New York State Elementary-level (Grade 5) and Intermediate-level (Grade 8) Science Tests

Standard Setting Report

Prepared for the New York State Education Department
by Pearson

October 2024

Copyright

Contents

List of Tables

List of Figures

Executive Summary

A standard setting meeting was conducted for the New York State Elementary-level (Grade 5) and Intermediate-level (Grade 8) Science Tests. The primary goal for this standard setting was to recommend cut scores that operationally define four performance levels: Level 1, Level 2, Level 3, and Level 4. The performance level designations are used by local, state, and federal accountability programs and are central to communicating with parents, teachers, and the public. This document provides a detailed description of the activities held at the meeting.

The standard setting meeting was held July 10–11, 2024, in Troy, New York. Panelists were trained in and followed the Modified Yes/No Angoff standard setting procedure, resulting in cut score recommendations that were brought to the New York State Education Department (NYSED).

In this report, panelists, materials, methodologies, and results are presented for the New York State Grade 5 and Grade 8 Science Tests standard setting.

1. Grades 5 and 8 Science Tests

The Office of State Assessment (OSA) at NYSED worked with NYS educators to develop the Grade 5 and Grade 8 Science Tests. The tests are designed to measure students' knowledge and understanding of the *NYS Grades 3–8 Science Learning Standards*, first adopted by NYS in 2016, which is part of the transition to the *Next Generation Science Standards* (NGSS) nationally. The Grade 5 Science Test assesses science standards for Grades 3–5, and the Grade 8 Science Test assesses science standards for Grades 6–8.

The new Grade 5 and Grade 8 Science Tests were first administered in spring 2024, and the standard setting activities used the test forms and data from this administration. Both tests are organized through four scientific domains that define the content to be covered on the exams. Table 1.1 presents the four domains along with the estimated percent of points for each domain. The tests are comprised of 1-point multiple-choice items along with 1-point constructed-response and 1-point technology enhanced items (TEIs). The TEIs include some graphing items, drag-and-drop items, multiple-select items, and grid items.

Table 1.1. Domain-level Operational Test Blueprint—Percent Ranges

Grade	Physical Science	Life Science	Earth and Space Science	Engineering, Technology and Applications of Science
5	34–40%	23–29%	27–33%	3–7%
8	32–38%	31–37%	21–27%	2–6%

All questions on the Grade 5 and Grade 8 Science Tests are organized into clusters of questions that follow an assessment storyline. An assessment storyline provides a coherent path toward building Science and Engineering Practices, Disciplinary Core Ideas, and Crosscutting Concepts attached to a phenomenon. In question clusters, each question that is answered may add to the developing explanation, model, or design solution. The group of questions in a cluster follow a theme or storyline grounded in a phenomenon that is focused on an anchor Performance Expectation. However, questions that address other related Performance Expectations can also be included in the cluster. Table 1.2 presents the test designs for the 2024 Grades 5 and 8 Science Tests.

Table 1.2. Grades 5 and 8 Science Test Designs

Grade	Number of Question Clusters	Total Number of Questions
5	7–9	36–43
8	10–12	56–62

2. Performance Level Descriptions

Performance level descriptions (PLDs) are the foundation of standard setting activities because they provide the explanation of how student performance differs from one performance level to the next (Perie, 2008). PLDs are of such influence that, in a well-run standard setting workshop, they determine the rigor of the performance and thus the decisions made about placement of the cut score (Perie et al., 2008). PLDs also serve multiple purposes in terms of communicating policy, facilitating test development, guiding standard setting, and providing score interpretation. Three types of PLDs (Egan et al., 2012) are used as an organizing framework for developing PLDs for the Science examinations:

- Policy PLD statements are designed to capture the vision an agency has for its performance levels. They specify the number of levels and the names for each level and summarize the expectations of student performance for a testing program, including any policy decisions being made at particular levels. Table 2.1 presents the Policy PLDs for the Grade 5 and Grade 8 Science Tests.
- Range PLDs are designed to describe the full range of performance for students at a given performance level. In other words, Range PLDs describe the aspects of test content or specific items that are indicative of a range of students at a specific performance level. Range PLDs can be informative in guiding item and test development as a testing program evolves. They are critical in that they are used to articulate the borderline descriptions, which are a key component for standard setting.
- Borderline descriptions (also known as threshold PLDs) are designed to articulate the transition points between the different ranges of performance defined by the Range PLDs. Specifically, they describe the knowledge and skills a student at the border between performance levels should know and be able to do. Because they articulate the specific performance that distinguishes levels of performance, borderline descriptions are typically used in standard setting activities. Range PLDs and borderline descriptions are interdependent, which necessitates that they be developed in conjunction with each other.

Table 2.1. New York State Science Tests Policy PLDs

Performance Level	Policy PLD
Level 4	Students performing at this level **excel** in standards for their grade. They demonstrate knowledge, skills, and practices embodied by the Learning Standards that are considered **more than sufficient** for the expectations at this grade.
Level 3	Students performing at this level are **proficient** in standards for their grade. They demonstrate knowledge, skills, and practices embodied by the Learning Standards that are considered **sufficient** for the expectations at this grade.
Level 2	Students performing at this level are **partially proficient** in standards for their grade. They demonstrate knowledge, skills, and practices embodied by the Learning Standards that are considered partial but insufficient for the expectations at this grade. Students performing at Level 2 are considered on track to meet current New York high school graduation requirements but are **not yet proficient** in Learning Standards at this grade.
Level 1	Students performing at this level are **below proficient** in standards for their grade. They may demonstrate **limited** knowledge, skills, and practices embodied by the Learning Standards that are considered **insufficient** for the expectations at this grade.

