California Department of Education

Executive Office

SBE-002 (REV. 11/2017)

memo-imb-adad-oct19item01

# **MEMORANDUM**

**DATE:** October 11, 2019

**TO:** MEMBERS, State Board of Education

**FROM:** TONY THURMOND, State Superintendent of Public Instruction

**SUBJECT:** Update on the California Science Test Standard Setting Plan.

## Summary of Key Issues

This Information Memorandum provides a summary of related activities conducted before and after the convening of the California Science Test (CAST) July/August 2019 Standard Setting Workshop, where educators collaborated to develop recommended threshold scores. With the completion of this standard setting workshop, the California Department of Education (CDE) will bring proposed CAST threshold scores to the California State Board of Education (SBE) for approval in November 2019. If these proposed threshold scores are approved by the SBE, Student Score Reports (SSRs) for the 2018–19 CAST administration will be available in January 2020, and summary results will be released shortly thereafter.

Briefly stated, there are four CAST achievement levels: Standard Exceeded, Standard Met, Standard Nearly Met, and Standard Not Met. The standard setting process used at the CAST workshop involved the panelists developing descriptions of what students can do at the beginning of three of these four CAST achievement levels—Standard Exceeded, Standard Met, Standard Nearly Met. Using the descriptions of what students can do, the CAST standard setting panelists made judgements about each item, and these judgements resulted in the setting of threshold scores at the following three levels: Level 2 (Standard Nearly Met), Level 3 (Standard Met), and Level 4 (Standard Exceeded). For more information on the methodology used by the panelists to arrive at the recommended CAST threshold scores, see Attachment 1 for the California Science Test Standard-Setting Plan.

## California Science Test Standard Setting

The testing contractor, Educational Testing Service (ETS), conducted a CAST Standard Setting Workshop July 31 through August 2, 2019, in Sacramento. This workshop was observed by CDE staff and SBE Vice President Ilene Straus, Ed.D.

### **History of CAST California State Board of Education Actions**

In March 2016, the CDE presented and the SBE approved the high-level test design for the CAST to measure the full range of the California Next Generation Science Standards (CA NGSS). That March 2016 presentation can be found at <https://www.cde.ca.gov/be/ag/ag/yr16/documents/mar16item02slides.pdf>. The CAST was designed with the following goals in mind:

* Promote improvements to teaching and learning.
* Encourage science instruction in all grades.
* Measure range and depth of the CA NGSS.
* Provide models of high-quality questions that reflect the CA NGSS.
* Minimize testing time and costs while providing accessibility for all students.

In November 2017, the SBE approved the CAST blueprint and general achievement level descriptors (ALDs). That Agenda Item can be found at <https://www.cde.ca.gov/be/ag/ag/yr17/documents/nov17item07.doc>. The CAST general ALDs (sometimes referred to as policy ALDs) provide high-level descriptions of student performance level expectations across the four levels and articulate the goals and rigor for the final performance standards. These general ALDs were used to inform the development of the grade- and content-specific descriptor range ALDs.

In preparation for the CAST July/August 2019 standard setting workshop, ETS and the CDE collaborated and consulted with national science and Technical Advisory Group (TAG) experts in the development of draft CAST range ALDs. Range ALDs are grade- and content-specific descriptors that describe the cognitive and content rigor that are encompassed within particular achievement levels. The range ALDs provide characteristics or behaviors that a student would exhibit if the student were in a specific achievement level for a specific science domain. They specifically describe the CA NGSS knowledge and skills expected of students in grades five and eight and high school at each achievement level.

In March 2019, the draft CAST range ALDs were vetted by 21 California science educators at a two-day meeting during which they provided ETS with their feedback and recommended edits. These range ALDs were used to inform the CAST July/August 2019 standard setting process. Table 1, below, provides a time line of the key activities related to the development of the CAST.

