APPENDIX M
2022 MCAS STANDARD SETTING REPORT

MCAS Standard Setting Meeting
Biology and Introductory Physics

August 2022

Pearson

Version 1.5

Table of Contents

Executive Report

MCAS Biology and Introductory Physics Standard Setting Meeting Executive Summary

August 2022

This report summarizes the process and results of setting achievement levels for the Massachusetts Comprehensive Assessment System (MCAS) assessments for Biology and Introductory Physics. The Massachusetts Department of Elementary and Secondary Education (ESE) partnered with Cognia and Pearson (the MCAS assessment contractors) to collect recommendations for cut scores associated with the achievement levels for the MCAS assessments.

MCAS Standard Setting Process and Results

Achievement levels are used to classify student achievement on an assessment. In order to classify student achievement into the four different levels, the following components are required: 1) policy-level definitions, 2) Achievement Level Descriptors (ALDs), and 3) cut scores. Policy-level definitions provide general descriptions of the knowledge, skills, and abilities students must demonstrate to be classified into each achievement level and apply to all courses or subject areas. ALDs illustrate the achievement levels in terms that are specific to a course or subject area. Cut scores represent the lowest boundary of each achievement level on the scale.

The process of recommending performance standards for the MCAS tests was based on standard setting procedures that were used for the MCAS tests for ELA, mathematics, and STE for grades 5 and 8, was in line with national best practice, and was conducted with review and approval of the MCAS Technical Advisory Committee (TAC). Results and details of the process are presented in the following sections.

Policy-level Definitions

Policy-level definitions for the MCAS achievement levels are shown in Table 1. The titles and descriptions of the achievement levels were defined to be part of a cohesive assessment system. The achievement levels indicate a student's ability to demonstrate proficiency in relation to subject- and grade-specific expectations, as indicators of a student's readiness for the next grade-level or college and career, as defined in the Massachusetts curriculum framework.

The Commissioner and the Board of Elementary and Secondary Education approved the final policy-level definitions for MCAS assessments in March 2017.

Table 1. Policy-level Definitions for MCAS Achievement Levels

Achievement Level	Policy-level Definition
Exceeding Expectations	A student who performed at this level exceeded grade-level expectations by demonstrating mastery of the subject matter.
Meeting Expectations	A student who performed at this level met grade-level expectations and is academically on track to succeed in the current grade in this subject.
Partially Meeting Expectations	A student who performed at this level partially met grade-level expectations in this subject. The school, in consultation with the student's parent/guardian, should consider whether the student needs additional academic assistance to succeed in this subject.
Not Meeting Expectations	A student who performed at this level did not meet grade-level expectations in this subject. The school, in consultation with the student's parent/guardian, should determine the coordinated academic assistance and/or additional instruction the student needs to succeed in this subject.

Achievement Level Descriptors (ALDs)

Draft sets of ALDs for the Biology and Introductory Physics, shown in Appendix A, indicate the knowledge and skills that students performing at a given achievement level should be able to demonstrate within each specific content area. Descriptors were developed for the *Partially Meeting Expectations, Meeting Expectations,* and *Exceeding Expectations* only. A student classified as *Not Meeting Expectations* has not demonstrated the knowledge, skills, and abilities necessary to achieve *Partially Meeting Expectations.*

A multi-step process was used to develop, review, and approve the ALDs for each test. Prior to the standard setting meeting, the DESE science test developers created the ALDs for each content area. Curriculum and instruction staff from DESE's Center for Instructional Support reviewed and commented on the ALDs. In addition, test developers from DESE's testing contractor, Cognia, reviewed and commented on the drafts. Finally, educators from DESE's Biology and Introductory Physics Assessment Development Committees (former and current members) reviewed and edited the drafts. These educators reviewed their content area ALDs and then reviewed the ALDs from the other content area. For example, Biology educators reviewed the Biology ALDs first and then reviewed the Introductory Physics ALDs. The educators first discussed their content area ALDs within their content area. This was followed by a whole group meeting of both sets of educators comparing the two sets of ALDs. The reason for the comparison between the content areas was to ensure similar rigor and expectations for each test at each achievement level. A final summary report of the ALD meeting will be included in the full standard setting report.

Teachers who participated in the standard setting committees had the opportunity to provide suggestions and edits to the draft set of ALDs. To produce the final ALDs, DESE science test developers will edit the draft ALDs based on suggestions generated by the participants in the standard setting meeting.

Cut Scores

The cut scores that were recommended for adoption for the MCAS assessments are based on a standardized set of procedures implemented during the standard setting meetings. General methods used during the meeting for obtaining the recommended cut scores are provided below.

Standard Setting Meeting

From August 9 to August 11, 2022, after the first year of operational administration in spring 2022, a standard setting meeting was conducted to obtain cut score recommendations for the next-generation high school science MCAS tests. There were two committees, with each recommending cut scores for one test:

1. Biology
2. Introductory Physics

Each committee was composed of 19 individuals, including teachers and non-teacher educators (e.g., administrators, curriculum specialists, professors of higher education). The participants were selected for the standard setting committee to provide content expertise during the committee meeting and to 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 method was used for the standard setting meeting (Davis & Moyer, 2015; Plake, Ferdous, Impara, & Buckendahl, 2005). This is a content- and item-based method that leads participants through a standardized process through which they consider student expectations, as defined by ALDs, and the individual items administered to students to recommend cut scores for each achievement level. The standardized process was used by the committees for each subject.

The process started with participants experiencing the test from the spring 2022 administration within the online testing system. Based on their experience with the test items and a review of the draft ALDs, panelists created borderline descriptions. During this process, participants worked within their committees to modify the draft ALDs to create descriptors of the knowledge, skills, and abilities that “borderline” students, or those students who just barely enter an achievement level, would be expected to demonstrate.

During the judgment process, participants reviewed each item on the test, referencing the borderline descriptions, and answered the following question for each achievement level:

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

The cut score recommendation for each individual participant was the expected raw score a borderline student at the respective achievement 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 3 students at the borderline level. Each recommended cut score from the standard setting committee was the median of the recommendations from the individual participants in

the committee.

Additionally, the percentage of students who would be classified into each achievement level based on committee recommendations—also known as impact data—was calculated. The impact data were determined using student data from the spring 2022 online administration. As part of the discussion of the round 2 judgments, the impact data were presented, based on the round 2 recommendations, so the participants could see the resulting student achievement level classifications prior to making their round 3 recommendations. This information was also presented after the round 3 cut score recommendations were calculated.

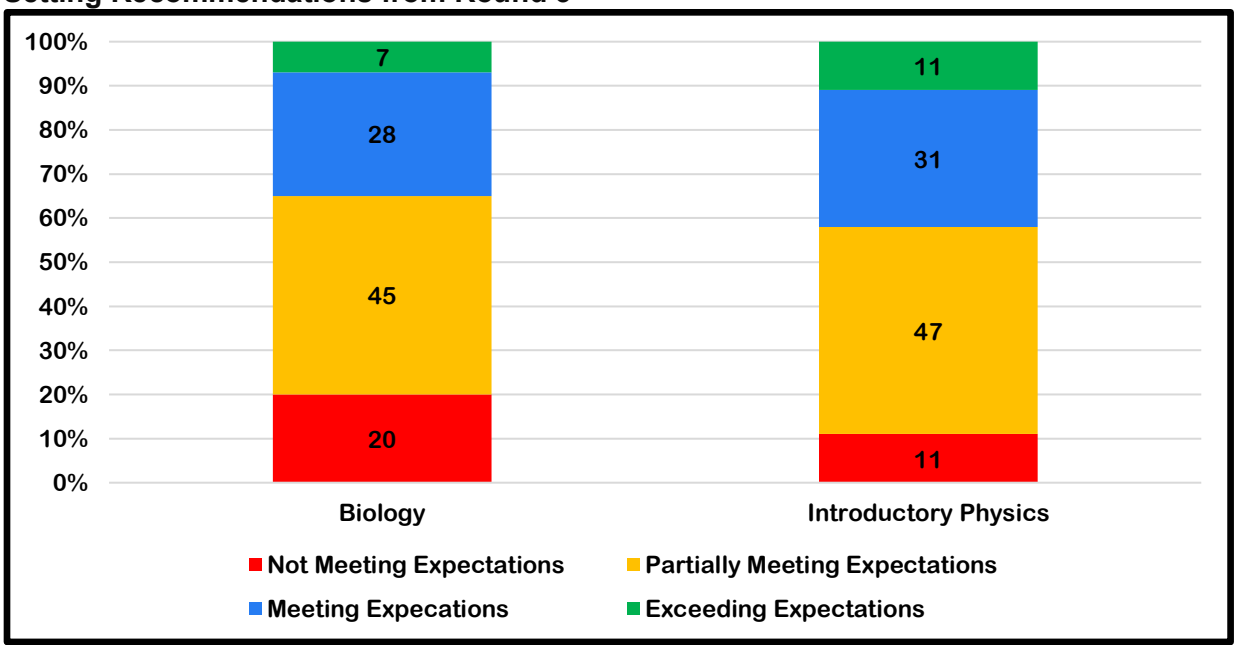
The results (Round 3 recommendations) from the standard setting meeting for the Biology and Introductory Physics panels are presented in Table 2.

Table 2. Standard Setting Recommendations for Biology and Introductory Physics (Round 3)

	Achievement Level							
	Not Meeting Expectations		Partially Meeting Expectations		Meeting Expectations		Exceeding Expectations	
Subject	Raw Score Range	% in Level	Raw Score Range	% in Level	Raw Score Range	% in Level	Raw Score Range	% in Level
Biology	0 to 16	20	17 to 36	45	37 to 51	28	52 to 60	7
Introductory Physics	0 to 15	11	16 to 36	47	37 to 50	31	51 to 60	11

Figure 1 presents the impact data from the final recommendations (Round 3) from the standard setting meeting as stacked bar graphs.

Figure 1. Impact Data for Biology and Introductory Physics Tests based on Standard Setting Recommendations from Round 3



Vertical and Horizontal Articulation Meeting

Subsequent to the standard setting meeting, on August 12, 2022, a vertical and horizontal articulation meeting was convened. The meeting consisted of one committee that reviewed the cut score recommendations from the Biology and Introductory Physics panels. The participants of the articulation meeting consisted of table leaders and other standard setting panel members selected prior to the standard setting meeting. The focus of the articulation meeting was to review the cut score recommendations from the standard setting meeting along with impact data to consider whether and to what extent adjustments to the recommended cut scores might be warranted based on both content and policy. In addition to the impact data for Biology and Introductory Physics, impact data for the Grade 8 STE test from the spring 2019 administration and matched data from 2022 Biology and Introductory Physics tests were presented to compare results both across grades and between subjects. The matched data was created using a statistical process to present impact data for both subjects based on students with statistically similar ability distributions. The adjustments to the recommendations made by the articulation committee were influenced by a desire to honor the content-based recommendations of the standard setting process, maintain high expectations for achievement across the MCAS assessments, and ensure the relationship among standards was coherent and defensible.

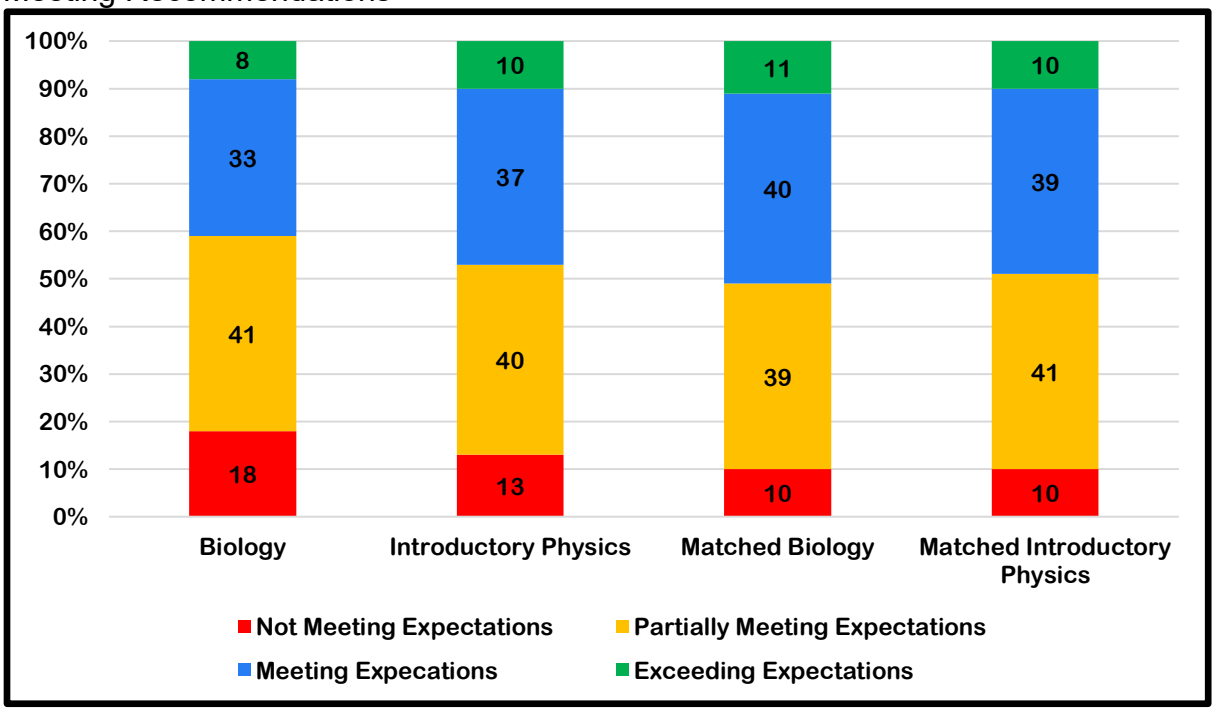
Table 3 presents the results from the vertical and horizontal articulation meeting.

Table 3. Recommendations for Biology and Introductory Physics from the Vertical and Horizontal Articulation Meeting Recommendations

	Achievement Level							
	Not Meeting Expectations		Partially Meeting Expectations		Meeting Expectations		Exceeding Expectations	
Subject	Raw Score Range	% in Level	Raw Score Range	% in Level	Raw Score Range	% in Level	Raw Score Range	% in Level
Biology	0 to 15	18	16 to 33	41	34 to 50	33	51 to 60	8
Introductory Physics	0 to 16	13	17 to 34	40	35 to 51	37	52 to 60	10

Figure 2 presents the impact data from the recommendations from the articulation meeting as stacked bar graphs, including the matched data for Biology and Introductory Physics.

Figure 2. Impact Data for Biology and Introductory Physics based on the Articulation Meeting Recommendations



Scaling

The process of determining the transformation rules from the Item Response Theory (IRT) scale to the reporting scale was guided by several principals identified by DESE:

1. The final cut scores achieved through the scaling solution should respect the cut score recommendations from the standard setting and articulation panels as closely as possible.
2. The impact data from the scaling solution should reflect a coherent assessment system across the grades.
3. The reporting scaled scores for the three achievement level cuts should be the same across grades and tests.
4. The scaling solution should involve a single linear transformation, from the IRT scale to the reporting scale.
5. The reporting scaled score range should be the same across grades and tests.

An iterative process involving Pearson, Cognia, and DESE was used to determine a scale and transformation rules for each test. First, based on the recommended raw score cuts for the three achievement levels, the IRT scale cuts were adjusted so that the differences between every two IRT scale cuts were the same, allowing for a single linear transformation rule. Based on the adjusted IRT cut scores, scaling constants for the linear transformation were determined. Using the scaling constants, look-up tables for each grade and test were created, displaying the relationship between the raw scores and reporting scaled scores. Based on the look-up tables, adjusted raw score cuts for each achievement level were determined. Finally, the resulting impact data based on the adjusted raw score cuts were calculated and reviewed to ensure a coherent system across grades.

The recommended reporting scale ranges from a lowest obtainable scale score of 440 to a highest obtainable scale score of 560. In order to create common points of reference across the assessments, the same scaled score cuts for each achievement level were defined, with a *Partially Meeting Expectations* cut of 470, a *Meeting Expectations* cut of 500, and an *Exceeding Expectations* cut of 530. While the cut scores were defined with the same scaled scores between the two tests, they are not identical, and direct comparisons through averaging and aggregation across grades should not be made without study and/or statistical adjustments. The scaled scores and distributions of students resulting from the cuts set for biology and introductory physics were not designed for direct comparison.

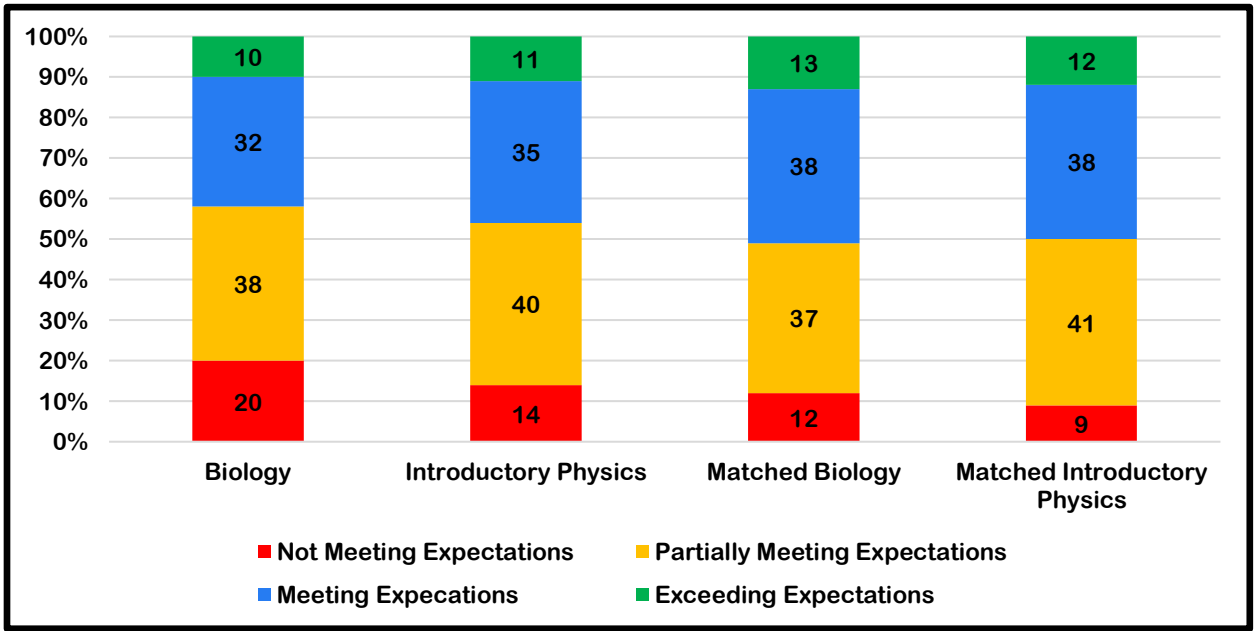
After the standard setting meeting, there was a discussion among DESE, Pearson, and Cognia staff about the results from the articulation and scaling. As a result of a need to bring the Biology more in line with content expectations from the standard setting committee, the Partially Meeting cut was raised to 17. As a result, the Exceeding cut was lowered to 50 and the Meeting cut was lowered to 34 to ensure proper scaling. Additionally, to bring the Introductory Physics more in line with the content expectations from the standard setting committee, the Partially Meeting cut was changed to 17. Table 4 presents the achievement level cut scores based on these changes.

Table 4. Final Recommendations for Biology and Introductory Physics Tests

Subject	Achievement Level							
	Not Meeting Expectations		Partially Meeting Expectations		Meeting Expectations		Exceeding Expectations	
	Raw Score Range	% in Level	Raw Score Range	% in Level	Raw Score Range	% in Level	Raw Score Range	% in Level
Biology	0 to 16	20	17 to 33	38	34 to 49	32	50 to 60	10
Introductory Physics	0 to 16	14	17 to 34	40	35 to 50	35	51 to 60	11

Figure 3 presents the impact data from the final recommendations as stacked bar graphs.

Figure 3. Impact Data for Biology and Introductory Physics based on Final Recommendations



The final approved result from this standard setting will be used for future administrations of the MCAS Biology and Introductory Physics tests to classify student results into achievement levels for reporting until it is determined that new standards need to be established for the MCAS by the DESE. To assist in this, the achievement level cut scores as raw scores are translated into achievement level cut scores on the item response theory (IRT) ability scale, as shown in Table 5. These values are uses to establish the scaling constants, A and B, to translate student scores on the IRT ability scale to the reporting scale.

Table 5. Achievement Level Cut Scores and Scaling Constants

	Cut Scores (Raw Score)			Cut Scores (IRT)			Scaling Constants	
Subject	**PME**	**ME**	**EE**	**PME**	**ME**	**EE**	**A**	**B**
Biology	17	34	50	-0.8500	0.2100	1.3000	27.90698	493.7209
Introductory Physics	17	35	51	-1.0100	0.1200	1.2600	26.43172	496.6960

Note: PME – Partially Meeting Expectations; ME – Meeting Expectations; EE – Exceeding Expectations

Interim Legacy Achievement Cut Score Validation

On the previous ("legacy") version of the Biology and Introductory Physics MCAS tests, a student was required for graduation to earn a Competency Determination (CD) by receiving a minimum scaled score of 220. As part of the transition to the next-generation MCAS, the Board of Elementary and Secondary Education voted to establish an interim CD standard for high school graduation. Interim standards would be defined as a similar level of achievement to the required standards on the legacy tests. Students in the classes of 2022 through 2025 taking the next-generation MCAS would be evaluated against the interim standards on each test.

The interim legacy achievement level standards were first identified through a statistical linking process. An equipercentile linking method was used to statistically establish an association between the raw scores from the spring 2019 and spring 2022 administrations of the MCAS tests. This was accomplished through determining the raw scores on the spring 2022 administration of the next-generation MCAS which would result in percentiles equal to those associated with the raw scores for each of the achievement levels from the spring 2019 administration of the legacy MCAS tests. Using the result of the equipercentile analysis, standard errors of measures for the raw scores, and statistical results from the test construction process, recommended ranges for raw scores associated were determined for each achievement level cut score.

After the standard setting panels completed their cut score recommendations, a subset of panelists was convened to recommend interim legacy MCAS achievement level cut scores from recommended ranges. The panelists reviewed the performance of students from the spring 2019 administration on the legacy MCAS to determine general descriptions of the achievement of students at the borderline of each legacy achievement level. The general descriptions were then used by the panelists to review the performance of students within the raw score ranges from the spring 2022 administration on the next-generation MCAS. Based on their review, the panelists completed a judgment survey where they answered the following question:

"Based on your review, which raw score within the recommended range for the achievement level on the Next-Generation MCAS test most closely represents a similar achievement level on the legacy assessment?"

Panelists provided individual recommendations for each achievement level, Needs Improvement, Proficient, and Advanced. The median of the committee recommendations was used as the committee recommendation for the achievement level. Table 6 displays the interim cut score recommendations for the legacy achievement levels on the next-generation MCAS.

Table 6. Recommended Cut Scores for the Legacy Achievement Levels

Subject	Legacy Achievement Levels		
	Needs Improvement	**Proficient**	**Advanced**
Biology	16	29	46
Introductory Physics	17	29	47

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.

Chapter 1 – Overview of the Standard Setting Process

This chapter provides an overview of the standard setting process used for the MCAS ELA and mathematics assessments for grade 10 and STE assessments for grades 5 and 8, and includes the following sections:

3. Goals of setting cut scores
4. MCAS achievement levels
5. MCAS cut score setting process

Goals of the Standard Setting Meeting

Once students are administered an assessment, various groups, including students, parents, educators, administrators and policy makers, want to know how the students performed on the assessment and how to interpret that performance. By establishing achievement levels associated with different student performance on the assessment, a frame of reference is developed for interpreting student scores. Setting the level of achievement on an assessment sufficient for student performance to be classified into each achievement level is one of the most critical steps in developing an assessment program.

For a criterion standards-based assessment, such as the next-generation MCAS program, achievement on the assessment is compared to a set of predefined content standards. The standards communicated within the *Massachusetts Curriculum Framework* define a set of knowledge, skills, and abilities the students taking the assessment are expected to demonstrate upon completion of each course or grade. The cut scores established represent the level of competence students are expected to demonstrate on the assessment to be classified into each achievement level.

MCAS Achievement Levels

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

The policy-level ALDs for the achievement levels provide general expectations for student achievement on the MCAS assessments to be classified into each achievement level. These do not differentiate student performance between content areas and grade levels. The achievement levels and policy ALDs for the next-generation MCAS assessments were developed with input from the Standard Setting Policy Committee. This 14-person committee is comprised of Massachusetts educators and policy makers representing K–12 education and higher education constituency groups (including MASS PTA, MASC and BESE, among others). Language for these levels was refined by the Massachusetts BESE at its monthly meeting in December 2016, and after eliciting public feedback, final Next-Generation MCAS Achievement Levels and Descriptors were adopted by BESE in March 2017.

The four achievement levels with their respective policy description are shown in Table 1.

Table 1. Policy Level Achievement Level Descriptors for the Next-Generation MCAS Tests

Label	Description
Exceeding Expectations	A student who performed at this level exceeded grade-level expectations by demonstrating mastery of the subject matter.
Meeting Expectations	A student who performed at this level met grade-level expectations and is academically on-track to succeed in the current grade in this subject.
Partially Meeting Expectations	A student who performed at this level partially met grade-level expectations in this subject. The school, in consultation with the student's parent/guardian, should consider whether the student needs additional academic assistance to succeed in this subject.
Not Meeting Expectations	A student who performed at this level did not meet grade-level expectations in this subject. The school, in consultation with the student's parent/guardian, should determine the coordinated academic assistance and/or additional instruction the student needs to succeed in this subject.

The MCAS Standard Setting Process

The recommendations by the standard setting committees represent the level of competence students are expected to demonstrate to be classified into each of the achievement levels. To establish the achievement levels for each assessment, the Extended Modified (Yes/No) Angoff Method (Davis & Moyer, 2015; Plake, Ferdous, Impara, & Buckendahl, 2005) was used to guide participants as they determined their achievement 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 achievement levels, including content standards and educator judgments about what students should know based on the *Massachusetts Curriculum Framework* and be able to demonstrate at each achievement level.

The following steps were used for the MCAS standard setting process.

1. Pre-meeting development – In anticipation of the standard setting meetings, various tasks were completed, including the development of draft ALDs for each grade and subject assessed, the development of materials for the participants, preparation of the Pearson standard setting website for participants and facilitators, presentation materials for the facilitators, and development of data analysis sources and procedures.
2. Standard setting meetings – Committees of participants referenced the grade- and subject-specific ALDs to make recommendations for cut scores that define the different achievement levels for each assessment.
3. Articulation meeting – The recommended cut scores for each assessment were reviewed for reasonableness and alignment of achievement-level expectations across the courses, Biology Introductory Physics, and with expectations from Grade 8 STE, by select members of the standard setting committees.
4. Competency determination validation meeting – The statistically determined cut scores associated with the previous MCAS assessments for Biology and Introductory Physics were reviewed for consistency of content expectations by select members of the standard setting committees.

5. Linear scaling – Using the recommended cut scores from the articulation meeting, a scaling transformation process was conducted to transform the IRT scale scores to MCAS scale scores.

The following chapters will describe the specific procedures and activities that occurred during each of these steps.

Chapter 2 – Pre-meeting Development

This chapter provides an overview of the work that was completed prior to the standard setting meetings for the next-generation MCAS Biology and Introductory Physics assessments, and includes the following sections:

6. MCAS achievement level descriptors
7. Development of participant materials
8. Development of presentation materials
9. Facilitator training
10. Preparation for data analysis during the meetings

MCAS Achievement Level Descriptors

ALDs are statements that articulate the knowledge, skills, and abilities that students classified into a particular achievement level should be able to do to demonstrate competency at that achievement level. All assessments within MCAS have four achievement levels, as defined in Table 1. The achievement levels range from Not Meeting Expectations, representing the lowest level of student achievement, to Exceeding Expectations, representing the highest level of student achievement. ALDs were not developed for the lowest achievement level, Not Meeting Expectations. The most accurate way to describe performance classified into the "Not Meeting Expectations" achievement level is as a student who has not demonstrated the knowledge and skills necessary to achieve "Partially Meeting Expectations."

The ALDs are associated with the achievement levels in the following way.

1. *Achievement levels* indicate a student's level of competency of the standards defined in the *Massachusetts Curriculum Framework* through classification of their achievement on an assessment for a specific grade and subject as *Not Meeting Expectations*, *Partially Meeting Expectations*, *Meeting Expectations*, and *Exceeding Expectations*.
2. *Achievement level descriptors* indicate the knowledge, skills, and abilities expected of students to demonstrate competency within each specific content area and at each grade level to be classified in each achievement level.
3. *Cut scores* partition the test scale and represent the minimum test score that a student must earn on an assessment for each subject and grade level to be classified into a given achievement level.

The use of a well-defined set of ALDs is critical to ensuring the validity of the standard setting process.

The ALDs were developed by DESE test development staff, in consultation with staff in the Center for Instructional Support and Cognia test development staff. In addition, educators from the Biology and Introductory Physics Assessment Development Committees (ADCs) were convened to review drafts of the ALDs prior to the standard setting meeting. The educators first reviewed ALDs for their specific content area in terms of the appropriateness of the abilities for each achievement level, and the clarity and logic of the progression across the achievement levels. After the validation or editing of the draft ALDs, the educators reviewed the ALDs from the other content area to compare the expectations of the ALDs for each performance level across the content areas. The resulting ALDs from this committee were used during the

standard setting meeting.

Pearson Standard Setting Website

The Pearson standard setting website is the online platform for meeting pre-work, facilitating the standard setting meeting and collecting panelist judgments throughout the standard setting process. Because the next-generation MCAS assessments are computer-delivered and the online test form were used for the standard setting process, the standard setting website provides panelists the opportunity to access online items within Pearson's secure online testing environment, TestNav 8. During the meeting, panelists accessed the website using a notebook computer provided by Pearson and set up specifically for the meeting.

Using a similar template to the websites used for the MCAS standard setting in 2017 and 2019, specific websites were created for each committee meeting by the Pearson standard setting team. The staff at DESE had the opportunity to review the website structure prior to finalizing the websites for the meeting. Additionally, members of the Pearson staff performed reviews of the websites to verify that the content on the website was correct.

Development of Participant Materials

The MCAS standard setting required a large number of materials be prepared for use by the participants during the standard setting meetings. The Pearson standard setting team worked with the content specialists at DESE to develop the materials and to ensure that all materials provided to meeting participants communicated correct information. The following materials, displayed in Table 2, were developed for use by participants during the meeting.

Table 2. Materials Prepared for Panelists

Panelist Material	Paper	Online
Meeting agenda	✓	✓
Panelist information survey		✓
Non-disclosure agreement		✓
Next-generation MCAS test forms/items		✓
"Experience the Test" response form	✓	
Test form item map/answer key		✓
Item comment form	✓	
Practice judgment form/items		✓
Practice judgment form item map/answer key		✓
Judgment round record form	✓	
Judgment round surveys		✓
Achievement level descriptors (ALDs)	✓	✓
ALD comment form	✓	
Process evaluations		✓

Using approved templates, documents were created for each specific committee meeting by the Pearson standard setting team. All documents developed for the website were reviewed and approved by DESE staff before being finalized for publication for the meetings. A sample set of materials for a committee are provided in Appendix C.

Development of Presentation Materials

PowerPoint presentations were developed to guide facilitators through the presentation of information and materials throughout the standard setting meetings. The Pearson standard setting team developed the initial PowerPoint presentations using the DESE presentation template. Staff from DESE 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.

1. Standard Setting General Session
2. Standard Setting Breakout Meeting presentations for each day
    1. Biology (Day 1, Day 2, and Day 3)
    2. Introductory Physics (Day 1, Day 2, and Day 3)
1. Articulation Meeting (Day 4)
    1. Biology and Introductory Physics with Grades 8 STE
    2. Biology with Introductory Physics
1. Competency Determination Validation Meetings (Day 4)
    1. Biology
    2. Introductory Physics

The PowerPoint presentations for the breakout meetings, Day 1 and Day 2, were customized to reflect the specific information for the subject and grades for each committee. Additionally, specific information was added to the notes section within each presentation to guide the facilitators through the presentations.

Facilitator Training

Procedures employed in the standard setting meeting are specific to the goals and objectives of the project. So, even though the facilitators for the MCAS standard setting meeting had prior experience in facilitating standard setting meetings, a training session was held to discuss the unique aspects of the MCAS standard setting and to walk through the process utilized for this meeting, demonstrate the use of the Pearson Standard Setting website, and display and discuss the PowerPoint presentations used during the standard setting meetings. The facilitator training meetings were held for 90 minutes each on July 15, 2022, and August 3, 2022. Additionally, there was a final training and discussion held on August 8, 2022, just prior to the meeting, to address any final topics.

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 independent analysis for each judgment round during the meeting, each analyst independently completed the programming necessary to conduct all analysis using the SAS statistical software. A trial was run with mock-data to ensure that each independent analysis generated the same results.

Prior to the standard setting meeting various data sets were generated for use prior to and during the standard setting meeting, including both the articulation and CD meetings. Table 3 presents the data in information that was generated and the purpose of the data/information.

Table 3. Data and Information Generated for the Standard Setting Process

Data/Information	Purpose
Frequency distribution of raw scores	This data was used during the standard setting meeting to provide reference data for panelists judgments. The impact data for standard setting was based on the whole tested student population after Cognia data cleaning.
Item performance data	The item mean and score distribution was calculated for each item on the standard setting form as reference data during the standard setting meeting.
Grade 8 STE impact data	The impact data for grade 8 STE from the spring 2019 administration was used during the articulation meeting as reference information for vertical articulation.
Frequency data for Biology and Introductory Physics using matched data	A matched sample was created for Biology and Introductory Physics from the spring 2022 administration to create impact data for the articulation meeting as reference information for the horizontal articulation.
Equipercentiles for the legacy achievement levels from the spring 2018, 2019, and 2021 administrations on the new MCAS spring 2022 form	The equipercentiles for each of the legacy achievement levels on the Biology and Introductory Physics will be used to inform the creation of the reasonable range for the CD validation meetings.
Student profile data on legacy Biology and Introductory Physics from spring 2019 administration	Student performance profiles showing item mean scores and distributions for operational items on the form at the legacy achievement level raw score cuts using the spring 2019 administration data. These score profiles will be used as information for the CD validation meeting.
Student profile data on new Biology and Introductory Physics from spring 2022 administration	Student performance profiles showing item mean scores and distributions for operational items on the form using the spring 2022 administration data. These score profiles will be used as information for the CD validation meeting.

Impact data look-up data sets were created for use during the standard setting meetings. Impact data are the percent of students that fall within an achievement level based on the recommended cut scores at the given judgment round for a particular grade, subject test, and testing mode. The impact data are provided to participants during the standard setting meeting to present the expected results of their recommendations on student achievement level

classifications. The analysis programs use impact data lookup tables to produce this output during the meetings, which need to be created prior to the standard setting meetings.

The impact data lookup tables were created using the data from students taking the online form of each subject and grade assessment during the spring 2022 administration. The impact data lookup tables were created using a sample of students that would be representative of the overall state student population, based on the following demographic variables:

- Gender
- Race/Ethnicity
- Economically disadvantaged
- Limited English Proficient (LEP)
- Special Education

The data analysts created the impact data lookup tables by calculating, for each possible raw score associated with the test, the percent of overall students in the sample that earned that specific raw score or greater.

As planned, a proof of concept of study was conducted with datafiles provided by DESE to evaluate the usability of matched data for the articulation meeting. Based on recommendations from the MA TAC, a bi-directional matching process was utilized, which matched both the Biology and Introductory Physics populations from the spring 2022 administration were matched to the testing population from the spring 2022 administration using a bi-directional matching process.

Bi-directional matching conducts the matching twice, each in a different direction. First of all, a subsample from the original “Biology” group is drawn to match the “Physics” group. The resulting matched sample is denoted as Physics-equivalent group. Second, a subsample from the original “Physics” group is drawn to match the original “Biology” group. The resulting matched sample is denoted as Biology-equivalent group. Then the original “Biology” group is combined with the Physics-equivalent group to form the matched “Biology” group; similarly, the original “Physics” group is combined with the Biology-equivalent group to form the matched “Online” group. To run this matching method, R package, called as “Matching” was used. The following table show the matching results of individual background variables used for R program.

Table 4. Summary Result from Bi-directional Matching Process

			Biology		Physics	
Variable	**Dummy Variable**	**Value**	**Frequency**	**Percent**	**Frequency**	**Percent**
Gender	**Female**	**0**	18734	51.51	18958	52.13
		1	17633	**48.49**	17409	**47.87**
Ethnicity	**Race_W**	**0**	11974	32.93	11925	32.79
		1	24393	**67.07**	24442	**67.21**
	Race_B	**0**	33256	91.45	33180	91.24
		1	3111	**8.55**	3187	**8.76**
	Race_A	**0**	34120	93.82	34170	93.96
		1	2247	**6.18**	2197	**6.04**
	Race_H	**0**	29394	80.83	29502	81.12
		1	6973	**19.17**	6865	**18.88**
Economic Status	**EDS**	**0**	25620	70.45	25424	69.91
		1	10747	**29.55**	10943	**30.09**
ELL Program	**ELL**	**0**	34242	94.16	34376	94.53
		1	2125	**5.84**	1991	**5.47**
Special Education	**Sp E.**	**0**	29879	82.16	29704	81.68
		1	6488	**17.84**	6663	**18.32**
Previous Science Score	**2021 Science**	**Mean**	496.50		496.56	
		SD	23.12		23.00	

Background

For the competency determination validation meetings for Biology and Introductory Physics, there were additional analyses that were performed in preparation for the meeting. Pearson worked with Cognia and DESE to statistically identify reasonable ranges around the interim cut scores for the legacy achievement levels, Needs Improvement (220), Proficient (240), and Advanced (260), for the CD validation meeting using an equipercentile process. This process determined cut scores on the next-generation MCAS tests which would result in similar impact data from the 2019 administration of the legacy MCAS.

Chapter 3 – Standard Setting Meetings

This chapter provides details about the cut score setting meeting process. The sections of this chapter include:

- Purpose of standard setting meetings
- Committee participant composition
- Standard setting meeting facilitators and staff
- Standard setting meeting proceedings
- Recommended achievement 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 performance expected for each achievement 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 judgments and to translate these into cut scores for the achievement levels. The purpose of the next generation MCAS standard setting meetings was to gather expert recommendations from groups of educators from across Massachusetts for the cut scores that define the different achievement levels on each MCAS assessment for Biology and Introductory Physics.

Student performance on each of the MCAS assessments is classified into one of four achievement levels. Each committee was asked to recommend three cut scores that would define the boundaries between the different achievement levels. These recommended cut scores represent the performance on each assessment that a student would need to meet or exceed to be classified into the specific achievement level.

Committee Participant Composition

All participants for the standard setting committees were selected by the DESE, then recruited and invited to participate in the standard setting meeting by Cognia. The process of selecting committee participants included selecting a sample of participants that would be as representative of the state as possible, including demographic variables (gender, race, etc.), geographic representation, and background (educational experience, education, etc.). When selecting participants, DESE 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 38 participants at the standard setting meetings, who were divided between two committees. Each committee focused on providing cut score recommendations for one assessment. The participants were assigned to the committee prior to the meetings based on their teaching experience. The tables in Appendix D summarize the characteristics and experience of the participants in each committee. These tables provide demographic information about the committee participants as well as information about the participant's current positions in education, their experience working with various types of student populations, and the types of districts they represent. Participant's responses to the gender and ethnicity questions were voluntary.

The participants 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 participants at each table were representative of the committee. The participants were placed into table groups to facilitate discussions during the standard setting meeting and ensure that each participant had the opportunity to fully engage in the process.

Prior to the standard setting meeting, individuals were selected from the participants to serve as table leaders for each committee. One table leader was assigned to each table group. The table leaders assisted the process facilitator during the meeting by helping to facilitate the table group discussions, ensuring that all participants had the opportunity to participate and that the discussion remained relevant to the meeting. To assist the table leaders in understanding and fulfilling their role during the meeting, Eric L. Moyer, Ph.D., the lead facilitator for the meeting, provided a table leader training on the first day of the standard setting,

Standard Setting Meeting Facilitators and Staff

Staff members from DESE, Cognia, and Pearson collaborated to conduct the MCAS standard setting meeting. These staff members worked in facilitative and observational roles and did not contribute to the cut score recommendations during the meeting.

Meeting Facilitators

The lead facilitator of the standard setting meeting was Eric L. Moyer, Ph.D., from Pearson. For each of the four breakout committees there were two facilitators assigned, a process facilitator and a content facilitator. The process facilitator was a member of the Pearson psychometric staff with experience in facilitating standard setting meetings and was responsible for leading the participants through the standard setting process. The content facilitators were content specialists familiar with the content for the MCAS assessment from DESE or Cognia was responsible for leading the participants through the information associated with the development of the test and procedures for scoring the items. Table 5 presents the process and content facilitators for each standard setting committee.

Table 5: Process and Content Facilitators for Standard Setting Committees

Committee	Facilitators	
Subject	**Process Facilitator**	**Content Facilitators**
Biology	Soo Ingrisone, Ph.D.	Katie Bowler Steven Long (DESE)
Introductory Physics	Scott Strickman, Ph.D.	Isadel Eddy (DESE) Phil Durham (Cognia)

Meeting Data Analysts

For the standard setting meeting, two data analysts performed all of the analysis for all four committees. The data analysts were Brian Wrobel and Michelle Anderson. During the meeting, the analysts collected participant judgment data, performed independent analysis to verify analysis results, and prepared participants' feedback. Brian Wrobel was the lead analyst and performed the analysis onsite, while Michelle Anderson was the replicator and completed the

analysis offsite.

DESE Staff

DESE staff members attended the standard setting meeting to observe the process, answer assessment and curriculum questions, and address policy questions. DESE staff also monitored the cut score recommendations for each achievement level throughout the standard setting meetings. DESE was represented at the cut score setting meeting by Michol Stapel, the Associate Commissioner for Student Assessment, and Robert Lee, the MCAS Chief Analyst. These were assisted by additional DESE staff to monitor the standard setting meeting, including content and assessment specialists.

Technical Advisors

A technical advisor, Dr. William Lorie, Ph.D., a staff member from the Center for Assessment, monitored the standard setting meetings for DESE. The technical advisor observed the standard setting meetings and gave his advice and findings to the DESE after the meeting. The technical advisor did not participate in the meeting or contribute to the cut score recommendations during the meeting.

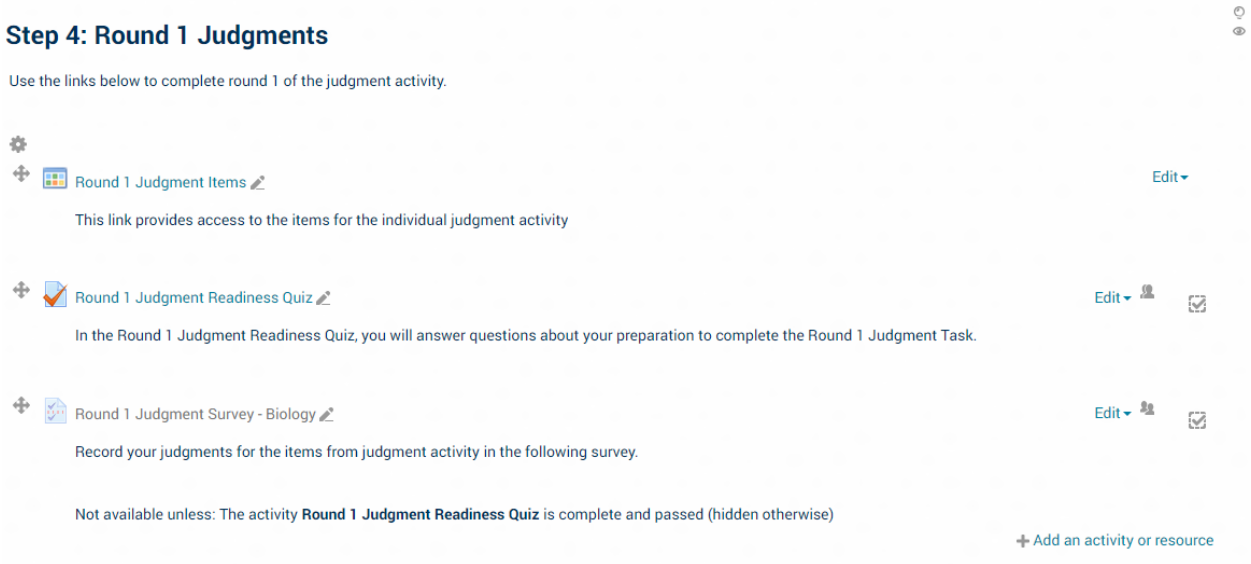
Materials

The following section describes the materials used by the committee members during the standard setting breakout sessions. Separate materials were developed for each committee.

