

Kentucky Science Assessment Systems



Standard Setting Meeting

Grade 11

July 2019

Version 1.0



Pearson

Table of Contents

| | |
|--|-----------|
| Executive Summary | 1 |
| Kentucky Science Assessment Standard Setting Process and Results | 1 |
| Policy Definitions..... | 1 |
| Performance Level Descriptors (PLDs) | 2 |
| Cut Scores | 2 |
| General Method | 3 |
| Results for Kentucky Science Assessments – Grade 11 | 4 |
| Chapter 1 – Overview of the Standard Setting Process | 5 |
| Goals of the Standard Setting Meeting | 5 |
| Kentucky Science Assessment Performance Levels | 5 |
| Kentucky Science Assessment Standard Setting Process..... | 6 |
| Chapter 2 – Preparations for the Standard Setting | 8 |
| Development of Participant Materials | 8 |
| Development of Presentation Materials | 9 |
| Facilitator Training | 9 |
| Preparation for Data Analysis During the Meetings..... | 9 |
| Chapter 3 – Standard Setting Meeting | 11 |
| Purpose of the Standard Setting Meetings | 11 |
| Committee Panelist Composition..... | 11 |
| Standard Setting Meeting Facilitators and Staff | 12 |
| Meeting Facilitator..... | 12 |
| Meeting Data Analysts | 12 |
| KDE Staff..... | 12 |
| Standard Setting Materials | 12 |
| Pearson Standard Setting Website | 12 |
| Committee Panelist Folders | 14 |
| Computers | 14 |
| Standard Setting Procedure | 14 |
| Standard Setting Meeting Proceedings..... | 15 |
| Standard Setting Meeting Pre-Work..... | 15 |
| Breakout Session..... | 16 |
| Recommended Cut Scores from Standard Setting Committee..... | 22 |
| Articulation | 23 |
| Articulation Process | 24 |
| Performance Level Descriptor Development..... | 25 |

| | |
|--|-----------|
| Meeting Process | 26 |
| Chapter 4 – Post-Standard Setting | 28 |
| Reasonableness Review | 28 |
| Chapter 5 – Evidence of Procedural Validity of the Standard Setting Process | 29 |
| Committee Representation | 29 |
| Committee Training | 29 |
| Perceived Validity of the Standard Setting | 32 |
| References | 34 |
| Appendix A – Panelist Meeting Materials | 35 |
| Appendix B – Committee Panelist Composition | 45 |
| Appendix C – Standard Setting Meeting Agenda | 49 |
| Appendix D – Presentations | 52 |
| Science Grade 11 Breakout Session – Day 1 | 52 |
| Science Grade 11 Breakout Session – Day 2 | 53 |
| Science Grade 11 Breakout Session – Day 3 | 54 |
| Appendix E – Examples of Feedback Data | 55 |
| Appendix F – Panelist Evaluation Results | 60 |
| Appendix G – Recommended Cut Scores by Judgment Round | 78 |
| Appendix H – Recommended Cut Score Summary Statistics by Judgment Round | 79 |
| Appendix I – Test-Level Participant Judgment Agreement | 80 |
| Appendix J – Performance Level Descriptors | 83 |

Executive Summary

This report summarizes the process and results of setting performance level standards for the Kentucky science assessments for Grade 11. The Kentucky Department of Education (KDE) and Pearson (Science assessment contractors) recommend the achievement levels shown in Table E2 of this report for adoption by KDE, the State Board of Education, and the Commissioner of Education.

Kentucky Science Assessment Standard Setting Process and Results

Performance levels are used to classify and describe student performance on an assessment. To classify student performance into different performance levels, the following components are generally required: 1) policy definitions, 2) Performance Level Descriptors (PLDs), and 3) cut scores. Policy definitions describe the performance levels in general terms that apply to all grades. PLDs illustrate the performance levels in terms that are specific to a grade. Cut scores represent the lowest boundary of each performance level on the scale.

The process of recommending performance standards for the Kentucky science assessments for grade 11 was in line with the process used for the Kentucky science assessments for grades 4 and 7 and national best practice for standard setting. Results and details of the process are presented in the following sections.

Policy Definitions

Policy level descriptors for the Kentucky science assessments are shown in Table ES1. The titles and descriptions of the achievement levels were defined to be part of a cohesive assessment system, and the achievement levels indicate a student's ability to demonstrate mastery on the Kentucky Academic Standards (KAS).

Table ES1. Policy level descriptors for the Kentucky Science Assessment

| Performance Level | Policy Level Descriptors |
|----------------------|---|
| Distinguished | A student performing at the <i>Distinguished</i> level has a comprehensive understanding of science concepts and practices. The student consistently communicates ideas in a sophisticated and complex manner, using thorough supporting detail and explicit examples. The student reasons and solves problems by using appropriate strategies in an insightful way. Connections between science concepts/ideas, when appropriate, are justified and insightful. |
| Proficient | A student performing at the <i>Proficient</i> level has a broad understanding of science concepts and practices. The student usually communicates ideas accurately using clear and appropriate examples, supporting or justifying those ideas with relevant details and evidence. Problem-solving and critical thinking skills are used effectively. Connections between science concepts/ideas, when present are reasonable and appropriate. |
| Apprentice | A student performing at the <i>Apprentice</i> level has a basic understanding of science concepts and practices. The student communicates ideas in a basic manner, but explanations, solutions or justifications may be unclear or ineffective. The student demonstrates some problem-solving and critical thinking skills, but they are not consistently applied. |
| Novice | A student performing at the <i>Novice</i> level has a minimal understanding of science concepts and practices. The student communicates ideas ineffectively or inaccurately, providing little detail and little or no support. Attempts at problem solving or critical thinking are minimal or inappropriate. |

Performance Level Descriptors (PLDs)

There are two types of performance level descriptors: Range PLDs and borderline descriptions. Borderline descriptions represent the knowledge and skills of a student with performance at the borderline of the performance level (i.e., one that is just-barely past the point-of-entry for the performance level). As part of the standard setting process, panelists developed borderline descriptions for the respective grade. These borderline descriptions were then used by the standard setting panelists to recommend cut scores. The range PLDs represent the range of knowledge and skills a typical student in the performance level would likely demonstrate. Panelists of the standard setting meeting developed range PLDs at the end of the standard setting using an anchored PLD development process. The recommended cut scores were used to divide science assessment content (i.e., items) into the four performance levels and the range PLDs were created using those item groupings. The range PLDs from the standard setting panelists were reviewed by KDE and will appear in the score reports release.

Cut Scores

To create a common point of reference across the science assessments in grades 4, 7, and 11, cut scores and measures of student performance on the Kentucky science assessments are translated to a scale that ranges from 100 to 300 points and has a Proficient cut of 210. The common value of 210 for the Proficient cut score across assessments does not mean that they reflect the same difficulty, or that achievement levels can be compared in difficulty through the scale values of their cut scores across grades. Similarly, the percentage of

students in a performance level is not directly comparable across grades. The population of students tested is different for each specific grade-level assessment. Performance levels from different tests are not comparable because the cut scores for these tests are criterion-referenced—they are based on content-specific expectations of what students should know and be able to do.

The cut scores recommended for adoption are shown in Table ES2. This table shows the scale score ranges corresponding to each performance level. The cut scores for the performance levels are the lowest cut score within each range. There is no cut score for Novice, given that 100 is the lowest attainable scale score a student can earn.

Table ES2. Cut Score Ranges for Kentucky Science Assessment Performance Levels

| Performance Level | Raw Score Ranges |
|-------------------|------------------|
| | Grade 11 |
| Distinguished | 231 to 300 |
| Proficient | 210 to 230 |
| Apprentice | 190 to 209 |
| Novice | 100 to 189 |

Details pertaining to the general method for obtaining the recommended cut scores are provided below.

General Method

From July 16 to July 18, 2019, after the first year of operational administration, a standard setting committee meeting was conducted to provide cut score recommendations for the Kentucky science assessment for grade 11. The committee was comprised of 11 individuals, including teachers and non-teacher educators. The panelists were selected for the standard setting committee to provide content and grade-level expertise during the committee meeting and be representative of the state teaching population, including geographic region, gender, ethnicity, educational experience, community size, and community socioeconomic status.

The Extended Modified (Yes/No) Angoff standard setting method was used at the standard setting meeting (Davis & Moyer, 2015; Plake, Ferdous, Impara, & Buckendahl, 2005). This is a content- and item-based method that leads panelists through a standardized process through which they consider expectations of student performance, as defined by the borderline descriptions, and the individual items administered to students to recommend cut scores for each performance level. Because items are presented in clusters during the assessment, the panelists used the same process to provide judgments for the item clusters.

The process started with panelists experiencing the science test for grade 11 from the spring 2019 administration using paper test books from the spring administration. Based on their experience with the test items and a review of the borderline descriptions, panelists reviewed each item on the test and answered the following question for each performance level:

“How many points would a student performing at the borderline of the [specific] performance level likely earn if he or she answered the question?”

The cut score recommendation for each individual panelist was the expected raw score a borderline student at the respective performance level would likely earn, calculated as the

sum of the individual item judgments. For the purposes of the standard setting, “likely” was defined as 2 out of 3 students at the borderline of the performance level. Each recommended cut score from the standard setting committee is the median of the recommendations from the individual panelists in the committee.

As part of the standard setting process, the panelists reviewed impact data resulting from their cut score recommendations for the grade 11 science assessment with the impact data from the performance level standards established for the grades 4 and 7 science assessment. After review of the data, the panelists discussed and recommended adjustments to the cut scores that would ensure the results were coherent with the Kentucky science assessment system and defensible.

Results for Kentucky Science Assessments – Grade 11

Table ES3 shows the percentage of students who took the Kentucky science assessments for grade 11 during the Spring 2018-2019 administration that would be classified into each performance level based on the cut score. The percentage of students in an achievement level is not directly comparable across grades and subjects. The population of students tested is different for each assessment. Achievement levels from different tests are not comparable because the cut scores for these tests are criterion-referenced—they are based on content-specific expectations of what students should know and be able to do.

Table ES3. Percentage of Students in Performance Levels

| Performance Level | Assessment |
|-------------------|------------------|
| | Science Grade 11 |
| Distinguished | 2% |
| Proficient | 28% |
| Apprentice | 50% |
| Novice | 20% |

Chapter 1 – Overview of the Standard Setting Process

Chapter 1 provides an overview of the standard setting process used for the Kentucky science assessment in grade 11.

Goals of the Standard Setting Meeting

Once an assessment is administered, various groups—including students, parents, educators, administrators, and policymakers—want to know how students performed on the assessment and how to interpret that performance. By establishing levels associated with different student performance on the assessment, a frame of reference is developed for interpreting scores. For a criterion, standards-based assessment, such as the Kentucky Science assessment program, performance on the assessment is compared to a set of predefined content standards. The standards communicated within the *Kentucky Academic Standards for Science* in grades K-12 define a set of performance expectations for what students should know and be able to do and are derived from the National Research Council's *Framework for K-12 Science Education*, also known as the *Next Generation Science Standards (NGSS)*. The cut scores established through the standard setting process represent the level of competence students are expected to demonstrate on the assessment to be classified into each performance level.

Kentucky Science Assessment Performance Levels

Federal statute requires that any statewide assessment used for accountability purposes includes at least three achievement or performance levels. The performance levels relate student achievement on the Kentucky science assessments directly to what students are expected to learn, based on the standards in the *Kentucky Science Assessments for Science*. Student achievement on all Kentucky science assessments is classified into four performance levels that delineate the knowledge, skills, and abilities for which students are able to demonstrate mastery.

The policy-level performance level descriptors (PLDs) provide general expectations for student performance to be classified into each performance level on the Kentucky science assessments. These do not differentiate student performance between grade levels. The policy-level PLDs for the Kentucky science assessments were developed prior to the standard setting meeting and approved by the Kentucky Department of Education (KDE) for use during the standard setting meeting.

The four performance levels and their respective policy descriptions are shown in Table 1.

Table 1. Policy-level Descriptors for the Kentucky Science Assessment

| Performance Level | Policy Level Descriptors |
|----------------------|---|
| Distinguished | A student performing at the <i>Distinguished</i> level has a comprehensive understanding of science concepts and practices. The student consistently communicates ideas in a sophisticated and complex manner, using thorough supporting detail and explicit examples. The student reasons and solves problems by using appropriate strategies in an insightful way. Connections between science concepts/ideas, when appropriate, are justified and insightful. |
| Proficient | A student performing at the <i>Proficient</i> level has a broad understanding of science concepts and practices. The student usually communicates ideas accurately using clear and appropriate examples, supporting or justifying those ideas with relevant details and evidence. Problem-solving and critical thinking skills are used effectively. Connections between science concepts/ideas, when present are reasonable and appropriate. |
| Apprentice | A student performing at the <i>Apprentice</i> level has a basic understanding of science concepts and practices. The student communicates ideas in a basic manner, but explanations, solutions or justifications may be unclear or ineffective. The student demonstrates some problem-solving and critical thinking skills, but they are not consistently applied. |
| Novice | A student performing at the <i>Novice</i> level has a minimal understanding of science concepts and practices. The student communicates ideas ineffectively or inaccurately, providing little detail and little or no support. Attempts at problem solving or critical thinking are minimal or inappropriate. |

Kentucky Science Assessment Standard Setting Process

The recommendations by the Kentucky science grade 11 standard setting committee represent the level of competence students are expected to demonstrate to be classified into each of the performance levels. To establish performance levels, the Extended Modified (Yes/No) Angoff Method (Davis & Moyer, 2015; Plake, Ferdous, Impara, & Buckendahl, 2005) was used to guide panelists as they determined their performance level cut score recommendations. This standard setting procedure is a systematic method for combining various considerations into the process for recommending cut scores for the different performance levels, including content standards and educator judgments about what students should know, based on the *Kentucky Academic Standards for Science*, and be able to demonstrate at each performance level.

The following steps were used for the Kentucky Science Assessment standard setting process.

- Pre-meeting development – In anticipation of the standard setting meeting, various tasks were completed, including development of materials for the panelists, preparation of the Pearson Standard Setting website for panelists and facilitators, presentation materials for the facilitators, and development of data analysis sources and procedures.

- Standard setting meeting – A committee of panelists worked with Science assessment content and referenced borderline descriptions to make recommendations for cut scores that define the different performance levels for each assessment.
- Articulation – The recommended cut scores for each assessment were reviewed for reasonableness and alignment of performance level expectations across science grades 4, 7, and 11 by the members of the grade 11 standard setting committee.
- Development of grade-specific range PLDs - The members of the science grade 11 standard setting committee used an anchored item approach to develop range PLDs following the standard setting meeting.
- Reasonableness review - Meetings were held by KDE to review the reasonableness of the recommended cut scores in light of additional external data.

The remaining chapters of the technical report describe specific procedures and activities that occurred during each phase of the standard setting.

Chapter 2 – Preparations for the Standard Setting

Chapter 2 provides an overview of the work completed prior to the standard setting meeting for the Kentucky science grade 11 assessment, including:

- Development of participant materials
- Development of presentation materials
- Facilitator training
- Preparation for data analysis during the meetings

Development of Participant Materials

The Kentucky science grade 11 standard setting required a large number of materials for use by the participants during the meetings. The Pearson standard setting team worked with KDE to develop the materials used during the meeting and to ensure that all materials provided to meeting participants communicated accurate information. The following materials were developed for use by participants during the meeting:

- Meeting agenda
- Panelist information survey*
- Meeting non-disclosure agreement
- Test form for grade 11
- Experience the test activity response form
- Test form answer key*
- Open-ended item rubrics and exemplars*
- Practice judgment items*
- Practice judgment items answer key*
- Practice judgment record form
- Practice judgment survey*
- Judgment round record form
- Judgment round survey* – rounds 1, 2, and 3
- Ordered item set
- Process evaluations*

Because the standard setting meetings utilized the Pearson Standard Setting website as a tool for facilitating the meeting, the website for each committee needed to be developed. Several of the documents developed, which are indicated with an asterisk (*), were presented online through the website. After initial development of the websites for the meetings, a complete quality control check was performed to verify that the information provided on the websites matched the information presented on the documents.

Using approved templates, documents were created for the science grade 11 committee meeting by the Pearson standard setting team. All documents developed for the website were reviewed and approved by KDE staff before being finalized for publication for the meetings. A sample set of materials used by the committee are provided in [Appendix A](#).

Development of Presentation Materials

PowerPoint presentations were developed to guide facilitators through the presentation of information and materials throughout the standard setting meeting. The Pearson standard setting team developed the initial PowerPoint presentations. Staff from KDE had the opportunity to review and provide suggested edits to the presentations, which were resolved by the Pearson standard setting team. The following PowerPoint presentations were created for the standard setting meetings.

- General Session Presentation and Standard Setting Overview
- Standard Setting Breakout Meeting – Day 1
- Standard Setting Breakout Meeting – Day 2
- Standard Setting Breakout Meeting – Day 3

Facilitator Training

The breakout session was facilitated by a psychometrician from Pearson with knowledge and experience facilitating standard setting meetings. The facilitator was responsible for ensuring appropriate processes were followed throughout all sections of the meeting and that panelists had a solid understanding of the tasks they were asked to complete.

The facilitator underwent an extensive program of training to prepare for leading the set of standard setting meeting. The facilitator training included:

- Use of the Pearson standard setting website—Because the standard setting website was used as a facilitation tool during the meeting, the facilitator needed to become familiar with the use of the platform. The website provided a framework for facilitating the standard setting process. Specific guidelines for modeling the website and providing access to the panelists were discussed.
- Kentucky Science Assessments—The facilitator was provided an overview of the Kentucky science assessment program, including the content areas assessed, different item types, scoring rules, performance levels, and scaling design.
- Standard setting process—The facilitator participated in a walkthrough of the standard setting meeting agenda, with a focus on specific issues, such as time management, the use of the online platform, and communicating feedback information.
- Presentation slides and script—As part of the walkthrough of the standard setting process, the facilitator also reviewed the standard setting training slides. The script provided along with the presentation slides offered the facilitator guidance throughout the presentation, including when specific language was to be used during the panelist training and use of the standard setting website.

Preparation for Data Analysis During the Meetings

Creation and testing of analysis programs and the calculation of impact data lookup tables were conducted prior to the standard setting meeting. To facilitate the analysis for each judgment round during the meeting, analysts independently completed the programming necessary to conduct all analysis using SAS statistical software. A trial analysis was run with mock data generated through the standard setting website to ensure each independent analysis produced the same results.

Impact data are the percentage of students classified within each performance based on the recommended cut scores for a given judgment round. Analysis programs developed prior to the standard setting meetings used impact data lookup tables to create impact data during the meetings.

The impact data lookup tables were designed to represent the expected impact from the four operational forms using panelists' recommendations based on only one form. For each form, a unique raw score-to-ability value conversion table was created using student responses from the spring 2019 administration. These conversion tables were used to assign each student administered the test an ability value, so all student scores were on the same scale. Using the raw score-to-ability value table for the selected form, the impact data lookup tables were constructed to select the percentage of students that had ability values equal to or greater than the ability value associated with each possible raw score value for the test.

In addition to the programming created to calculate impact data, Pearson analysts developed programs to generate all feedback handouts, plots, and tables needed during the standard setting meeting. For example, following a round of judgment, the analyst produced:

- Individual panelist feedback – a listing of the judgments made by a panelist to ensure they were recorded and analyzed accurately (given to all panelists).
- Committee-level feedback – a summary of judgments from all panelists, including a frequency distribution of judgments and the mean, median, and range of the committee's cut score recommendation by performance level (given to facilitators and KDE; presented to panelists using tables and histograms in PowerPoint slides).
- Impact data (*after* judgment rounds 2 and 3) – the percentage of students, not disaggregated by demographic groups, in each performance level according to the recommended cut scores for that round (displayed to panelists as stacked bar graphs in digital presentations).

Chapter 3 – Standard Setting Meeting

Chapter 3 provides details about the process used for the Kentucky science grade 11 standard setting meeting. The sections of this chapter include:

- Purpose of standard setting meeting
- Committee panelist composition
- Standard setting meeting facilitator and staff
- Standard setting materials
- Standard setting procedure
- Standard setting meeting proceedings
- Recommended performance level cut scores

Purpose of the Standard Setting Meetings

Standard setting is based, to a large degree, on the judgment of educators. Committees of educators make expert recommendations about the level of achievement expected for each performance level based on their experience with different groups of students and knowledge of the assessed content. A specific process, or standard setting method, is used to capture the educator’s judgments and to translate them into cut scores for the performance levels. The purpose of the standard setting meeting was to gather expert recommendations from groups of educators from across Kentucky for the cut scores that define the different performance levels on the science grade 11 assessment.

Student performance on the Kentucky science grade 11 assessment was classified into one of four performance levels. Each committee was asked to recommend three cut scores that defined the boundaries between the different performance levels. The committee’s recommended cut scores represented the performance a student would need to meet or exceed to be classified into the specific performance level on the assessment.

Committee Panelist Composition

KDE started the process of selecting panelists for the standard setting meeting by requesting volunteers from across Kentucky. All panelists for the standard setting committee were selected by KDE from the individuals who volunteered and were then invited to participate in the standard setting meeting. The process of selecting committee panelists involved selection of a sample of panelists that would be as representative of the state as possible, including demographic variables (e.g., gender, race), geographic representation, and background (e.g., educational experience, education). When selecting panelists, KDE placed an emphasis on those educators who had relevant content knowledge as well as experience with a variety of student groups.

There was a total of 11 panelists at the standard setting meeting. The tables in [Appendix B](#) summarize the characteristics and experience of the panelists in the committee, including demographic information, current positions in education, experience working with various types of student populations, regional representation, and the types of districts they represent. Responses to the gender and ethnicity questions was voluntary.

The panelists in each committee were assigned to table groups. The table groups were selected prior to the meeting to ensure that, to the greatest extent possible, the panelists at each table were representative of the committee. The panelists were placed into table groups to facilitate discussions during the standard setting meetings and ensure that each participant had the opportunity to fully engage in the process.

Standard Setting Meeting Facilitators and Staff

Staff members from Pearson and KDE collaborated to conduct the standard setting meeting. Both groups worked in facilitative and observational roles during the meeting and took special care not to influence the committee's cut score recommendations.

Meeting Facilitator

The facilitator of the standard setting meeting was Mark Robeck, Ph.D., from Pearson. Mark Robeck, Ph.D. is a member of the Pearson psychometric staff who possesses experience facilitating standard setting meetings.

Meeting Data Analysts

For the standard setting meeting, two data analysts performed all analyses. The data analysts were Trey Heideman, who was on-site for the duration of the meeting, and Mike Watson. Both are members of the Statistical Analyst staff at Pearson. During the meeting, the analysts collected panelist judgment data from the Pearson standard setting website, performed independent analysis to verify results, and prepared panelist feedback reports.

KDE Staff

KDE staff members attended the standard setting meeting to observe the process, answer assessment and curriculum questions, and address policy questions. KDE staff also monitored the cut score recommendations for each performance level throughout the standard setting meetings.

Standard Setting Materials

The following section describes the materials used by the committee members during the standard setting breakout session. Separate materials were developed for the anchored PLD meeting, which will be discussed later in the report.

Pearson Standard Setting Website

The Pearson standard setting website was used as the online platform during the standard setting meeting. The website provided panelists access to the standard setting meeting materials and tools used to collect panelist judgments (see Figure 1).

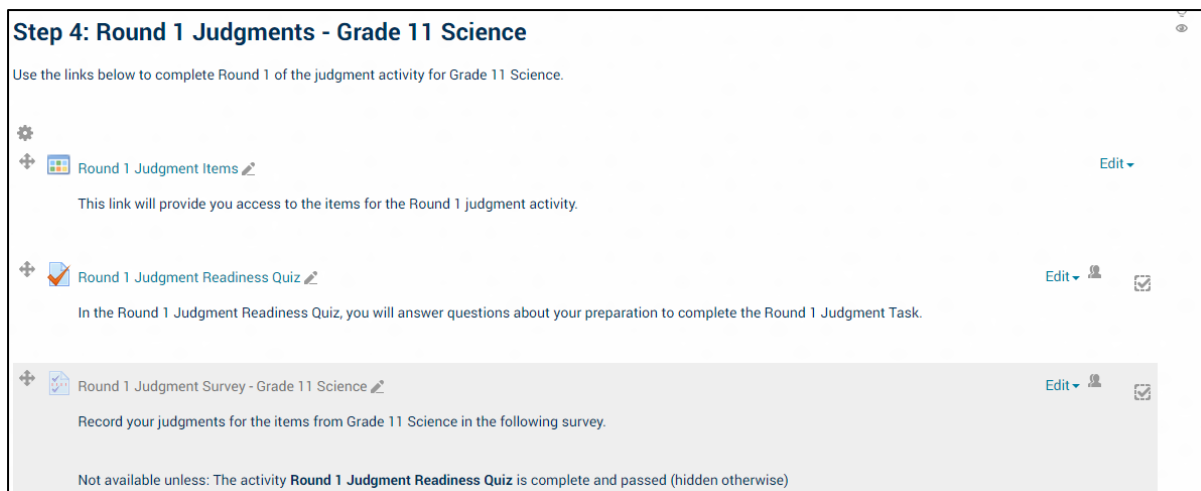


Figure 1. Example website interface with links to standard setting materials

The website was built using Moodle, an online, open source collaboration and learning tool. Each panelist was given unique login credentials that allowed secure access to the website. Panelists' access was restricted to only sections of the website associated with the standard setting meeting. Because the Kentucky science assessments are computer-delivered using TestNav 8, the standard setting website allowed panelists to view items as students did during the spring 2019 administration.

The website enabled participants access to online documents that provided background information about the Kentucky science assessments prior to the standard setting meeting. The preparation materials on the website included:

- Standard setting orientation video
- Kentucky science academic standards
- Kentucky science policy-level performance level descriptors
- Appendix F (Science and Engineering Practices) and Appendix G (Crosscutting Concepts) of the Next Generation Science Standards
- Kentucky standard setting non-disclosure agreement

The website also provided panelists access to materials and tools necessary for completing activities during the standard setting meeting. The standard setting materials and tools on the website included:

- Test item map and answer key
- Borderline descriptions worksheet
- Practice judgment activity items
- Practice judgment readiness survey
- Practice judgment survey
- Judgment items for rounds 1, 2, and 3
- Judgment readiness quiz for rounds 1, 2 and 3
- Judgment survey for rounds 1, 2, and 3
- Judgment feedback folders for rounds 1 and 2
- Articulation (Round 4) spreadsheet
- PLD development ordered item map
- PLD development ordered items
- PLD development worksheet
- Process evaluations 1, 2, and 3

A unique course site was created for the Kentucky science grade 11 committee in the Pearson standard setting website. The meeting facilitator controlled panelist access to each section of the website. Website access was disabled at the end of each meeting day to prevent panelists from viewing secure website materials outside of designated meeting times. Following the meetings, the online materials were archived.