Table 1. Key Activities—California Science Test

| **Activity** | **Date** |
| --- | --- |
| SBE approved the high-level test design | March 2016 |
| SBE approved the general ALDs and blueprint for CAST | November 2017 |
| First operational CAST administered | January−July 2019 |
| ETS conducted the CAST Standard Setting Workshop with California educators | August 2019 |
| SBE considers for approval the CAST threshold scores | November 2019 |
| ETS provides CAST SSRs to LEAs for distribution to parents (contingent on approval of threshold scores in November) | January 2020 |

### **Purpose of the CAST Standard Setting**

The 46 educators serving as panelists were a diverse group, familiar with the breadth and depth of the CA NGSS and experienced in teaching science to students in grades three through twelve, with an emphasis in grades five and eight and high school. Workshop facilitators trained these educators for the standard setting process and the use of the range ALDs as a tool to guide them in developing panel-recommended threshold scores.

Standard setting is an empirical process that uses a panel of experts to determine threshold scores for each of the achievement levels. During standard setting, educators use range ALDs to create definitions of threshold ALDs or the borderline student definitions. Using a consensus building process, panelists make judgements that define what a student “can do” at the entry to each ALD. The threshold scores are then used to provide the test results to students and parents.

For the CAST standard setting workshop, ETS adhered to the guidelines and best practices proposed in standard setting literature, using two widely accepted standard setting approaches—the modified Angoff and extended Angoff methods. The modified Angoff method was used for the CAST one-point items, and the extended Angoff method was used for the CAST two-point items. ETS calculated the average judgement or rating made by the panelists for each item, and then summed the averages to obtain a panel-recommended raw threshold score. Panelists in the CAST standard setting workshop made three rounds of judgements for the borderline students at Level 2, Level 3, and Level 4 of the ALDs. They also provided feedback and discussed their judgements between each round. This standard setting process led to the panel-recommended threshold scores.

Following the CAST July/August 2019 standard setting workshop, ETS reported the educator panel-recommended threshold scores to the CDE for review. Psychometricians from the CDE and selected CAASPP TAG members will complete a review of these recommendations, along with other data, which will inform the recommended threshold scores that will be presented to the SBE in November 2019.

For detailed information on the standard setting process used in this standard setting workshop, see Attachment 1.

### **Next Steps for CAST Threshold Scores**

In November 2019, the CDE will bring the recommended CAST threshold scores to the SBE for approval. If the SBE approves these recommended CAST threshold scores, LEAs will receive electronic SSRs to send to parents/guardians and students. Summary results will be made available on the public reporting website at a future date.

In addition to the SSRs, LEAs also will be able to access student results in a downloadable student-level data file in the Test Operations Management System. The CDE has created a CAST—Key Messages web page to assist LEAs with consistent key messages about the CAST. The web page is located at <https://www.cde.ca.gov/ta/tg/ca/castkeymessages.asp>. Once available, the statewide-, county-, LEA-, and school-level summary results by student groups will be displayed on the new CDE Public Web Reporting website.

## Attachment(s)

* Attachment 1: California Science Test Standard-Setting Plan (17 Pages)



**California Science Test** **Standard-Setting Plan**

**Contract #CN150012**

**Prepared for the California Department of Education by Educational Testing Service**

**Revised July 15, 2019**



**Table of Contents**

[**Background 3**](#_Toc14070672)

[**Description of the Standard Setting 4**](#_Toc14070673)

[Purpose 4](#_Toc14070675)

[General Description of the Process 4](#_Toc14070676)

[Panelists 4](#_Toc14070677)

[**Standard-Setting Materials 5**](#_Toc14070678)

[Test Familiarization Materials 5](#_Toc14070679)

[CAST General and Range ALDs 5](#_Toc14070680)

[Judgment Materials 6](#_Toc14070682)

[Impact Data 6](#_Toc14070683)

[Evaluation Forms 6](#_Toc14070684)

[**Standard-Setting Process 7**](#_Toc14070685)

[Test Familiarization 7](#_Toc14070686)

[Defining the Borderline Student 8](#_Toc14070687)

[**Standard-Setting Methodology 9**](#_Toc14070688)

[Item Scoring, Judgments, and Rating Scales 9](#_Toc14070689)

