Standard Setting Technical Report for the California Science Test

Prepared for the California Department of Education by Educational Testing Service

Presented November 18, 2019

Table of Contents

List of Tables

List of Figures

Acronyms and Initialisms Used in the *Standard Setting Technical Report for the California Science Test*

Abbreviation	Definition
CAST	California Science Test
CDE	California Department of Education
CSEM	conditional standard errors of measurement
ETS	Educational Testing Service
ALD	achievement level descriptor
SBE	State Board of Education
SEJ	standard error of judgment
SS Scale	Standard Setting Scale
SSPI	State Superintendent of Public Instruction

This page is left blank intentionally.

Chapter 1: Introduction

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 students in grades five and eight and once in high school (grade ten, eleven, or twelve). The CAST includes stand-alone, or discrete, items and performance tasks (PTs). The discrete item types include, for example, selected response, constructed response, table, fill-in, and graphing. The PTs measure a student’s ability to integrate knowledge and skills across multiple standards through extended activities.

The CAST uses the current 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 CAST will report four achievement levels—Levels 1 through 4. These are described by achievement level descriptors (ALDs) of four types: general ALDs, range ALDs, threshold ALDs, and reporting ALDs. Prior to the standard setting, CAST general ALDs were presented and approved at the November 2017 State Board of Education (SBE) meeting. The general ALDs describe expectations for each achievement level; they are descriptions at a very high level and are typically used by policy makers. 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

From March 26–28, 2019, 21 California educators convened in Sacramento to define range ALDs. They reviewed and provided input on the draft Range ALDs, which are descriptions of the CA NGSS knowledge and skills necessary for students in grades five, eight, and high school at each of four achievement levels. As one of the most critical parts of the standard setting process, participants referred to the range ALDs to define the threshold ALDs, which is the set of knowledge and skills expected of *borderline students* who are at the entry-point of each achievement level.

The reporting ALDs provide descriptions of the achievement levels on the student score reports; they are developed based on both the general and range ALDs. Reporting ALDs will be finalized in November 2019.

Figure 1 provides the SBE-approved score-reporting hierarchy, which applies to the CAST, for grades five, eight, and high school (CDE, 2017). The CDE recommended that additional scores should be investigated after the 2018–19 CAASPP administration.

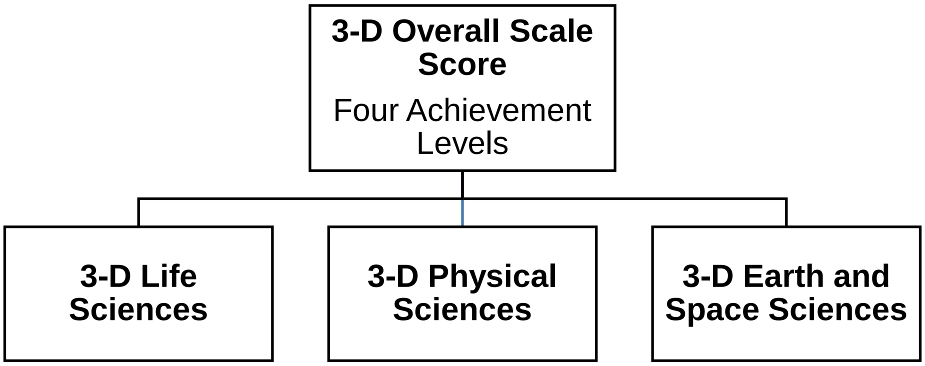

Figure 1. CAST Student Score Reporting Structure

To develop threshold-score recommendations aligned to the score-reporting hierarchy, Educational Testing Service (ETS) conducted a standard setting workshop for the three CAST assessments (grade five, grade eight, and high school) in Sacramento, California, on August 6–9, 2019. All items in the CAST item pool were considered in the process of standard setting. The Modified Angoff and Extended Angoff standard setting methods were applied, as appropriate.

For each grade, the standard setting panel recommended threshold scores to indicate the score that must be earned for a student to reach the beginning (i.e., threshold) of three of the four achievement levels (Levels 2, 3, and 4) for the CAST total score. California educators utilized the CA NGSS, the CAST General ALDs (CDE, 2017), and the range ALDs. The general ALDs were approved by the SBE on November 8, 2017; CAST range ALDs were reviewed and approved by the CDE following the educator panel review on May 24, 2019. A standard setting plan was approved by the CDE on July 26, 2019, in preparation for the meetings.

This document provides the following information:

- The purpose of the standard setting workshop and a discussion of the work conducted prior to the workshop
- An overview of the standard setting methods implemented, including discussions of the Modified and Extended Angoff methods used to develop the overall score thresholds
- A description of the panels and materials used in the approach, an overview of the process before and during the workshop, and a discussion of the training
- The results, including summary data from the panel judgments and evaluations by the panelists.

Purpose and General Description of the Standard Setting Workshop

The purpose of standard setting for the CAST was to collect recommendations for the placement of the CAST threshold scores for the CDE to review, with final approval by the SBE. For each assessment, there are four achievement levels (Levels 1 through 4). A threshold score defines the beginning of a higher level of performance or achievement. A review of the standard setting literature supports the need for attention to best practices (Brandon, 2004; Hambleton & Pitoniak, 2006; Tannenbaum & Katz, 2013), which include the following:

- A careful selection of panel members
- A sufficient number of panel members to represent varying perspectives and provide for replication
- Sufficient time devoted to developing a common understanding of the assessment domain
- Adequate training of panel members
- Development of a description of each achievement level
- Multiple rounds of judgments
- Inclusion of data, where appropriate, to inform judgments

The approach used in this study adheres to the guidelines and best practices; specifically, the Modified Angoff and Extended Angoff standard setting methods. These methods allowed for the collection of panelist judgments for each item administered in 2019–20, thereby providing flexibility in the development of threshold scores for reporting the overall score and possible domain scores.

Chapter 2: Method

Chapter 2 includes the following:

- Descriptions of the Modified and Extended Angoff Methods of standard setting
- Descriptions of the standard setting panels

Modified and Extended Angoff Methods

The Modified Angoff method (Brandon, 2004; Hambleton & Pitoniak, 2006) is a probability-based standard setting method. For one-point items, each panelist judged the item on the likelihood that the borderline student would answer the item correctly. Panelists made judgments using the following rating scale: 0, .05, .10, .20, .30, .40, .50, .60, .70, .80, .90, .95, 1. The lower the value, the less likely it is that the borderline student would answer the item correctly because the item is difficult for the borderline student. The higher the value, the more likely it is that the borderline student would answer the item correctly.

An Extended Angoff method (Cizek & Bunch, 2007; Hambleton & Plake, 1995) was used for the two-point items. For these items, the task was to decide on the assigned score value that would most likely be earned by the borderline student for each constructed-response item. Panelists were 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.

In standard setting, the critical components involve having a standard setting panel of experts who can provide appropriate consideration and judgments. The panel begins by becoming familiar with the test and considering the content assessed and the relative difficulty of the items. The test-familiarization stage also allows the panelists to experience the test in a manner that is similar to an operational test administration, which allows the panelists to get a sense of the test taker’s experience. After independently reviewing the assessment, the panelists discuss the content measured and the relative difficulty of the items.

Following a discussion about the test content and the students who would take the test, the panelists consider the different achievement level descriptors. The panelists work together in small and large groups to draft and reach consensus on the Level 3 borderline student definition followed by the borderline student definition for Level 2. These definitions are the operational description of the threshold scores and are used by the panelists as they make three rounds of judgments.

Prior to making judgments, panelists are trained and have an opportunity to practice using training materials. Once the training is completed and all panelists have indicated on the training evaluations a readiness to proceed, the first round of independent judgments takes place without discussion. Before the Round 2 and final Round 3 judgments take place, panelists are presented with feedback data on the panel judgments. Before Round 3, panelists also see impact data. Once the data is presented, panelists engage in table- and room-level discussions about the reactions to the data. The panelists also discuss the rationales behind the judgments as the next round of judgments is made. Presenting more information prior to each round of judgments allows the panelists to become more informed judges. Any subsequent adjustments the panelists' make to the judgments are refinements informed by new information, including the rationales of colleagues.

At the conclusion of the standard setting, a final evaluation is administered to obtain feedback concerning the panelists' perspectives on the standard setting procedures, instructions, and materials. In additional to procedural feedback, the panelists also provide their opinions of the final recommended scores.

Standard Setting Panels

A diverse group, representative of science educators in grades five and eight and high school in California, were recruited to participate as panelists in the standard setting sessions. In recruiting panelists, the goal was to include California educators who are familiar with the breadth and depth of the California Next Generation Science Standards (CA NGSS) and who have experience in the science education of the students in grades five and eight and high school (grade ten, eleven, or twelve) who will take the California Science Test (CAST). It was important to include teachers working with these students, as these teachers provided a perspective on the knowledge, skills, and learning goals for the students taking the CAST.

Panelists were assigned to one of three panels of educators; each panel focused on one grade-level CAST: grade five, grade eight, or high school. The targeted number of panelists from this population of educators was 15 per panel, for a total of 45 educators.