Committee Panelist Folders

In addition to the online resources delivered through the website, panelists were provided a folder to organize a variety of hard copy materials they used throughout the meeting. The materials supplied to committee panelists in their folders included:

- Meeting agenda
- Non-disclosure agreement
- Kentucky science policy-level PLDs
- “Experience the assessment” activity response form
- Practice judgment form
- Rounds 1, 2, and 3 Judgment Record form

The panelist folders were prepared in advance of the standard setting meetings. Panelists were required to check-in at the start of each day and to return their folders and check-out at the end of each day of their meetings. Panelists were provided additional materials throughout the meeting, which they were instructed to insert into their folders.

Computers

Each panelist was provided a laptop computer in his or her meeting room to access the online resources through the website. The laptops were Dell latitudes with 15.6” screens, standard keyboards with a full-size number pad, and an external mouse. Panelists were not provided with external keyboards, numeric keypads, or monitors. Panelists were seated in table groups and provided enough space to freely work with the computer and folder materials. Power supplies for the computers were centrally located at the base of each table. The panelists used Google Chrome to access the standard setting website. Each computer was programmed with a whitelist of websites that restricted use to work associated with the standard setting meeting.

Standard Setting Procedure

To set performance standards, the Extended Modified (Yes/No) Angoff method was used. This standard-setting procedure operates as both a content- and item-based method that leads panelists through a standardized process in which they consider student expectations, as defined by the PLDs, and the individual items administered to recommend cut scores for each performance level.

The design of the Kentucky science assessment, which involves independent items clustered into item sets with associated stimuli, led to a modification of this method. For Rounds 1, panelists were asked to review each independent item from the operational administration and answer the following question:

“How many points would a student performing at the borderline of the [specific] performance level likely earn if he or she answered the question?”

For Round 2, in addition to the item-level judgment, panelists were asked to also review each cluster of items associated with the same stimuli and answer the following question:

“How many points would a student performing at the borderline of the [specific] performance level likely earn if he or she answered all the items associated with the cluster?”

In Round 3, panelists only provided judgments for each cluster of items on the assessment. For the standard setting meeting, “likely” was defined statistically as the student having at least a 67% chance of earning the number of points. The panelists completed the task for each performance level of the assessment. Between the judgment rounds, panelists were provided feedback information.

Standard Setting Meeting Proceedings

The standard setting meetings were conducted across three days, July 16-18, 2019, in Lexington, Kentucky. The complete agenda for the meetings is available in [Appendix C](#).

The remaining sections of Chapter 3 will describe the steps used to guide panelists through the entire standard setting process.

Standard Setting Meeting Pre-Work

The standard setting participants completed a set of activities prior to attending the meeting. The purpose of the pre-work was to expedite training by providing panelists an opportunity to become familiar with the information that would be used throughout the meeting. The pre-work included:

- Pearson standard setting website – The pre-work was provided via documentation or links embedded within the secure Pearson standard setting website developed for the meeting. The panelists were provided their unique login and temporary password through an email sent to the email address they provided during registration. The panelists were instructed to login to the website to complete the pre-work activities, which also gave them an opportunity to experience the website and navigate through the pre-work sections and activities.
- Participant information survey – Panelists completed a survey to document their demographic information as well as current teaching position, experience, and school information. Panelists were able to access the survey before and during the meeting.
- Standard setting orientation video – A short video was uploaded to the website to introduce panelists to the purpose and concepts associated with the Kentucky science standard setting meeting.
- Borderline descriptions development – Panelists developed draft borderline descriptions for a specific set of performance expectations as part of their pre-work activities. They were provided an instructional video describing the concepts of performance level descriptors, the expectations associated with borderline performance, and the steps needed to complete the activity. Panelists used their individual borderline description worksheets to identify the specific performance expectations and develop borderline expectations for each performance level.
- Security and Non-disclosure agreement – A Security and Non-disclosure agreement was uploaded to the website for panelists to review prior to the standard setting meeting. The intention was to familiarize panelists with security protocol in advance of the meeting so they would be familiar with expectations when requested to sign the agreement at the meeting.

Breakout Session

Because standards were being set for only science grade 11, a single breakout session was held that spanned three days. During the breakout session, the committee was responsible for providing cut score recommendations for each of the performance levels of the science grade 11 test. An overview of the activities conducted each day of the breakout session is provided in Table 2. The presentation slides used during the breakout session are available in [Appendix D](#).

Table 2. Overview of Activities During Breakout Sessions

| Day 1 Activities | Day 2 Activities | Day 3 Activities |
|---|--|----------------------------------|
| Introductions and process overview | Round 1 judgments | Articulation (Round 4 judgments) |
| 'Experience the Assessment' activity | Discussion of Round 1 judgments and feedback | Anchored PLD development |
| Review of standards and policy-level PLDs | Round 2 judgments | Evaluation and closing remarks |
| Development of borderline descriptions | Discussion of Round 2 judgments and feedback | |
| Standard setting training | Round 3 judgments | |
| Practice judgment activity and discussion | Discussion of Round 3 judgments and feedback | |

Introductions and Overview. To begin the breakout session, individuals in the room—the facilitator, panelists, and observers—introduced themselves by sharing the following:

- Name
- Area of the state
- Experience in current field
- Role and any courses taught
- Experience with KAS test committees

After introductions, the facilitator discussed the security and non-disclosure expectations for the meeting. The panelists then individually reviewed, agreed to, and signed the security and non-disclosure agreement.

The facilitator also distributed folders containing secure and essential materials for the meeting. The facilitator reviewed the documents and materials in the folder, on the standard

setting website, and how the resources would be used during the standard setting process. The panelists were given an opportunity to ask questions before proceeding.

The overview concluded with a presentation of the Kentucky science assessment system, including the use of classroom-embedded assessments, ‘through course tasks,’ and a statewide summative test. Alignment between the Kentucky Academic Standards and Next Generation Science Standards was shown to panelists as well. Lastly, the purpose of the statewide summative assessment was shared and a description of the test design was provided.

Experience the Assessment. Panelists experienced one of the four operational test forms administered to students during the spring 2019 administration. The panelists experienced the test through the same online system used by students. The ‘Experience the Assessment’ activity allowed panelists to interact with the test items and develop insight regarding the knowledge and skills required to correctly answer the test items. Panelists were trained on specific scoring rules used for particular items on the test. For constructed-response items, the panelists were introduced to rubrics and notes used to score student responses as well as student exemplars that demonstrated responses receiving different scores.

Panelists recorded their responses to the ‘Experience the Assessment’ items on a separate form, which was provided in their folder. After the panelists completed the activity, they were given information about how the assessment for their assigned subject is scored. A test map, or online answer key, on the standard setting website provided information about each item, including the unique item number, correct response for the item, maximum number of points, associated learning standard (i.e., KAS), and the associated science engineering practice(s), disciplinary core idea(s), and crosscutting concept(s). Panelists were given an opportunity to review the correct responses and score their test using the test map on the website.

Borderline Descriptions. An essential component to the standard setting process is the development of borderline descriptions. As part of their pre-work activities, panelists individually developed draft borderline descriptions for select performance expectations and wrote them in interactive worksheets on the standard setting website. Prior to the breakout meeting, the facilitator collected the draft borderline descriptions and grouped them into worksheets by table that the panelists used during the breakout session.

To help inform the borderline descriptions development activity, the facilitator began by reviewing the performance levels and the policy-level PLDs for the Kentucky science grade 11 assessment with panelists. The review provided panelists with a common understanding of the knowledge, skills, and abilities typical students should demonstrate within each performance level. Additionally, panelists participated in a group discussion regarding the differences between the expectations at the various performance levels.

The panelists were then introduced to the difference between a student with *typical* performance and a student with performance at the *borderline* of a performance level. A student with performance at the borderline was described as one who possessed “just-barely” enough knowledge, skills, and abilities to be classified into a specific performance level.

The table groups reviewed the draft borderline descriptions from the pre-work activity they completed and made revisions to more clearly define the expectations for students at the borderline of each performance level. The borderline descriptions from each table were then compiled into a master document and reviewed by the whole committee. Edits were made to the master document during the whole-group discussion to create a common set of

borderline descriptions. The final list of borderline descriptions was printed and given to each panelist as a reference for subsequent activities.

Judgment Process Training. The process facilitator provided panelists with training on the Extended Modified (Yes/No) Angoff standard setting procedure and how to use the website to record their individual judgments. Panelists were instructed to review each item from the assessment, consider the knowledge and skills necessary to answer the question, and consult the borderline descriptions during the judgment process. Based on their review of the item and the related materials, panelists answered the following question for each of the three performance levels:

“How many points would a student performing at the borderline of the [specific] performance level likely earn if he or she answered the question?”

Significant time was spent describing the thought process the panelists should go through using parts of the question.

- “Would...”—When considering expected student response to an item, the panelists were asked to consider how a student would respond. Where “should” is an aspirational expectation, “would” is a more realistic expectation of a student response to an item.
- “...a student performing at the borderline of the [specific] performance level...”—The panelists were reminded to reference the borderline descriptions for the specific performance level to determine how a student performing at the borderline of that performance level would be expected to respond.
- “...likely...”—In this context, likely was defined as 2 out of 3 times, or 67%. To make this concrete for panelists, facilitators asked them to think about 3 students at the borderline of a performance level. If a panelist believed 2 out of 3 students with performance at the borderline would correctly answer the item, they would respond “yes” to the question. If a panelist did not believe 2 out of 3 students with performance at the borderline would correctly answer the item, they would respond “no” to the question.
- “...earn if he or she answered the question.”—Panelists selected the number of points a student with performance at the borderline would be expected to earn if he or she answered the item.

Panelists were instructed to review each item and make a judgment for each performance level, starting with *Apprentice* and then proceeding to *Proficient* and *Distinguished*. Panelists were trained to check their judgments for expected patterns across performance levels, which included multiple examples with different judgment patterns. The judgments made by panelists were recorded in the judgment survey via the standard setting website. Figure 2 shows an example item from the judgment survey on the website.

| Item: SCHS1626_01 | | | | | |
|-------------------|--------|----------|---|--------------------------|--------------------|
| Key | Points | KAS | SEP | DCI | CC |
| | 1 | HS-PS2-1 | 1 Using Mathematics and Computational Thinking 2 Analyzing and Interpreting Data | PS2.A: Forces and Motion | 2 Cause and Effect |

| | 0 Points | 1 Point |
|---------------|-----------------------|-----------------------|
| Apprentice | <input type="radio"/> | <input type="radio"/> |
| Proficient | <input type="radio"/> | <input type="radio"/> |
| Distinguished | <input type="radio"/> | <input type="radio"/> |

Figure 2. Example item from the judgment survey in the website

Panelists also kept a record of their judgments on their paper Judgment Record Sheet, which was provided as part of the materials in their folder. The Judgment Record Sheet included the unique item number, KAS, correct response, maximum score, and judgment for each performance level. Panelists were shown how to use the unique item number to ensure they referenced the correct item on both the paper Judgment Record Sheet and online judgment survey.

Practice Judgment Activity. Panelists completed a practice judgment activity prior to beginning the actual judgment rounds. The goals of this activity were to:

- Give panelists experience reviewing and making judgments about different types of items.
- Familiarize panelists with the judgment survey on the standard setting website.
- Build confidence in their understanding of the task to be completed.

Eight items were selected for the practice activity. The practice items were a subset of those panelists ultimately reviewed in the actual judgment rounds and included examples of different item types, difficulty, and score points. After all panelists completed their practice judgments, the facilitator presented item-level judgment results interactively through the standard setting website. Group discussion was initiated to review the judgment process and panelist responses, demonstrate how their judgments are used to determine a cut score recommendation, and answer any questions.

Judgment Rounds. After receiving training on the standard setting process, the panelists worked through three rounds of judgments. Before starting each of the three judgment rounds, the facilitator reviewed the judgment process, including explicit instructions on which materials were needed for the judgment task. Panelists were required to complete a readiness survey in the website prior to each round, which indicated they understood the task and process used to complete the judgments. The panelists were required to answer “yes” to all readiness survey questions before continuing with the judgment round. If a panelist responded “no” to any question, he/she was asked to notify the facilitator for additional assistance. The readiness survey included the following questions:

- Do you understand your task for the judgment activity? (Rounds 1, 2, and 3)
- Are you ready to begin the judgment activity? (Rounds 1, 2, and 3)

An example of the readiness survey panelists completed before starting the judgment task is presented in Figure 3.

Readiness Survey:
 Before starting the activity, select a response for each of the following questions.
 Do you understand your task for the Judgment activity?

Select one:

Yes

No

Are you ready to begin the Judgment activity?

Select one:

Yes

No

Figure 3. Example readiness survey

After panelists finished the readiness survey, they were provided access to the judgment survey for the respective round.

During the first judgment round, panelists made individual judgments for each item, based on the borderline descriptions and knowledge and skills required by the item. Panelists answered the question, “How many points would a student performing at the borderline of the [specific] performance level likely earn if he or she answered the question?” Panelists completed judgments on both the paper Judgment Record Sheet and in the judgment survey for all performance levels before moving onto the next item.

In Round 2, panelists were asked to review each item and each cluster of items associated with the same stimuli. Panelists first made their judgments for each individual item of a cluster, just as was done in Round 1. Once they reached the final item of a cluster, panelists were asked “How many points would a student performing at the borderline of the [specific] performance level likely earn if he or she answered all the items associated with the cluster?” An example of the prompt displayed to panelists for the cluster judgments in Round 2 is shown in Figure 4.

“How many points would a student performing at the borderline of the [specific] performance level likely earn if he or she answered all the items associated with the cluster?”

Cluster: SCHS1622

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| Apprentice | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Proficient | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Distinguished | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

Figure 4. Example of the cluster judgment prompt in Round 2

In Round 3, panelists only provided cluster-level judgments. The items were grouped into clusters in the judgment survey for panelists in Round 3 and they had the opportunity to review each item and the stimuli associated with each cluster.

Feedback and Discussion. The panelists were given feedback after each judgment round. The feedback was based on each individual's current cut score recommendations, the recommendations of others in their committee, and relevant information from actual student results on the assessment. Feedback data included the following:

- Information about panelists' cut scores for each performance level:
 - Individual cut scores: Judgments were summed across items to obtain a cut score for each level. The panelists were provided individual paper handouts showing their judgments and recommended cut score for each performance level.
 - Committee cut score recommendations and statistics: Committee-level recommendations were the median cut score across all panelists for each performance level. Panelists were provided the committee-level cut score recommendations and cut score statistics (minimum, maximum, median, mean, and standard deviation) for each performance level.
 - Panelist agreement data: Bar graphs showing the frequency of individual recommended cut scores for each performance level and across adjacent performance levels.
- Item-level judgment agreement across panelists: Distribution of panelist judgments for each item and performance level.
- Item means (p-values) and score-point distributions: Average score earned by students for each item and the distribution of score points, for polytomously scored items, calculated from operational test data.
- Impact data: Percentage of students that would be classified into each performance level, based on the committee's current recommended cut scores and the results of students who took the assessment during the Spring 2019 administration.

Some information was provided only after certain rounds. The feedback information shared with panelists after each judgment round is shown in Table 3.

Table 3. Feedback Data Provided to Panelists after Each Judgment Round

| Feedback Data | | Judgment Round | | |
|----------------------------|------------------------------|----------------|---------|---------|
| | | Round 1 | Round 2 | Round 3 |
| Item-Level Feedback | Panelist Agreement Data | ✓ | ✓ | |
| | Item Means | ✓ | | |
| | Score Point Distributions | ✓ | | |
| Test-Level Feedback | Individual Cut Score | ✓ | ✓ | ✓ |
| | Committee Cut Score | ✓ | ✓ | ✓ |
| | Panelist Item Agreement Data | ✓ | ✓ | |
| | Impact Data | | ✓ | ✓ |

[Appendix E](#) provides examples of each of the feedback data provided to participants, along with a brief description of the feedback presented.

Before the discussions of feedback data, panelists were given guidance regarding the independence of their judgments. That is, they were encouraged to listen to other panelists and consider the rationales given for their judgments, but they should not feel pressured to reach consensus. Following Rounds 1 and 2, panelists shared the rationale for their judgments during table-group and whole-group discussions. Items with the highest level of disagreement amongst the committee were revisited for each performance level. Committee members discussed a range of topics, such as item difficulty, student strategies when responding to the items, their individual rationale for a judgment, and, importantly, the borderline descriptions the group crafted. The goal of discussions was to demonstrate to panelists how their judgments compared to the rest of the committee and to guide them toward a common and shared understanding of the borderline descriptions and judgment task. After Round 2, panelists also participated in a whole-group discussion about the impact data and whether it matched expectations, given the student population.

Process Evaluations. The validity of standard setting outcomes relies on procedural validity. Evidence of procedural validity was gathered through three evaluation surveys administered throughout the standard setting meeting. The evaluations focused on the processes and procedures of the meeting, including the panelists’ overall views of the standard-setting process, training, materials, meeting facilitation, and ultimately how they feel about the final results. The evaluations were kept anonymous. The results from the evaluations were aggregated and can be found in [Appendix F](#). All panelists were also allowed to provide any additional information concerning their evaluation of the process of the standard setting meeting through an open-response question.

Recommended Cut Scores from Standard Setting Committee

The median cut score recommendation from the committee was used to establish the cut score for each performance level. The cut score recommendations resulting from the Round 3 judgments were considered the committee’s final recommendations for the standard

setting meeting. The recommended cut scores for each performance level based on the Round 3 recommendations are displayed in Table 4.

Table 4. Round 3 Cut Score Recommendations from Standard Setting Committee

| Grade | Maximum Score | Apprentice | | Proficient | | Distinguished | |
|-------|---------------|------------|-----------|------------|-----------|---------------|-----------|
| | | Raw Score | % Correct | Raw Score | % Correct | Raw Score | % Correct |
| 11 | 48 | 12 | 25.0% | 22 | 45.8% | 38 | 79.2% |

The estimated impact data after judgment Round 3 are illustrated in Figure 5 for each performance level.

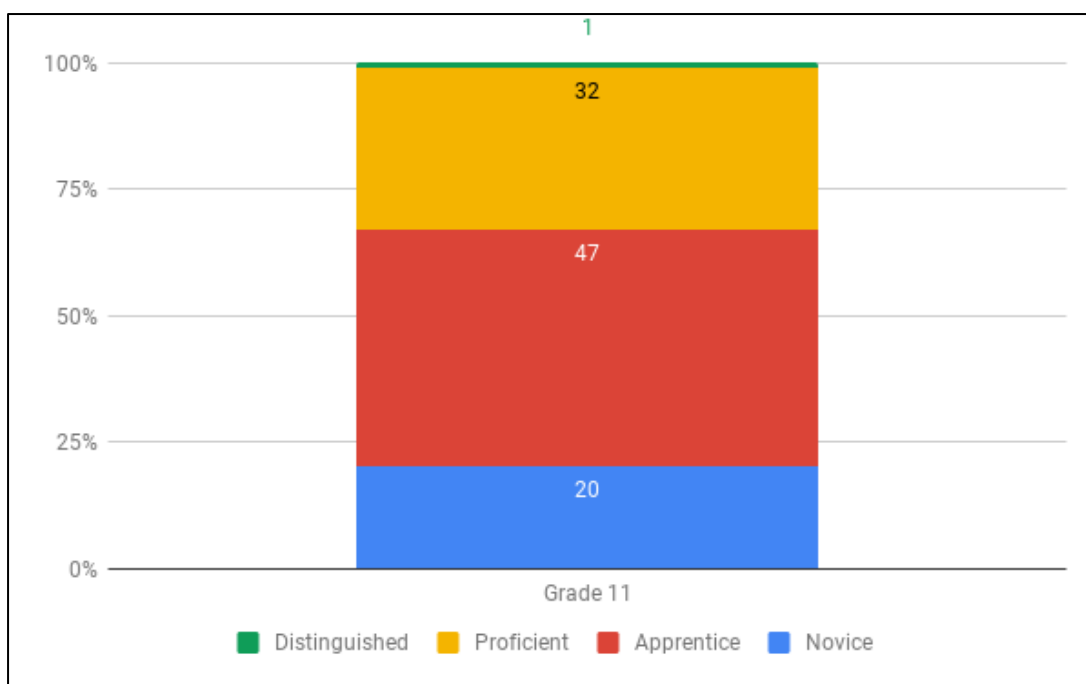


Figure 5. Impact data from Round 3 cut score recommendations

The recommended cut scores at the end of each judgment round are presented for all performance levels in [Appendix G](#). Summary statistics for the recommended cut scores for at the end of each judgment round are shown in [Appendix H](#). Panelist agreement data after each judgment round are displayed by performance level in [Appendix I](#).

Articulation

The purpose of the articulation, which was referred to as ‘Round 4 judgments’ at the meeting, was to review and evaluate the reasonableness of the committee’s cut score recommendations. In the first three judgment rounds, the recommendations from the standard setting committee were made with a specific focus on the respective content for science grade 11. The focus of the articulation was to evaluate whether the cut score recommendations across grades resulted in a cohesive assessment system.

The panelists were guided through a specific process in which they reviewed and discussed the cut scores established for Kentucky science grades 4 and 7 in relation to the recommendations for science grade 11. Then, the committee was tasked with determining if

changes to the cut score recommendations were necessary to represent a coherent set of student expectations across science grades.

The participants in the articulation process included all panelists who were in the science grade 11 standard setting committee. The articulation occurred on Thursday, July 18, 2019, the final day of the standard setting meeting. The session was led by Mark Robeck, Ph.D., who was also the process facilitator of the standard setting meeting. The participants remained in the same table groups they were assigned during the breakout session.

Articulation Process

The articulation process involved two steps:

- Review and discussion of the cross-grade impact data and ACT science data
- Discuss necessary adjustments to recommended cut scores

At the beginning of the articulation process, the panelists were informed of the purpose of the task, which was to review the adopted standards of the science grades 4 and 7 standard setting meetings as well as the recommended cut scores from the science grade 11 meeting. In the standard setting breakout session, panelists were focused primarily on the content related to their committee, whereas during the articulation process, they were asked to consider the adopted standards and cut score recommendations from a policy perspective to ensure the results represented a cohesive assessment system.

The panelists were shown cross-grade impact data charts that reflected the results from the adopted standards for science grades 4 and 7, ACT Science data, and the Round 3 judgments of the and 11 standard setting committee. Impact data using the adopted standards for science grades 4 and 7 are presented in Figure 6. Impact data using the cut score recommendations for science grade 11 can be found in Figure 5.

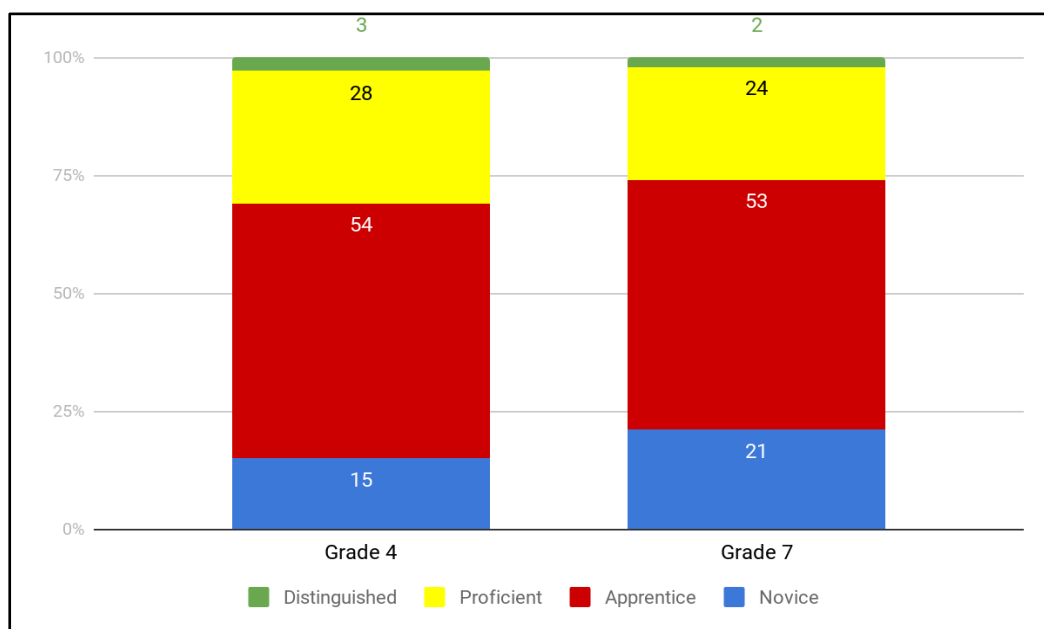


Figure 6. Impact data using adopted standards for science grades 4 and 7

After the impact data were presented, the panelists engaged in table-group discussion of the results and how they aligned with their initial expectations. Following discussion, the panelists were provided an opportunity to investigate changes to the Round 3 recommended cut scores for grade 11 using an interactive spreadsheet. The interactive spreadsheet used

as part of the articulation was accessed through the Pearson standard setting website and is presented in Figure 7, which includes the impact data after Round 4.

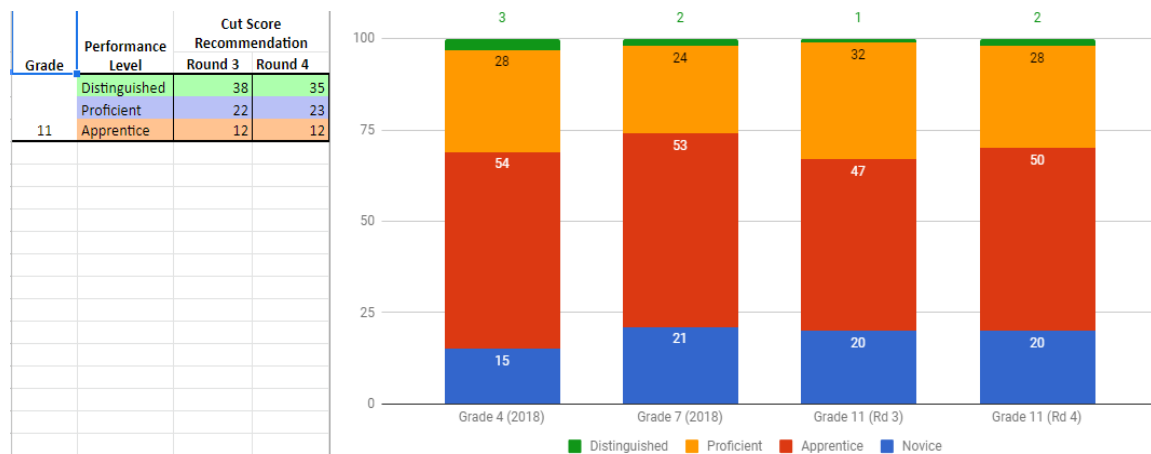


Figure 7. Interactive spreadsheet used for Round 4 judgments

The interactive spreadsheet allowed panelists to view how possible modifications to the current cut score recommendations resulted in changes to the impact. The committee was given an opportunity to discuss and recommend changes to cut scores for the performance levels if they noticed a misalignment in the impact data. When a cut score change was recommended, the meeting facilitator input the changes into the interactive spreadsheet for the entire committee to review the resulting impact data.

