory systems and behavioral output function to support animal survival, growth, and reproductive success. Can develop a model of sensory systems showing how animals' memories Can impact future behavior, survival, and reproduction.

Standards: Students use science and engineering practices, crosscutting concepts, and an understanding of **physical science** disciplinary core ideas to make sense of phenomena and solve problems.	Sometimes uses given evidence to describe the relative speed of an object (e.g., faster vs. slower).	Can describe that the speed of a given object is related to the energy of the object (e.g., the faster an object is moving, the more energy it possesses).	Can interpret simple quantitative data to support the idea that the speed of a given object is related to the energy of the object (e.g., the faster an object is moving, the more energy it possesses).	Can use evidence and reasoning to construct an explanation for how a given phenomenon affects the speed and related energy of an object.
	Sometimes identifies examples showing a transfer of energy.	Can describe the purpose of an investigation of a phenomenon related to energy transfer (e.g., moving objects, sound, light, heat, electric currents).	Can plan and conduct an investigation that fairly tests a phenomenon involving the transfer of energy from place to place (e.g., moving objects, sound, light, heat, electric currents).	Can obtain and evaluate evidence from multiple sources to design a solution to a problem related to the transfer of energy.
	Sometimes identifies a device that converts energy from one type to another (e.g., a light bulb to convert electrical energy into light energy).	Can identify a possible solution to a given problem involving the conversion of energy from one form to another.	Can design an evidence-based improvement to local systems (e.g. transportation, energy grid) to reduce the environmental impact of the conversion of energy from one form to another.	Can analyze and interpret evidence gathered from testing a device that converts energy from one form to another and use the results of the test to address problems in the design or improve its functioning.
	Sometimes identifies a phenomenon in which waves Can cause an object to move.	Can compare waves in phenomena in terms of amplitude and wavelength.	Can develop a model of a phenomenon related to wave behavior that describes wave amplitude, wavelength, or motion of objects (e.g., wave models of loud vs soft sound).	Can design a solution to transfer information over a distance, comparing methods using waves (e.g. sound, light) to other methods using patterns in addressing particular criteria and constraints.

Standards: Students use science and engineering practices, crosscutting concepts, and an understanding of **earth and space science** disciplinary core ideas to make sense of phenomena and solve problems.	Sometimes uses fossil evidence to infer a basic feature of what an environment used to be like (e.g. marine fossils indicate that in the past a landscape was covered in water).	Can ask cause and effect questions about rock layers, fossils, and geological features that could lead to productive investigations about these phenomena.	Can use a diagram of rock layers and fossils, as well as other geological features such as Canyons, to help explain how an environment has changed over time.	Can design a solution that addresses particular criteria and constraints, to prevent water, ice, or wind from impacting a particular landscape.
	Sometimes recognizes which type of maps Can be used to best locate different land and water features on Earth.	Can use evidence from given topographic maps to identify various Earth features (e.g., mountain ranges, ocean trenches, ocean floor structures, fault lines, volcanoes).	Can use patterns in a map as evidence to explain where geologic processes are likely to occur (e.g., earthquakes, erosion, volcanoes).	Can evaluate a map of a fantasy land to describe where it does or does not show reasonable patterns of geologic features.
	Sometimes identifies examples of natural resources that humans use for energy.	Can use given evidence to identify cause and effect relationships between the use of a natural resource and its likely impact on the environment.	Can analyze and interpret patterns in evidence to describe that energy and fuels are derived from natural resources (e.g., fossil fuels, solar, wind, water) and their uses Can have various effects on the environment.	Can design a solution based on evidence from multiple sources to a problem related to the use of natural resources and their effects on the environment.
	Sometimes identifies possible negative impacts to humans from a natural Earth process (e.g., an earthquake, volcano, flood, landslide).	Can use evidence to design a possible solution to reduce the impacts of natural Earth processes on humans.	Can use evidence to describe how one Earth process Can have a greater negative impact compared to another Earth process in a given area or region.	Can use evidence to generate multiple possible solutions to reduce the impacts of natural Earth processes on humans and then evaluate which one best addresses criteria and constraints.

Standards: Students use science and engineering practices, crosscutting concepts, and an understanding of **engineering, technology and applications of science** disciplinary core ideas, to make sense of phenomena and solve problems.	Sometimes uses scientific understanding to define criteria for a simple design problem that includes responding to a human need or want. Sometimes uses scientific understanding to identify which tools and methods could be used to collect data for a given investigation.	Can use scientific understanding to define a problem related to local phenomena that Can be solved with the development of a new or improved object, tool, process, or system. Can use scientific understanding to make conclusions related to how well a model works or a prototype performs against given criteria and constraints.	Can use given scientific information and information about an everyday situation or phenomenon to design a solution to a problem that includes responding to needs or wants of humans. Can conduct fair tests in which variables are controlled and possible failure points are considered when designing a prototype to solve problems related to a particular local phenomenon.	Can use scientific and engineering practices and understanding to evaluate multiple possible solutions to a problem and describe how well the solution addresses the constraints within which the problem must be solved. Can use scientific and engineering practices and understanding to evaluate a range of new technologies to determine how they may change how people live and interact with each other.