[Modified and Extended Angoff Judgments 9](#_Toc14070690)

[Feedback and Discussion 11](#_Toc14070692)

[**Recommendations and Technical Report 13**](#_Toc14070693)

[**Staffing, Logistics, and Security of Panel Meetings 14**](#_Toc14070694)

[**Appendix A. Sample Agenda 15**](#_Toc14070695)

[Day 1 15](#_Toc14070696)

[Day 2 15](#_Toc14070697)

[Day 3 16](#_Toc14070698)

[**References 17**](#_Toc14070699)

**Background**

The California Science Test (CAST) is an online assessment aligned with the California Next Generation Science Standards (CA NGSS) (California Department of Education [CDE], 2019).The CAST is required for all eligible students in grades five and eight and once in high school in grade ten, eleven, or twelve. (Students assigned to take an alternate assessment based on their Individualized Education Program needs should take the California Alternate Assessment for Science.) The CAST includes stand-alone, or discrete, items and performance tasks (PTs). The discrete item types include selected response, constructed response, grid, match, inline choice, zone select, etc. The discrete items measure a student’s ability to apply scientific knowledge and skills across the standards. The PTs also include a variety of item types; PTs measure a student’s ability to integrate knowledge and skills across multiple standards through extended activities.

The CAST uses the California Assessment of Student Performance and Progress (CAASPP) test delivery system. The first operational administration of the CAST occurred during the 2018–19 CAASPP administration. Standard setting is required so that threshold scores and achievement levels will be available for the fall 2019 release of results in CAST score reports.

The general, or policy, achievement-level descriptors (ALDs) were approved by the State Board of Education (SBE) in November 2017. General ALDs describe the expectations for student achievement at each reporting level across all grades tested. From March 26–28, 2019, 18 California educators convened in Sacramento to review and provide input in the draft descriptions of the range ALDs. The range ALDs are grade- and content- specific descriptors that incorporate the three dimensions (i.e., Disciplinary Core Ideas [DCIs], Science and Engineering Practices [SEPs], and Crosscutting Concepts [CCCs]) of the CA NGSS. These range ALDs are organized by DCI strands and include the relevant SEPs and CCCs in their descriptions. Additionally, these descriptions include the cognitive and content rigor that are encompassed within particular achievement levels. The range ALDs specifically describe the CA NGSS knowledge and skills expected of students in grades five, eight, and high school at each achievement level. The range ALDs will be used to inform the standard-setting process.

Threshold ALDs, also called borderline student definitions, are developed and used during the standard-setting process. Panelists describe what is expected at the beginning of three achievement levels (Levels 2, 3, and 4), and use these ALDs to make standard-setting judgments. Three threshold scores allow four achievement levels to be reported. Achievement levels for the CAST are as follows:

Level 4—Standard Exceeded

Level 3—Standard Met

Level 2—Standard Nearly Met

Level 1—Standard Not Met

The reporting ALDs describe the meaning of the levels on score reports and are anticipated to be finalized in November 2019, following the SBE meeting.

**Description of the Standard Setting**

**Purpose**

The CAST standard setting will result in recommendations for the CAST threshold scores, which are the minimum scores at Level 2, Level 3, and Level 4. The CDE will review these recommendations, along with additional data, and present them for approval to the SBE in November 2019.

**General Description of the Process**

The approach used in this study adheres to the guidelines and best practices recommended in the standard-setting literature; specifically, the modified Angoff and extended Angoff standard-setting methods (Cizek & Bunch, 2007; Hambleton & Plake, 1995) will be implemented. Panelists will make three rounds of judgments, with feedback and discussion occurring between each round.

**Panelists**

A diverse group, representative of science educators in grades five and eight and high school in California, will be recruited to participate as panelists in the standard-setting sessions. In recruiting panelists, the goal will be to include California educators who are familiar with the breadth and depth of the CA NGSS and who have experience in the science education of the students in grades three through twelve, with an emphasis on those with direct experience teaching students in grades five and eight and high school (i.e., grades ten through twelve) who will take the CAST. It is important to include teachers working with these students, as they will provide a perspective on the students’ knowledge and skills related to the CA NGSS.