Ultimately, the three types of PLDs are designed to describe the competencies of each performance level in relation to grade-level content standards while addressing their different functions. PLDs play a critical role in the standard setting process.

3. Standard Setting

Standard setting is the process whereby a group of educators is convened to recommend the cut scores (also known as performance or achievement standards) that separate an assessment's score scale into performance levels (i.e., a cut score is the minimum score students must receive to be classified into a certain performance level). Cut scores for the Grade 5 and Grade 8 Science Tests were recommended by two panels of 14 NYS educators each over a two-day standard setting meeting. The Modified Yes/No Angoff procedure (Impara & Plake, 1997; Plake, Ferdous, Impara, & Buckendahl, 2005) of determining cut scores was used in a multi-round process of performance judgments, feedback data, and discussions.

3.1. PANELISTS

The panelists, recruited by NYSED, represented the major geographic regions of NYS, as shown in Table 3.1. As shown in Table 3.2, a high percent of the panelists were classroom teachers, with those not serving as teachers indicating roles such as Curriculum Instruction or Academic Coordinator. In Table 3.3, the variety of settings for the panelists can be observed, with panelists coming from across Rural, Suburban, and Urban settings. Appendix B presents additional details on the demographic characteristics of the panelists.

Table 3.1. Number of Panelists by Geographic Location

Geographic Location	Grade 5	Grade 8
Capital District	3	3
Central	1	2
Long Island	2	1
Lower Hudson	2	1
Mid-Hudson	1	0
North Country/Adirondacks	1	1
NYC	3	3
Southern Tier	1	1
Western	0	2

Table 3.2. Number of Panelists by Current Role

Role	Grade 5	Grade 8
Classroom Teacher	12	10
Other (e.g. Curriculum/Learning Director)	2	4

Table 3.3. Number of Panelists by Current Setting

Setting	Grade 5	Grade 8
Rural	5	3
Suburban	6	4
Urban	3	7

3.2. METHODOLOGY

The Modified Yes/No Angoff standard setting method was used for the standard setting meeting. This is a content- and item-based method that leads participants through a standardized process through which they consider student expectations, as defined by PLDs, and the individual items that could be administered to students to recommend cut scores for each performance level. The process that was followed by the panel to establish their cut score recommendations involved the following steps:

- Review and familiarize themselves with the test form
- Review the current PLDs and develop borderline PLDs for each cut score
- Review and receive training on the Modified Yes/No Angoff methodology
- Complete independent Round 1 ratings and discuss with group after receiving feedback
- Complete independent Round 2 ratings and discuss with group after receiving feedback
- Complete independent Round 3 ratings

Once all three rounds of ratings were completed, the panelists completed an evaluation survey and concluded the meeting activities.

The standard setting process focused on students *just barely* at each performance level, or *threshold* (borderline) students. Therefore, the judgments provided by the panelists for each item and performance level were considered in terms of the success of borderline students. For example, *"Would a student with knowledge and skills at the borderline of the performance level be likely to answer the item correctly?"*

3.3. PRE-WORKSHOP

To engage in the judgment process of standard setting, there must be an understanding of content expectations for each performance level. Prior to the standard setting workshop, panelists were provided some pre-workshop tasks through the Pearson standard setting website, including an introductory standard setting training video, and copies of the Policy and Range PLDs. These tasks were provided ahead of the workshop to set the context for standard setting. Panelists were also asked to review the Educator Guide that includes some sample test items—items available to the public as practice items—to understand what students had to do on the test. Panelists were also asked to review and sign a non-disclosure agreement prior to the workshop and complete a brief demographic survey.

3.4. WORKSHOP

The standard setting workshop was held in Troy, New York, from July 10–11, 2024. Appendix A presents the workshop agenda. The workshop began with a welcome from NYSED, introductory remarks about the Grade 5 and Grade 8 Science Tests, and the goals for setting performance standards on the tests. The lead facilitator provided an overview of the standard setting process, explaining the different types of contextual information used (e.g., PLDs, test content), the standard setting judgment process, and the different types of feedback data that would be presented throughout the workshop. After the general orientation, including workshop logistics, the panelists split into their separate panels and began their work by first reviewing an online version of an operational test form for their grade level.

3.5. PEARSON STANDARD SETTING WEBSITE

The Pearson standard setting website (Moodle) was used as the online platform for meeting pre-work, facilitating the standard setting meeting, and collecting panelist judgments throughout the standard setting process. Each panelist was provided a unique user identification and password that provided secure access to the site. Panelist access was restricted to the section of the site associated with the specific exam assigned to their panel. The standard setting website provided panelists the opportunity to access all resource materials within a secure environment. The website also allowed for streamlining of the data collection from the individual judgment process.

3.6. TEST REVIEW

The panelists were provided access to the spring 2024 computer-based tests that included the full operational test. This provided them with an opportunity to review the multiple-choice items, constructed-response items, and technology-enhanced items to better understand what students were asked to do on the tests. The Rating Guide was provided via the standard setting website to provide the key idea assessed for each multiple-choice item, the answer key for the multiple-choice items, and the scoring rubrics for constructed-response or technology-enhanced items.