Reminder: These Performance Level Descriptors are intended as examples of the type of the types of skills and understanding students would have at each performance level. Not all 4th grade crosscutting concepts, science and engineering practices, and disciplinary core ideas, or combinations thereof, are contained within this document.

Grade 8

Wisconsin Science PLDs

These Performance Level Descriptors are intended as examples of the type of the types of skills and understanding students would have at each performance level. Not all 4th grade crosscutting concepts, science and engineering practices, and disciplinary core ideas, or combinations thereof, are contained within this document.

	Performance Level Descriptors (PLDs)			
	Below Basic	**Basic**	**Proficient**	**Advanced**
Policy	Student demonstrates minimal understanding of and ability to apply the knowledge and skills for their grade level that are associated with college content-readiness.	Student demonstrates partial understanding of and ability to apply the knowledge and skills for their grade level that are associated with college content-readiness.	Student demonstrates adequate understanding of and ability to apply the knowledge and skills for their grade level that are associated with college content-readiness.	Student demonstrates thorough understanding of and ability to apply the knowledge and skills for their grade level that are associated with college content-readiness.
Standards: Students use science and engineering practices, crosscutting concepts, and an understanding of **life science** disciplinary core ideas, to make sense of phenomena and solve problems.	**A student at this level** Sometimes makes observations to use as evidence that an object is a living or non-living thing.	**A student at this level** Can collect and analyze data during an investigation to determine whether existence of cells Can be a distinguishing characteristic of living things.	**A student at this level** Can conduct an investigation to provide evidence that tissues and organs are made of cells with specialized functions in the body system.	**A student at this level** Can use arguments based on scientific reasoning and evidence from multiple sources to support the idea that a body is a system of interacting subsystems composed of various cells.
	Sometimes identifies evidence supporting the claim that some animal behaviors Can affect the survival of another species.	Can use patterns to predict how a certain animal behavior will likely affect the reproductive success of multiple other species.	Can develop a model showing how genetic and/or environmental factors Can affect an organism's growth and reproductive success.	Can analyze the validity and reliability of given evidence to solve problems related to biological and environmental factors affecting organisms.
	Sometimes asks questions that could guide an investigation into differences between how plants and animals obtain food.	Can use a scientific model to describe how food molecules are rearranged through chemical reactions to form new molecules that support growth and/or to release energy as this matter moves through an organism.	Can construct a scientific explanation based on evidence or use a model to describe the role of photosynthesis in the cycling of matter and flow of energy into and out of ecosystems.	Can develop a model to track the changes of energy and matter into, out of, and within a system or ecosystem, noting the limitations of the model.
	Sometimes engages in the understanding of phenomena associated with sense receptors responding to different inputs (i.e., mechanical, chemical, electromagnetic).	Can construct an explanation, based on cause and effect evidence, of a phenomenon that results in sensory receptors sending signals to the brain.	Can develop and use a model to show sensory receptors responding to stimuli by sending messages to the brain for immediate behavior or storage as memories.	Can synthesize information that provides evidence of causal relationships between information received by sensory receptors and behavior, at various time scales.

	Sometimes describes how a change in the amount or availability of a natural resource Can result in changes in a population of organisms.	Can analyze and interpret graphical displays of data (e.g., graphs, charts) to provide evidence of the relationships between resource availability and organism abundance.	Can make sense of phenomena related to patterns of interactions among organisms to help explain relationships within ecosystems.	Can use reasoning and evidence in an explanation of interactions within an ecosystem and predict future interactions based on patterns in that evidence.
	Sometimes identifies examples of producers, consumers, and/or decomposers within an ecosystem and what effects they have in that system.	Can construct a model of how matter and energy are transferred between producers, consumers, and decomposers within an ecosystem.	Can construct an evidence-based argument describing the conservation of matter within and flow of energy into and out of an ecosystem.	Can develop and use a model to explain the transfer of matter (atoms) and energy between living and nonliving parts of the ecosystem at various levels within the system, and define the boundaries of that system.
	Sometimes recognizes that structural changes to genes (i.e., mutations) may result in observable changes in organisms.	Can use a model (i.e., Punnett square) to show that more genetic variation occurs in organisms that reproduce sexually compared to organisms that reproduce asexually.	Can use evidence related to a phenomenon (e.g., pictures of a litter of puppies or diagrams of fern spores and amoeba) to predict the differences in genetic variation resulting from sexual and asexual reproduction.	Can use a model of a protein to explain how changes to protein structure Can lead to changes in its function that may cause beneficial, neutral, or harmful changes in the structures or functions of organisms.
	Sometimes identifies a similar anatomical feature shared by organisms that suggests they are likely to be more closely related than other organisms that do not share the similar anatomical feature.	Can construct an argument based on evidence that some organisms sharing a pattern of anatomical features are likely to be more closely related than organisms that do not share a pattern of anatomical features.	Can apply scientific ideas and evidence to construct an explanation for the anatomical similarities and differences between modern and fossil organisms that describes evolution over time.	Can connect multiple sources of evidence comparing modern living animals and fossilized animals to support an argument for past connections of multiple lines of descent of different species.
	Sometimes recognizes that data shows that some organisms will better survive and reproduce from generation to generation due to traits that are	Can use patterns as evidence to support claims that some traits have advantages that make it more probable that an organism will be able to survive and	Can evaluate different explanations about natural selection within a population of organisms to determine which is better supported by evidence.	Can analyze data to trace the increase or decrease of particular traits in a population over time and make claims about how those changes were likely