Panelists will be assigned to one of three panels of educators; each panel will focus on one grade-level CAST: grade five, grade eight, or high school. The targeted number of panelists from this population of educators is 15 per panel, for a total of 45 educators. The decision on the panelists’ selection and panel assignment will be made by the CDE. Panelists will be notified after they are approved, and the Sacramento County Office of Education (SCOE), which is assisting with meeting logistics, will assist with their travel arrangements, if requested. Panelists will be required to sign a security agreement notifying them of the confidentiality of the materials used in the standard setting and prohibiting the removal of the meeting materials from the meeting rooms.

**Standard-Setting Materials**

The materials listed will be provided to panelists for each CAST assessment. The manner in which these materials are used are described more fully in the [Standard-Setting Process](#_Standard-Setting_Process) and [Standard-Setting Methodology](#_Standard-Setting_Methodology) sections.

Workshop agenda ([appendix A](#_Appendix_A._Sample))

Test familiarization materials

Printed format of operational CAST items administered in 2018–19 (for writing answers and notes)

Answer key with scoring rules and rubrics, where appropriate (described later

CAST grade- and content-specific range ALDs (excerpts provided in the premeeting materials; the full set of ALDs provided at the workshop)

Judgment materials

Survey forms on tablets, one per panelist

Printed judgment form

Training materials (e.g., printed format of CAST training items)

Impact data based on the 2018–19 administration of the CAST

Standard-setting Evaluation forms

Training evaluation form

Final evaluation form

**Test Familiarization Materials**

Panelists will receive materials to become familiar with the test content and the items they will be evaluating during the standard-setting study. All items administered in 2018–19 will be evaluated in the standard-setting process. The facilitator will share the difference between what panelists will see and the structure and length of the test students took in 2018–19. The panelists will “take the test,” reviewing all items without an answer key, and will later self-score using a provided answer key. During the test-familiarization process, the computer-administered version of the test will be displayed as panelists follow along on a printed version, make notes, and respond to the questions.

Answer keys differ by task type; specifically, some items include one or more selected responses; others require an extended response. Panelists will be provided rubrics for extended-response items as well as the scoring rules for items that have multiple parts, requiring different response types (e.g., students must respond to selected-response and grid parts of a single item).

**CAST General and Range ALDs**

The general ALDs provide the meaning of achievement levels across grades. Range ALDs define the achievement levels specific to each CAST grade level and science domain.

**Judgment Materials**

During the training round, the panelists will make their judgments on items from the CAST training test. There will be a variety of task types provided during the training round to familiarize the panelists with the types of judgments they will make. The training process, which includes learning how to make judgments on online forms using tablet computers, is described in the [Standard-Setting Methodology](#_Standard-Setting_Methodology) section.

During the operational rounds of judgments, panelists will again enter their judgments on online forms, using tablet computers. For both the training round and the operational judgment rounds, materials will include a printed form where panelists can make notes prior to entering judgments on the survey.

**Impact Data**

After the second and third rounds of judgment, panelists will learn the percentage of students from the 2018–19 administration who would fall into each of the four achievement levels, if the recommendations at that point were applied. This impact data will be based on the panel-recommended threshold scores for Round 2 and Round 3 judgments.

**Evaluation Forms**

Following the training round, the panelists will be asked to complete a training evaluation, where they can indicate their understanding of the judgment-making process. At the conclusion of the workshop, panelists will be asked to complete a final evaluation of the entire process.

**Standard-Setting Process**

A preworkshop assignment, consisting of two parts, will be given to the panelists approximately two weeks before the in-person workshop. For the first part, panelists will be provided with a link to the CAST training test and asked to take a training test to become familiar with the item types and task types on the CAST. The assignment will include directions for finding the answer key to the test items. The second part of the assignment will include reviewing the general ALDs, reading excerpts from the rangeALDs for the panelists’ assigned grade, and making notes on a notetaking form.