Pearson Standard Setting Website

The Pearson standard setting website served as the online platform during the standard setting meetings. The website provided panelists access to the standard setting meeting materials and tools used to collect panelist judgments (see Figure 1). 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, as defined by their assigned subject area. Because the next-generation MCAS assessments are computer-delivered using TestNav 8, the standard setting website allowed panelists to view items as students did during the spring 2022 administration.

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



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

- Standard setting orientation video
- MCAS curriculum framework for each grade level
- Subject- and grade-level ALDs

1. MCAS 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:

- Subject- and grade-level ALDs
- 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 survey for rounds 1, 2 and 3
- Judgment survey for rounds 1, 2, and 3
- Judgment feedback folders for rounds 1, 2, and 3
- Process evaluations 1 and 2
- Participant information survey

A unique course site was created for each assessment associated with the 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 provided through the website, panelists were given a meeting folder to organize a variety of hard copy materials they used throughout the meeting. The materials provided to committee panelists in their folders included:

1. Meeting agenda
2. Non-disclosure agreement
1. Subject- and grade-specific ALDs
2. ALD comment form
3. “Experience the assessment” activity response form
4. Item comment form
1. Practice judgment record form
2. 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 participant was provided a laptop computer in his or her meeting room to access the online resources through the Pearson standard setting website. Additionally, participants were provided an external monitor, so they would be able to access the online materials with limited switching between online windows. Participants were seated in table groups in pod configuration to provide each participant with enough space to work with the computer and binder materials. The power supplies were centrally located in the middle of each table. The participants used Google Chrome to access the Moodle site, which was programmed with a whitelist of websites to restrict participants use of the computers to work associated with the cut score setting meeting.

Procedure

The Extended Modified (Yes/No) Angoff Method (Davis & Moyer, 2015; Plake, Ferdous, Impara, & Buckendahl, 2005) was used during the standard setting meeting to assist participants in recommending achievement level cut scores for each assessment. 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 ALDs, and the individual items administered to recommend cut scores for each performance level. This method asked participants to review each item from the operational administration and answer the following question:

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

For the standard setting meeting, “likely” was defined statistically as the student having at least a 2/3 chance of earning the number of points. The participants completed the task for each achievement level.

The same standardized process was used by all committees and resulted in cut score recommendations. Participants completed three rounds of item judgments. Between the item

judgment rounds they were provided feedback information including data relative to participant agreement, student performance on the items, and student performance on the test as a whole.

Standard Setting Meeting Proceedings

The standard setting meetings were conducted across three days, August 9-11, 2022, in Wakefield, Massachusetts. Appendix E includes the complete agenda for the standard setting meetings. The following sections will describe the steps used to guide the participants through the entire standard setting process.

Standard Setting Meeting Pre-Work

The standard setting meeting participants were provided access to a set of activities prior to attending the onsite meetings. The purpose of the pre-work was to expedite the training of the participants by providing the participants an opportunity to familiarize themselves with information that would be used throughout meeting. The pre-work included:

1. Standard setting website – The pre-work was provided via documentation or links embedded within the secure website developed for the standard setting meeting. This allowed participants to access the website and gain some familiarity with navigation in the site prior to the meeting.
2. Participant information survey – Participants were provided a survey to document their demographic information as well as current teaching position, experience, and school information. Participants were able to access this survey before and during the meetings.
3. *MCAS Curriculum Framework* – Participants were provided access to the current version of the *MCAS Curriculum Framework* for the subject associated with their meeting.
4. ALDs - Participants reviewed policy level and achievement level descriptors for the specific grade and course, which is a key set of information that was used throughout the cut score setting meeting.
5. Security and Non-disclosure – Participants were provided access to the security and non-disclosure agreement for the standard setting meeting so they would be familiar with its content before signing the agreement at the meeting.

To provide the participants access to the pre-work materials prior to the meeting, they were supplied their unique login and a temporary password for the website to the email they provided when they registered for the meeting. This login provided them access to the specific section of the website associated with the standard setting meeting for which they were registered. Participant access was restricted to only the respective site for the standard setting meeting they were attending.

General Session

The purpose of the general session was to welcome the participants, provide background information about the next-generation MCAS assessment system, and introduce the standard setting process. A single general session including all 38 standard setting participants was conducted on Monday morning at the beginning of the standard setting meeting. Rob Curtin, Chief Officer for Data, Assessment, and Accountability provided a welcome to the Massachusetts educators and an overview of history of the MCAS assessment program. The official charge for the meeting along with a review of related student performance statistics was provided by Michol Stapel and Katie Bowler. An overview of the cut score setting process was provided by Eric Moyer, the lead research scientist from Pearson facilitating the standard setting process.

Breakout Session

After the general session, participants moved into subject-specific breakout sessions for the remainder of the standard setting meeting. Each committee was responsible for providing recommendations for cut scores for each of the achievement levels for the test associated with the committee. The committee provided recommendations using each of the activities described below.

Experience the Test. Participants experienced the specific operational test form that the students were administered during the spring 2022 administration. The participants experienced the test just as students did, online administered through the TestNav 8 system, which was accessed through the standard setting website.

Since the version of the online testing system used during the standard setting meetings did not store and score participant responses, participants recorded their responses on a separate item response form, provided in the participant folder. During this activity, if the participants wanted to provide any comments regarding items on the test form, they were asked to record the comments on an Item Comment Form, which was collected at the end of the meeting.

After the participants completed the Experience the Test activity, the content facilitators provided instruction on how to score the items based on the scoring rules used for MCAS. A test map document, accessed through the standard setting website, provided information about each item on the test, including a unique item number, reporting category, maximum possible score, the correct response for the item, and any specific scoring rules for the item. For open-ended items, the test map provided a reference to the open-ended item rubric and exemplar documents so the participant could see what was expected to earn each possible score point. Participants were also provided training on characteristics that make an item difficult, in addition to how to use the rubric to score responses for open-response items and how these corresponded to the student exemplar response scores.

Borderline Achievement Level Descriptions. An essential component to the standard setting process is the development of borderline descriptions. The purpose of the borderline descriptions activity was for panelists within a committee to develop a common understanding of student achievement at the threshold, or borderline, of each achievement level.

To help inform this activity during the standard setting meeting, the process facilitators reviewed the achievement levels and the achievement level descriptors for the respective grade and subject. Panelists were informed that the ALDs provided a snapshot of the typical characteristics at each achievement level, including the breadth and depth of the knowledge, skills, and abilities expected to be demonstrated by students within each level. The participants reviewed the course-specific ALDs, providing them with a common understanding of expectations for what students should demonstrate within each achievement level for the respective assessment.

The participants were then introduced to the difference between a *typical* student performance and *borderline* student performance within an achievement level. The borderline student performance was described as the performance to be minimally qualified to be classified within a particular achievement level, possessing just enough knowledge, skills, and abilities to achieve the specific achievement level classification. The facilitator then led the panelists through a modeling activity. A collaborative and guided approach was used to draft one or two borderline statements for the *Meets Expectations* achievement level that served as examples

for the committee. The facilitator asked guiding questions during the modeling activity to help panelists develop an appropriate understanding of how to create borderline descriptions.

Panelists were then split into their table groups to review the ALDs for a specific reporting category within each achievement level. Each small group created draft borderline descriptions for their specific reporting category using a borderline descriptions worksheet accessed through the standard setting website. The borderline descriptions from each group were collected into a master document and reviewed/discussed together by the whole committee. Revisions to the master document were made during the whole-group discussion to create a common set of borderline descriptions.

The final list of borderline descriptions were printed and provided to each participant to place in his or her folder as a reference for subsequent activities.

Item Judgment Process Training. The process facilitator for the committee provided the participants with training on the Extended Modified (Yes/No) Angoff standard setting process (Davis & Moyer, 2015; Plake, Ferdous, Impara, & Buckendahl, 2005) and how to record their individual item judgments within the standard setting website. They were instructed to review each item from the assessment, which was accessed through the website, review the borderline descriptions, the answer key, and, if needed, the rubric and student exemplars for the item. Based on their review of the item and the related materials, the participants answered the following question for each achievement level:

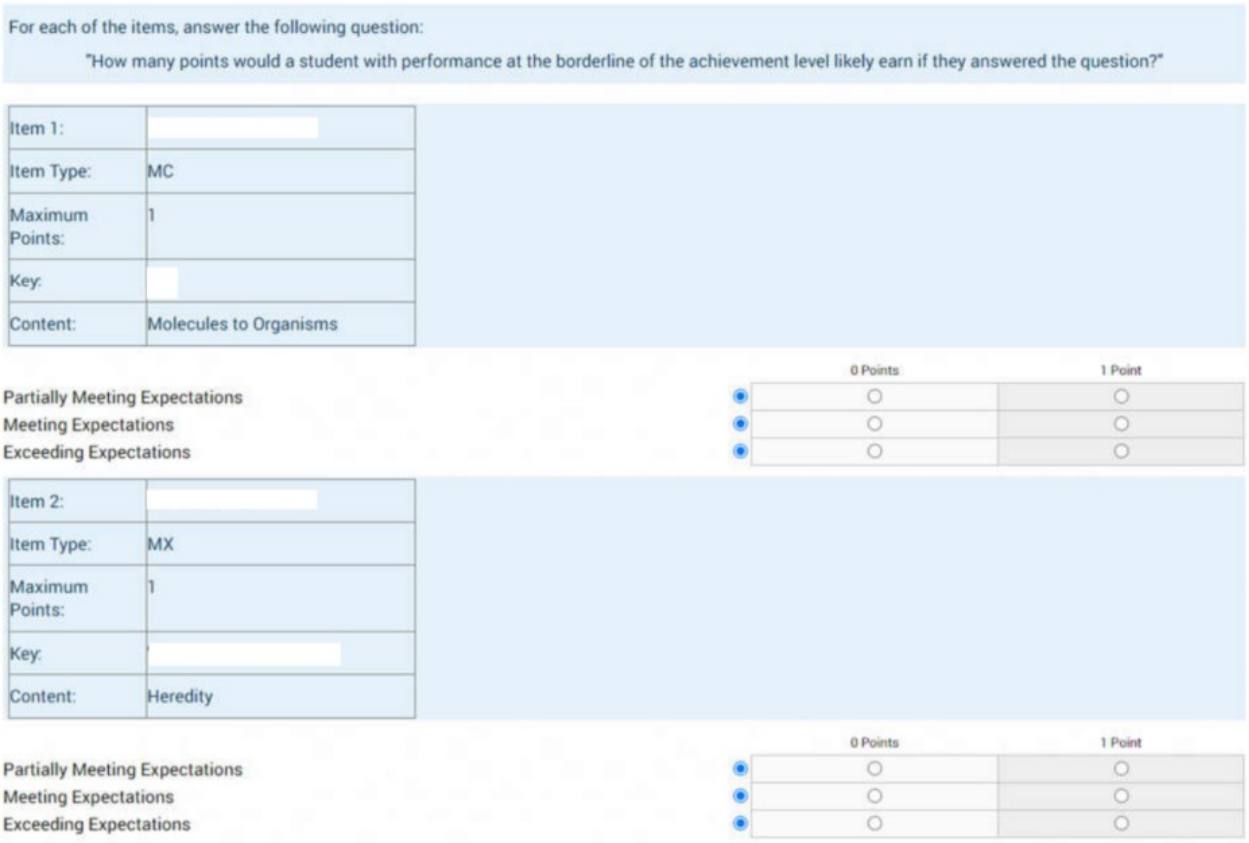
"How many points would a borderline student at the [specific] achievement level likely earn if he or she answered the question?"

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

1. "would…" — When envisioning 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.
2. "...a student performing at the borderline of the [specific] performance level…" — The panelists were reminded to reference the borderline descriptions to determine how a student performing at the borderline of that performance level would be expected to respond.
3. "...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 three students at the borderline of a performance level. If a panelist believed 2 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 of 3 students with performance at the borderline would correctly answer the item, they would respond "no" to the question.
4. "...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.

The response to judgment question for each item was recorded within the judgment survey within the website. Figure 2 presents an example item judgment survey within the website. Participants completed the item judgments for each achievement level for an item before moving on to the next item.

Figure 2: Example Item Judgment Survey from Pearson Standard Setting Website



The participants also kept a record of their item judgments on the Judgment Record Sheet. This document was provided to them as part of the materials in their folder. It included the unique item number, reporting category, and maximum possible points for the item. The participants were shown how to use the unique item number to ensure that they were referencing the correct item on all documents within the judgment survey and in the online system.

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

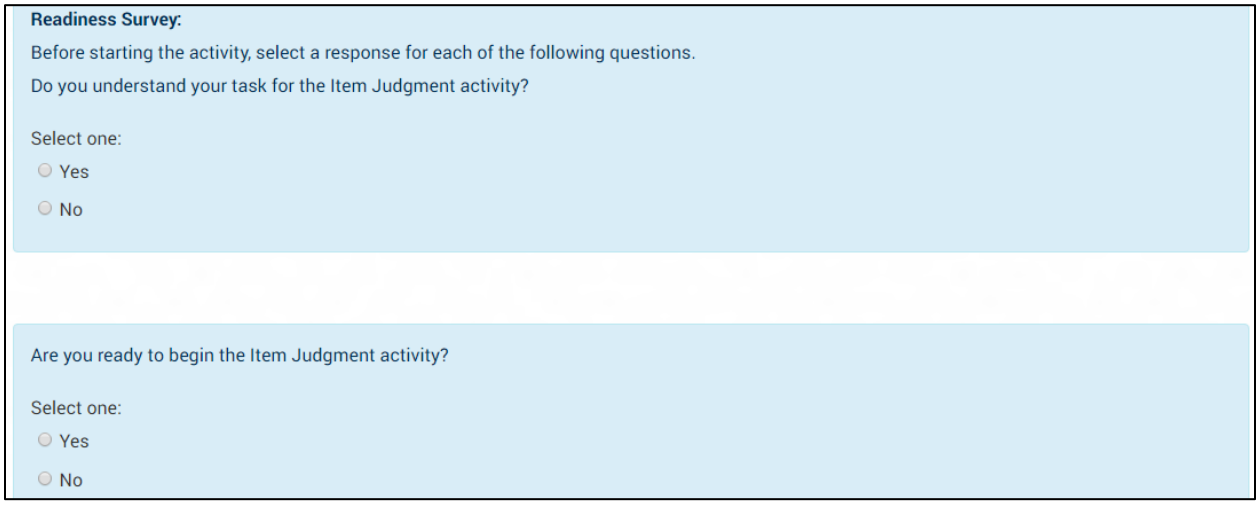
1. Give panelists experience reviewing and making judgments about different types of items.
2. Familiarize panelists with the judgment survey on the standard setting website.
3. Build confidence in panelists' understanding of the task to be completed.

The practice items selected for the activity 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.

Item Judgment Rounds. After receiving training on the standard setting process, the participants participated in three rounds of judgments. Before starting each of the three

judgment rounds, the participants were required to complete a readiness survey in the website indicating that they understood the task and process used to complete the item judgments. The participants had to answer "yes" to all readiness survey questions before continuing with the judgment round. If they responded "no" to any question, they were asked to notify a facilitator for additional assistance. Figure 3 presents an example of the readiness quiz participants needed to complete before starting the judgment task.

Figure 3: Example Readiness Quiz Before Judgment Task



Each judgment round consisted of a review of the judgment process by the process facilitator, with explicit instruction on which materials would be needed to complete the task, followed by participants working independently on their item judgments. Participants were required by the website to provide judgments for each item before they could submit their judgment survey.

Judgment Feedback. Once all the participants had completed their item judgments, data analysts from Pearson collected the data from the website and performed the analysis to determine an aggregate recommendation for the committee. The participants were provided feedback after each judgment round which could be used to inform subsequent judgments. Feedback data included the following:

1. Individual item judgment record: A record of each panelists' individual item judgments for each achievement level. This was provided for the panelists to check their individual judgments against what was recorded in the website survey.
2. Information about panelists' cut scores for each achievement level:
    1. 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 achievement level.
    2. Committee cut score recommendations and statistics: Committee-level recommendations were the median cut score across all panelists for each achievement level. Panelists were provided the committee-level cut score recommendations and cut score statistics for each achievement level.
    3. Panelist agreement data: Bar graphs showing the frequency of individual recommended cut scores for each achievement level and across adjacent achievement levels.

3. Item-level judgment agreement across panelists: Distribution of panelist judgments for each item and achievement level.
4. Item means (p-values) and score-point distributions: The average score earned by students for each item and the distribution of score points, for polytomously scored items, calculated from operational test data.
5. Impact data: Percentage of students that would be classified into each achievement level, based on the committee's current recommended cut scores and the results of students who took the assessment during the spring 2022 administration.

Table 6 displays the type of feedback that was provided to participants after each round of judgments.

Table 6: Feedback Data Provided to Participants After Each Judgment Round

	Round		
Feedback	**1**	**2**	**3**
Individual item-level judgment record	Yes	Yes	Yes
Individual test-level recommendations	Yes	Yes	No
Table test-level recommendations	Yes	Yes	No
Committee test-level recommendations	Yes	Yes	Yes
Item-level participant agreement	Yes	Yes	No
Test-level participant agreement	Yes	Yes	No
Item score mean and score distribution	Yes	Yes	No
Impact data	Yes	Yes	Yes

Appendix F 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 achievement 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 the 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. Since the round 3 judgments were the participants' final judgments, the feedback data was provided to facilitate the participants' evaluation of the final recommendation by the committee.

Process Evaluations. The validity of standard setting outcomes relies partially on the procedural validity of the meeting. Evidence of the procedural validity was gathered through evaluation surveys administered during the standard setting. An evaluation survey was administered within the website in each committee after the practice judgment activity and the after round 3 judgments. The purpose of these surveys was to collect information about each participants' experience in recommending cut scores for the achievement levels associated with

the MCAS assessments. The survey asked participants to provide feedback on the following:

1. The level of success of the various components of the meeting
2. The usefulness of the activities conducted during the meeting
3. The adequacy of the various components of the meeting
4. The adequacy of opportunities to ask questions, etc., at the meeting
5. How confident participants were that the recommended cut scores accurately reflected student performance at each achievement level
6. Whether committee members thought that their judgments and opinions were treated with respect by facilitators and fellow participants

All participants 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 MCAS Cut Scores from Standard Setting Committees

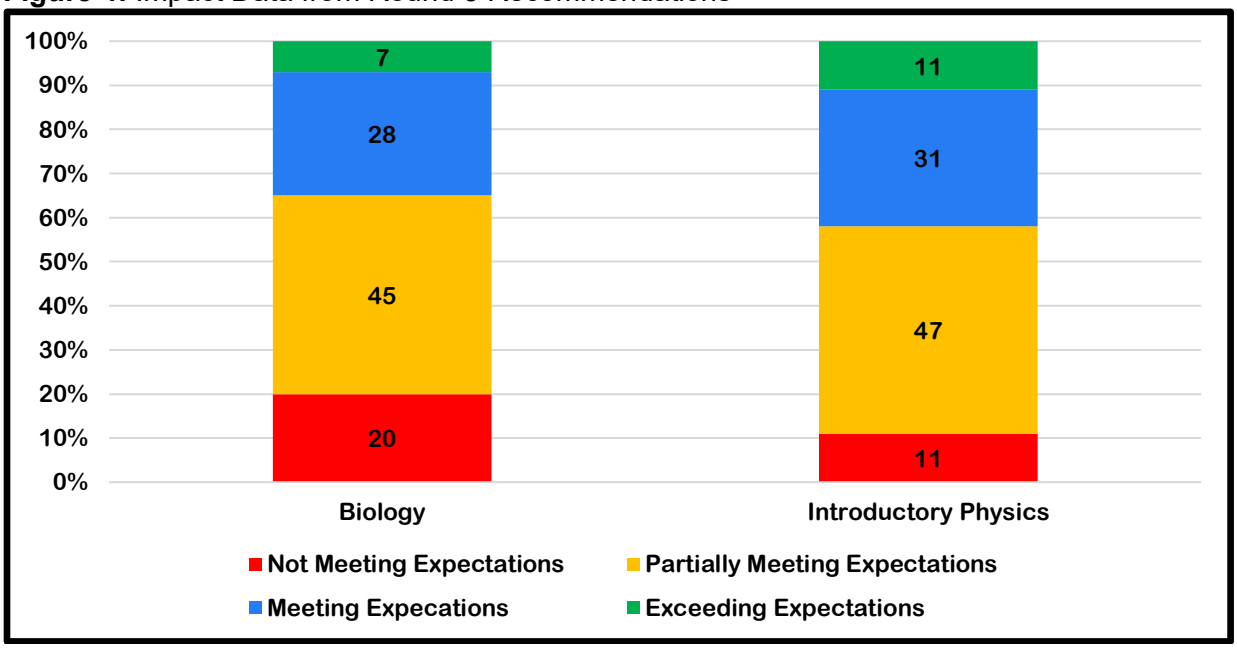
During the cut score setting meeting, it was expected that there would be variation between participants' cut score recommendations for each achievement level. To determine a single cut score recommendation for an achievement level for a committee, the cut score recommendations for the achievement level were averaged across participants. Specifically, the median cut score from a set of participants' cut score recommendations was used to determine the recommended cut score for an achievement level for the committee. The recommendation resulting from the round 3 judgments was considered as the committee's recommendation for the standard setting meeting. Table 7 displays the recommended cut scores for each achievement level based on the round 3 recommendations for each course and subject. Figures 6 and 7 display the impact data for Biology and Introductory Physics, respectively, based on the recommended cuts scores from round 3 from each committee.

The recommended cut scores for each achievement level from the three judgment rounds for each standard setting committee, represented as raw scores, are presented in Appendix G. The summary statistics for the recommended cut scores for each achievement level from the three judgment rounds for each standard setting committee are shown in Appendix H. The participant agreement data for each performance level for judgment rounds 1 and 2 for each standard setting meeting are shown in Appendix I. The estimated impact data after judgment round 3 for each achievement level for each standard setting committee are shown in Appendix J.

Table 7: Cut Score Recommendations from Standard Setting Committees

Subject	Maximum Score	Partially Meeting Expectations		Meeting Expectations		Exceeding Expectations	
		Raw Score	% Correct	Raw Score	% Correct	Raw Score	% Correct
Biology	60	17	28.33	37	61.67	52	86.67
Introductory Physics	60	16	26.67	37	61.67	51	85

Figure 4: Impact Data from Round 3 Recommendations



Chapter 4 – Post-Standard Setting

This chapter provides details about the work completed after the standard setting committee meetings. The sections of this chapter include:

- Articulation process
- Linear scaling process
- Competency Determination Validation

Articulation Process

The purpose of the articulation meeting was to review the cut score recommendations from the standard setting committees within a content area and evaluate the reasonableness of the recommendation. Where the recommendations from the standard setting committees were made with a specific focus on the respective content for this committee, the focus of the articulation committee was to view the cut score recommendations across the courses, Biology and Introductory Physics, and within a content area, including grade 8 STE, to evaluate whether the recommendation resulted in a cohesive assessment system. The participants of the articulation were guided through a specific process where they would review the recommendations from the standard setting committee and, if necessary, recommend and review changes to the recommendation, resulting in a set of recommended cut scores from the vertical articulation committee.

For the Biology and Introductory Physics committees, the vertical articulation occurred with half of the committee participants after the round 3 judgment recommendations. The lead facilitator, Eric Moyer, Ph.D., was the facilitator for the articulation meeting.

Meeting Process

The articulation process involved the following steps:

1. ALD cross-subject and grade review activity
2. Review and discussion of the cross-subject impact data
3. Review and discussion of the cross-grade impact data
4. Review and recommendation to recommended cut scores

At the beginning of the process, the participants were instructed to the purpose of the articulation process, as the opportunity to review the recommended cut scores from the standard setting meetings across courses within the same subject, ensuring that they represented a cohesive assessment system. In the previous standard setting meetings, they were focused primarily on the content related to the grade within their committees, where in this meeting they would review the recommendation from across grades from a policy perspective.

To start the vertical articulation process, the participants were provided the opportunity to independently review the ALDs across courses. The instructions for this activity were to look for differences or similarities in student expectation across courses that could be used to explain the articulation of student impact across grades. After looking at the ALDs independently, the participants had the opportunity to discuss the ALDs as table groups. During a whole group discussion, the participants discussed what their expectation would be of the articulation of the impact data across courses. The focus of this discussion was to establish a content-based expectation for the impact data across grades.

The participants were then presented with the cross-grade impact data chart reflecting the results from the round 3 judgments of all standard setting committees for Biology and Introductory Physics. Additionally, the panelists were presented with impact data using the results from the matched sample process between Physics and Introductory Physics. The panelists were instructed that the matched sample process was performed to statistical remove differences between two populations. For Biology and Introductory Physics, the participants were presented with the impact data from round 3 along with the final impact data for grade 8 for STE from the 2019 standard setting meeting. The groups had the opportunity to discuss how the results looked across grades based on their initial expectations.

Based on their expectations of student impact relative to their review of the ALDs, the participants were provided the opportunity to investigate changes to the recommended cut scores from round 3 using an interactive spreadsheet, which was accessed through the standard setting website.

The interactive spreadsheet allowed participants to investigate possible changes to the cut scores from their committee by adjusting the current cut scores and simultaneously viewing the change to the impact data. The participants were instructed to investigate changes to the recommended cuts scores if they felt that the pattern of the impact data across grades was inconsistent with what they expected, based on their review of the ALDs and their understanding of a cohesive assessment system. The changes would be made directly at the cut score level and did not involve changes to the item level judgments. The range of individual participant's cut score recommendations from round 3 were used as a guide when evaluating how much change would be reasonable to make. The participants were aware of the need to honor the work the standard setting committees had done and were judicious in making changes.

The committee had the opportunity to recommend changes to cut scores for achievement levels for the grades which they determined had inconsistent results. When a change in cut score was recommended, it was entered into a master interactive spreadsheet by the meeting facilitator for the entire committee to view the change in cut score and pattern of impact data across grades and achievement levels. One recommended change at a time was viewed, discussed, and then either accepted or rejected by the vertical articulation committee. This process was repeated until all recommended changes were discussed and the vertical articulation committee agreed with the entire set of cut score recommendation across all grades.

Participants were aware of the need to honor the work the standard setting committees had done and were selective in making changes so that the number and magnitude of changes were limited to only those changes necessary to support the articulation across grades. Table 8 displays the changes made to the recommended cut scores from the standard setting committees.

Table 8: Changes to the Cut Score Recommendations by the Vertical Articulation Process

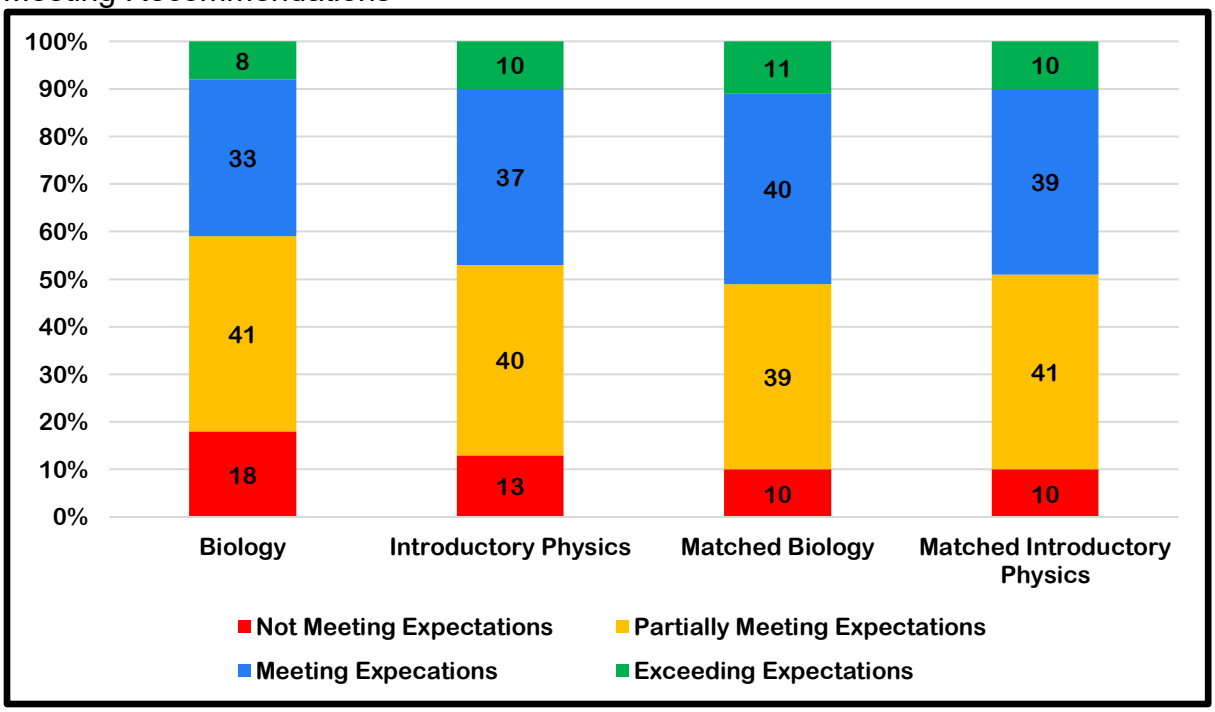
Course	Partially Meeting Expectations	Meeting Expectations	Exceeding Expectations
Biology	-1	-3	-1
Introductory Physics	+1	-2	0

Table 9 displays the recommended cut scores for each achievement level based on the final vertical articulation recommendations for each course and subject. Figure 5 displays the impact data for Biology and Physics, based on the recommended cuts scores from the vertical articulation process.

Table 9: Cut Score Recommendations from the Vertical Articulation Process

Subject	Maximum Score	Partially Meeting Expectations		Meeting Expectations		Exceeding Expectations	
		Raw Score	% Correct	Raw Score	% Correct	Raw Score	% Correct
Biology	60	16	26.67	34	56.67	51	85
Introductory Physics	60	17	28.33	35	58.33	52	86.67

Figure 5: Impact Data for Biology and Introductory Physics based on the Articulation Meeting Recommendations



Process Evaluation Survey

At the end of the vertical articulation process, participants were asked to complete a process evaluation survey within the website. The purpose of the evaluation was to collect information about each participants' experience in the vertical articulation meeting. The evaluation asked participants to provide feedback on the following:

- The level of success of the various component of the meeting

- The usefulness of the activities conducted during the meeting
- The adequacy of the various components of the meeting
- The level of support the participants had in setting the recommended cut scores for each achievement level across all grades

All participants were also allowed to provide any additional information concerning their evaluation of the process of the vertical articulation meeting through an open response question.

Linear Scaling Process

The recommendations from the standard setting and vertical articulation committees were cut scores in terms of raw scores on the test. Student results are not reported as raw scores, since the overall difficulty of tests may change from year to year, so results would not be able to be compared across years. To address this, student results on the MCAS are reported using scale scores, which are comparable across administration years. After the vertical articulation process, a method was implemented to determine the process for transforming the raw scores from the spring 2022 administration to MCAS scale scores.

The process of determining the rules for transforming the raw scores to the final MCAS reporting scale was guided by several principles identified by DESE:

- Respect the cut score recommendations provided by the vertical articulation committee by preserving the final cut scores while also establishing a coherent system of measurement across grades
- The impact data from the final scaling solution should reflect a coherent assessment system across the grades
- The reporting MCAS scaled scores for the three achievement level cuts should be the same across grades and tests
- The scaling solution should involve a single linear transformation, from the underlying IRT scale to the reporting MCAS scale
- The reporting MCAS scaled score range should be the same across grades and tests.

This process, involving Pearson, Cognia, and DESE, was used to determine a final reporting scale and transformation rules for each test. A more extensive description of the development of the scaling process will be included in the overall MCAS technical report.

The following iterative process was used to determine the final cut scores for the achievement levels for the MCAS assessments, starting with the raw score cuts recommended from the vertical articulation meeting:

- The raw score cuts for the three achievement levels were translated to cuts on the IRT scale using the raw score to theta (IRT) lookup table for the specific assessment.
- The cuts on the IRT scale were adjusted so that the differences between consecutive cuts were the same, allowing for the use of a single linear transformation rule.
- Based on the adjusted IRT cut scores, scaling constants for the linear transformation from the IRT cuts to MCAS scale score cuts were determined.
- Using the scaling constants, lookup tables for each grade and test were created, displaying the relationship between the raw scores and reporting MCAS scaled scores.
- Based on the lookup tables, adjusted raw score cuts for each achievement level were determined.
- Finally, the resulting impact data based on the adjusted raw score cuts was calculated

and reviewed to ensure a coherent system across grades.

This process was repeated several times until a final scaling solution was determined, which met, as closely as possible, DESE's requirements.

The recommended reporting scale ranges from a lowest obtainable scale score of 440 to a highest obtainable scale score of 560. In order to create common points of reference across the assessments, the same scaled score cuts for each achievement level were defined, with a *Partially Meeting Expectations* cut of 470, a *Meeting Expectations* cut of 500, and an *Exceeding Expectations* cut of 530. While the cut scores were defined with the same scaled scores between the two tests, they are not identical, and direct comparisons through averaging and aggregation across grades should not be made without study and/or statistical adjustments. The scaled scores and distributions of students resulting from the cuts set for biology and introductory physics were not designed for direct comparison. Table 10 displays the changes made to the recommended cut scores from the vertical articulation committees.

Table 10: Changes to the Cut Score Recommendations for Linear Scaling

Subject	Partially Meeting Expectations	Meeting Expectations	Exceeding Expectations
Biology	+1	0	-1
Introductory Physics	0	0	-1

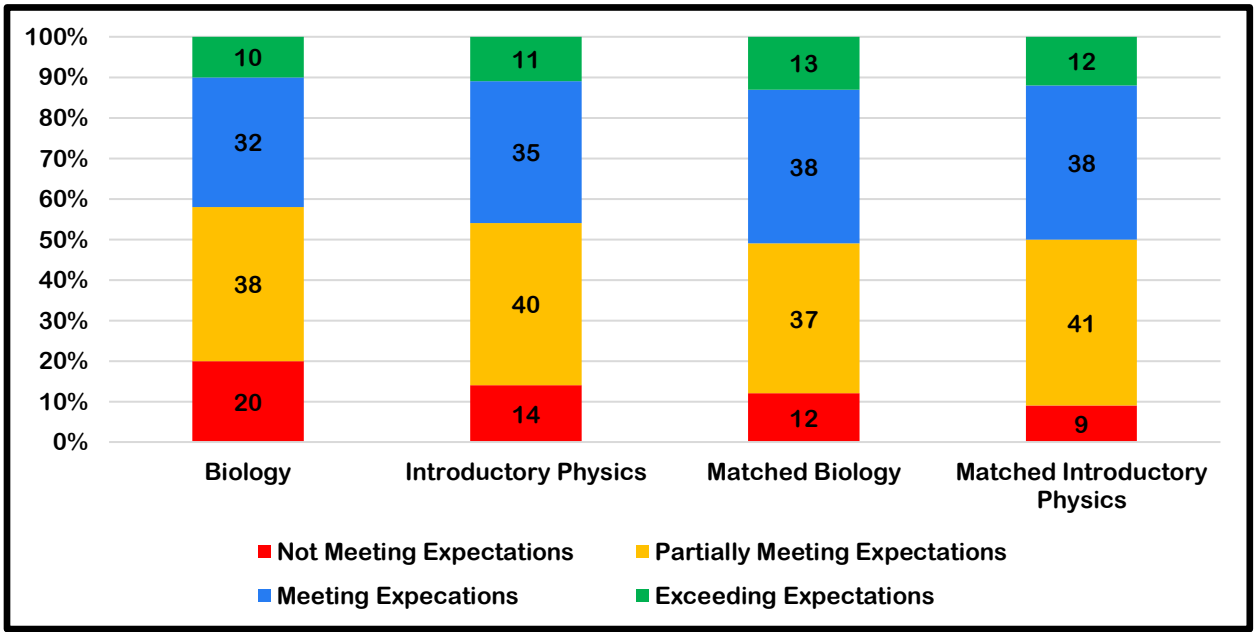
Table 11 displays the final recommended from the Linear Smoothing for each achievement level based on the results of this process for each course and subject.

Table 11: Final Cut Score Recommendations from the Linear Scaling

Subject	Maximum Score	Partially Meeting Expectations		Meeting Expectations		Exceeding Expectations	
		Raw Score	% Correct	Raw Score	% Correct	Raw Score	% Correct
Biology	60	17	28.3%	34	56.7%	50	83.3%
Introductory Physics	60	17	28.3%	35	58.3%	51	85.0%

Figure 6 presents the impact data from the final recommendations as stacked bar graphs.

Figure 6. Impact Data for Biology and Introductory Physics based on Final Recommendations



Competency Determination Validation

A competency determination validation meeting was convened to review and either validate or adjust competency determination cuts on the next-generation MCAS assessments for Biology and Introductory Physics. The competency determination cuts on the next-generation MCAS are interim cut scores that correspond to the scale score cuts for each of the achievement levels on the previous (legacy) MCAS assessments for Biology and Introductory Physics. The identification and validation of the interim competency determination cuts was legislatively mandated to provide students, parents, and educators with sufficient time to become familiar with the new assessment and expectations before the next-generation passing standards are established.

In addition to the Competency Determination (220) score, the legacy cut scores and scaled scores will be used during the transition period to determine eligibility for John and Abigail Adams Scholarship.

Prior to the competency determination meetings, the Pearson standard setting team worked to statistically identify interim cuts for the achievement levels on the legacy MCAS assessments, Needs Improvement (220), Proficient (240), and Advanced (260). An equipercentile process with a matched sample was used to statistically identify the interim cut scores for each achievement level. The statistically determined interim cut score ranges were established so they result in similar impact data on the spring 2019 administration of the next-generation MCAS and on the spring 2019 administration of the legacy MCAS. Table 12 provides the statistically defined interim cut score ranges for each subject.

Table 12: Statistically Defined Interim Cut Scores

Subject	Needs Improvement	Proficient	Advanced
Biology	15 to 19	28 to 33	44 to 50
Introductory Physics	15 to 19	25 to 32	45 to 50

There were two competency determination validation meetings, one for Biology and one for Introductory Physics. The competency determination committees were convened as a separate meeting after the standard setting committee concluded. The panelists for the competency determination meetings were a subset of the panelists from the standard setting committee, including some of the committee table leaders. The facilitators for the Biology and the Introductory Physics competency determination meeting were Soo Ingrisone and Scott Strickman, respectively.

Meeting Process

The competency determination validation process involved three steps:

1. Determine content expectations for previous (legacy) MCAS achievement levels
2. Review student performance on next-generation MCAS within reasonable ranges for the interim cut scores
3. Provide individual judgments about interim cut scores for each achievement level

For the panelists to review whether the content expectations defined by the interim cut scores on the next-generation MCAS was similar to the expectation on the legacy MCAS assessment, they first had to define the content expectations for each achievement level. Prior to the meeting, the Pearson standard setting team created score profiles for each scale score associated with the achievement level cut scores. The score profile presented student performance on each of the items for students that received the associated scale score. Figure 7 displays an example of a score profile for grade 10 ELA.

Participants reviewed the score profiles for each achievement level on the legacy MCAS in table groups to create an outline of student expectations for each achievement level. For each item on the score profile, the participants were provided item keys and scoring information, accessed through the standard setting website. Based on the panelist review of the items and the score profiles, the facilitator guided the group through a discussion to develop an outline of student expectations for each achievement level.

The panelists then reviewed score profiles for each interim cut score range on the next-generation MCAS assessment. The score profiles were based on student performance on the next-generation MCAS assessment administered in spring 2022. Access to the score profiles for the interim cut score ranges, items and scoring information for items was provided to the panelists through the standard setting website. For each score profile, the participants were comparing the expectations defined by student performance on the items and how they compared to the content expectations defined for the legacy achievement level.

Figure 7: Example student profile for legacy assessment

Scale Score: 220

Raw Score Distribution:

Raw score	21	22
Number of students	183	162
Percentile	2nd	2nd

Composition Score Distribution

Score	0	2	3	4	5	6	7	8	9	10	11	12
Topic Dev.	39%	7%	4%	31%	8%	11%	0%	0%	0%	0%	0%	0%
Conv.	39%	1%	2%	45%	6%	6%	0%	0%				

Open Repsonse Score Distribution

Item Sequence	Item ID	Max Points	0	1	2	3	4
9	EL315010	4	26%	50%	22%	3%	0%
18	EL316592	4	21%	50%	23%	5%	0%
27	EL309870	4	35%	51%	13%	1%	0%
36	EL315350	4	18%	51%	27%	4%	0%

Selected Response Score Distribution:

Item Sequence	Item ID	Max Points	% correct
Percentage of students with correct response greater than 75%			
Percentage of students with correct response between 60% and 74%			
Percentage of students with correct response between 40% and 59%			
1	EL314958	1	40%
3	EL314971	1	45%
4	EL314992	1	56%
5	EL314981	1	40%
6	EL314984	1	44%
10	EL316630	1	41%
11	EL316629	1	57%
12	EL316631	1	47%
13	EL316636	1	47%

Based on the panelists' review of the score profiles associated with each interim cut scores and the scores around them, the panelists then provided an individual judgment for each achievement level. For each achievement level, the panelist responded to the following question:

Based on your review, which raw score within the range on the Next-Generation MCAS represents similar expectations to the performance level expectations on the spring 2019 Legacy MCAS?

Panelists selected a raw score within the reasonable range for each achievement level through an online judgment survey. The median of the committee recommendations was used as the committee recommendation for the achievement level. Table 13 displays the interim cut score recommendations for the legacy achievement levels on the next-generation MCAS.

Table 13. Recommended Cut Scores for the Legacy Achievement Levels

Subject	Legacy Achievement Levels		
	Needs Improvement	Proficient	Advanced
Biology	16	29	46
Introductory Physics	17	29	47

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

This chapter details various evidence for the validity of process used during the standard setting meetings. The sections in this chapter include the following:

- Committee representation
- Committee training
- Participants' perceived validity of the meeting
- Technical advisors' perceived validity of the meeting

Committee Representation

As part of the standard setting evaluation, participants completed a demographic survey that collected information about their background relevant to educational experience. The results of the self-reported demographic characteristics of the participants are documented in Appendix D.

As part of the survey, their current position (Table D.1) and the number of years teaching a course related to their standard setting meeting (Table D.2). A majority of the participants of each committee were teachers in grades K–12. The majority of panelists in both committees had more than 10 years of experience.

The experience of the teachers in the committees included experience teaching different populations of students, as displayed in Table D.3 A large majority of participants of each committee had experience teaching general education, mainstream special education, and English language learners.

A large majority of participants were currently working in school districts, as presented in Table D.7. The participants that worked within school districts represented the various types of districts across the state, including size, type, and socioeconomic status. Teachers representing schools from a rural area were the least represented, with only one teacher in the Biology committee from a rural school. The set of participants for this standard setting was well selected for representing the teachers across the state in this process, which was noticed consistently by the facilitators of the meeting.