	advantageous in a specific environment.	reproduce in a specific environment.		the result of particular historical phenomena that changed their environment.
Standards: Students use science and engineering practices, crosscutting concepts, and an understanding of **physical science** disciplinary core ideas, to make sense of phenomena and solve problems.	Sometimes uses simple models of different atoms and molecules as support for why different substances have different properties.	Can plan experiments to identify different substances based on their characteristic physical properties (e.g., density, melting point).	Can analyze data to identify changes in physical and chemical properties of substances before and after an interaction to make a claim about whether a chemical reaction occurred.	Can use simple models to provide evidence for an argument that a change in properties of substances Can be related to the rearrangement of atoms in a chemical reaction.
	Sometimes makes sense of a phenomenon where a collision occurs by describing how a force was exerted by the first object on the second object.	Can use evidence to model the components within the system that are involved in a collision between two objects.	Can address a given problem involving a collision of two objects by detailing a process or system that helps solve the problem.	Can evaluate data on two designs for solving a problem involving a collision of two objects to determine which design better meets the criteria and constraints of the situation.
	Sometimes uses information from a model to recognize that an object subjected to balanced forces does not change its motion and an object subjected to unbalanced forces does change its motion.	Can conduct an investigation involving the change in motion of an object and gather evidence identifying various factors affecting the object's motion.	Can use mathematical thinking to explain how changes in an object's motion Can be due to the degree of balanced or unbalanced forces acting on the object as well as the mass of the object.	Can plan an investigation to provide evidence that the change in an object's motion depends on specific variables such as initial motion of the object, total forces acting on the object, and mass of the object.
	Sometimes identifies the relative magnitude and direction of the forces between objects in a given system.	Can identify evidence from a given phenomenon to support the idea that gravitational forces are attractive and mass dependent.	Can construct an argument based on evidence that gravitational interactions are attractive and dependent on the masses of interacting objects.	Can evaluate evidence on how well it supports the idea that gravitational forces are attractive and mass dependent, identifying strengths and weaknesses of the evidence, and including possible alternative interpretations of it.
	Sometimes recognizes that energy increases if either the mass or the speed of the object	Can interpret graphical displays of data to describe the relationships of kinetic energy to	Can construct graphical displays or other models to communicate the idea that the	Can design a solution to a problem identified by analyzing data from multiple sources that

	increases and that energy decreases if either the mass or the speed of the object decreases. Sometimes identifies that the interactions of two objects at a distance Can cause a transfer of energy between the objects. Sometimes identifies examples of mechanical waves that need a medium through which they are transmitted. Sometimes identifies examples of waves interacting with materials by being reflected, absorbed, or transmitted.	the mass of an object and to the speed of an object. Can analyze and interpret data to provide evidence that as the relative position of two objects changes, the potential energy of the system changes (e.g. an object higher off the ground has more gravitational potential energy). Can describe a simple model for waves (repeating pattern) that includes how the amplitude of a wave is related to the energy in a wave. Can use a given model to make sense of given phenomena involving reflection, absorption, or transmission properties of different materials for light and matter waves.	mass and speed of an object affect its kinetic energy. Can develop a model to describe that when the arrangement of objects interacting at a distance changes, different amounts of potential energy are stored in the system. Can use mathematical representations to evaluate a model for waves on whether it include evidence for how the amplitude of a wave is related to the energy in a wave. Can develop and use a model about phenomena involving light and/or matter waves to describe the differences between how light and matter waves interact with different materials.	show the relationships of kinetic energy to the mass and speed of an object. Can plan and conduct an investigation to make sense of a given phenomenon involving two objects interacting at a distance. Can develop a model to compare mechanical and electromagnetic waves and explain similarities and differences in how energy is transferred by each type of wave. Can design a solution to a problem using an understanding of waves, how they interact with different materials, and their applications in technologies for information transfer.
Standards: Students use science and engineering practices, crosscutting concepts, and an understanding of **earth and space**	Sometimes uses evidence to describe how the Earth's rotation causes day/night cycle in Wisconsin. Sometimes recognizes that gravity causes a pattern of smaller and less massive objects	Can use a model of the Earth-Moon-Sun system to explain patterns in lunar phases. Can use a model to help describe gravity as an attractive force between solar system and galaxy	Can develop and use an Earth-Moon-Sun model to explain solar and lunar eclipses. Can ask scientific questions to clarify the role of gravity in the motions within galaxies and	Can develop and use a model of the Earth-Sun system, including the earth's atmosphere and tilt, to describe the cyclic patterns of seasons. Can use computational thinking to develop and refine a qualitative or quantitative model