One disciplinary core idea (DCI) strand from each domain, including Physical Sciences, Life Sciences, or Earth and Space Sciences, will be provided for the assignment. Panelists will be asked to consider the expected performance of a student at each of the achievement levels, take notes about the knowledge and skills of students *at the beginning* of Level 3, and bring those notes to the standard-setting workshop.

Once on site at the workshop, all panelists will attend a general session that will include an overview of the CAST and the standard-setting procedure. At the conclusion of the general session, panel facilitators experienced in working with educators in standard setting will provide in-depth training and practice on the method in the panel rooms. Panelists will then complete three rounds of judgments. Feedback and discussion will take place after each round of judgment. Each panel will complete the standard-setting process on one grade-level assessment.

**Test Familiarization**

Immediately following the general training session, panelists will break into their assigned panels associated with the grade-level assessment for which they will be setting standards. Panelists will become familiar with the CAST items and discuss the test content before making any judgments.

During the test-familiarization process, standard setting staff will present the computer-administered assessment to the room. The panelists, using their paper forms, will follow along and record their responses to the items; as if they were “taking the test,” without discussion. Next, panelists will independently score their responses against the answer key. They will then discuss the content measured on the test as a whole, what content they think might be particularly challenging for students, and what might be less difficult. The goal of this activity is for panelists to begin to think about and articulate their perception of the general difficulty of the tested content for students.

Once the panelists are familiar with the content of the assessment, they will begin discussing the notes taken in the preworkshop assignment, specifically articulating the knowledge and skills necessary for students to reach achievement Level 3, using the range ALDs. The initial focus will be on one DCI strand that was included in the preworkshop assignment and represents the initial training for panelists to define the borderline students.

**Defining the Borderline Student**

Developing definitions of borderline students is a critical component of any standard-setting workshop. The process to arrive at borderline student definitions will involve small-group discussions and the development of draft borderline-student definitions, followed by a whole-panel discussion of the draft definitions, in order to reach a panel consensus of what is expected.

For the CAST, three definitions will be needed for three thresholds—the Level 2 borderline, Level 3 borderline, and Level 4 borderline student definitions. Panels will work first on the Level 3 borderline, because this is the point at which students will be classified as having *met standard,* demonstrating that they can apply their knowledge and skills of the CA NGSS to problems in each of the CAST domains.

Panelists will refer to the range ALDs that describe the full range for three levels. The borderline Level 3 student will be defined by considering what is expected of students in Level 2, compared with expectations in Level 3, and describing what more the student just entering Level 3—the borderline Level 3 student—can do compared to the highest-performing student in Level 2. Educational Testing Service (ETS) facilitators will instruct panelists on the process and will provide an exemplar in the form of a few bullets to exemplify the level of detail expected in a borderline student definition. Panelists will be asked to review and discuss the exemplar and modify it as needed. Instructions to panelists will include a goal to limit the definitions of their borderline students to a sufficient, but not all-encompassing, description.

After the borderline Level 3 student definition is drafted and consensus is reached, panelists will complete the borderline Level 2 and Level 4 student definitions. Once the borderline student definitions are completed, printed versions of the definitions will be distributed to the panelists for use when making judgments.

**Standard-Setting Methodology**

Following the completion of the borderline student definitions, panelists will be trained in the judgment-making procedure. They will have an opportunity to practice making judgments prior to the start of actual standard-setting rounds of judgments, as described in the next subsection.

During training, panelists will make judgments on online forms using tablet computers; two ETS staff members will be available to help panelists as needed in the use of the online survey. There will be a variety of items and judgment types in the training test; the panel facilitator will provide feedback on the panel’s judgments, and panelists will discuss rationales they used for making judgments. In this exercise the standard-setting panel facilitator is listening for rationales and comments that are consistent with the training and will redirect if comments are inconsistent with the instructions provided. After training, panelists will be asked to sign a training evaluation form confirming their understanding and readiness to proceed; additional training will be provided as needed.