Committee Training

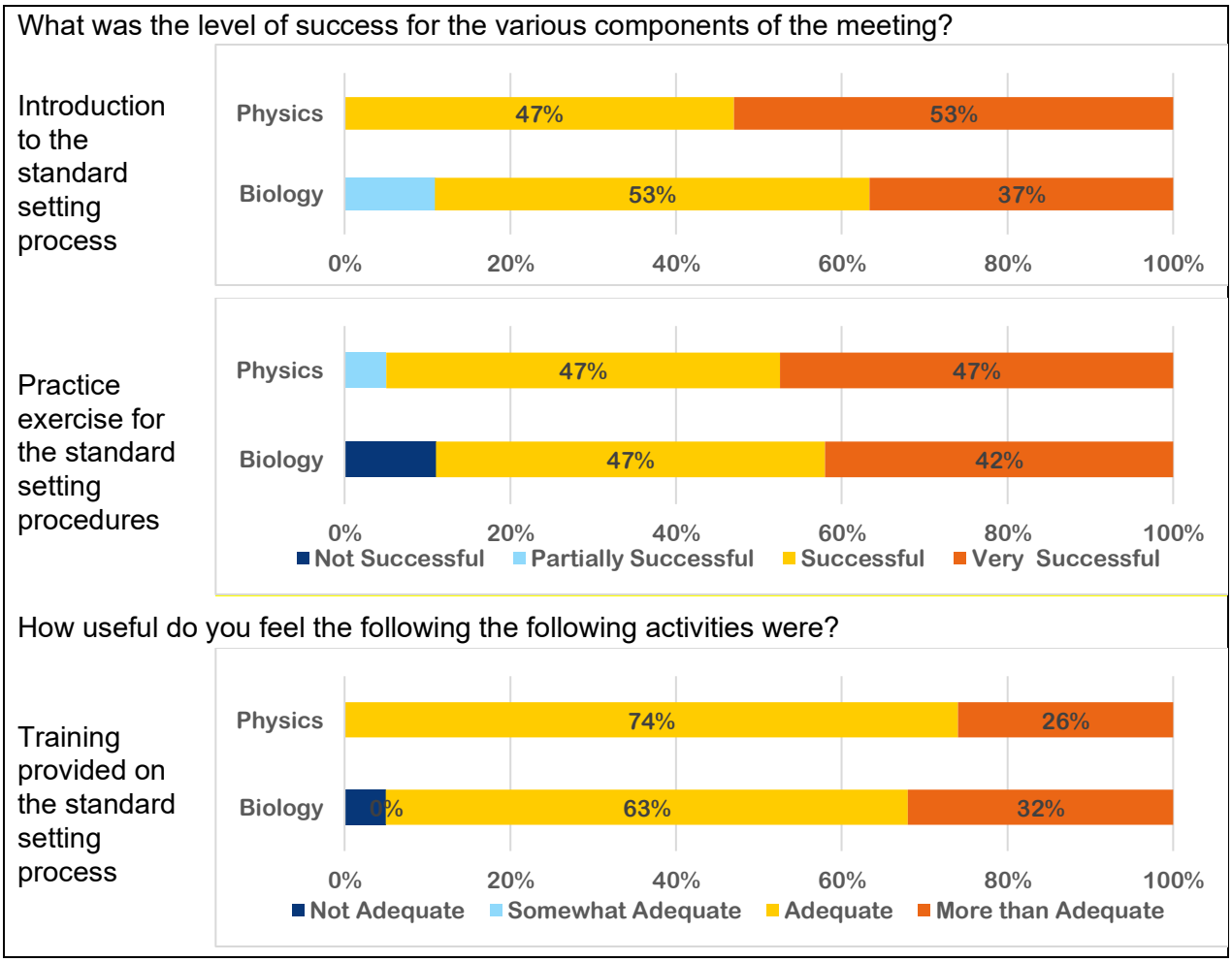
During the cut score setting meeting, it was essential that participants understood how to make judgments as part of the Extended Modified (Yes/No) Angoff standard setting methodology. The training on the standard setting methodology was provided during the general session and in the individual standard setting committees. The training on the implementation of the standard setting process was standardized across committees through the PowerPoint training slides.

Participants participated in a practice judgment round as an opportunity to implement the standard setting methodology without consequence, including making judgments within the standard setting website. During the practice judgment round, the participants reviewed a reduced set of items and provided judgments for the three achievement levels, *Partially Meeting Expectation*, *Meeting Expectations*, and *Exceeding Expectations*. After the practice round, the process facilitator led a whole-group discussion to identify and respond to any questions or issues participants encountered while implementing the standard setting process. Before each

judgement round, participants responded to a readiness survey that asked whether participants were prepared for making their judgments. Participants were not able to continue to the judgment survey unless they answered yes to both questions on the readiness survey. They were encouraged to ask the facilitator questions if they responded "no" to either question.

At various points within the standard setting meeting, participants completed a process evaluation survey to record their impressions of the effectiveness of the materials and methods employed throughout the process. Figure 8 displays the results of the evaluation survey across subject-level committees for several questions related to the training on the standard setting process. The results of these process evaluations for each individual committee are presented in Appendix K.

Figure 8: Evaluation results on standard setting process training activities

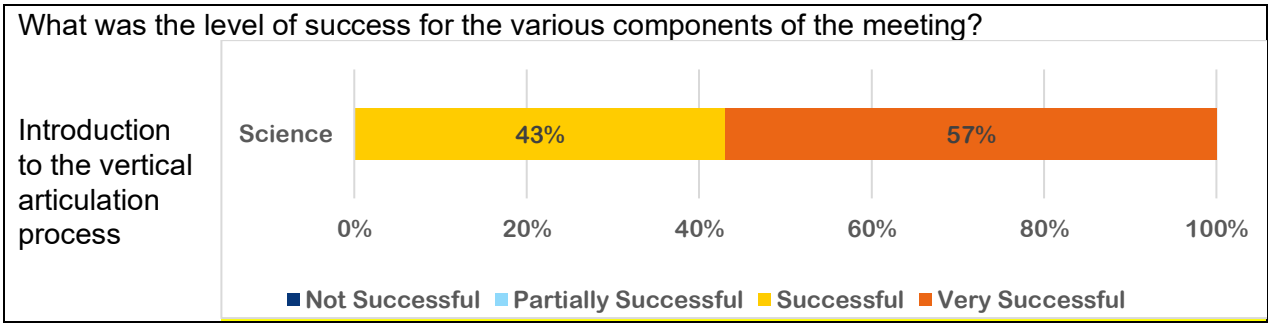


As part of the evaluation survey, the participants were specifically asked about the effectiveness of the training they received on the standard setting process. One question asked participants to rate the level of success of the initial introduction to the standard setting process during the general session. Overall, the initial introduction to the standard setting process was perceived as successful with at least 90 percent of participants in the committees responding that it was either *Successful* or *Very Successful*. The perception of the training on the standard setting process in the breakout groups was also good, where more than 90 percent of participants in the committees responded that it was either *Useful* or *Very Useful*. More than 80 percent of

participants in the committees indicated that the practice judgment activity for the standard setting process was either *Successful* or *Very Successful*. These responses indicate that, overall, most participants believed that the training provided prepared them to implement the standard setting procedure, providing cut score recommendations for each assessment for which they were responsible.

During the vertical articulation meeting for Biology and Introductory Physics, the participants were provided training on the process and tools used during the meeting. At the end of the meeting, the participant completed a process evaluation form to record their opinion on the training provided. The results of this process evaluation are presented in Appendix K. All participants indicated that the introduction to the vertical articulation process was either *Successful* or *Very Successful.*

Figure 9: Evaluation results on vertical articulation process training activities



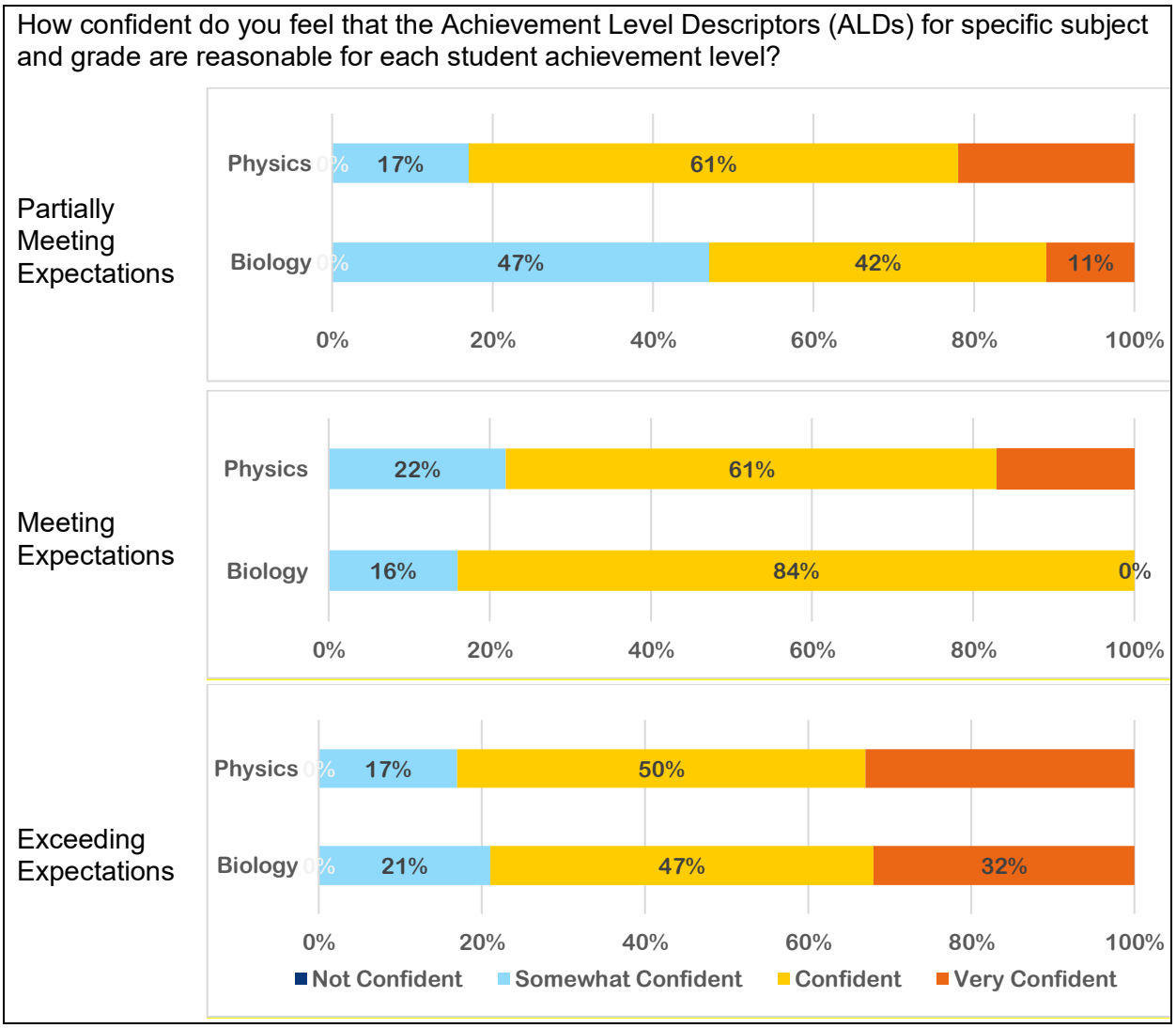
Perceived Validity of the Workshop

Participants and reviewers communicated their perceived validity of the workshop and the recommended cut scores. Participants indicated their perceived validity of the workshop as part of the workshop evaluation. Evaluations are important evidence for establishing the validity of recommended cut scores for the achievement levels.

Participant Evaluations

Generally, the participants were satisfied with their recommendations and with the workshop as a whole. As part of the process evaluation from each committee, the participants had the opportunity to indicate their confidence that the Achievement Level Descriptors were reasonable for each of the achievement levels. Figure 10 displays the results of the evaluation survey across subject-level committees and indicates that the ALDs were reasonable for each of the achievement levels. The results for each subject and grade are presented in Appendix K.

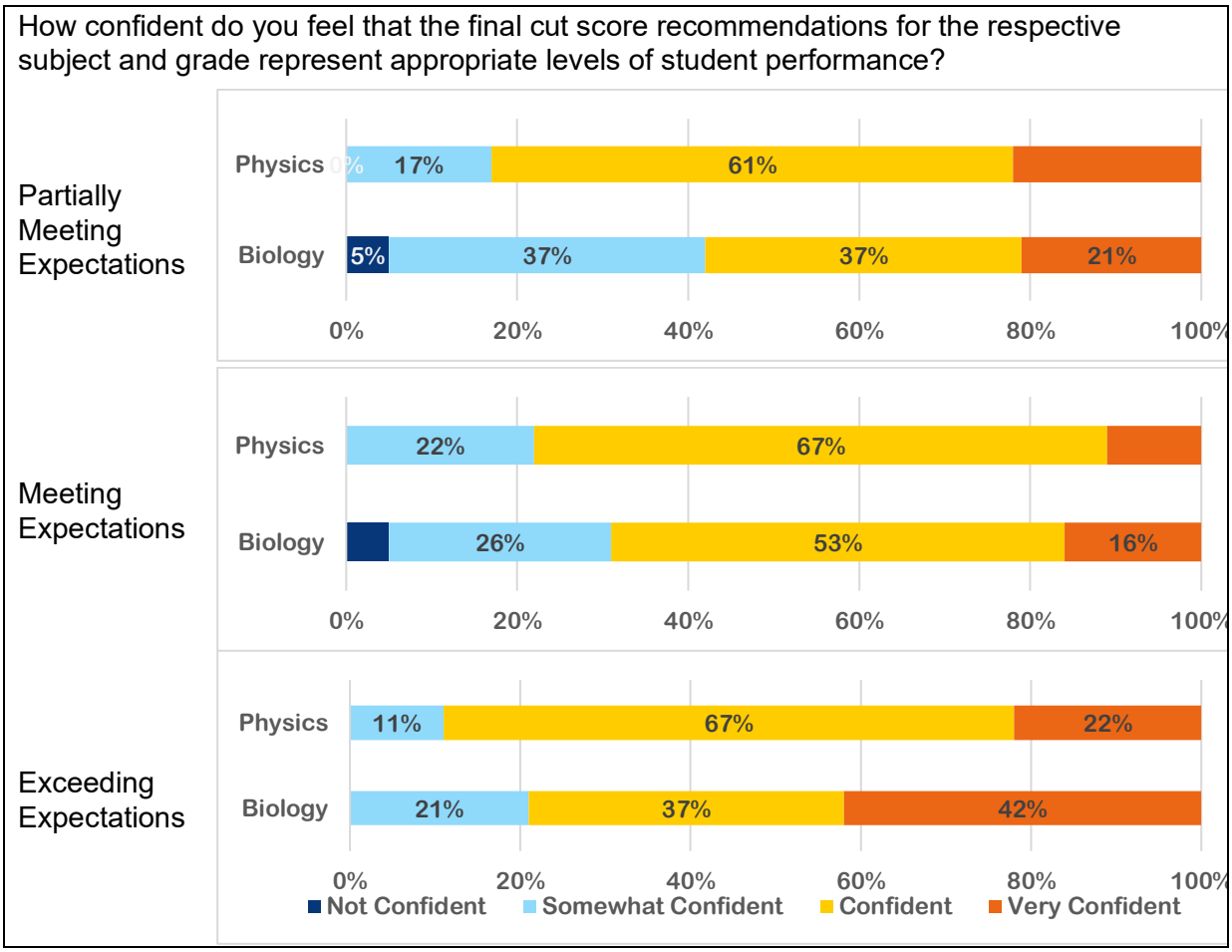
Figure 10: Evaluation results on reasonableness of the ALDs for each achievement level



Overall, the majority of panelists had at least some confidence that the ALDs were reasonable for each of the achievement levels. In the majority of committees, at least 50 percent of the participants were *Confident* or *Very Confident* that the ALDs were reasonable for the achievement levels. The panelists from the physics committee had the greatest level of confidence, with greater than 75% of the panelists indicating they were *Confident* or *Very Confident*. These responses provide evidence that, overall, the ALDs, a foundation for the standard setting process, were perceived by the participants as providing reasonable expectations for each achievement level.

The participants were also provided the opportunity to indicate their confidence in the cut scores recommended by the standard setting committees. Figure 11 displays the results of the evaluation survey across committees for their confidence in the recommended cut scores. The results for each science course are presented in Appendix K.

Figure 11: Evaluation results on reasonableness of the cut scores for each achievement level



As with the ALDs, the majority of participants indicated that they had at least some confidence that the recommended cut scores represented appropriated levels of student performance for each achievement level. There seemed to be a difference between the level of confidence in the cut score recommendations for the different subjects. The physics participants demonstrated a greater confidence in the cut score recommendations, with greater than 60 percent of panelists selecting *Confident* or *Very Confident* for all achievement levels. Although the biology panelists indicated lower confidence, at least 50 percent of panelists indicated *Confident* or *Very Confident* for Meeting Expectations and Exceeding Expectations.

Overall, this feedback from the cut score setting participants provides evidence for the validity of the cut score recommendations for each of the achievement levels from the standard setting committee.

The participants in the vertical articulation meetings were also provided the opportunity to provide their opinion concerning the cut score recommendations for each achievement level resulting from the vertical articulation process. Based on the results, shown in Appendix K, the large majority of participants, at least 90 percent of panelists from the science vertical articulation committee, indicated that they were *Very Confident* or *Confident* of the cut score recommendations from the vertical articulation process. These results provide further evidence for the validity of the process and the results used to create the cut scores for achievement levels for each assessment.

Technical Reviewer Evaluations

After the standard setting meeting, a technical advisor, William Lorie, Ph.D., provided a written review of the standard setting process used during the meetings.

References

Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice*, *23*(4), 31-50.

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.

Hambleton, R. K., & Plake, B. S. (1995). Extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education*, *8*, 41-56.

Iacus, S. M., King, G., & Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, *20*(1), 1-24. doi: 10.1093/pan/mpr013.

Plake, B. S., & Cizek, G. J. (2012). Variations on a Theme. In G.J. Cizek (Ed.), *Setting performance standards* (2nd ed.). New York: Routledge.

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 AA – Achievement Level Descriptors

Biology

Student results on the MCAS tests are reported according to four achievement levels: *Exceeding Expectations, Meeting Expectations, Partially Meeting Expectations,* and *Not Meeting Expectations.* The descriptors below illustrate the knowledge and skills students demonstrate on MCAS at each level. **Knowledge and skills are cumulative at each level.** No descriptors are provided for the *Not Meeting Expectations* achievement level because students work at this level, by definition, does not meet the criteria of the *Partially Meeting Expectations* level.

	Partially Meeting Expectations *On MCAS, a student at this level:*	**Meeting Expectations** *On MCAS, a student at this level:*	**Exceeding Expectations** *On MCAS, a student at this level:*
Understanding and Application of Disciplinary Core Ideas	Demonstrates a partial understanding of some scientific concepts and processes by identifying and sometimes describing or providing evidence for these concepts and processes. Uses some basic scientific terms in common scientific examples.	Demonstrates a solid understanding of many scientific concepts and processes by mostly describing, explaining, and providing evidence for these concepts and processes. Mostly applies appropriate scientific terms in a variety of applications, including common science examples and some novel situations.	Demonstrates a comprehensive, in-depth understanding of many scientific concepts and processes by consistently describing, explaining, and providing evidence for these concepts and processes. Consistently applies scientific terms in appropriate contexts in both common science examples and many novel situations.
Understanding and Application of Scientific and Engineering Practices	Identifies a testable, scientific question for an investigation. Completes a simple, commonly used model. Uses simple graphs or data to draw general conclusions about a familiar scientific investigation or phenomena. Identifies evidence to support a claim.	Develops some testable, scientific questions for an investigation. Completes or uses a model and describes some strengths and weaknesses of the model. Analyzes multiple sources of data, including graphs and tables, to draw conclusions about a familiar scientific investigation or phenomena. Provides some evidence to support a claim and constructs basic explanations for	Consistently develops testable, scientific questions for an investigation. Creates a model, consistently describes the strengths and weaknesses of the model, and provides information for how to improve the model. Analyzes multiple sources of data, including graphs and tables, to draw conclusions about a novel or complex scientific investigation or phenomena. Provides several pieces of evidence to support a claim and constructs thorough

	Describes a benefit or drawback of simple design features given a familiar device or prototype.	scientific phenomena or results from an investigation. Analyzes design features of a familiar device or prototype and describes a benefit or drawback of the design.	explanations for scientific phenomena or results from an investigation. Analyzes design features of a novel device or prototype and constructs an explanation for how the design features meet criteria for success or are limited by constraints.

LS1. From Molecules to Organisms: Structures and Processes		
Partially Meeting Expectations ***On MCAS, a student at this level:***	**Meeting Expectations** ***On MCAS, a student at this level:***	**Exceeding Expectations** ***On MCAS, a student at this level:***
Identifies some of the most common elements that make up organic macromolecules. Describes a basic function of a type of organic macromolecule (carbohydrate, lipid, nucleic acid, or protein). Identifies the source of energy and the major reactants and products of photosynthesis by their names or chemical formulas. Describes ATP as a source of usable energy and that it is produced in mitochondria. Describes some major events of the cell cycle (including interphase, mitosis, cytokinesis) and their purposes. Recognizes that chromosomes are separated during mitosis and that mitosis is responsible for tissue growth and repair. Identifies complementary base pairs for a DNA sequence and for an mRNA sequence. Identifies that a gene codes for a protein and describes one function of a protein. Completes a basic model to generally describe how a body system works.	Analyzes models to classify most organic macromolecules and identifies all common elements for a given example. Analyzes models of monomers to identify some types of organic macromolecules and describes some basic functions of these macromolecules. Constructs or completes models of photosynthesis using the names or chemical formulas of reactants and products and describes the importance of photosynthesis. Constructs or completes models of cellular respiration using the names or chemical formulas of reactants and products and describes the importance of cellular respiration. Completes a model to describe how major events of the cell cycle, including DNA replication, allow a cell to grow and survive. Describe the number of chromosomes in a body cell and its daughter cells. Describes the structure of DNA and how its structure affects its function. Describes how genes code for proteins through transcription and translation and describes several functions of proteins. Recognizes that all cells within the same organism have the same genes.	Analyzes models of monomers to consistently identify their organic macromolecules and describes the functions of these molecules. Constructs an explanation about the important uses of the products of photosynthesis for both plants and animals. Analyzes data to determine the relative amount of ATP that is generated by organisms under different conditions. Explains how ATP is used in a variety of ways by both animal and plant cells. Constructs an explanation about how the sequence of events of the cell cycle allows organisms to grow and survive. Explains the importance of mitosis and cytokinesis in an organism. Describes specific functions of several proteins, including enzymes, hormones, and structural proteins. Calculates the percentage of one type of nitrogenous base for a DNA molecule using complementary base pairs. Analyzes and creates models of DNA, RNA, and amino acid chains to describe the products of replication, transcription, or translation.

Describes one way the body maintains homeostasis.	Describes several functions of proteins. Describes the functions of structures and organs of body systems. Interprets models to draw a conclusion about the way the human body maintains homeostasis.	Analyzes data to determine when a gene is expressed and to determine whether replication, transcription, or translation occurs. Constructs an explanation about why different types of cells express different genes, which results in different cell functions. Analyzes data to draw conclusions about how body systems work together to support life functions. Constructs an explanation about how body systems work to restore homeostasis through a sequence of events.

LS2. Ecosystems: Interactions, Energy, and Dynamics		
Partially Meeting Expectations ***On MCAS, a student at this level:***	**Meeting Expectations** ***On MCAS, a student at this level:***	**Exceeding Expectations** ***On MCAS, a student at this level:***
Describes birth and immigration as factors that increase population size, and death and emigration as factors that decrease population size. Identifies some basic ecological relationships (such as predation, competition, mutualism), when given an example. Interprets a basic food web to identify simple ecological relationships. Analyzes a food web to identify the trophic level of a species. Recognizes that less energy is available at higher tropic levels in an energy pyramid. Identifies some carbon cycle processes and recognizes that carbon is released or stored in the environment depending on the process. Recognizes that the biodiversity of an ecosystem is affected by the number of species in the ecosystem. Identify some characteristics of invasive species. Describes one way invasive species can impact other species in an ecosystem. Identifies human impacts (climate change, pollution, habitat destruction) on an	Describes how various biotic and abiotic factors affect a population's birth rate, death rate, immigration rate, or emigration rate. Describes several ecological relationships and determines evidence that supports claims about ecological relationships. Analyzes a food web to describe changes to populations resulting from an increase or decrease of another population. Uses an energy pyramid to calculate the amount of energy that is expected to be stored in different trophic levels. Completes a carbon cycle model showing how carbon is moved through both biotic and abiotic parts of an ecosystem. Describes how the biodiversity of an ecosystem is affected by the number of individuals within a species (genetic diversity is lower in smaller populations). Describes how characteristics of invasive species can affect other species in an ecosystem. Analyzes data to determine the human impact on an ecosystem and describes several ways to reduce the impact of human activity on the ecosystem.	Analyzes multiple factors (such as species interactions, human activities, and natural phenomena) to solve problems relating to population size and carrying capacity of an ecosystem. Analyzes complex food webs and constructs explanations about various interactions in the food web as the sizes of populations change. Constructs an explanation for why only about 10% of the energy stored in one trophic level will be available to the next higher trophic level and how having less energy available reduces the number of organisms that can be supported at higher trophic levels. Constructs an explanation for how several carbon cycle processes interact within an ecosystem and how changes in the environment can disrupt the cycle. Explains how biodiversity of an ecosystem can be impacted by both the number of species in that ecosystem as well as the number of individuals within a species. Constructs thorough explanations for how and why invasive species can affect an ecosystem. Evaluates several solutions for either reducing the impact of human activity on an ecosystem or restoring an ecosystem and explains the benefits and drawbacks of these solutions.

ecosystem and describes some ways to address them.		

LS3. Heredity		
Partially Meeting Expectations ***On MCAS, a student at this level:***	**Meeting Expectations** ***On MCAS, a student at this level:***	**Exceeding Expectations** ***On MCAS, a student at this level:***
Identifies the general purpose of meiosis, that gametes come from two parents, and that egg and sperm combine to produce offspring. Recognizes that inherited traits are a result of genetic information encoded in an organism's DNA and RNA. Completes a simple model to show how a mutation in a DNA sequence can change an mRNA codon. Identifies that only mutations in a gamete can be passed from parent to offspring and that mutations can be a source of genetic diversity. Interprets information to determine when traits show dominant-recessive and codominant inheritance patterns. Identifies genotypes for a certain trait, completes a Punnett square for a given cross, and calculates the expected percentage of offspring for a given genotype or phenotype. Identifies the genotype of an individual in a basic pedigree when the inheritance pattern is given.	Analyzes and completes a basic model of meiosis. Describes the product of fertilization as a zygote (a diploid cell) containing genetic information from both parents. Describes how mutations in DNA can lead to the production of different amino acids and therefore different proteins. Interprets a model of crossing over and concludes that genetic variability increases as a result of crossing over. Interprets information to describe how a trait is inherited by incomplete dominance, sex-linked, multiple alleles, and polygenic inheritance patterns. Constructs and completes Punnett squares and calculates the expected percentages of genotypes and phenotypes of crosses for a given scenario. Analyzes a pedigree to determine the inheritance pattern of a trait. Describes how environmental factors can influence the expression of some inherited traits.	Constructs an explanation of why meiosis is important for maintaining the number of chromosomes from one generation to the next. Explains how crossing over, independent assortment, and random pairing of gametes contribute to the genetic diversity of offspring. Constructs an explanation for how a mutation in a DNA code may or may not result in a phenotypic (trait) change. Analyzes Punnett squares to determine the expected genotype and phenotype percentages for sex-linked traits. Analyzes a complex pedigree to determine genotypes and phenotypes of individuals and to make predictions about future offspring of parents in the pedigree. Uses data to explain the likelihood that a certain trait will be more influenced by genetics or by the environment.

LS4. Evolution		
Partially Meeting Expectations ***On MCAS, a student at this level:***	**Meeting Expectations** ***On MCAS, a student at this level:***	**Exceeding Expectations** ***On MCAS, a student at this level:***
Identifies some types of evidence (genomes, amino acids, fossils, homologous structures) that support the process of evolution. Recognizes that individuals with certain traits survive and produce more offspring than individuals without those traits. Describes that, in general, two organisms from the same species are able to mate and produce offspring. Recognizes that isolated populations generally have a smaller gene pool than larger populations. Recognizes that viruses are unable to reproduce outside of a host cell and that bacteria reproduce through asexual reproduction.	Explains how evolution can be supported by evidence that demonstrates common ancestry. Completes a cladogram to show the evolutionary relationships among several species. Describes how an advantageous heritable trait allows individuals in a population to survive and reproduce more than individuals without that trait. Describes how to determine whether two organisms are closely related and/or from the same species. Describes the role of genetic drift or gene flow in the speciation or extinction of a population. Describes how bacteria and viruses adapt quickly to changing environments due to their high mutation rate and the ability to quickly reproduce.	Constructs an explanation based on a model, such as a cladogram, to support a claim about the evolutionary relatedness of species and explains why comparing genomes provides the best evidence that two species are closely related. Constructs a thorough explanation about evolution, including conditions (heritable variation, differential fitness) that need to be met for evolution to occur and how there will be changes in the frequency of alleles (or traits) within a population over time. Analyzes a situation to determine evidence of selection pressures that could influence the evolution of a population. Constructs explanations based on data for how genetic drift, gene flow, mutations, and natural selection can play a role in the speciation or extinction of a population. Analyzes the results of an investigation to determine conditions that will support the growth of bacteria or viruses.

Introductory Physics

Student results on the MCAS tests are reported according to four achievement levels: *Exceeding Expectations, Meeting Expectations, Partially Meeting Expectations,* and *Not Meeting Expectations.* The descriptors below illustrate the knowledge and skills students demonstrate on MCAS at each level. **Knowledge and skills are cumulative at each level.** No descriptors are provided for the *Not Meeting Expectations* achievement level because students work at this level, by definition, does not meet the criteria of the *Partially Meeting Expectations* level.

	Partially Meeting Expectations ***On MCAS, a student at this level:***	**Meeting Expectations** ***On MCAS, a student at this level:***	**Exceeding Expectations** ***On MCAS, a student at this level:***
Understanding and Application of Disciplinary Core Ideas	Demonstrates a partial understanding of some scientific concepts and processes by identifying and sometimes describing or providing evidence for these concepts and processes. Uses some basic scientific terms in common scientific examples.	Demonstrates a solid understanding of many scientific concepts and processes by mostly describing, explaining, and providing evidence for these concepts and processes. Mostly applies appropriate scientific terms in a variety of applications, including common science examples and some novel situations.	Demonstrates a comprehensive, in-depth understanding of many scientific concepts and processes by consistently describing, explaining, and providing evidence for these concepts and processes. Consistently applies scientific terms in appropriate contexts in both common science examples and many novel situations.

Understanding and Application of Scientific and Engineering Practices	Identifies a testable, scientific question for an investigation. Completes a simple, commonly used model. Uses simple graphs or data to draw general conclusions about a familiar scientific investigation or phenomena. Identifies evidence to support a claim. Describes a benefit or drawback of simple design features given a familiar device or prototype.	Develops some testable, scientific questions for an investigation. Completes or uses a model and describes some strengths and weaknesses of the model. Analyzes multiple sources of data, including graphs and tables, to draw conclusions about a familiar scientific investigation or phenomena. Provides some evidence to support a claim and constructs basic explanations for scientific phenomena or results from an investigation. Analyzes design features of a familiar device or prototype and describes a benefit or drawback of the design.	Consistently develops testable, scientific questions for an investigation. Creates a model, consistently describes the strengths and weaknesses of the model, and provides information for how to improve the model. Analyzes multiple sources of data, including graphs and tables, to draw conclusions about a novel or complex scientific investigation or phenomena. Provides several pieces of evidence to support a claim and constructs thorough explanations for scientific phenomena or results from an investigation. Analyzes design features of a novel device or prototype and constructs an explanation for how the design features meet criteria for success or are limited by constraints.

PS1. Matter and Its Interactions		
Partially Meeting Expectations ***On MCAS, a student at this level:***	**Meeting Expectations** ***On MCAS, a student at this level:***	**Exceeding Expectations** ***On MCAS, a student at this level:***
Interprets a model to determine that energy is released during the processes of fission, fusion, and radioactive decay.	Analyzes a model to determine whether fission, fusion, or a radioactive decay (alpha, beta, or gamma) process occurred.	Analyzes incomplete models of fission, fusion, and radioactive decay and describes the results of each in terms of energy and products.

PS2. Motion and Stability: Forces and Interactions		
Partially Meeting Expectations ***On MCAS, a student at this level:***	**Meeting Expectations** ***On MCAS, a student at this level:***	**Exceeding Expectations** ***On MCAS, a student at this level:***

Solves simple problems involving average speed, velocity, and acceleration. Interprets a position vs. time graph to determine how far an object is from its starting location. Interprets a scenario to determine the relative magnitude of a force. Determines a net force using Newton's 2nd law or by interpreting a free-body force diagram with two colinear forces. Solves simple momentum and change in momentum (impulse) problems. Interprets a model to determine whether two charges will attract or repel. Describes how the magnitude of charges or the distance between charges affects electrostatic forces. Describes how the masses of objects or the distance between objects affects gravitational forces. Solves simple problems using Ohm's Law when given two of the three variables (current, voltage, or resistance). Identifies a schematic symbol for a simple circuit element and generally explains its role.	Solves problems involving acceleration, velocity, and change in position for a given time. Analyzes motion graphs and their slopes to solve for and compare speeds, velocities, accelerations, and net forces. Analyzes free-body force diagrams to determine which diagram represents a given system. Solves for an unknown force by interpreting a model with two or more colinear forces when also given the net force. Solves for the total momentum or change in momentum of a system. Interprets a model to determine the direction an object will move after a collision. Compares the magnitude and the direction of the forces that two objects exert on each other when they collide. Compares models of pairs of masses or charges to order the magnitude of the gravitational or electrostatic forces. Completes a model to represent electrostatic forces between charges. Interprets a model to support a claim that an electric current produces a magnetic field or a claim that a changing magnetic field produces an electric current. Describes how a change to a circuit affects current, voltage, or resistance.	Solves a motion problem by analyzing a model and then applying information from the model to solve for velocity or acceleration. Explains how changing a system would affect an object's velocity or acceleration. Solves force problems by analyzing motion graphs and then models the forces involved using free-body force diagrams. Analyzes a motion graph and then applies information from the graph to solve a momentum problem. Describes that the total momentum of a system stays the same during a collision and solves for velocity or mass by applying conservation of momentum. Explains how forces involved in a collision can be minimized. Applies proportional reasoning to solve for how changing the distance between a pair of masses or a pair of charges affects the forces between the pair. Applies proportional reasoning when multiple variables are changed to determine the forces between a pair of masses or charges. Describes the effect of a gravitational or electrostatic force between two objects by solving for the force using either

	Interprets a series circuit diagram with several circuit elements and solves for current, resistance, or voltage. Interprets simple series or parallel circuit diagrams and explains which circuit elements will have the same current through them and which elements will have the same voltage drop across them.	Newton's law of gravitation or Coulomb's law. Explains that the interplay of electric and magnetic forces is the basis for electric motors and generators. Analyzes series and parallel circuit diagrams with multiple circuit elements to compare and solve for current, voltage, and resistance.

PS3. Energy		
Partially Meeting Expectations ***On MCAS, a student at this level:***	**Meeting Expectations** ***On MCAS, a student at this level:***	**Exceeding Expectations** ***On MCAS, a student at this level:***
Solves for gravitational potential energy when given the height and mass of an object. Describes an example of energy being converted from one form to another. Interprets a model to determine a location where gravitational potential energy or kinetic energy is either the greatest or the least. Solves simple problems for work when given the force and distance. Solves efficiency problems when given energy in and energy out. Interprets a simple graph to determine when thermal equilibrium is reached. Recognizes that heat flows from a substance with a higher temperature to a substance with a lower temperature. Recognizes the relationship between average molecular motion and temperature. Describes the relative amount of force between two magnets as they are moved closer together or farther apart.	Analyzes a model of a system and then uses information from the model to calculate kinetic energy or gravitational potential energy. Describes that energy cannot be created or destroyed, but energy may enter or leave a system. Compares an object's kinetic energy at two positions or an object's potential energy at two positions when mechanical energy is conserved. Analyzes data to solve mechanical energy problems. Interprets a model of a device and explains how to increase the efficiency of the device. Explains how the temperatures in two substances change as the substances reach thermal equilibrium. Describes how changing the mass of a substance affects the energy required to cause a temperature change. Analyzes electric field diagrams and determines the direction and relative strength of the electric field around two charges. Explains how the energy stored in a field between two magnets or two charges changes when they are moved different distances apart.	Constructs an explanation for how kinetic energy and potential energy change over time in a given model. Explains how the mechanical energy of a system can change, due to work being done on the system by a force, while maintaining the law of conservation of energy. Solves complex work problems, including first solving for initial and final mechanical energy. Analyzes a graph to compare the energy efficiency of multiple devices. Explains how the average molecular motion of molecules in two substances changes as the substances reach thermal equilibrium, and how energy is conserved in a system as thermal equilibrium is reached. Analyzes a model and solves problems for the amount of heat transferred in a system, the specific heat of a substance, or the initial or final temperature of a substance. Interprets a model to describe the motion of a freely moving charged particle and the energy stored in the field between two charged particles.

PS4. Waves and Their Applications in Technologies for Information Transfer		
Partially Meeting Expectations ***On MCAS, a student at this level:***	**Meeting Expectations** ***On MCAS, a student at this level:***	**Exceeding Expectations** ***On MCAS, a student at this level:***
Solves simple wave problems for velocity/speed, wavelength, or frequency when given two of these three variables. Identifies the wavelength of a wave on a model. Solves simple wave problems involving period and frequency when given one of the variables. Identifies differences between mechanical waves and electromagnetic waves. Recognizes the relationships between frequency and energy of a light particle. Identifies evidence of light behaving like a wave or light behaving like a particle. Interprets simple models of the photoelectric effect. Interprets simple models of common wave behaviors, including resonance, diffraction, refraction, and interference.	Analyzes data to determine additional information needed to solve wave problems. Describes how the particles in a medium move when a longitudinal or transverse wave travels through the medium. Describes several properties of mechanical waves and electromagnetic waves. Compares electromagnetic radiation in terms of frequency, energy, and wavelength. Analyzes a model and explains the causes of resonance and refraction. Analyzes a model of a technology or device and describes how wave behaviors or the photoelectric effect are used in the technology or device.	Analyzes models of waves and uses information from the models to solve problems. Interprets a graph with relative speeds of mechanical waves to determine the states of matter of various media. Constructs an explanation with evidence about how light can behave like a wave and how it can behave like a particle. Explains the relationship between photon energy and the electrons ejected by the photoelectric effect. Analyzes a model of constructive and destructive interference and determines the amplitude of a wave pulse that results from the interference. Analyzes how a technology or device uses waves and describes how changing the properties of the waves would influence the device.

Appendix BB – Final Recommended Cut Scores on IRT Scale and Scaling Constants

Table B.1: Final Recommended Cut Scores on IRT Scale

	Cut Scores (Raw Score)			Cut Scores (IRT)			Scaling Constants	
Subject	**PME**	**ME**	**EE**	**PME**	**ME**	**EE**	**A**	**B**
Biology	17	34	50	-0.8500	0.2100	1.3000	27.90698	493.7209
Introductory Physics	17	35	51	-1.0100	0.1200	1.2600	26.43172	496.6960

Note: PME – Partially Meeting Expectations; ME – Meeting Expectations; EE – Exceeding Expectations

Appendix CC – Participant Meeting Materials

The materials developed for the Biology standard setting committee are provided as an example of the materials developed and provided to the participants. Since the materials provided to participants contained secure information, any place where secure information would be provided, that information would be removed. Additionally, the following materials will not be provided within the appendix:

- Test form – This was presented to participants through the online testing platform used during the spring 2022 administration, TestNav 8.
- Open-ended item rubrics – These documents presented the scoring rubrics and notes and student-produced response examples for each open-ended item presented to participants.
- Practice item judgment set – This was presented to participants through the online testing platform used during the spring 2022 administration, TestNav 8.

Participant Agenda

MCAS Standard Setting Meeting
August 2022

Agenda

Day 1 – August 9, 2022

8:30 am *General Session*
Welcome
Overview of MCAS STE Assessments
Standard Setting Overview

10:00 am *Break*

10:10 am *Breakout Sessions (Biology and Introductory Physics)*

Welcome and Introductions
Experience the Assessment Activity

11:30 am *Lunch*

Experience the Assessment Activity (cont.)
Achievement Level Descriptors Discussion

1:50 pm *Break*

Borderline Descriptions Development Training
Borderline Description Development – Meeting Expectations

4:30 pm *End-of-Day*

Day 2 – August 10, 2022

8:30 am *Breakout Session (Biology and Introductory Physics)*

Borderline Descriptions Development – Partially Meeting and Exceeding

10:00 am *Break*

Achievement Level Setting Training
Practice Judgment Activity and Discussion

11:30 am *Lunch*

Round 1 Judgments

1:30 pm *Break*

Round 1 Judgment Feedback and Discussion

3:15 pm *Break*

Round 2 Judgments

4:30 pm *End-of-Day*

Day 3 – August 11, 2022

8:30 am *Breakout Session (Biology and Introductory Physics)*

Round 2 Judgment Feedback and Discussion
Round 3 Judgments

10:45 am *Break*

Round 3 Judgment Feedback and Discussion
Next Steps and Closing

12:00 pm *End-of-Day*

MCAS Non-disclosure Agreement

Massachusetts Department of Elementary and Secondary Education

75 Pleasant Street, Malden, Massachusetts 02148-4906
338-3000

Telephone: (781)
TTY: N.E.T. Relay
1-800-439-2370

Jeffrey C. Riley
Commissioner

Massachusetts Comprehensive Assessment System
NON-DISCLOSURE AGREEMENT

In order to preserve and ensure the security, validity, and integrity of Massachusetts Comprehensive Assessment System (MCAS) tests, the Massachusetts Department of Elementary and Secondary Education (the Department) requires that all individuals whom the Department authorizes to participate in the development, review, and production of MCAS tests and reports accept the terms of the following non-disclosure agreement.

- With the exception of test items released by the Department for informational purposes, all MCAS test items are deemed secure instruments. The materials are specifically excluded from the Massachusetts Public Records Law. (**G. L. c. 4, § 7(26) (l)**) As a result, I agree not to reproduce, discuss, or in any way release or distribute test items and associated materials to unauthorized persons (i.e., persons not specifically authorized by the Department to have access to secure MCAS materials and information).
- All information about MCAS English language arts passages and English language arts, mathematics, history and social science, and science and technology/engineering graphics under consideration for inclusion in current or future MCAS tests is confidential. Therefore, I agree not to share this information in any way with unauthorized persons.
- Details about MCAS test construction, including the positions of items in test forms, must be kept secure. Consequently, I agree not to share MCAS test blueprints or any information related to MCAS test blueprints with unauthorized persons.
- Discussions and materials related to all technical aspects of the MCAS program, including possible new models and future directions, are confidential. Therefore, I agree not to reveal information regarding discussions and deliberations that take place in committee meetings to unauthorized persons.
- I further understand and agree that all MCAS test items, ideas for items, and related test materials developed, reviewed, and produced by authorized persons working in collaboration with the Department are and will forever remain the exclusive property of the Massachusetts Department of Elementary and Secondary Education.