The decision on the panelists' selection and panel assignment was made by the California Department of Education (CDE). Panelists were notified after they were approved, and staff from the Sacramento County Office of Education (SCOE) assisted with any travel arrangements. Panelists were 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 materials from the meeting area.

The CAST panels were assembled into grade- and test-specific panel rooms for much of the standard setting work. Educators teaching students in the grade level or adjacent grade levels served on those panels. Educators teaching across the high school grade span served on the high school panel. Panelists were seated at two tables, with seven or eight educators at each table.

The number of panelists is displayed in Table 1.

Table 1. Number of Panelists

Panel	N
Grade Five	15
Grade Eight	15
High School	16
Total	**46**

Because standard setting is based on expert judgment—informed by student performance data—it is important that panelists collectively reflect the diversity of the educators working with students who take the assessment. Special efforts were made to assemble panels that were representative of the geographic and socioeconomic diversity of California in general and the CAST educator population in particular. Table 2 provides the distribution of the panel by gender; all panels included at least three male educators.

Table 2. Panelist Gender

Gender	Grade Five	Grade Eight	High School	Total
Female	12	10	7	29
Male	3	4	9	16
No Response	0	1	0	1

Table 3 provides the educators' responses regarding personal ethnic or racial background. The two largest groups represented were Hispanic (n = 10) and White (n = 18). All panelists except one responded to the question, "What is your primary ethnicity/race?"

Table 3. Panelist Primary Ethnicity/Race

Ethnicity/Race	Grade Five	Grade Eight	High School	Total
American Indian/Alaska Native	0	0	0	0
Asian	1	2	5	8
Black or African American	1	1	1	3
Filipino	0	1	0	1
Hispanic or Latino	3	5	2	10
Native Hawaiian or Other Pacific Islander	0	0	0	0
White	8	6	4	18
Two or More Races	1	0	4	5
No Response	1	0	0	1

Table 4 presents the location in which California educators are teaching. A majority of the educators reported working in the southern region of California. All panelists responded.

Table 4. Geographical Region of Panelists

Region	Grade Five	Grade Eight	High School	Total
Central	1	1	5	7
Northern	3	3	3	9
Southern	11	11	8	30

Table 5 presents the teaching experience of the educators in each panel and across the standard-setting workshop by the number of years taught. A majority of the educators indicated having had more than ten years of experience teaching science. All panelists responded.

Table 5. Panelist Years Experience Teaching Science

Experience	Grade Five	Grade Eight	High School	Total
1 to 3 years	0	0	0	0
4 to 6 years	1	3	3	7
7 to 10 years	1	1	1	3
10+ years	13	11	12	36

Table 6 presents the subject or subjects the educators currently teach; multiple responses were permitted. The responses indicate that all panels were primarily comprised of teachers currently teaching science. The grade five panel has the highest number of multiple subject teachers.

Table 6. Panelist Subject(s) Currently Teaching

Subject	Grade Five	Grade Eight	High School	Total
All Subjects	5	0	1	6
Mathematics	2	1	1	4
Science	8	12	15	35
Social Studies	0	0	0	0
English	1	0	0	1
Other	4	4	0	8

Table 7 presents the grade or grades educators currently teach; multiple responses were permitted. The responses show that all panels included a majority of educators who were currently teaching the grade (or the adjacent grade) corresponding to the CAST test specific panel to which they were assigned.

Table 7. Panelist Grade(s) Currently Taught

Grade	Grade Five	Grade Eight	High School	Total
3	3	0	2	5
4	4	0	2	6
5	7	0	2	9
6	1	1	2	4
7	0	8	2	10
8	0	6	2	8
9–12	0	2	16	18
Other	10	3	2	15

Table 8 presents the educators' years of experience working with the CA NGSS. The responses indicate that all panels included educators who have been working with the standards for more than four years.

Table 8. Panelist Experience Working with the CA NGSS

Years	Grade Five	Grade Eight	High School	Total
Not Applicable	0	0	0	0
1–2	1	3	3	7
3–4	1	2	4	7
4+ years	13	10	9	32

Educators were asked about teaching experience with students across the general education, English learner, and special education populations. Table 9 indicates that in all panels, educators have teaching experience with students in general education, students who are English learners, and students in special education.

Table 9. Does your teaching experience include students from these populations?

Population	Grade Five	Grade Eight	High School	Total
General education	15	15	16	46
English learners	15	15	16	46
Special education	15	15	14	44

Chapter 3: Materials

At the standard setting workshop, panelists received training materials and a set of operational materials. Items were kept secure by assigning panelists an individual identification number and giving each panelist a set of materials marked with that same number. Each panelist was asked to sign a nondisclosure agreement, check the material out and in each day, and accept responsibility for controlling all documents labeled with the panelist's identification number. The Sacramento County Office of Education (SCOE) and Educational Testing Service (ETS) staff monitored each room to ensure that materials remained in the rooms and that no room was left unattended when unlocked.

Materials and data were based on the pool of items administered in 2018–19. For each California Science Test (CAST) assessment, the following materials were provided to each panelist:

- CAST general and range achievement level descriptors (ALDs)
- Test familiarization materials
  - Paper assessment for entering answers and notes
  - Answer key with scoring rules where appropriate
  - Rubrics
- Judgment materials
  - Survey forms on tablets, one per panelist
- Practice and training materials
- Impact data based on the 2018–19 administration of the CAST
- Evaluation forms
  - Training evaluation form
  - Final evaluation form
- Workshop agenda

Panelists developed borderline student definitions in the workshop; refer to attachment B in appendix 1. The test familiarization materials, judgment materials, and impact data are described more fully in the next subsection.

Test Familiarization Materials

Panelists received materials to become familiar with the test content and were instructed to "take the test" and then self-score using a provided answer key. Operational items administered in 2018–19 were used for all grades. During the test-familiarization process, the computer-administered version of the test was displayed as panelists followed along on a printed version, made notes, and responded to the questions.

Panelists received the answer key for all items. CAST answer keys differ by task type; specifically, some items include one or more selected responses; others require an extended response. Panelists were provided with rubrics for extended-response items as well as the scoring rules for items that have multiple parts, and received instructions on how to use the key, rubrics, and scoring rules.

Judgment Materials

During the practice round, the panelists completed entering judgments on a practice judgment form that included items from the CAST training test. A variety of task types were included in the practice round to familiarize the panelists with the types of judgments that would be made in the operational rounds.

Following the practice round, the panelists were asked to complete an evaluation of the training on how to make judgments and then began making operational judgments. During the operational rounds of judgments, panelists entered the judgments on online forms using tablet computers.

Impact Data

Between the second and third rounds of judgment, panelists learned the percentage of students from the 2018–19 administration that would fall into each of the four achievement levels, if the recommendations at that point were applied. The data was based on the panel-recommended score for Round 2 and Round 3 judgments.

Evaluation Forms

It is important to collect information from the panelists to document procedural validity (Cizek & Bunch, 2007; Hambleton & Pitoniak, 2006). Panelists received evaluation forms at two points during the process to gauge the panelists' understanding of the procedures and to gather other information (refer to attachment C in appendix 1 for the evaluation forms). Evaluations included questions about the following:

- Training
- Understanding the tasks
- The influence of different aspects of the standard setting process
- Panelists' beliefs about the final recommended threshold scores

Because ETS was interested in knowing as soon as possible if panelists were not satisfied with the level of training provided, the first evaluation form was given to the panelists at the end of the training to gauge the panelists' current understanding of the process and comfort level with the tasks to be performed. These training evaluation forms were analyzed immediately, and responses were reviewed by the panel facilitator and lead facilitator so the facilitators could review any tasks or materials that appeared to be unclear. In two panels, no additional review and discussion was needed; panelists indicated sufficient comfort with the process and a readiness to proceed. In the high school panel, two panelists asked for and received additional information and then indicated a sufficient comfort level and readiness to proceed. An overview of the results obtained through the evaluation forms is included in the Results section of this report.

Chapter 4: Process

This section of the report describes what occurred prior to and during the standard setting workshop.

Preparation and Training

Prior to the standard setting, a preworkshop assignment, consisting of two parts, was given to the panelists two weeks before the in-person workshop (refer to attachment A). For the first part, panelists were provided with a link to the California Science Test (CAST) training test and asked to take a training test to become familiar with the task types on the CAST. The assignment included directions for finding the answer key to the training test items. The second part of the assignment included reading excerpts from the range achievement level descriptors (ALDs) for the panelists' assigned grade, the general ALDs, and the California Next Generation Science Standards (CA NGSS).

One disciplinary core idea (DCI) strand for each domain, including Physical Sciences, Life Sciences, and Earth and Space Sciences, was provided for the assignment. Panelists were asked to consider the expected performance of a student in 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 attended a general session that included 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 provided in-depth training and practice on the method in each panel room. Panelists then completed three rounds of judgments, with feedback and discussion after each round of judgment. Each panel completed the standard setting process on one grade-level assessment.

General Session Training

Panelists were trained in various aspects of the process throughout the course of the workshop; training was often followed immediately by doing the task addressed in the training. On the first day, a general orientation session was held for the entire group where the need for threshold scores was explained, as were the roles and responsibilities of the staff from the California Department of Education (CDE), Educational Testing Service (ETS) and the Sacramento Office of Education (SCOE). Staff from the CDE were available throughout the process to answer questions about the policies surrounding the test; staff from ETS were available to answer questions about the test and the standard setting procedures.