The panelists discussed the differences in impact data across the grades, ACT scores, and their impressions of the Round 3 cut score recommendations for grade 11. Based on the discussion, the committee recommended adjustments to the cut scores for the Proficient and Distinguished performance levels.

At the end of articulation, the panelists were reminded of the review and approval process before cut score implementation. Panelists also completed an evaluation of the articulation process on the standard setting website.

Performance Level Descriptor Development

PLDs are statements that articulate the knowledge, skills, and abilities that students classified into a particular performance level should be able to demonstrate. All Kentucky science assessments have four performance levels, as defined in Table 1. The performance levels range from Novice, representing the lowest level of student performance, to Distinguished, representing the highest level of student performance.

The PLDs are associated with the performance levels in the following way:

- *Performance levels* indicate a student’s level of competency of the standards, defined in the Kentucky Academic Standards for Science through classification of their performance on an assessment for the specific grade as *Novice*, *Apprentice*, *Proficient*, or *Distinguished*.
- *PLDs* indicate the knowledge, skills, and abilities students should demonstrate in each grade level to be classified into a performance level.
- *Cut scores* partition the test scale and represent the minimum test score a student must earn on each grade-level assessment to be classified into a given performance level.

Development of the PLDs for the Kentucky science grade 11 assessment was done after the articulation activity, when the committee cut score recommendations were finalized for each performance level. The panelists of the standard setting meeting worked in their table groups to develop the PLDs. The next section will describe the process panelists were led through to develop the PLDs for grade 11.

Meeting Process

Because the PLDs were drafted at the end of the standard setting meeting, an anchored development process was applied to facilitate the development of the PLDs. The anchored development process used the ability level associated with the committee's cut score recommendations to align the items for the assessment with each performance level. The knowledge, skills, and abilities needed to respond to each set of items were then used to define the expectations communicated by the PLDs for each performance level—*Apprentice*, *Proficient*, and *Distinguished*. There were no PLDs developed for the *Novice* performance level.

The set of ordered items was essential to the implementation of the anchored PLD development process. The ordered item set presented the test items from the spring 2019 administration of the Kentucky science grade 11 assessment from easiest to most difficult. The order of item difficulty was based on the Rasch item parameters for each item, which was determined using student data from the spring 2019 administration. Each multiple-choice item was represented one time in the item set. Polytomously-scored items, which including short answer and extended response items and have a maximum score of greater than one, were represented in the set one time for each non-zero score point. Polytomous machine-scored items (i.e., multiple-select items) were not represented in the item set. An item map was created to communicate the order of the items in the set to panelists, along with other item information, such as answer keys and associated learning standards.

The items associated with each performance level were determined using the cut score recommendations for each performance level from the articulation activity (i.e., Round 4 judgments). The ability level associated with each cut score recommendation was determined using a raw score-to-ability level conversion table. For each item in the ordered set, the ability value associated with a 67% probability of providing a correct response was determined, which is also known as an RP67 value. The item with an RP67 value closest to but greater than or equal to the cut score ability level was the first item in the set associated with the performance level. The item set for a performance level included the first item associated with the performance level until the first item associated with the next performance level. The item range associated with each performance level is displayed in Table 5.

Table 5. Item Ranges Associated with each Performance Level

| Grade | Novice | | Apprentice | | Proficient | | Distinguished | |
|-------|--------|-----|------------|-----|------------|-----|---------------|-----|
| | Min | Max | Min | Max | Min | Max | Min | Max |
| 11 | 1 | 7 | 8 | 35 | 36 | 86 | 87 | 120 |

The facilitator introduced the ordered item set to the panelists by discussing its construction and the relationship between the items in the set and the performance levels. The panelists were provided with the item sequences that separated the item sets associated with each performance level. The panelists worked in their table groups to review the items associated with each performance level assigned to their group. Based on their item review, the panelists defined the knowledge, skills, and abilities that represented a reasonable set of expectations for students classified into each performance level.

After the table groups created their draft PLDs, the expectations were collected into a master document and shared with the whole group for final review. The participants suggested and discussed edits for the draft PLDs of each performance level. The facilitator led the discussion by revising to the master document as the panelists made suggestions. The whole-group PLD review was intended to ensure consistency in student expectations across performance levels within grade 11.

The final PLDs from the grade 11 committee were reviewed by KDE and revisions were made to ensure comparability in the PLDs across grades. The final grade 11 PLDs review are available in [Appendix J](#).

Chapter 4 – Post-Standard Setting

Chapter 4 provides details about the process used after the standard setting to approve the performance level standards for the Kentucky science grade 11 assessment. The sections of this chapter include:

- Reasonableness review

Reasonableness Review

Following the standard setting meeting, an executive summary was provided to KDE to facilitate a review of the science grade 11 cut score recommendations. The executive summary included a brief overview of the methodology and process used to obtain the cut score recommendations, the panelist's cut score recommendations for each performance level, and the impact data associated with the recommended cut scores. This summary was provided to KDE on Thursday, July 18, 2019.

Using the executive summary, KDE reviewed the reasonableness of the cut score recommendations for the Kentucky science assessments. The purpose of this review was to evaluate the reasonableness and alignment of the recommendations with other data, expectations for alignment across grades, and usefulness in the communication of results within the context of the state accountability system. Members from KDE along with technical advisors for the science assessment program participated in the reasonableness review discussion.

The recommendation from the reasonableness review was to approve the standard setting committee's recommended standards.

Chapter 5 – Evidence of Procedural Validity of the Standard Setting Process

Chapter 5 details evidence supporting the validity of the process used for the standard setting meeting. The sections in Chapter 5 include the following:

- Committee representation
- Committee training
- Perceived validity of the standard setting

Committee Representation

As part of the recruitment process, KDE collected demographic information about panelists' background relevant to educational experience and representativeness of the teaching population in Kentucky. Full results of the demographic information collected from panelists is available in Appendix B, including current position (Table B.1), professional experience teaching in education (Table B.2), professional experience teaching in science grade 11 (Table B.3), experience with different student populations (Table B.4), education (Table B.5), region of Kentucky (Table B.6), gender (Table B.7), ethnicity (Table B.8), race (Table B.9), if currently working in a school district (Table B.10), size of school district (Table B.11), type of school district (Table B.12), and socioeconomic status of school district (Table B.13).

A majority of the panelists in each committee were classroom teachers in K-12. Most panelists had at least 10 years teaching experience and nearly half had 20 years or more and there was a mix of experience teaching science grade 11, specifically. The committee members had a diverse array of experience teaching different student populations, including general education, special education, English language learners, and vocational technical education.

Over 90 percent of panelists had at least a Master's degree. The panelists were representative of the different regions of Kentucky and the different types of school districts across the state. A large majority of panelists were currently working school districts and reflected the state, including size, type, and socioeconomic status.

Committee Training

It was essential that panelists understood how to make judgments as part of the Extended Modified (Yes/No) Angoff standard setting methodology. Training on the standard setting methodology was provided throughout the process. Training and implementation of the standard setting process was standardized through the PowerPoint training slides, script, and materials used.

Panelists completed practice judgment round as an opportunity to apply the standard setting methodology without consequence. During the practice judgment round, the panelists reviewed a reduced set of items and provided judgments for three performance levels, *Apprentice*, *Proficient*, and *Distinguished*. After the practice round, a whole-group discussion

was led by the process facilitator to identify and respond to any questions or issues panelists encountered while implementing the standard setting process. Before each judgment round, panelists responded to a readiness survey that confirmed they were prepared to make their judgments. Panelists were not permitted to begin the judgment survey unless they answered “Yes” to all questions on the readiness survey and were encouraged to ask the facilitator for clarification if they responded “No” to any question.

Panelists completed an evaluation survey at the end of the standard setting meeting to record their impressions of the effectiveness of the materials and methods employed through the process. As part of the process evaluation survey, the panelists were asked to provide their thoughts about the effectiveness of a few different components of the training they received for the standard setting. All panelists believed the training provided on the standard-setting process was either *Adequate* or *More than Adequate*, as shown in Figure 8.

Training provided on the standard-setting process

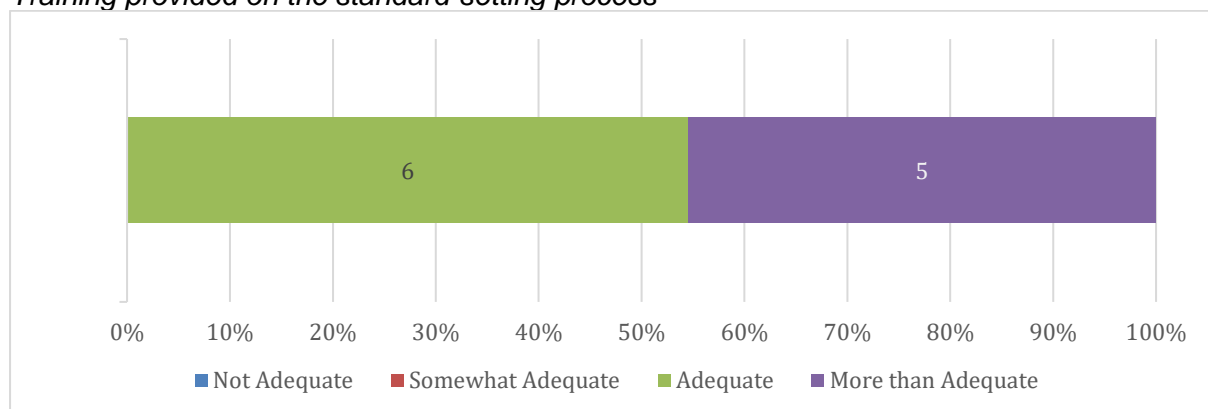


Figure 8. Process evaluation results regarding clear explanation for purpose of standard setting

Likewise, all panelists felt the amount of time to complete and discuss the borderline descriptions and results of the practice judgment activity was *Adequate* or *More than Adequate*. These results are shown in Figures 9 and 10, respectively.

Total amount of time to create and discuss borderline descriptions

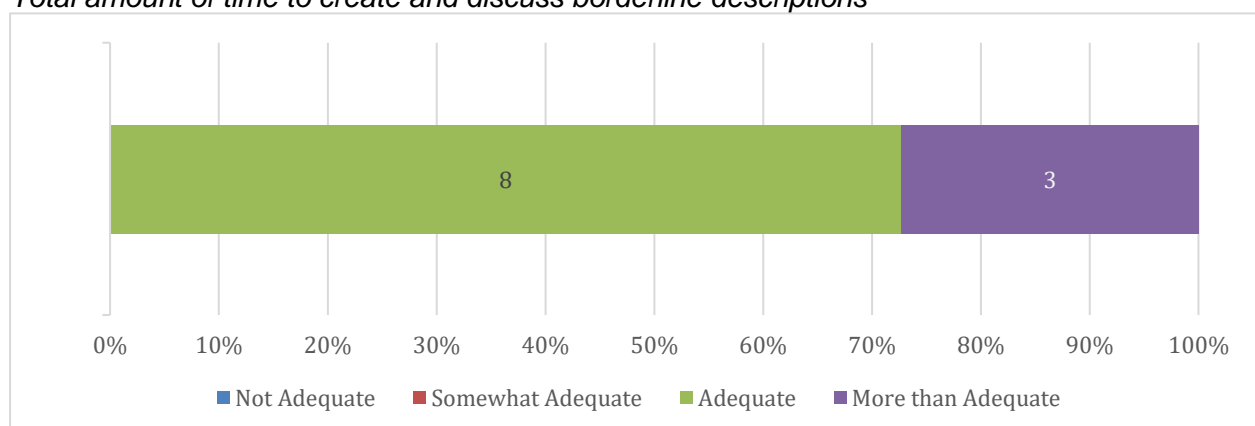


Figure 9. Process evaluation results regarding amount of time spent on borderline descriptions activity

Total amount of time to discuss the practice judgments

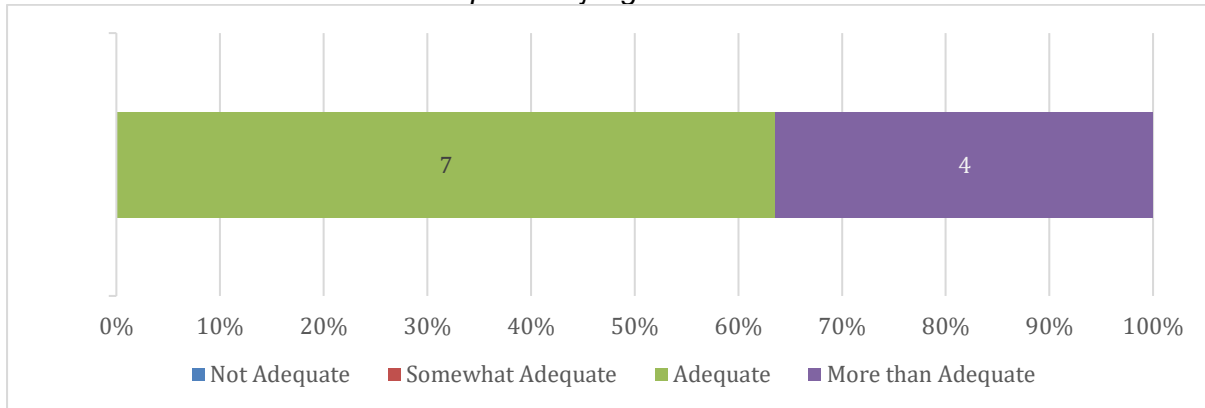


Figure 10. Process evaluation results regarding amount of time spent on practice judgment activity

Overall feedback from the panelists about the standard setting was positive. Most indicated that it was a valuable experience and that it was facilitated well. Below are select comments from the free-response question in the evaluation:

“This was a most enjoyable learning experience. To see how cut scores and descriptors were created really help me to better understand how the NGSS is connected. The facility was clean and cool. The facilitator was organized and personable. Thank you.”

“I am very pleased with all of the facilitators and the process. I think everything was explained very thoroughly and that we were able to make decisions as a whole group without personal opinions getting in the way.”

“WOW! Great job Pearson and the opportunity for teachers to lead and drive this process is appreciated THANK you KDE for this great learning experience :) Rae you are a superstar!”

Full results from the process evaluations are presented in [Appendix F](#).

Perceived Validity of the Standard Setting

Panelists communicated their perceived validity of the standard setting and the recommended cut scores as part of the process evaluation. Generally, the panelists were satisfied with their cut score recommendations and the standard setting process, as a whole. Results from the evaluation survey, displayed in Figures 11, 12 and 13, indicated most panelists had confidence in the committee’s recommended cut score for all performance levels.

How confident do you feel that the final cut score recommendations for Science Grade 11 represent appropriate levels of student performance?

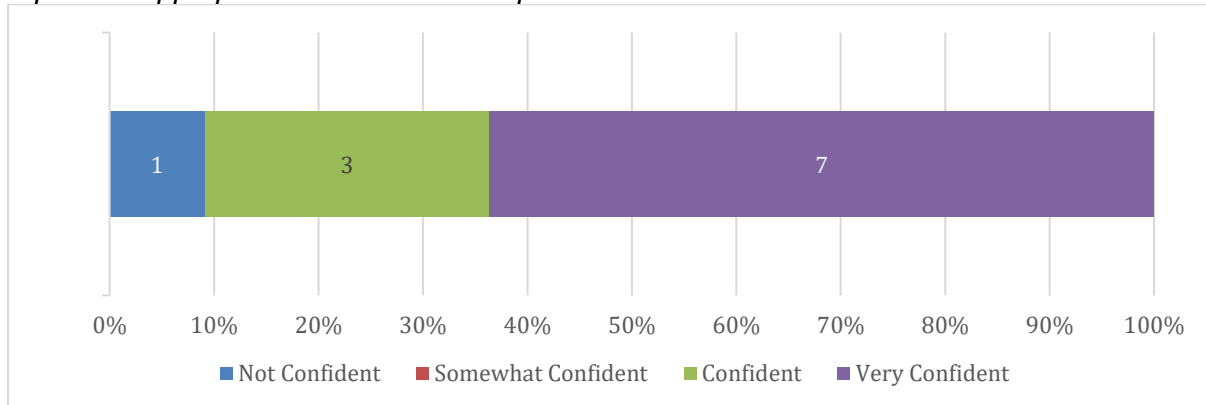


Figure 11. Process evaluation results regarding the final recommended cut scores for Apprentice

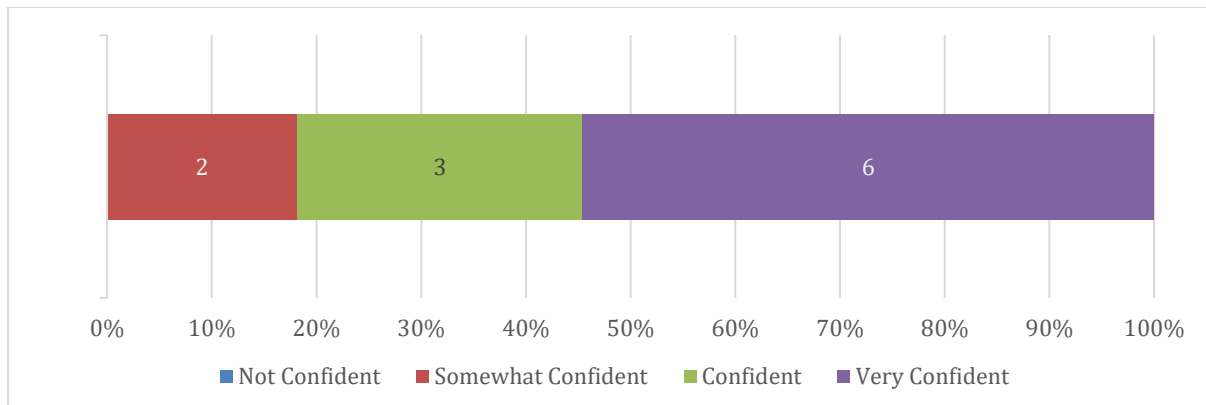


Figure 12. Process evaluation results regarding the final recommended cut scores for Proficient

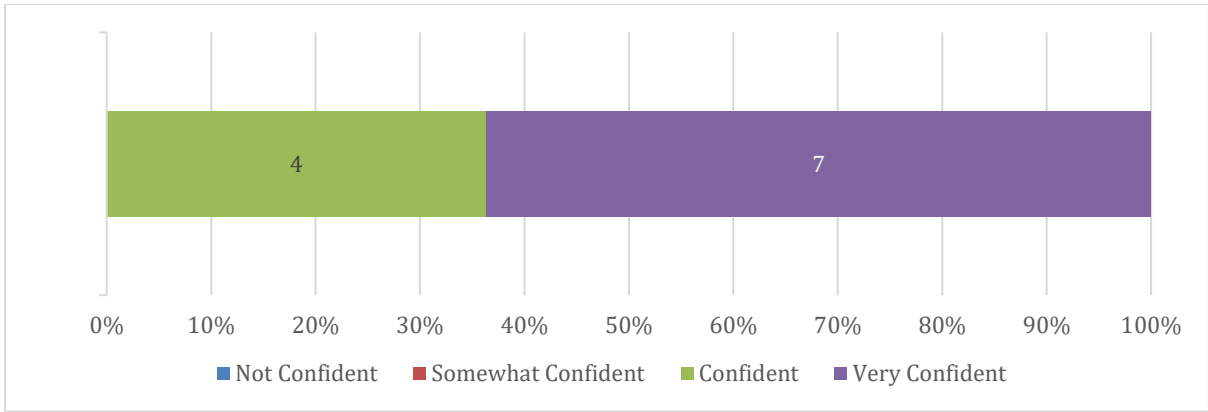


Figure 13. Process evaluation results regarding the final recommended cut scores for Distinguished

Overall, feedback from the standard setting panelists provides evidence for the validity of the cut score recommendations for each of the performance levels.

References

- Davis, L. L. & Moyer, E. L. (2015). PARCC performance level setting technical report. Available from Partnership for Assessment of Readiness for College and Careers (PARCC), Washington, D.C.
- Plake, B. S., Ferdous, A. A., Impara, J. C., & Buckendahl, C. W. (2005). *Setting Multiple Performance Standards Using the Yes/No Method: An Alternative Item Mapping Method*. Meeting of the National Council on Measurement in Education. Montreal, Canada.

Appendix A – Panelist Meeting Materials

The materials developed for the Kentucky science grade 11 standard setting committee are provided as an example of what was shared with the panelists. Because the materials provided to participants contained secure information, not all documents will be presented in Appendix A. Specifically, the following materials will not be available in the appendix:

- Test form – This was presented to panelists using actual student test books from the spring 2019 administration.
- Open-ended item rubrics – These documents presented the scoring rubrics and scoring notes for each open-ended item presented to panelists.
- Student exemplars – These documents presented student-produced responses for each open-ended item presented to panelists.
- Practice judgment items – This was presented to panelists through the Pearson Standard Setting website as a pdf document.

Kentucky Science Assessment Standard Setting Meeting Grade 11

Agenda

Day 1 - Tuesday, July 16

Introductions and Meeting Orientation

Standard Setting Overview

Experience the Assessment

Kentucky Standards Performance Levels

Lunch

Borderline Descriptions

Standard Setting Training

Practice Judgment Activity

Day 2 - Wednesday, July 17

Round 1 Judgments

Round 1 Judgment Feedback and Discussion

Round 2 Judgments

Lunch

Round 2 Judgment Feedback and Discussion

Round 3 Judgments

Round 3 Judgment Feedback and Discussion

Day 3 - Thursday, July 18

Articulation (Round 4 Judgments)

Performance Level Descriptor (PLD) Development

Next Steps and Evaluations



Kentucky State-Required Assessments Nondisclosure Agreement Form

Kentucky state-required assessments requires that all materials used during the standard setting process remain secure. To protect the security of the test items, only authorized persons are permitted to work with or view the materials. All test items and/or components of items, draft or final, and all supporting assessment materials or notes, student responses, and feedback from the standard setting process are to be regarded as secure documents. Thus, they may not be reproduced, discussed, or in any way released or distributed to unauthorized personnel during or after the standard setting process. As a member of the standard setting committee, you may not use any information gleaned from the standard setting process to gain/provide an unfair advantage to schools/districts.

The undersigned is an employee, contractor, consultant, advisory committee member or person otherwise authorized to view material associated with the standard setting process, and hereby agrees to be bound to the terms of this agreement restricting the disclosure of said materials.

Name (printed)

Signature

Date

Kentucky Science Assessment Standard Setting Meeting

Participant Information Survey Grade 11

Professional Experience

What is your current position?

- Teacher (K-12 Education)
- Teacher (Higher Education)
- Administrator (School)
- Administrator (District)
- Other Position:

How many years of professional experience in education do you have?

- None
- 1 to 5 years
- 6 to 10 years
- 11 to 15 years
- 16 to 20 years
- More than 20 years

How many years of professional experience do you have teaching Science grade 11?

- None
- 1 to 5 years
- 6 to 10 years
- 11 to 15 years
- 16 to 20 years
- More than 20 years

For which of the following populations do you have educational experience with?

(Check all that apply.)

- Students receiving mainstream special education services
- Students receiving self-contained special education services
- Students who are English language learners
- Students who are receiving general education instruction
- Students who are receiving vocational technical instruction

What is the highest degree you have completed?

- High School Diploma
- Associates degree (A.A., A.S.)
- Bachelors degree (B.A., B.S.)
- Masters degree (M.A., M.S.)
- Doctoral degree (Ph.D., Ed.D.)

Demographic Information

The map displays different regions of Kentucky.



In which region of Kentucky do you reside?

- West
- North West
- South West
- North Central
- South Central
- North East
- South East

What is your gender?

- Male
- Female
- No answer

What is your ethnicity?

- Hispanic or Latino
- Not Hispanic or Latino
- No answer

What is your race?

- American Indian or Alaskan Native
- Asian
- Black or African American
- Native Hawaiian or Pacific Islander
- White
- No answer

Do you currently work in a school district?

- Yes
- No

Kentucky Science Standard Setting Meeting July 2019

Experience the Assessment Record Sheet Grade 11 Science

| Sequence | Item ID | Passage | KAS by Topic | Max Points | Response/Notes |
|----------|-------------|-----------------|------------------------------|------------|----------------|
| 1 | SCHS1626_01 | Arrestor Cables | HS-PS2-1 | 1 | |
| 2 | SCHS1626_02 | | 08-PS3-1, HS-PS2-1, HS-PS2-3 | 2 | |
| 3 | SCHS1626_03 | | HS-PS2-1 | 1 | |
| 4 | SCHS1626_04 | | HS-PS2-3 | 1 | |
| 5 | SCHS1626_06 | | 08-PS3-1 | 1 | |
| 6 | SCHS1626_07 | | 08-PS3-1 | 1 | |
| 7 | SCHS1626_10 | | 08-PS3-1 | 4 | |
| 8 | SCHS1626_09 | | HS-PS2-1 | 1 | |
| 9 | SCHS1622_09 | Sustainability | HS-ESS3-1 | 1 | |
| 10 | SCHS1622_02 | | HS-ESS3-1 | 1 | |

Note: Only the first page of this document is presented as an example.

Kentucky Science Grade 11 Test Map

| Seq | UIN | Key | Points | KAS | SEP | DCI | CC |
|-----|-------------|-----|--------|------------------------------|---|---|-----------------------------------|
| 1 | SCHS1626_01 | | 1 | HS-PS2-1 | 1 Using Mathematics and Computational Thinking 2 Analyzing and Interpreting Data | PS2.A: Forces and Motion | 2 Cause and Effect |
| 2 | SCHS1626_02 | | 2 | 08-PS3-1, HS-PS2-1, HS-PS2-3 | 2 Analyzing and Interpreting Data 3 Using Mathematics and Computational Thinking | PS2.A: Forces and Motion | 2 Cause and Effect |
| 3 | SCHS1626_03 | | 1 | HS-PS2-1 | 2 Analyzing and Interpreting Data 3 Using Mathematics and Computational Thinking | PS2.A: Forces and Motion | 2 Cause and Effect |
| 4 | SCHS1626_04 | | 1 | HS-PS2-3 | 4 Constructing Explanations and Designing Solutions | PS2.A: Forces and Motion | 2 Cause and Effect |
| 5 | SCHS1626_06 | | 1 | 08-PS3-1 | 1 Analyzing and Interpreting Data | PS3.A: Definitions of Energy PS3.B: Conservation of Energy and Energy Transfer | 1 Scale, Proportion, and Quantity |
| 6 | SCHS1626_07 | | 1 | 08-PS3-1 | 1 Analyzing and Interpreting Data | PS3.A: Definitions of Energy PS3.B: Conservation of Energy and Energy Transfer | |

Note: Only these rows of the test map are presented as an example.