3.7. PERFORMANCE LEVEL DESCRIPTIONS

After the test review, the facilitator discussed the Range PLDs and their use during the standard setting process. Panelists were given 15 minutes to discuss the Range PLDs in their table groups, focusing on key differences between the performance levels. The facilitator then provided an explanation for how to derive borderline descriptions from the Range PLDs. Prior to the standard setting, the PLDs were unpacked to highlight the multi-dimensional nature of the standards. For each PLD, the Crosscutting Concepts (CCCs), the Disciplinary Core Ideas (DCIs), and Scientific and Engineering Principles (SEPs) were included within the PLD statements. Using the unpacked PLDs, the facilitator led the full panel through the process of creating borderline descriptions for a small set of PLDs. Following the initial development, panelists split into smaller table discussions and proceeded with the development of the remaining PLDs. To complete the work, the panels first focused on the development of borderline descriptions for the Level 3 cut. After completing the Level 3 borderline descriptions, panelists proceeded to complete the Level 2 and Level 4 descriptions.

After the panelists drafted the borderline descriptions within their table, the facilitator organized the draft descriptions from each table group into a master Google doc. The facilitator then led the whole group through a review of the descriptions and captured any group-approved edits into the master document. The borderline descriptions were printed and shared with the panelists to reference during the judgment activities.

3.8. MODIFIED YES/NO ANGOFF JUDGMENT TRAINING

The panelists were provided thorough training on how to make their recommendations as part of the standard setting meeting. They were instructed on using the Modified Yes/No Angoff method. All items on the test were scored dichotomously. Because all items were scored dichotomously, the essential question that panelists were asked to address was, "*Would a student with knowledge and skills at the borderline of the performance level be likely to answer the item correctly?"* Panelists were instructed to review this question for each of the three cut scores for each item. Significant time was spent on describing the thought process the panelists should go through using parts of the question:

- "*Would*…"— When considering the expected student response to an item, the panelists needed to consider how a student would respond rather than how they should respond. Where "should" is an aspirational expectation, "would" is a more realistic expectation of a student response to the item.
- "*...a student with knowledge and skills at the borderline of the performance level*…"— Panelists should reference the borderline descriptions for each performance level to determine how a student with knowledge and skills at the borderline would be expected to respond.
- "...*be likely answer the question correctly*?"—The panelists will review the knowledge and skills necessary to provide a correct response to the item compared to the expected PLDs for the borderline performance level student. In this context, "likely" is defined as 2 out of 3 times, or 67%. To make this concrete for panelists, facilitators asked them to think about three students at the borderline of a performance level.

Panelists were then instructed to answer the judgment question using the thought process and determine a Yes or No answer for each of the three cut scores for each item. An illustration of the rating form is shown in Figure 3.1.

Figure 3.1. Available Response Options to Judgment Question for Multiple-Choice, CR, and TE Items

	A	B	C	E	F	G
1	**New York State**	**Science Grade 8**				
2	**Round 1 Rating Sheet**					
3	**Test**	**Sequence#**	**Item Type**	**Level 2**	**Level 3**	**Level 4**
4	Grade 8 Science	1	MC			
5	Grade 8 Science	2	MC			
6	Grade 8 Science	3	MC			
7	Grade 8 Science	4	CR (TEI-match)			
8	Grade 8 Science	5	CR			
9	Grade 8 Science	6	MC			

Another step in the standard setting process is a practice judgment task to give the panelists the opportunity to practice making judgments prior to beginning the actual judgment rounds. A set of five practice items was selected from the NYS Question Sampler for use in this activity. However, this activity did not take place during the actual standard setting, given that the borderline description development activity took more time than anticipated. As a result, NYSED and Pearson made the decision to forgo the practice activity and move directly to making the actual judgments in the standard setting rounds in an effort to manage the panelists' time as effectively as possible.

3.9. STANDARD SETTING ROUNDS

Prior to starting each judgment round, panelists were asked a series of readiness questions (via a survey on the website, as shown in Appendix C) to verify that they understood their task and were ready to begin:

- Do you understand your task for the item judgment activity?
- Are you ready to begin the item judgment activity?

Following the readiness survey, the facilitator reviewed the responses. If a panelist were to have responded “no” to either of the questions in the readiness survey, the facilitator would have provided additional training and support as needed to the panelist. Once the facilitator ensured that all panelists were ready to proceed, panelists were asked to make judgments for the first item starting at the lowest performance level based on the borderline descriptions and the knowledge and skills required by the item. The panelists then made judgments for the same item for the rest of the performance levels before proceeding to the next item. Judgments were recorded in an Excel rating form available through the Pearson standard setting website. Once the panelists completed making judgments for all items, they notified their facilitator, who then aggregated all ratings for all panelists. After all panelists completed each judgment activity, the facilitators gathered the item judgments, performed the necessary analysis of the data, and created feedback data that were provided to the panelists.

For the purposes of this workshop, the ratings for all panelists were determined by summing up the number of items that panelists indicated “Yes” for each performance level. This score represented the raw score recommendation for each panelist. However, within the test form being reviewed and rated, there were some field test, or pretest items, that were not used in the calculation of scores for candidates. In order to keep the location of the pretest items confidential, feedback to panelists was not provided at a raw score level. Instead, a pseudo scale score was created for each exam. All feedback to panelists was provided using the pseudo scale score.

Using the pseudo scale score prevented the panelists from easily calculating their personal cut scores, which could have alerted the panelists to the location of the field test items. Instead, a linear transformation of the raw scores was completed to arrive at the pseudo scale score for each cut score recommendation. The linear transformation was designed to create a unique scale used only for the standard setting meeting with a minimum score of 570 and a maximum score of 790. Panelists were provided the guideposts provided in Table 3.4 to aid in their interpretation of the pseudo scale. A copy of the raw score to pseudo scale score is included in Appendix C.