By signing below, I, as a member of the MCAS Bias and Sensitivity Committee, Standard Setting Committee, Assessment Development Committee, or Technical Advisory Committee, acknowledge and accept that I am bound by the terms of this agreement prohibiting the disclosure of information regarding secure materials and discussions. I also acknowledge and accept that my failure to abide by any term of this non-disclosure agreement will result in serious consequences, including but not limited to action to limit or revoke my Massachusetts educator license.

NAME: _____

COMMITTEE (include subject & grade): _____

AFFILIATION: _____

SIGNATURE: _____ **DATE:** _____

Experience the Test Response Record Form

MCAS Standard Setting Meeting
August 2022

Experience the Assessment Notes Sheet
Biology

Sequence	Item Notes
1	
2	
3	
4	
5	
6	
7	

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

Item Judgment Round Record Form

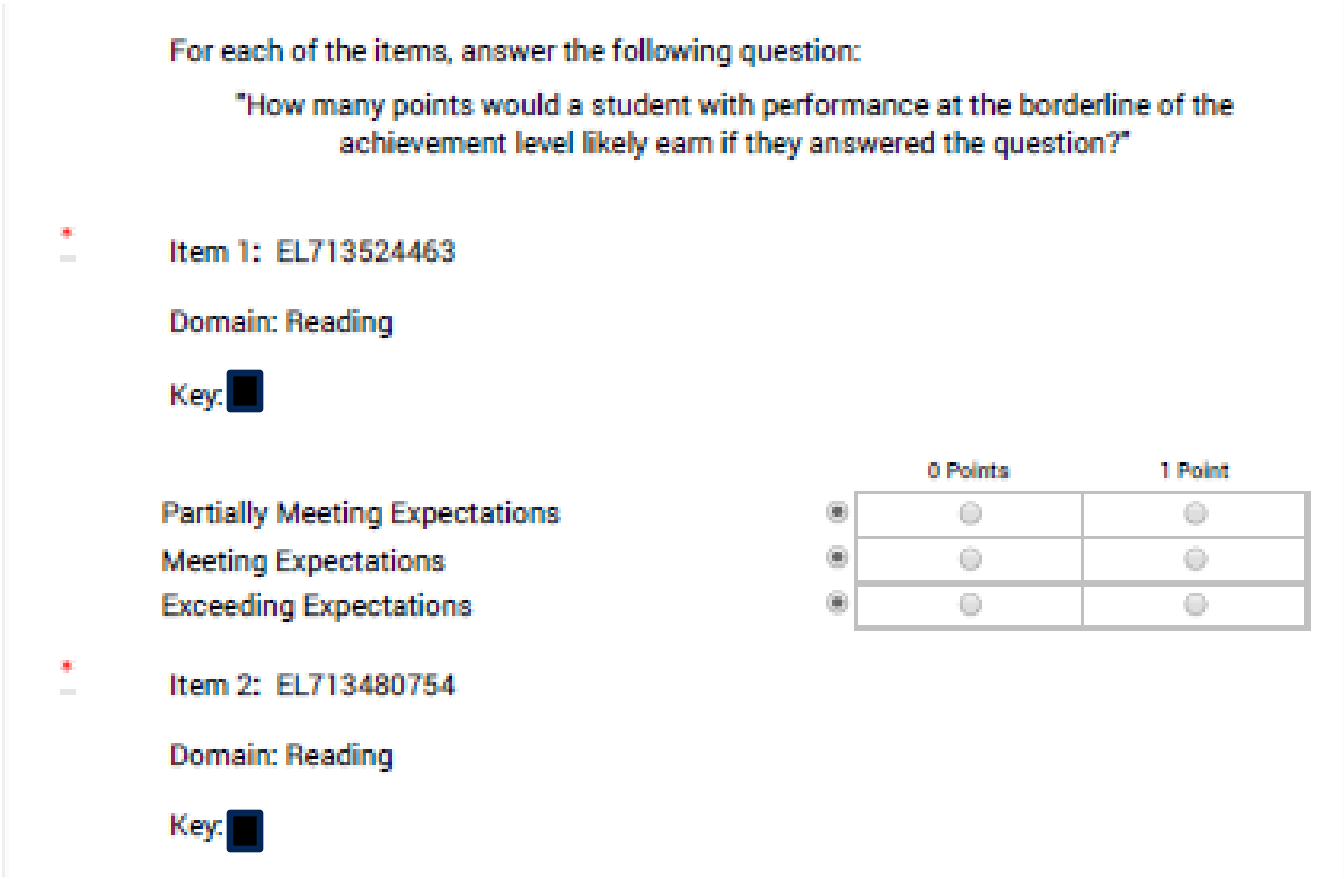
MCAS Standard Setting Meeting
August 2022

Judgment Record Sheet
Biology

Seq.	Item	Item Maximum Score	Judgment Round								
			1			2			3		
			PME	ME	EE	PME	ME	EE	PME	ME	EE
1	SC626969020	1									
2	SC800159954	1									
3	SC721652006	1									
4	SC735277981	1									
5	SC801968916	1									
6	SC800133220	2									
7	SC802464161	1									
8	SC723341794	1									
9	SC316130	1									
10	SC802252224	1									

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

Item Judgment Survey



Note: The survey for only the first two items is shown.

Process Evaluation #1

Process Evaluation Day 1

Next Generation Massachusetts Comprehensive Assessment System (Next-Gen MCAS)

Standard Setting Meeting

Process Evaluation Survey #1

The purpose of this evaluation is to collect information about your experience in recommending cut scores associated with the achievement levels for the MCAS assessments. Your opinions provide an important part of our evaluation of this meeting.

* 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.

	Not Successful	Partially Successful	Successful	Very Successful
Meeting pre-work	○	○	○	○
General session training	○	○	○	○
Overview of the MCAS assessments	○	○	○	○
Introduction to the standard setting process	○	○	○	○
Experiencing the actual assessment	○	○	○	○
Discussion of the scoring of items on the assessment	○	○	○	○
Discussion of achievement level descriptors (ALDs)	○	○	○	○
Development and discussion of the borderline descriptions	○	○	○	○
Overview of the standard-setting procedure	○	○	○	○
Practice exercise for the standard-setting procedure	○	○	○	○

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

	Very Useful	Useful	Somewhat Useful	Not Useful
Achievement Level Descriptors (ALDs)	○	○	○	○
Borderline Descriptions	○	○	○	○

* How adequate were the following elements of the session?

	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate
Total amount of time to create and discuss borderline descriptions	○	○	○	○
Training provided on the standard-setting process	○	○	○	○
Amount of time spent training	○	○	○	○
Total amount of time to discuss the practice judgment activity	○	○	○	○

Close this window

Process Evaluation #2

Next Generation Massachusetts Comprehensive Assessment System (Next-Gen MCAS)

Standard Setting Meeting

Process Evaluation Survey #2

The purpose of this evaluation is to collect information about your experience in recommending cut scores associated with the achievement levels for the MCAS assessments. Your opinions provide an important part of our evaluation of this meeting.

* 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.

	Not Successful	Partially Successful	Successful	Very Successful
Judgment rounds	○	○	○	○
Judgment round feedback - committee-level statistics	○	○	○	○
Judgment round feedback - panelist cut score agreement data	○	○	○	○
Judgment round feedback - panelist judgment agreement data	○	○	○	○
Judgment round feedback - impact data	○	○	○	○
Discussions after each round	○	○	○	○

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

	Very Useful	Useful	Somewhat Useful	Not Useful
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	○	○	○	○

* How adequate were the following elements of the session?

	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate
Amount of time to make judgments	○	○	○	○
Visual presentation of the feedback provided	○	○	○	○
Number of judgment rounds	○	○	○	○

* In applying the standard-setting method, you were asked to recommend cut scores (separating four achievement levels) for student performance on MCAS assessments.

How confident do you feel that the Achievement Level Descriptors (ALDs) for grade 10 ELA are reasonable for each student achievement level?

	Not Confident	Somewhat Confident	Confident	Very Confident
Exceeding Expectations	○	○	○	○
Meeting Expectations	○	○	○	○
Partially Meeting Expectations	○	○	○	○

* How confident do you feel that the final cut score recommendations for grade 10 ELA represent appropriate levels of student performance?

	Not Confident	Somewhat Confident	Confident	Very Confident
Exceeding Expectations	○	○	○	○
Meeting Expectations	○	○	○	○
Partially Meeting Expectations	○	○	○	○

Page 2

* How adequate were the following elements of the session?

	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate

Facilities used for the general session	○	○	○	○
Facilities used for the breakout session	○	○	○	○
Computers used during the meetings	○	○	○	○
Standard Setting website for accessing materials and making judgments	○	○	○	○
Materials provided in the folder	○	○	○	○
Work space in table groups during the meeting	○	○	○	○

* Did you have adequate opportunities during the session to:

	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate
Express your opinions about student achievement 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	○	○	○	○

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

	No	Sometimes	Yes
Fellow panelists	○	○	○
Facilitators	○	○	○

* Please use the space below to provide any additional comments you have regarding the standard setting process, facilitators, materials, etc.

Paragraph

Path: p

Close this window

Appendix DD – Committee Participant Composition

Table D.1: Participant Position

	STE	
	Biology	**Physics**
Teacher (K–12)	14	18
Teacher (Higher Ed.)	1	1
Administrator (School)	2	0
Administrator (District)	0	0
Coordinator/Coach	2	1
Total	19	20

Table D.2: Years of Relevant Teaching Experience

	STE	
	Biology	**Physics**
1 to 5 years	1	3
6 to 10 years	3	4
11 to 15 years	4	5
16 to 20 years	5	5
More than 20 years	6	3
Total	19	20

Table D.3: Experience Teaching Student Populations

	STE	
	Biology	**Physics**
Mainstream special education	19	18
Self-contained special education	7	2
English language learners (ELL)	19	19
General education	19	20
Vocational technical education	5	5

Table D.4: Demographic: Gender

	STE	
	Biology	**Physics**
Female	12	6
Male	7	12
Other	0	2

Table D.5: Demographic: Race and Ethnicity

	STE	
	Biology	**Physics**
Asian	1	3
Black or African American	0	1
Hispanic	0	1
Middle Eastern	0	2
White	19	14

Table D.6: Currently Work in a School District

	STE	
	Biology	**Physics**
Yes	17	19
No (Higher Ed)	2	1

Table D.7: Size of School District

	STE	
	Biology	**Physics**
Small	3	8
Medium	6	7
Large	8	4

Table D.8: Type of School District

	STE	
	Biology	**Physics**
Rural	1	0
Metropolitan/Urban	6	6
Suburban	10	13

Table D.9: Socioeconomic Status of School District

	STE	
	Biology	**Physics**
Low	7	5
Moderate	7	8
High	3	6

Appendix EE – Standard Setting Meeting Agenda

MCAS High School STE Standard Setting

Standard Setting Meeting – Agenda

Day 1 (Tuesday – August 9, 2022)

Start time	End time	Activity
8:00 am	8:30 am	Breakfast
General Session		
8:30 am	9:00 am	Welcome introductions, materials orientation, and security
9:00 am	10:00 am	Standard Setting Overview
10:00 am	10:10 am	*Break*
Breakout Session (Biology and Introductory Physics)		
10:10 am	10:30 am	Welcome, Introductions, and Orientation
10:30 am	11:30 am	Experience the Assessment Overview of Science Assessments Orientation to activity Individual Activity
11:30 am	12:15 pm	*Lunch*
12:15 pm	1:00 pm	Experience the Assessment Individual Activity (cont.)
1:00 pm	1:20 pm	Review of Scoring Materials
1:20 pm	1:50 pm	Item Difficulty Comparison
1:50 pm	2:00 pm	*Break*
2:00 pm	2:30 pm	Achievement Level Descriptors (ALDs) Introduction to ALDs Table-group discussions Whole-group discussions
2:30 pm	3:00 pm	Borderline Descriptions Training Introduction to Borderline Descriptions Modeling of borderline descriptions development
3:00 pm	4:30 pm	Borderline Descriptions Development – Meeting Expectations Table-group discussion Whole-group discussion

Day 2 (Wednesday – August 10, 2022)

Start time	End time	Activity
8:00 am	8:30 am	Breakfast
8:30 am	10:00 am	Borderline Descriptions Development – Partially Meeting and Exceeding Table-group discussion Whole-group discussion
10:00 am	10:15 am	Break
10:15 am	10:45 am	Achievement Level Setting Training
10:45 am	11:30 am	Practice Judgment Activity Practice Judgment Activity Group Discussion
11:30 am	12:15 pm	Lunch
12:15 pm	1:30 pm	Round 1 Judgments Round 1 readiness form Independent round 1 judgments
1:30 pm	2:00 pm	Break
2:00 pm	3:15 pm	Round 1 Judgment Feedback and Discussion Introduction to feedback data Whole-group discussion
3:15 pm	3:30 pm	Break
3:30 pm	4:30 pm	Round 2 Judgments Round 2 readiness form Independent round 2 judgments

Day 3 (Thursday – August 11, 2022)

Start time	End time	Activity
8:00 am	8:30 am	Breakfast
8:30 am	10:00 am	Round 2 Judgment Feedback and Discussion Whole-group discussion Impact data
10:00 am	10:45 am	Round 3 Judgments Round 3 readiness form Independent round 3 judgments
10:45 am	11:15 am	Break
11:15 am	11:45 am	Round 3 Judgment Feedback and Discussion
11:45 am	12:00 pm	Close-out and Evaluations

Appendix FF – Examples of Feedback Data

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

Individual Item—Level Judgments

This provided the participant with the actual item-level judgments that were recorded in Moodle for the participant. This was provided so that the participant could check that the system recorded the judgments correctly.

Biology - Individual Rating - Round 1

Table=1 Name=

SeqNo	UIN	PME	ME	EE
1MC	SC626969020	0	1	1
2MX	SC800159954	1	1	1
3XI	SC721652006	1	1	1
4MX	SC735277981	0	1	1
5XI	SC801968916	1	1	1
6MX	SC800133220	0	1	2

Individual Test—Level Recommendation

This provided the participant with the recommendations for test-level cut scores based on their item judgments for the Partially Meeting Expectations, Meeting Expectations, and Exceeding Expectations achievement levels.

Biology - Individual Cut Scores - Round 1

Table=1 Name=

PME Raw Score	ME Raw Score	EE Raw Score
17	43	57

Overall Test—Level Recommendations

This provided the participant with the aggregate test-level recommendation, based on the individual participants in the committee, including the number of participants, the mean recommendation, the median recommendation, the minimum and maximum recommendation, and the first and third quartiles for each achievement level.

Biology Round 1 Summary Statistics - Overall

	N	Mean	Median	Min	Max	Q1	Q3
PME Raw Score	19	19.79	20.00	11.00	31.00	17.00	23.00
ME Raw Score	19	43.79	44.00	30.00	54.00	41.00	49.00
EE Raw Score	19	57.68	59.00	52.00	60.00	56.00	60.00

Item-level Judgment Agreement

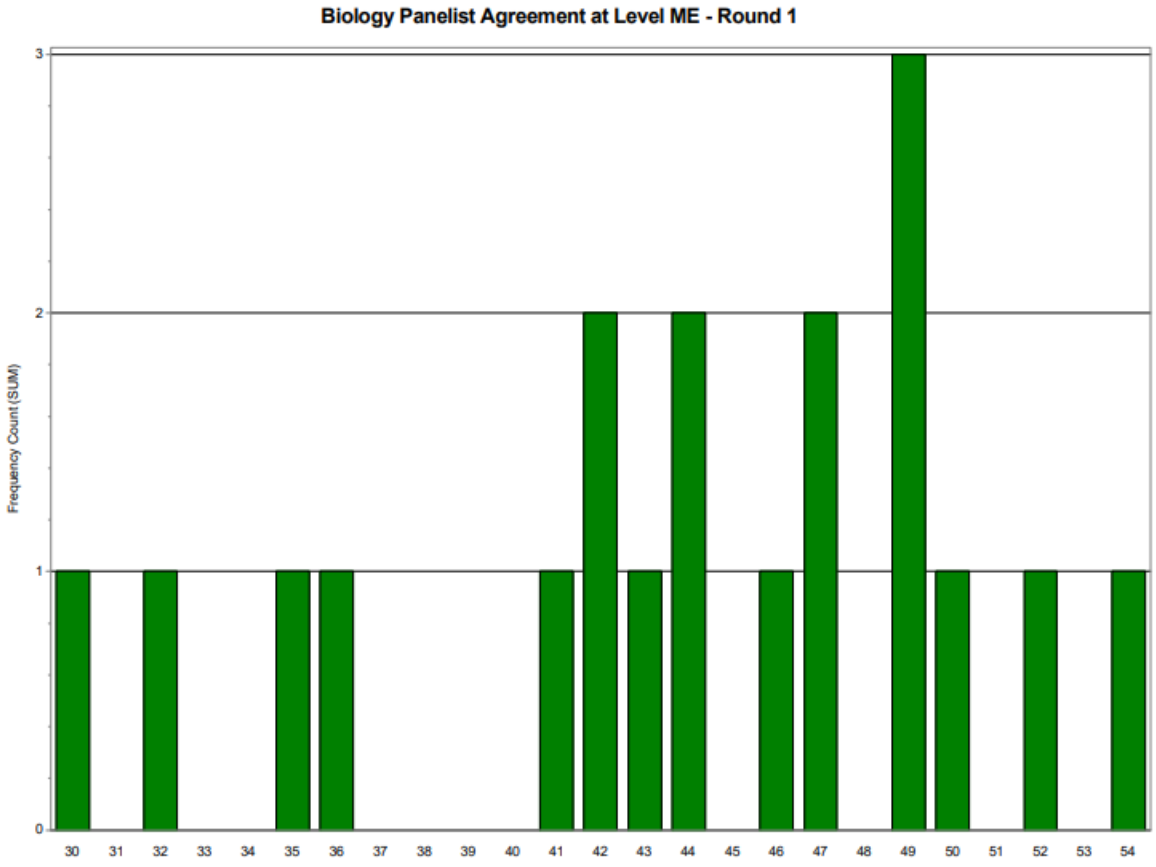
This provided the participants with item-level judgment distributions for the committee for each item. Additionally, for each achievement level, the items with the greatest level of judgment disagreement were identified.

Biology Round 1 Level ME

SeqNo	UIN	Max Points	0	1	2	3	4
1MC	SC626969020	1	.	100%	.	.	.
2MX	SC800159954	1	11%	89%	.	.	.
3XI	SC721652006	1	.	100%	.	.	.
4MX	SC735277981	1	.	100%	.	.	.
5XI	SC801968916	1	5%	95%	.	.	.
6MX	SC800133220	2	11%	68%	21%	.	.
7MC	SC802464161	1	5%	95%	.	.	.

Test-level Participant Recommendation Agreement

This feedback was presented to participants by the facilitator. It presented bar graphs displaying the distribution of participant recommendations for the cut score, by raw score, for each achievement level: Partially Meeting Expectation, Meeting Expectations, and Exceeding Expectations. Graphs displaying consecutive achievement levels (Partially Meeting Expectations and Meeting Expectations) on the scale graph were also presented.



Item Score Mean and Score Distribution

This provided, for each item, the mean score and the distribution of scores received by students during the Spring 2017 administration. The results presented were based on the sample of data used to create the impact data.

Item Performance Data

Biology

Question No	UIN	Item Type	Max Score points	Item Mean	Point Distribution				
					0 pts	1 pt	2 pts	3 pts	4 pts
1	SC626969020	MC	1	0.684	32%	68%			
2	SC800159954	MX	1	0.358	64%	36%			
3	SC721652006	XI	1	0.676	32%	68%			
4	SC735277981	MX	1	0.530	47%	53%			
5	SC801968916	XI	1	0.363	64%	36%			
6	SC800133220	MX	2	0.732	43%	42%	16%		
7	SC802464161	MC	1	0.710	29%	71%			

Appendix GG – Committee Recommended Cut Scores by Round

Table G.1: Biology

Achievement Level	Maximum Score	Rounds			Vertical Articulation
		1	2	3	
Partially Meeting Expectations	60	20	16	17	16
Meeting Expectations		44	38	37	34
Exceeding Expectations		59	53	52	51

Table G.2: Introductory Physics

Achievement Level	Maximum Score	Rounds			Vertical Articulation
		1	2	3	
Partially Meeting Expectations	60	12	16	16	17
Meeting Expectations		34	39	37	35
Exceeding Expectations		50	53	51	51

Appendix HH – Recommended Cut Score Summary Statistics

Biology

Round	Statistic	Achievement Level		
		Partially Meeting Expectations	Meeting Expectations	Exceeding Expectations
1	Mean	19.79	43.79	57.68
	Minimum	11	30	52
	Q1	17	41	56
	Median	**20**	**44**	**59**
	Q3	23	49	60
	Maximum	31	54	60
2	Mean	15.42	36.89	52.37
	Minimum	7	24	46
	Q1	13	31	50
	Median	**16**	**38**	**53**
	Q3	18	41	55
	Maximum	22	52	58
3	Mean	17	35.95	51.42
	Minimum	11	26	44
	Q1	15	33	49
	Median	**17**	**37**	**52**
	Q3	19	40	54
	Maximum	25	44	56

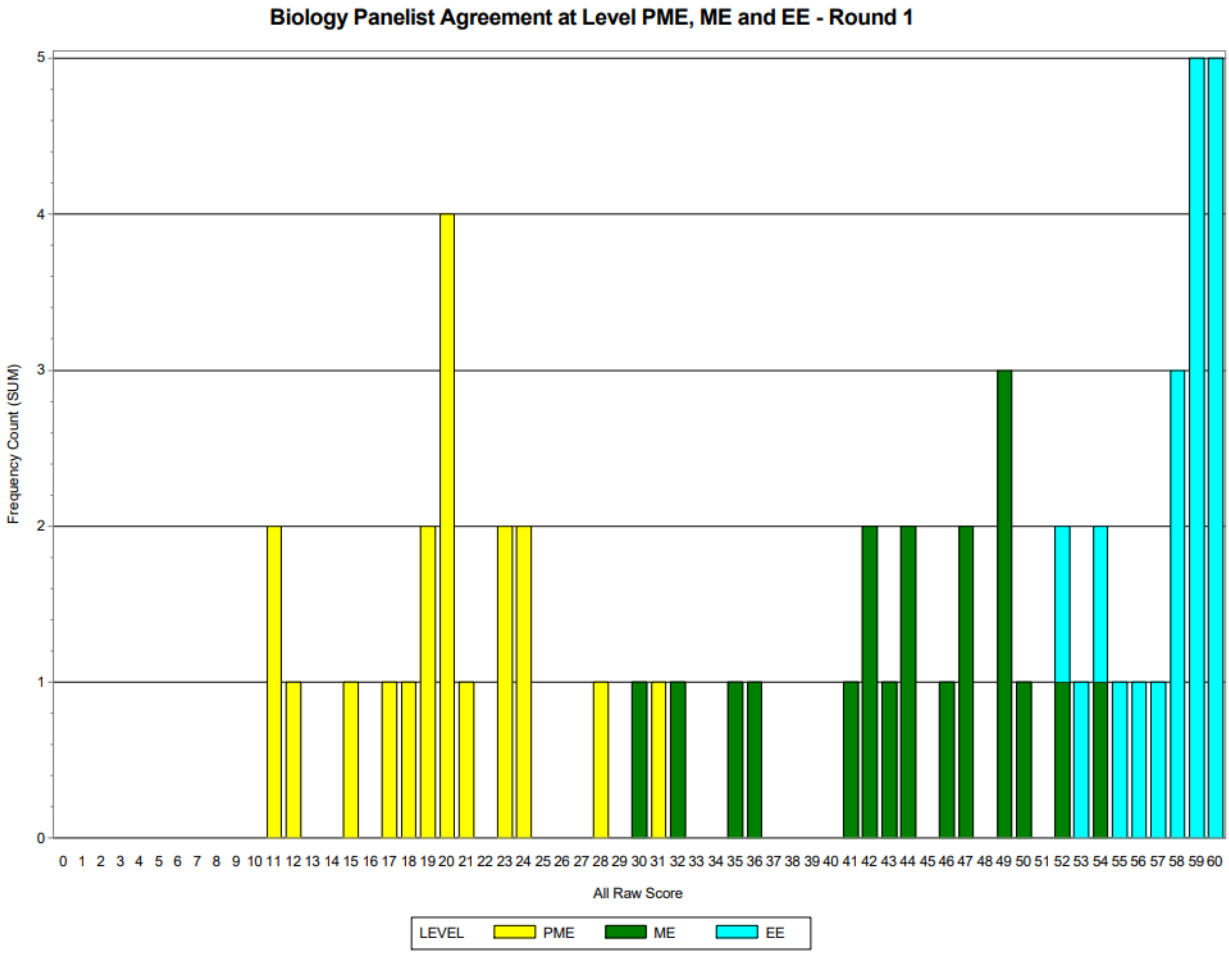
Physics

Round	Statistic	Achievement Level		
		Partially Meeting Expectations	Meeting Expectations	Exceeding Expectations
1	Mean	12.67	33.39	49.11
	Minimum	8	25	38
	Q1	10	29	45
	Median	**12**	**34**	**50**
	Q3	14	36	53
	Maximum	23	44	55
2	Mean	16.11	39.17	51.61
	Minimum	11	33	42
	Q1	15	37	50
	Median	**16**	**39**	**53**
	Q3	18	41	54
	Maximum	21	48	59
3	Mean	15.26	37.21	51.26
	Minimum	12	33	46
	Q1	13	35	49
	Median	**16**	**37**	**51**
	Q3	17	39	54
	Maximum	21	44	56

Appendix II – Test-Level Participant Judgment Agreement

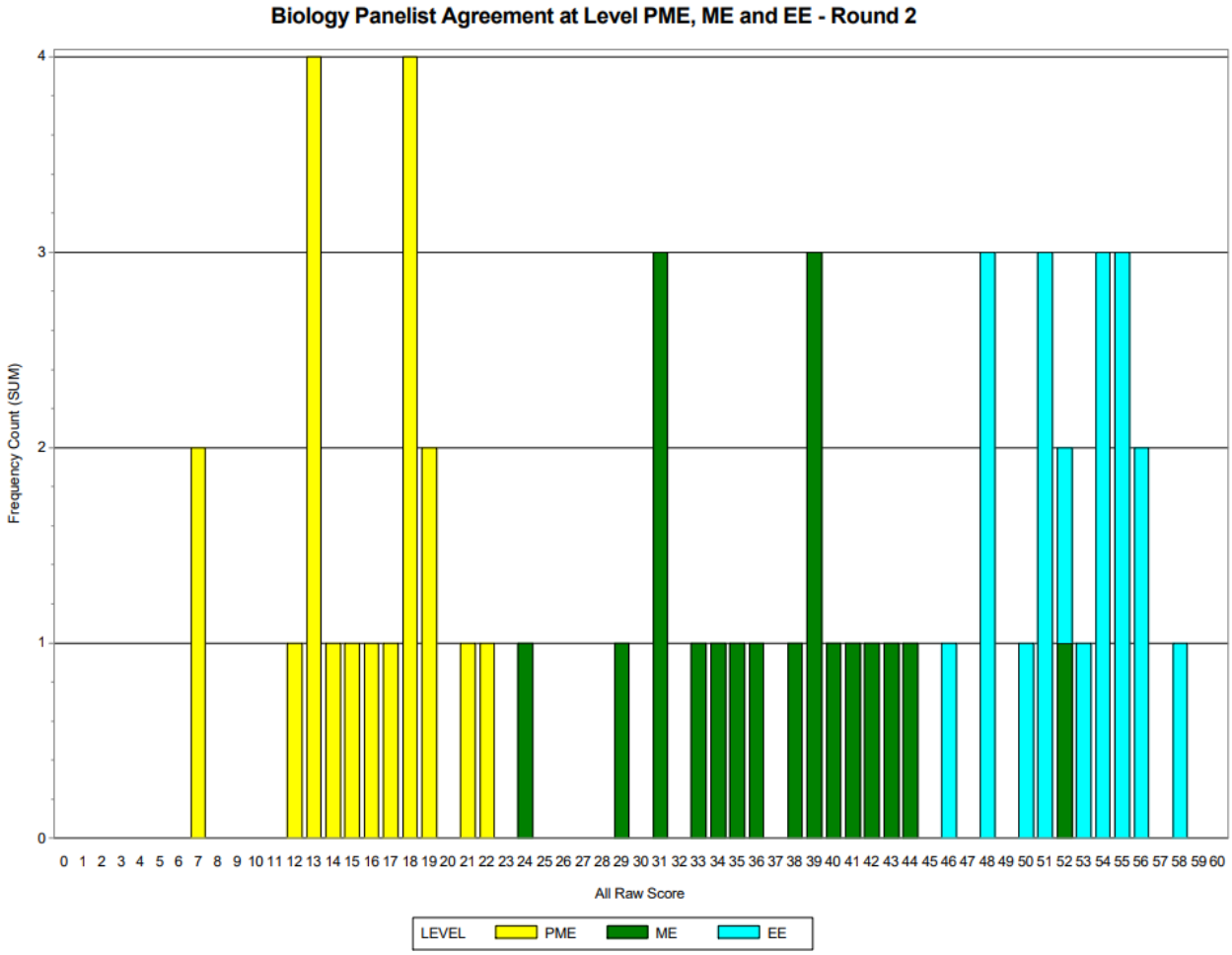
Biology

Round 1:



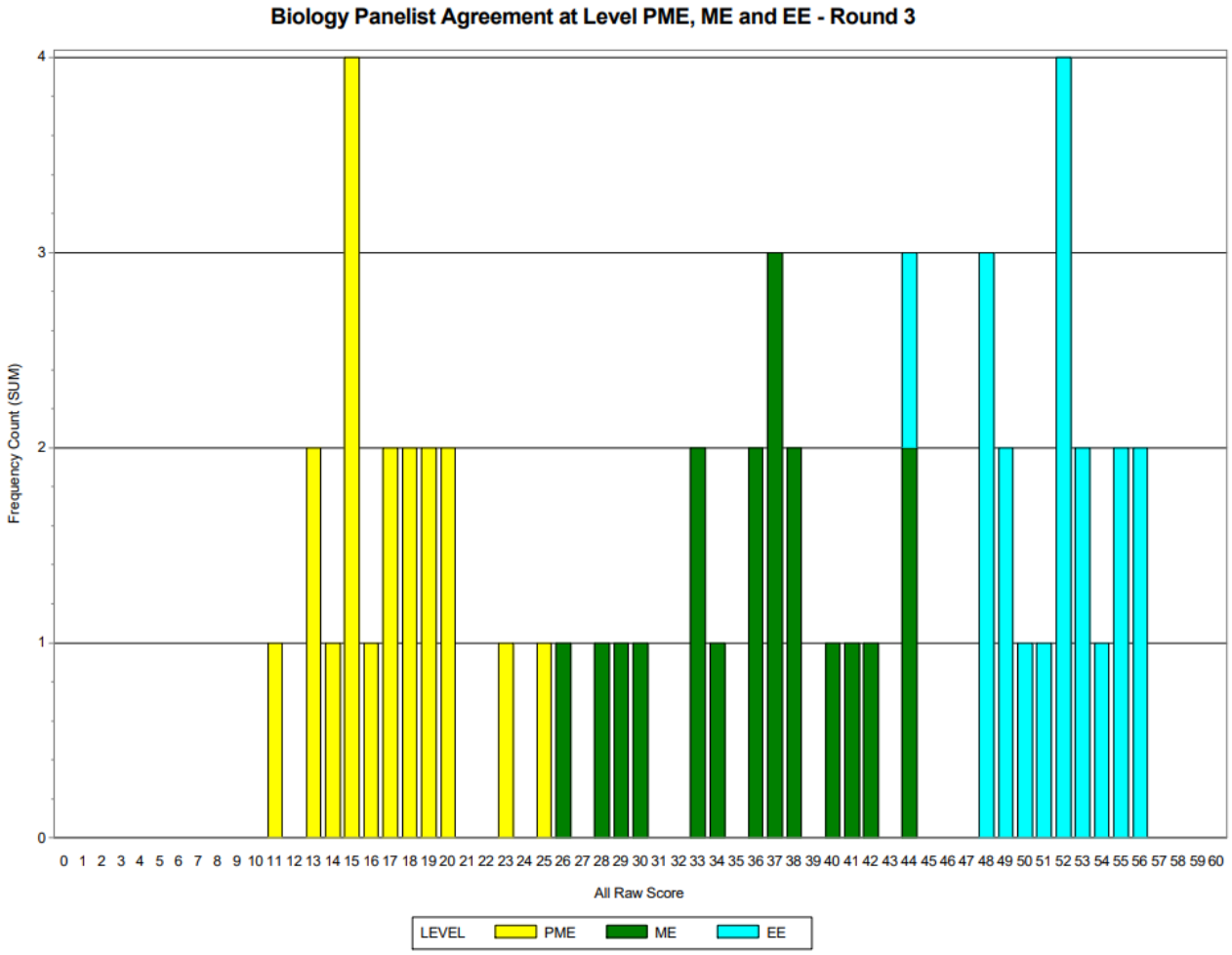
All Three Achievement Levels Concurrently

Round 2:



All Three Achievement Levels Concurrently

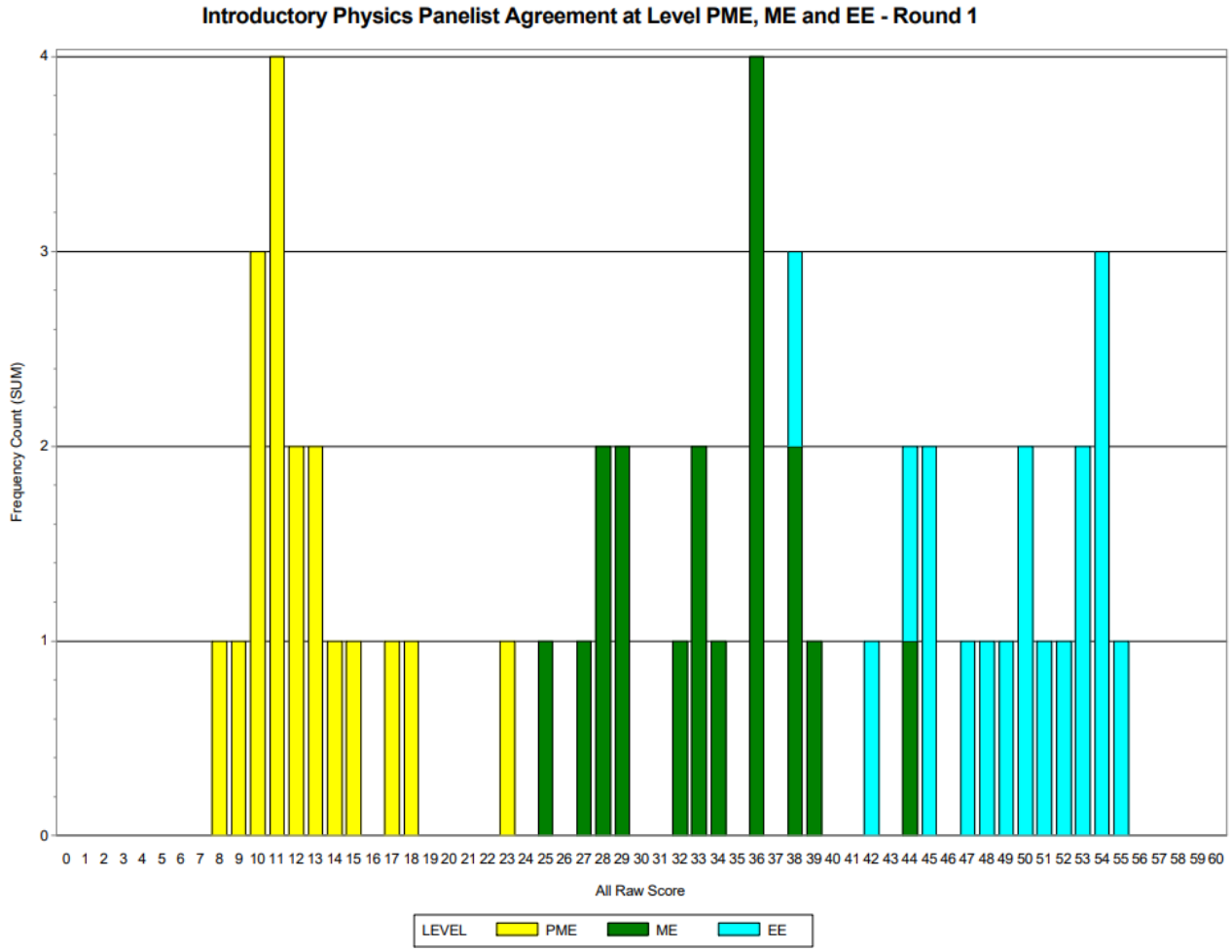
Round 3:



All Three Achievement Levels Concurrently

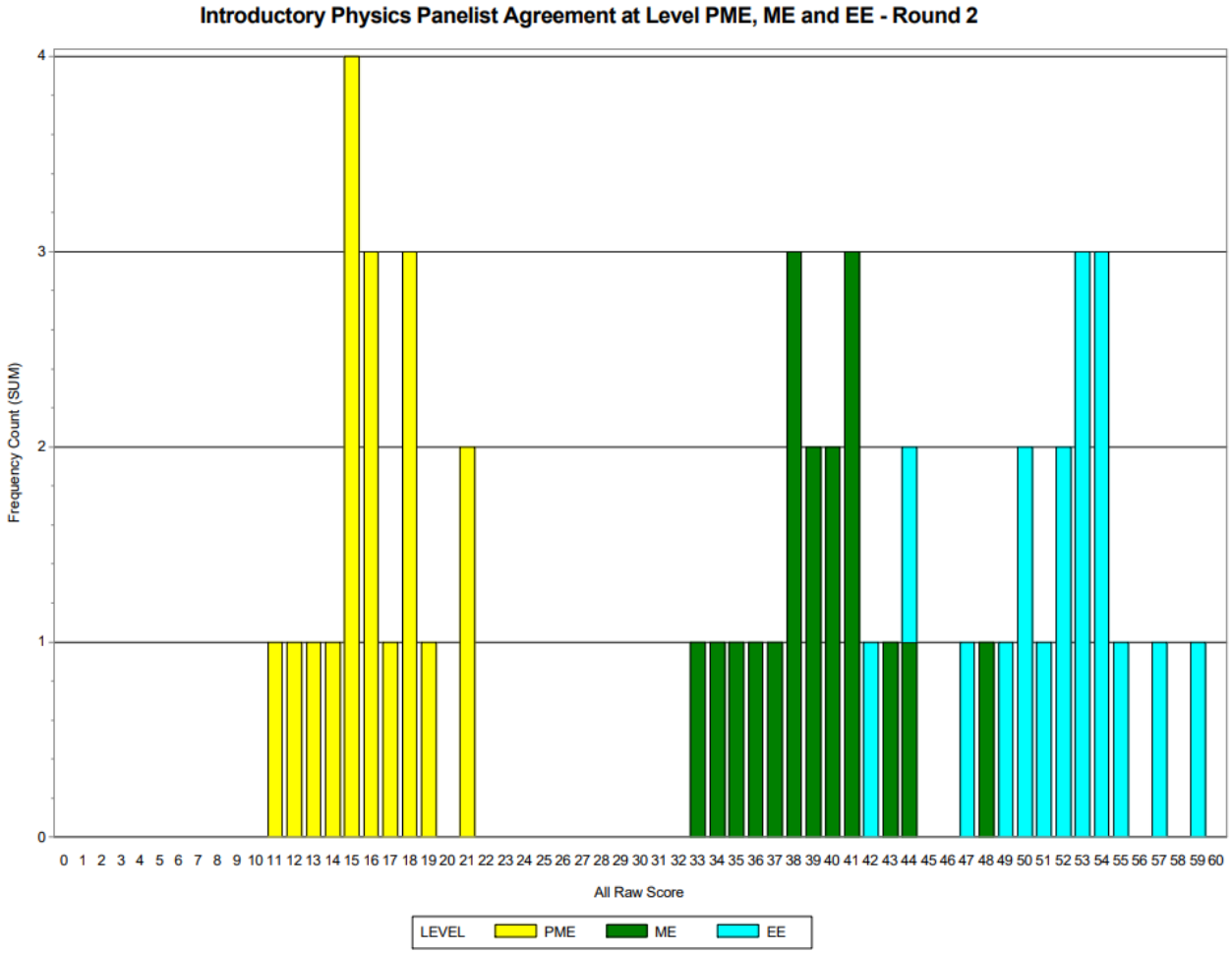
Physics

Round 1:



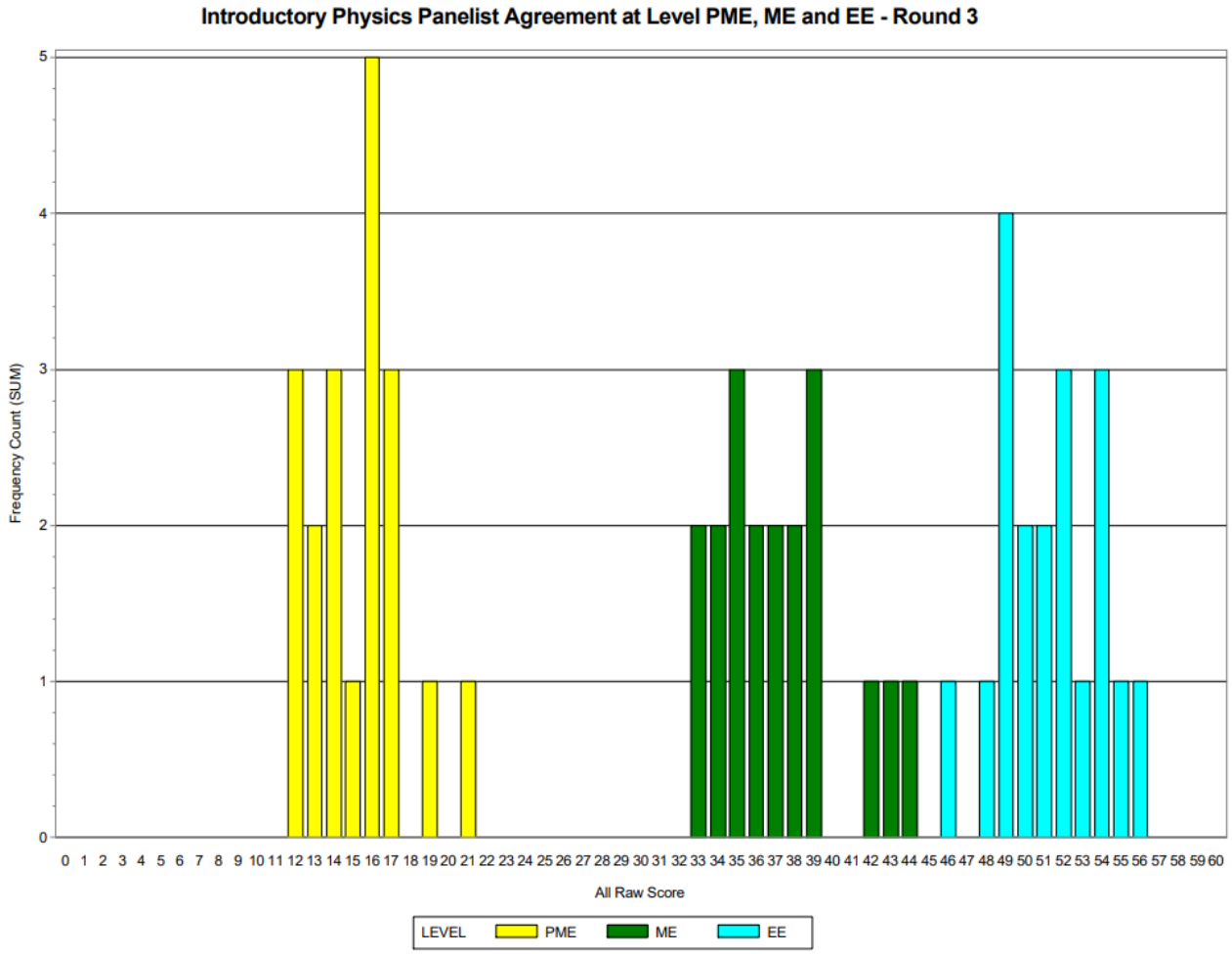
All Three Achievement Levels Concurrently