science disciplinary core ideas to make sense of phenomena and solve problems.	orbiting around larger and more massive objects.	objects that increases as the mass of the interacting objects increases, or decreases as the distances between objects increases.	the solar system.	that describes the role of gravity in the motions within galaxies and the solar system.
	Sometimes organizes given data on solar system objects (e.g., surface features, object layers) from various Earth- and space-based instruments to allow for analysis and interpretation.	Can use quantitative analyses to describe similarities and differences among solar system objects by describing features of those objects at different scales.	Can use patterns in given data at varying scales to make conclusions about the identifying characteristics of different categories of solar system objects (e.g., planets, meteors, asteroids, comets) based on their features, composition, and locations within the solar system.	Can describe how advances in solar system science have been made possible by improved engineering (e.g., knowledge of the evolution of the solar system from lunar exploration and space probes) and new developments in engineering made possible by advances in science (e.g., space-based telescopes using different wavelengths).
	Sometimes describes how newer rock layers sit on top of older rock layers, allowing for a relative ordering in time of the formation of the layers (assuming no disturbance of the layers).	Can construct an explanation of how the fossil record Can provide relative dates of change over time based on the appearance or disappearance of organisms (e.g., within fossil layers).	Can obtain, evaluate, and synthesize information about local geological features to use as evidence to construct an explanation about the relative order of events and relative ages of rock units.	Can make sense of a major event in Earth's history by constructing an argument supported by evidence from specific changes in fossils and geologic features over time (e.g., volcanic eruptions, glaciations, asteroid impacts).
	Sometimes identifies examples of how humans Can protect themselves from natural disasters.	Can identify patterns by mapping the history of natural hazards in a region and understanding related geological forces.	Can construct an explanation based on evidence for how the availability of natural resources, occurrence of natural hazards, and changes in climate have influenced human activity.	Can analyze and interpret data on natural hazards to forecast future catastrophic events and inform the development of technologies to mitigate their effects.

Standards: Students use science and engineering practices, crosscutting concepts, and an understanding of **engineering, technology and applications of science** disciplinary core ideas, to make sense of phenomena and solve problems.	Sometimes identifies examples of how humans Can positively and negatively impact the environment.	Can define criteria and constraints for a solution to a local community problem that Can be addressed through engineering.	Can analyze data from tests to determine similarities and differences among several design solutions for solving a community problem to identify the best characteristics of each that Can be combined into a new, optimized solution to better meet the criteria for success.	Can analyze and interpret data from multiple sources to determine patterns among several design solutions to identify the best approach to a new solution to a community problem to better meet the criteria for success within the given the constraints.
	Sometimes identifies limitations and possible negative consequences of the use of technologies.	Can design a technology-based solution to monitor or minimize negative effects that human activities have on the environment.	Can develop a model (i.e., prototype) of a proposed object, tool, or process that Can generate data through iterative testing so that an optimal design Can be achieved.	Can refine a prototype to generate data for iterative testing and modification of a proposed object, tool, or process such that an optimal design Can be achieved.
	Sometimes identifies problems that could be solved with the help of technology.	Can ask questions to clarify evidence of which solution best solves a local problem.	Can use a model of relevant systems to define a problem and show how possible solutions may help solve a problem.	Can develop and refine models of local and global systems to define a problem and show how possible solutions may help solve a local and/or global problem.
	Sometimes identifies individual or societal needs and wants that Can have environmental impacts.	Can use scientific principles to identify relationships between human activity and likely negative environmental impacts and determine whether a particular design will mitigate those impacts.	Can use evidence and scientific reasoning to design a solution that addresses the negative environmental impacts of a particular human activity.	Can plan and conduct investigations to evaluate how well designs monitor and minimize negative effects that human activities have on the environment.

Reminder: These Performance Level Descriptors are intended as examples of the type of the types of skills and understanding students would have at each performance level. Not all crosscutting concepts, science and engineering practices, and disciplinary core ideas, or combinations thereof, are contained within this document.

Appendix H

Participant Evaluations of the Workshop

Participant Survey

Thank you for completing this survey. We gather this information to demonstrate the level of expertise of the participants at our standard setting events. When done, please return your survey to a facilitator.