Table 3.4. Guidance Provided to Panelists on the Interpretation of the Pseudo Scale

Minimum	25% Correct	50% Correct	75% Correct	Maximum
570	640	680	720	790

After Round 1, the facilitator provided cut scores generated from the panelists’ item-level judgments. Each panelist was able to see their recommended cut score in the Excel rating form. The facilitator then presented a summary of the overall ratings. These feedback in the minimum and maximum values received for each cut score, along with the mean and median across all panelists. Panelists were also shown a histogram that indicated the number of people who provided each cut score recommendation. Using this information, panelists could compare their own cut scores to those from the overall panel and consider if their cut scores matched their level of expectations. The facilitator then led a discussion with the panel regarding their ratings and how they fit within the overall distribution and if panelists felt comfortable with their overall ratings.

After this review, the facilitator led a discussion of the ratings for specific items. Using the panels’ Round 1 judgments, items were flagged that witnessed significant disagreements for any of the cut scores. These disagreements could be reflected with a wide range of ratings, with some panelists rating an item fairly easy (the borderline level 2 would get the item correct) and still others rating it fairly difficult (the borderline level 4 would not get the item correct). The facilitator led the panel in a discussion of the items and panelists discussed what characteristics or features

of the items moved them to rate as they did. During this discussion, the facilitator also had available estimates for item difficulty. The item difficulty estimates were not shared directly with the panelists, but the facilitator did share difficulty estimates at a broad level (easy item, hard item, medium difficulty).

After this discussion concluded, panelists completed their Round 2 ratings. Round 2 of standard setting was performed just as Round 1 had been. Panelists were instructed to revisit their judgments from Round 1 and make a new set of judgments, keeping their judgments from Round 1 or making revisions as they felt necessary. After Round 2 judgments, panelists were provided with another set of individual and panel-level cut score information. The facilitator also led a discussion of another set of items where significant disagreement on the ratings were observed. The facilitator led the discussion for both the feedback and the specific items reviewed.

The facilitator also displayed impact data, or the distribution of students among performance levels based on the panel's overall cut scores. Presenting these data during the standard setting process gave the panelists the opportunity to see the consequences of their judgments and whether these consequences fit their expectations. The panelists were reminded that the data should not drive their judgments; rather, their judgments should be driven by content expectations. A discussion was led by the facilitator to discuss whether the impact data aligned with their content expectations.

Following the discussion of the Round 2 feedback data, the panelists provided one final round of judgments. This round was performed just as the previous two rounds. Once the results for Round 3 were complete, panelists were shown the final recommended cut scores and corresponding impact data. As a final task, the panelists completed a workshop evaluation that asked questions ranging from how comfortable they were with specific workshop activities to how comfortable they were with the final recommended cut scores. Table 3.5 presents the types of feedback data and at what round they were provided to the panelist. Appendix D presents examples of the feedback.

Table 3.5. Feedback Data by Judgment Round

Level	Feedback	Round 1	Round 2	Round 3
Item Level	Panelist Agreement Data	✓	✓	
	Score Point Distributions	✓		
Test Level	Individual Cut Scores	✓	✓	
	Committee Scores	✓	✓	✓
	Panelist Agreement Data	✓	✓	
	Impact Data		✓	✓

3.10. CUT SCORES AND IMPACT DATA

Cut scores were generated after each round of judgments. The median value of the individual panelists' cut scores, per performance level, was used as the recommended cut score of the standard setting panel. The standard error of judgment (SEJ) was also calculated for the final recommended cut scores to serve as additional information. Table 3.6 and Table 3.7 present a summary of the cut scores for all three rounds. Figure 3.2 presents the impact data for the third and final round of ratings from the panelists.

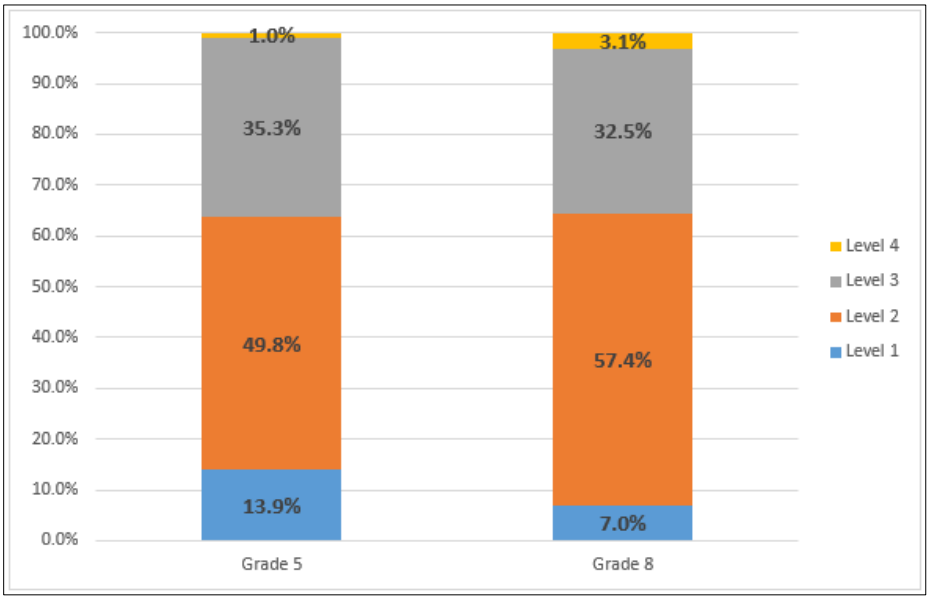
Table 3.6. Recommended Cut Scores Across Rounds—Grade 5