Dr. Patricia Baron, ETS standard setting director, introduced the Modified Angoff and Extended Angoff Methods for setting threshold scores and presented the agenda and expectations for panel members' participation. Dr. Baron then continued the general session with initial training on the methods, after which panelists moved into test-specific groups, where the panel facilitators continued with training and guided the panelists through the rest of the standard setting activities.

Test Familiarization

Immediately following the general training session, panelists split into the assigned panels associated with the grade-level assessment for which standards were being set.

During the test-familiarization process, standard setting staff presented the computer-administered assessment to the room. The panelists, using the paper forms, followed along and recorded responses to the items. After independently completing the test, panelists checked the responses against the answer key. Next the group discussed the content measured, what the panelists thought might be particularly challenging for students, and what panelists thought might be less difficult. The goal of this activity was for panelists to begin to think about and articulate the perceived general difficulty of the tested content for students.

The next step was a facilitated discussion, starting with the preworkshop assignment, to begin to articulate the knowledge and skills necessary for students to reach achievement Level 3, using the range ALDs. The focus was on one DCI strand that was included in the preworkshop assignment; this activity represents the initial training for panelists to define the borderline students. Once the process was familiar to the group, the group continued to define the borderline students in smaller groups, followed by a room-level consensus-building process.

Borderline Student Definitions

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

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

Panelists reviewed the general ALDs in smaller groups and, after the facilitator familiarized the panel with the task, worked as a whole group to describe one aspect of the Level 3 borderline student. By working on one bullet as a whole group, the process was modeled, and the facilitator provided guidance to ensure that the focus was on the differentiation between Level 2 and Level 3 ALDs. The panel then worked in two small groups to complete the borderline student definitions for Level 3.

Panelists referred to the range ALDs that describe the full range for three levels. The borderline Level 3 student was 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. ETS facilitators instructed panelists to limit the definitions of the borderline students to a sufficient, but not all-encompassing, description.

After the borderline Level 3 student definition was drafted and consensus was reached, panelists completed the borderline Level 2 and Level 4 student definitions.

The borderline student definitions are provided in attachment B in appendix 1.

Panelist Judgments

Prior to the start of actual standard setting rounds of judgments, as described in the next subsection, panelists were trained and then practiced making judgments on six one-point and two-point item types. The practice round included a summary of the judgments and a discussion of rationales. After training, panelists were asked to sign a training evaluation form confirming an understanding of the procedures and a readiness to proceed; additional training was provided as needed. The standard setting process continued once the readiness of all panelists was confirmed.

Panelists made three rounds of judgments, rounds 1, 2, and 3, for each of the achievement levels. Round 1 judgments were made independently, without discussion. Following Round 1 judgments, panelists received feedback and participated in a discussion. As part of the post-Round 1 feedback, panelists reviewed the individual judgments in the context of the range of judgments across the panel, and the facilitator shared feedback on the similarity and differences of the panel judgments on the CAST items. Panelists discussed with the other panelists the rationales for the independent judgments.

The feedback and discussion from the Round 1 judgment data then informed Round 2 judgments. Panelists engaged in another round of feedback based on the Round 2 data, having table-level and room-level discussions before making a third and final round (Round 3) of judgments. After each round, panelists' judgments were collected, analyzed, summarized and shared with the panel.

Each panelist was seated at one of two tables of educators to facilitate discussion. This table format provided an environment more conducive to panelists sharing opinions and rationales, as some panelists might have been less inclined to speak or have had less opportunity to be heard in a large group.

During table-level discussions, each educator participated in a discussion of the rationales for the individual judgments. A room-level discussion followed, where panelists shared individual perspectives and the themes from the table-level conversation with the rest of the panel. This process provided an opportunity for the varied perspectives of the experts serving on the panel to be heard.

Item Scoring, Judgments, and Rating Scales

There are multiple item types on the CAST, which solicit a variety of response types. The CAST one-point items include discrete item types, and the two-point items include extended-response and composite items with scoring rules that indicate how a student will obtain a score of 0, 1, or 2.

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 panelist's decision. Instructions and judgments are more intuitive for the panelists when the ratings are aligned to the scoring rules.

ETS implemented the standard setting using two standard setting methods: Modified Angoff judgments for one-point items and Extended Angoff judgments for extended-response items and items with complex scoring rules, which required panelists to consider how students can obtain either one or two points. Using these two judgment types allowed panelist judgments to align with the scoring of the item. Scoring rubrics were provided to panelists, along with the answer key.

Modified and Extended Angoff Judgments

For items scored as one-point items, the Modified Angoff method (Brandon, 2004; Hambleton & Pitoniak, 2006) was used; for two-point items, the Extended Angoff method (Cizek & Bunch, 2007; Hambleton & Plake, 1995) was used. One-point items included discrete item types such as selected- or constructed-response items, such as table or sentence completion, and graphing. The two-point items included extended response and composite items with complex scoring rules that indicate how a student will obtain a score of 0, 1, or 2.

The Modified Angoff method is a probability-based standard setting method. For one-point items, each panelist was asked to judge the item on the likelihood that the borderline student would answer the item correctly. Panelists made a judgment using the following rating scale: 0, .05, .10, .20, .30, .40, .50, .60, .70, .80, .90, .95, 1. On this scale, the lower the value, the less likely it is that the borderline student would answer the item correctly because the item is difficult for the borderline student. The higher the value, the more likely it is that the borderline student would answer the item correctly.

The facilitator suggested to the panelists that the judgment process be approached in two stages. The first stage involved reviewing both the description of the borderline student and the item and then considering the probability that the borderline student would answer the question correctly. The facilitator encouraged the panelists to use the following rules of thumb to guide this decision:

- 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.

In the second stage, the task was to refine the judgment within the range. For example, if a panelist thought that there is a high chance that the borderline student would answer the question correctly, the initial decision would be in the .70 to 1 range. The second decision for the panelist was to judge if the likelihood of a borderline student answering it correctly is .70, .80, .90, .95, or 1.

Panelists were asked to make three judgments for each item. The overall instructions included a reminder that, when making Level 2 (L2), Level 3 (L3), and Level 4 (L4) judgments, it was 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 was .30, then the borderline L3 judgment had to be .30 or higher. For Extended Angoff judgments, the same applied: the borderline L3 judgment had to be the same, or higher, than the L2 judgments. Note that the judgments were made on tablets and the software required that judgments be the same or higher as the level before.

An Extended Angoff method (Cizek & Bunch, 2007; Hambleton & Plake, 1995) was used for the two-point items. For these items, a panelist decided on the assigned score value that would most likely be earned by the borderline student for each constructed-response item. Panelists were 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 defined, holistically, the quality of the evidence that would merit a response earning a particular score. The scoring rules for two-point composite items described what responses are required to achieve one point and what responses are required to achieve two points.

During this review, each panelist independently considered 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 decided on the score most likely to be earned by each borderline student from the possible values a student can earn. Panelists were 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 needed to have the same or higher expectation.

Feedback and Discussion

The purposes of feedback and discussion was to allow panelists to hear the rationales of the other panelists, receive empirical information about student performance, and arrive at a mutual understanding of the expectations of borderline students' performance on this test. The process of judgment, feedback, and discussion was repeated over the entire standard setting workshop until all threshold score recommendations were collected.

Panelists were provided with the judgments for all items and received feedback after Round 1 judgments were collected and summarized. The mean, minimum, maximum, and range of the panel judgments (from low to high) were projected in the room. Panelists reviewed item-level judgment information for each item; the facilitator projected, for the room, a presentation that identified where panelists were closer to consensus in the judgments for some items and where judgments were more diverse for other items. Panelists were encouraged to discuss the judgments and rationales. Panelists made notes on the judgment forms and entered independent Round 2 judgments for any items which the panelist wanted to change.

After making Round 2 independent judgments, results were again projected in each panel room, including summary statistics of the panel's threshold scores and the panel's range of judgments. After the panelists discussed the data, the student performance data showing the impact, or consequence, data for the Round 2 judgments was presented. This performance data was based on 2018–19 CAST student performance. The feedback showed what percentage of students would fall into each level based on these decisions.

After the room-level discussions, panelists were invited to continue with table-level discussions as needed.

Once all discussions were concluded, panelists were asked to make a final round of judgments. The results from the Round 3 judgments were considered the final threshold score recommendations from the standard setting panel. Panelists reviewed Round 3 feedback and responded to a final, confidential evaluation form.

Chapter 5: Results

This section describes the results from the workshop, which include the item judgments and total score recommendations, the impact data based on student performance, and an evaluation of the process based on questionnaires completed by the panelists.