1. What is your gender?

- O Female
- O Male

2. What is your ethnicity?

- O American Indian/Alaska Native
- O Asian
- O Hawaiian or Pacific Islander
- O Black
- O Hispanic
- O Mixed (Two or more races)
- O Caucasian

3. What is your current assignment?

- O Classroom teacher
- O Educator, non-teacher
- O Higher education
- O Other (please describe):

____

4. What is your work setting?

- O Urban
- O Suburban
- O Rural

5. How many years, in total, have you been teaching?

- O Fewer than 5 years
- O 5–10 years
- O 11–15 years
- O 16–20 years
- O 21–25 years
- O More than 25 years

6. What is your highest level of education?

- O High school diploma
- O Bachelor's degree
- O Master's degree
- O Doctoral degree

7. What is the name of your school district?

____

8. Which of these groups do you have experience teaching?

- O Special education (in a self-contained classroom)
- O Special education (in a mainstream classroom)
- O English language learners
- O Vocational education
- O Alternative education

9. In what group did you participate in this workshop?

- O Science, Grade 4
- O Science, Grade 8

10. In which grades and subjects (and for how many years) have you taught?

Example: Grade 8 Math (5 years)

____

____

____

____

____

11. What professional development have you taken or experienced in the last two years?

____

____

____

____

____

____

1. What is your gender?

Response	Frequency	Percent	Mean: 1.19
Female	22	81.48	
Male	5	18.52	

2. What is your ethnicity?

Response	Frequency	Percent	Mean: 6.67
American Indian/Alaska Native	0	0.00	
Asian	1	3.70	
Hawaiian or Pacific Islander	0	0.00	
Black	0	0.00	
Hispanic	2	7.41	
Mixed -Two or more races	0	0.00	
Caucasian	24	88.89	

3. What is your current assignment?

Response	Frequency	Percent	Mean: 1.85
Classroom teacher	14	51.85	
Educator non-teacher	6	22.22	
Higher education	2	7.41	
Other	4	14.81	
No Response	1	3.70	

4. What is your work setting?

Response	Frequency	Percent	Mean: 2.07
Urban	10	37.04	
Suburban	5	18.52	
Rural	12	44.44	

5. How many years, in total, have you been teaching?

Response	Frequency	Percent	Mean: 3.41
Fewer than 5 years	1	3.70	
5-10 years	8	29.63	
11-15 years	7	25.93	
16-20 years	5	18.52	
21-25 years	2	7.41	
More than 25 years	4	14.81	

6. What is your highest level of education?

Response	Frequency	Percent	Mean: 2.93
High school diploma	0	0.00	
Bachelor's degree	3	11.11	
Master's degree	23	85.19	
Doctoral degree	1	3.70	

8. Which of these groups do you have experience teaching?

Response	Frequency	Percent	Mean: -
Special education -in a self-contained classroom	3	11.11	
Special education -in a mainstream classroom	21	77.78	
English language learners	17	62.96	
Vocational education	2	7.41	
Alternative education	4	14.81	
No Response	3	11.11	

9. In what group did you participate in this workshop?

Response	Frequency	Percent	Mean: 1.56
Science Grade 4	11	40.74	
Science Grade 8	14	51.85	
No Response	2	7.41	

Wisconsin Forward Exam 2019 Standard Setting Evaluation *for* Grades 4 and 8 Science

The purpose of this evaluation is to help document the process used to recommend performance standards. Your opinions and comments are important, as they will provide a basis for judging the quality of this process.

Please do not put your name on this form. While we need the information to examine the success of the various steps in the process, we want your comments to remain anonymous. This information will be reported only in the aggregate. **When you have completed the evaluation, please give it to a facilitator. *Thank you!***

	Part 1: ABOUT THE STANDARD SETTING Please *consider the statements below* and mark the level of agreement or disagreement you have with each statement. Please bubble ***only one*** of the five options for each statement.	**Strongly Disagree**	**Disagree**	**Neutral**	**Agree**	**Strongly Agree**
Training & PLDs	1. The training provided a clear description of the workshop goals.	O	O	O	O	O
	2. The training session leader clearly explained the Bookmark Procedure.	O	O	O	O	O
	3. The training session leader clearly explained the materials used in the bookmark process.	O	O	O	O	O
	4. The training addressed many of my questions and concerns.	O	O	O	O	O
	5. The practice exercises were useful.	O	O	O	O	O
	6. The opening session provided a clear overview of the standard setting process.	O	O	O	O	O
	7. My role in the standard setting was well described.	O	O	O	O	O
	8. After the training, I felt confident I was prepared to complete the standard setting task.	O	O	O	O	O
	9. The performance level descriptors (PLDs) were clear.	O	O	O	O	O
	10. Adequate information was provided regarding the PLDs.	O	O	O	O	O
	11. Enough time was provided to read and understand the PLDs.	O	O	O	O	O
	12. The PLDs communicate a reasonable profile of students' performance at each level.	O	O	O	O	O

	Please indicate your opinion regarding the usefulness of the following **materials** used. Please bubble ***only one*** of the four options for each material.	**Not Useful**	**Somewhat Useful**	**Useful**	**Very Useful**
Materials	13. Performance level descriptors (PLDs)	O	O	O	O
	14. Ordered item booklets (OIBs)	O	O	O	O
	15. Operational test books	O	O	O	O
	16. Item information sheets / item maps	O	O	O	O
	17. Item separation charts	O	O	O	O
	18. Impact data	O	O	O	O