Kentucky Science Standard Setting Meeting

July 2019

Judgment Round

Record Sheet

Science Grade 11

"How many points would a student performing at the borderline of the [specific] performance level likely earn if he or she answered the question?"

| Seq | Item ID | KAS by Topic | Answer Key | Max Score | Judgment Round | | | | | | | | | |
|--------------------------------|-------------|------------------------------|------------|-----------|-------------------------------------|---|---|---|---|---|---|---|---|--|
| | | | | | 1 | | | 2 | | | 3 | | | |
| | | | | | A | P | D | A | P | D | A | P | D | |
| 1 | SCHS1626_01 | HS-PS2-1 | | 1 | | | | | | | | | | |
| 2 | SCHS1626_02 | 08-PS3-1, HS-PS2-1, HS-PS2-3 | | 2 | | | | | | | | | | |
| 3 | SCHS1626_03 | HS-PS2-1 | | 1 | | | | | | | | | | |
| 4 | SCHS1626_04 | HS-PS2-3 | | 1 | | | | | | | | | | |
| 5 | SCHS1626_06 | 08-PS3-1 | | 1 | | | | | | | | | | |
| 6 | SCHS1626_07 | 08-PS3-1 | | 1 | | | | | | | | | | |
| 7 | SCHS1626_10 | 08-PS3-1 | | 4 | | | | | | | | | | |
| 8 | SCHS1626_09 | HS-PS2-1 | | 1 | | | | | | | | | | |
| CLUSTER SCHS1626 SCORE: | | | | 12 | Panelist's CLUSTER Judgment: | | | | | | | | | |

A=Apprentice; P=Proficient; D=Distinguished

Note: Only the first page of the Judgment Round Record Sheet is shown as an example.

Kentucky Science Assessment Standard Setting Meeting Grade 11

Practice Judgment Survey

You are now ready to begin!

For each Practice Judgment item, do the following for each performance level:

- Review the item in the online system.
- Review the information provided about the item in the item map and answer key. For open response items, review the information in the rubric and exemplars.
- Review the borderline descriptions for the performance level.
- Answer the following question:

"How many points would a student performing at the borderline of the [specific] performance level likely earn if he or she answered the question?"

- Record your response to the question for the performance level for the specific item on the judgment record sheet and in the online survey.


Continue reviewing the items until you have provided judgments for each performance level for all of the items.

You will now start the Judgment Process for the practice items.

For each of the items, answer the following question:

"How many points would a student performing at the borderline of the [specific] performance level likely earn if he or she answered the question?"


Item: SCHS1626_01

| Key | KAS | SEP | DCI | CC |
|---|----------|---|--------------------------|--------------------|
|  | HS-PS2-1 | 1 Using Mathematics and Computational Thinking 2 Analyzing and Interpreting Data | PS2.A: Forces and Motion | 2 Cause and Effect |

Apprentice
Proficient
Distinguished

| | 0 Points | 1 Point |
|---------------|-----------------------|-----------------------|
| Apprentice | <input type="radio"/> | <input type="radio"/> |
| Proficient | <input type="radio"/> | <input type="radio"/> |
| Distinguished | <input type="radio"/> | <input type="radio"/> |

Item: SCHS1626_02

| Key | KAS | SEP | DCI | CC |
|---|------------------------------|---|--------------------------|--------------------|
|  | 08-PS3-1, HS-PS2-1, HS-PS2-3 | 2 Analyzing and Interpreting Data 3 Using Mathematics and Computational Thinking | PS2.A: Forces and Motion | 2 Cause and Effect |

Apprentice
Proficient
Distinguished

| | 0 Points | 1 Point | 2 Points |
|---------------|-----------------------|-----------------------|-----------------------|
| Apprentice | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Proficient | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Distinguished | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

Note: Only two items are displayed as an example.

Appendix B – Committee Panelist Composition

Table B.1: Current Position

| | Science Grade 11 |
|--------------------------|------------------|
| Teacher (K–12) | 7 |
| Teacher (Higher Ed.) | 1 |
| Administrator (School) | 0 |
| Administrator (District) | 2 |
| Other | 1 |

Table B.2: Years of Professional Experience in Education

| | Science Grade 11 |
|--------------------|------------------|
| None | 0 |
| 1 to 5 years | 1 |
| 6 to 10 years | 2 |
| 11 to 15 years | 1 |
| 16 to 20 years | 2 |
| More than 20 years | 5 |

Table B.3: Years of Teaching Experience in Science Grade 11

| | Science Grade 11 |
|--------------------|------------------|
| None | 1 |
| 1 to 5 years | 3 |
| 6 to 10 years | 2 |
| 11 to 15 years | 0 |
| 16 to 20 years | 2 |
| More than 20 years | 3 |

Table B.4: Experience Teaching Student Populations

| | Science Grade 11 |
|----------------------------------|------------------|
| Mainstream special education | 10 |
| Self-contained special education | 2 |
| English language learners (ELL) | 6 |
| General education | 11 |
| Vocational technical education | 6 |

Table B.5: Highest Education Degree

| | Science Grade 11 |
|---------------------|------------------|
| High School Diploma | 0 |
| Associates degree | 0 |
| Bachelor's degree | 1 |
| Master's degree | 9 |
| Doctoral degree | 1 |

Table B.6: Demographic: Regions of Kentucky

| | Science Grade 11 |
|---------------|------------------|
| West | 2 |
| North West | 1 |
| South West | 1 |
| North Central | 4 |
| South Central | 1 |
| North East | 1 |
| South East | 1 |

Table B.7: Demographic: Gender

| | Science Grade 11 |
|-----------|------------------|
| Male | 4 |
| Female | 7 |
| No answer | 0 |

Table B.8: Demographic: Ethnicity

| | Science Grade 11 |
|------------------------|------------------|
| Hispanic or Latino | 0 |
| Not Hispanic or Latino | 11 |
| No answer | 0 |

Table B.9: Demographic: Race

| | Science Grade 11 |
|-------------------------------------|------------------|
| American Indian or Alaskan Native | 0 |
| Asian | 0 |
| Black or African American | 1 |
| Native Hawaiian or Pacific Islander | 0 |
| White | 10 |
| No answer | 0 |

Table B.10: Currently Work in a School District

| | Science Grade 11 |
|-----|------------------|
| Yes | 10 |
| No | 1 |

Table B.11: Size of School District

| | Science Grade 11 |
|--------|------------------|
| Small | 3 |
| Medium | 4 |
| Large | 3 |

Table B.12: Type of School District

| | Science Grade 11 |
|--------------------|------------------|
| Rural | 6 |
| Metropolitan/Urban | 2 |
| Suburban | 2 |

Table B.13: Socioeconomic Status of School District

| | Science Grade 11 |
|----------|------------------|
| Low | 5 |
| Moderate | 5 |
| High | 0 |

Appendix C – Standard Setting Meeting Agenda

Kentucky Science Assessment Standard Setting Meeting Grade 11 Agenda

Day 1

| | |
|------------------|---|
| 8:00 – 8:45 am | Welcome and Orientation Welcome and meeting purpose Introductions Materials orientations Meeting security |
| 8:45 – 9:15 am | Standard Setting Overview |
| 9:15 – 10:30 am | Experience the Assessment |
| 10:30 – 10:45 am | <i>Break</i> |
| 10:45 – 11:30 am | Review and Discuss Standards and Policy Level Descriptors |
| 11:30 – 12:15 pm | <i>Lunch</i> |
| 12:15 – 12:45 pm | Borderline Descriptions Training |
| 12:45 – 1:30 pm | Borderline Descriptions Table Discussion |
| 1:30 – 1:45 pm | <i>Break</i> |
| 1:45 – 3:00 pm | Borderline Descriptions Whole-Group Discussion |
| 3:00 – 3:30 pm | Standard Setting Training |
| 3:30 – 4:30 pm | Practice Judgment Activity and Discussion |

Day 2

| | |
|------------------|--|
| 8:00 – 9:15 am | Round 1 Judgments (Item level judgments) Round 1 Readiness Form Panelists work independently to make Round 1 judgments |
| 9:15 – 9:45 am | <i>Break</i> |
| 9:45 – 10:15 am | Round 1 Judgment Feedback Item Level – Item means and distributions Test Level – Cut score recommendations; Panelist agreement |
| 10:15 – 10:45 am | Table Discussion – Round 1 Feedback Panelists discuss feedback data at their tables |
| 10:45 – 11:45 am | Round 2 Judgments (Item and cluster level judgments) Round 2 Readiness form Panelists work independently to make Round 2 judgments |
| 11:45 – 12:30 pm | <i>Lunch</i> |
| 12:30 – 12:45 pm | Round 2 Judgment Feedback Item Level – Item means and distributions Test Level – Cut score recommendations; Panelist agreement |
| 12:45 – 1:15 pm | Table Discussion – Round 2 Feedback |
| 1:15 – 2:00 pm | Whole-Group Discussion – Round 2 Feedback |
| 2:00 – 2:45 pm | Round 3 Judgments (Cluster judgments) Round 3 Readiness form Panelists work independently to make Round 3 judgments |
| 2:45 – 3:15 pm | <i>Break</i> |
| 3:15 – 3:30 pm | Round 3 Judgment Feedback and Discussion Test level – Cut score recommendations Impact data |
| 3:30 – 4:30 pm | Whole-Group Discussion – Round 3 Feedback |

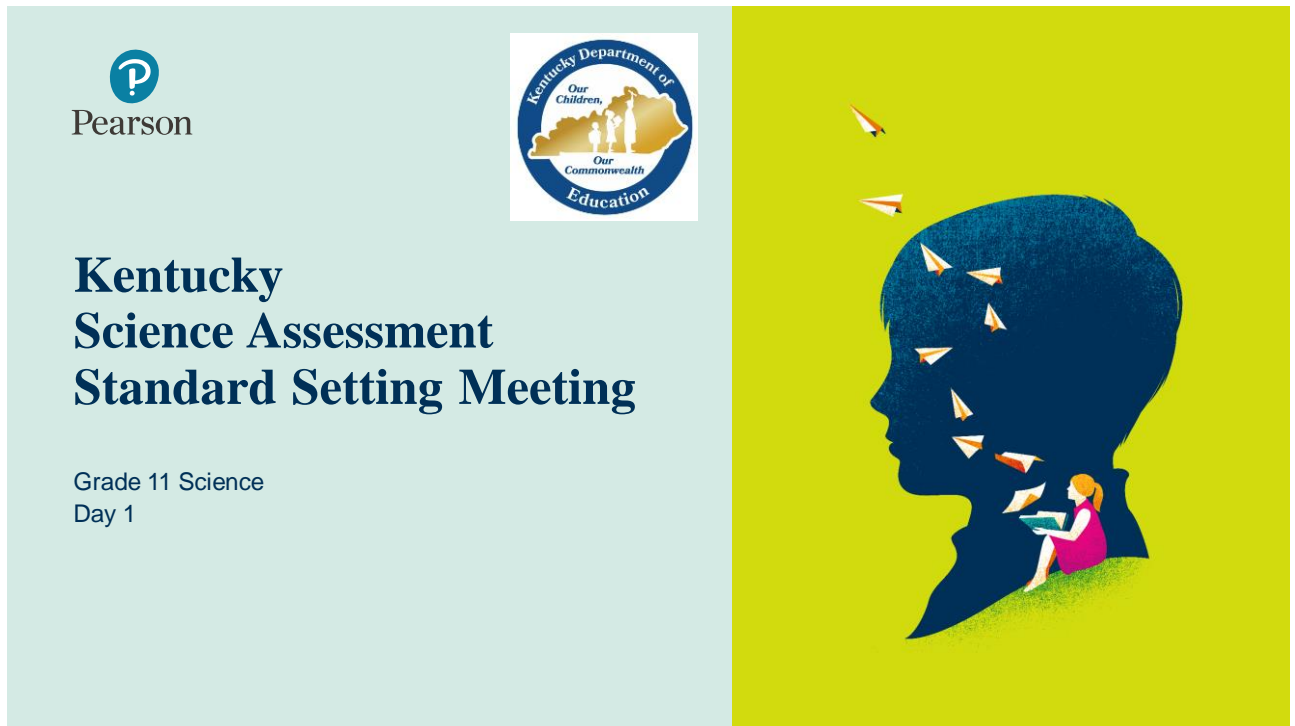
Day 3

| | |
|------------------|--|
| 8:00 – 9:30 am | Articulation (Round 4 Judgments) Impact data review Cut score change recommendations Review final cut score recommendations |
| 9:30 – 9:45 am | <i>Break</i> |
| 9:45 – 10:15 am | Performance Level Descriptor (PLD) Development Process Introduction to the anchor process Orientation to the ordered item book (OIB) |
| 10:45 – 11:30 am | Table Discussion for Proficient PLDs |
| 11:30 – 12:15 am | Whole-Group Discussion for Proficient PLDs |
| 12:45 – 1:00 pm | <i>Lunch</i> |
| 1:00 – 1:45 pm | Table Discussion for Apprentice PLDs |
| 1:45 – 2:30 pm | Whole-Group Discussion for Apprentice PLDs |
| 2:30 – 2:45 pm | <i>Break</i> |
| 2:45 – 3:30 pm | Table Discussion for Distinguished PLDs |
| 3:30 – 4:15 pm | Whole-Group Discussion for Distinguished PLDs |
| 4:15 – 4:30 pm | Next Steps, Process Evaluation, and Close Out |

Appendix D – Presentations

The presentations for each day of the standard setting are embedded in Appendix D. Double-click the cover slide to view the full presentation for a given day.

Science Grade 11 Breakout Session – Day 1

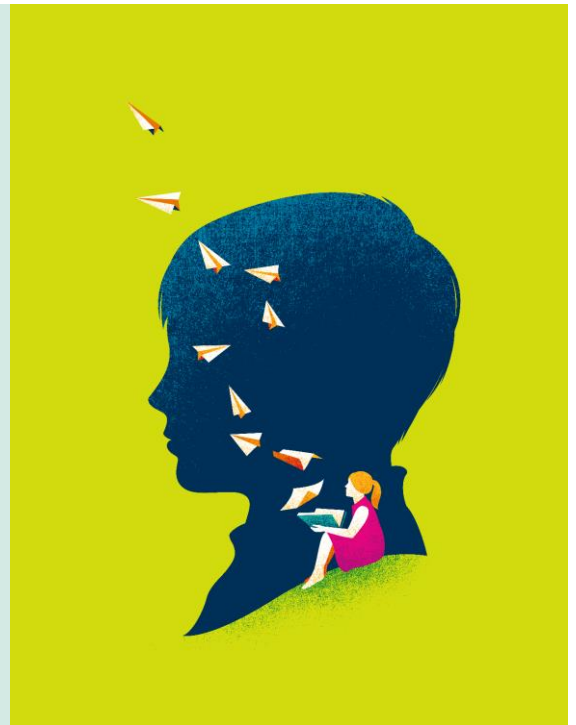


Science Grade 11 Breakout Session – Day 2



Kentucky Science Assessment Standard Setting Meeting

Grade 11 Science
Day 2

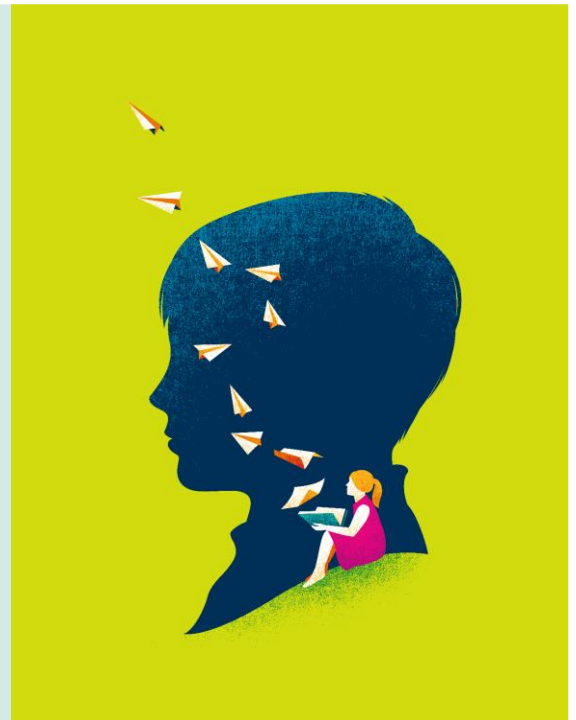


Science Grade 11 Breakout Session – Day 3



Kentucky Science Assessment Standard Setting Meeting

Grade 11 Science
Day 3



Appendix E – Examples of Feedback Data

Feedback data were provided to participants after each judgment round. The following are examples of feedback data provided to participants.

Individual Item-Level Judgments

The graphic below shows an example of the item-level judgments recorded in the judgment survey during Rounds 1 and 2. The individual item-level judgments were provided to panelists so they could verify the system accurately recorded their judgments for each performance level -- Apprentice (A), Proficient (P) and Distinguished (D).

| UIN | A | P | D |
|-------------|---|---|---|
| SCHS1626_01 | 0 | 1 | 1 |
| SCHS1626_02 | 0 | 1 | 2 |
| SCHS1626_03 | 0 | 1 | 1 |
| SCHS1626_04 | 0 | 1 | 1 |
| SCHS1626_06 | 1 | 1 | 1 |
| SCHS1626_07 | 0 | 1 | 1 |

Individual Cluster-Level Judgments

Panelists were also given handouts of their cluster-level judgments recorded in the judgment survey for Rounds 2 and 3. The individual cluster-level judgments were provided to panelists so they could verify the system accurately recorded their judgments for each performance level -- Apprentice (A), Proficient (P) and Distinguished (D).

| UIN | A | P | D |
|----------|---|---|----|
| SCHS1626 | 1 | 7 | 10 |
| SCHS1622 | 3 | 5 | 9 |
| BI1708 | 3 | 8 | 10 |
| BI1718 | 2 | 9 | 11 |

Individual Test-Level Recommendation

Each panelist was provided their test-level cut score recommendations, which was the sum of their judgments for the Apprentice (A), Proficient (P), and Distinguished (D) performance levels.

| A Raw Score | P Raw Score | D Raw Score |
|-------------|-------------|-------------|
| 16 | 38 | 45 |

Table-Level and Overall Test-Level Recommendation

Panelists were provided with both their table's and the overall committee's aggregate test-level cut score recommendations, including the number of participants, the mean, median, minimum, maximum, and the first and third quartile cut score recommendations for each performance level.

| | N | Mean | Median | Min | Max | Q1 | Q3 |
|-------------|----|-------|--------|-------|-------|-------|-------|
| A Raw Score | 11 | 8.09 | 8.00 | 4.00 | 16.00 | 6.00 | 9.00 |
| P Raw Score | 11 | 26.00 | 27.00 | 17.00 | 33.00 | 25.00 | 29.00 |
| D Raw Score | 11 | 39.18 | 41.00 | 31.00 | 43.00 | 38.00 | 42.00 |

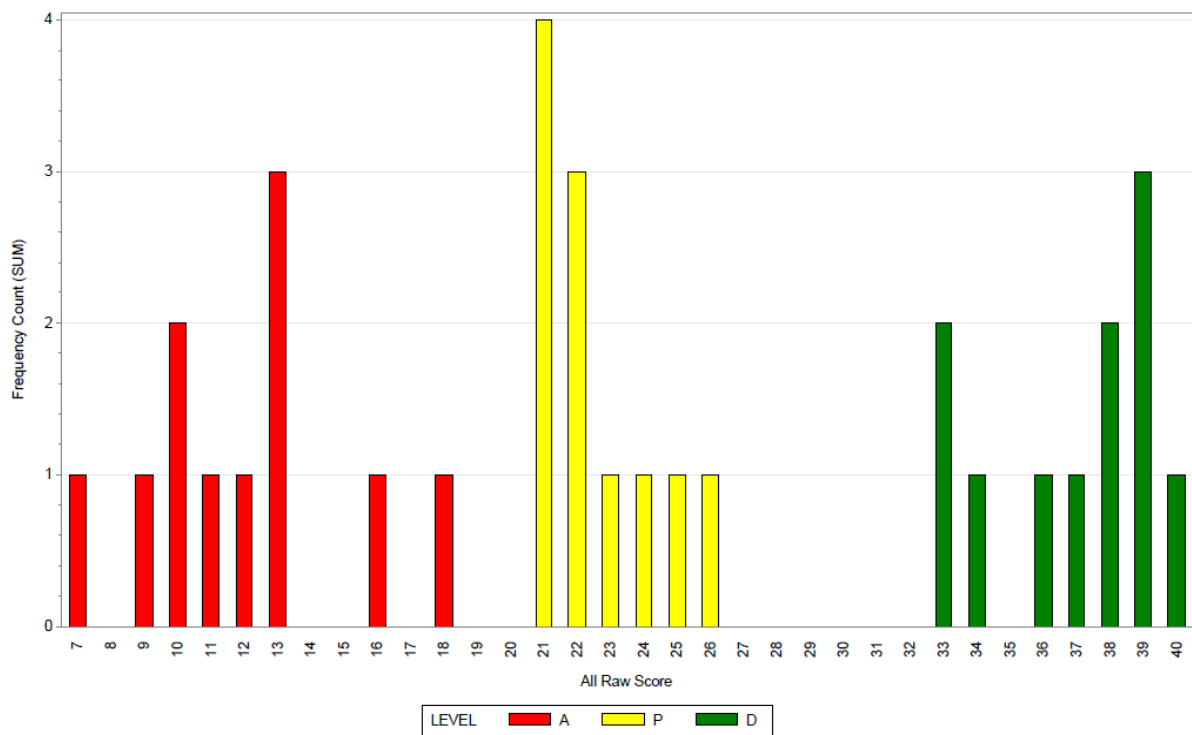
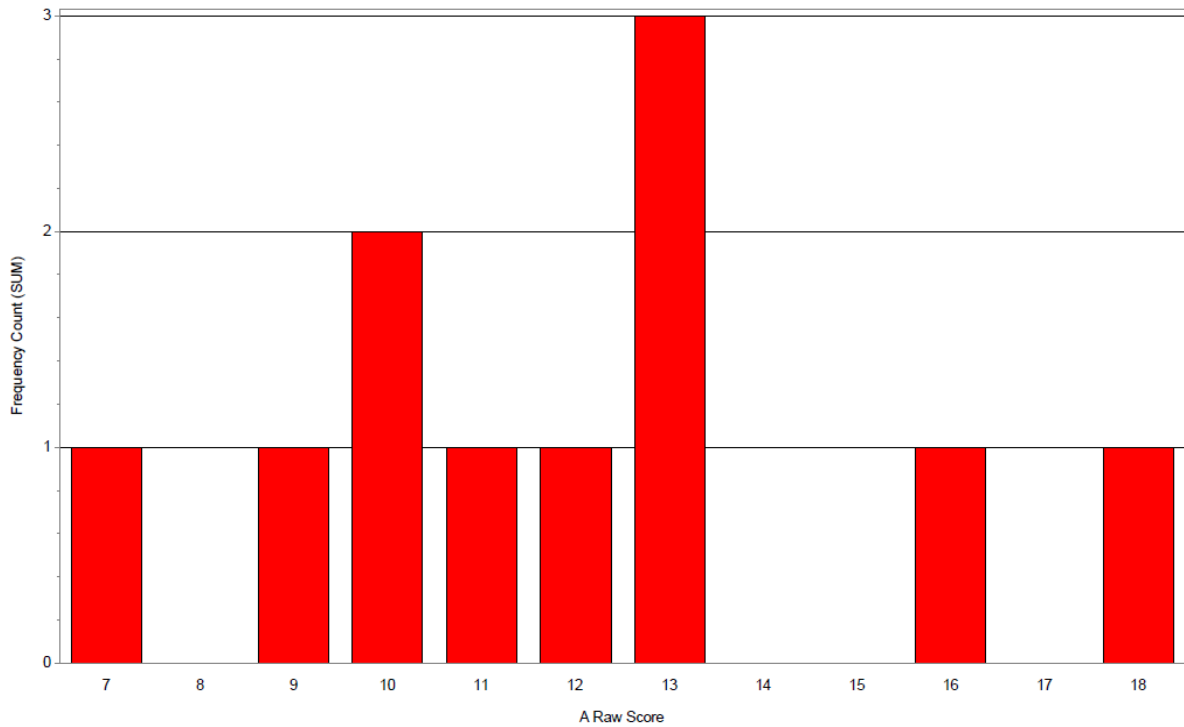
Item-Level Judgment Agreement

Item-level judgment distributions for the committee were provided to panelists for each item and performance level judgment. Additionally, for each performance level, the items with the greatest level of judgment disagreement were identified and discussed as a committee.

| UIN | Max Points | 0 | 1 | 2 | 3 | 4 |
|-------------|------------|-----|-----|---|---|---|
| BI1718_07 | 1 | 55% | 45% | . | . | . |
| BI1718_09 | 1 | 55% | 45% | . | . | . |
| BI1718_03 | 1 | 55% | 45% | . | . | . |
| SCHS1622_04 | 1 | 55% | 45% | . | . | . |
| SCHS1622_09 | 1 | 55% | 45% | . | . | . |

Test-Level Cut Score Recommendations Agreement

The facilitator presented bar graphs to the panelists that displayed the distribution of cut score recommendations, by raw score, for each performance level: Apprentice (A), Proficient (P), and Distinguished (D). A graph with all performance levels on the scale was also presented.