Data for each panel is presented in this section. Six types of tables are presented; a general description of the six types follows:

Five Types of Data Tables

1. **Mean threshold scores, by round.** Mean raw score threshold scores are presented. The range of possible scores is equal to the number of possible points in the pool of items administered in 2018–19 to students in each grade or grade span; the test includes one-point and two-point items.
2. **Standard errors of judgment (SEJs), by round.** SEJs are presented in the raw score metric, based on the panelists' judgments, in Table 13 through Table 15.
3. **Round 3 raw score judgments +/-1 SEJ and +/-2 SEJs**. The range around the final recommended threshold score is presented in Table 16 through Table 18. SEJ is one way of estimating the reliability or consistency of a panel's standard setting judgments and may be used as guidance when evaluating the appropriateness of threshold scores.
4. **Projected distribution of 2018–19 CAST students, shown as the percent, at each level based on the Round 2 recommended threshold scores on the standard setting scale score metric.** Panelists were provided with the impact data after Round 2 for consideration prior to the final, Round 3 judgments. Refer to Table 19 through Table 21.
5. **Projected distribution of 2018–19 CAST students, shown as the percent, at each level based on the Round 3 recommended threshold scores on the total standard setting scale score metric.** Panelists were provided with the impact data after Round 3, the final round of judgments. Refer to Table 22 through Table 24.
6. **Projected percentage of 2018–19 CAST students at and above the Round 3 recommended threshold score, +/-1 conditional standard errors of measurement (CSEM), and +/-2 CSEM scores on the standard setting scale score metric.** Refer to Table 25 through Table 27.

Data Presentation

Panel threshold score recommendations were presented to panelists first as a raw score. The raw score is based on judgments on all items administered in the 2018–19 CAST operational administration. The test administration model for CAST included field test items, which meant that not all students were administered the same set of items. When panelists considered the impact data, this feedback was provided on a scale that was based on the underlying theta distribution of the total score for all students who took the CAST. Educational Testing Service transformed the raw scale to a scale score unique to each grade or grade span, with a range of approximately 100 points, via a linear translation of the theta scale. All scale score information included in this technical report is based on the working scale—the Standard Setting Scale (SS Scale).

CAST Threshold Score Results

Table 10 through Table 12 display the mean threshold scores after each round for each test. These raw scores were based on the complete item pool and are rounded up to the nearest whole number. The tables show how panelists moved the judgments across rounds. Lower numbers indicate a lower threshold score. Higher numbers translate to a higher threshold score; a higher threshold score means that more is required for a student to be included in the level.

For all three CAST tests, the mean raw threshold score decreased from Round 1 to Round 2 for all three levels. Round 2 to Round 3 changes differed by test. The grade five mean did not change from Round 2 to Round 3 for Level 2 and Level 3; however, the mean for Level 4 decreased one point. In grade eight, the Level 2 mean increased one point from Round 2 to Round 3, there was no change to the Level 3 threshold score, and there was a one-point decrease to the Level 4 threshold score. The high school mean threshold scores decreased by two points for Level 2 and Level 4, from Round 2 to Round 3, and by one point for Level 3.

Table 10. Mean Raw Score Threshold Scores at the End of Each Round: Grade Five

Level	Round 1	Round 2	Round 3
Level 2	24	19	19
Level 3	49	39	39
Level 4	66	55	54

Table 11. Mean Raw Score Threshold Scores at the End of Each Round: Grade Eight

Level	Round 1	Round 2	Round 3
Level 2	22	15	16
Level 3	56	41	41
Level 4	74	57	56

Table 12. Mean Raw Score Threshold Scores at the End of Each Round: High School

Level	Round 1	Round 2	Round 3
Level 2	21	15	13
Level 3	42	31	30
Level 4	55	46	44

Table 13 through Table 15 provide estimates of the standard error of judgment (SEJ) after each round by panel. The SEJ is one way of estimating the reliability or consistency of a panel's standard setting judgments. Lower numbers from Round 1 to Round 3 indicate the convergence of panelists' judgments over rounds during the process. Ideally, the SEJ should decrease across rounds; although, occasionally, the introduction of impact data will result in the SEJ increasing, as panelists have different reactions to the normative data. An SEJ assumes that panelists are randomly selected and that standard setting judgments are independent. It is seldom the case that panelists are randomly sampled, and only the first round of judgments may be considered independent. The SEJ, therefore, likely underestimates the uncertainty of passing scores (Tannenbaum & Katz, 2013).

In all three panels, for Level 2 and Level 3, the SEJs decreased or remained within a 0.1 increase over three rounds; however, for Level 4, the SEJ increased in Round 2 and then decreased in Round 3.

Table 13. Raw Score SEJs by Round: Grade Five

Level	Round 1	Round 2	Round 3
Level 2	1.34	1.05	1.02
Level 3	1.50	1.38	1.19
Level 4	1.07	1.59	1.44

Table 14. Raw Score SEJs by Round: Grade Eight

Level	Round 1	Round 2	Round 3
Level 2	1.98	1.14	0.96
Level 3	1.77	1.06	0.74
Level 4	0.73	1.85	1.47

Table 15. Raw Score SEJs by Round: High School

Level	Round 1	Round 2	Round 3
Level 2	2.01	1.15	0.81
Level 3	1.87	1.45	1.27
Level 4	0.88	1.38	1.26

Table 16 through Table 18 provide the panel-recommended threshold score +/- 1 SEJ and +/- 2 SEJs by panel. Ranges around the panel-recommended threshold score are rounded to one decimal.

Table 16. Grade Five Round 3 Recommended Threshold Scores +/- 1 SEJ and +/- 2 SEJs

Threshold	Level 2 Raw Score	Level 3 Raw Score	Level 4 Raw Score
-2 SEJ	17.0	36.6	51.1
-1 SEJ	18.0	37.8	52.6
Panel Recommended	19.0	39.0	54.0
+1 SEJ	20.0	40.2	55.4
+2 SEJ	21.0	41.4	56.9

Table 17. Grade Eight Round 3 Recommended Threshold Scores +/- 1 SEJ and +/- 2 SEJs

Threshold	Level 2 Raw Score	Level 3 Raw Score	Level 4 Raw Score
-2 SEJ	14.1	39.5	53.1
-1 SEJ	15.0	40.3	54.5
Panel Recommended	16.0	41.0	56.0
+1 SEJ	17.0	41.7	57.5
+2 SEJ	17.9	42.5	58.9

Table 18. High School Round 3 Recommended Threshold Scores +/- 1 SEJ and +/- 2 SEJs

Threshold	Level 2 Raw Score	Level 3 Raw Score	Level 4 Raw Score
-2 SEJ	11.4	27.5	41.5
-1 SEJ	12.2	28.7	42.7
Panel Recommended	13	30	44
+1 SEJ	13.8	31.3	45.3
+2 SEJ	14.6	32.5	46.5

Table 19 through Table 21 present the data shown to the panelists after Round 2 judgments were calculated. The tables show the percent of students who would be placed in each achievement level based on Round 2 mean threshold score recommendations on the standard setting scale. There is no threshold score needed for Level 1, so this is not applicable and indicated by "NA." These projected distributions are based on the 2018–19

CAST score distributions and are typically referred to as impact data. This information is presented to the panels as feedback to consider when making Round 3 judgments.

Table 19. Projected Distribution of 2018–19 Students Based on Round 2 Recommendations: Grade Five

Achievement Level	Threshold Score	Percentage
Level 1	NA	18.1
Level 2	177	44.4
Level 3	207	27.1
Level 4	231	10.4

Table 20. Projected Distribution of 2018–19 Students Based on Round 2 Recommendations: Grade Eight

Achievement Level	Threshold Score	Percentage
Level 1	NA	7.8
Level 2	169	55.4
Level 3	209	27.5
Level 4	232	9.3

Table 21. Projected Distribution of 2018–19 Students Based on Round 2 Recommendations: High School

Achievement Level	Threshold Score	Percentage
Level 1	NA	24.6
Level 2	179	50.5
Level 3	215	22.3
Level 4	241	2.6

Table 22 through Table 24 present Round 3 results, based on panel-recommended threshold scores on the standard setting scale, and the projected distribution based on the 2018–19 CAST administration, which were displayed to each panel. The impact of a change in the threshold score can be demonstrated by considering the two tables for grade five: Table 19 and Table 22. There was no change to the grade five panel recommendations for Level 2 and Level 3 and therefore no change to the percentages in Level 1 and Level 2. However, the Level 4 threshold score decreased from 231 to 229, resulting in a decrease in the percent of students in Level 3 (from 27.1 to 25.2) and an increase in the percent in Level 4 (from 10.4 to 12.3). Panelists used this final feedback on Round 3 threshold scores when responding to the last questions in the final evaluation form.

Table 22. Projected Distribution of 2018–19 Students Based on Round 3 Recommendations: Grade Five

Achievement Level	Threshold Score	Percentage
Level 1	NA	18.1
Level 2	177	44.4
Level 3	207	25.2
Level 4	229	12.3

Table 23. Projected Distribution of 2018–19 Students Based on Round 3 Recommendations: Grade Eight

Achievement Level	Threshold Score	Percentage
Level 1	NA	8.8
Level 2	170	54.4
Level 3	209	26.7
Level 4	231	10.1

Table 24. Projected Distribution of 2018–19 Students Based on Round 3 Recommendations: High School

Achievement Level	Threshold Score	Percentage
Level 1	NA	16.5
Level 2	174	55.8
Level 3	213	23.3
Level 4	238	4.4

The data displayed in Table 25 through Table 27 presents the final round threshold score recommendations converted to rounded scale scores and the scale scores at +/- 1 and +/- 2 conditional standard errors of measurement (CSEMs) at each recommended threshold score. Every test has measurement error, and the CSEM presents the error surrounding one particular score—the recommended threshold score. The CSEM is a way to take into consideration the reliability of test scores. More specifically, this statistic is an indication of the degree of uncertainty at each scale score and is sometimes used for guidance when evaluating the appropriateness of threshold scores.