Round	Performance Level	Min.	Max.	Average	Median	SE	25th	75th	% Impact
Round 1	Level 1	--	--	--	--	--	--	--	52.1
	Level 2	9	21	14.6	13.5	1.1	11.25	18.75	47.6
	Level 3	21	34	28.6	29.5	0.9	26.25	31	0.3
	Level 4	31	34	33.6	34.0	0.2	34	34	0.0
Round 2	Level 1	--	--	--	--	--	--	--	45.9
	Level 2	3	22	12.4	12.0	1.4	9	16	50.2
	Level 3	12	31	23.9	24.5	1.3	23	26	3.8
	Level 4	26	32	30.0	30.5	0.5	29.25	31	0.1
Round 3	Level 1	--	--	--	--	--	--	--	13.9
	Level 2	0	20	7.7	7.0	1.3	5.25	9.75	49.8
	Level 3	11	28	16.6	15.0	1.4	12.5	19.25	35.3
	Level 4	21	32	27.2	27.0	0.9	25.25	30.5	0.1

Table 3.7. Recommended Cut Scores Across Rounds—Grade 8

Round	Performance Level	Min.	Max.	Average	Median	SE	25th	75th	% Impact
Round 1	Level 1	--	--	--	--	--	--	--	81.2
	Level 2	4	42	26.1	26.0	2.4	21.5	32	18.2
	Level 3	12	51	42.7	43.5	2.6	42.25	48	0.5
	Level 4	29	53	50.7	53.0	1.7	52	53	0.0
Round 2	Level 1	--	--	--	--	--	--	--	48.6
	Level 2	0	22	12.7	16.0	2.0	6.25	17.75	43.4
	Level 3	6	38	26.6	32.0	3.2	17.75	34	7.5
	Level 4	27	48	40.9	43.5	1.8	37.5	46	0.5
Round 3	Level 1	--	--	--	--	--	--	--	7.0
	Level 2	0	16	7.7	8.5	1.3	5.25	9.75	57.4
	Level 3	7	53	21.8	20.5	3.0	13.5	24.75	32.5
	Level 4	20	47	36.0	37.0	1.9	33.25	41.25	3.1

Figure 3.2. Impact Data Based on Round 3 Ratings



3.11. WORKSHOP EVALUATION

Once the standard setting process was complete and the final recommended cut scores were shown, panelists completed a workshop evaluation on the various materials and activities of the standard setting process and the final recommended cut scores. The intent of this survey was to gather how well panelists understood the process and the materials used and how comfortable they felt about the final recommended cut scores. For the survey questions covering recommended cut scores, panelists were able to express how they would modify the percent of students classified into each performance level if they were somewhat uncomfortable with the overall final recommendation. Most survey questions used a Likert scale, with different scales of affect (e.g., not confident to very confident, not adequate to very adequate, not useful to very useful) across the evaluation.

A complete summary of the evaluation results for both grade levels can be found in Appendix E. One question assessed panelists' confidence in the final cut scores. More specifically, the panelists were asked to rate:

- Please indicate your opinion regarding whether you feel the group's final recommended cut scores were too low, about right, or too high for each cut score. Please bubble *only* one of the three options for each cut score.

As shown in Table 3.8 and Table 3.9, the panelists generally felt comfortable with the cut score recommendations they had developed. There were some panelists in both panels that felt the Level 4 cut score was too high, but a clear majority still rated it as "About Right."

Table 3.8. Ratings from Grade 5 Panel

Performance Level	Too Low	About Right	Too High
Level 2	--	13	1
Level 3	--	14	--
Level 4	1	9	4

Table 3.9. Ratings from Grade 8 Panel

Performance Level	Too Low	About Right	Too High
Level 2	5	8	1
Level 3	1	11	2
Level 4	1	9	4

3.12. FINAL RECOMMENDATIONS

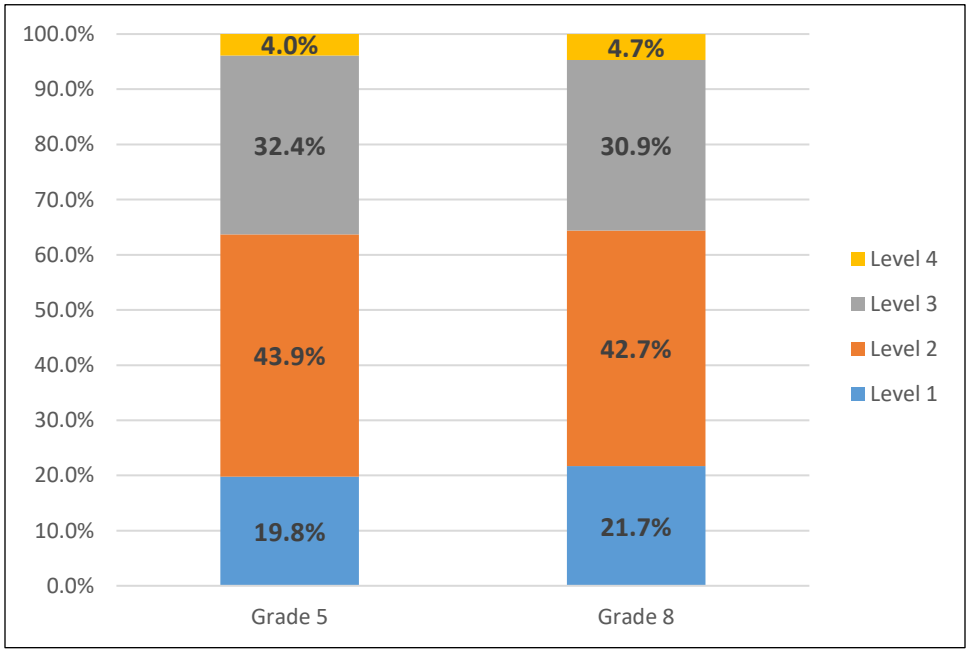
The goal of the standard setting meeting was to identify performance level cut scores consistent with the PLDs and state policy directives using a standardized procedure called the Modified Yes/No Angoff method. The meeting reflected best practice as articulated in the *Standards for Educational and Psychological Measurement* (AERA et al., 2014) and proceeded according to plans reviewed by the New York State Technical Advisory Committee. The panelists were diverse and representative of the state, and the group followed, without incident, instructions delivered by the standard setting facilitator. All activities were formally overseen by the OSA senior management and psychometric staff.