Panelists will make three rounds of judgments for each of the achievement levels. The first round of judgments, Round 1, will be made independently, without discussion. Panelist judgments made in Round 2 and Round 3 will include consideration of data provided as feedback, as well as panelist discussions of the data and their rationales, as described in the [Feedback and Discussion](#_Feedback_and_Discussion) section. The steps in this process, including training and practice, judgments, and feedback, may also be found in the agenda (refer to [appendix A](#_Appendix_A._Sample)).

**Item Scoring, Judgments, and Rating Scales**

There are multiple item types on the CAST, which solicit a variety of response types. One important goal in standard setting is to reduce the cognitive complexity of making judgments; instructions to the panelists need to be clear and understandable. The more difficult the judgment task, the less accurate (and meaningful) is the panelists’ decision. Instructions and judgments will be more intuitive for the panelists if the ratings are aligned to the scoring rules.

For the CAST, the standard-setting process will include both the Modified Angoff judgment procedure (Brandon, 2004; Hambleton & Pitoniak, 2006) and the Extended Angoff procedure (Cizek & Bunch, 2007; Hambleton & Plake, 1995). Using these two judgment types allows panelist judgments to align with the scoring of the item. Rubrics and scoring rules will be provided to panelists, along with the answer key.

**Modified and Extended Angoff Judgments**

For all one-point items, the Modified Angoff method will be implemented. For items scored as two-point items, the Extended Angoff method will be implemented. One-point items include discrete item types such as selected- or constructed-response items (e.g., grid, fill-in-blank, and graphing). The two-point items include extended response and composite items with complex scoring rules that indicate how a student will obtain a score of 0, 1, or 2.

***Modified Angoff***

The Modified Angoff method is a probability-based standard setting method. For one-point items, each panelist will judge the item on the likelihood that the borderline student would answer the item correctly. Panelists will make their judgments using the following rating scale: 0, .05, .10, .20, .30, .40, .50, .60, .70, .80, .90, .95, 1. If an item is judged as difficult for the borderline student, the panelist would select a lower value, meaning that the probability the borderline student would answer the item correctly is low. If another item is judged as being easier for the borderline student, a higher value would be selected, indicating it is more likely that the borderline student would answer the item correctly.

Panelists will receive the printed version of the borderline student definitions and will be instructed to refer to those definitions in comparison to the demands of the item for each judgment. They will be advised to approach the Modified Angoff judgment-making process in two stages. First, they will consider what is the most likely probability range that the borderline student would answer the question correctly. The facilitator will encourage the panelists to consider the following rules of thumb to guide their decisions:

Items in the 0 to .30 range are those that the borderline student would have a low chance of answering correctly.

Items in the .40 to .60 range are those that the borderline student would have a moderate chance of answering correctly.

Items in the .70 to 1 range are those that the borderline student would have a high chance of answering correctly.

Second, panelists will decide how to refine their judgment within the range. For example, if a panelist thinks that there is a high chance that the borderline student would answer the question correctly, the initial decision will be in the .70 to 1 range. The second decision for the panelist will be to judge if the likelihood of answering it correctly is .70, .80, .90, .95, or 1.

Panelists will be asked to make three judgments for each item. The overall instructions will include a reminder that, when making Level 2 (L2), Level 3 (L3), and then Level 4 (L4) judgments, it is expected that each judgment value must be at least the same as the value of the level below, for each item. For example, if the borderline L2 judgment is .30, then the borderline L3 judgment must be .30 or higher. For Extended Angoff judgments, the same applies: the borderline L3 judgment must be the same, or higher, than the L2 judgments. Note that the judgments are made on tablets, and the software requires that judgments are the same or higher as the level before.

***Extended Angoff***

An Extended Angoffmethod (Cizek & Bunch, 2007; Hambleton & Plake, 1995) will be used for the two-point items. For these items, a panelist will decide on the assigned score value that would most likely be earned by the borderline student for each constructed-response item. Panelists will be asked to first review the definition of the borderline student and then to review the item and its scoring rubric. The rubric for an extended-response item defines, holistically, the quality of the evidence that would merit a response earning a particular score. The scoring rules for two-point composite items describe what responses are required to achieve one point and what responses are required to achieve two points.