Round 2:



All Three Achievement Levels Concurrently

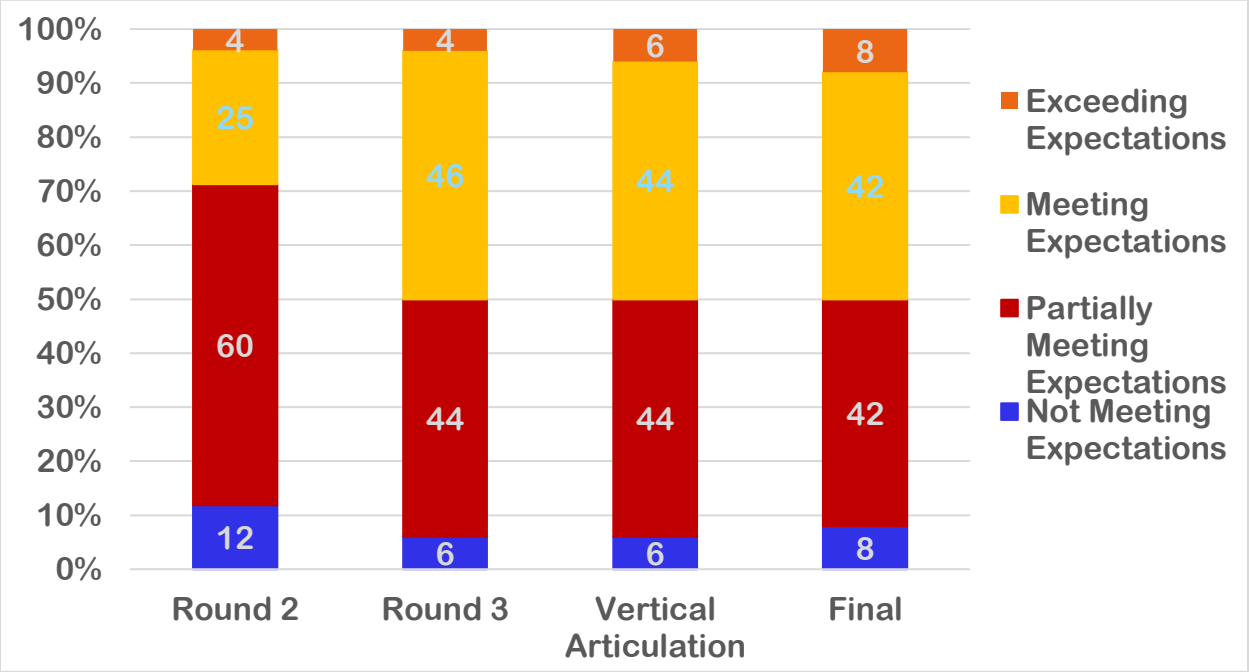
Round 3:



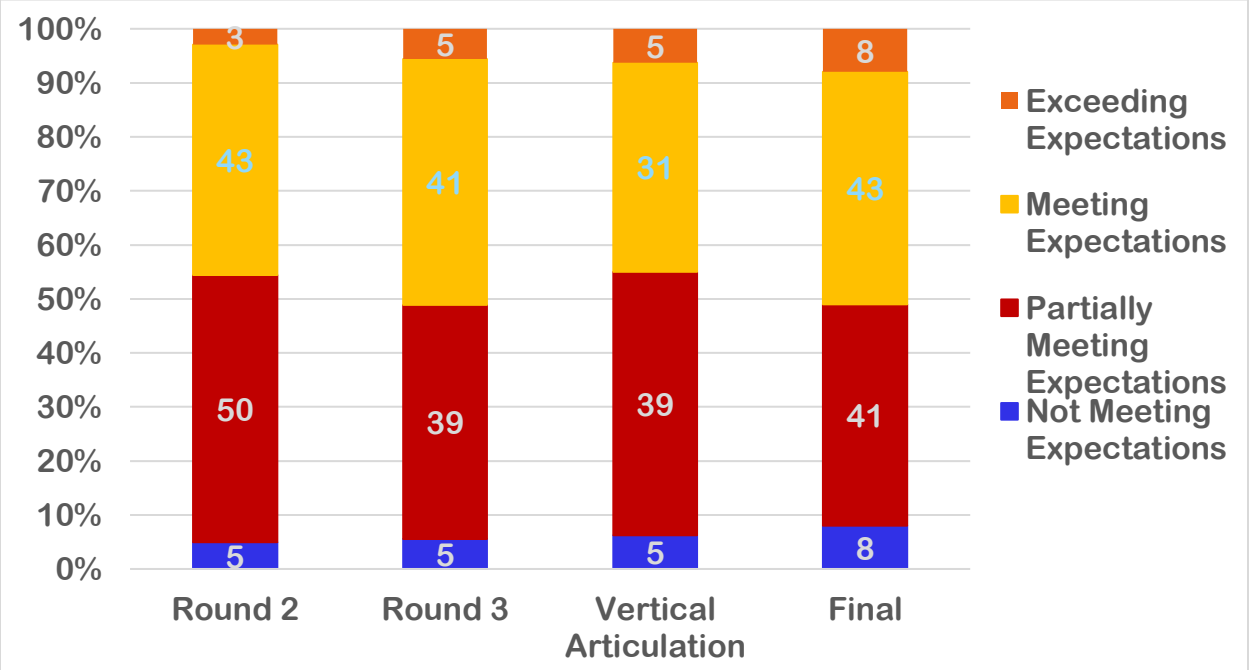
All Three Achievement Levels Concurrently

Appendix JJ – Impact Data

Biology



Introductory Physics

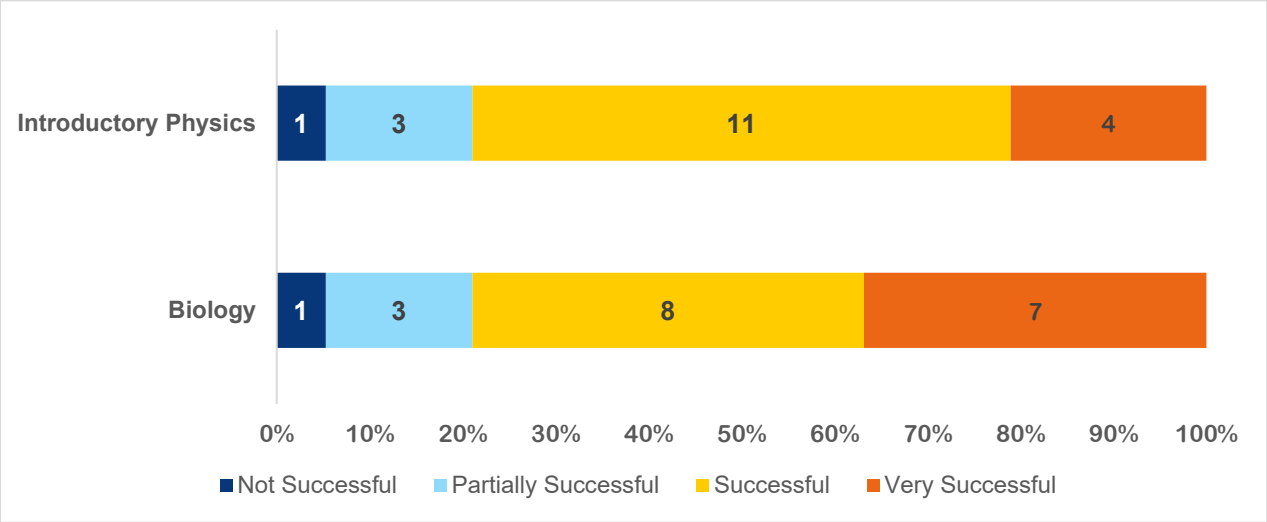


Appendix KK – Participant Evaluation Results

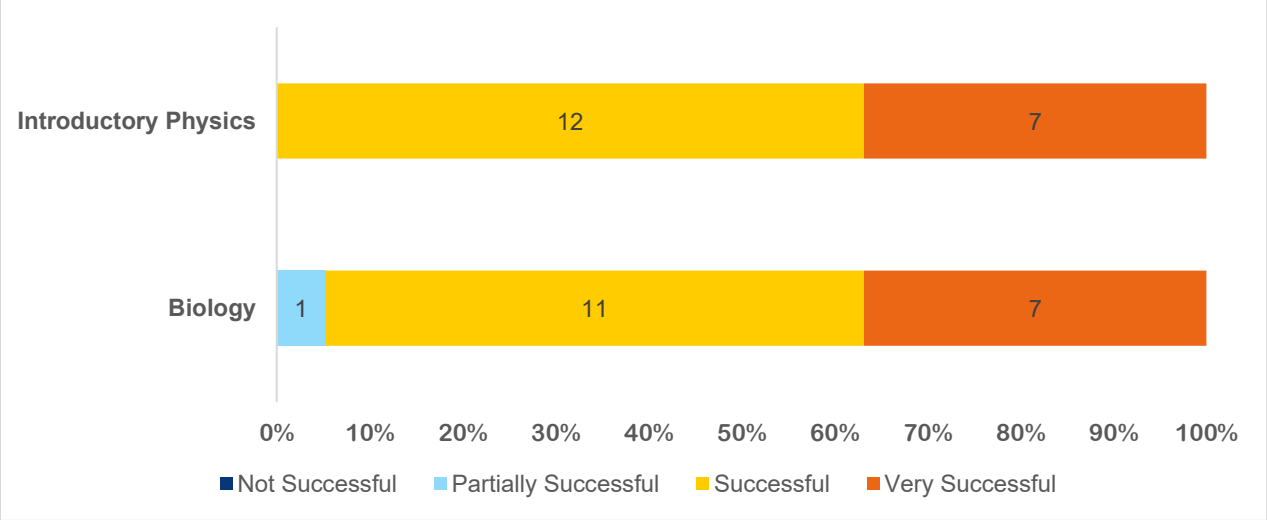
Breakout Session Process Evaluation

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.

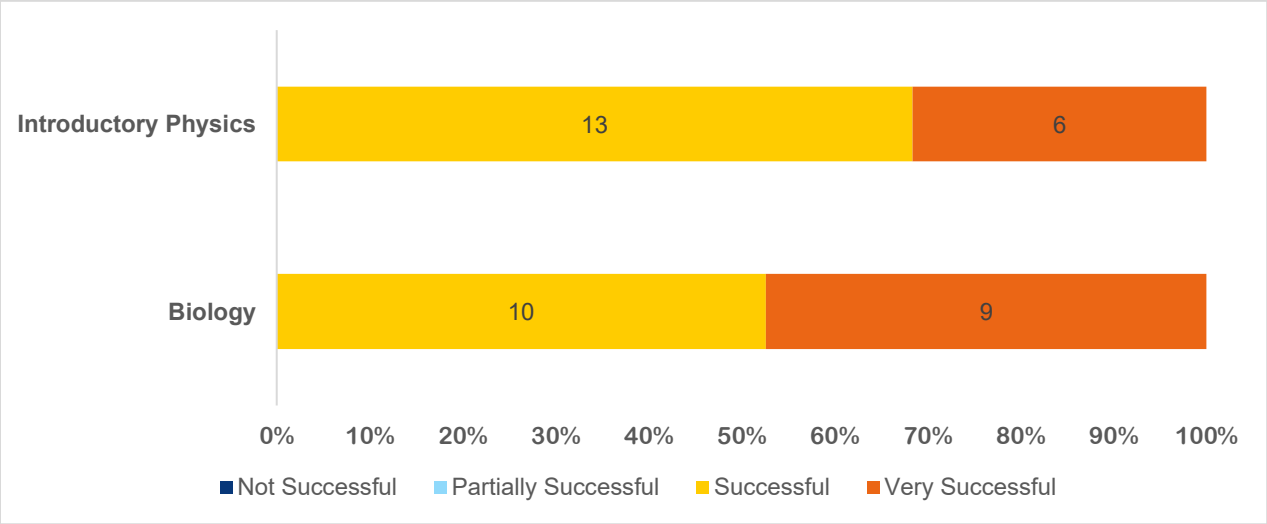
Meeting pre-work



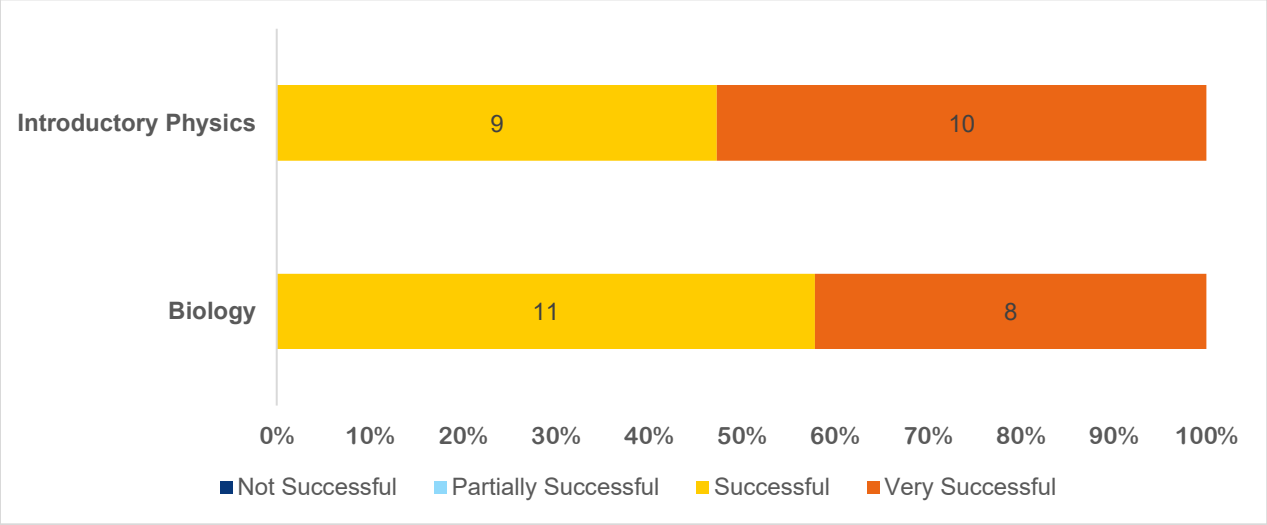
General session training



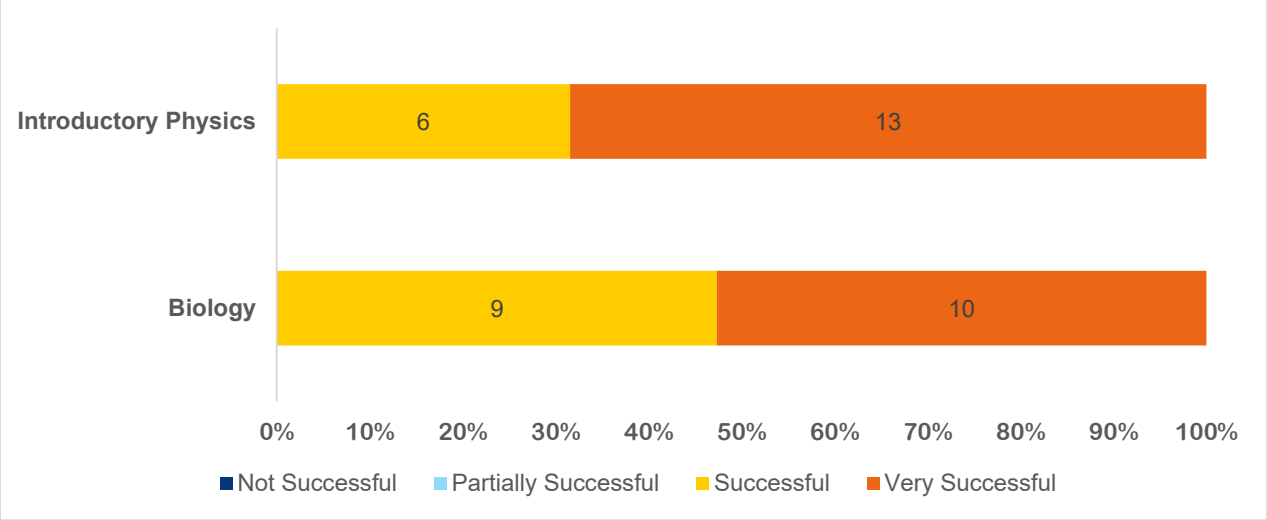
Overview of the MCAS assessments



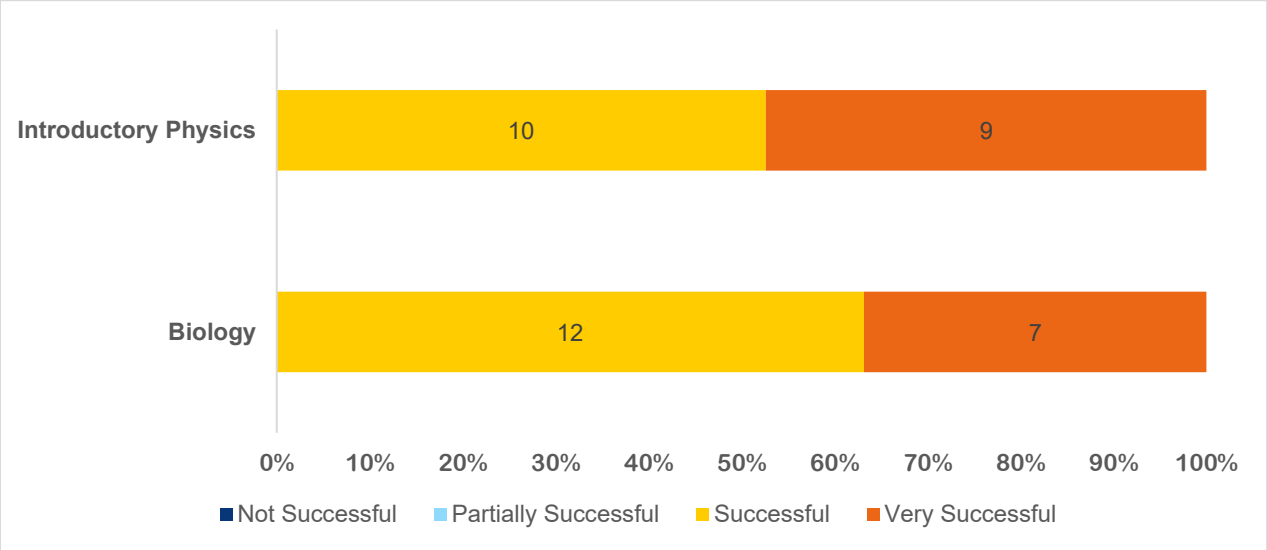
Introduction to the standard setting process



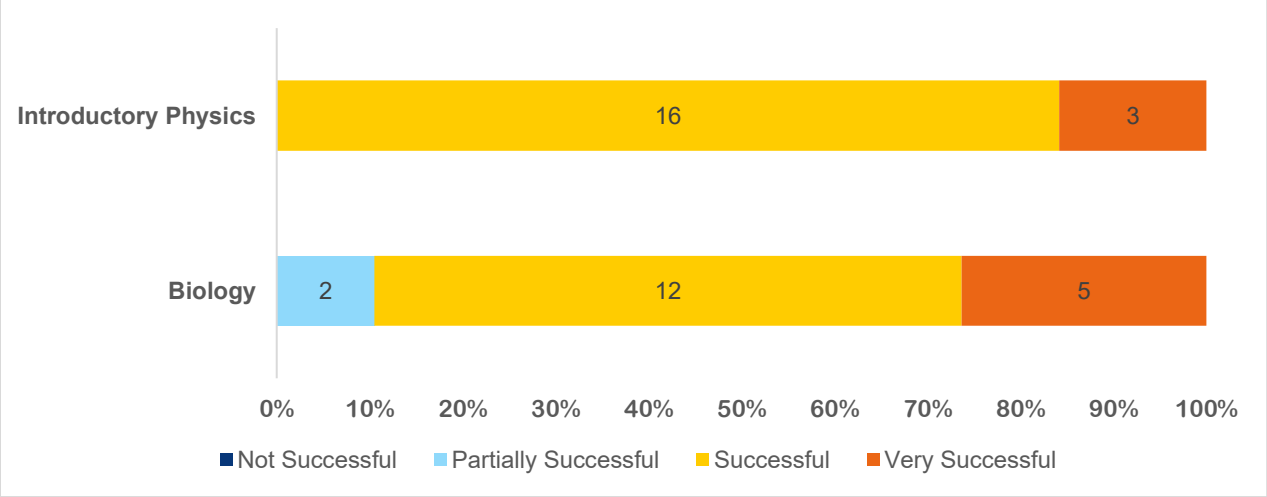
Experiencing the actual assessment



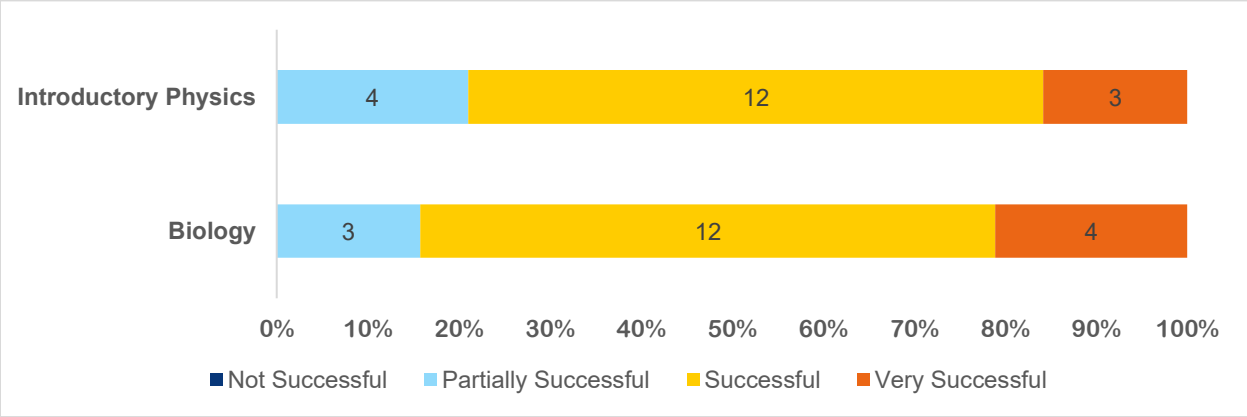
Discussion of the scoring of items on the assessment



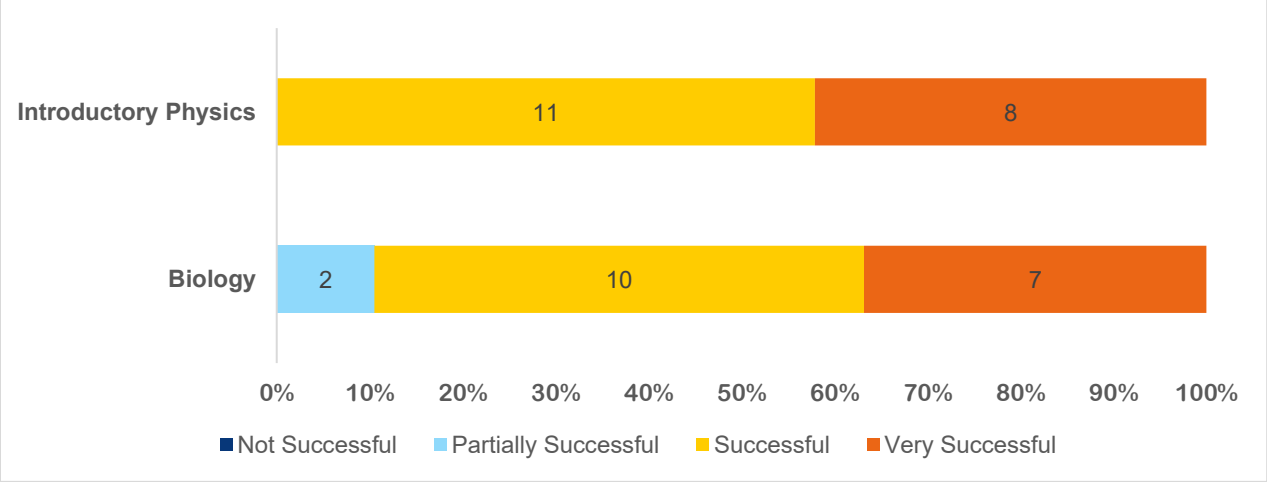
Discussion of achievement level descriptors (ALDs)



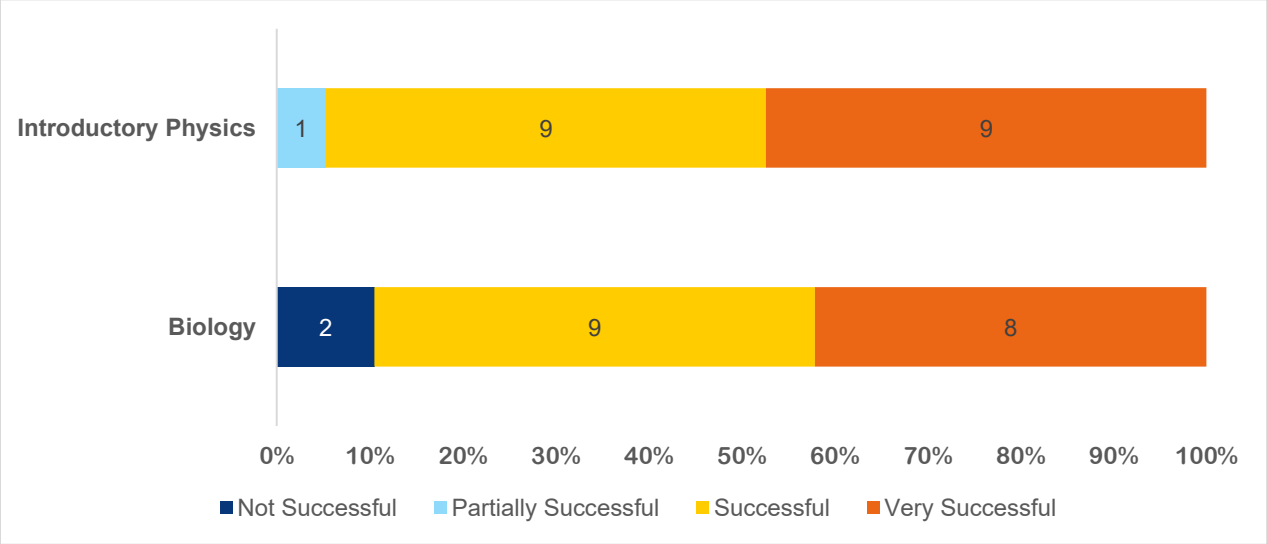
Development and discussion of the borderline descriptions



Overview of the standard-setting procedure

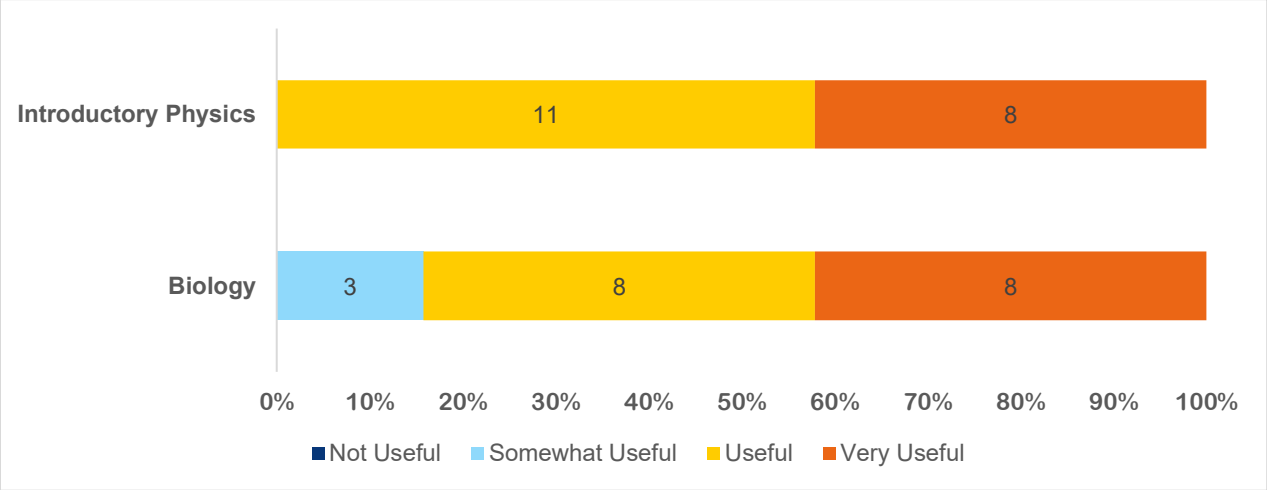


Practice exercise for the standard-setting procedure

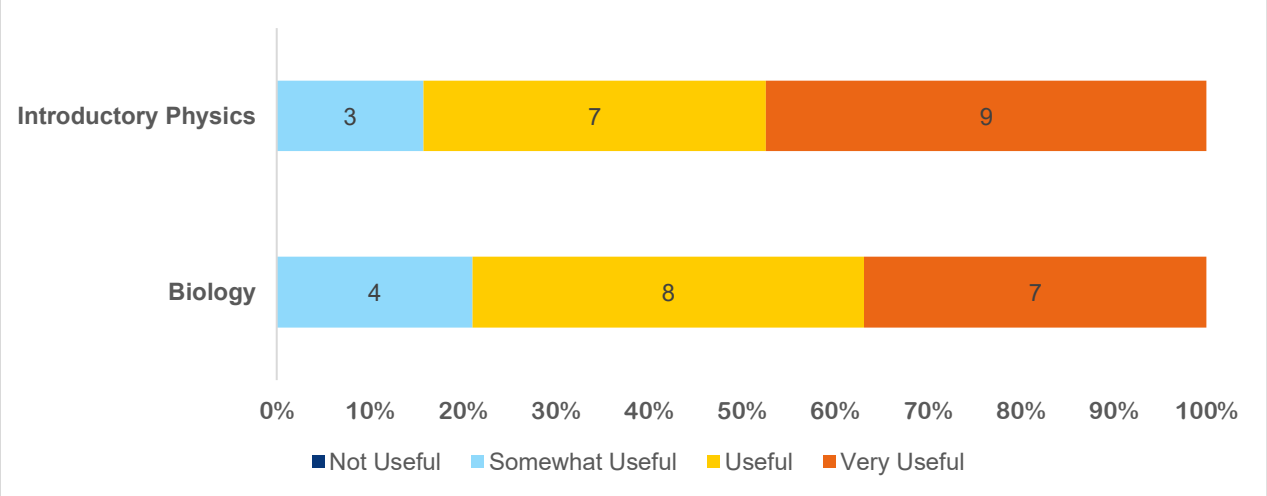


Question 2: How useful do you feel the following activities or information were in assisting you to make your recommendations?

Achievement Level Descriptors (ALDs)

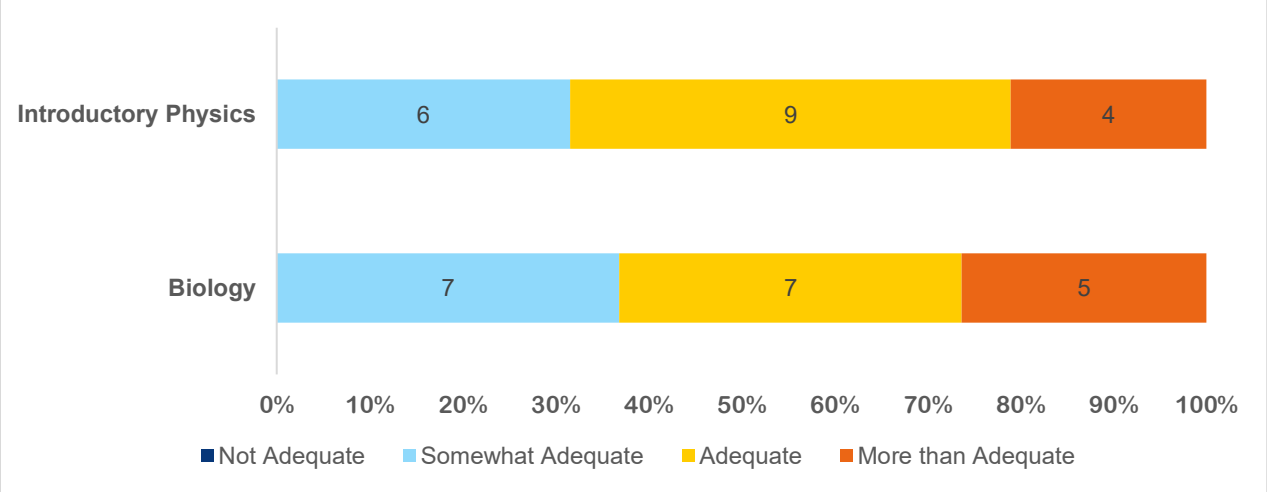


Borderline Descriptions

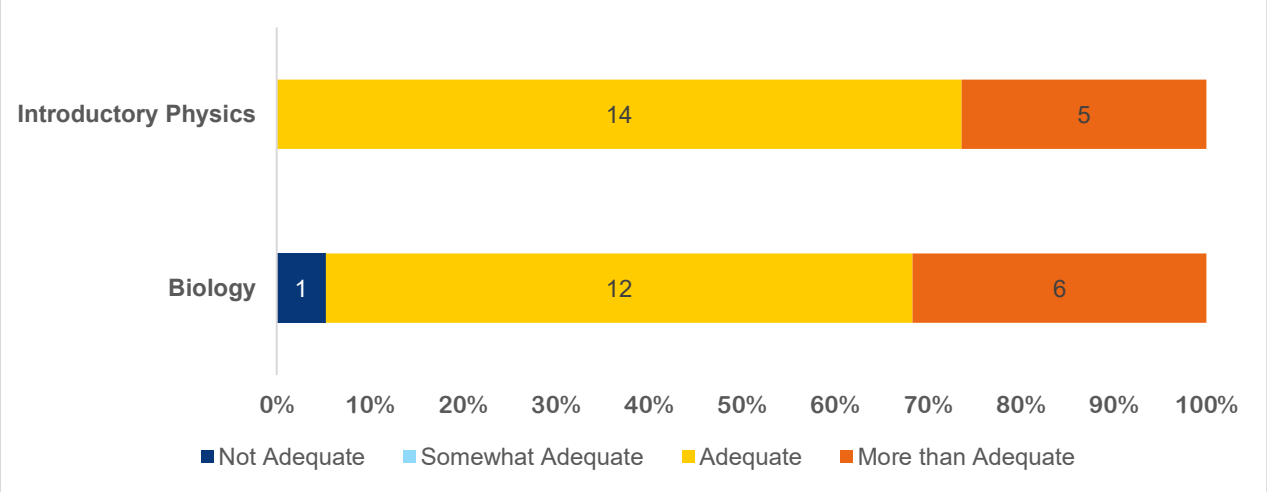


Question 3: How adequate were the following elements of the session?

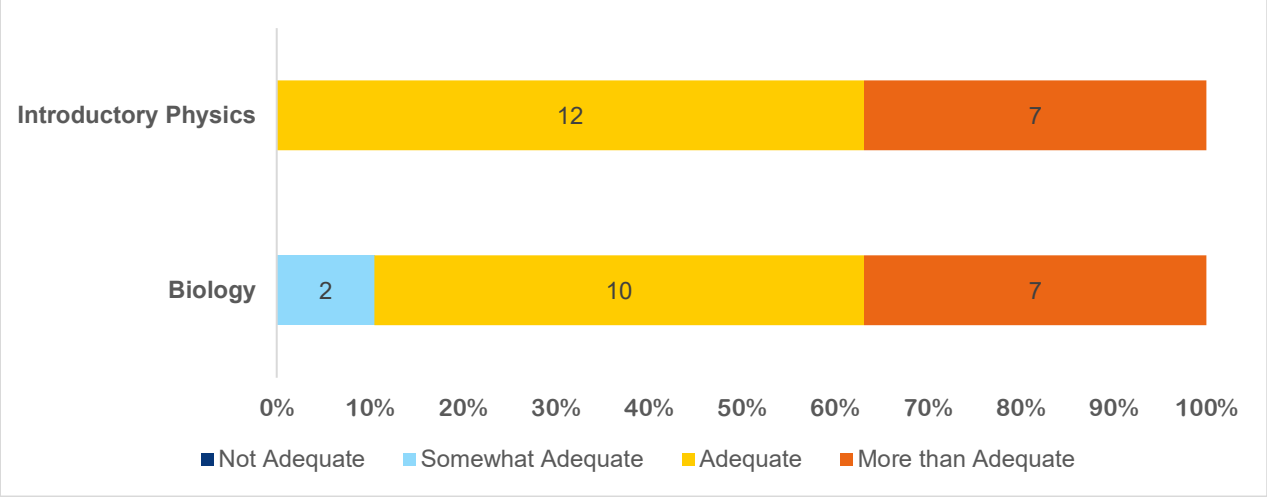
Total amount of time to create and discuss borderline descriptions



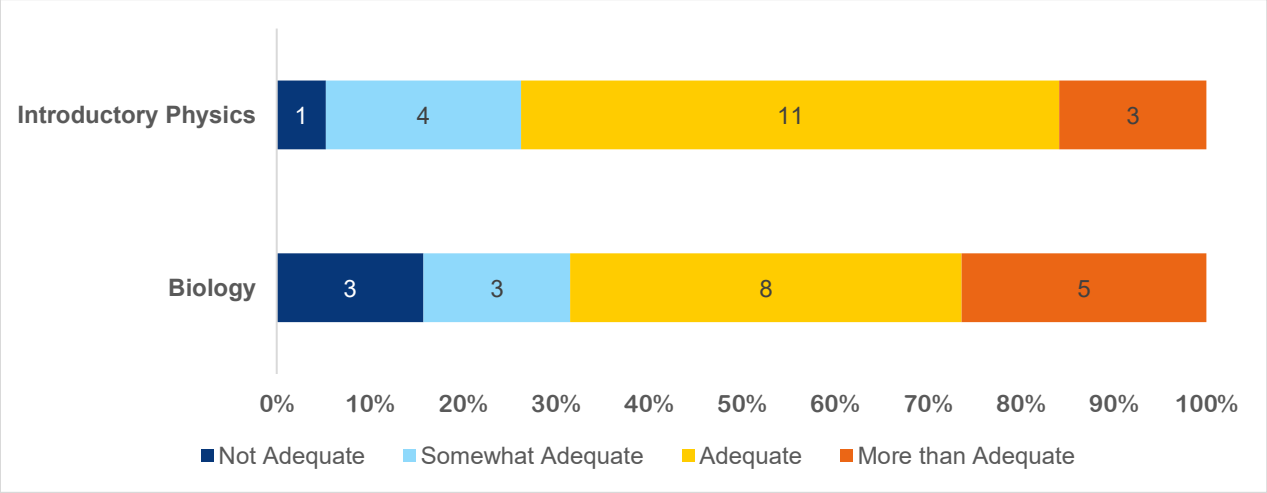
Training provided on the standard-setting process



Amount of time spent training

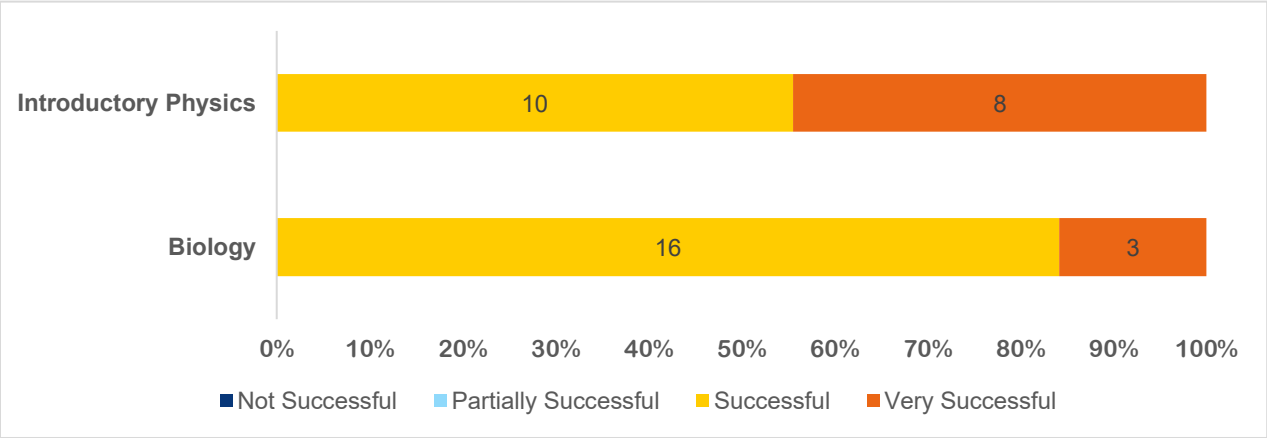


Total amount of time to discuss the practice judgment activity

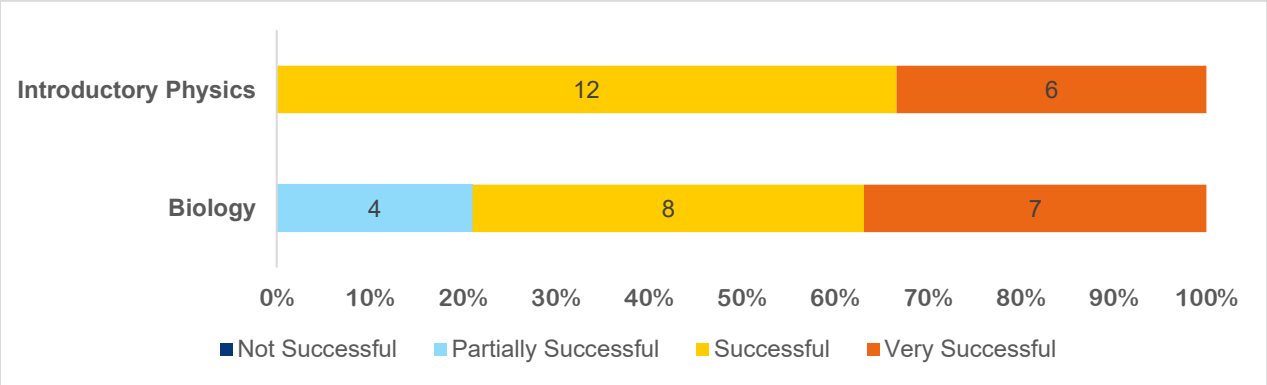


Question 4: Select the option that best reflects your opinion about the level of success of the various components of the meeting in which you participated. The activities were designed to help you both understand the process and be supportive of the recommendations made by the committee.