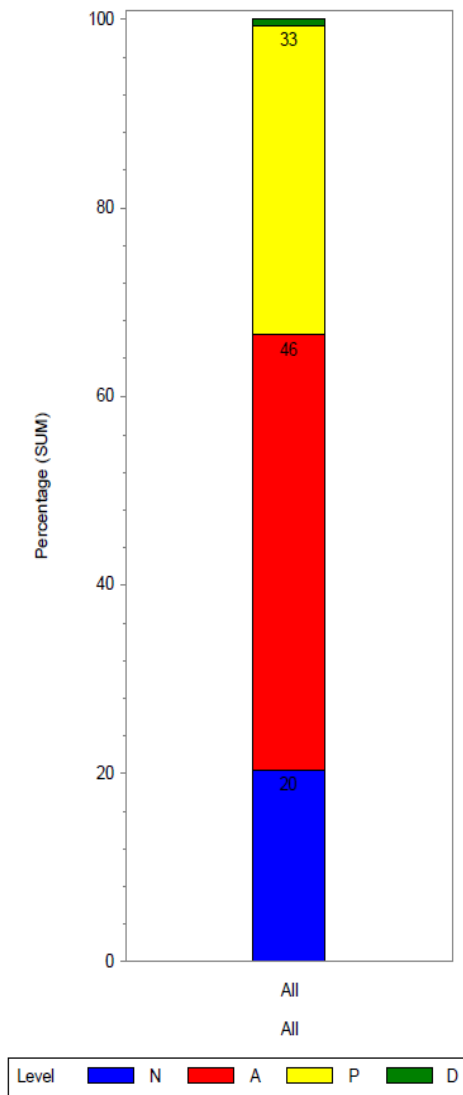
Item Score Mean and Score Distribution

The mean and distribution of scores received by students during the Spring 2019 administration was provided to panelists for each item.

| Sequence | Item | Maximum Points | Score Mean | Score Distribution | | | | |
|----------|-------------|----------------|------------|--------------------|------|-------|-------|-------|
| | | | | 0 pts | 1 pt | 2 pts | 3 pts | 4 pts |
| 1 | SCHS1626_01 | 1 | 0.32 | | | | | |
| 2 | SCHS1626_02 | 2 | 1.04 | 36% | 25% | 40% | | |
| 3 | SCHS1626_03 | 1 | 0.34 | | | | | |
| 4 | SCHS1626_04 | 1 | 0.21 | | | | | |
| 5 | SCHS1626_06 | 1 | 0.38 | | | | | |

Impact Data

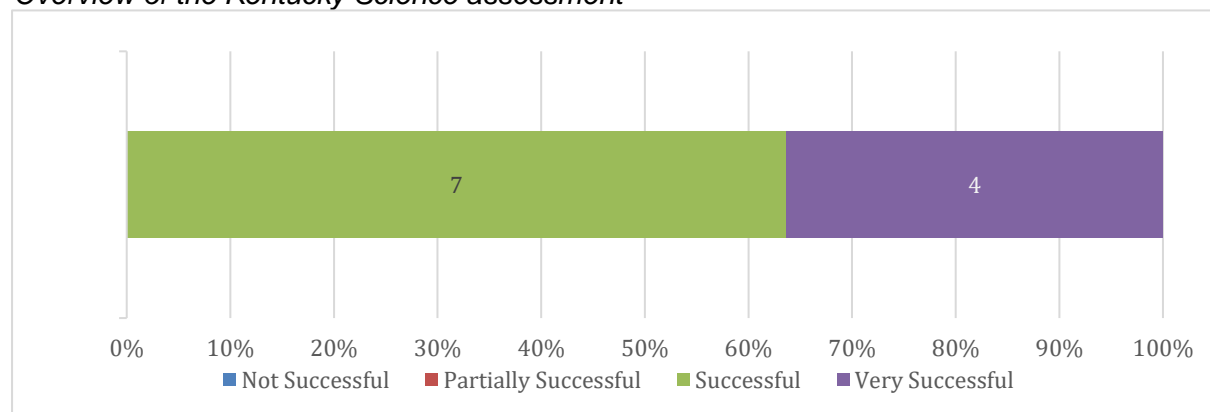
After Rounds 2 and 3, panelists were shown the percentage of students expected to be classified into each performance level—Novice (N), Apprentice (A), Proficient (P), and Distinguished (D)—based on the committee’s test-level cut score recommendations for that round. The impact data results were based on the sample of student data from the Spring 2019 administration of the Kentucky science grade 11 assessment.



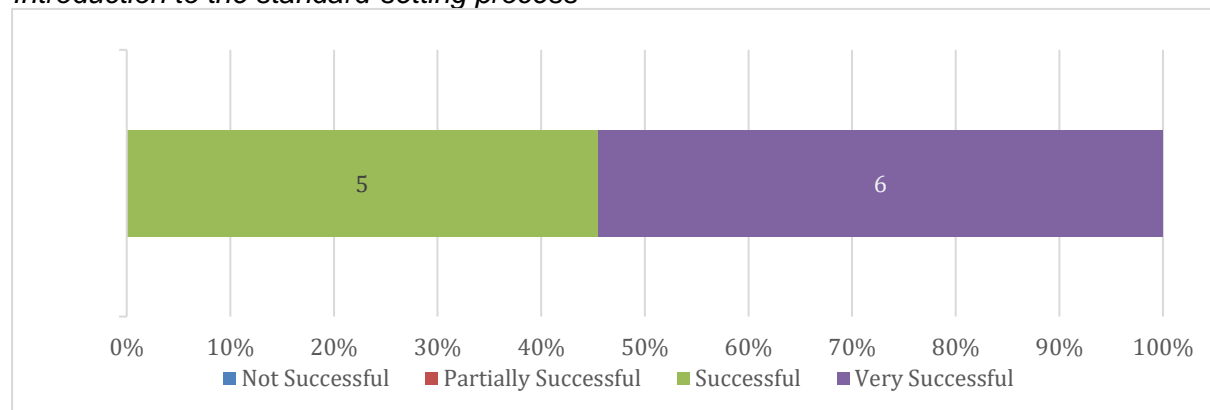
Appendix F – Panelist Evaluation Results

Question 1: Select the option that best reflects your opinion about the level of success of the various components of the meeting in which you participated. The activities were designed to help you both understand the process and be supportive of the recommendations made by the committee for Science Grade 11.

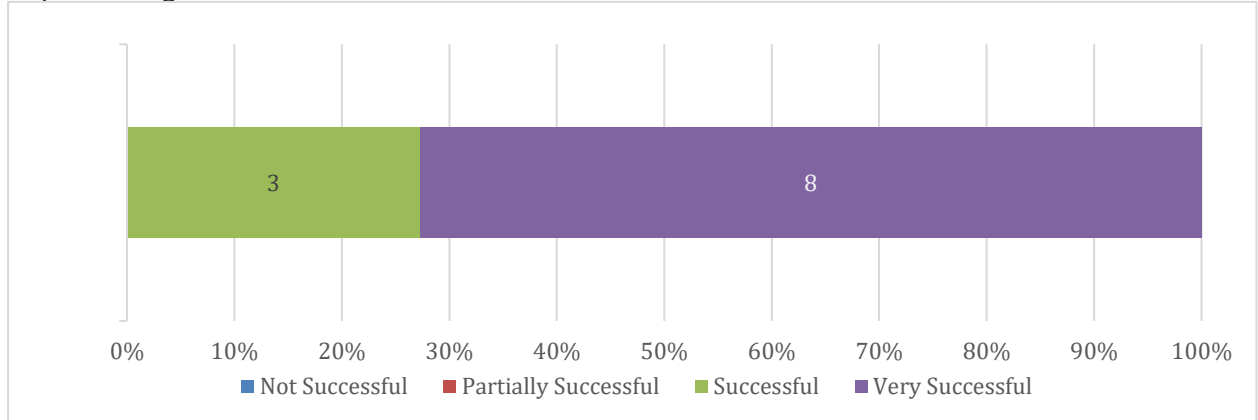
Overview of the Kentucky Science assessment



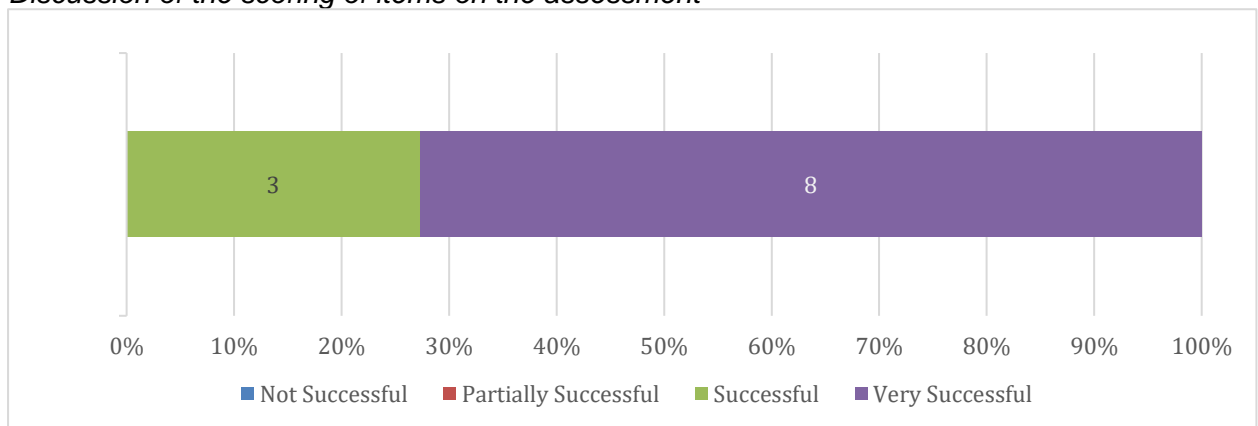
Introduction to the standard-setting process



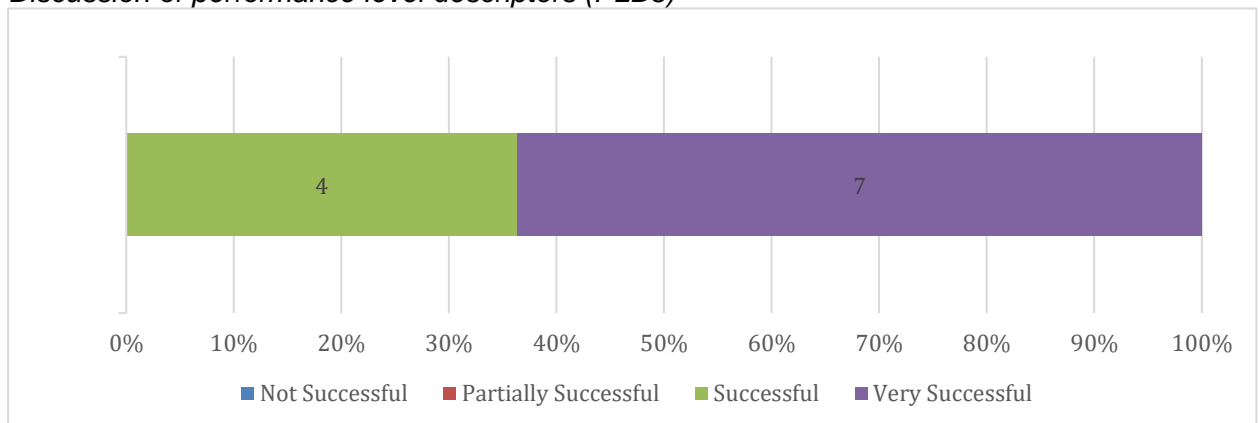
Experiencing the actual assessment



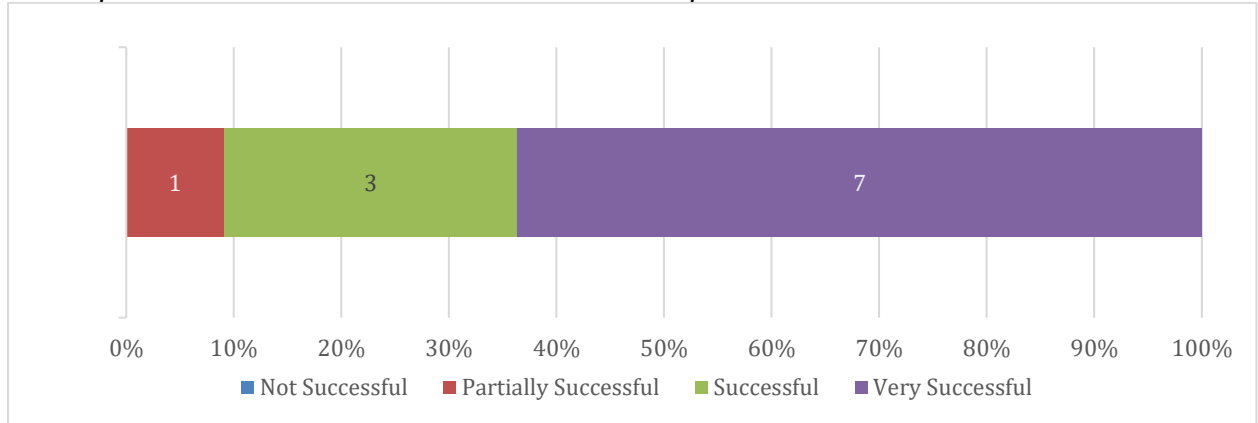
Discussion of the scoring of items on the assessment



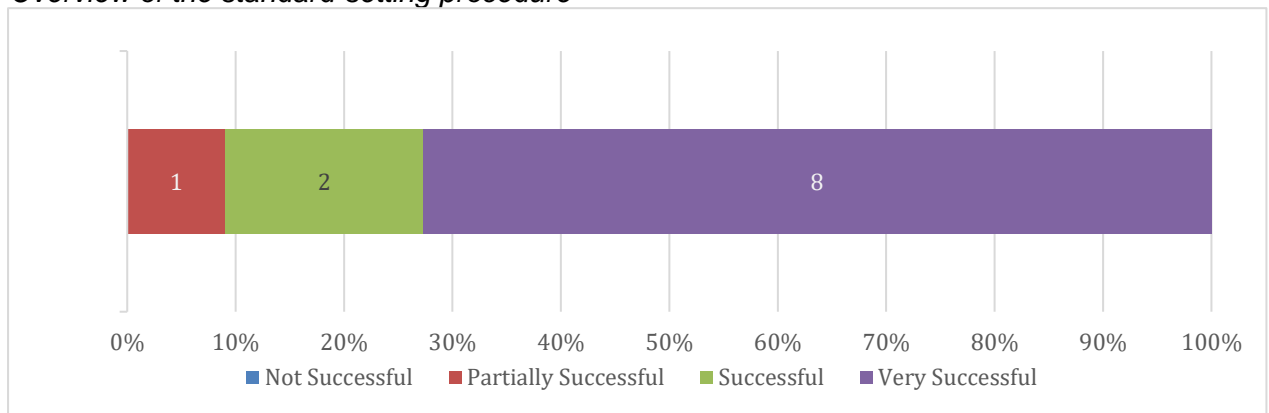
Discussion of performance level descriptors (PLDs)



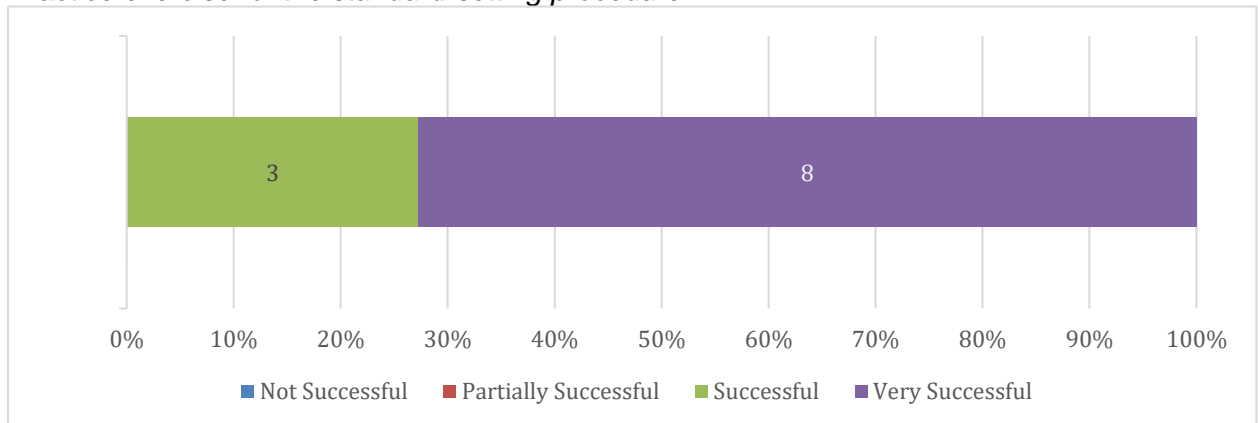
Development and discussion of the borderline descriptions



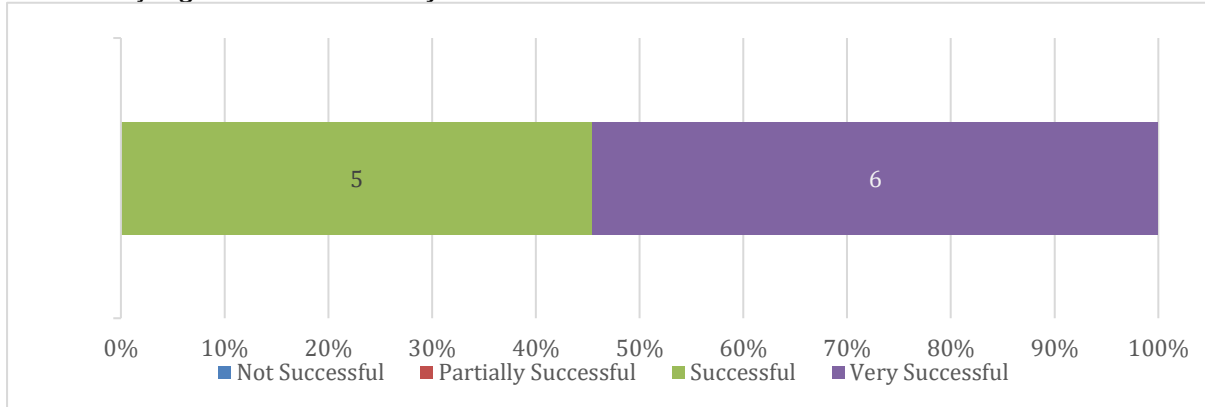
Overview of the standard-setting procedure



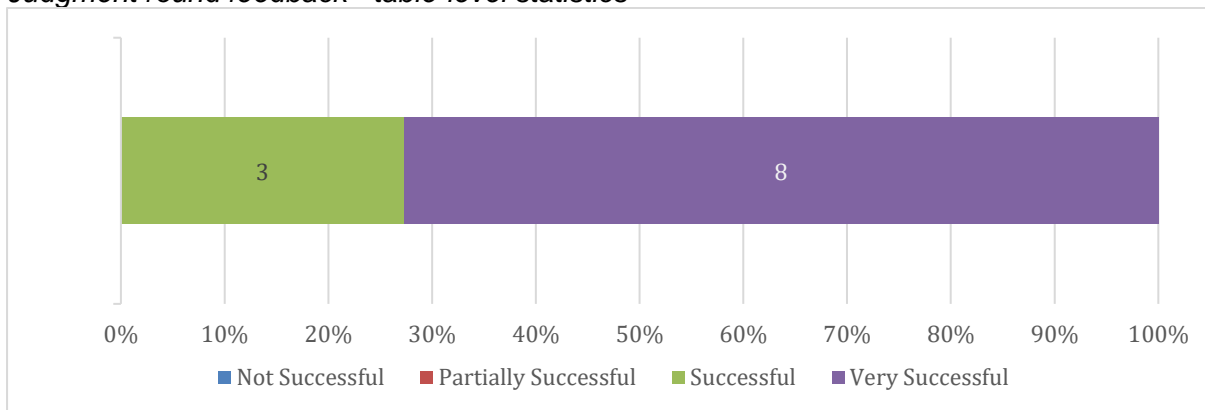
Practice exercise for the standard-setting procedure



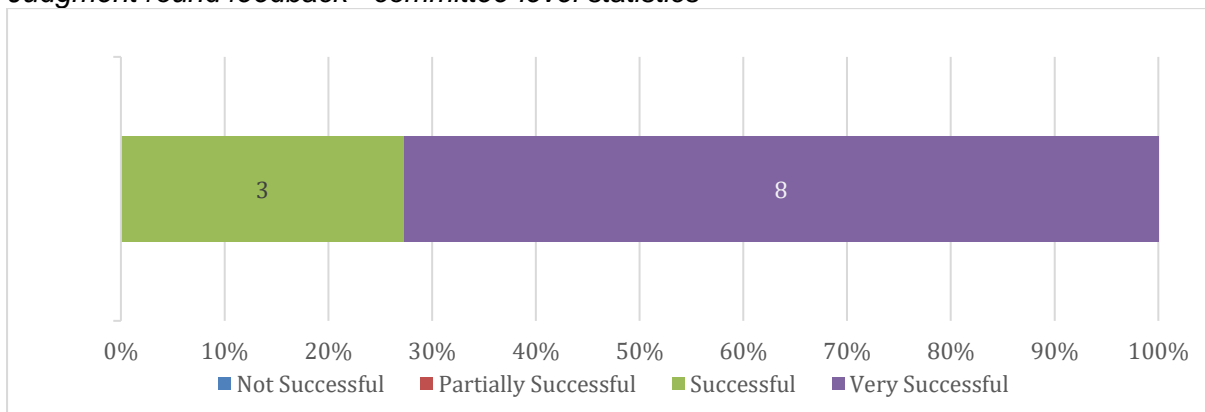
Individual judgment round activity



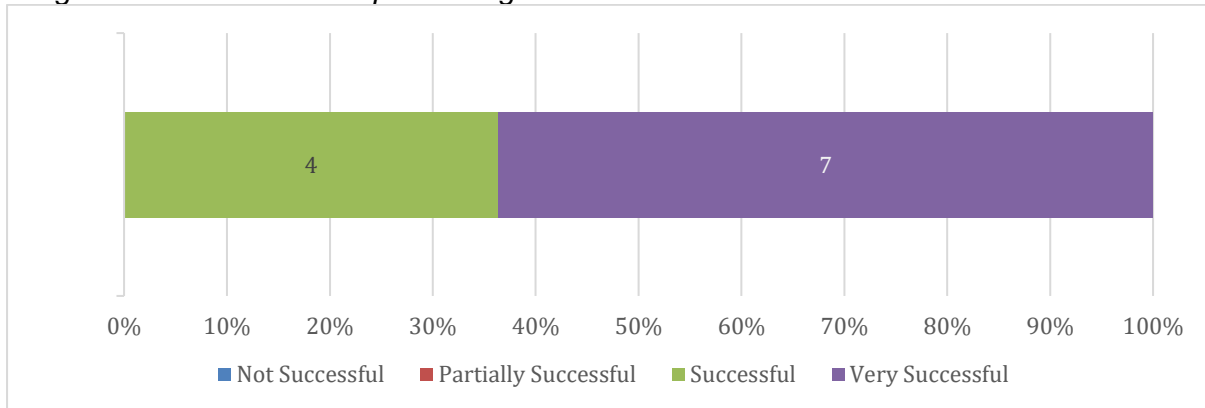
Judgment round feedback - table-level statistics



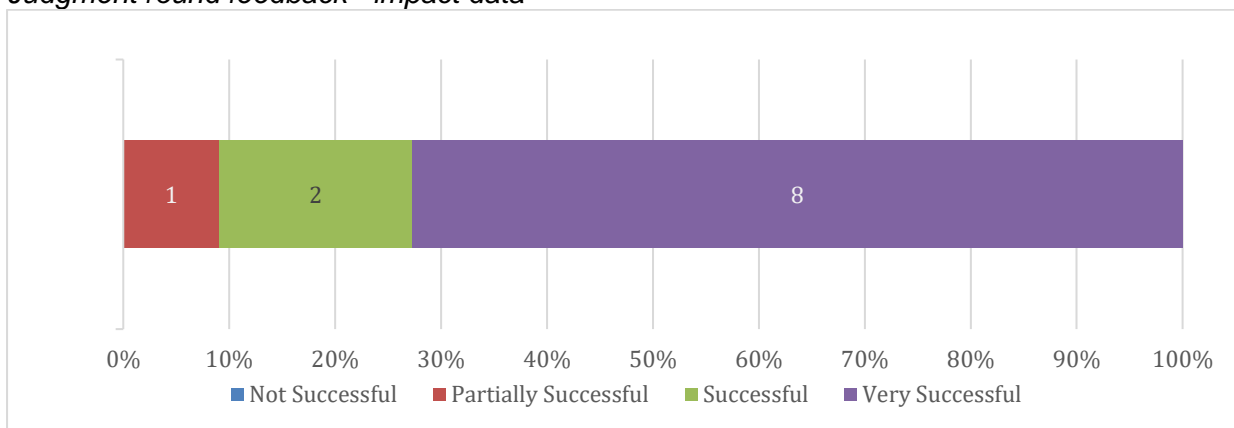
Judgment round feedback - committee-level statistics



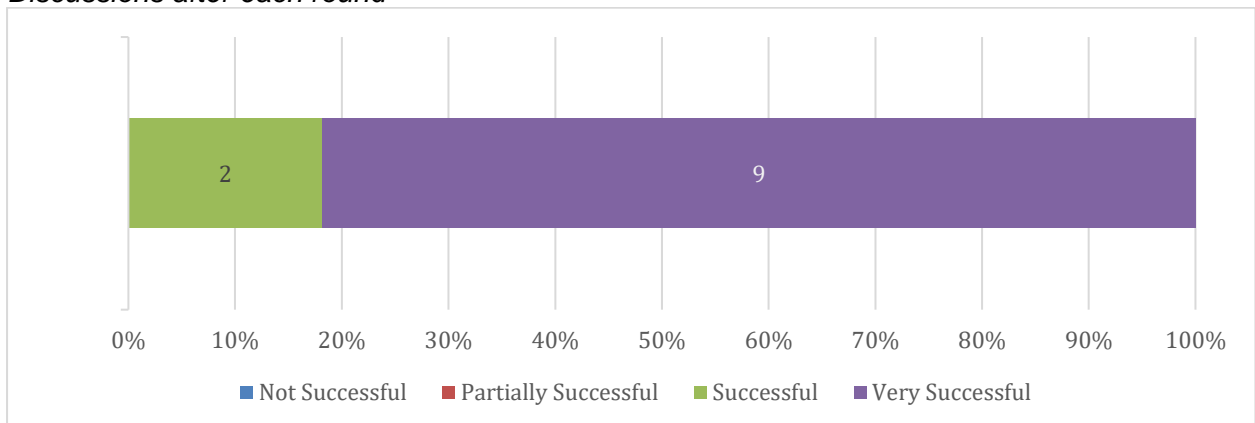
Judgment round feedback - panelist agreement data



Judgment round feedback - impact data

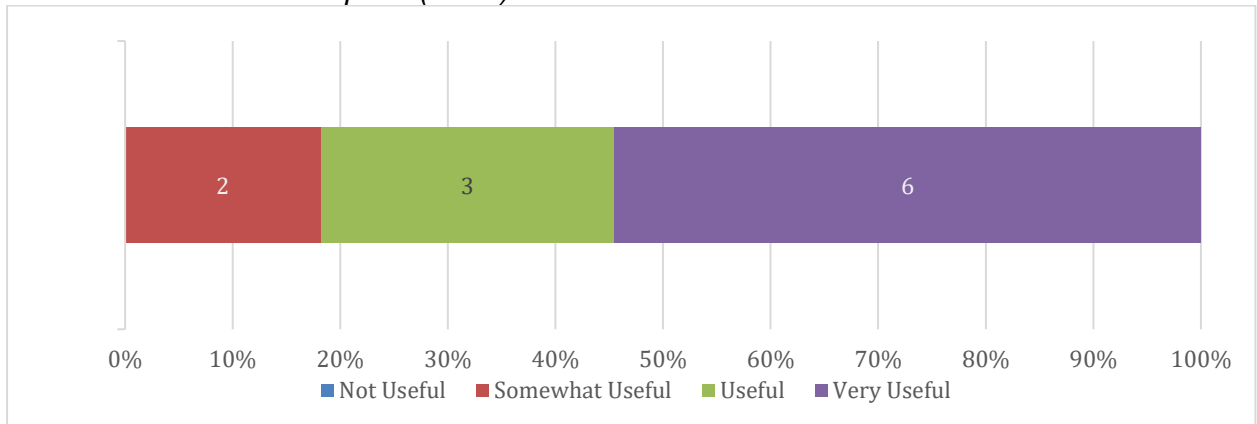


Discussions after each round



Question 2: How useful do you feel the following activities or information were in assisting you to make your recommendations for Science Grade 11?