Table 25. Projected Percentage of 2018–19 Students at or Above the Recommended Threshold Score, +/-1 CSEM and +/-2 CSEM for Grade Five

Threshold	Level 2 Scale Score	Level 2 Percent at or Above	Level 3 Scale Score	Level 3 Percent at or Above	Level 4 Scale Score	Level 4 Percent at or Above
-2 CSEM	167	95.3	195	54.4	219	22.7
-1 CSEM	172	89.6	201	45.9	224	17.4
Panel Recommended	177	81.9	207	37.5	229	12.3
+1 CSEM	182	74.7	213	29.5	234	7.6
+2 CSEM	187	67.0	219	22.7	239	4.1

Table 26. Projected Percentage of 2018–19 Students at or Above the Recommended Threshold Score, +/-1 CSEM and +/-2 CSEM for Grade Eight

Threshold	Level 2 Scale Score	Level 2 Percent at or Above	Level 3 Scale Score	Level 3 Percent at or Above	Level 4 Scale Score	Level 4 Percent at or Above
-2 CSEM	160	99.1	197	52.4	221	22.0
-1 CSEM	165	96.3	203	44.9	226	16.0
Panel Recommended	170	91.2	209	36.8	231	10.1
+1 CSEM	175	84.4	215	28.9	236	5.8
+2 CSEM	180	76.8	221	22.0	241	2.3

Table 27. Projected Percentage of 2018–19 Students at or Above the Recommended Threshold Score, +/-1 CSEM and +/-2 CSEM for High School

Threshold	Level 2 Scale Score	Level 2 Percent at or Above	Level 3 Scale Score	Level 3 Percent at or Above	Level 4 Scale Score	Level 4 Percent at or Above
-2 CSEM	162	96.8	199	45.8	230	9.7
-1 CSEM	168	91.8	206	36.7	234	6.8
Panel Recommended	174	83.5	213	27.7	238	4.4
+1 CSEM	180	75.4	220	19.8	242	1.9
+2 CSEM	186	65.8	227	13.3	246	0.6

Incorporating Additional Considerations in Setting Threshold Scores

In standard setting, policymakers sometimes wish to reduce the number of examinees who fall below the panel-recommended threshold scores due to random error. In addition to measurement error metrics (e.g., CSEM, SEJ), policymakers should consider the likelihood of classification error; that is, when adjusting a threshold score, policymakers should consider whether it is more important to minimize a false-positive decision or to minimize a false-negative decision.

A false-positive decision occurs when a test taker's score suggests one level of knowledge and skills, but the student's actual level is lower (i.e., the student does not possess the required skills). A false-negative decision occurs when a test taker's score suggests that the student does not possess the required skills, but that student nevertheless actually does possess those skills.

In order to reduce the number of false negatives, policymakers will decide to lower the threshold score(s). On the other hand, they may desire to reduce the number of test takers who attain a score above the recommended threshold score because of random error at each level in order to reduce the number of false positives and thus raise the threshold score(s).

Raising threshold scores reduces false positives but increases false negatives; the reverse occurs when threshold scores are lowered. Policymakers need to consider which decision error to minimize; it is not possible to eliminate both types of decision errors simultaneously.

Evaluation of the Standard Setting Process

Each panelist was asked at two points over the course of the workshop to rate the following:

1. The panelist's understanding of the process
2. The usefulness of different training exercises
3. The influence of various factors on the judgments

Panelists' ratings were collected using evaluation forms. The purpose of the first evaluation form, completed prior to judgments being made on the operational items, was to provide an early check on the level of panelists' understanding of the task and to identify any areas of confusion. Assessing the level of clarity prior to beginning the judgment process is essential to validating the overall standard setting process. The second and final evaluation form contained additional questions used to analyze the whole process, including the following:

- Training
- Usefulness of materials and procedures
- Influence of policy documents and work products
- Individual and group perceptions
- Student performance data
- Discussion

Results from the evaluation forms are panel-based and are specific to each panel. There was no cross-panel discussion during the process of the standard setting workshop; therefore, any comparisons across panels should acknowledge the independence of the panels.

Evaluation Results from the CAST Standard Setting Final Evaluations

On the grade five and grade eight panels, no panelists indicated on the initial training evaluation that additional training or review was needed. In the high school panel, two panelists had questions on specific aspects of the standard setting process, and after discussion with the panel facilitator, indicated a readiness to proceed. The evaluation forms are in appendix 1: attachment C.

Table 28 through Table 42 provide the results of final evaluations. The results provide information about panelists' thoughts as to the usefulness and influence of materials and other aspects of the three-day process. It also provides insight into each panelist's stated belief as to the appropriateness of the threshold-score recommendations and whether the panelist could support them.

In the final evaluation, the majority of panelists indicated having a clear understanding of the standard setting process and indicated that the materials and processes were somewhat or very useful. Panelists overall indicated that most of the process materials, data, and discussion were somewhat or very influential. Five panelists in the grade five panel and one panelist in the high school panel indicated that "Completing the pre-workshop assignment" was not at all useful (refer to attachment A). In some panels, one panelist indicated that one aspect was not influential (e.g., "the percent of students in each achievement level was not at all influential").

The majority of panelists indicated that the amount of time for different components of the process was about right. However, panelists' responses to the questions about the appropriate amount of time allowed for each step varied somewhat. In all three panels, one or more of the panelists indicated that there was either too little or too much time allotted to some aspect of the process. For example, in the grade five panel, two panelists indicated there was too little time for group discussion, and two panelists indicated there was too much time for group discussion. Experience indicates that variability in panelists' sense of the training and process is expected and dependent on the characteristics and interactions of the panel.

Panelists provided independent judgments on the standard setting forms and were given another opportunity to provide opinions when asked, in the final evaluation, if the recommended threshold scores were too low, about right, or too high, based on the Round 3 panel mean judgments (refer to Table 31, Table 36, and Table 41). Generally, panelists were comfortable with the threshold-score recommendations; a majority of panelists in all three panels indicated that the final recommendations were "about right." Where panelists indicated disagreement with the threshold scores, there were some trends. In the grade five panel, four to five panelists thought the threshold scores were too low. Four panelists in the grade eight panel indicated that the Level 2 and Level 4 threshold scores were too low. In the high school panel, one panelist indicated that the Level 2 threshold score was too low, and all panelists agreed that the Level 3 and Level 4 threshold scores were "about right." The last question on the evaluation asked panelists to confirm support for the final recommendations (refer to Table 32, Table 37, and Table 42.) The majority of panelists in all three panels responded "Yes;" support was confirmed.

Table 28. Final Evaluation Grade Five on the Usefulness of Materials

How *useful* was each of the following materials or procedures in completing the standard setting process?	Not at All Useful *N*	Not at All Useful %	Somewhat Useful *N*	Somewhat Useful %	Very Useful *N*	Very Useful %
Completing the pre-workshop assignment	5	33	6	40	4	27
Taking the test before making judgments	0	0	0	0	15	100
Defining the borderline students	0	0	3	20	12	80
Practicing the procedure	0	0	3	20	12	80
Group discussions	0	0	1	7	14	93
Impact information (percent of students in each achievement level)	1	7	7	47	7	47

Table 29. Final Evaluation Grade Five on the Influence of Process Components

How *influential* was each of the following in making your judgments?	Not at All Influential *N*	Not at All Influential %	Somewhat Influential *N*	Somewhat Influential %	Very Influential *N*	Very Influential %
Achievement level descriptors	1	7	0	0	14	93
Borderline student definitions	0	0	3	20	12	80
My perception of the difficulty of the items and tasks	0	0	5	33	10	67
My experience with the students	0	0	4	27	11	73
Group discussions	0	0	3	20	12	80
Judgments and rationales of other panelists	1	7	6	40	8	53
Percent of students in each achievement level	2	13	10	67	3	20
My sense of what students need to know to be proficient	0	0	5	33	10	67

Table 30. Final Evaluation Grade Five on Timing

How appropriate was the *amount of time* you were given to complete the different components of the process?	**Too Little Time *N***	**Too Little Time %**	**About Right *N***	**About Right %**	**Too Much Time *N***	**Too Much Time %**
Training in the procedure (Angoff)	0	0	14	93	1	7
Training in the procedure (Extended Angoff)	1	7	13	87	1	7
Test familiarization	3	20	12	80	0	0
Group discussion	2	13	11	73	2	13

Table 31. Final Evaluation Grade Five on the Appropriateness of the Final Recommendations