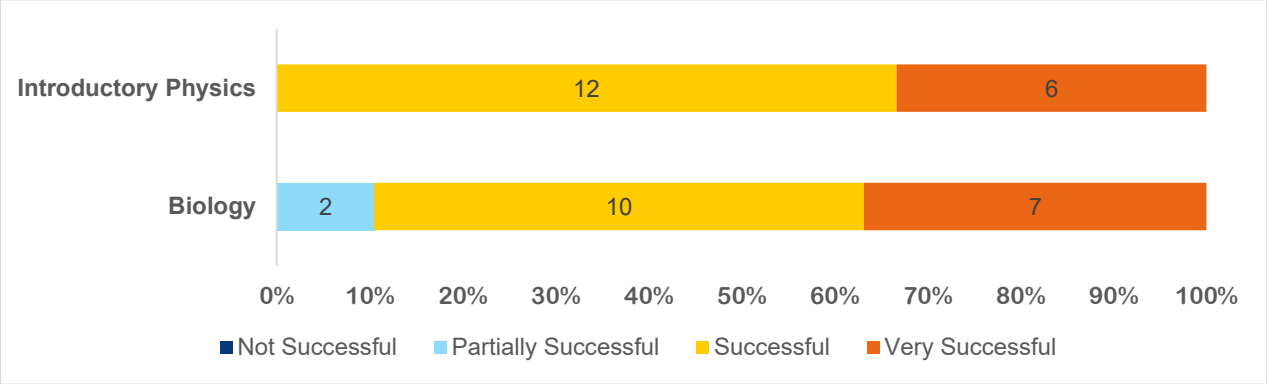
Judgment rounds



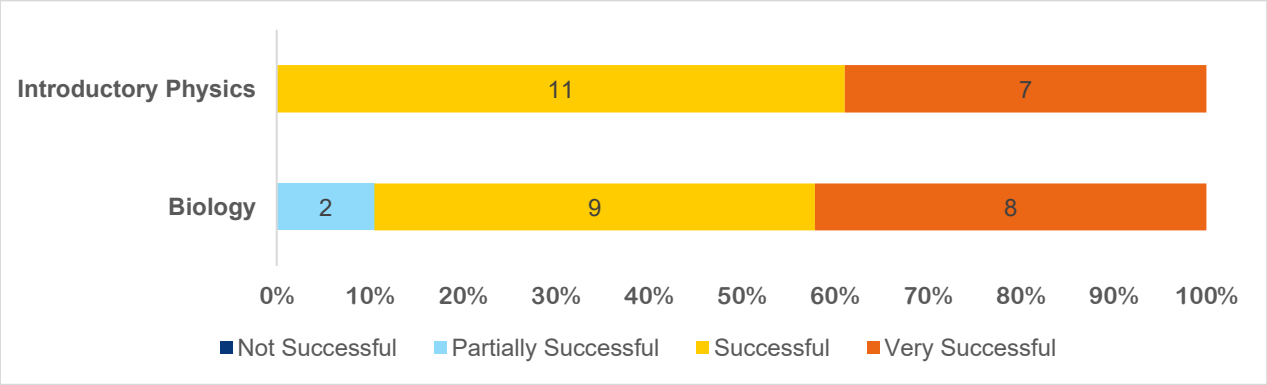
Judgment round feedback – committee-level statistics



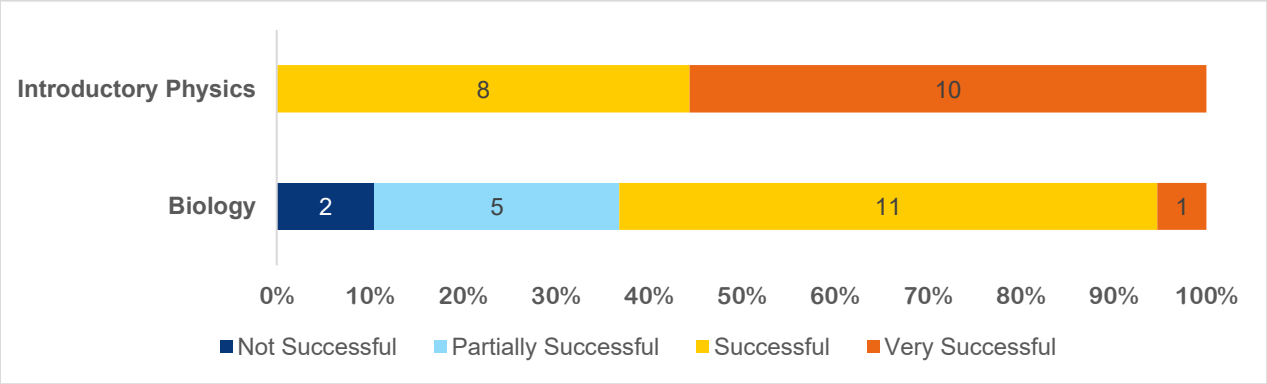
Judgment round feedback – panelist cut score agreement data



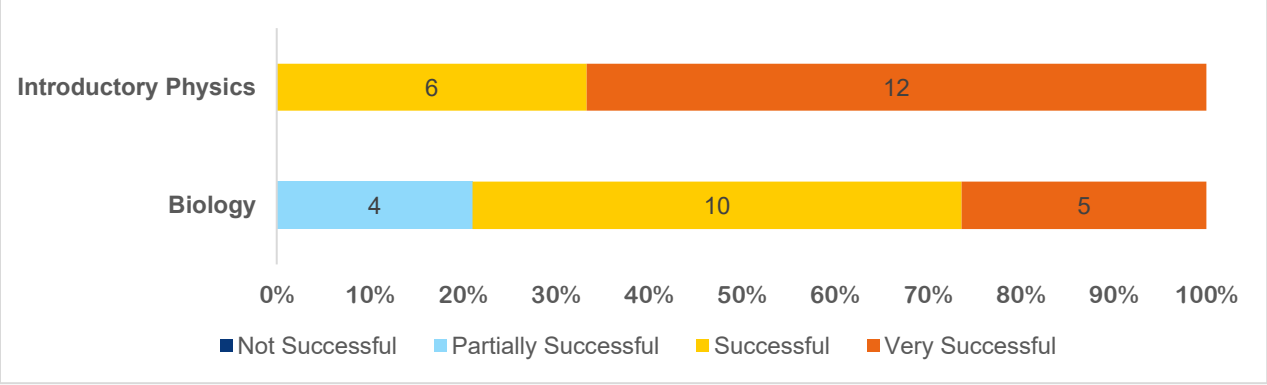
Judgment round feedback – panelist judgment agreement data



Judgment round feedback – impact data

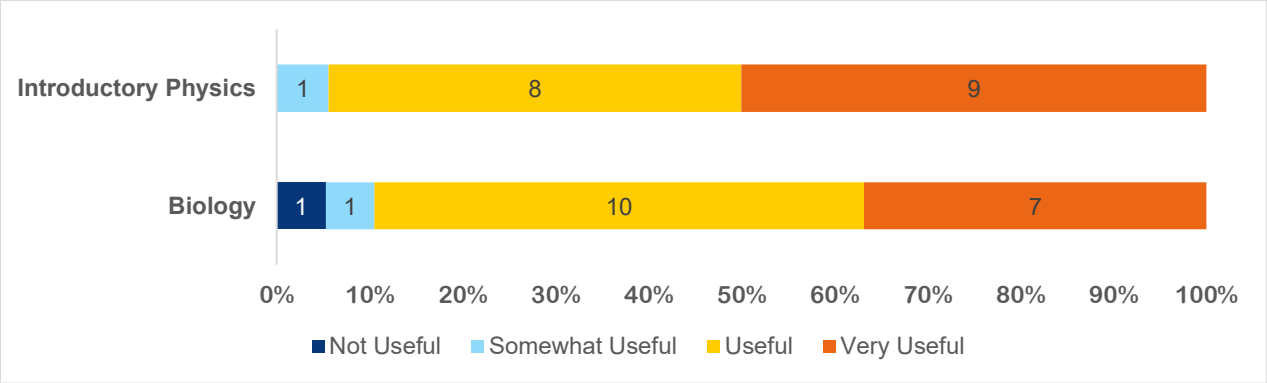


Discussions after each round

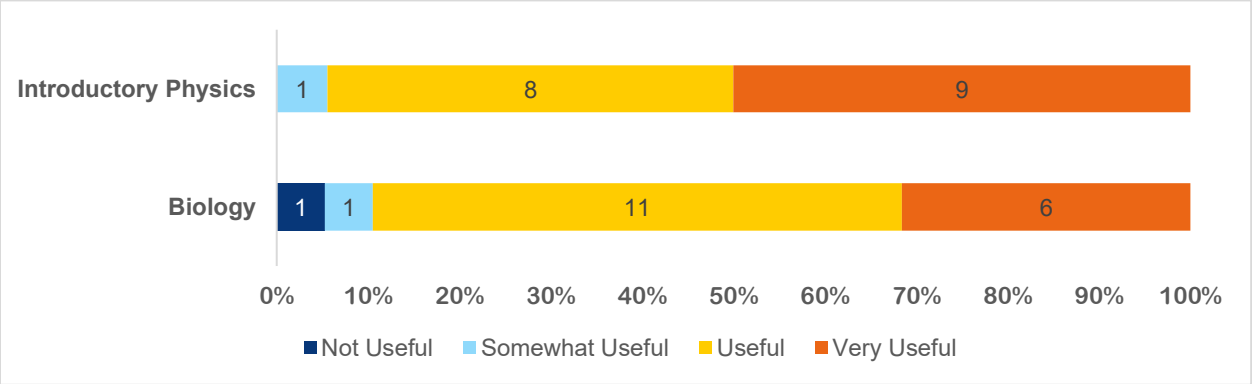


Question 5: How useful do you feel the following activities or information were in assisting you to make your recommendations?

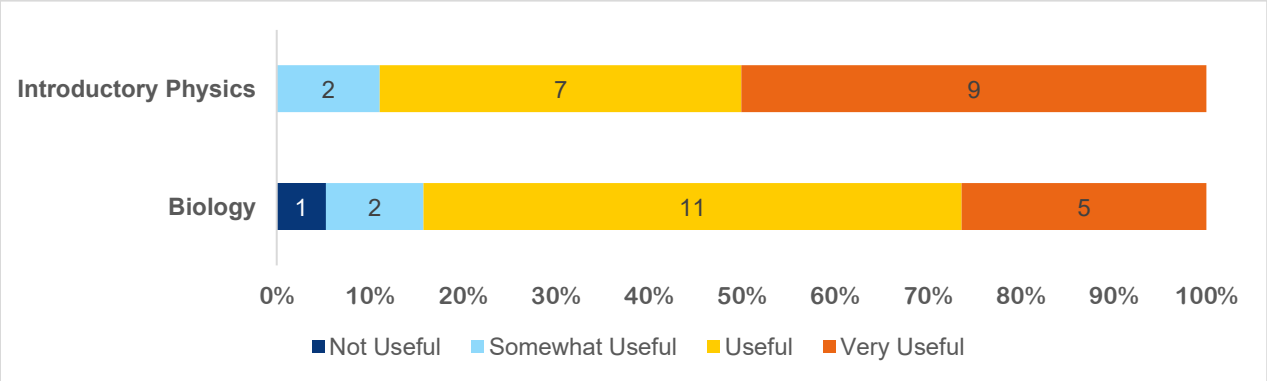
Committee-level statistics after Rounds 1 and 2



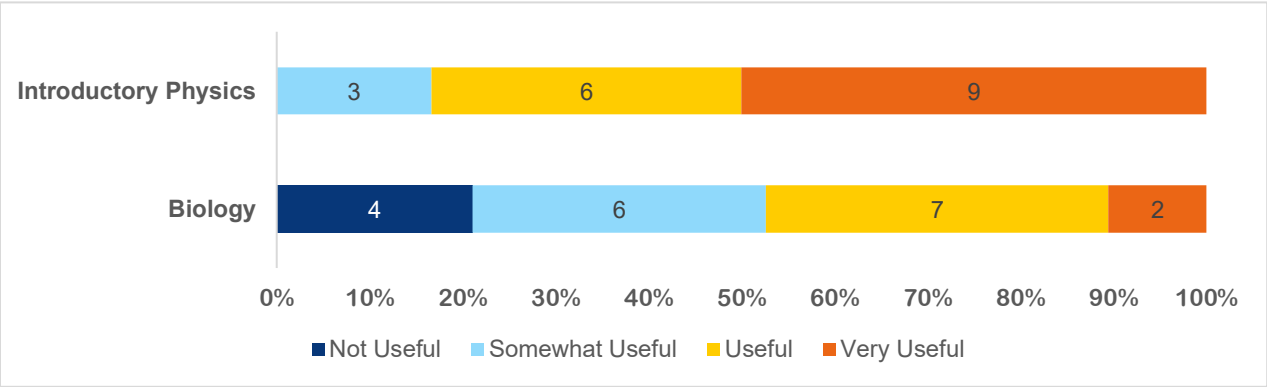
Panelist agreement data provided after Round 1



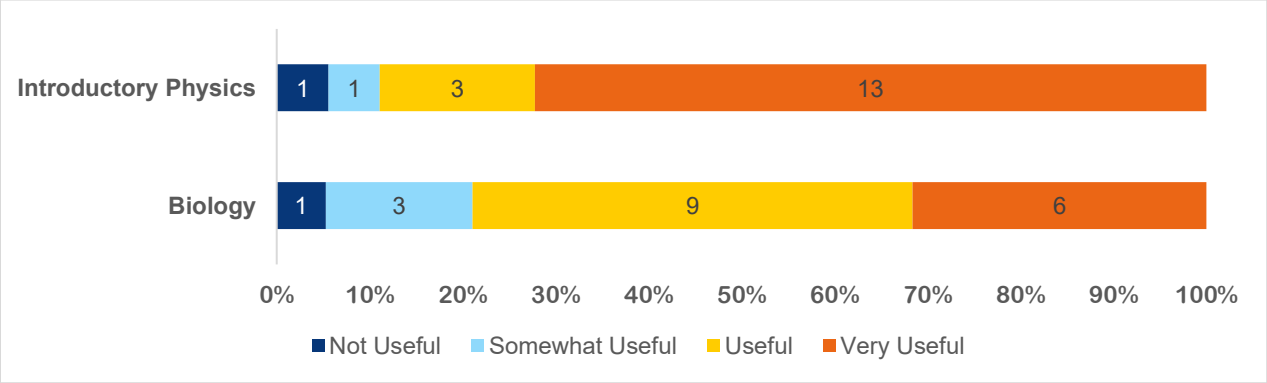
Panelist agreement data provided after Round 2



Impact data after Round 2

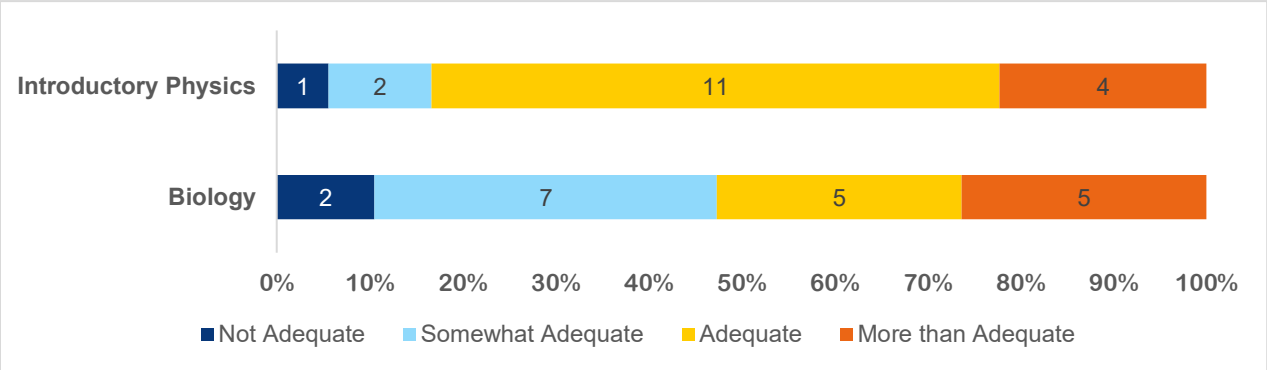


Discussion after each judgment round

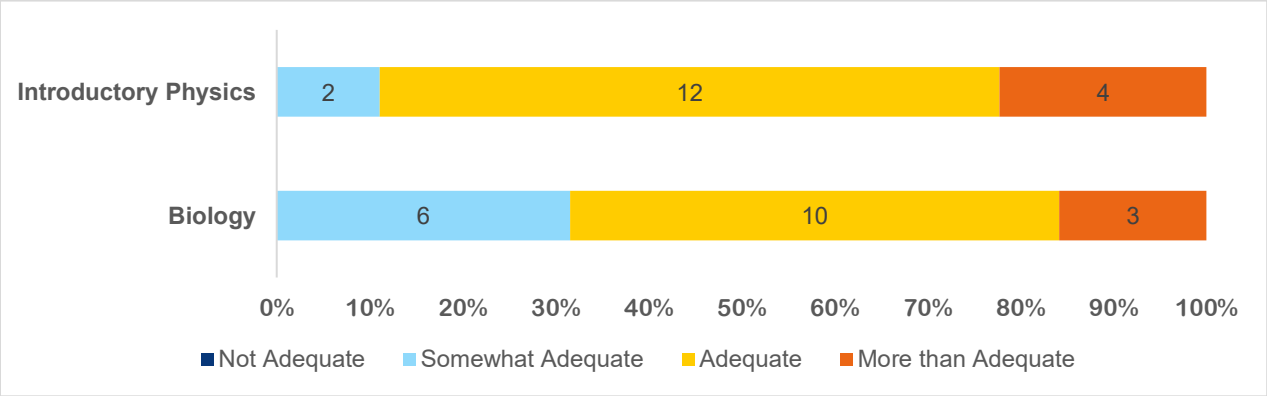


Question 6: How adequate were the following elements of the session?

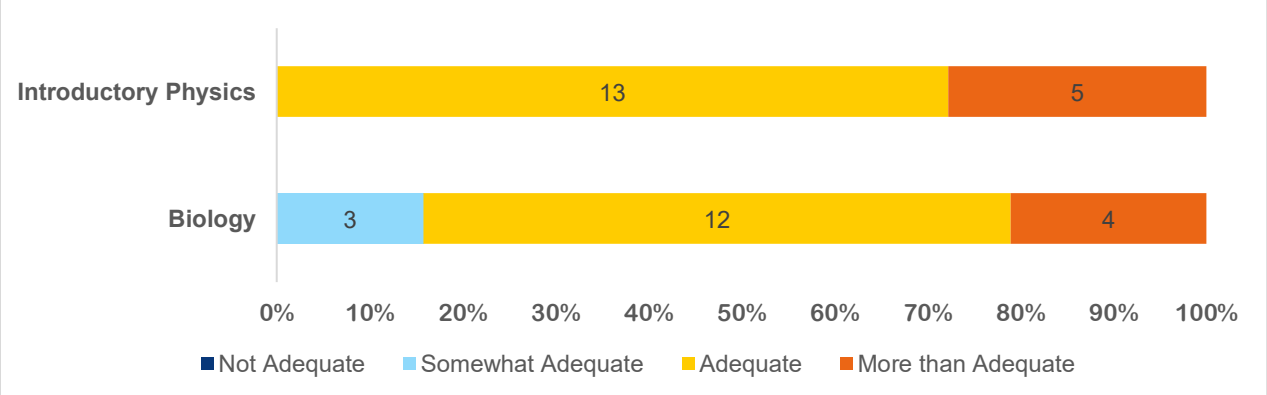
Amount of time to make judgments



Visual presentation of the feedback provided



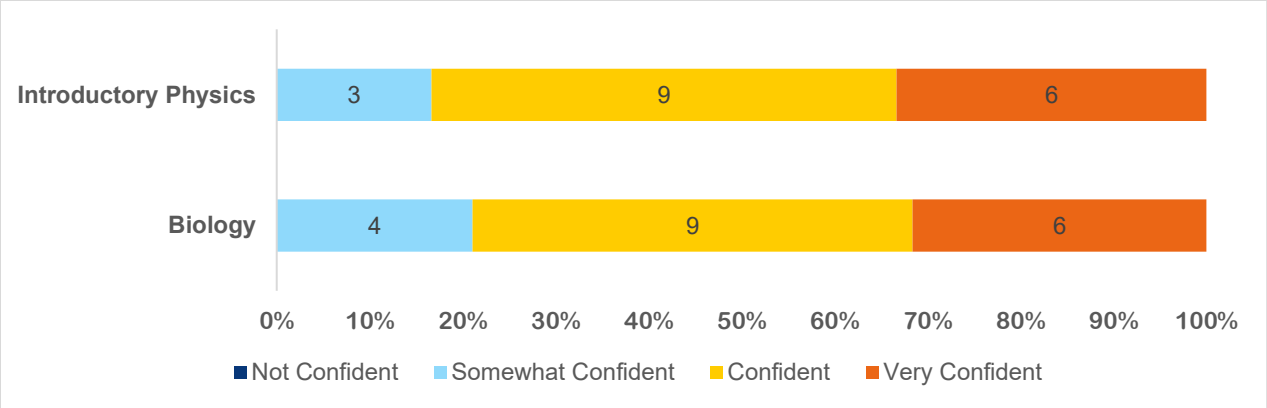
Number of judgment rounds



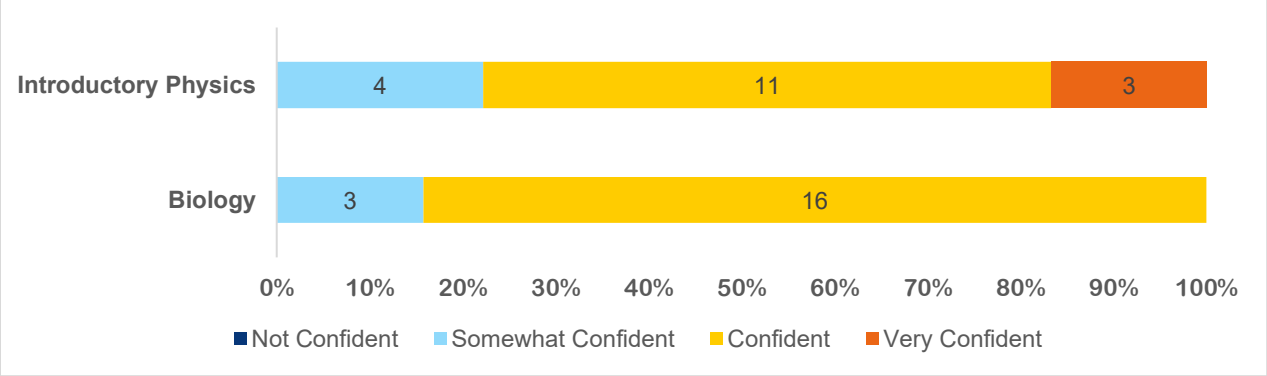
Question 7: In applying the standard-setting method, you were asked to recommend cut scores (separating four achievement levels) for student performance on MCAS assessments.

How confident do you feel that the Achievement Level Descriptors (ALDs) for the specific subject and grade are reasonable for each student achievement level?

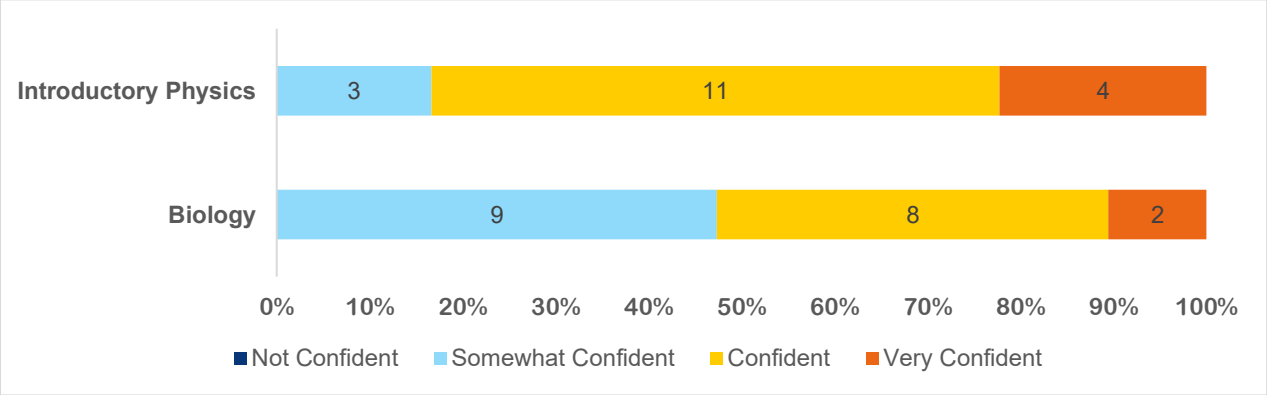
Exceeding Expectations



Meeting Expectations

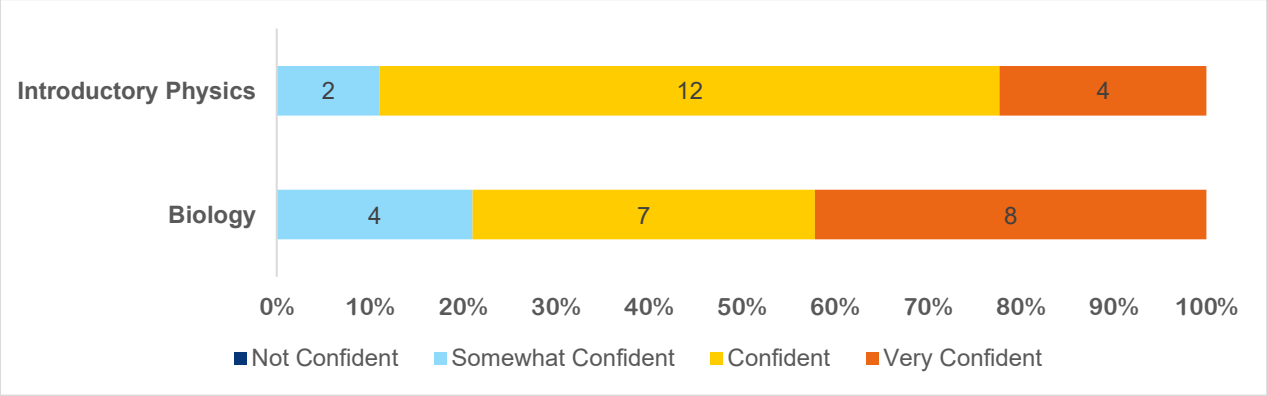


Partially Meeting Expectations

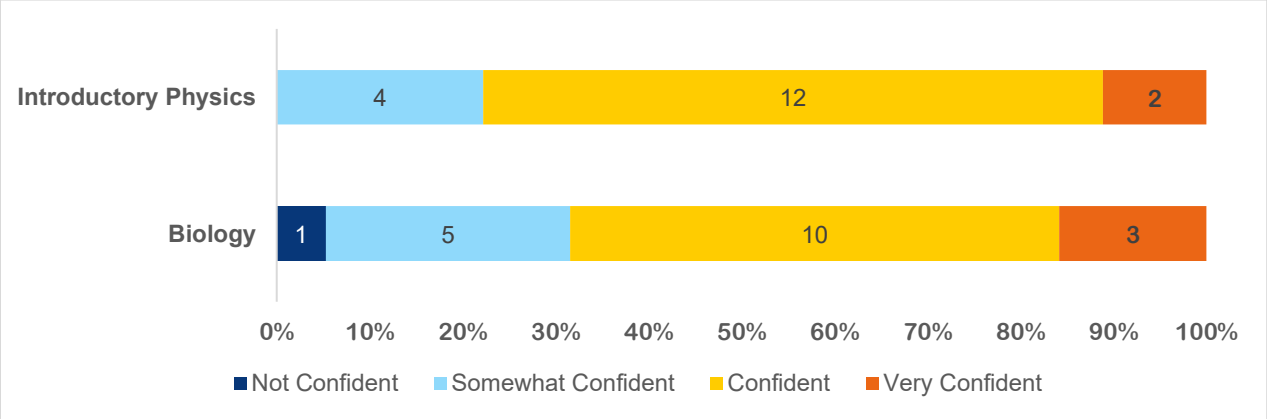


Question 8: How confident do you feel that the final cut score recommendations for the specific subject and grade represent appropriate levels of student performance?

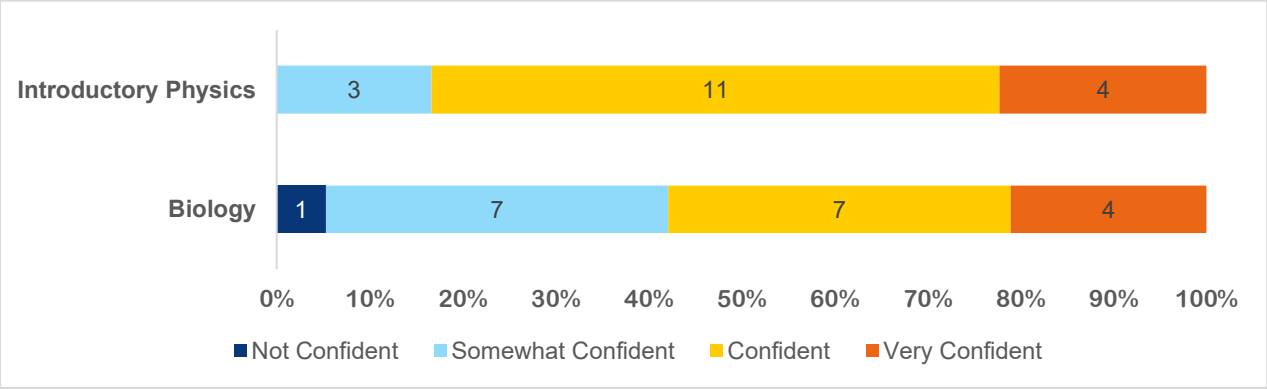
Exceeding Expectations



Meeting Expectations

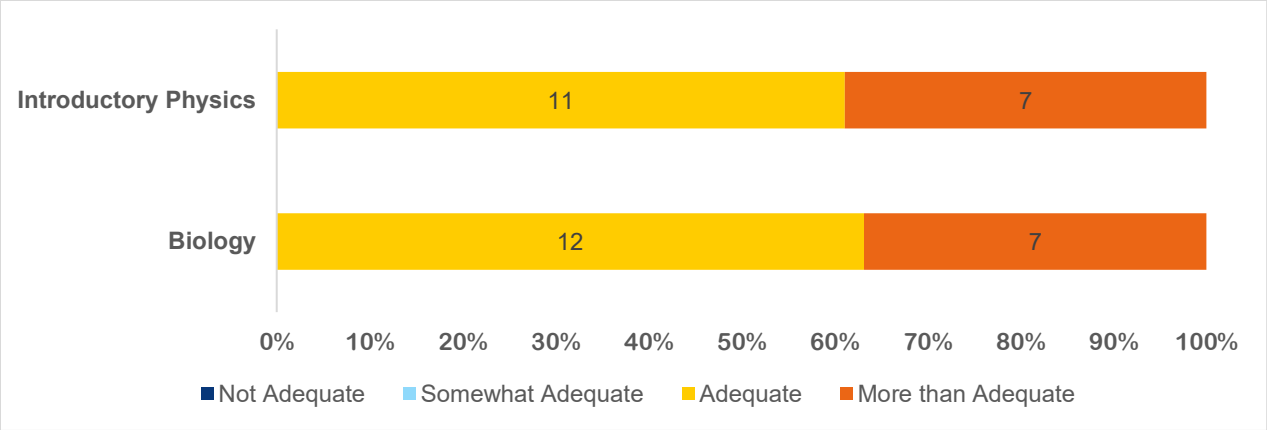


Partially Meeting Expectations

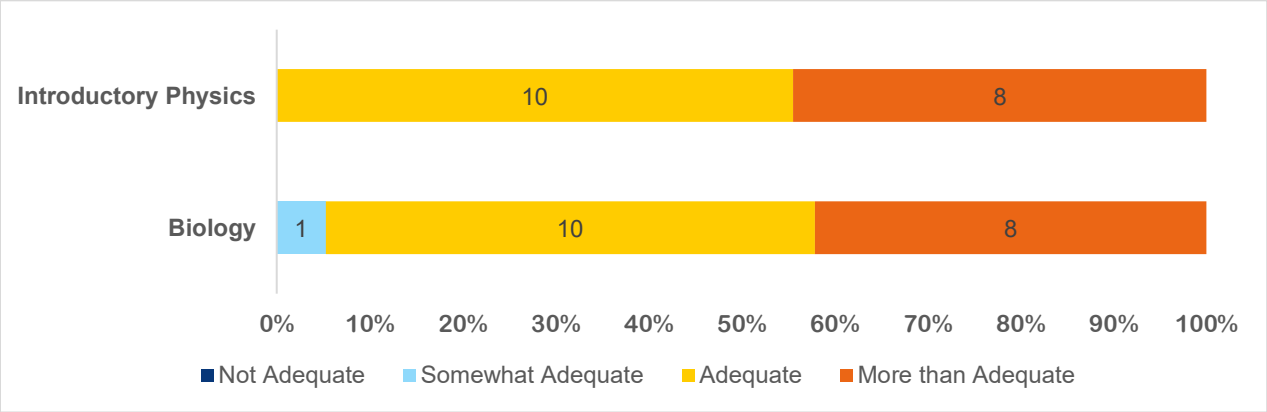


Question 9: How adequate were the following elements of the session?

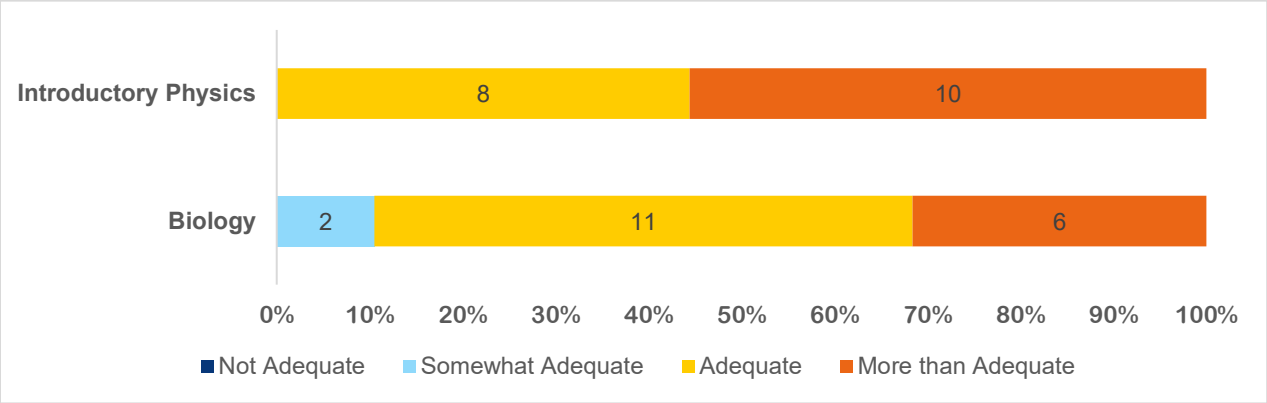
Facilities used for the general session



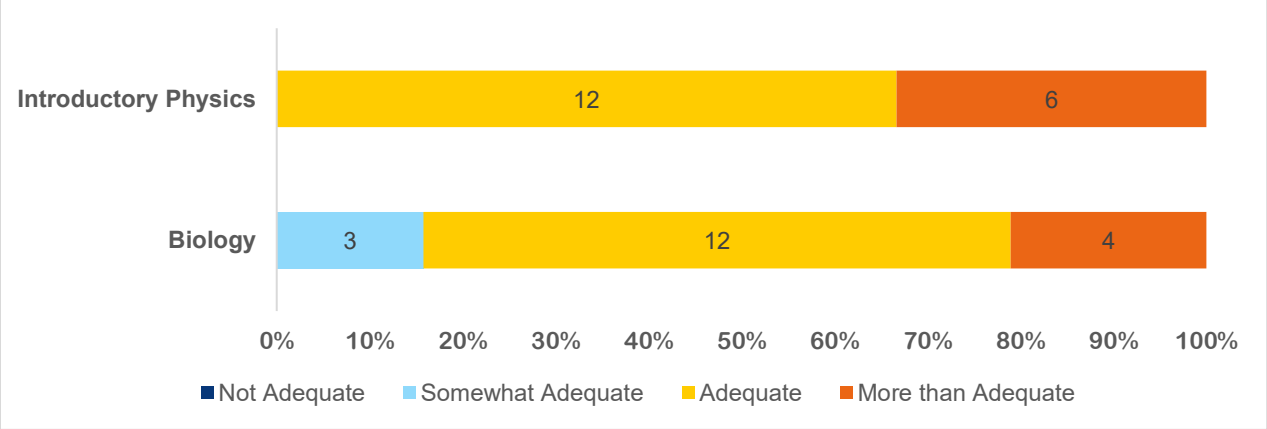
Facilities used for the breakout session



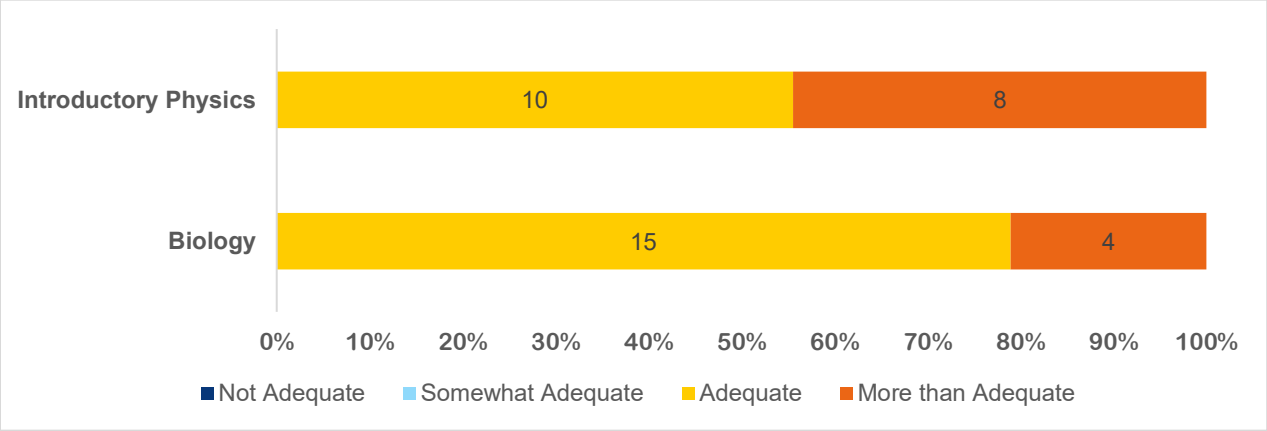
Computers used during the meetings



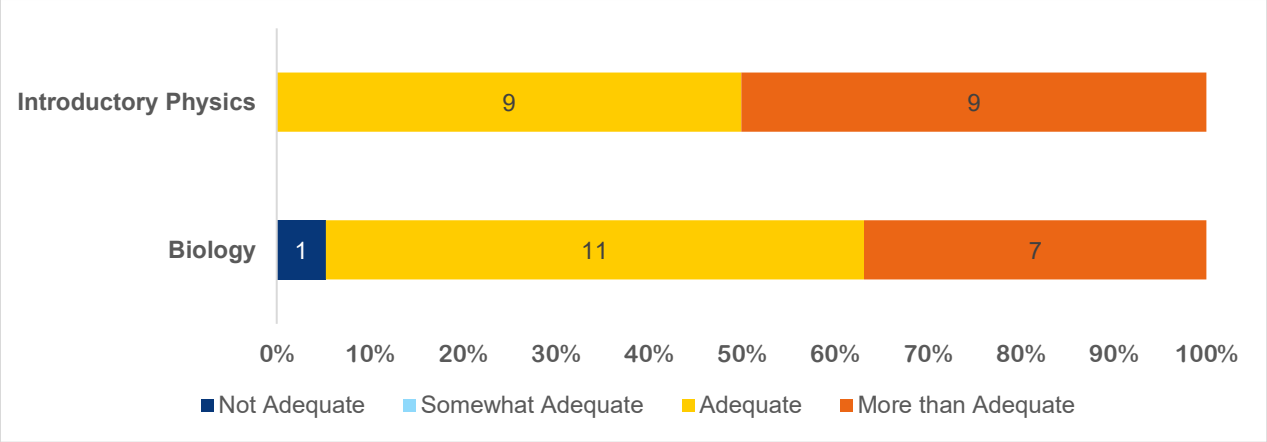
Standard Setting website for accessing materials and making judgments



Materials provided in the folder

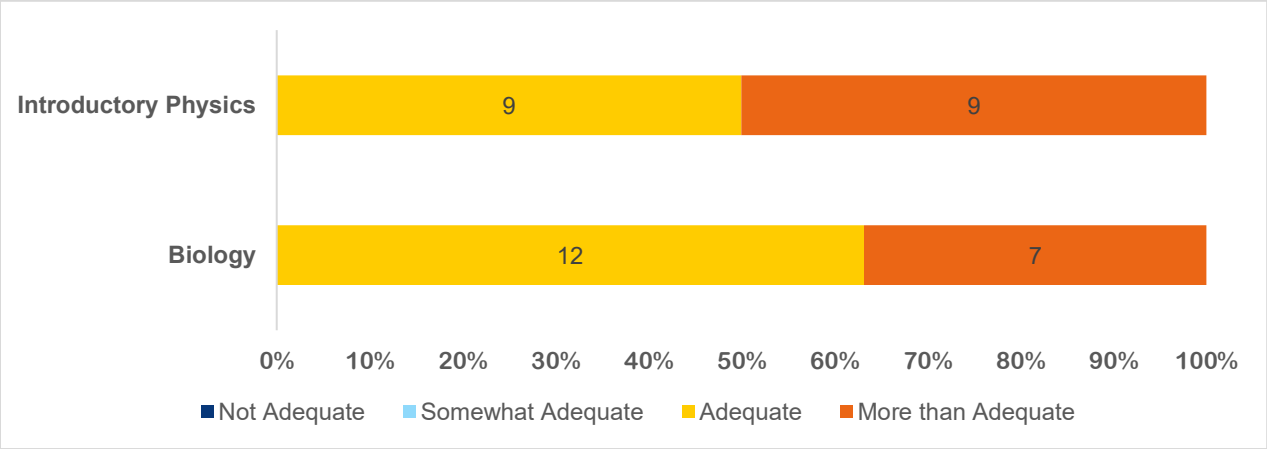


Work space in table groups during the meeting

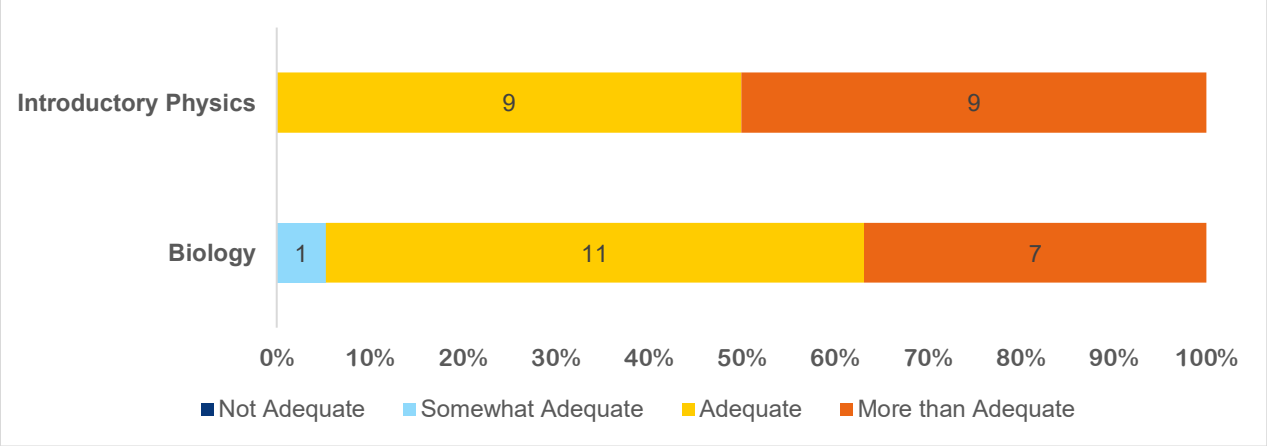


Question 10: Did you have adequate opportunities during the session to:

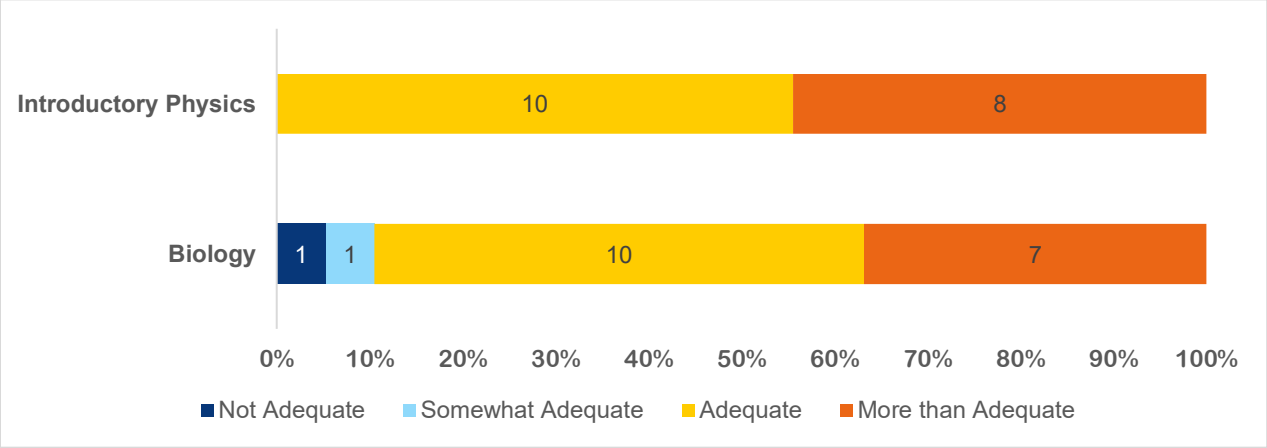
Express your opinions about student achievement levels



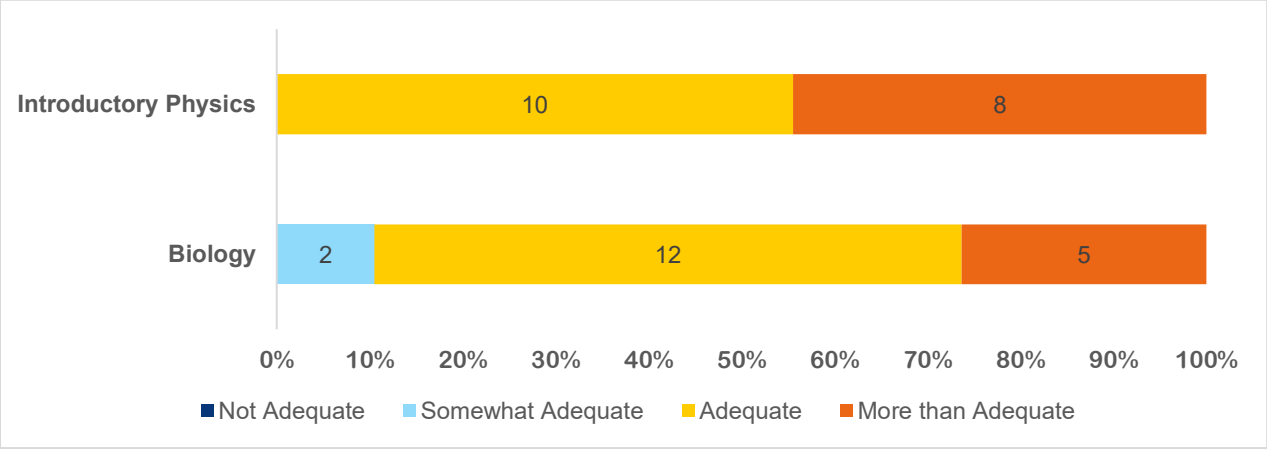
Ask questions about the cut scores and how they will be used



Ask questions about the process of making cut score recommendations

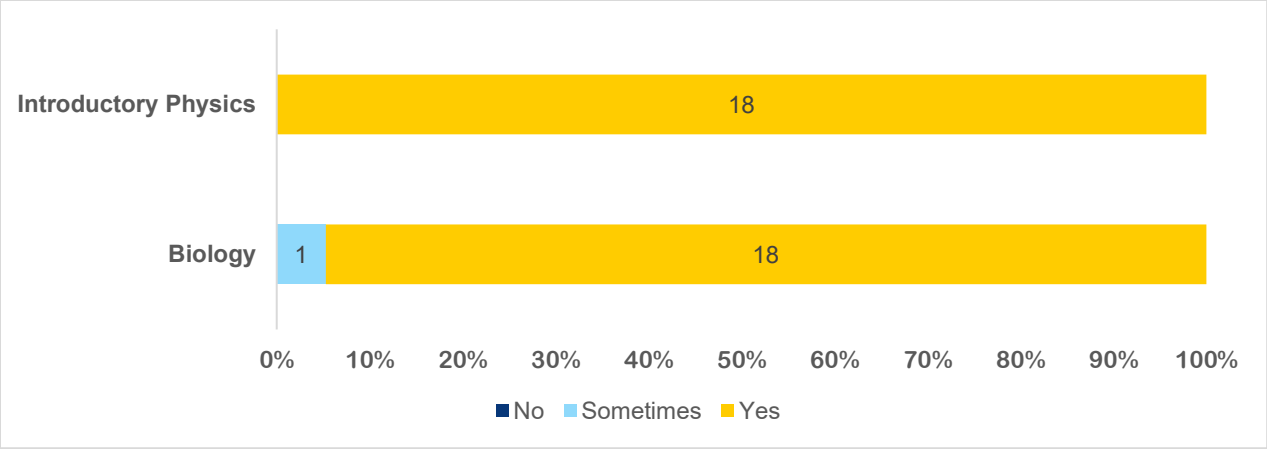


Interact with your fellow panelists

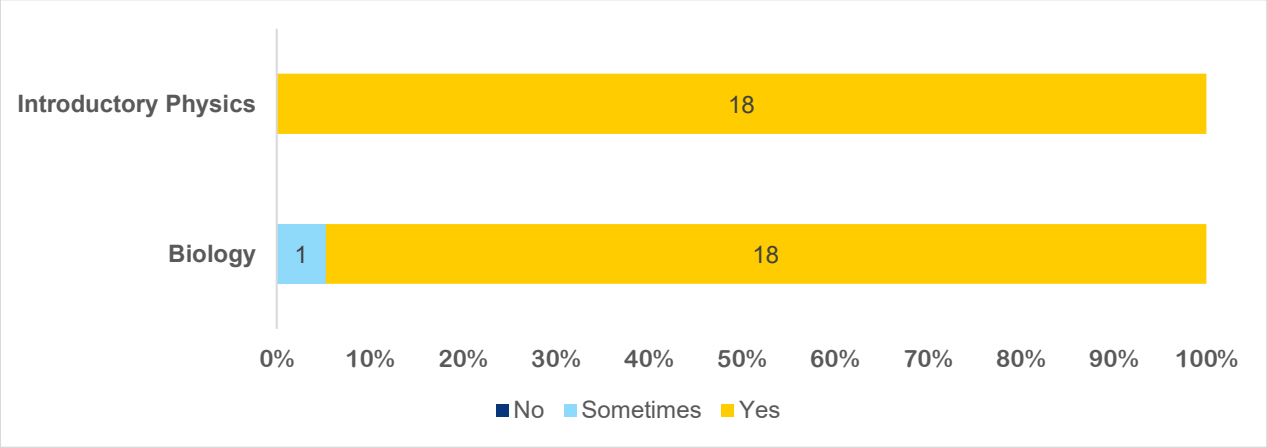


Question 11: Do you believe your opinions and judgments were treated with respect by:

Fellow panelists



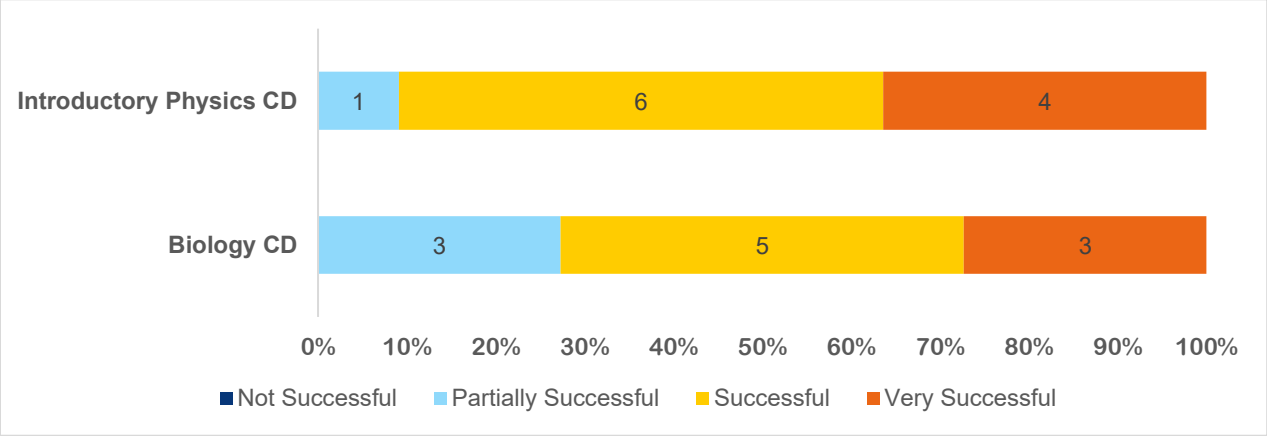
Facilitators



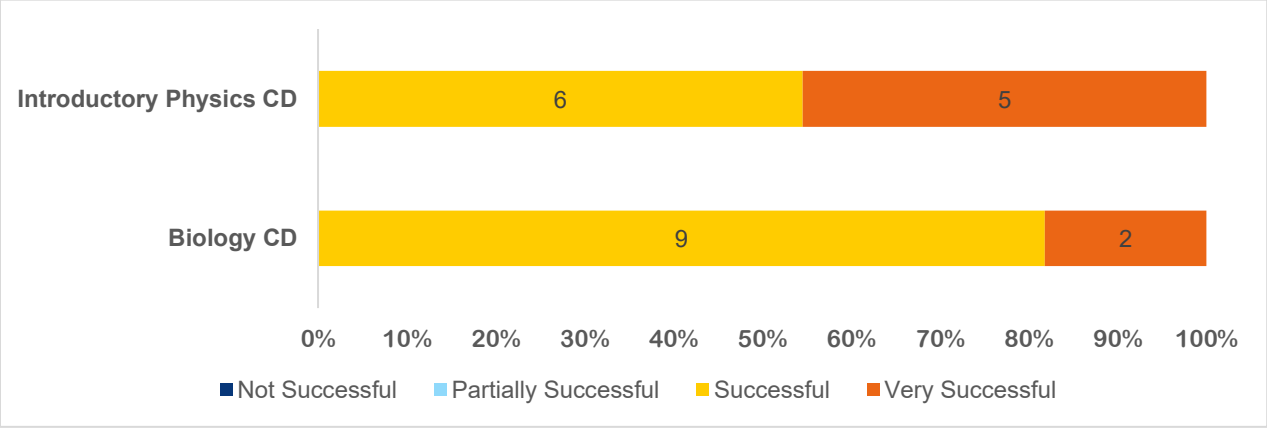
CD Validation Session Process Evaluation

Q.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.

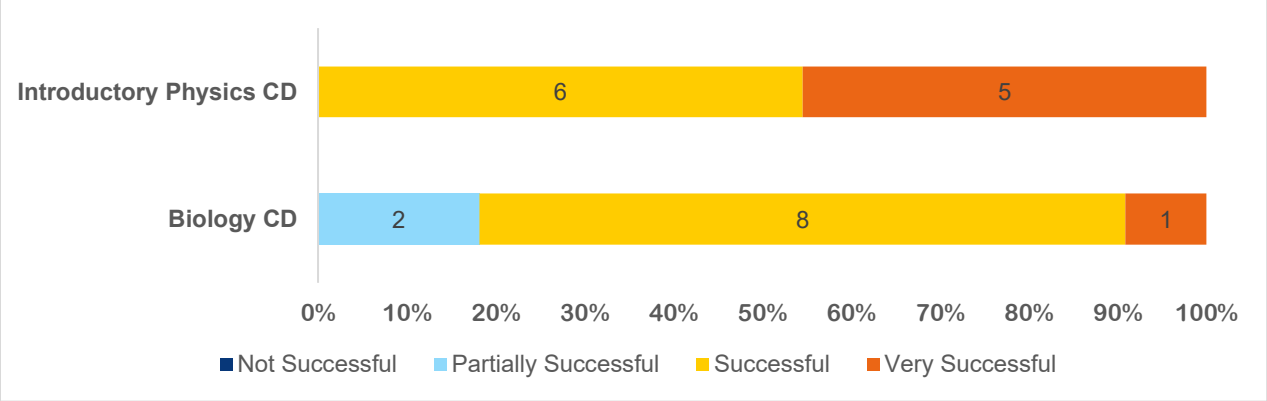
Introduction to CD validation process



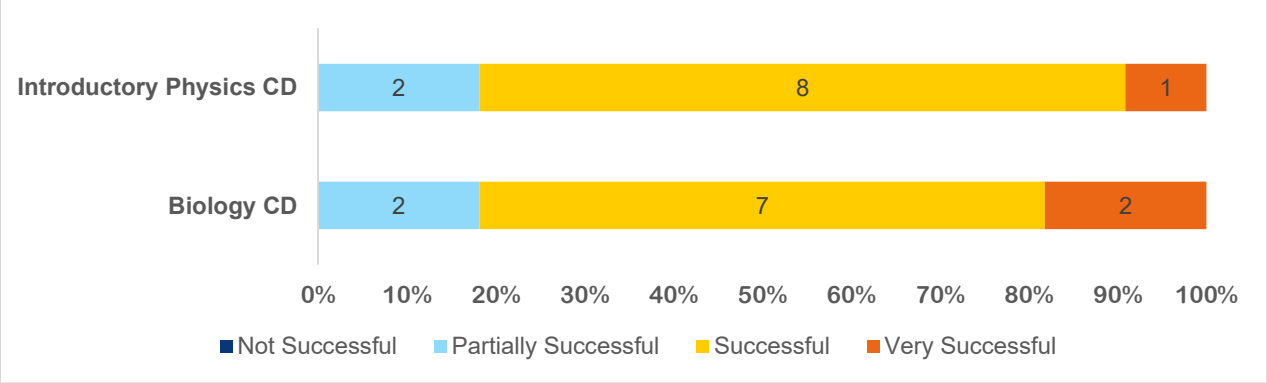
Review of the Legacy assessment and student profiles



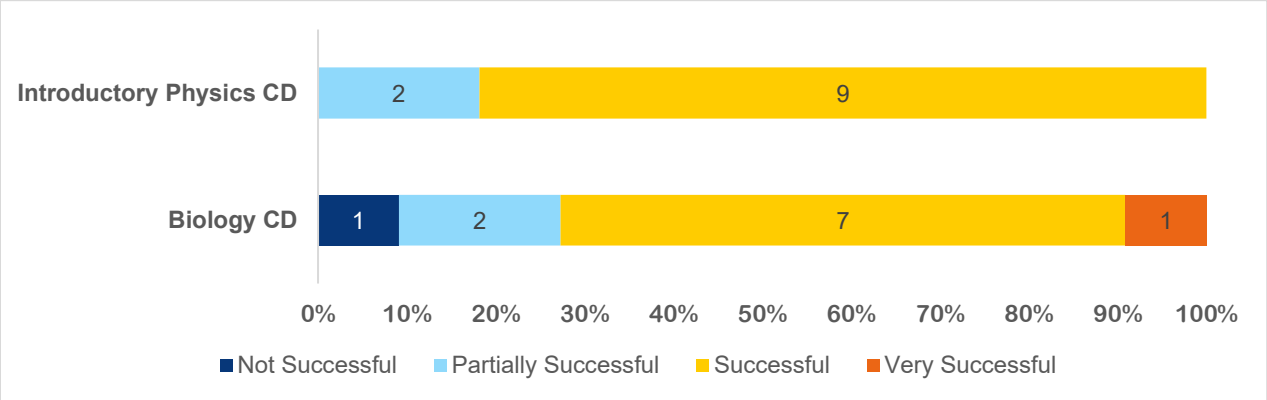
Review of the legacy assessment achievement expectations



Review of the next-generation assessment and student profiles

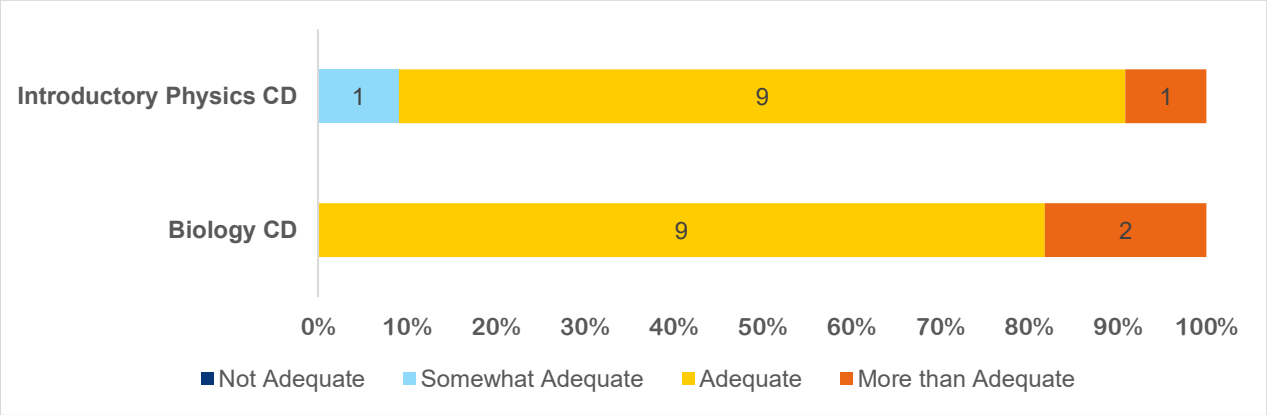


Judgment process of Interim cut score recommendations

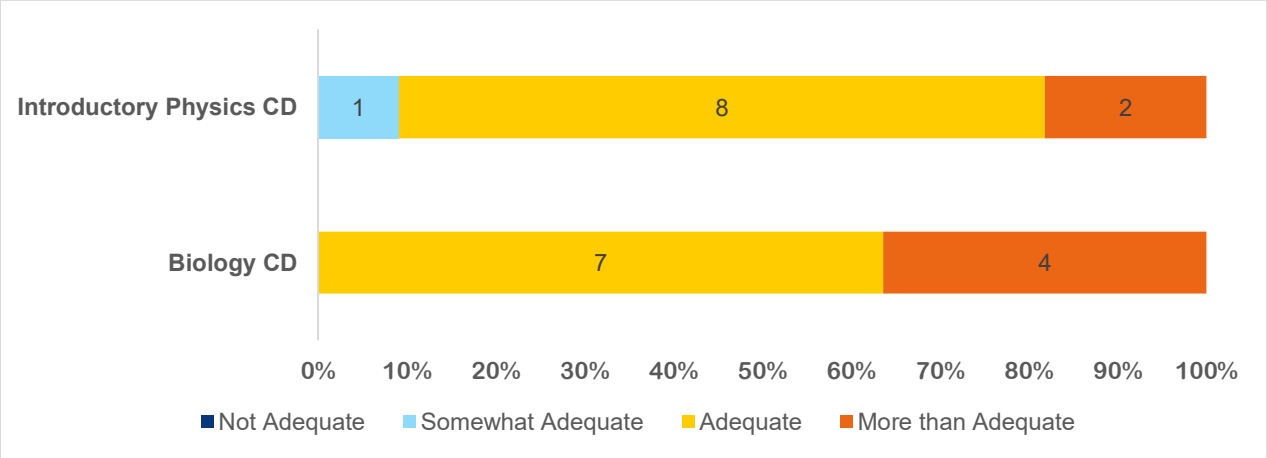


Q.2. How adequate were the following elements of the session?

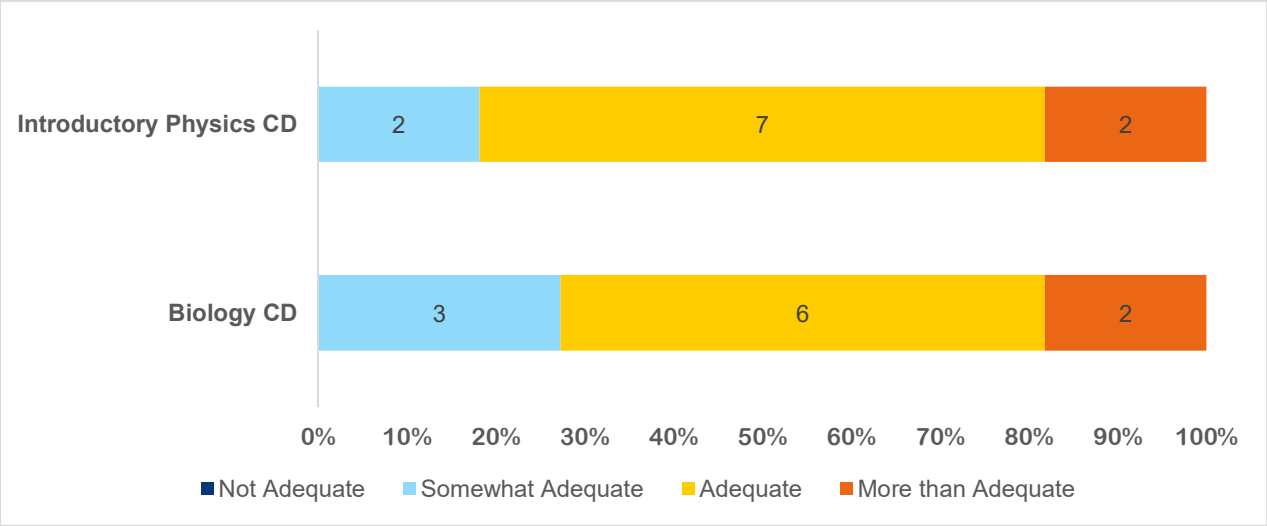
Amount of time spent reviewing the Legacy assessment



Amount of time discussing the achievement level definitions

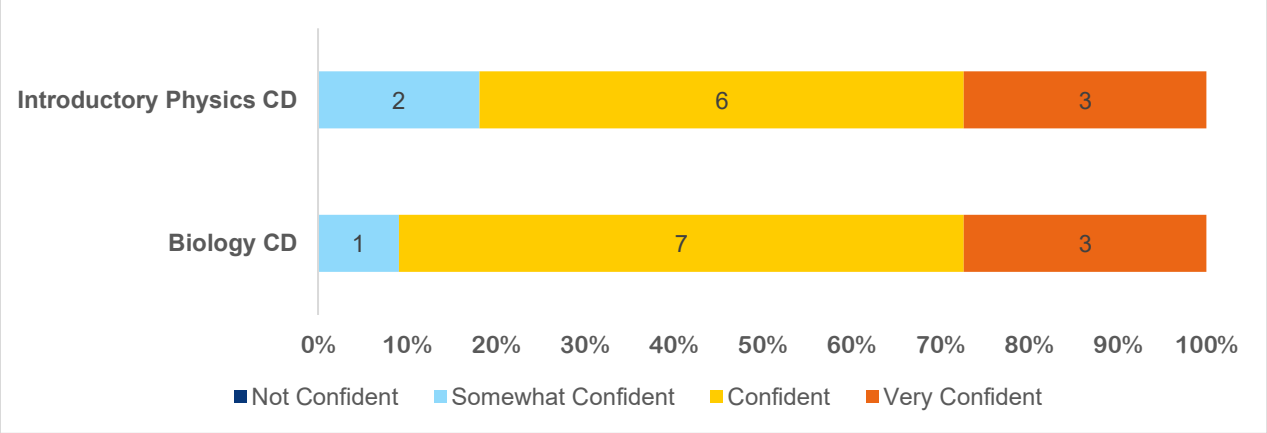


Amount of time spent reviewing the Next-Generation assessment

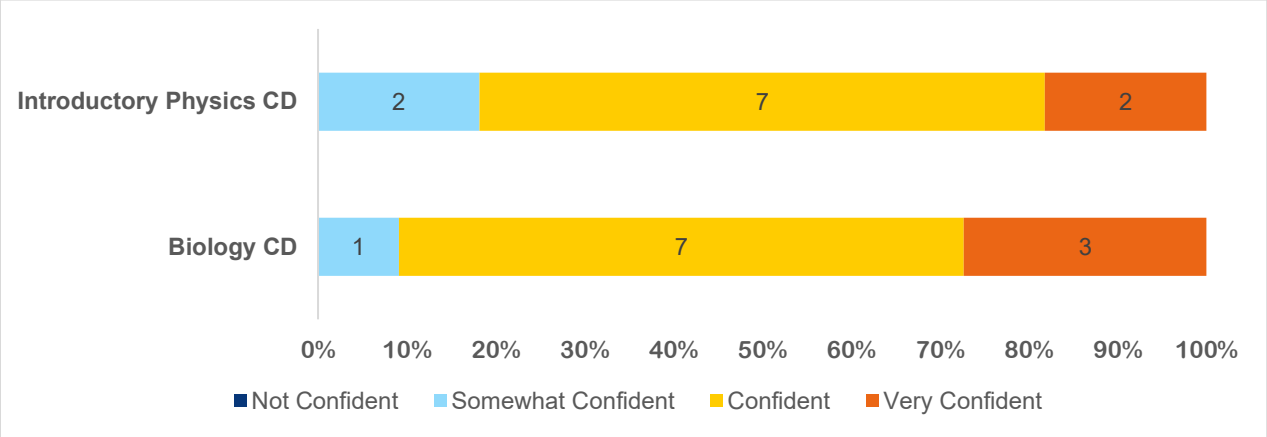


Q.3. How confident do you feel that the final interim cut score recommendations for the subject represent appropriate levels of student achievement?

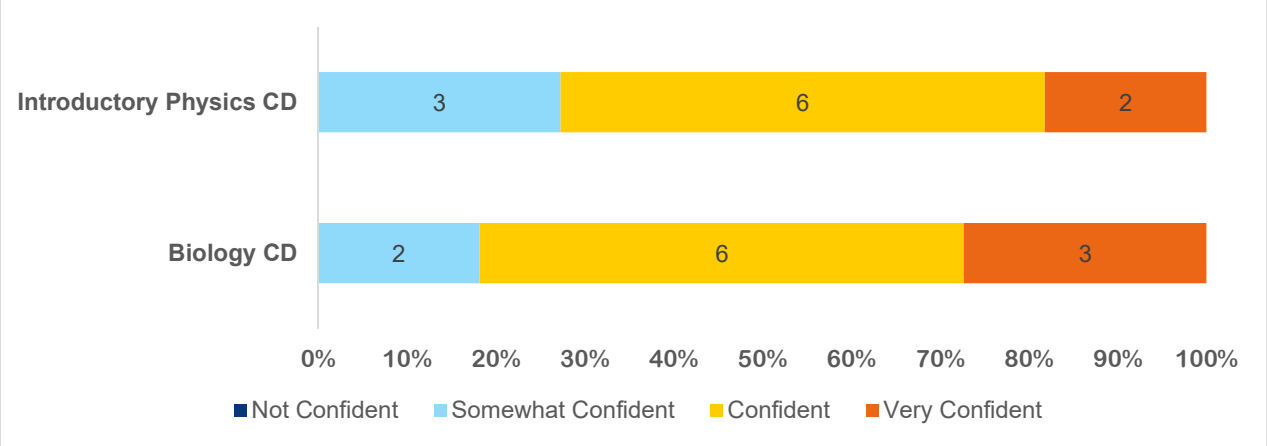
Needs improvement



Proficient



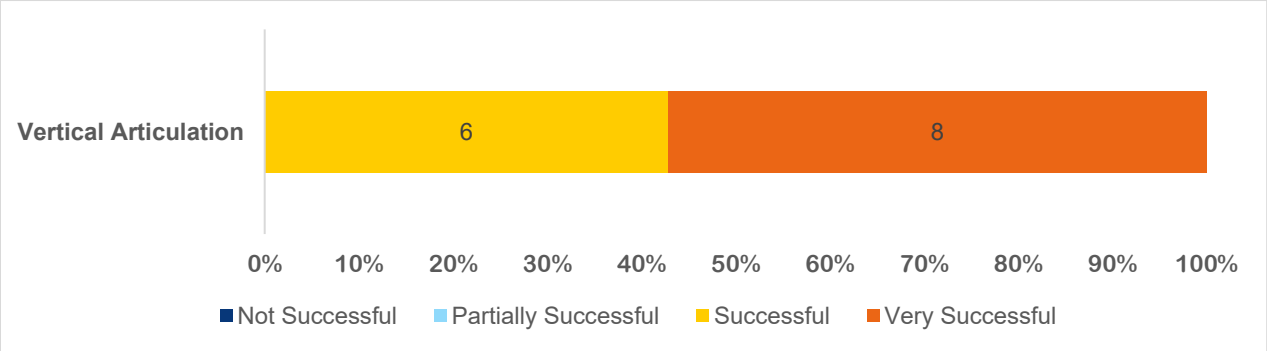
Advanced



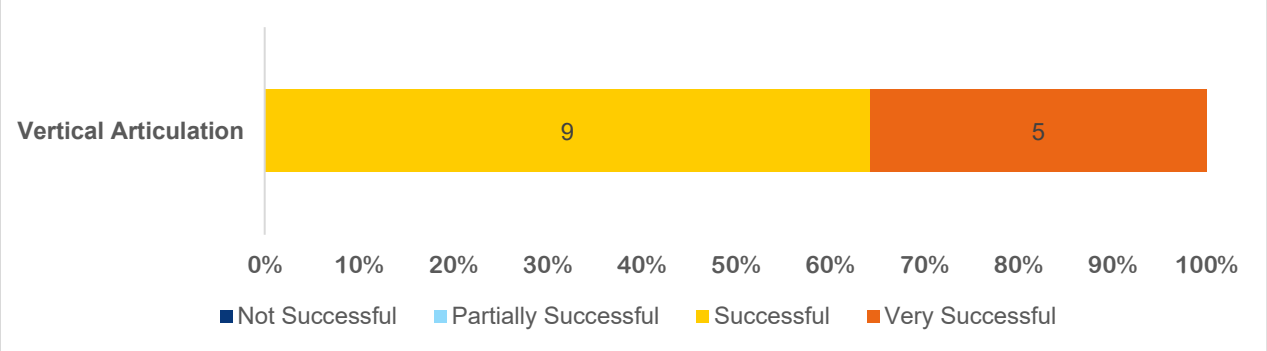
Vertical Articulation Session Process Evaluation

Q.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.

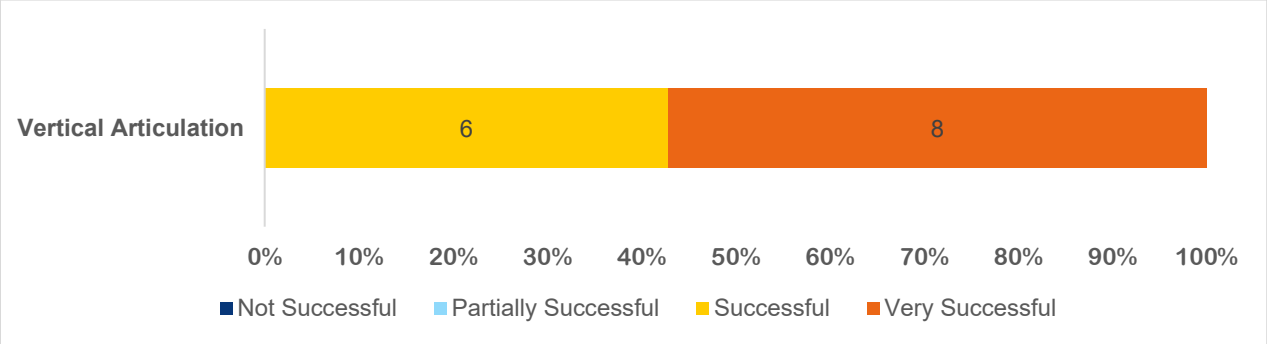
Introduction to vertical articulation process



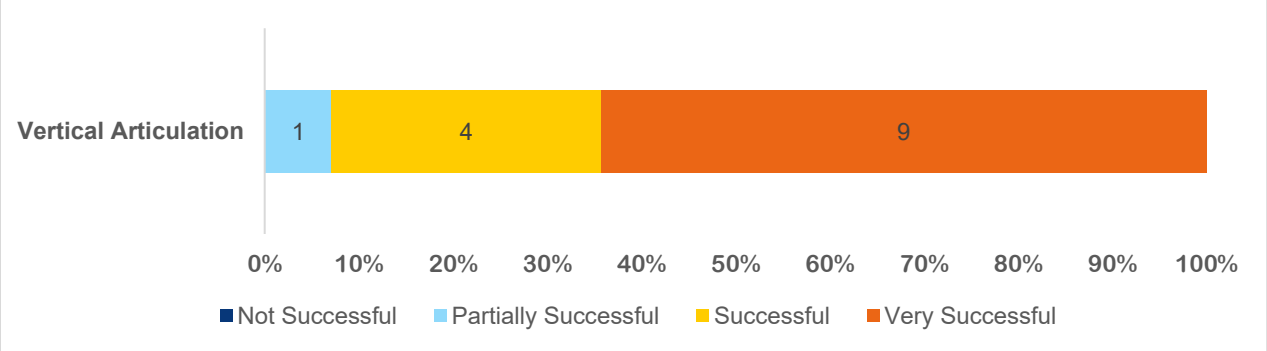
Review of the Achievement Level Descriptors



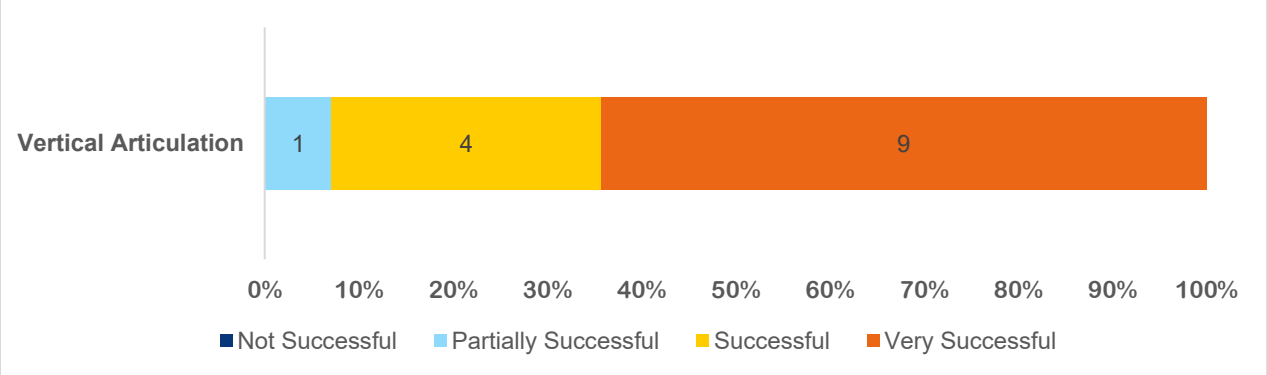
Review of the cross-grade impact data



Use of interactive vertical articulation spreadsheet

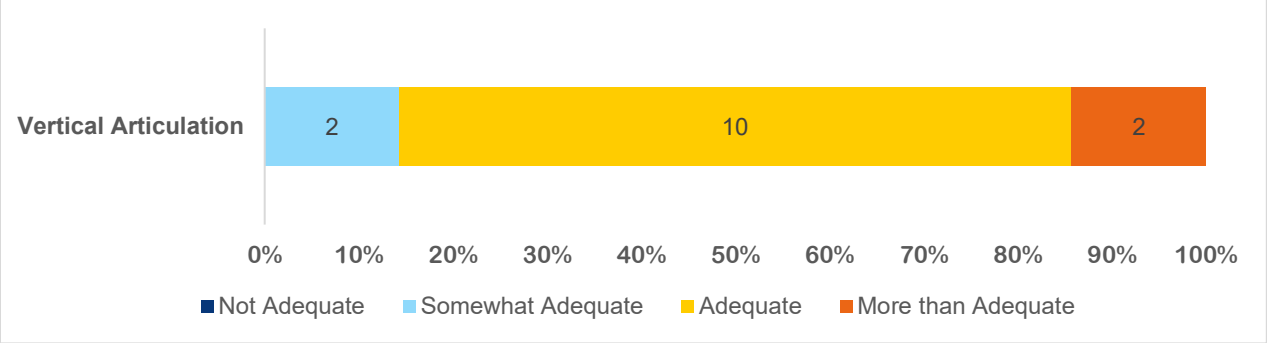


Discussion of recommended changes

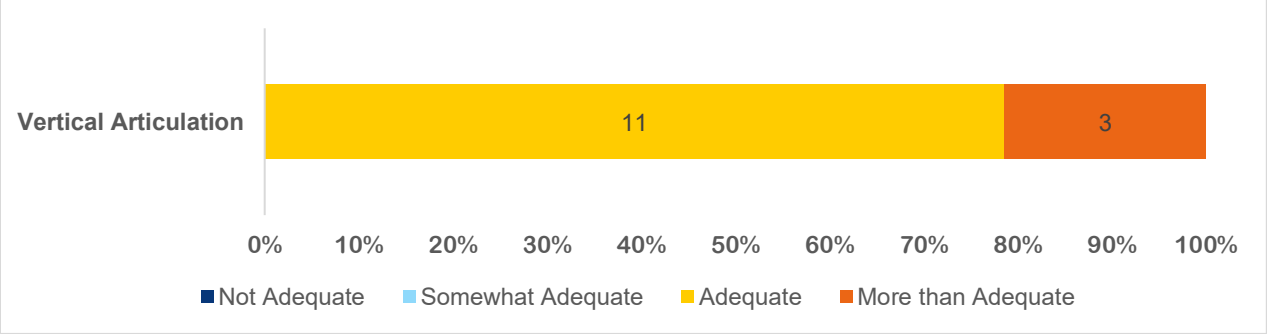


Q.2. How adequate were the following elements of the session?

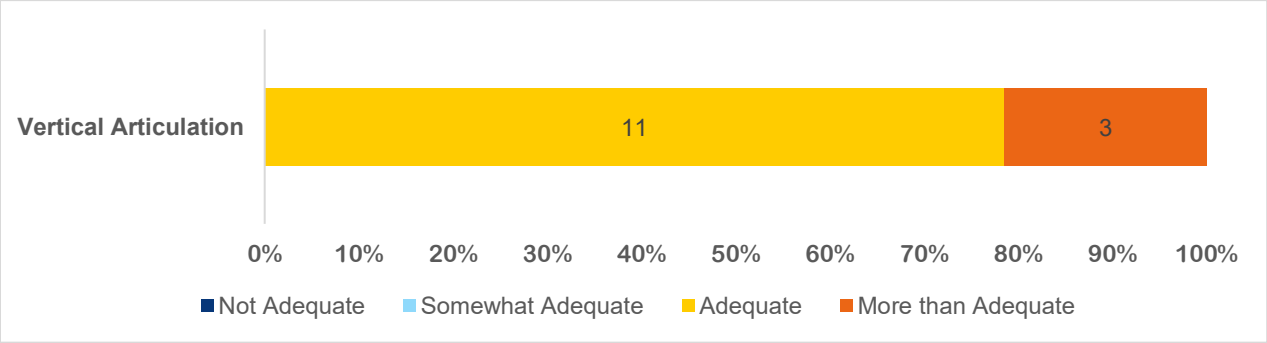
Amount of time spent reviewing the ALDs



Amount of time discussing the impact data

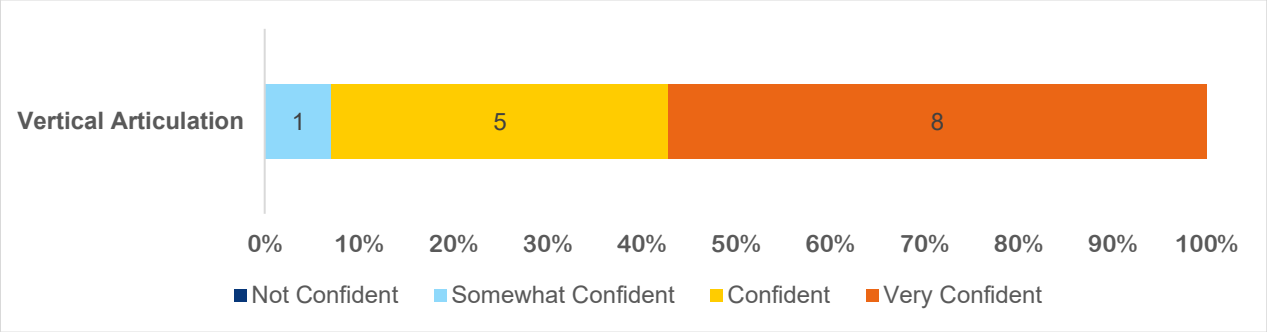


Amount of time working with the interactive spreadsheet

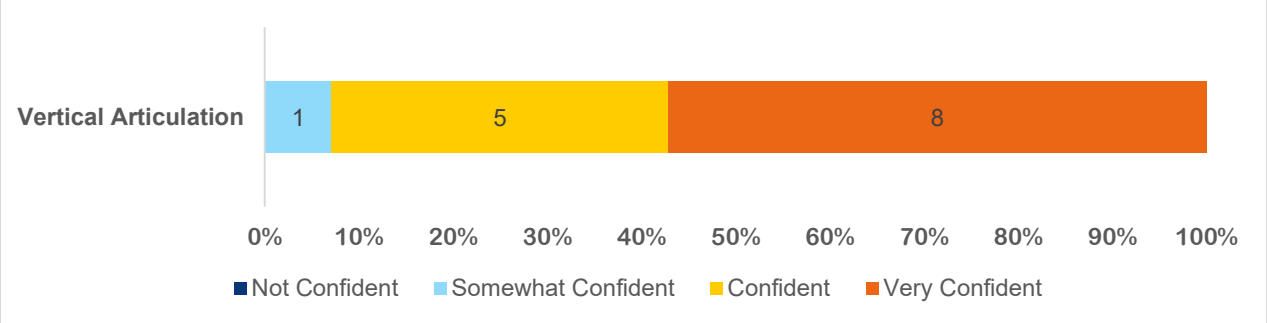


Q.3. How confident do you feel that the final cut score recommendations for Biology and Introductory Physics represent appropriate levels of student achievement?