Performance Level Descriptors (PLDs)



Borderline descriptions

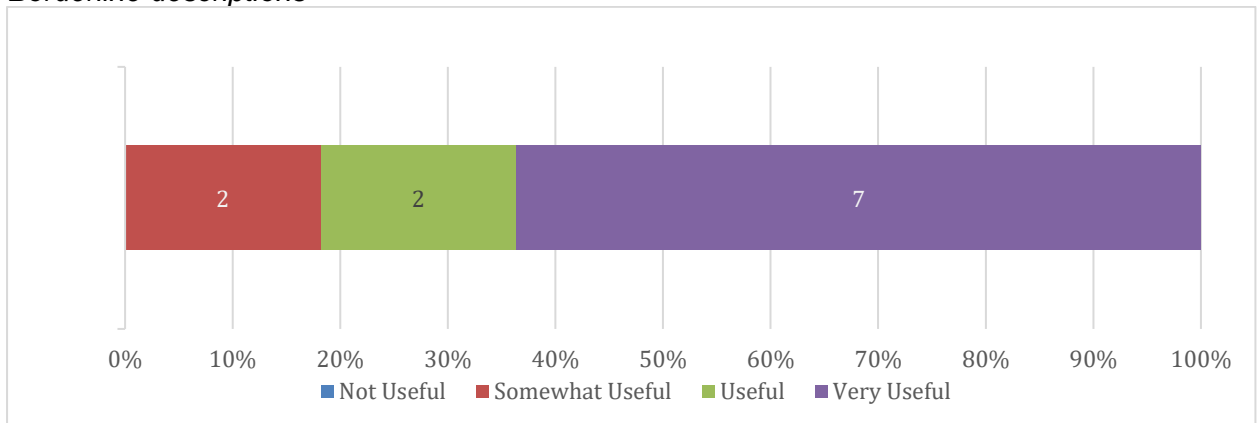
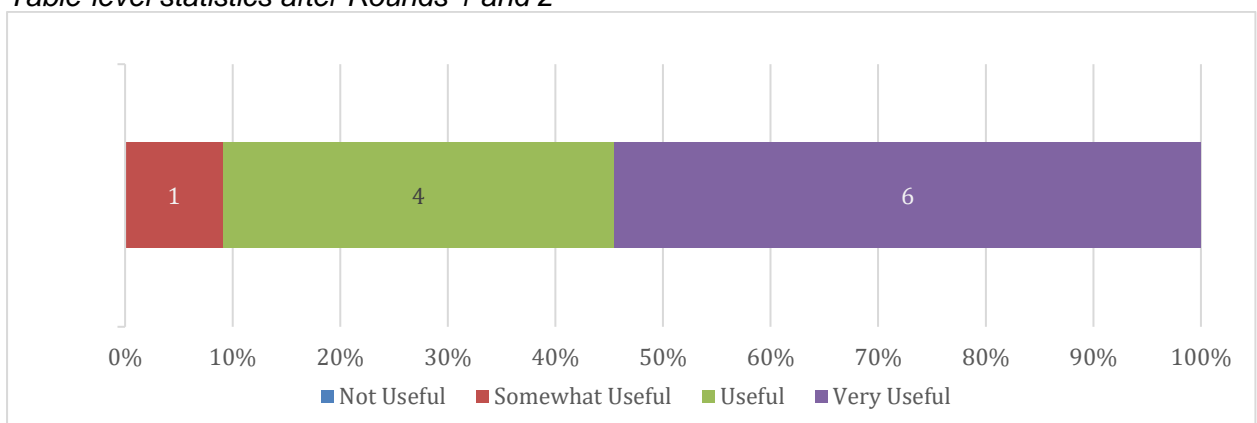
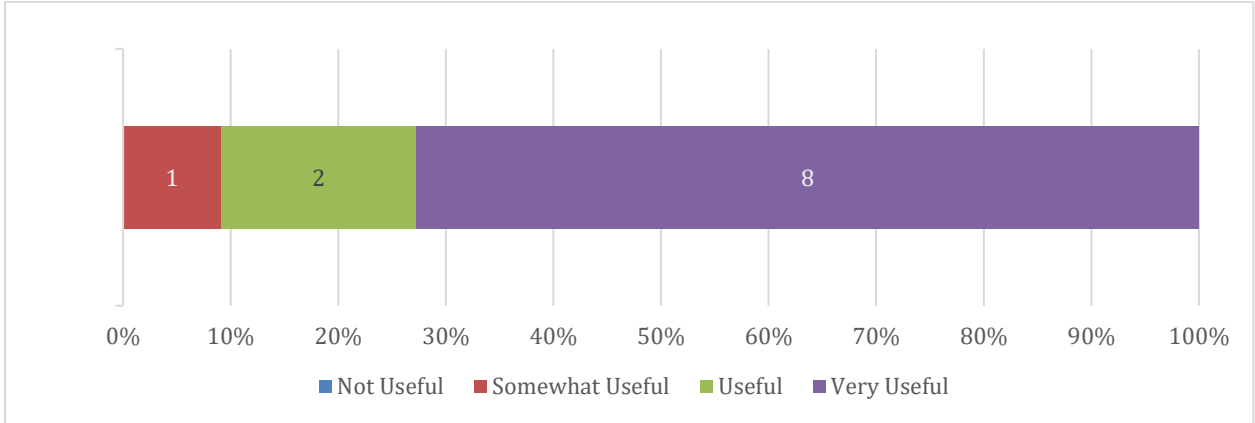


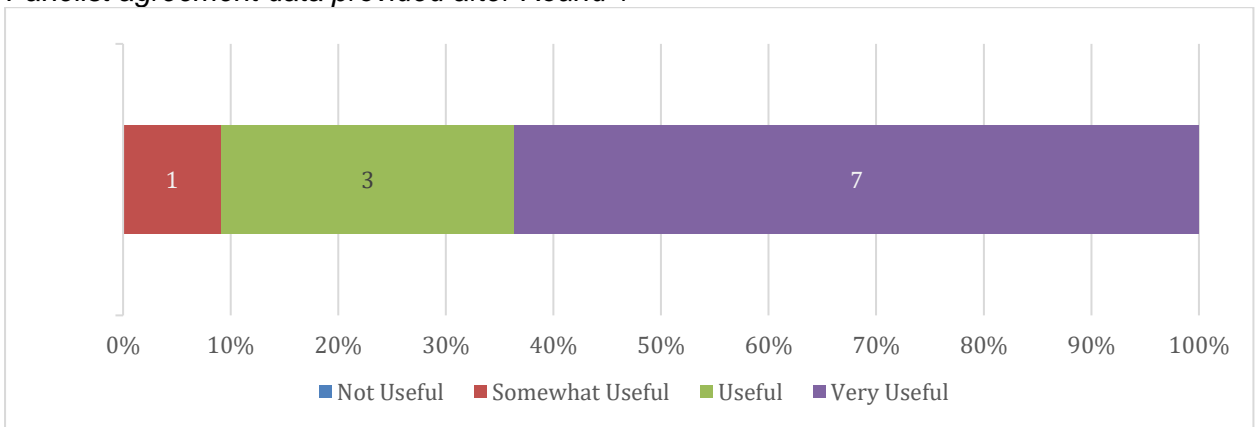
Table-level statistics after Rounds 1 and 2



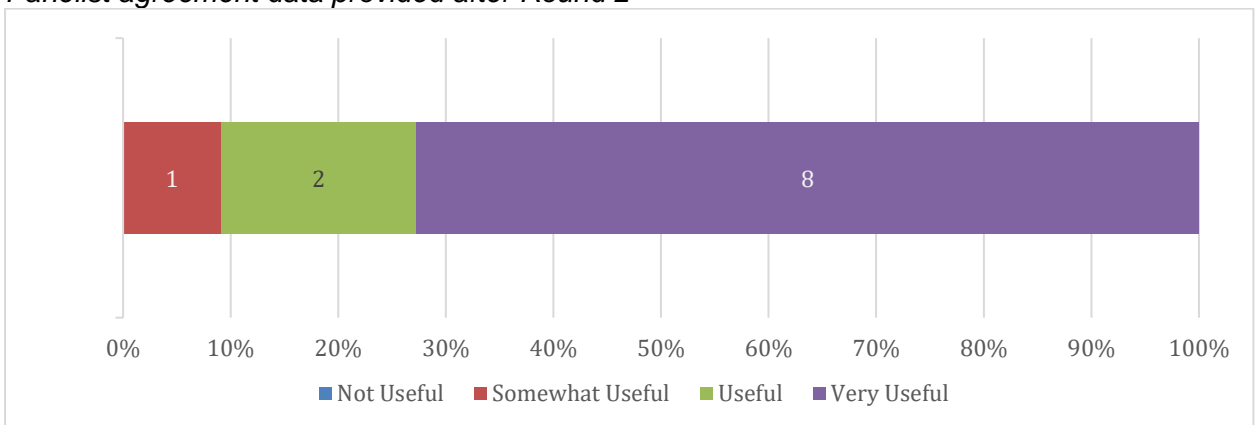
Committee-level statistics after Rounds 1 and 2



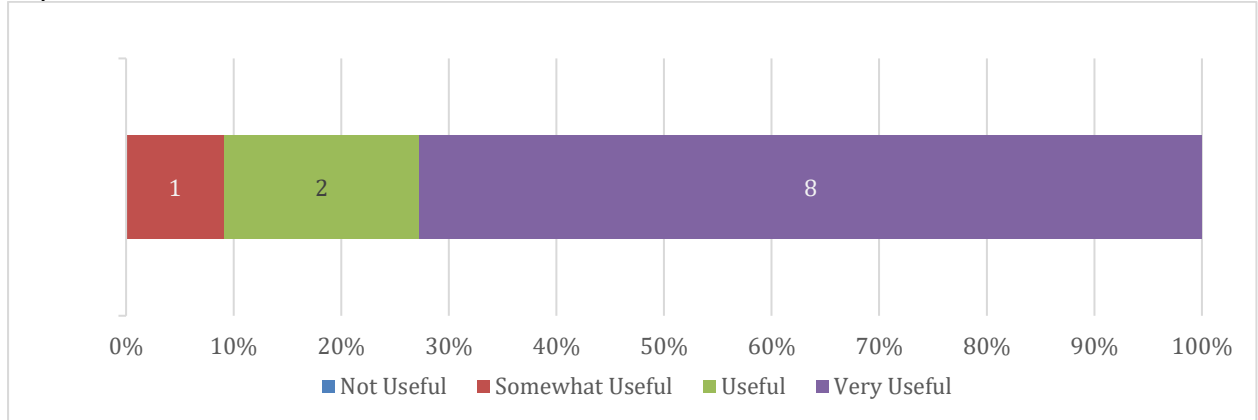
Panelist agreement data provided after Round 1



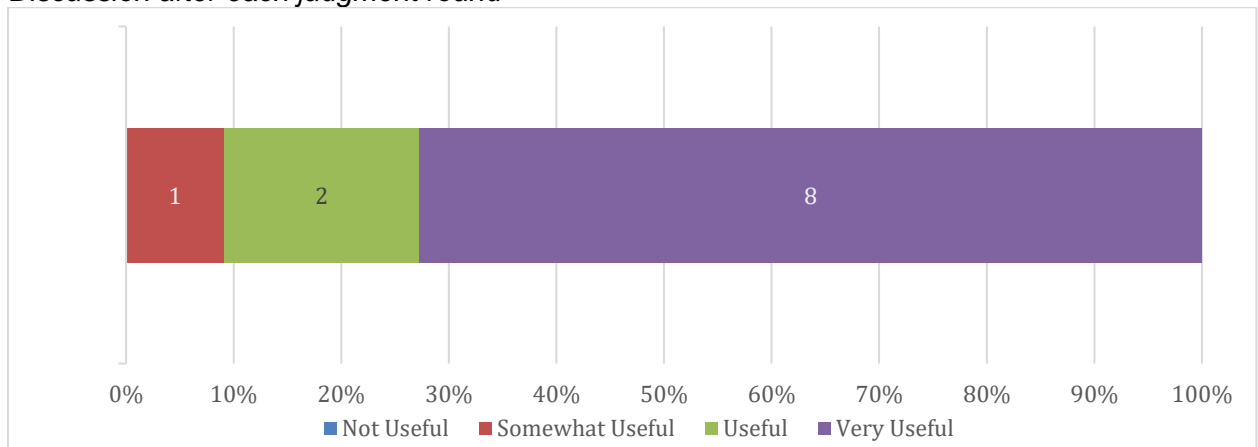
Panelist agreement data provided after Round 2



Impact data after Round 2

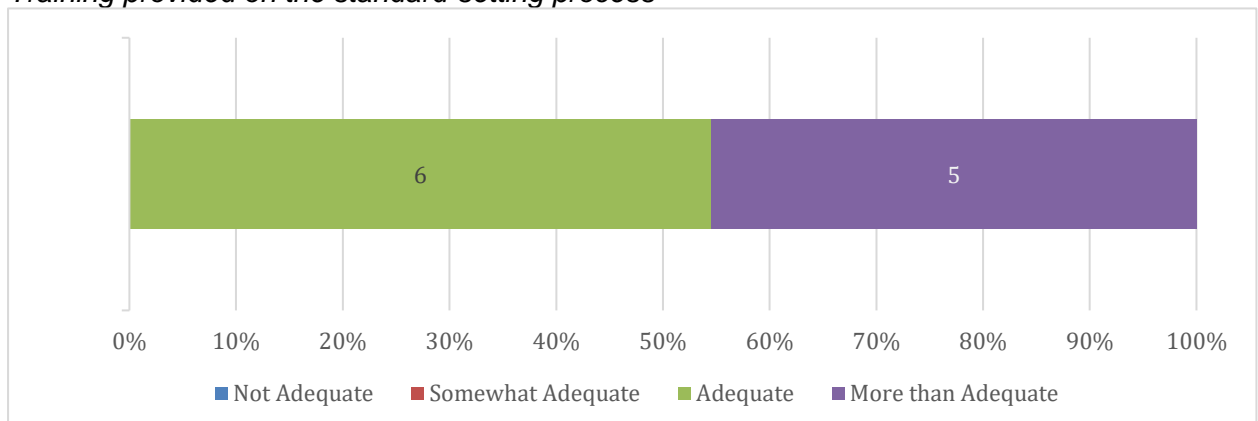


Discussion after each judgment round

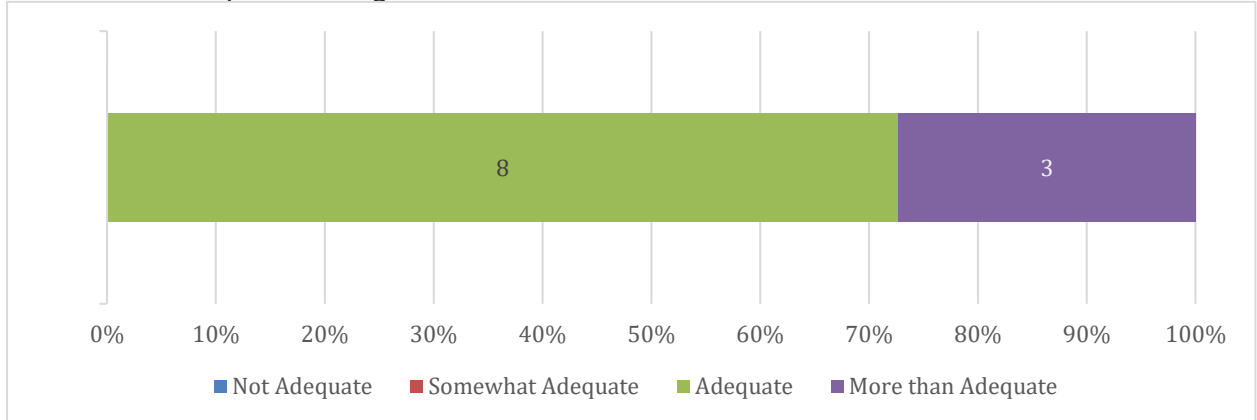


Question 3: How adequate were the following elements of the session for Science Grade 11?

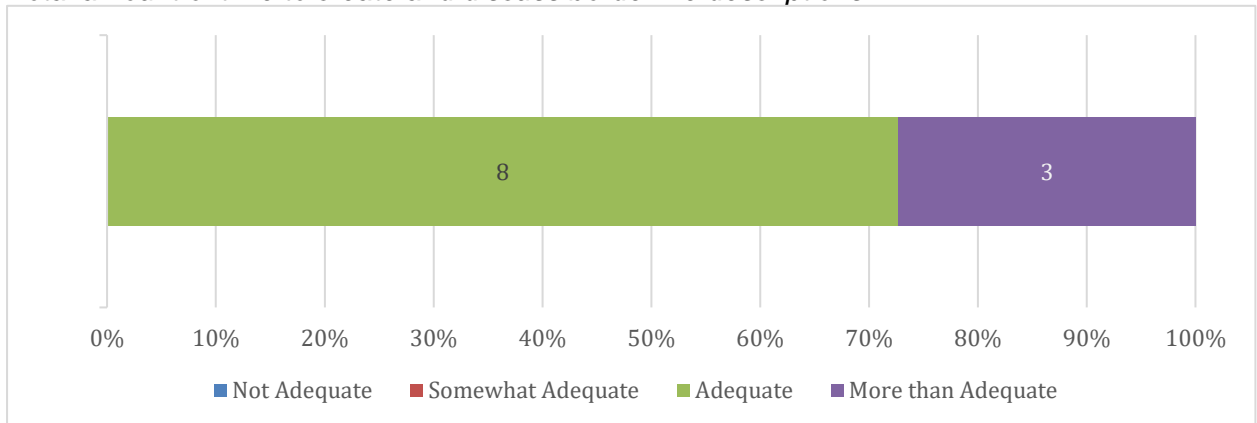
Training provided on the standard-setting process



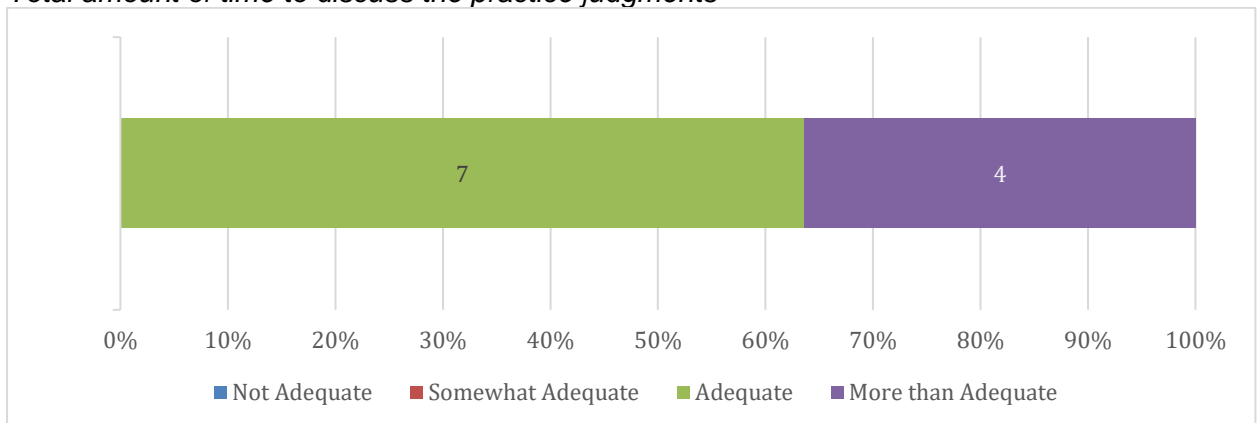
Amount of time spent training



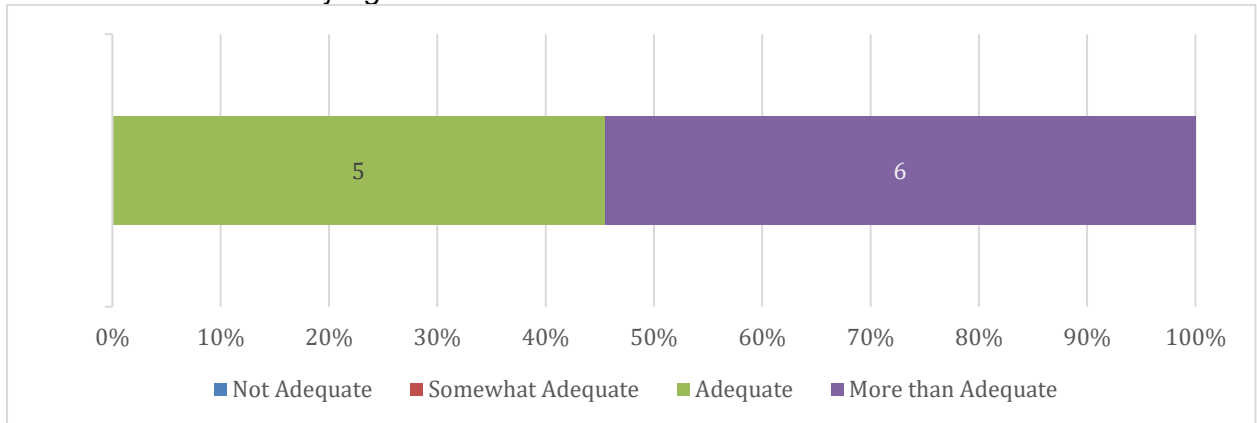
Total amount of time to create and discuss borderline descriptions



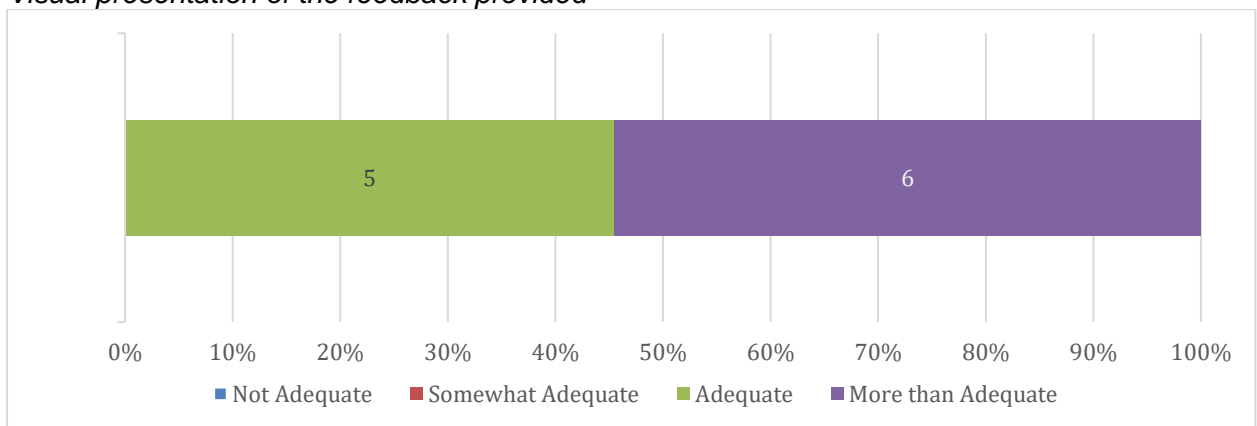
Total amount of time to discuss the practice judgments



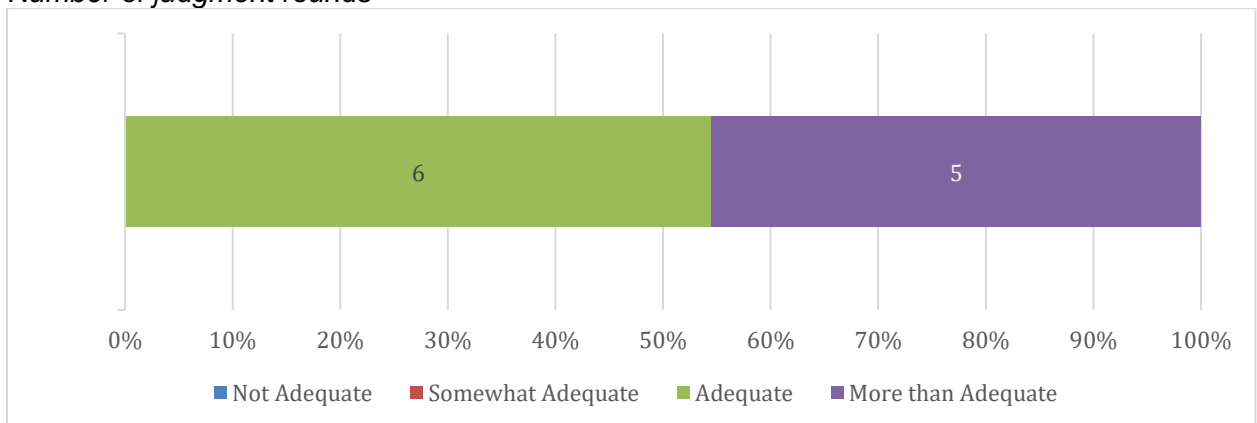
Amount of time to make judgments



Visual presentation of the feedback provided

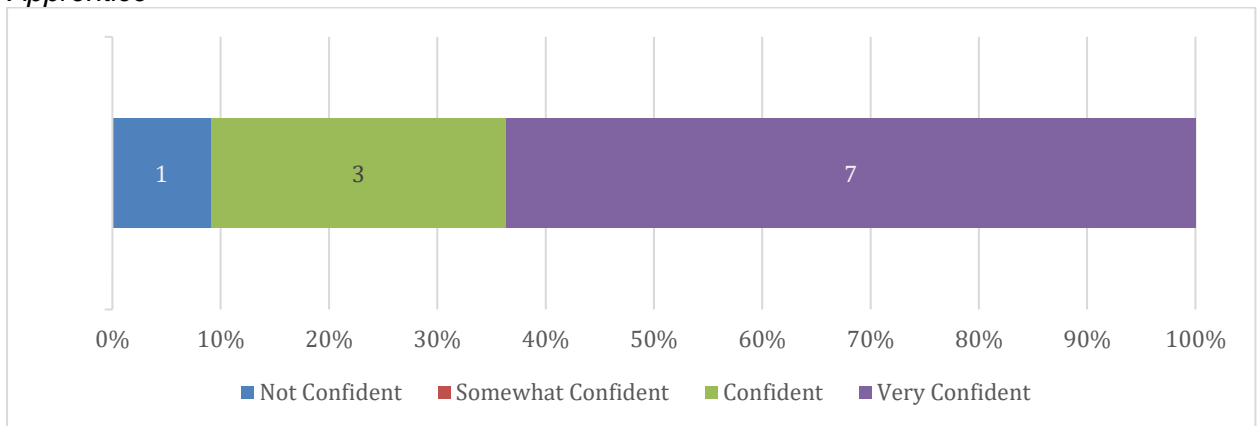


Number of judgment rounds

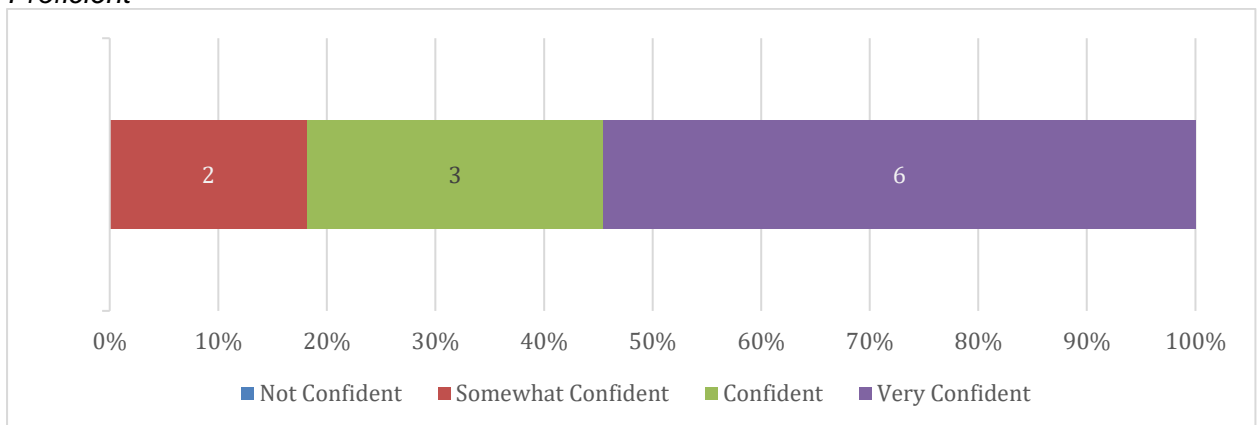


Question 4: How confident do you feel that the final cut score recommendations for Science Grade 11 represent appropriate levels of student performance?