After careful consideration of the nature of the new examination, the rigor of the new curricula, the transitional and aspirational aspects of the NYSED policy directives, and the role of the assessment in student learning, the standard setting committee made recommendations on the cut scores to the Commissioner of Education. The Commissioner of Education subsequently made adjustments to the recommendations based on the committee feedback from the survey, standard errors of judgement, and historical data. The final approved cut scores were implemented within the scale of measurement used to report student performance on the New York State Grade 5 and Grade 8 Science Tests. Table 3.10 presents the approved cuts scores, with subsequent impact data provided in Figure 3.3.

Table 3.10. Final Approved Cut Scores

Grade	Level 2 Cut	Level 3 Cut	Level 4 Cut
5	8	15	24
8	11	20	35

Figure 3.3. Impact Data Based on Final Approved Cut Scores



4. References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for Educational and Psychological Testing*. AERA.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (pp.508–600). American Council on Education.

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). Routledge.

Perie, M. (2008). A guide to understanding and developing performance-level descriptors. *Educational Measurement: Issues and Practice*, *27*(4), 15–29.

Perie, M., Hess, K., & Gong, B. (2008). Writing performance level descriptors: Applying lessons learned from the general assessment to alternate assessments based on alternate and modified achievement standards. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York.

Appendix A: Standard Setting Agenda

Standard Setting Meeting

New York State Elementary-level Science (Grade 5)
and Intermediate-level Science (Grade 8)
Agenda

Day 1- July 10, 2024

7:30 – 8:00am	*Breakfast*
8:00 – 8:30am	Welcome and Standard Setting Overview
****	Break into Grade-level panels ****
8:30 – 8:45am	Introductions, logins, material orientation, meeting security
8:45 – 9:30am	Experience the Assessment
9:30 – 9:45am	*Break*
9:45 – 10:15am	Review and Discuss Performance Level Descriptions
10:15 – 10:45am	Borderline Performance Level Descriptors Training [Includes modeling]
10:45 – 11:45am	Borderline PLD Level 3 Creation
	Table Discussion
	Group Discussion
11:45 – 12:30pm	*Lunch*
12:30 – 2:00pm	Borderline PLD Levels 2 and 4 Creation
	Table Discussion
	Group Discussion
2:00 – 2:30pm	Standard Setting Training
2:30 – 3:00pm	Practice Judgment Activity and Discussion
3:00 – 3:15pm	Break
3:15 – 4:30pm	Round 1 Judgments

Day 2 - July 11, 2024

7:30 – 8:30am	*Breakfast*
****	Break into Grade-level panels ****
8:30 – 8:45am	Round 1 Judgment Feedback
	Item Level - Item means and distributions
	Test Level - Cut score recommendations; Panelist agreement
8:45 – 9:30am	Table Discussion - Round 1 Feedback
	Panelists discuss feedback data at their tables
9:30 – 9:45am	Whole Group Discussion – Item Disagreement Data
9:45 – 10:45am	Round 2 Judgments
	Round 2 Readiness form
	Panelists work independently to make Round 2 judgments
10:45 – 11:00am	*Break*
11:00 – 11:15am	Round 2 Judgment Feedback
	Item Level - Item means and distributions
	Test Level - Cut score recommendations, Panelist agreement
11:15 – 11:45am	Table Discussion - Round 2 Feedback
11:45 – 12:30pm	*Lunch*
12:30 – 1:30pm	Whole Group Discussion - Round 2 Feedback
	Impact Data
1:30 – 2:15pm	Round 3 Judgments
	Round 3 Readiness form
	Panelists work independently to make Round 3 judgments
2:15 – 2:45pm	*Break*
2:45 – 3:15pm	Round 3 Feedback, Evaluation, and Workshop Wrap-up

Appendix B: Panelist Demographics

Panelists responded to an information survey to provide demographic and other pertinent information for validity evidence of the standard setting. A total of 28 panelists participated in the standard setting. The survey results have been tabulated below.

Table B.1. What is your current position?

Answer Option	Grade 5	Grade 8
Classroom Teacher	12	10
Other (e.g. Curriculum/Learning Director)	2	4

Table B.2. How many years have you been in the education field?

Answer Option	Grade 5	Grade 8
1–5 years	--	1
6–10 years	--	2
11–15 years	1	3
16–20 years	4	4
More than 20 years	9	4

Table B.3. What is the highest educational degree you have earned?

Answer Option	Grade 5	Grade 8
Master's degree (M.A., M.S.)	13	14
Doctoral degree (Ph.D., Ed.D.)	1	--

Table B.4. What is your gender?