During this review, each panelist will independently consider the level of knowledge and skill required to respond to the item as well as the features of a response that would earn a particular score as defined by the scoring rubric. Each panelist will decide on the score most likely to be earned by each borderline student from the possible values a student can earn. Panelists will be reminded to refer to the knowledge and skills of the borderline student definition and the scoring rules and not to expect the three levels to match to the three possible scores. For the three judgments—L2, L3, and L4—each higher level should have the same or higher expectation.

**Feedback and Discussion**

The purposes of feedback and discussion are to allow panelists to see the range of judgments across the panel, hear the rationales of the other panelists, receive total score information about student performance (impact data), and arrive at a mutual understanding of the expectations of borderline students’ performance on this test. Discussions are encouraged at the panel level, as well as at each table. Each panelist will be seated at one of two tables per panel. This table format facilitates small group discussion and provides an opportunity for more panelists to provide input. The process of judgment, feedback, and discussion will be repeated during the entire standard-setting workshop until all threshold score recommendations have been collected.

***Round 1 Feedback and Discussion, and Round 2 Judgments***

The feedback and discussion from the Round 1 judgment data will inform Round 2 judgments. As part of the post-Round 1 feedback, panelists will receive their own judgments in the context of the range of judgments across the panel, and the facilitator will share feedback on the similarity and differences of the panel judgments on the CAST items.

The feedback on judgments is displayed as a summary table with the mean, minimum, and maximum judgment for each level and also as three ranges, that is, the number of panelists rating an item from 0 to .30, the number of panelists rating an item from .40 to .60, and the number of panelists rating an item from .70 to 1.0. This is helpful for panelists to see where there is agreement and is also consistent with the instructions on how to think about the full range of 0 to 1.0 when making judgments.

Feedback for item judgments based on Extended Angoff judgments will be displayed as 0, 1, and 2. For both types of judgments, items for which two thirds, or 67 percent, of panelists are in the same range for a level will be considered as close to agreement; items with a more widespread distribution of judgments will be targeted for further discussion.

Panelists will be encouraged to discuss their judgments and rationales. During discussion, panelists will make notes on their own judgment forms and enter independent Round 2 judgments for any judgments they want to change at the end of the discussion.

***Round 2 Feedback and Discussion, and Round 3 Judgments***

After the Round 2 data has been analyzed, panelists will then engage in another round of feedback and discussions, based on the Round 2 data, before making their third and final round of judgments in Round 3.

Panel results will again be projected in each panel room, including summary statistics of the panel’s threshold scores and the panel’s range of judgments. After the panel has discussed this data, the facilitator will present impact data, based on student performance on the 2018–19 CAST administration. This feedback data will show what percentage of students would fall into each level based on the Round 2 panel mean threshold scores. After the room-level discussions, panelists will be invited to continue with table-level discussions as needed.

***Final Feedback***

Once all discussions have concluded, panelists will be asked to make their final round of judgments. The results from the Round 3 judgments will be considered the final threshold score recommendations from the standard-setting panel. Panelists will review Round 3 feedback, in the same format as the Round 2 feedback, and will respond to a final, confidential evaluation form.

**Recommendations and Technical Report**

ETS will deliver the recommended threshold scores and the data files containing score distributions for grades five and eight and the high school grade span to the CDE on Monday, August 12, 2019. In addition to the recommended threshold score tables designed for presentation to the SBE, additional tables may be developed by the CDE and verified by data analysts at ETS. Further discussion between ETS and the CDE to define the composition of the tables and a timeline for delivery will be required prior to the standard setting.

ETS will produce and deliver the final technical report for the standard setting by Friday, October 4, 2019. The technical report will contain a description of the process used to set standards, a description of the panelists’ qualifications, results presented during the standard-setting process, and statistical information related to the threshold score judgments (i.e., two standard errors of judgment and two standard errors of measurement above and below the panel-recommended threshold score).