	Please indicate the extent of your satisfaction with the following **roles**. Please bubble ***only one*** of the four options for each role.	**Not Satisfied**	**Partially Satisfied**	**Satisfied**	**Very Satisfied**
Roles	19. DRC trainer	O	O	O	O
	20. DRC facilitator(s)	O	O	O	O
	21. DRC content specialist	O	O	O	O
	22. Other DRC staff	O	O	O	O

	Please indicate your opinion regarding the amount of **time allotted** for each activity. Please bubble ***only one*** of the three options for each activity.	**Too Little Time**	**About Right**	**Too Much Time**
Time Allotted	23. Training	O	O	O
	24. PLD discussion	O	O	O
	25. Round 1 bookmarks	O	O	O
	26. Discussion after Round 1	O	O	O
	27. Round 2 bookmarks	O	O	O
	28. Discussion after Round 2	O	O	O
	29. Round 3 bookmarks	O	O	O
	30. Discussion of final recommendations	O	O	O

	Please *consider the statements below* and mark the level of agreement or disagreement you have with each statement. Please bubble ***only one*** of the five options for each statement.	**Strongly Disagree**	**Disagree**	**Neutral**	**Agree**	**Strongly Agree**
Bookmarks	31. I understood how to make my bookmarks.	O	O	O	O	O
	32. I had adequate time to make my bookmarks.	O	O	O	O	O
	33. I considered the threshold students when making my bookmarks.	O	O	O	O	O
	34. There was adequate time provided for discussion.	O	O	O	O	O
	35. Discussing the threshold students helped me make my bookmarks.	O	O	O	O	O
	36. I considered the standards when I placed my bookmarks.	O	O	O	O	O
Overall	37. Overall, I believe my opinions were considered and valued by my group.	O	O	O	O	O
	38. My group's work was reflected in the presentation of recommendations across grades.	O	O	O	O	O
	39. The group leader in my breakout room provided clear instructions.	O	O	O	O	O
	40. Overall, I valued the workshop as a professional development experience.	O	O	O	O	O
Rooms	41. The food and service at the facility met my expectations.	O	O	O	O	O
	42. The breakout rooms had appropriate accommodations to facilitate our work.	O	O	O	O	O

Grade	Please indicate the level of confidence you had in **recommending the cut scores** for each performance level. Please bubble ***only one*** of the four options for each cut score. ***Important:*** **Only complete this section for the grade(s) you worked on.**	Not Confident	Partially Confident	Confident	Very Confident
4	43. Basic cut score	O	O	O	O
	44. Proficient cut score	O	O	O	O
	45. Advanced cut score	O	O	O	O
8	46. Basic cut score	O	O	O	O
	47. Proficient cut score	O	O	O	O
	48. Advanced cut score	O	O	O	O

Part 2: ABOUT YOU

49. In which group did you work?

O Science Grade 4

O Science Grade 8

Part 3 below

Part 3: YOUR TURN

In this box, please feel free to add comments about any of your responses, make suggestions to improve future workshops, or tell us what you liked and did not like about this workshop. ***Thank you!***

1. The training provided a clear description of the workshop goals.

Response	Frequency	Percent	Mean: 4.58
Strongly Disagree	1	3.85	
Disagree	0	0.00	
Neutral	0	0.00	
Agree	7	26.92	
Strongly Agree	18	69.23	

2. The training session leader clearly explained the Bookmark Procedure.

Response	Frequency	Percent	Mean: 4.62
Strongly Disagree	1	3.85	
Disagree	0	0.00	
Neutral	0	0.00	
Agree	6	23.08	
Strongly Agree	19	73.08	

3. The training session leader clearly explained the materials used in the bookmark process.

Response	Frequency	Percent	Mean: 4.69
Strongly Disagree	1	3.85	
Disagree	0	0.00	
Neutral	0	0.00	
Agree	4	15.38	
Strongly Agree	21	80.77	

4. The training addressed many of my questions and concerns.

Response	Frequency	Percent	Mean: 4.35
Strongly Disagree	1	3.85	
Disagree	0	0.00	
Neutral	1	3.85	
Agree	11	42.31	
Strongly Agree	13	50.00	

5. The practice exercises were useful.

Response	Frequency	Percent	Mean: 4.35
Strongly Disagree	1	3.85	
Disagree	0	0.00	
Neutral	2	7.69	
Agree	9	34.62	
Strongly Agree	14	53.85	

6. The opening session provided a clear overview of the standard setting process.

Response	Frequency	Percent	Mean: 4.50
Strongly Disagree	1	3.85	
Disagree	0	0.00	
Neutral	0	0.00	
Agree	9	34.62	
Strongly Agree	16	61.54	

7. My role in the standard setting was well described.

Response	Frequency	Percent	Mean: 4.50
Strongly Disagree	1	3.85	
Disagree	0	0.00	
Neutral	0	0.00	
Agree	9	34.62	
Strongly Agree	16	61.54	