Answer Option	Grade 5	Grade 8
Female	13	10
Male	--	4
No response	1	--

Table B.5. What is your race/ethnicity?

Answer Option	Grade 5	Grade 8
Asian	1	1
Black or African American	--	2
Hispanic or Latino	--	1
Multi-racial	2	--
White	10	8
No response	1	2

Table B.6. In what type of school district do you work?

Answer Option	Grade 5	Grade 8
Rural	5	3
Metropolitan/Urban	6	4
Suburban	3	7

Appendix C: Raw Score to Pseudo Scale for the Workshop

Table C.1. Grade 5 Raw Score to Pseudo Scale Score

Raw Score	Pseudo Scale
0	571
1	576
2	580
3	584
4	588
5	592
6	596
7	600
8	606
9	612
10	618
11	623
12	627
13	632
14	636
15	641
16	645
17	649
18	654
19	658
20	662
21	667
22	671
23	676
24	681
25	687
26	693
27	700
28	704
29	708
30	712
31	716
32	720
33	724
34	729

Table C.2. Grade 8 Raw Score to Pseudo Scale Score

Raw Score	Pseudo Scale
0	573
1	578
2	582
3	586
4	591
5	595
6	600
7	607
8	614
9	619
10	624
11	629
12	633
13	637
14	641
15	645
16	649
17	652
18	655
19	659
20	662
21	665
22	668
23	671
24	674
25	677
26	680
27	683
28	686
29	688
30	691
31	694
32	697
33	700
34	703
35	706
36	710
37	713
38	716
39	720
40	724
41	728
42	733
43	738
44	743
45	750
46	755
47	759
48	763
49	768
50	772
51	777
52	781
53	786

Appendix D: Sample Feedback

Table D.1. Rating Summary (provided all rounds)

	Panelists	Average Rating	Median Rating	Min	Max
Level 1	--	--	--	--	--
Level 2	14	605.5	600.0	571	662
Level 3	14	648.3	641.0	623	704
Level 4	14	698.3	700.0	667	720

Figure D.1. Cut Score Rating Distribution (provided all rounds)

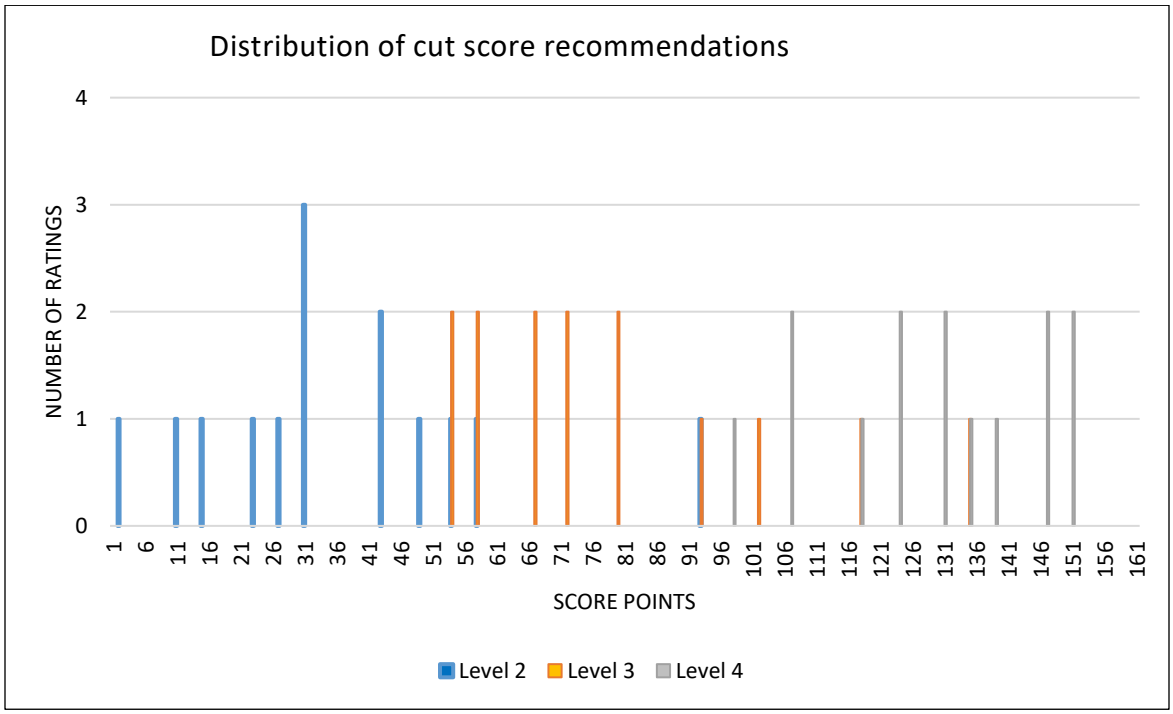
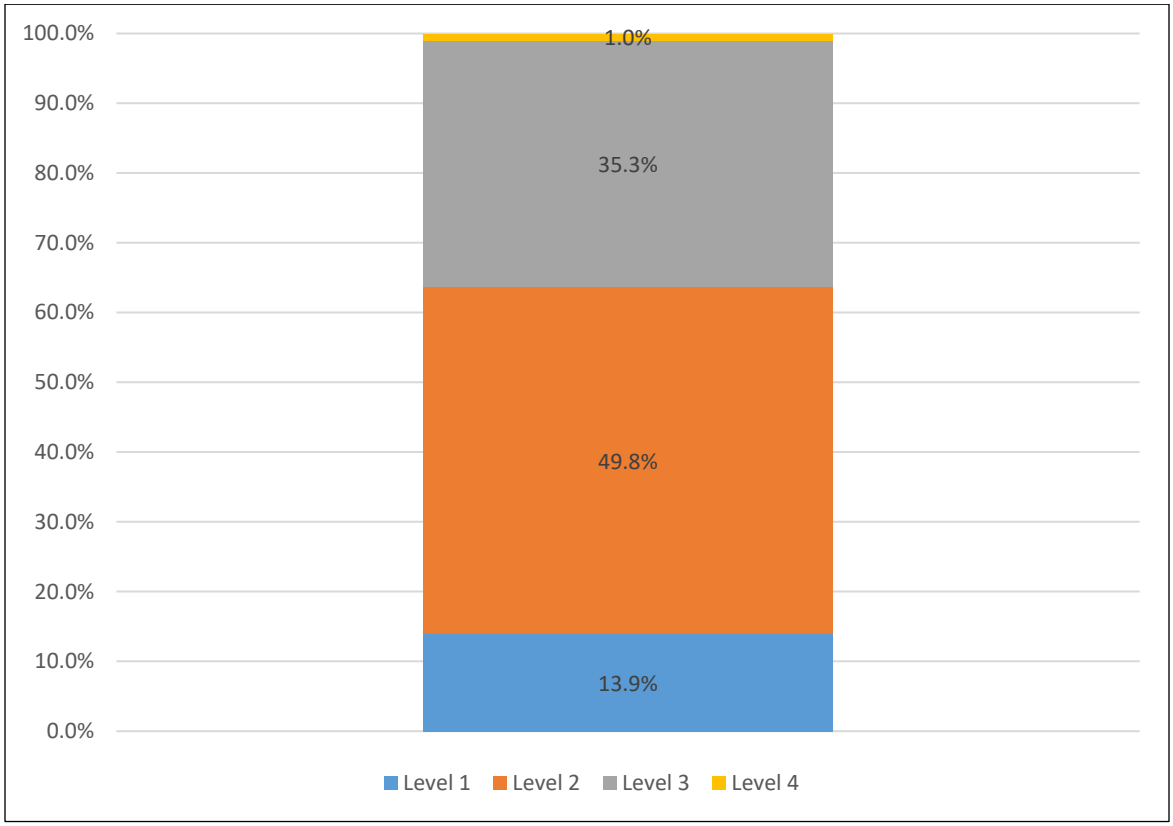