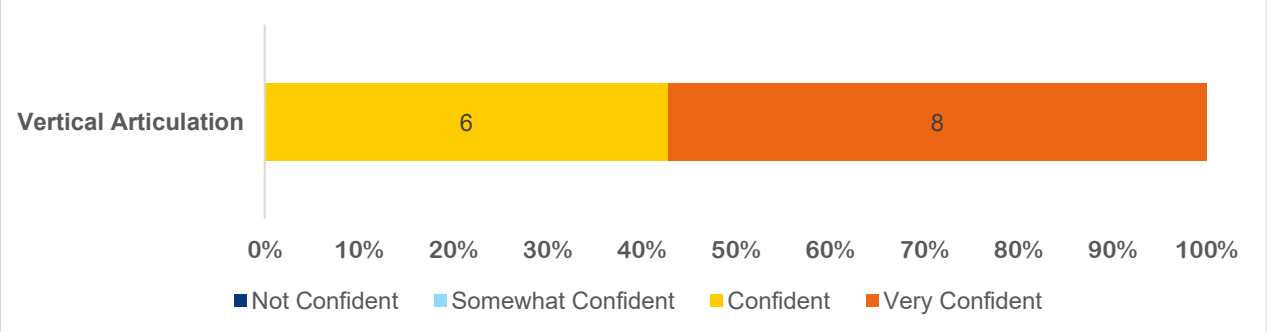
Partially Meeting Expectations



Meeting Expectations



Exceeding Expectations

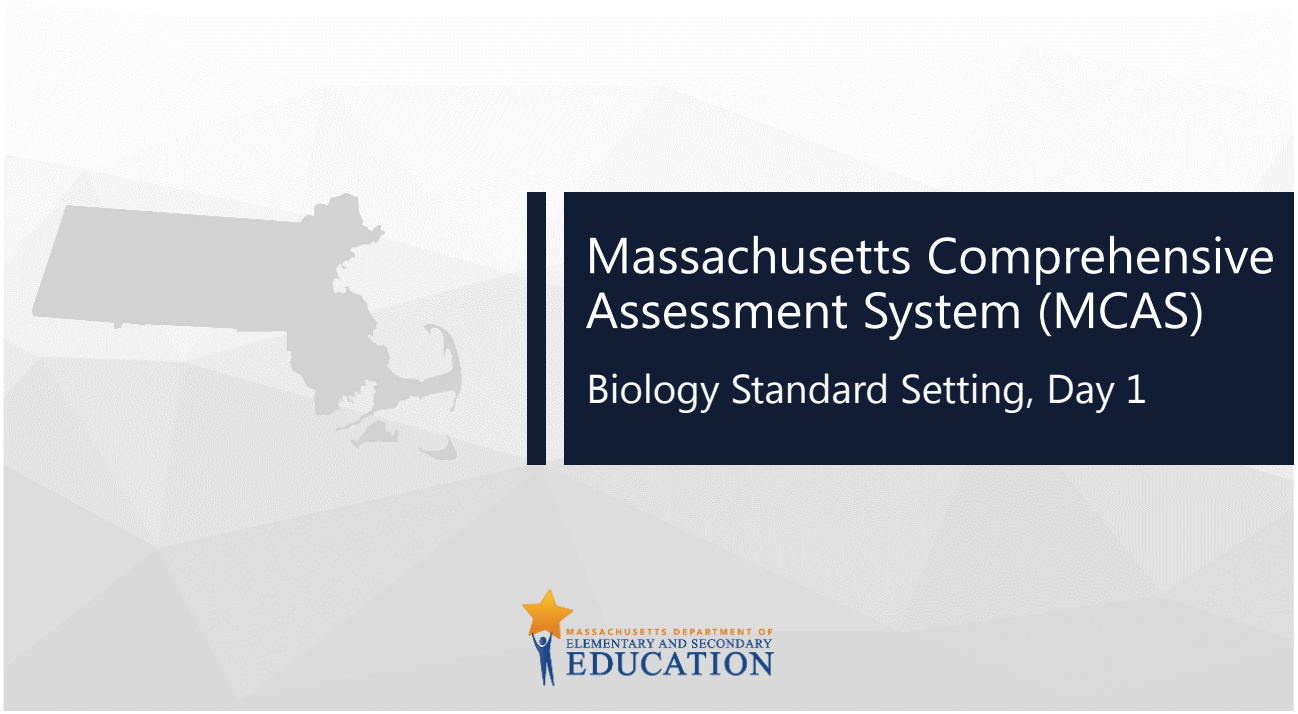


Appendix LL – PowerPoint Presentations

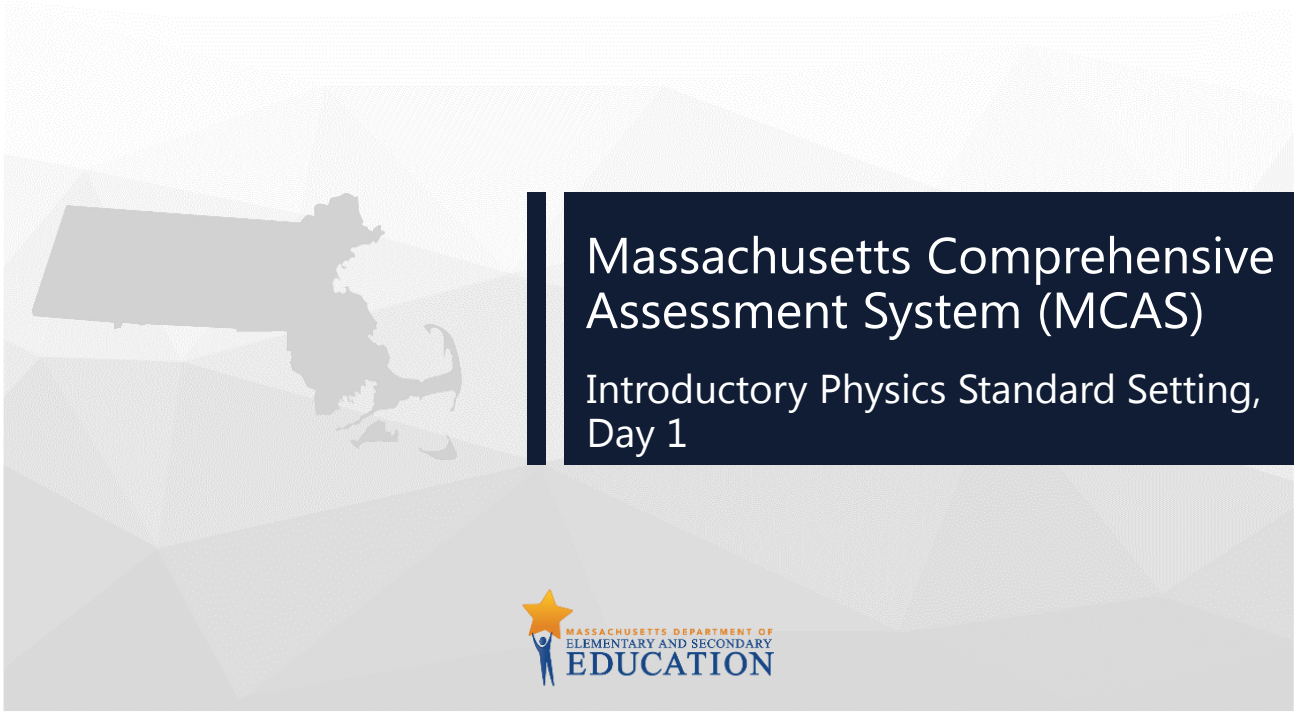
A sampling of presentations from the General Session and Breakout sessions by day are presented below. The full presentations may be accessed via the attachment paperclip on the left side of the pdf reader.

General Session

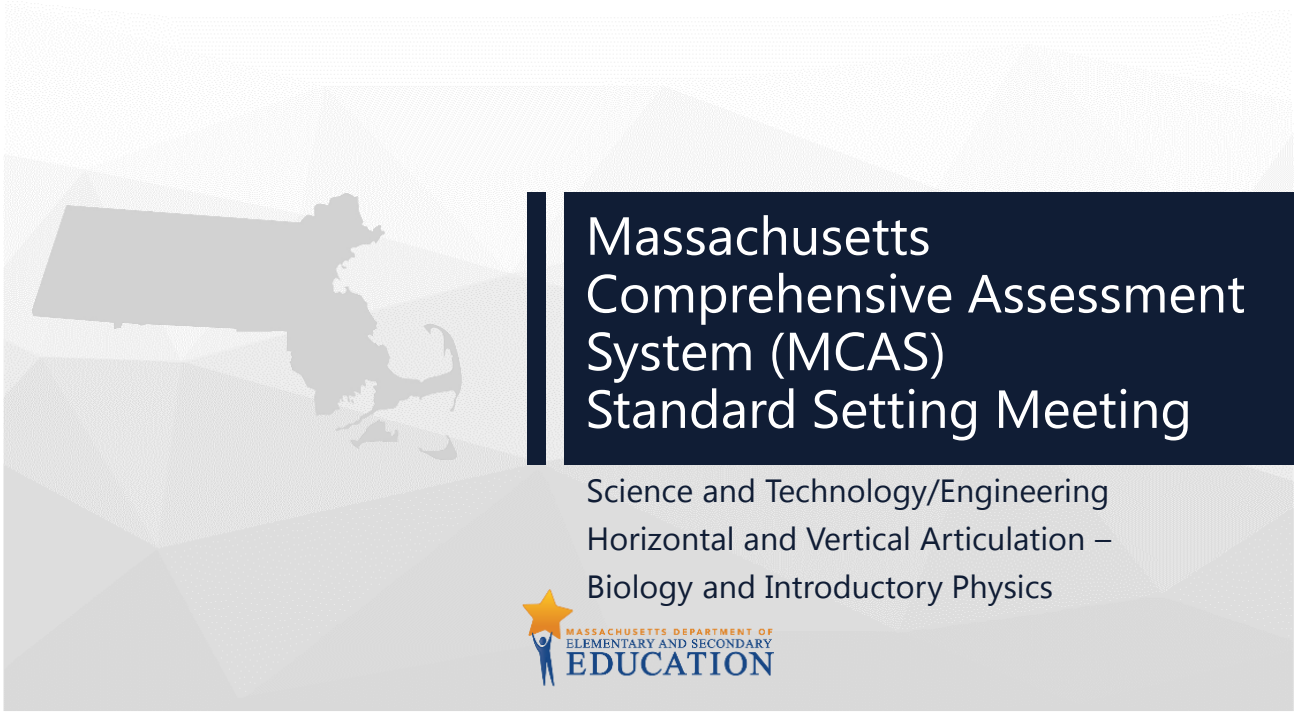
MCAS Biology Breakout Days 1-3



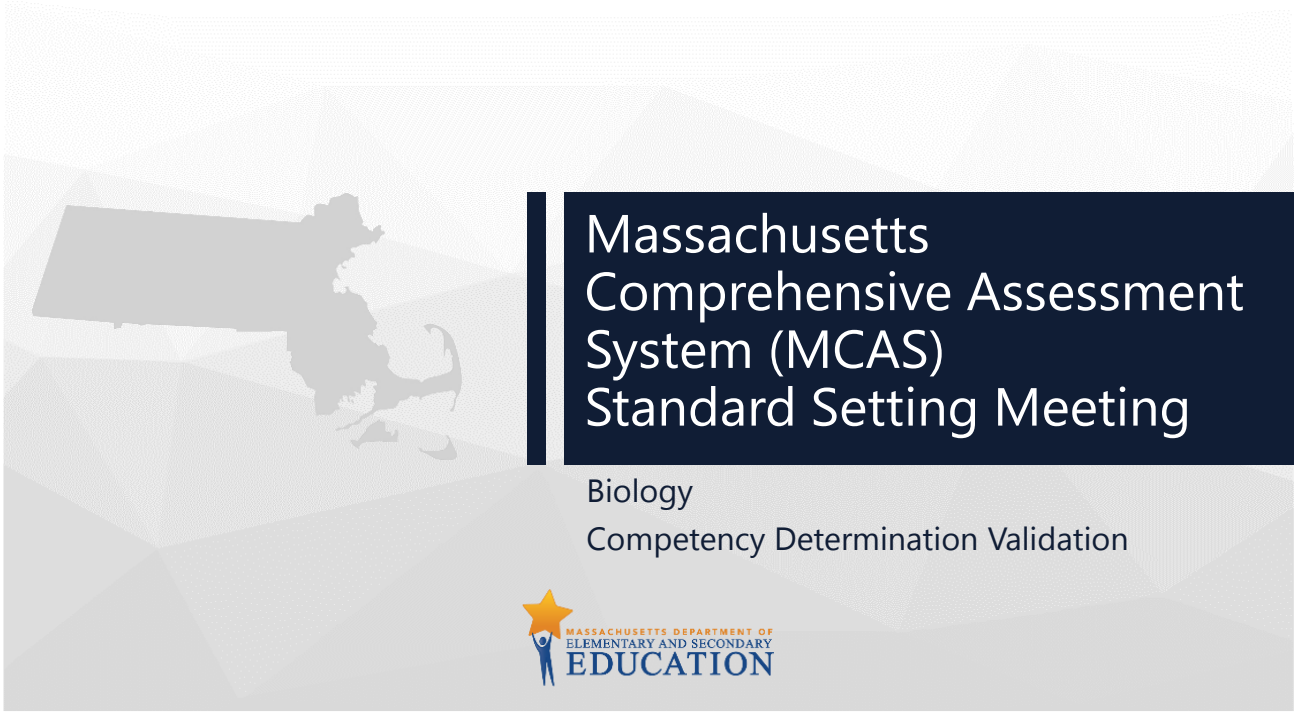
MCAS Introductory Physics Days 1-3



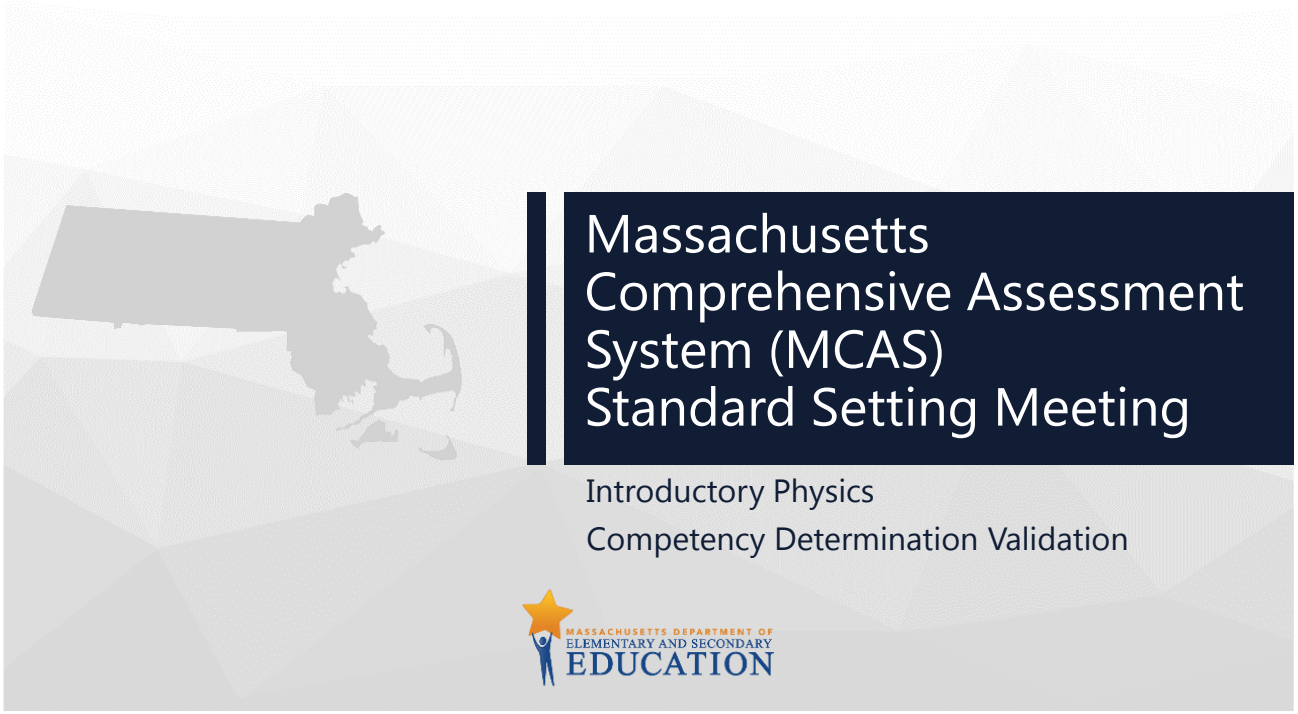
Vertical Articulation



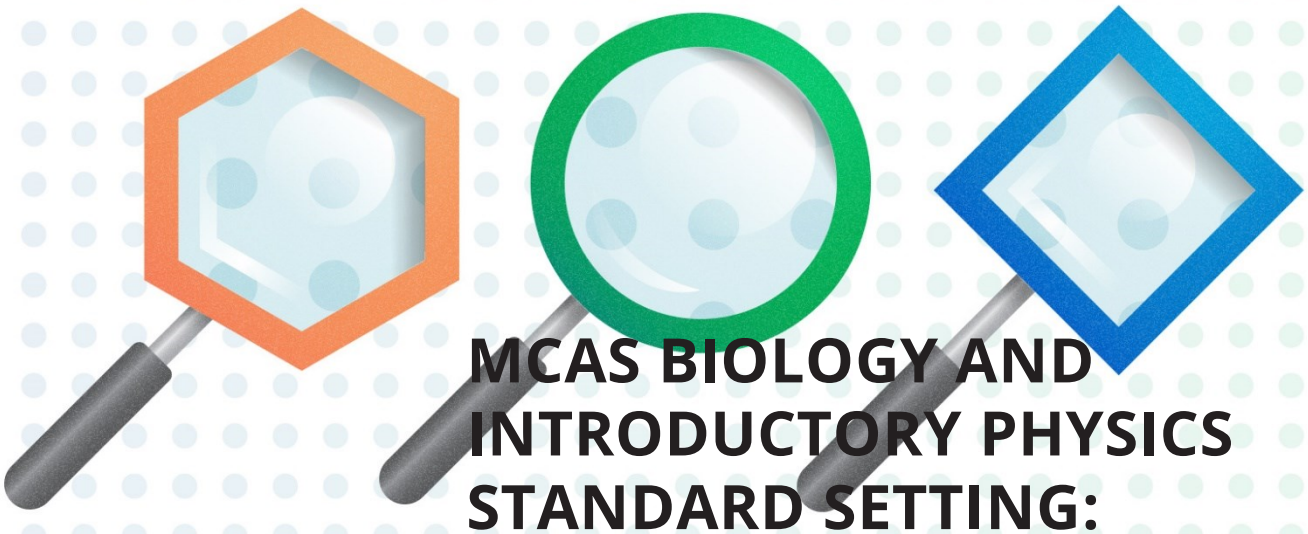
Competency Determination Validation Biology



Competency Determination Validation Introductory Physics



Appendix MM – External Reviewer Report



MCAS BIOLOGY AND INTRODUCTORY PHYSICS STANDARD SETTING: OBSERVATION REPORT

September 2022

Will Lorié, Ph.D.
Senior Associate
Center for Assessment

Center for Assessment

National Center for the Improvement
of Educational Assessment
Dover, New Hampshire

Table of Contents

Pearson and Cognia, under contract with the Massachusetts Department of Elementary and Secondary Education (DESE), held a standard-setting meeting on August 9-12, 2022, in Wakefield, MA. The purpose of the meeting was to establish cut scores for the Massachusetts Comprehensive Assessment System (MCAS) Next Generation tests for High School Biology and Introductory Physics (collectively, "high school science"). The meeting included vertical articulation (from 8th-grade science to high school science) and horizontal articulation (between the high school science subjects) components. In addition, the meeting included a scale anchoring procedure to translate the competency determination (CD) cuts from the MCAS legacy assessments in high school science to the Next Generation assessments.

DESE contracted with the National Center for the Improvement of Educational Assessment (Center for Assessment) to observe the standard-setting meetings. Will Lorié was the Center for Assessment observer and the author of this observation report.

Workshop Description

Pearson designed a workshop in which it would implement the Extended Modified Yes/No Angoff method (Davis & Moyer, 2015; Plake, Ferdous, Impara, & Buckendahl, 2005) to recommend achievement level cut scores for each assessment. These cut scores will classify students into four levels: "does not meet expectations," "partially meets expectations," "meets expectations," and "exceeds expectations." The workshop ran over four days. Each content area panel consisted of 19 content experts, principally high school teachers, from across the state. As planned, the panelists experienced the test, reviewed content standards and achievement level descriptions (ALDs), defined borderline expectations, made item judgments over three rounds, and participated in either an articulation or competency determination meeting.

Participants

Several people attended the standard-setting meetings, including Pearson and DESE facilitators, educators who served as panelists, DESE representatives and observers, and an outside observer from the Center for Assessment.

Eric Moyer (Senior Research Scientist, Pearson) facilitated the study. He was assisted by process facilitators Soo Ingrisone (Biology) and Scott Strickman (Introductory Physics), Senior Research Scientists at Pearson. Content facilitators led large-group content conversations in each room. Content facilitators were Katie Bowler (Director of Test Development, DESE) and Steve Long (Test Development, DESE) in Biology and Phil Slauzis-Durham (Test Development, Cognia), and Isadel Eddy (Test Development, DESE) in Introductory Physics. DESE representatives, including Michol Stapel (Associate Commissioner, Student Assessment) and Rob Curtin (Chief Officer for Data, Assessment, and Accountability), provided context in the general session or served as process observers. Ha Phan (Research Scientist, Pearson) also served as an observer.

DESE invited 38 participants to attend the standard setting – 19 panelists for Biology and 19 panelists for Introductory Physics. Throughout the standard-setting meeting, participants were engaged with the task, posed questions to the facilitators, and sought clarification from each other.

Observation Notes

Day 0: Monday, August 8, 2022.

Attendees: Eric Moyer, Michol Stapel, Frank Padellaro (Psychometrics Manager, Cognia), Chris Clough (Lead Program Manager, Cognia), Phil Slauzis-Durham, Dawn Cope (Test Development, Cognia), Scott Strickman, Soo Ingrisone, Ha Phan, Steve Long, Isadel Eddy, Katie Bowler, and Will Lorié.

Shortly after 3 PM and per the schedule, Eric Moyer opened this meeting. Eric indicated that he would provide some updates at this meeting. He said that a new section was added to Pearson's online standard-setting system to assist with a part of the standard-setting meeting in which groups would discuss the relative difficulties of pairs of items. Eric highlighted that there should be two aspects to the discussion – the first would be about the items' content, and the second would be about item "features" that can influence the relative difficulties of the items. Eric provided examples of such features: single-part versus two-part, stand-alone versus part of a module, and TEI versus open response. Eric indicated that other than this, there should be no major changes to the Day 1 process.

For Day 2, Eric said that a slide would be added to illustrate the process of making a standard-setting judgment on a multipoint item. He noted that content specialists would be making updates to some item keys. Eric added that handwritten responses had been turned into typed text.

We then had general introductions.

Eric asked process and content facilitators to be in the large-room presentation, scheduled to begin at 8:30 the following day so that panelists could see them and connect names to faces.

Eric provided guidance for how far each group should get by the end of Day 1 – drafting borderline ALDs of students at the meeting expectations achievement level. He said that if participants completed that activity by 3:45, then it was OK to move on to drafting borderline ALDs of students at the "partially meets expectations" achievement level.

Eric said there would be a debrief at the end of each day, from 4:30 to 5:00. He adjourned the meeting.

After the meeting, there was some discussion between Eric and DESE (Michol and Katie) regarding the CD validation activity. Katie raised concerns that the starting point for the CD validation implies, from a policy perspective, cut scores that would be too low. I did not observe the resolution of this discussion, but this issue was raised at least a couple more times before Day 4 (when the CD validation activity would take place). The range of raw score points for CD validation was ultimately expanded prior to Day 4.

Day 1: Tuesday, August 9, 2022.

Attendees: Michol Stapel, Rob Curtain, Eric Moyer, process facilitators, content facilitators, observers, and all panelists.

At 8:30 AM on Day 1, Michol opened the general orientation session of the standard setting with a welcome and an introduction to the lead staff. Katie followed with additional orientation in which she highlighted the diversity and collective experience of the participating educators. Rob then addressed the role of Massachusetts educators in the MCAS test development process, drawing particular attention to the challenges raised by the COVID-19 pandemic. He told the panelists that it is "really important that we do not lower our expectations for our students." He thanked them and said he hoped they would enjoy the experience. Michol then recounted the history of MCAS standard settings. She said that the shift from the legacy MCAS to the Next Generation MCAS implied a "new definition of how we talk about student achievement." She then turned the session over to Eric, who provided an overview of the Next Generation MCAS standard-setting meeting.

Eric reviewed the schedule, ground rules, and the role of the panelists. He emphasized that they will set standards for the entire state – not (just) for their students, schools, or districts. He provided test security dos and don'ts. Among the latter: Not using one's phone in the room, not taking equipment or materials out of the room, and not discussing – outside of the meetings – the specifics of in-room discussions. He encouraged them to share their experience in general terms with their colleagues. He drew attention to the Cognia NDA, which he then asked participants to read and sign if they had not already done so.

Eric then explained the standard setting process at a high level, from panelists to commissioner approval, to incorporating the final cut scores for reporting test results. He explained that standard-setting meetings bring together student expectations (in the form of ALDs), content expertise (provided by panelists), and assessment to answer the question, "How much is enough?" He presented a visual showing how cut scores classify students. He then turned to the task of the panelists, defining basic terms such as content standards, achievement levels, ALDs, cut scores, standard-setting process, and feedback data.

Eric then introduced the Extended Modified (yes/no) Angoff process. He said it was the same process used in other Massachusetts state tests. Eric explained that it is a content-based standard-setting method in which panelists make item-centered judgments. Eric said that panelists would be looking at an item and thinking about the knowledge and skills needed to get the item correct. To illustrate the process for the lowest cut that panelists will be considering and for a simple multiple-choice item, he said that panelists would ask themselves: If a "partially meets expectations" student, as defined by the ALDs, encounters this item, will they get it correct? They will answer that question with either yes or no.

Eric said that the standard-setting process would be iterative, occurring over several rounds. The first round would be about "you and the content," referring to the panelists working individually. Feedback would be provided, comparing one's judgments with those

of others, followed by discussion among all panelists. There will then be a second and third round.
Eric then outlined the standard-setting process at a higher level, which included experiencing the test, reviewing standards and ALDs, defining borderline expectations, rendering item judgments, and participating in articulation and competency determinations.
Eric returned the judgment process in more detail, using a hypothetical example of setting standards on a test designed to elicit the degree to which the test taker was a fan of the New England Patriots (U.S. football team).
Eric emphasized that in the actual standard setting, all judgments must be individual. He briefly turned to the agenda for the Friday meetings (vertical and horizontal articulation and competency determinations). He ended with an overview of the agenda before opening it up for questions.
A panelist asked if this standard-setting process is done for every test. Eric replied, "We don't do this every year for every test. We take your cut score recommendation and place it on the test scale. This applies to every test that is created based on the scale."
Another panelist asked how tests differ from year to year. Eric said that tests vary yearly in specific content but not in the representation across the reporting categories and approximate difficulty.
After no further questions, Eric adjourned the general session and asked the panelists to join their content area groups.
I observed the Introductory Physics group for a few minutes. Scott, the process facilitator, asked content facilitators, panelists, and observers to introduce themselves. After they did so, Scott returned to address group norms and security. Next, they moved into the "experience the assessment" activity, in which panelists took the test.
I moved to the Biology room, where panelists were also taking the assessment.
Back in the Introductory Physics room, Scott closed the "experience the assessment" portion of the program at around 1:10 PM. He asked for reflections on the experience. Panelists made several comments, including:

- Many questions, especially module questions, require students to ignore information[1]
- Not many problems were simple one-step problems. Many involved two or three steps
- It was possible to (easily) eliminate distractors on many questions

[1] This refers to information in item stimuli or stems, which is not needed to answer a specific item. Including extraneous information for some items by design can be a legitimate way to ensure students demonstrate their knowledge on the assessed content.

- [Remark by a college-level teacher educator] Students should / need to receive instruction on how to use different tools, especially the equation editor
- No words were bolded or italicized. (Isadel responded that italics are not used due to universal design considerations, but that bolding is still used when necessary and that DESE has guidelines for bolding.)

Scott then moved to the item difficulty comparison activity, showing two items on the room screen. He facilitated discussion about factors that make one or the other item more difficult. Panelists referenced several factors: content, format, cognitive load, computational, and reading load. A second item pair was presented, followed by discussion. Then a third item pair, followed by discussion.
A similar discussion was taking place in the Biology room.
Around 2:30 PM in both rooms, Scott and Soo oriented panelists to the next scheduled activity, discussing and drafting borderline ALDs. Working from existing (range) ALDs (prepared by a separate educator group before the standard-setting meeting), panelists worked in groups for about 45 minutes to create borderline ALDs. These were to consist of 3-5 borderline descriptions per disciplinary core idea (DCI), identifying key characteristics that the “just barely” student in the “meets expectations” category would be able to demonstrate. Facilitators displayed a slide with some guiding questions on the room screens.
Since the groups worked in tables while drafting borderline ALDs, a large-group discussion was necessary for the entire group to share and contribute these draft descriptions. Content facilitators in each room led this discussion, writing the draft borderline descriptions on large sheets of paper. This activity continued until the scheduled close of the panelists’ work, at 4:30 PM, when panelist meetings adjourned.
Eric held a debrief at the end of Day 1. All non-panelist participants attended. Eric addressed several points, beginning with panelist Q&As:

5. Question: Wouldn’t the CD activity affect the borderline descriptions?

   Answer: No.

6. Question: How should we consider the pandemic (in our borderline descriptions and judgments)? Answer: Don’t.

7. Question: If our school made a conscious decision not to teach (a particular standard), the answer to the 2/3rds question (the operational definition of "likely to get an item correct") would always be no for that standard, right?

   Answer: No. You must think about the state, not your school.

8. Question: They are thinking about potential items when they are thinking about the borderline description, and they shouldn't, right?

   Answer: Correct, they should be thinking about content, not items. "We are not going to be able to write borderline descriptions that apply to every single item. What we're trying to get them to do is get them to speak the same language – not a list, but a comprehensive judgment. The working definition will get refined as we move along."

Eric encouraged the process and content facilitators, as they facilitate discussions, to listen for "should versus would," "meets versus borderline," and "2/3rds of the time." These were critical phrases from the training. After a brief discussion of reimbursement logistics, the types of feedback that panelists would receive, and expectations for where panelists should be at the end of Day 2, the debrief was adjourned.

Day 2: Wednesday, August 10, 2022.

Attendees: Michol Stapel, Eric Moyer, process facilitators, content facilitators, observers, and all panelists.

Day 2 began with continued discussion and formulation of the borderline ALDs for Biology and Introductory Physics. In both groups, a facilitator made it clear to panelists that if they felt that a particular range ALD statement adequately conveyed some aspect of the borderline description for that achievement level, they were free to take that statement without modification into the borderline ALD. This clarification helped prevent unproductive discussion about how to revise a range ALD component when it might already be in "minimal" (that is, borderline) form.

Process facilitators began judgment task training in both rooms after participants drafted all three borderline ALDs. (Meanwhile, content facilitators typed and printed the text of the borderline ALDs for distribution to panelists.)

Participants next engaged in a practice round of the judgment activity and had an opportunity to ask questions and to individually register, via survey, their comfort with instructions and the judgment task.

Participants next entered Round 1 of their judgments individually. This phase entailed reviewing every operational item on the Spring 2022 test form. After all panelists in a room had made their judgments, process facilitators provided judgment feedback, which consisted of individual and committee-level feedback. Each panelist received their individual item judgments and their individual cut scores. The latter was expressed in raw score units and computed by Pearson. The committee-level feedback consisted of the mean, median, minimum, maximum, and second and third quartiles of the cut score recommendations across all panelists. A panelist item judgment agreement stacked bar graph showing the number of panelists recommending a particular raw score for a specific level, with the level judgments color-coded in such a way to allow panelists to visually inspect the degree of overlap in the distribution of judgments for the different achievement level cuts; and item statistics (item p-value or distribution of point-score for multipoint items).

Next, process facilitators initiated a whole group discussion for each achievement level cut. The discussion was driven by items flagged statistically as exhibiting the most disagreement[2] across panelists.

In reviewing sample flagged items, panelists had discussions justifying their ratings based on their descriptions of the borderline student at the level being discussed. During this discussion in Introductory Physics, one panelist drew on another's reasoning to change her rating, consistent with guidelines about the conditions under which panelists should change their ratings: As a result of dialogue and information exchange, not unrelated social influence factors (Fitzpatrick, 1989).

Two potentially problematic issues emerged in the Biology room, as evidenced by the discussion following Round 1. The major one was the perception among some panelists that the borderline "exceeds expectations" student would get all items correct, which led to a high recommended cut score for that level for Round 1. The second issue stemmed from a component of the "partially meets expectations" borderline ALD. The panelists had agreed that the borderline student at that level would know about one body system, not all of them. However, the test had items addressing several different body systems. Thus, it was not readily apparent how to implement this borderline ALD in rendering judgments. One panelist chose an item-by-item strategy, which resulted in her providing a judgment of "yes" on each body system item meeting the borderline ALD description, independent of her judgments on other body system items. This led to her providing more "yeses" to body system items for that achievement level that she would have if she considered them collectively. Another panelist adopted a different strategy: He reasoned that the digestive system was the simplest and, therefore, the one most likely to be the one body system familiar to a student at the borderline of partially meeting expectations. Thus, he rendered his judgments according to this criterion – that is, "yeses" to digestive system items wherever appropriate, "noes" to the others.

[2] As Eric explained to me, if a multiple-choice item was such that half the panelists judged a borderline "meeting expectations" student to get it correct (with the other half rendering a "no" judgment), then panelists could not disagree more on that item (for that level), and it would be among the items flagged.

Eric addressed the first issue by reminding the Biology panelists that the concept of "just barely" meeting the requirements of an achievement level also applies to the highest level. So it does not follow that a borderline "exceeds expectations" student would get every point on every item. After his explanation, the panelists appeared to understand and internalize this notion.

Panelists adopted a workaround strategy post-Round 2 that would result in cut score recommendations unaffected by this artifact of item-by-item methods for borderline ALD statements that followed a "one of N" formulation. Each panelist chose a specific system to operationalize the borderline ALD statement and rendered their judgments according to that choice.

Panelists in both groups concluded their Round 1 discussion and moved to Round 2. Biology room panelists post-Round 2 drew heavily on item statistics and the disparity between panelists' expectations of how students would do on the items and how they had performed. This prompted Eric to remind the panelists that the item statistics they are seeing are somewhat depressed from what they would have been had there not been a pandemic. This was especially the case in Biology, where item data included students who took Biology in 9th grade but, because of schooling disruptions, took the test a full year later, in 10th grade, when they might have forgotten much of the material. Eric told the panelists to consider the item statistics "with a large grain of salt."

At the close of day debriefing, Eric and others discussed the reasonableness of the initial cuts and the resulting impact data. Eric suggested that process and content facilitators emphasize the COVID effect to contextualize the item statistics and impact data for panelists. At the debrief, one person said that some panelists set their "partially meets expectations" cut below the test's guessing level in Biology. Eric said the solution to this is to have a conversation with the group about guessing to have them think about whether their recommendation implies a cut below the test's guessing level.

Day 3: Thursday, August 11, 2022.

Attendees: Michol Stapel, Eric Moyer, process facilitators, content facilitators, observers, and all panelists.

In the Introductory Physics room, Day 3 began with Round 2 feedback. Scott presented items on which there was the most disagreement in judgments and elicited opinions from the panelists on their reasons for their judgments. Most panelists drew on the borderline descriptions or the ALDs. Some brought in considerations about guessing, cognitive processes (such as the number of steps required to arrive at the correct answer), and the Spring statistics. Some panelists articulated how they were explicitly weighing these different considerations to arrive at their judgments.

In the Biology room, as panelists moved into a discussion of Round 2 feedback, Eric addressed the impact data extensively. He used a thermometer analogy, saying that the panelists' job was to operationalize the equivalent of 98.6 degrees Fahrenheit concerning the standard, as determined by the content. Eric also emphasized that panelists were setting standards not just for 2022 but also for several years into the future. He said competency determination is a different matter, which panelists will consider later in the

standard setting. Eric later had a similar discussion with the Introductory Physics panelists when they saw the impact data post-Round 2.

The item-by-item whole group discussion in the Biology room was predominantly content-based. Not all items on the flagged list were discussed, which the process facilitator noted was due to time. However, she said that if anyone wanted a particular item to be addressed, to let her know. (She also focused on items that they had not already discussed.)

After panelists in both groups indicated they were ready, they moved to Round 3. As before, feedback was provided at the end of the round.

The panelists in each group were split approximately equally, with half assigned to articulation and the other half to competency determinations. The panelist meetings were then adjourned for the day.

The Day 3 debrief took place after lunch. At the debrief, Eric outlined what would happen on Day 4. He advised the facilitators in the articulation groups not to begin vertical articulation by showing the data "because the panelists will create a story to explain it." He said that instead, they should start by asking the panelists what they expected the impact data to look like across the grades, based on the content.

For horizontal articulation, Eric said that the discussions and patterns until that point suggest that panelists will think that the impact should be the same across the grades. Eric said that we would bring the matched data in for this discussion.

Eric then turned to the competency determination validation plan. He illustrated how to browse legacy items in the panelists' online portal. He said the "key pieces" of information were the profiles at the three legacy cuts. These profiles showed, for each legacy cut,

3. The distribution of points on multipoint items; and

4. The conditional percent correct on dichotomous items, grouped into four categories:

    a. Category 1: 0-39% ("about guessing"),

    b. Category 2: 40-59% ("a little above guessing"),

    c. Category 3: 60-74% ("this matches the 2/3rds -this gives us an idea of what students at this level know and are able to do"), and

    d. Category 4: 75%+

Eric explained that the goal was to have panelists take the items in this third category and have them build a description of what these students know and can do. They were to do this for the legacy achievement levels: "needs improvement," "proficient," and "advanced."

Eric said to aim for about five statements per level; he clarified that panelists are not to look at or draw on the legacy ALDs for this task.
The next question for the CD group would be to answer, for each legacy borderline ALD that they crafted, what raw score in the NextGen assessment best describes that legacy borderline ALD? At that point, Eric explained, panelists will review a range of profiles (each conditional on a different raw score) to answer that question. Eric instructed facilitators to have panelists start in the middle of the profiles group, then work up or down.
Eric then reviewed the judgment form with the debrief group, and the group then discussed whether to plan to have a second round. The consensus was to look at the spread of panelist judgments and decide.
Concerning the legacy profiles, Eric drew particular attention to the profile for the Biology "needs improvement" level, as there was only one dichotomous item in category 3, and none in category 4. Eric said that in this case, to instruct panelists to look at the category 2 items that assess more or less the same content as the item in category 3 and to craft a statement about "emerging understanding" at the Biology "needs improvement" level.

Day 4: Friday, August 12, 2022.

Attendees: Michol Stapel, Eric Moyer, process facilitators, content facilitators, observers, and all panelists.
On Day 4, three groups worked simultaneously: the Biology competency determinations (CD) group, the Introductory Physics CD group, and the articulations group. There were 11, 12, and 15 panelists in these groups, respectively. In the articulations group, there were seven Introductory Physics and eight Biology panelists.
The process in the CD groups proceeded as planned. Notably, the Biology CD group dealt effectively with the most challenging profile ("needs improvement" for the Biology legacy assessment). The group reviewed the category 3 items with the highest conditional p-values to develop a description of the "emerging" knowledge and skills of the borderline "needs improvement" student. By 11 AM, both CD groups were well into describing the borderline "advanced" student.
Eric facilitated the articulations group process by first asking panelists to consider what they would expect in terms of general patterns from 8th-grade science to high school Biology or Physics, and across the two content areas in high school. Eric provided the interquartile ranges (IQRs) of the standard setting judgments for each of the cuts to panelists, indicating that these ranges represented the leeway that panelists had to adjust cuts. They were not allowed to adjust any cut, however, without the approval of the panelists from that content area. The articulation group ultimately adjusted cuts mostly downward to align with their general expectations about how students would be doing statewide. Two of these expectations, for example, were that approximately 50% of students would be at "meets expectations" or above in high school, and about 10% would be in "exceeds expectations." Panelists adjusted four cut scores to reflect this, three down and one up. These changes were all within the IQRs.
In the Introductory Physics CD room, a panelist voiced a concern, after completing the needs improvement CD judgment, that the task was challenging because no Next

Generation test profile corresponded to the group's borderline ALD for the (legacy test) "needs improvement" level. The panelist said she was very unsure of her recommendation. Another did not feel so strongly but felt the methodology was "a little hand-wavy."
Eric intervened, emphasizing that there would be no one-to-one mapping, and acknowledged how difficult the task was. A third panelist expressed similar frustration and proposed they discuss the matter as a group. (Up to this point, panelists had been working individually.) Eric agreed that they could discuss their process with their table colleagues but to refrain from sharing specific raw score judgments. This change in the process was to help them complete the task. Eric had a similar discussion with the Biology CD group after panelists in this group expressed similar concerns.
Both CD groups were shown the distribution of their raw score judgments. They were asked if they'd like an additional round of judgments. Nobody said they'd move their judgments for reasons other than to reach more agreement with their peers, so there was no extra round.
The final meeting of the day, which began at 2:30 PM, involved half the Introductory Physics CD group and half the Biology CD group. Eric asked them what they expected to see in the impact data if they had Biology and Introductory Physics next to each other. One panelist said he'd expect the distributions to be similar. Another echoed this idea. Eric showed the impact data on the matched data, with cuts based on the CD validation activity. He asked the group if they were OK with forwarding the recommended cuts to DESE. The panelists said yes.

Conclusion

The Massachusetts Department of Elementary and Secondary Education (DESE), with the assistance of Pearson and Cognia, held a standard-setting meeting on August 9-12, 2022, in Wakefield, MA. The purpose of the meeting was to establish cut scores for the Massachusetts Comprehensive Assessment System (MCAS) Next Generation tests for High School Biology and Introductory Physics (collectively, "high school science"). The meeting included vertical articulation (from 8th-grade science to high school science) and horizontal articulation (between the high school science subjects) components. In addition, the meeting included a scale anchoring procedure to translate the competency determination (CD) cuts from the MCAS legacy assessments in high school science to the Next Generation assessments.
Pearson designed a workshop in which it would implement the Extended Modified Yes/No Angoff method (Davis & Moyer, 2015; Plake, Ferdous, Impara, & Buckendahl, 2005) to recommend achievement level cut scores for each assessment to distinguish students into four levels. Consistent with Pearson's plan for the workshop, the standard setting included content experts, principally high school teachers, from across the state. The workshop ran over four days. As planned, the panelists experienced the test, reviewed standards and achievement level descriptions (ALDs), defined borderline expectations,

made item judgments over three rounds, and participated in articulation and competency determination meetings.

My observations confirm that Pearson and Cognia followed through on their plan. Consistent with generally accepted guidelines for conducting standard settings, the two organizations implemented a defensible process whereby suitably qualified individuals can bring their expertise to operationalizing expectations expressed in ALDs. Thus, the results of this standard-setting meeting should be accepted as the informed and considered judgment of Massachusetts educators as to the Next Generation MCAS Biology and Introductory Physics assessment cut scores and competency determinations.

Implications for Standard Setting Practice

Two aspects of this standard-setting meeting have, in my opinion, implications for future standard-setting practice. Neither of these implications should be construed as reasons for modifying or second-guessing the results of the process I observed.

5. Item-by-item standard setting methods, including the Extended Modified Yes/No Angoff method (Davis & Moyer, 2015; Plake, Ferdous, Impara, & Buckendahl, 2005), fail to accommodate some legitimate formulations of borderline ALDs. The prime example I observed in this standard-setting implementation was the borderline ALD component referencing body systems for the Biology “partially meets expectations” level. As explained in my observation report, the borderline ALD for this level asserted that the borderline student would know about only one such system. The borderline ALD did not (and need not) specify the system. However, this lack of specification implies that items addressing body systems should be considered as a group. As discussed in my notes, the group found a defensible workaround, the result of which is the same as if the items had been considered as a group. But a more robust standard-setting methodology would allow panelists to render their judgments without such ad hoc considerations.

6. How panelists should take “guessing” into account in making judgments on multiple-choice questions is an unresolved issue in standard-setting practice. The Pearson/Cognia implementation of the MCAS Next Generation high school science assessments assumes that panelists consider guessing when they make their judgments. However, it is not impossible for panelists taking guessing into account, appropriately and consistently, to make judgments that result in cut scores outside of reasonable boundaries – for example, below the guessing level of a test or at or near the ceiling. Both occurred in Biology at Round 1. Pearson did the reasonable thing and communicated the consequences of such “extreme” judgments to panelists. But this begs the question: What if panelists are correct in their “extreme” judgments? What if the test does not span the range it should, at least concerning the ALDs? This potential disconnect between ALDs and the range of the test should be avoided in a more systematic manner, but it is not clear how to do so.

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.

Fitzpatrick, A. R. (1989). Social influences in standard setting: The effects of social interaction on group judgments. *Review of Educational Research*, *59*(3), 315-328.

Plake, B.S., & Cizek, G. J. (2012). Variations on a Theme. In G.J. Cizek (Ed.), Setting performance standards (2nd ed.). New York: Routledge.

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.

Observation of the MCAS Biology and Introductory Physics Standard Setting

This report was prepared for the Massachusetts Department of Elementary and Secondary Education (DESE) by Will Lorié, Ph.D., Senior Associate at the National Center for the Improvement of Educational Assessment (Center for Assessment). For any questions related to this report, please contact him at wlorie@nciea.org.