8. After the training, I felt confident I was prepared to complete the standard setting task.

Response	Frequency	Percent	Mean: 4.38
Strongly Disagree	1	3.85	
Disagree	0	0.00	
Neutral	0	0.00	
Agree	12	46.15	
Strongly Agree	13	50.00	

9. The performance level descriptors (PLDs) were clear.

Response	Frequency	Percent	Mean: 3.88
Strongly Disagree	1	3.85	
Disagree	1	3.85	
Neutral	5	19.23	
Agree	12	46.15	
Strongly Agree	7	26.92	

10. Adequate information was provided regarding the PLDs.

Response	Frequency	Percent	Mean: 4.23
Strongly Disagree	1	3.85	
Disagree	0	0.00	
Neutral	1	3.85	
Agree	14	53.85	
Strongly Agree	10	38.46	

11. Enough time was provided to read and understand the PLDs.

Response	Frequency	Percent	Mean: 4.27
Strongly Disagree	1	3.85	
Disagree	1	3.85	
Neutral	3	11.54	
Agree	6	23.08	
Strongly Agree	15	57.69	

12. The PLDs communicate a reasonable profile of students' performance at each level.

Response	Frequency	Percent	Mean: 4.04
Strongly Disagree	1	3.85	
Disagree	0	0.00	
Neutral	3	11.54	
Agree	15	57.69	
Strongly Agree	7	26.92	

13. Performance level descriptors (PLDs)

Response	Frequency	Percent	Mean: 3.54
Not Useful	0	0.00	
Somewhat Useful	2	7.69	
Useful	8	30.77	
Very Useful	16	61.54	

14. Ordered item booklets (OIBs)

Response	Frequency	Percent	Mean: 3.85
Not Useful	0	0.00	
Somewhat Useful	1	3.85	
Useful	2	7.69	
Very Useful	23	88.46	

15. Operational test books

Response	Frequency	Percent	Mean: 3.45
Not Useful	0	0.00	
Somewhat Useful	3	11.54	
Useful	5	19.23	
Very Useful	12	46.15	
No Response	6	23.08	

16. Item information sheets/item maps

Response	Frequency	Percent	Mean: 3.62
Not Useful	0	0.00	
Somewhat Useful	1	3.85	
Useful	8	30.77	
Very Useful	17	65.38	

17. Item separation charts

Response	Frequency	Percent	Mean: 3.46
Not Useful	0	0.00	
Somewhat Useful	1	3.85	
Useful	12	46.15	
Very Useful	13	50.00	

18. Impact data

Response	Frequency	Percent	Mean: 3.69
Not Useful	0	0.00	
Somewhat Useful	1	3.85	
Useful	6	23.08	
Very Useful	19	73.08	

19. DRC trainer

Response	Frequency	Percent	Mean: 3.85
Not Satisfied	0	0.00	
Partially Satisfied	0	0.00	
Satisfied	4	15.38	
Very Satisfied	22	84.62	

20. DRC facilitator(s)

Response	Frequency	Percent	Mean: 3.77
Not Satisfied	0	0.00	
Partially Satisfied	0	0.00	
Satisfied	6	23.08	
Very Satisfied	20	76.92	

21. DRC content specialist

Response	Frequency	Percent	Mean: 3.69
Not Satisfied	0	0.00	
Partially Satisfied	0	0.00	
Satisfied	8	30.77	
Very Satisfied	18	69.23	

22. Other DRC staff

Response	Frequency	Percent	Mean: 3.60
Not Satisfied	0	0.00	
Partially Satisfied	0	0.00	
Satisfied	10	38.46	
Very Satisfied	15	57.69	
No Response	1	3.85	

23. Training

Response	Frequency	Percent	Mean: 2.12
Too Little Time	1	3.85	
About Right	21	80.77	
Too Much Time	4	15.38	

24. PLD discussion

Response	Frequency	Percent	Mean: 2.35
Too Little Time	2	7.69	
About Right	13	50.00	
Too Much Time	11	42.31	

25. Round 1 bookmarks

Response	Frequency	Percent	Mean: 1.92
Too Little Time	3	11.54	
About Right	22	84.62	
Too Much Time	1	3.85	

26. Discussion after Round 1

Response	Frequency	Percent	Mean: 1.92
Too Little Time	2	7.69	
About Right	24	92.31	
Too Much Time	0	0.00	

27. Round 2 bookmarks

Response	Frequency	Percent	Mean: 2.08
Too Little Time	0	0.00	
About Right	23	88.46	
Too Much Time	2	7.69	
No Response	1	3.85	

28. Discussion after Round 2

Response	Frequency	Percent	Mean: 2.04
Too Little Time	2	7.69	
About Right	21	80.77	
Too Much Time	3	11.54	

29. Round 3 bookmarks

Response	Frequency	Percent	Mean: 2.08
Too Little Time	0	0.00	
About Right	24	92.31	
Too Much Time	2	7.69	

30. Discussion of final recommendations

Response	Frequency	Percent	Mean: 2.05
Too Little Time	0	0.00	
About Right	21	80.77	
Too Much Time	1	3.85	
No Response	4	15.38	