Figure D.2. Impact Data (Provided after Round 2 and Round 3)



3695831.401

Appendix E: Workshop Evaluation Results

The purpose of this evaluation is to help document the process used to recommend cut scores for New York State's Grades 5 and 8 Science Tests.

GRADE 5

Table E.1. Training Process—Grade 5

Response Option	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Before Round 1 began, I was comfortable with the item rating procedure.	2	4	6	3	--
I understood the cut-score summary data that was presented between the rounds.	--	1	--	10	3
I understood the impact data that were presented after Round 2.	--	--	--	9	5
By the end of Round 3, I was comfortable with the item rating procedure.	--	--	--	4	10
Overall, I believe my opinions were considered and valued by my group.	--	--	--	3	11

Table E.2. Influence—Grade 5

Response Option	Not Influential	Somewhat Influential	Influential	Very Influential
The Performance Level Descriptions (PLDs)	--	1	4	9
The descriptions of students demonstrating borderline performance.	--	--	8	6
My perception of the difficulty of the items	--	--	8	6
My experiences with students	--	1	7	6
Discussion within my group	--	--	5	9
The item ratings of other participants	1	6	5	2
The percent of students in each performance level (the impact data)	1	4	8	1
My sense of what a student needs to know to be identified at Level 2.	--	1	5	8
My sense of what a student needs to know to be identified at Level 3	--	1	5	8
My sense of what a student needs to know to be identified at Level 4	--	1	5	8

Table E.3. Cut Scores—Grade 5

Response Option	Too Low	About Right	Too High
Level 2 cut score	--	13	1
Level 3 cut score	--	14	--
Level 4 cut score	1	9	4

GRADE 8

Table E.4. Training Process—Grade 8

Response Option	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Before Round 1 began, I was comfortable with the item rating procedure.	1	5	3	4	1
I understood the cut-score summary data that was presented between the rounds.	--	--	1	9	4
I understood the impact data that were presented after Round 2.	--	--	--	4	10
By the end of Round 3, I was comfortable with the item rating procedure.	--	--	1	3	10
Overall, I believe my opinions were considered and valued by my group.	--	--	--	4	10

Table E.5. Influence—Grade 8

Response Option	Not Influential	Somewhat Influential	Influential	Very Influential
The Performance Level Descriptions (PLDs)	--	2	5	7
The descriptions of students demonstrating borderline performance.	1	1	5	7
My perception of the difficulty of the items	--	--	7	7
My experiences with students	--	--	4	10
Discussion within my group	--	1	6	7
The item ratings of other participants	--	4	8	2
The percent of students in each performance level (the impact data)	--	4	4	6
My sense of what a student needs to know to be identified at Level 2.	--	2	5	7
My sense of what a student needs to know to be identified at Level 3	--	1	6	7
My sense of what a student needs to know to be identified at Level 4	--	1	6	7

Table E.6. Cut Scores—Grade 8

Response Option	Too Low	About Right	Too High
Level 2 cut score	5	8	1
Level 3 cut score	1	11	2
Level 4 cut score	1	9	4