**Staffing, Logistics, and Security of Panel Meetings**

The CAST standard-setting workshop will be held from July 31 to August 2, 2019, at SCOE in Mather, California. Dr. Patricia Baron, the standard-setting director at ETS, will conduct a walk-through of the process for the CDE on July 16, 2019.

To allow the standard-setting meetings to run smoothly, all groups will be led by trained and experienced standard-setting facilitators who will conduct the training, facilitate the process, and keep the discussions on track. The panel facilitators will respond to any standard-setting process questions.

Dr. Baron will lead the introductory training session and will oversee the workshop process. In addition, ETS will provide two assessment development specialists who will be available to respond to questions about items and other materials. ETS will also provide two data analysts and a lead facilitator experienced in the Angoff standard setting process, as well as a psychometrician for the duration of the workshop. Representatives of ETS’ Program Management staff, Mary Anne Arcilla, Executive Director, and Kelly Bolton, Associate Director, also will attend the sessions and be available to the CDE as needed.

All logistics and panelists’ travel concerns will be addressed by SCOE. ETS understands that CDE staff will be present during the standard-setting sessions to hear discussions, observe the process, and address any policy-level issues, as appropriate.

Groups will be provided with numbered materials on the first day at the time of registration and as needed during the three-day process. At the end of the process, ETS staff will collect and securely destroy all confidential materials.

**Appendix A. Sample Agenda**

All times are approximate and may change based upon the panel’s progress.

**Day 1**

|  |  |
| --- | --- |
| **Time** | **Event** |
| 7:30 a.m. | Registration and breakfast |
| 8:15 a.m. | Welcome and general session |
| 9:45 a.m. | Morning break |
| 10 a.m. | Training and test familiarization |
| 12:15 p.m. | Lunch |
| 1:15 p.m. | Develop draft borderline student definitions |
| 3:00 p.m. | Afternoon break |
| 5 p.m. | Reach consensus on borderline student definitions |
| 5 p.m. | End of Day 1 |

**Day 2**

|  |  |
| --- | --- |
| **Time** | **Event** |
| 7:30 a.m. | Sign-in and breakfast |
| 8:15 a.m. | Goals for today |
| 8:30 a.m. | Complete training and practice on Modified Angoff and Extended Angoff methods; begin Round 1 standard setting judgments |
| 9:45 a.m. | Morning break |
| 10 a.m. | Continue Round 1 standard-setting judgments |
| 12:30 p.m. | Lunch |
| 1:30 p.m. | Complete Round 1 standard-setting judgments |
| 2:45 p.m. | Afternoon break |
| 3 p.m. | Review Round 1 feedback and begin considering Round 2 judgments |
| 5 p.m. | Finish Day 2 |

**Day 3**

|  |  |
| --- | --- |
| **Time** | **Event** |
| 7:30 a.m. | Sign-in and breakfast |
| 8:15 a.m. | Goals for today |
| 8:30 a.m. | Complete Round 1 feedback and complete Round 2 judgments |
| 9:30 a.m. | Break |
| 10 a.m. | Begin Round 2 feedback (including impact data) and begin Round 3 judgments |
| 12:30 p.m. | Lunch |
| 1:30 p.m. | Complete Round 2 feedback and complete Round 3 judgments |
| 2:45 p.m. | Afternoon break |
| 3:15 p.m. | Review feedback on Round 3 recommended threshold scores |
| 4 p.m. | Complete final evaluation |
| 4:15 p.m. | Wrap up the meeting, including security steps |
| 4:30 p.m. | End of CAST standard setting |

**References**

Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard-setting topics. *Applied Measurement in Education*, 17, 59–88.

California Department of Education. (2019). *NGSS for California public schools, K–12.* Sacramento, CA: California Department of Education. Retrieved from <https://www.cde.ca.gov/pd/ca/sc/ngssstandards.asp>

Cizek, G. J. (Ed.). (2012). *Setting performance standards: foundations, methods, and innovations.* Abingdon-on-Thames, UK: Routledge.

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

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.). *Educational Measurement* (4th ed., pp. 433–70). Westport, CT: Praeger.

Hambleton, R. K., & Plake, B.S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education*, 8, 41–55.