Do you believe that the final recommended threshold score for entering each of the achievement levels is too low, about right, or too high?	**Too Low *N***	**Too Low %**	**About Right *N***	**About Right %**	**Too High *N***	**Too High %**
Level 2	5	33	8	53	2	13
Level 3	5	33	9	60	1	7
Level 4	4	27	9	60	2	13

Table 32. Final Evaluation Grade Five on Panelists’ Support of Recommendations

Question	**Yes *N***	**Yes %**	**No *N***	**No %**
Do you support the final recommendations of the committee?	14	93	1	7

Table 33. Final Evaluation Grade Eight on the Usefulness of Materials

How *useful* was each of the following materials or procedures in completing the standard setting process?	**Not at All Useful *N***	**Not at All Useful %**	**Somewhat Useful *N***	**Somewhat Useful %**	**Very Useful *N***	**Very Useful %**
Completing the pre-workshop assignment	0	0	9	60	6	40
Taking the test before making judgments	0	0	0	0	15	100
Defining the borderline students	0	0	3	20	12	80
Practicing the procedure	0	0	1	7	14	93
Group discussions	0	0	1	7	14	93
Impact information (percent of students in each achievement level)	0	0	5	33	10	67

Table 34. Final Evaluation Grade Eight on the Influence of Process Components

How *influential* was each of the following in making your judgments?	**Not at All Influential *N***	**Not at All Influential %**	**Somewhat Influential *N***	**Somewhat Influential %**	**Very Influential *N***	**Very Influential %**
Achievement level descriptors	0	0	1	7	14	93
Borderline student definitions	0	0	3	20	12	80
My perception of the difficulty of the items and tasks	0	0	2	13	13	87
My experience with the students	0	0	4	27	11	73
Group discussions	0	0	4	27	11	73
Judgments and rationales of other panelists	0	0	4	27	11	73
Percent of students in each achievement level	1	7	8	53	6	40
My sense of what students need to know to be proficient	0	0	4	27	11	73

Table 35. Final Evaluation Grade Eight on Timing

How appropriate was the *amount of time* you were given to complete the different components of the process?	Too Little Time *N*	Too Little Time %	About Right *N*	About Right %	Too Much Time *N*	Too Much Time %
Training in the procedure (Angoff)	0	0	15	100	0	0
Training in the procedure (Extended Angoff)	0	0	15	100	0	0
Test familiarization	5	33	9	60	1	7
Group discussion	0	0	11	73	4	27

Table 36. Final Evaluation Grade Eight on the Appropriateness of the Final Recommendations

Do you believe that the final recommended threshold score for entering each of the achievement levels is too low, about right, or too high?	Too Low *N*	Too Low %	About Right *N*	About Right %	Too High *N*	Too High %
Level 2	4	27	11	73	0	0
Level 3	1	7	14	93	0	0
Level 4	4	27	10	67	1	7

Table 37. Final Evaluation Grade Eight on Panelists’ Support of Recommendations

Question	Yes *N*	Yes %	No *N*	No %
Do you support the final recommendations of the committee?	13	87	2	13

Table 38. Final Evaluation High School on the Usefulness of Materials

How *useful* was each of the following materials or procedures in completing the standard setting process?	Not at All Useful *N*	Not at All Useful %	Somewhat Useful *N*	Somewhat Useful %	Very Useful *N*	Very Useful %
Completing the pre-workshop assignment	1	6	11	69	4	25
Taking the test before making judgments	0	0	1	6	15	94
Defining the borderline students	0	0	1	6	15	94
Practicing the procedure	0	0	3	19	13	81
Group discussions	0	0	1	6	15	94
Impact information (percent of students in each achievement level)	0	0	5	31	11	69

Table 39. Final Evaluation High School on the Influence of the Process Components

How *influential* was each of the following in making your judgments?	Not at All Influential *N*	Not at All Influential %	Somewhat Influential *N*	Somewhat Influential %	Very Influential *N*	Very Influential %
Achievement level descriptors	0	0	7	44	9	56
Borderline student definitions	0	0	2	13	14	88
My perception of the difficulty of the items and tasks	0	0	4	25	12	75
My experience with the students	0	0	2	13	14	88
Group discussions	0	0	1	6	15	94
Judgments and rationales of other panelists	0	0	3	19	13	81
Percent of students in each achievement level	1	6	10	63	5	31
My sense of what students need to know to be proficient	0	0	4	25	12	75

Table 40. Final Evaluation High School on Timing

How appropriate was the *amount of time* you were given to complete the different components of the process?	Too Little Time *N*	Too Little Time %	About Right *N*	About Right %	Too Much Time *N*	Too Much Time %
Training in the procedure (Angoff)	0	0	11	69	5	31
Training in the procedure (Extended Angoff)	0	0	12	75	4	25
Test familiarization	3	19	13	81	0	0
Group discussion	1	6	11	69	4	25

Table 41. Final Evaluation High School on the Appropriateness of the Final Recommendations

Do you believe that the final recommended threshold score for entering each of the achievement levels is too low, about right, or too high?	Too Low *N*	Too Low %	About Right *N*	About Right %	Too High *N*	Too High %
Level 2	1	6	15	94	0	0
Level 3	0	0	16	100	0	0
Level 4	0	0	16	100	0	0

Table 42. Final Evaluation High School on Panelists' Support of Recommendations

Question	Yes *N*	Yes %	No *N*	No %
Do you support the final recommendations of the committee?	16	100	0	0

Chapter 6: Post Standard Setting Results

The 2018–19 administration of the California Science Test (CAST) included operational field test items, which were analyzed using classical and item response theory (IRT) item analyses. These analyses were completed after the standard- setting workshop. Results from the item analysis indicated that two items in the grade five test and two items in the grade eight test did not perform as expected, and therefore will not be included in the reported total score. All items in the high school test functioned as expected.

Educational Testing Service (ETS) recalculated the recommended threshold scores with the two items in grades five and eight excluded. Results tables similar to those found in chapter 5 can be found in appendix 2.

Conclusion

At the request of the California Department of Education (CDE), ETS conducted a standard setting workshop for the CAST, grade five, grade eight, and high school, from July 31–August 2, 2019. The Modified Angoff and Extended Angoff Methods were applied. The process was implemented as planned. Three rounds of judgments with feedback and discussion were completed, and evidence of internal procedural validity was collected via the panelists' evaluations.

The results of the evaluations indicated that the panelists understood the process and the tasks they were asked to complete, found the instructions easy to follow and the training and materials sufficient and clear, and had adequate time to complete the various tasks. In all panels, the majority of panelists judged the final recommended threshold scores to be appropriate (not too high or too low), although there was an indication that some grade five panelists had some disagreement as to the recommended threshold scores.

Immediately following the workshop, preliminary results were provided to the CDE in the form of recommended threshold scores for each achievement level for the total score for all three grades or grade spans. Data files were provided to the CDE on August 5, 2019. The final standard setting report presented here provides details about panelists, materials, and processes that were not included in the preliminary results table.

References

Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard setting topics. *Applied Measurement in Education, 17*(1), 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.

California Department of Education. (2018). *CAST achievement level descriptors*. Retrieved from https://www.cde.ca.gov/be/ag/ag/yr18/documents/nov18item08.docx.

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

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.

Tannenbaum, R. J. & Katz, I. R. (2013). Standard setting. In K. F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology,* (Vol. 3, pp. 455–77). Washington, DC: American Psychological Association.

Appendix 1: Attachments

Attachment A: Panelist Invitation to Participate and Pre-Workshop Assignment, Grade Five Sample

Dear Standard Setting Panelist,

Thank you once again for agreeing to serve as a member of a standard setting panel for the California Science Test (CAST). We have selected you based upon your expertise to make the necessary recommendations, given that you know the California Next Generation Science Standards (CA NGSS), you are familiar with the CAST, and you are working with students who will be taking the CAST.

You will be working with other panelists who represent science educators across the state. You have been assigned to **panel** that will work on **grade 5**. Your grade assignment is also on the top of your notetaking form.

During the standard setting workshop, you will work with your fellow panelists and participate in training for the procedure to develop threshold scores that define four CAST achievement levels. Standard setting facilitators and assessment specialists from Educational Testing Service (ETS) will guide you through the process, and the California Department of Education (CDE) will be present to answer any policy questions you may have. ETS will present the results of the workshop to the CDE. In November 2019, the CDE will present for the California State Board of Education's approval the threshold score recommendations.

An important part of your work will be to define the knowledge and skills at the entry point of the achievement levels. To help you become familiar with the expectations for the CAST, we have attached the general achievement level descriptors (ALDs) as well as an excerpt from the range ALDs. You will notice that the first page of the range ALD document includes a statement that links the general ALD language with the grade-specific range ALD language. Both the general ALDs and the range ALDs will be used in the task described in this email.

To help you prepare for the workshop, please complete the following two-part task. The purpose of the first part is for you to become familiar with the item types and tasks that we will work with during standard setting. The second part will give you some familiarity with another important resource for setting standards.

Part 1: Take your assigned grade CAST training test. The training test consists of several standalone questions and one performance task. The training test provides examples of items that may be found on the operational CAST. The scoring guide includes details about the items, answer key, and scoring rubrics. Take the following steps to access the CAST training test:

1. Go to the Online Practice and Training Tests Portal web page.
2. Select the [**Student Interface Practice and Training Tests**] button.
3. Select the [**Sign In**] button (at the bottom of the web page).
4. Select your grade (5 or 8) or grade level (High School) from the *Grade* drop-down list.
5. On the *Your Tests* web page that appears, move down the page and then select the [**Start CAST *Grade X* Training Test**] button.