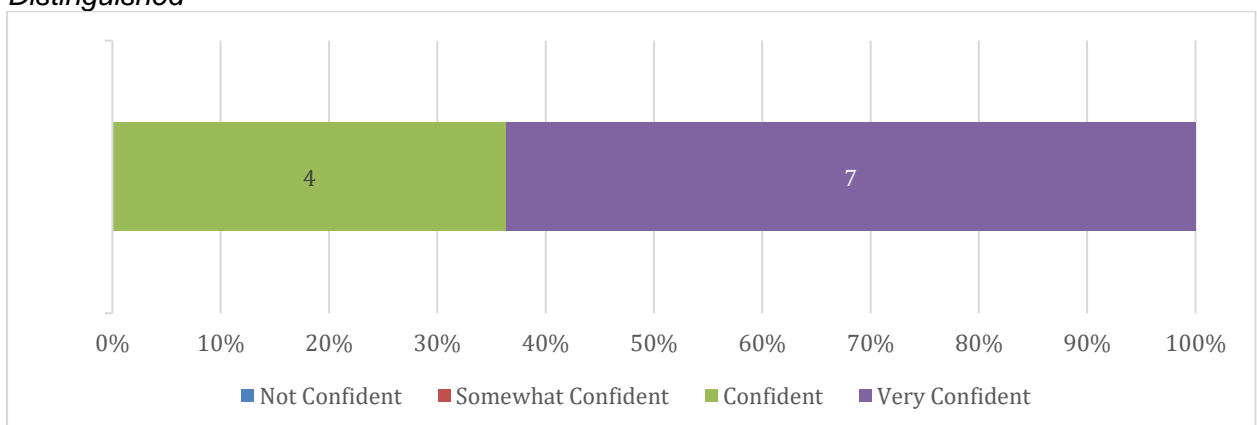
Apprentice



Proficient

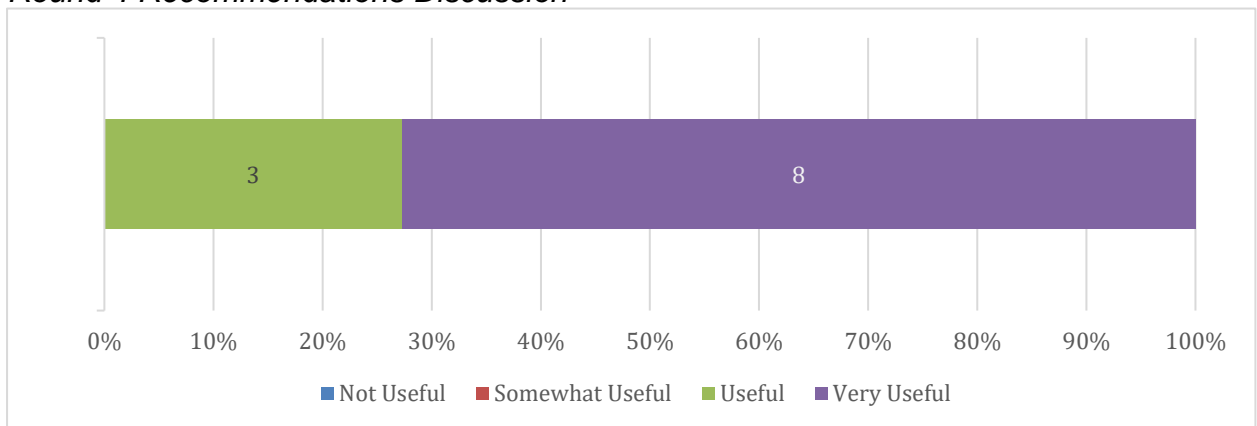


Distinguished

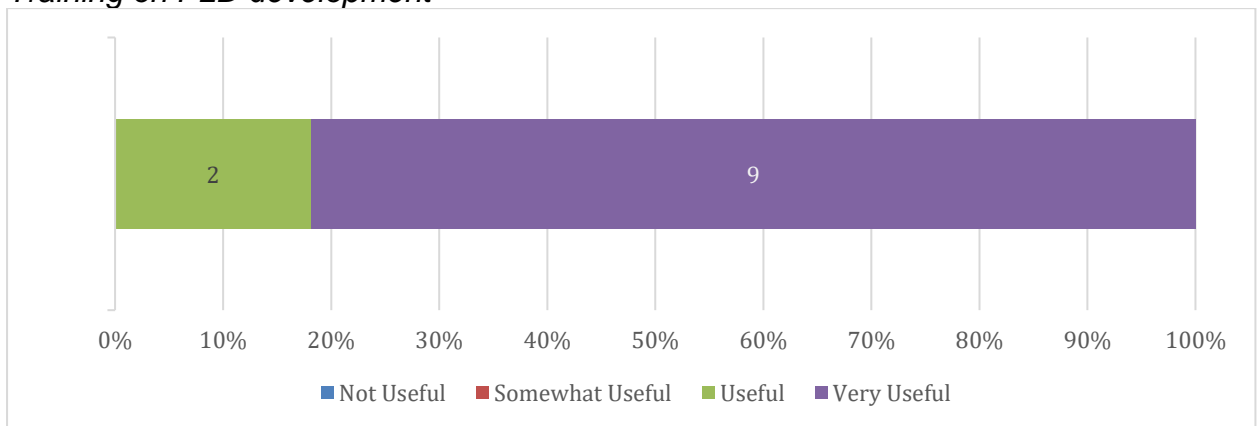


Question 5: How useful do you feel the following activities or information were in assisting you to make your recommendations?

Round 4 Recommendations Discussion

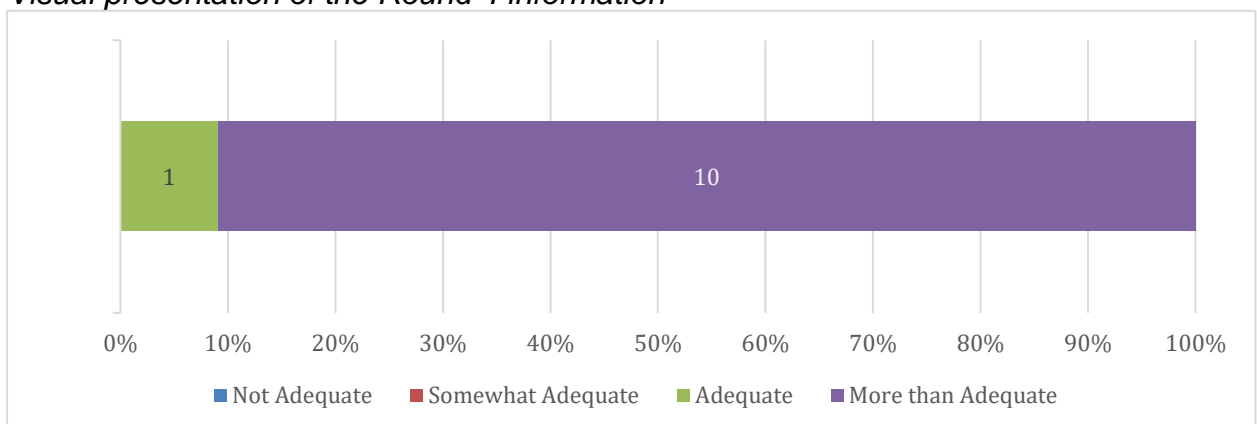


Training on PLD development

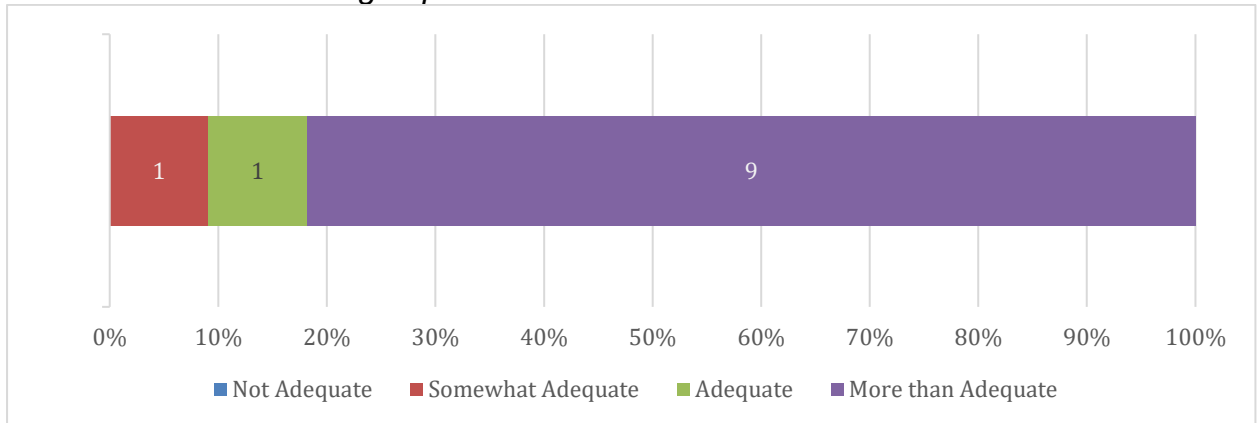


Question 6: How adequate were the following elements of the session?

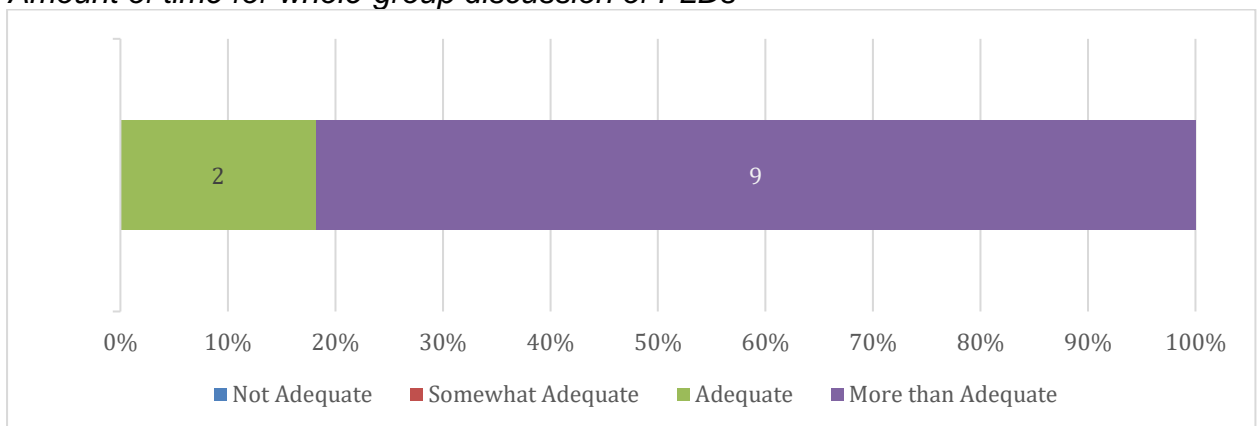
Visual presentation of the Round 4 information



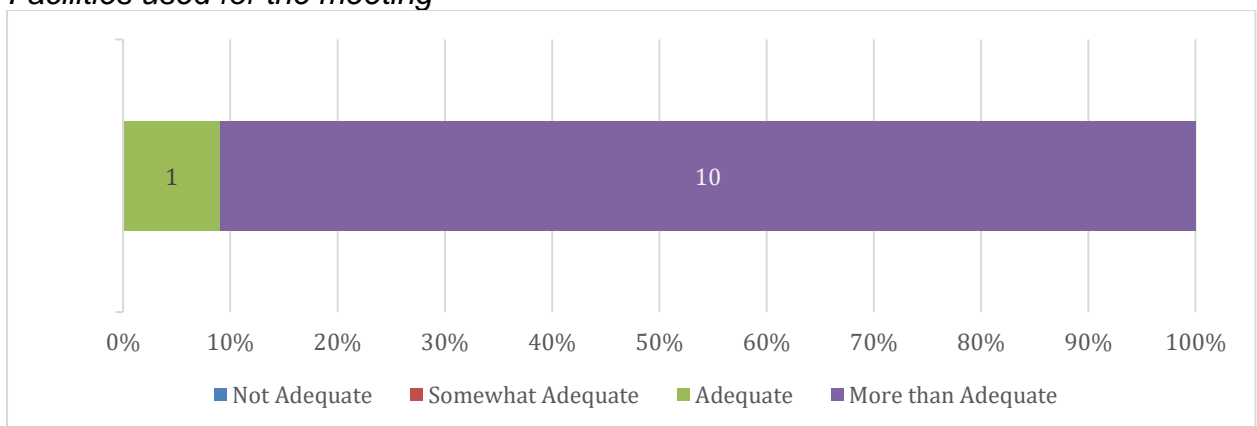
Amount of time for table-group discussion of PLDs



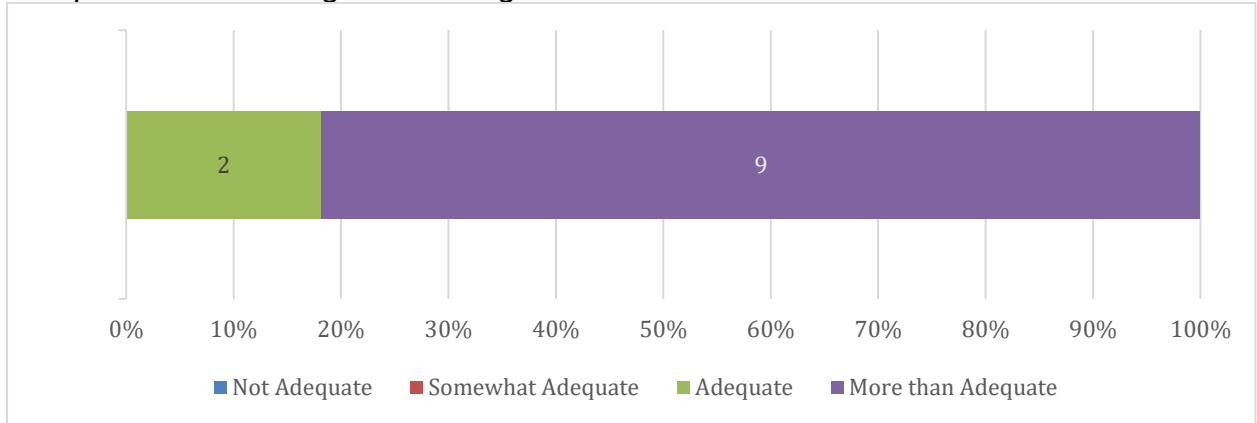
Amount of time for whole-group discussion of PLDs



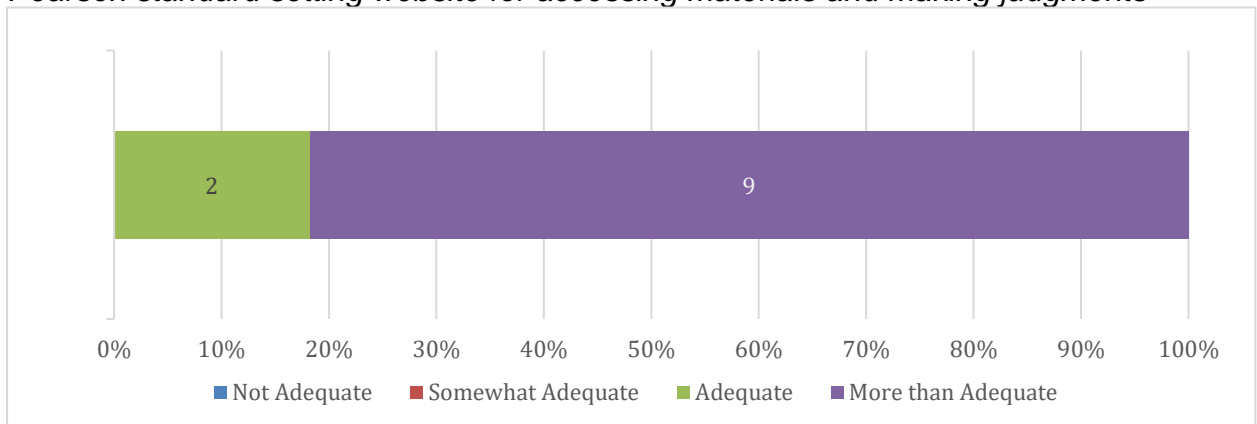
Facilities used for the meeting



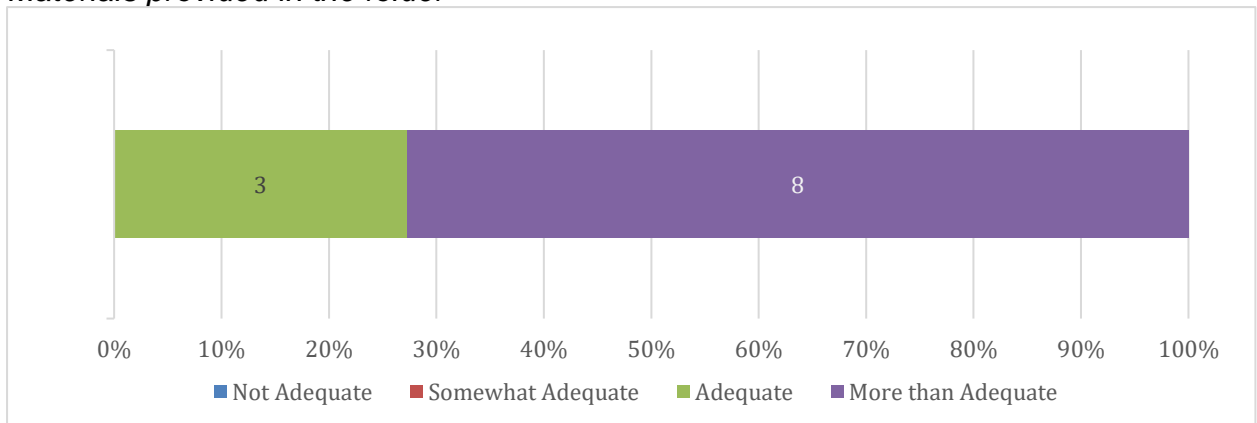
Computers used during the meeting



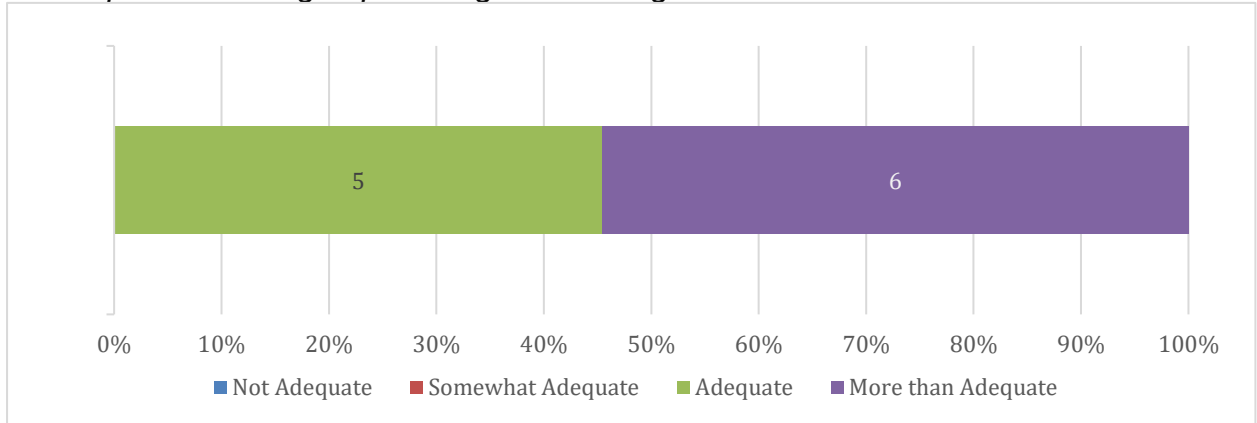
Pearson standard setting website for accessing materials and making judgments



Materials provided in the folder

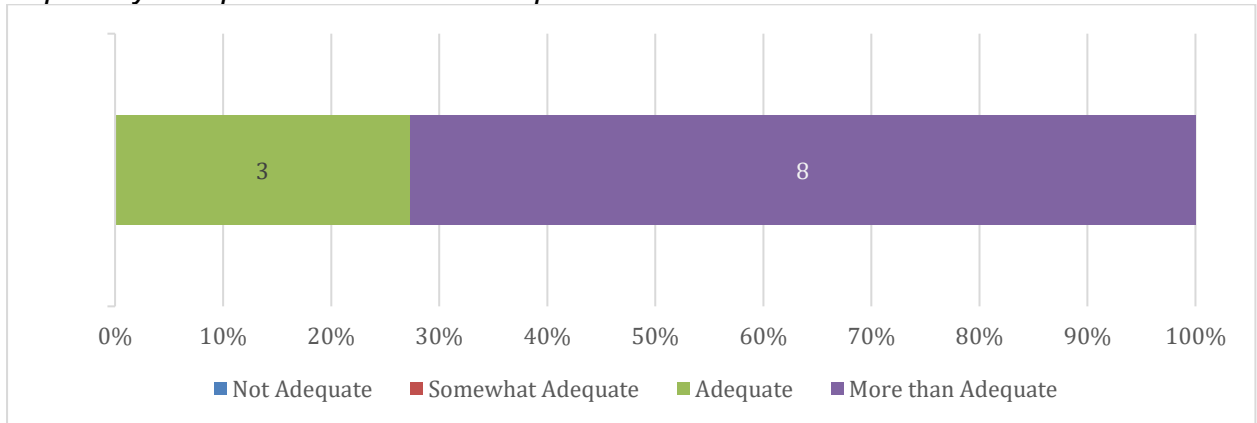


Work space in table groups during the meeting

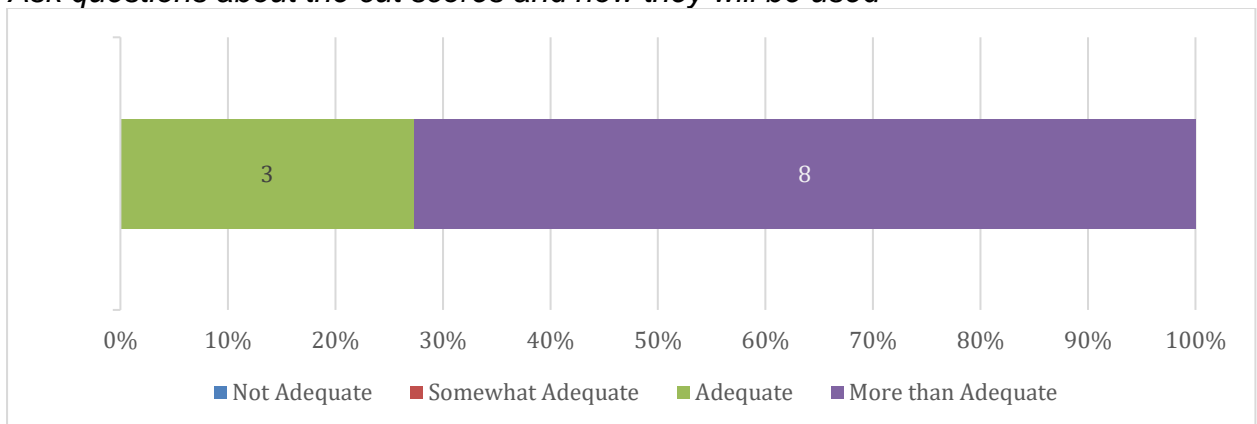


Question 7: Did you have adequate opportunities during the session to:

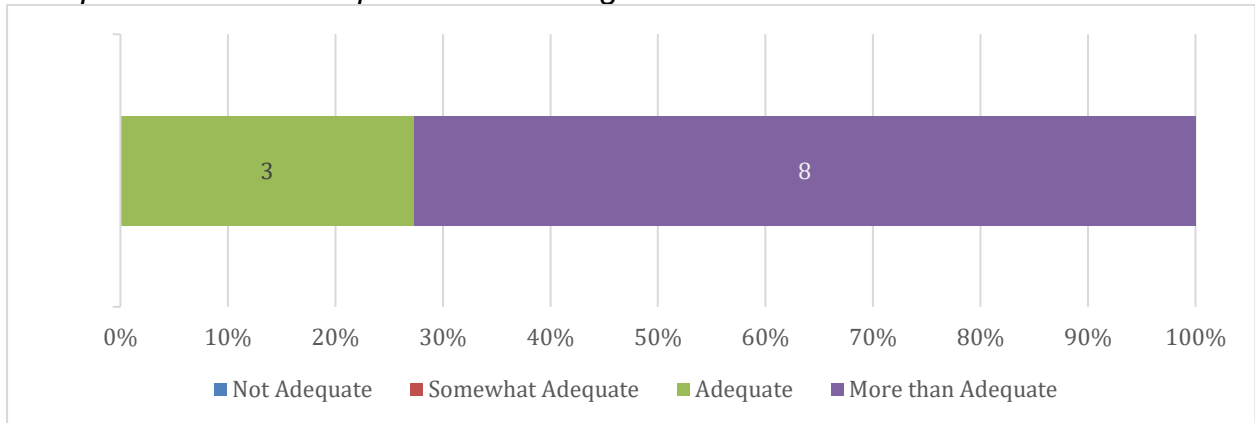
Express your opinions about student performance levels