31. I understood how to make my bookmarks.

Response	Frequency	Percent	Mean: 4.62
Strongly Disagree	0	0.00	
Disagree	0	0.00	
Neutral	0	0.00	
Agree	10	38.46	
Strongly Agree	16	61.54	

32. I had adequate time to make my bookmarks.

Response	Frequency	Percent	Mean: 4.69
Strongly Disagree	0	0.00	
Disagree	0	0.00	
Neutral	1	3.85	
Agree	6	23.08	
Strongly Agree	19	73.08	

33. I considered the threshold students when making my bookmarks.

Response	Frequency	Percent	Mean: 4.73
Strongly Disagree	0	0.00	
Disagree	0	0.00	
Neutral	0	0.00	
Agree	7	26.92	
Strongly Agree	19	73.08	

34. There was adequate time provided for discussion.

Response	Frequency	Percent	Mean: 4.42
Strongly Disagree	0	0.00	
Disagree	2	7.69	
Neutral	1	3.85	
Agree	7	26.92	
Strongly Agree	16	61.54	

35. Discussing the threshold students helped me make my bookmarks.

Response	Frequency	Percent	Mean: 4.69
Strongly Disagree	0	0.00	
Disagree	0	0.00	
Neutral	0	0.00	
Agree	8	30.77	
Strongly Agree	18	69.23	

36. I considered the standards when I placed my bookmarks.

Response	Frequency	Percent	Mean: 4.73
Strongly Disagree	0	0.00	
Disagree	0	0.00	
Neutral	0	0.00	
Agree	7	26.92	
Strongly Agree	19	73.08	

37. Overall, I believe my opinions were considered and valued by my group.

Response	Frequency	Percent	Mean: 4.85
Strongly Disagree	0	0.00	
Disagree	0	0.00	
Neutral	0	0.00	
Agree	4	15.38	
Strongly Agree	22	84.62	

38. My group's work was reflected in the presentation of recommendations across grades.

Response	Frequency	Percent	Mean: 4.73
Strongly Disagree	0	0.00	
Disagree	0	0.00	
Neutral	0	0.00	
Agree	7	26.92	
Strongly Agree	19	73.08	

39. The group leader in my breakout room provided clear instructions.

Response	Frequency	Percent	Mean: 4.73
Strongly Disagree	0	0.00	
Disagree	0	0.00	
Neutral	1	3.85	
Agree	5	19.23	
Strongly Agree	20	76.92	

40. Overall, I valued the workshop as a professional development experience.

Response	Frequency	Percent	Mean: 4.81
Strongly Disagree	0	0.00	
Disagree	0	0.00	
Neutral	0	0.00	
Agree	5	19.23	
Strongly Agree	21	80.77	

41. The food and service at the facility met my expectations.

Response	Frequency	Percent	Mean: 4.85
Strongly Disagree	0	0.00	
Disagree	0	0.00	
Neutral	0	0.00	
Agree	4	15.38	
Strongly Agree	22	84.62	

42. The breakout rooms had appropriate accommodations to facilitate our work.

Response	Frequency	Percent	Mean: 4.46
Strongly Disagree	0	0.00	
Disagree	2	7.69	
Neutral	2	7.69	
Agree	4	15.38	
Strongly Agree	18	69.23	

43. Grade 4 Basic cut score

Response	Frequency	Percent	Mean: 3.64
Not Confident	0	0.00	
Partially Confident	0	0.00	
Confident	5	19.23	
Very Confident	9	34.62	
No Response	12	46.15	

44. Grade 4 Proficient cut score

Response	Frequency	Percent	Mean: 3.43
Not Confident	0	0.00	
Partially Confident	0	0.00	
Confident	8	30.77	
Very Confident	6	23.08	
No Response	12	46.15	

45. Grade 4 Advanced cut score

Response	Frequency	Percent	Mean: 3.36
Not Confident	0	0.00	
Partially Confident	2	7.69	
Confident	5	19.23	
Very Confident	7	26.92	
No Response	12	46.15	

46. Grade 8 Basic cut score

Response	Frequency	Percent	Mean: 3.29
Not Confident	0	0.00	
Partially Confident	0	0.00	
Confident	10	38.46	
Very Confident	4	15.38	
No Response	11	42.31	
Multiple	1	3.85	

47. Grade 8 Proficient cut score

Response	Frequency	Percent	Mean: 2.87
Not Confident	0	0.00	
Partially Confident	3	11.54	
Confident	11	42.31	
Very Confident	1	3.85	
No Response	11	42.31	

48. Grade 8 Advanced cut score

Response	Frequency	Percent	Mean: 3.07
Not Confident	0	0.00	
Partially Confident	2	7.69	
Confident	10	38.46	
Very Confident	3	11.54	
No Response	11	42.31	

49. In which group did you work?

Response	Frequency	Percent	Mean: 1.50
Science Grade 4	12	46.15	
Science Grade 8	12	46.15	
No Response	2	7.69	