6. Select the [**Select**] button without making any changes.
7. Play the video and then, if you can hear the video, select the [**I could play the video and sound**] button.
8. Select [**Continue**].
9. Select [**Begin Test Now**] and take the training test.

To score yourself, select the link for your grade to score your training test.

- **Grade five**

Part 2: Review the excerpt of the range ALDs for your assigned grade. Use the attached notetaking form to help you structure your thoughts as you review the excerpt of the ALDs. Please focus on your assigned grade when using your notetaking form and bring these notes with you to the standard setting workshop. You do not have to bring the ALDs; we will have printed ALDs as well as test materials for your reference at the workshop.

We have found that by completing this preworkshop task, panelists feel more prepared at the workshop. If you have any questions or concerns regarding standard setting, please contact me at PBaron@ets.org. Thank you in advance for your involvement in this very important work, and we look forward to seeing you in Sacramento.

Sincerely,

Patricia Baron, Ed.D.
Standard Setting Director
Educational Testing Service

Attachments

CALIFORNIA SCIENCE TEST (CAST)

NOTETAKING TASK

The CAST achievement level descriptors (ALDs) reflect expected performance for a range of students at each achievement level. Figure 2 represents students ordered according to the students' science proficiency in each grade tested. Three achievement levels are indicated. In each level, the student at the beginning of a level is the **borderline student**. The Level 3 borderline student (in solid pink) has slightly more knowledge than the highest-performing student in Level 2 (in sage green plaid).

In this task, you will focus only on one part of each of the disciplinary core idea (DCI) for each science domain (Physical Sciences, Life Sciences, and Earth and Space Sciences). Your task is to think about one borderline student for each science domain: the Level 3 borderline student. The task on the following pages will allow you to become familiar with the ALDs and with the type of comparisons we will be making at the standard setting workshop.

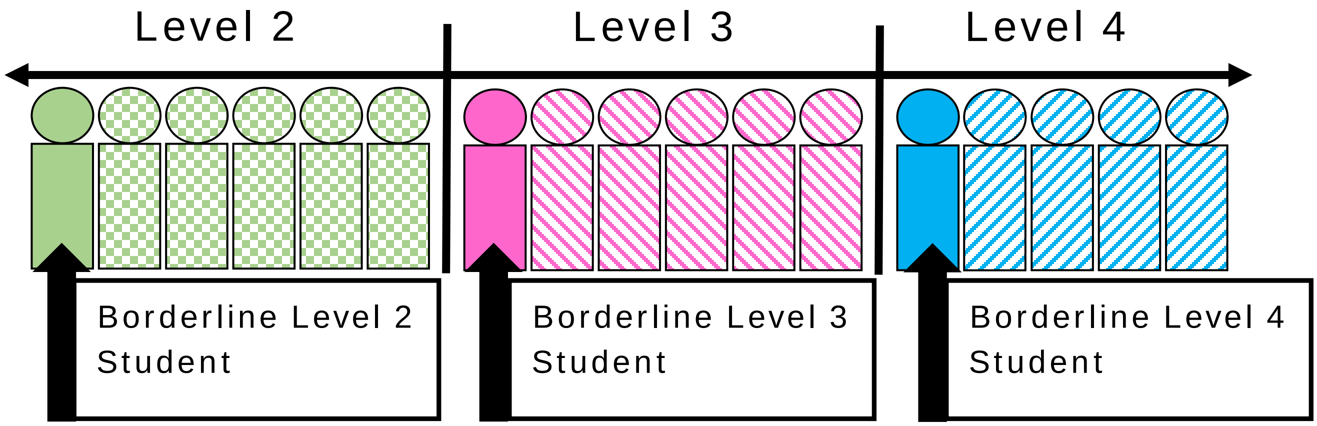
Level 2
Level 3
Level 4
Borderline Level 2 Student
Borderline Level 3 Student
Borderline Level 4 Student

Figure 2. Borderline Student Definitions

CALIFORNIA SCIENCE TEST (CAST) GRADE FIVE

HOMEWORK

California Science Test (CAST) Range Achievement Level Descriptors (Excerpt)

Table 43. CAST Range Achievement Level Descriptors

Nearly Met Standard	Met Standard	Exceeded Standard
Students at Level 2 **consistently** apply their knowledge and skills of the CA NGSS to problems of **low complexity**, demonstrating a **partial understanding** of the Physical Sciences; Life Sciences; Earth and Space Sciences; and Engineering, Technology, and Application of Sciences.	Students at Level 3 **consistently** apply their knowledge and skills of the CA NGSS to problems of **medium complexity**, demonstrating an **adequate understanding** of the Physical Sciences; Life Sciences; Earth and Space Sciences; and Engineering, Technology, and Application of Sciences.	Students at Level 4 **consistently** apply their knowledge and skills of the CA NGSS to problems of **high complexity**, demonstrating a **thorough understanding** of the Physical Sciences; Life Sciences; Earth and Space Sciences; and Engineering, Technology, and Application of Sciences.

Table 44. CAST Grade Five Three–Dimensional Physical Sciences Achievement Level Descriptors

Physical Sciences: DCI Strands	**Nearly Met Standard Level 2**	**Met Standard Level 3**	**Exceeded Standard Level 4**
Matter and Its Interactions (PS1)	Students can • **use a model to identify** that matter is made of particles too small to be seen, • **identify or observe** properties of materials, • **use measurements** of matter such as weight and temperature to make observations that matter is conserved during physical changes, and • **identify** whether the mixing of substances produces a new substance.	Students can • **develop a model to describe** that matter is made of particles too small to be seen, • **make observations and measurements** to identify materials by their properties, • **measure and graph** quantities to provide evidence that matter is conserved during physical changes, and • **investigate** whether the mixing of substances produces a new substance.	Students can • **develop a model to explain** that particles too small to be seen can account for one or more phenomena, • **plan an investigation** using an independent variable to identify materials based upon their properties, • **evaluate evidence to substantiate a claim** that matter is conserved during physical or chemical changes, and • **use evidence to plan a new investigation** to determine whether the mixing of substances produces a new substance.

Table 45. CAST Grade Five Three–Dimensional Life Sciences Achievement Level Descriptors

Life Sciences: DCI Strands	**Nearly Met Standard** **Level 2**	**Met Standard** **Level 3**	**Exceeded Standard** **Level 4**
From Molecules to Organisms: Structures and Processes (LS1)	Students can • **use a model to describe** that organisms have unique life cycles but all have in common birth, growth, reproduction, and death; • **identify evidence** that plants and animals have internal and external structures that function to support survival, growth, behavior, and reproduction; • **identify components in a model that describes** how animals receive different types of information through their senses, process the information in their brain, and respond to the information in different ways; and • **identify** that plants get the materials they need for growth chiefly from air and water.	Students can • **develop models to describe** that organisms have unique life cycles but all have in common birth, growth, reproduction, and death; • **construct an argument** that plants and animals have internal and external structures that function to support survival, growth, behavior, and reproduction; • **use a model to describe** that animals receive different types of information through their senses, process the information in their brain, and respond to the information in different ways; and • **support an argument using evidence** that plants get the materials they need for growth chiefly from air and water.	Students can • **develop and use models to explain** that organisms have unique life cycles but all have in common birth, growth, reproduction, and death; • **construct an argument** that plants and animals have internal and external structures that function as systems to support survival, growth, behavior, and reproduction; • **develop a model to explain** that animals receive different types of information through their senses, process the information in their brain, and respond to the information in different ways; and • **construct an argument using reasoning and data** to show that plants get the materials they need for growth chiefly from air and water.

Table 46. CAST Grade Five Three–Dimensional Earth and Space Sciences Achievement Level Descriptors

Earth and Space Sciences: DCI Strands	**Nearly Met Standard** **Level 2**	**Met Standard** **Level 3**	**Exceeded Standard** **Level 4**
Earth's Place in the Universe (ESS1)	Students can • **identify simple patterns** in rock formations or fossils, • **use data to identify** the relative distances of stars, and • **use data to identify** daily changes in length and direction of shadows, day and night, and the seasonal appearance of some stars in the night sky.	Students can • **identify evidence from patterns** in rock formations and fossils to support an explanation for changes in a landscape over time, • **support an argument** that differences in the apparent brightness of the sun and stars are due to their relative distances from Earth, and • **graph data to show patterns** of daily changes in length and direction of shadows, day and night, and the seasonal appearance of some stars in the night sky.	Students can • **use reasoning to explain patterns** in rock formations and fossils in a landscape over time • **use a model to support an argument** that differences in the apparent brightness of the sun and stars are due to their relative distances from Earth, and • **use graphical data to explain patterns** in daily changes in length and direction of shadows, day and night, and the seasonal appearance of some stars in the night sky.

Attachment B: Final Borderline Student Definitions

CAST Borderline Student Definitions Grade Five

Borderline Level 2 Student Physical Sciences

The grade five borderline Level 2 student can. . .