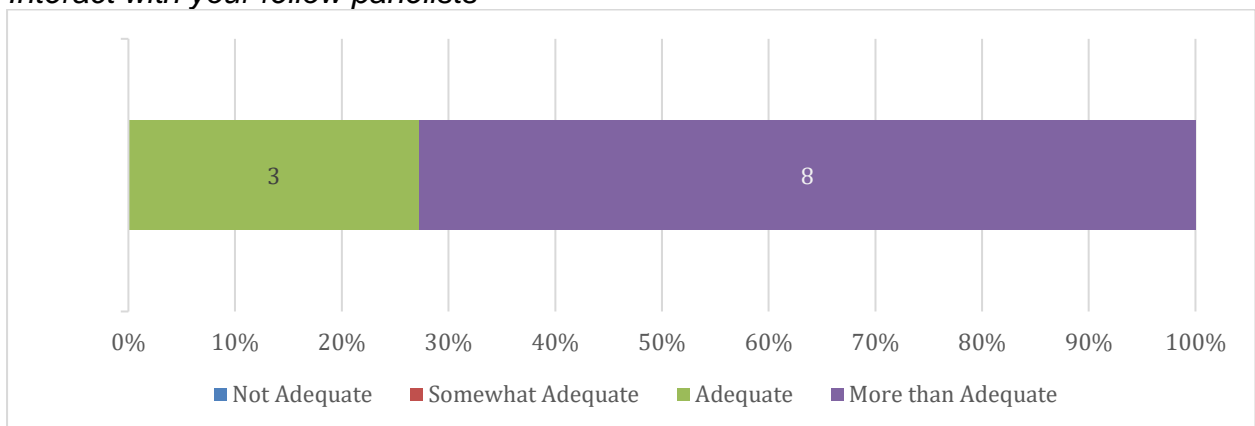
Ask questions about the cut scores and how they will be used



Ask questions about the process of making cut score recommendations

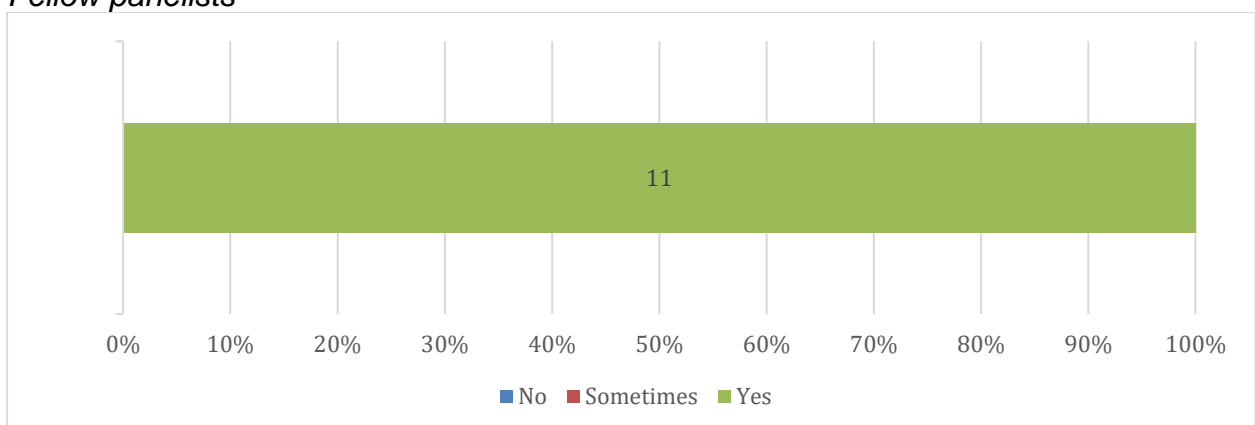


Interact with your fellow panelists

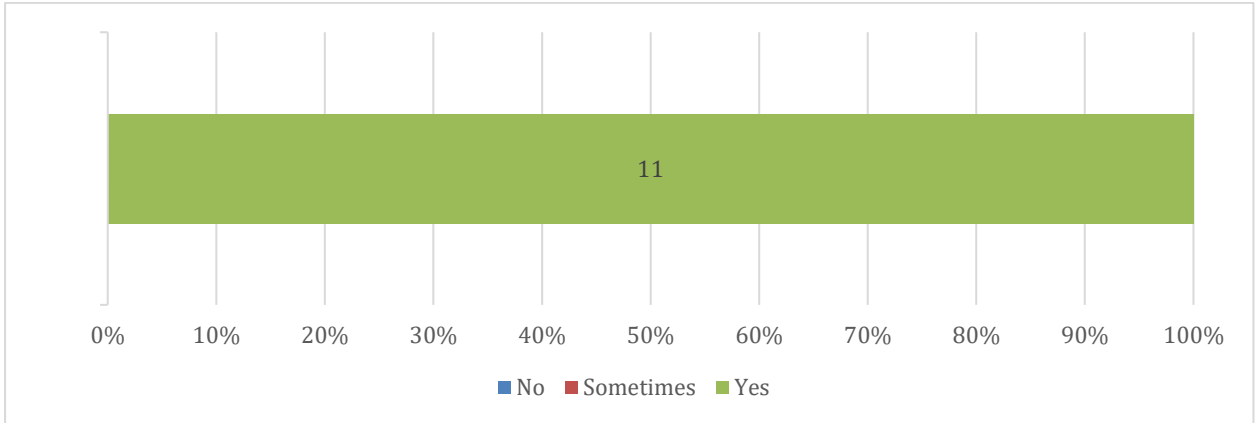


Question 8: Do you believe your opinions and judgments were treated with respect by:

Fellow panelists

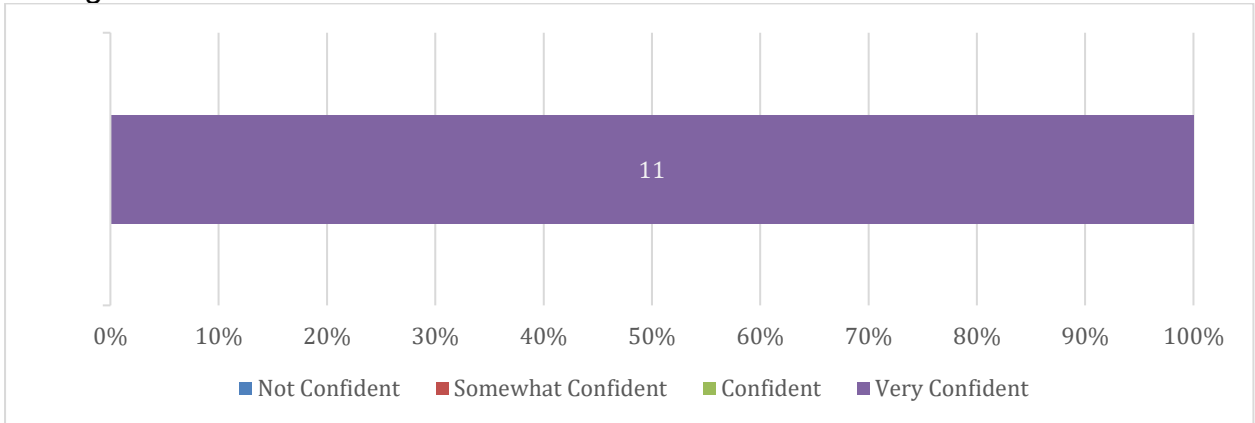


Facilitators

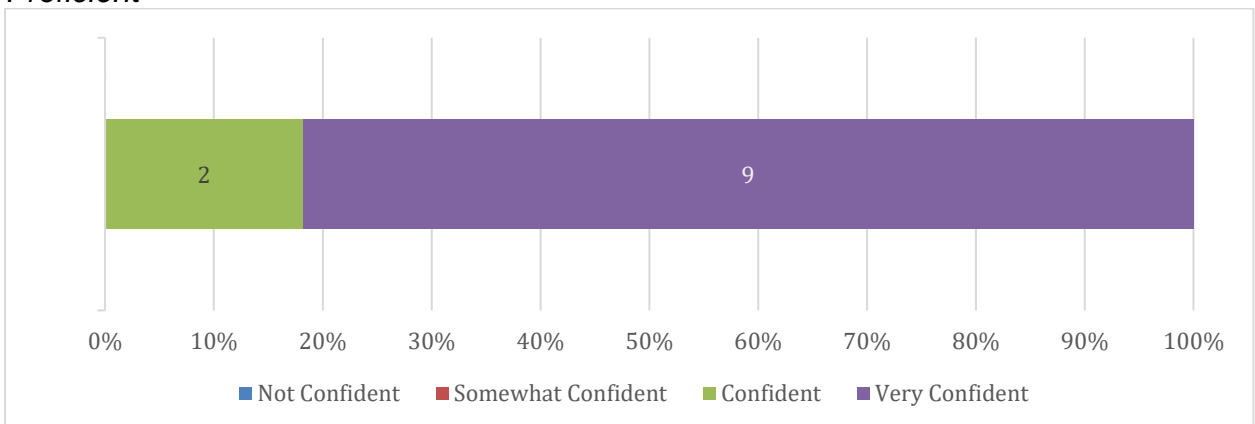


Question 9: How confident do you feel that the performance level descriptors (PLDs) you developed for Grade 11 Science are reasonable for each student performance level?

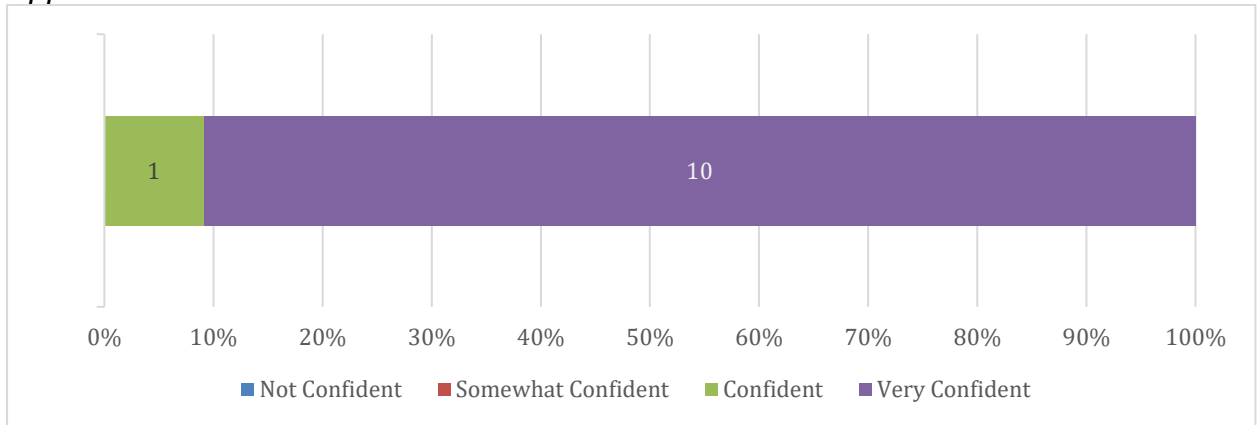
Distinguished



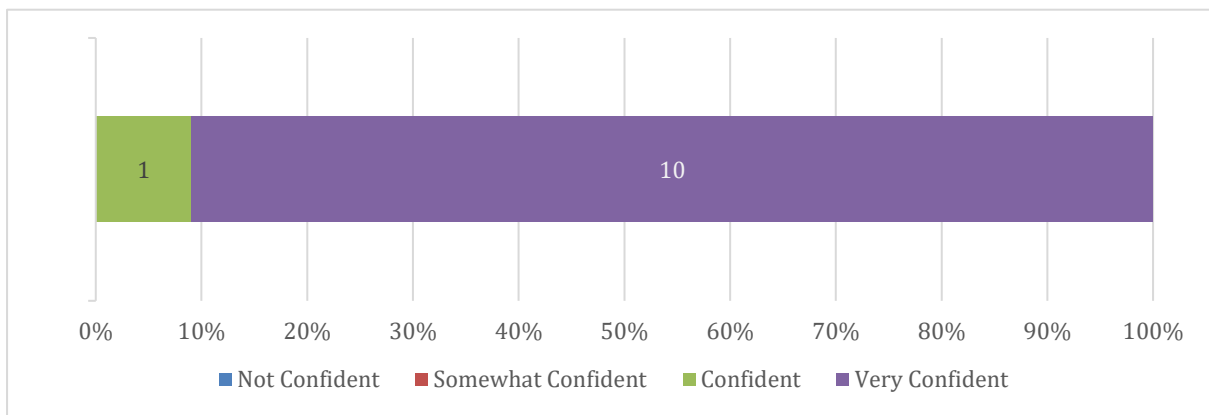
Proficient



Apprentice



Novice



Appendix G – Recommended Cut Scores by Judgment Round

Table G.1: Science Grade 11

| Performance Level | Maximum Score | Rounds | | |
|-------------------|---------------|--------|----|----|
| | | 1 | 2 | 3 |
| Apprentice | 48 | 11 | 8 | 12 |
| Proficient | | 34 | 27 | 22 |
| Distinguished | | 45 | 41 | 38 |

Appendix H – Recommended Cut Score Summary Statistics by Judgment Round

Table H.1: Science Grade 11

| Round | Statistic | Performance Level | | |
|-------|-----------|-------------------|------------|---------------|
| | | Apprentice | Proficient | Distinguished |
| 1 | Mean | 12.36 | 34.64 | 44.73 |
| | Minimum | 4 | 25 | 43 |
| | Q1 | 8 | 31 | 44 |
| | Median | 11 | 34 | 45 |
| | Q3 | 18 | 40 | 46 |
| | Maximum | 24 | 42 | 46 |
| 2 | Mean | 8.09 | 26.00 | 39.18 |
| | Minimum | 4 | 17 | 31 |
| | Q1 | 6 | 25 | 38 |
| | Median | 8 | 27 | 41 |
| | Q3 | 9 | 29 | 42 |
| | Maximum | 16 | 33 | 43 |
| 3 | Mean | 12.00 | 22.55 | 36.91 |
| | Minimum | 7 | 21 | 33 |
| | Q1 | 10 | 21 | 34 |
| | Median | 12 | 22 | 38 |
| | Q3 | 13 | 24 | 39 |
| | Maximum | 18 | 26 | 40 |

Appendix I – Test-Level Participant Judgment Agreement

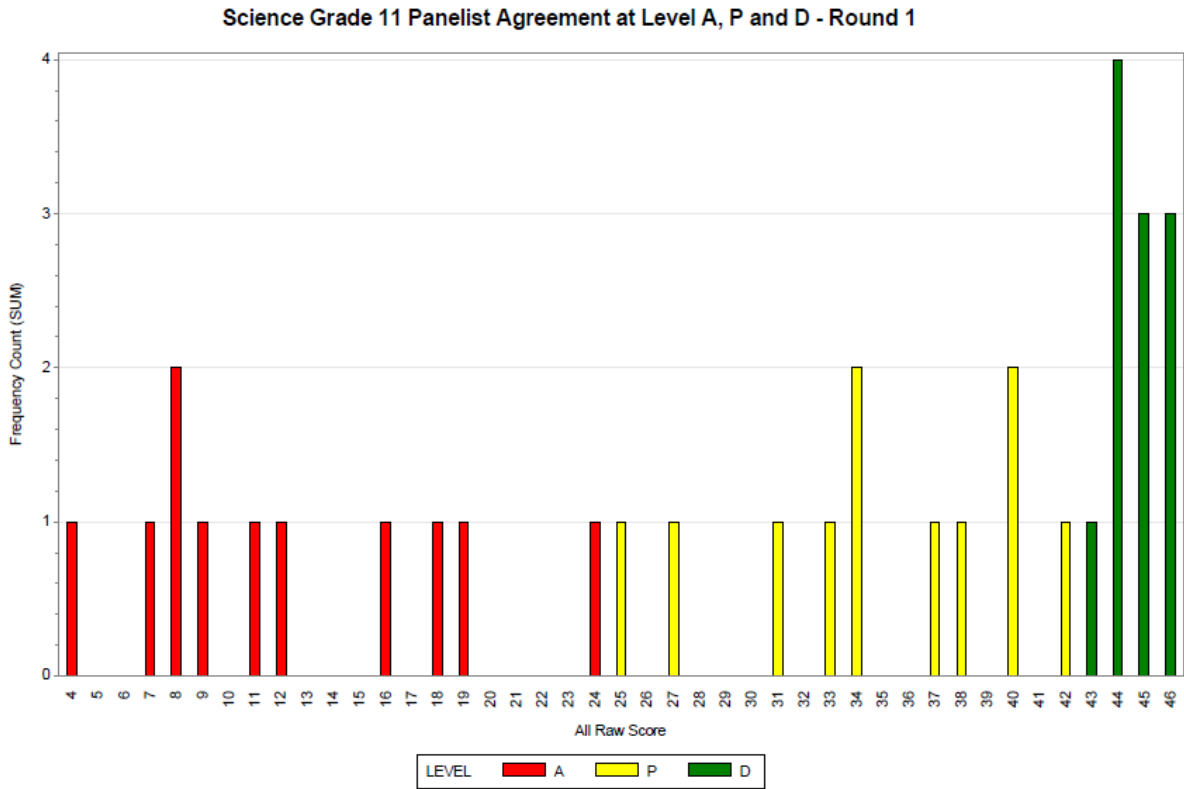


Figure I.1: Grade 11 Science Round 1 Panelist Agreement

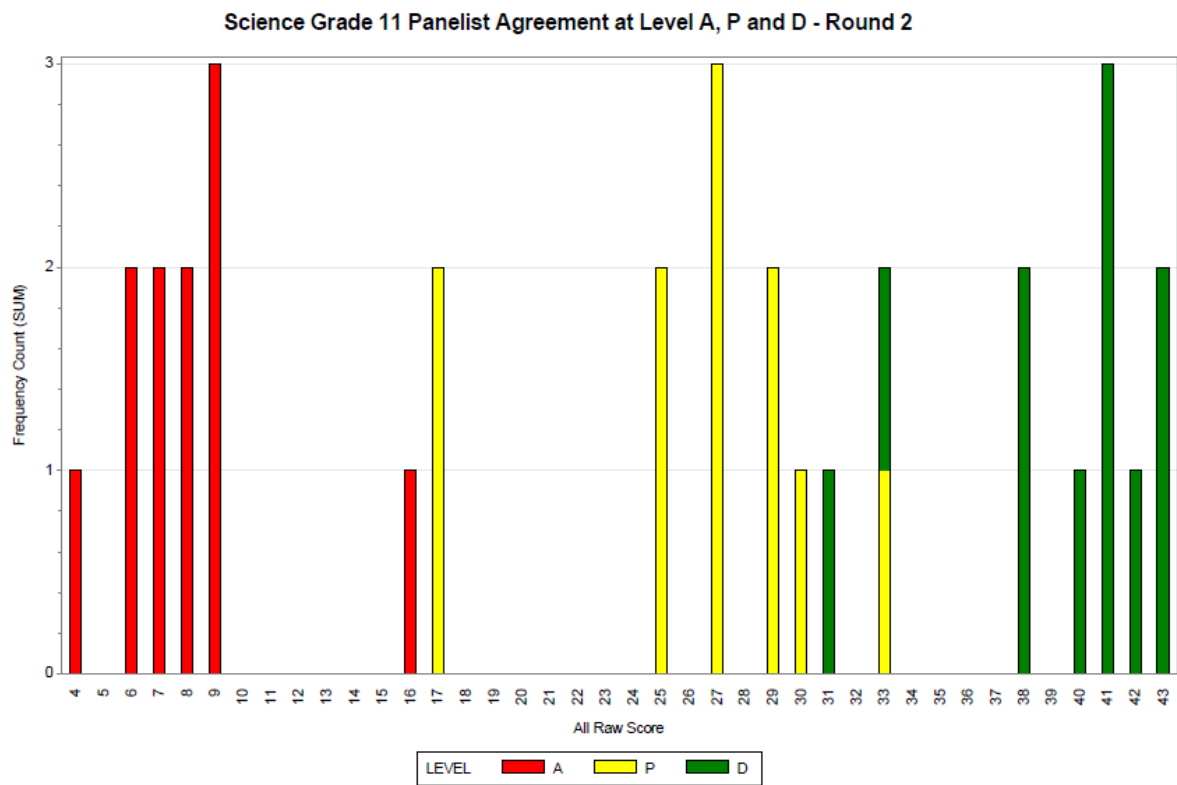


Figure I.2: Grade 11 Science Round 2 Panelist Agreement

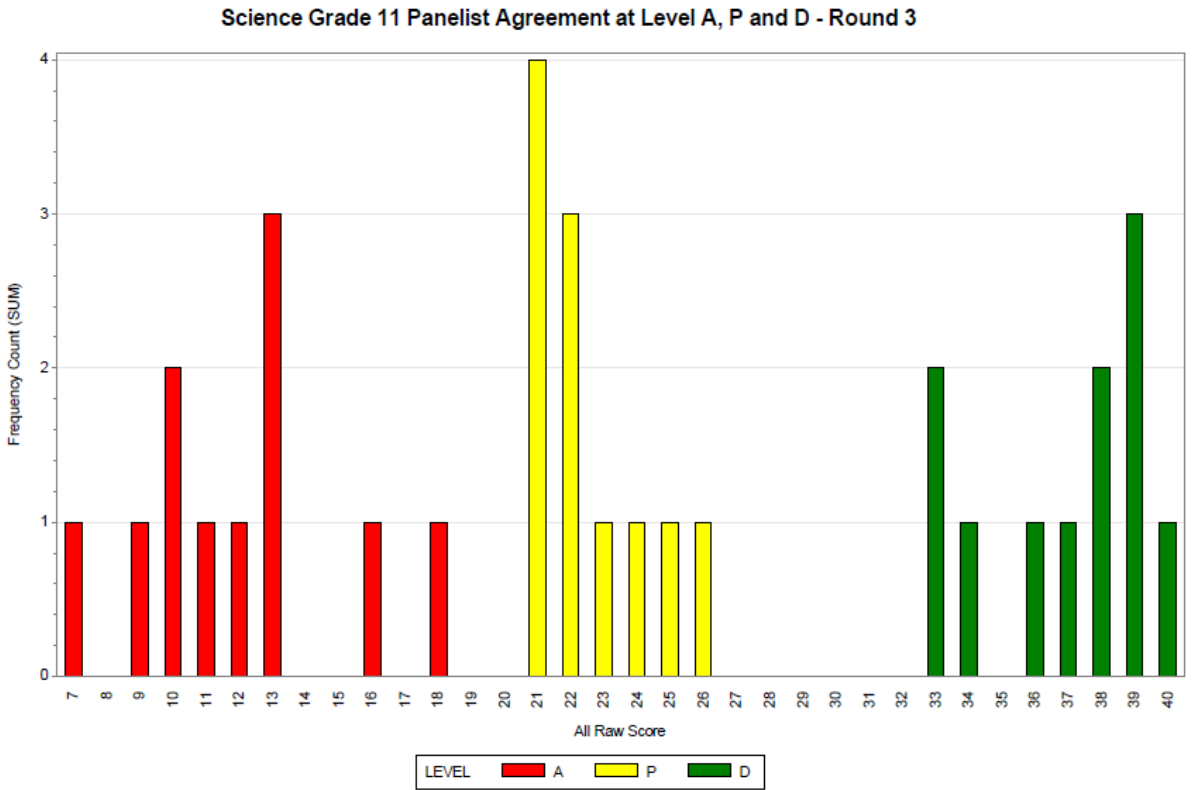


Figure 1.3: Grade 11 Science Round 3 Participant Agreement

Appendix J – Performance Level Descriptors

Kentucky Science Assessments Performance Level Descriptors (PLD) Science - Grade 11

Distinguished Performance Level

A student performing at the **Distinguished** performance level for grade 11 science has a comprehensive understanding of the three dimensions of the science and engineering concepts and practices incorporated in the Kentucky Academic Standards for science up through grade 11. The student consistently communicates ideas in a sophisticated and complex manner, using thorough supporting detail and explicit examples. The student reasons and solves problems by using appropriate strategies in an insightful way. Connections between concepts/ideas from different areas of science, when appropriate, are justified and insightful.

The student at the **Distinguished** performance level will demonstrate knowledge, skills, and abilities related to the Kentucky Academic Standards for grade 11 science such as:

1. Can evaluate and revise a sophisticated argument based on evidence to determine causal or correlational relationships.
2. Can make insightful predictions based on patterns evaluated in mathematical representations and computer simulations of phenomena.
3. Can consistently make and defend a claim based on valid information, or construct effective counter arguments.
4. Can predict complex cause and effect relationships from observed patterns within a system.
5. Can develop, evaluate, and revise a complex investigation.
6. Can use detailed models to justify a claim.

7. Can clearly identify non-obvious relationships within complex systems.
8. Can manipulate, evaluate and revise complex or incomplete models, including testing the reliability and merits and limitations of the model.
9. Can identify and evaluate solutions to a problem by constructing insightful explanations based on evidence.

Proficient Performance Level

A student performing at the **Proficient** performance level for grade 11 science has a broad understanding of the three dimensions of the science and engineering concepts and practices incorporated in the Kentucky Academic Standards for science up through grade 11. The student usually communicates ideas accurately using clear and appropriate examples, supporting or justifying those ideas with relevant details and evidence. Problem-solving and critical thinking skills are used effectively. Connections between concepts/ideas from different areas of science, when present, are reasonable and appropriate.

The student at the **Proficient** performance level will demonstrate knowledge, skills, and abilities related to the Kentucky Academic Standards for grade 11 science, such as:

1. Can construct a relevant argument based on evidence to determine causal or correlational relationships.
2. Can make predictions based on patterns identified in mathematical representations and computer simulations of phenomena.
3. Can make and defend a claim based on valid information, or construct counter arguments.
4. Can predict cause and effect relationships from observed patterns within a system.
5. Can plan and evaluate a complex investigation.
6. Can develop or use models to support a claim.
7. Can identify important relationships within systems.
8. Can manipulate, evaluate and revise models, including testing the reliability and merits and limitations of the model
9. Can identify and evaluate solutions to a problem by constructing appropriate explanations based on evidence.

Apprentice Performance Level

A student performing at the **Apprentice** performance level for grade 11 science has a basic understanding of the three dimensions of the science and engineering concepts and practices incorporated in the Kentucky Academic Standards for science up through grade 11. The student demonstrates some problem-solving and critical thinking skills, but they are not consistently applied. The student communicates ideas in a basic manner, but explanations, solutions or justifications may be unclear or ineffective.

The student at the **Apprentice** performance level will demonstrate knowledge, skills, and abilities related to the Kentucky Academic Standards for grade 11 science such as:

1. Can identify an argument based on limited evidence demonstrating a basic understanding of relationships.
2. Can identify patterns in mathematical representations and computer simulations of phenomena.
3. Can attempt to make a claim based on valid information.
4. Can select appropriate tools to collect and record data.
5. Can use simplistic or incomplete models to support a claim.
6. Can identify limited or basic relationships within systems.
7. Can perform limited evaluation or manipulation of a model which may include testing the reliability and merits and limitations of the model.
8. Can identify potential solutions to a problem based on evidence.

Novice Performance Level

A student performing at the **Novice** performance level for grade 11 science has a minimal understanding of the three dimensions of the science and engineering concepts and practices incorporated in the Kentucky Academic Standards for science up through grade 11. The student communicates ideas ineffectively or inaccurately, providing little detail and little or no support. Attempts at problem solving or critical thinking are minimal or inappropriate.

The student at the **Novice** performance level does not demonstrate the knowledge, skills, and abilities to be classified into the Apprentice performance level.