1. Identify or observe properties of materials.
2. Take measurements of an object’s motion.
3. Describe the speed or energy of an object.
4. Use a model to identify patterns in wave properties.

Borderline Level 2 Student: Life Sciences

The grade five borderline Level 2 student can. . .

1. Identify evidence that plants or animals have structures that function to support survival, growth, behavior, or reproduction.
2. Identify examples to show that some animals form groups that help members survive.
3. Identify that plants or animals have traits inherited from parents.
4. Identify examples that in a particular habitat some organisms can survive well, some survive less well, and some cannot survive at all.

Borderline Level 2 Student: Earth and Space Sciences

The grade five borderline Level 2 student can. . .

1. Identify simple patterns in rock formations or fossils.
2. Identify that climates differ in different regions of the world.
3. Identify one way that individual communities might use science ideas to protect Earth’s resources and environment.

Borderline Level 2 Student: Engineering, Technology, and Applications of Science

The grade five borderline Level 2 student can. . .

1. Identify a solution to a problem.

Borderline Level 3 Student Physical Sciences

The grade five borderline Level 3 student can. . .

1. A. Develop a model to describe that matter is made of particles too small to be seen.

   B. Make observations to identify materials by their properties.
2. A. Make observations or measurements to identify patterns in an object’s motion.

   B. Investigate the effects of balanced and unbalanced forces on the motion of an object.
3. Make observations to provide evidence that energy can be transferred.
4. Develop a model to describe patterns in the properties of one type of wave.

Borderline Level 3 Student Life Sciences

The grade five borderline Level 3 student can. . .

1. Construct an argument that plants or animals have structures that function to support survival, growth, behavior, or reproduction.
2. Develop a model to describe the movement of matter among plants, animals, decomposers, and the environment.
3. Use evidence to explain that traits can be influenced by the environment.
4. Use evidence to explain how variation in individuals may affect survival in a particular habitat.

Borderline Level 3 Student Earth and Space Sciences

The grade five borderline Level 3 student can. . .

1. Graph data to show patterns of daily changes in length and direction of shadows, or day and night, or the seasonal appearance of some stars in the night sky.
2. A. Combine information to describe climates in different regions of the world.

   B. Use observations and measurements to identify the effects of weathering or the rate of erosion.
3. Compare multiple solutions to reduce the impacts of natural Earth processes on humans.

Borderline Level 3 Student Engineering, Technology, and Applications of Science

The grade five borderline Level 3 student can. . .

1. Describe a simple design problem reflecting a need or a want that includes specified criteria for success and constraints on materials, time, or cost.

Borderline Level 4 Student Physical Sciences

The grade five borderline Level 4 student can. . .

1. Develop a model to explain that particles too small to be seen can account for one or more phenomena.
2. Use evidence to ask new questions about the effects of balanced and unbalanced forces on the motion of an object.
3. A. Make a prediction using evidence about the relationship between the speed of an object and the energy of that object.

   B. Develop a model to explain that energy in animals’ food was once energy from the sun.
4. Use a model to explain a phenomenon about light reflecting from objects and entering the eye allowing object to be seen.

Borderline Level 4 Student Life Sciences

The grade five borderline Level 4 student can. . .

1. Develop and use models to explain that organisms have unique life cycles but all have in common birth, growth, reproduction, and death.
2. Revise a model that reflects changes in the cycling of matter and the systems or interactions within the ecosystem among plants, animals, decomposers, and the environment.
3. Use evidence to predict how traits might be influenced by changes in the environment.
4. Use evidence to predict which organisms will survive well in a particular habitat.

Borderline Level 4 Student Earth and Space Sciences

The grade five borderline Level 4 student can. . .

1. Use graphical data to explain patterns in daily changes in length and direction of shadows or day and night, or the seasonal appearance of some stars in the night sky.
2. Develop a model to describe multiple ways in which the geosphere, biosphere, hydrosphere, and atmosphere interact.
3. Evaluate the ways individual communities use science ideas to protect Earth’s resources and environment.

Borderline Level 4 Student Engineering, Technology, and Applications of Science

The grade five borderline Level 4 student can. . .

1. Define a complex design problem reflecting a need or want that includes specified criteria for success and constraints on materials, time, or cost.

CAST Borderline Student Definitions Grade Eight

Borderline Level 2 Student Physical Sciences

The grade eight borderline Level 2 student can. . .

1. Use a model to identify the atomic composition of simple molecules.
2. Identify that the change in an object’s motion depends on the sum of the forces on the object.
3. Identify evidence that energy is transferred to and from an object.
4. Use a model to describe that waves are reflected, absorbed, or transmitted through various materials.

Borderline Level 2 Student Life Sciences

The grade eight borderline Level 2 student can. . .

1. Make observations that living things are made of either one cell or different numbers and types of cells.
2. Identify the effects of resource availability on organisms and populations of organisms in an ecosystem.

3. Use a model to describe that mutations may affect the structure and function of the organism.
4. Describe how genetic variations of traits in a population increase some individuals' probability of surviving and reproducing in a specific environment.

Borderline Level 2 Student Earth and Space Sciences

The grade eight borderline Level 2 student can. . .

1. Use a model to identify the motions of objects within galaxies and the solar system.
2. Use a model to describe the cycling of Earth's materials and the flow of energy that drives this process.
3. Describe an example of how humans impact the environment.

Borderline Level 2 Student Engineering, Technology, and Applications of Science

The grade eight borderline Level 2 student can. . .

1. Describe a problem that needs to be solved using the design process.

Borderline Level 3 Student Physical Sciences

The grade eight borderline Level 3 student can. . .

1. Develop a simple model to describe how the total number of atoms does not change in a chemical reaction of simple molecules and interpret data to know when a chemical reaction has occurred.
2. Explain how the change in an object's motion depends on the sum of the forces on the object and the mass of the object.
3. Construct a graph, model, or argument to support the claim that energy can be transferred to or from an object.
4. A. Develop and use a simple model to describe the behavior of waves.

   B. Describe the relationship between the amplitude and energy of a wave.

Borderline Level 3 Student Life Sciences

The grade eight borderline Level 3 student can. . .

1. Construct an argument supported by one piece of simple evidence for how the body is a system of interacting subsystems composed of groups of cells.
2. Develop a model showing the cycling of matter and flow of energy among living and nonliving parts of an ecosystem.
3. Develop and use a model to describe that mutations may result in harmful, beneficial, or neutral effects to the structure and function of an organism.
4. Explain (using one piece of evidence) how genetic variations of traits in a population increase some individuals' probability of survival and reproducing in a specific environment.

Borderline Level 3 Student Earth and Space Sciences

The grade eight borderline Level 3 student can. . .

1. Develop and use a model of the Earth-Sun-Moon system to describe the cyclic patterns of lunar phases, eclipses of the sun and moon, and seasons.
2. Develop a model to describe the cycling of Earth's materials and the flow of energy that drives this process (rock, water, carbon, nitrogen cycle) and analyze how geoscience processes have changed Earth's surface.
3. Construct an argument supported by one piece of evidence for how the increase in human population and consumption of natural resources impact Earth's system and describe a basic solution to the problem.

Borderline Level 3 Student Engineering, Technology, and Applications of Science

The grade eight borderline Level 3 student can. . .

1. Compare competing design solutions based on how well they meet the criteria and constraints of the problem.

Borderline Level 4 Student Physical Sciences

The grade eight borderline Level 4 student can. . .

1. Explain the law of conservation of mass and analyze and interpret data to identify patterns in properties of substances to determine if a chemical reaction has occurred.
2. Plan an investigation about how the change in an object's motion depends on the sum of the forces on the object and the mass of the object.
3. Construct an argument to support the claim that energy takes different forms when energy is transferred to or from an object when the kinetic energy of the object changes.
4. Use mathematical representations to describe patterns in a simple wave model that include that the energy of the wave is proportional to the square of the amplitude.

Borderline Level 4 Student Life Sciences

The grade eight borderline Level 4 student can. . .

1. Construct an argument to support an explanation of the cause-and-effect relationship among multiple species that have structures and functions that are specialized, which increases the probability of success.
2. Analyze and interpret data to identify cause-and-effect relationships between resource availability and the organisms and populations of organisms in an ecosystem.
3. Develop a model to explain why sexual reproduction results in offspring with genetic variation.
4. Use proportional reasoning to construct an explanation based on evidence that describes the cause-and-effect relationship between the genetic variation of traits (natural or artificial selection) in a population and the probability of an individual surviving and reproducing in a specific environment.

Borderline Level 4 Student Earth and Space Sciences

The grade eight borderline Level 4 student can. . .

1. Develop a model of the Earth-Sun-Moon system to predict the occurrence of lunar phases, eclipses of the sun and moon, and seasons.
2. Construct an explanation based on evidence for how geoscience processes have changed Earth's surface at varying time and spatial scales.
3. Design and evaluate a method based on how well it meets the criteria and constraints for monitoring and minimizing a human impact on the environment.

Borderline Level 4 Student Engineering, Technology, and Applications of Science

The grade eight borderline Level 4 student can. . .

1. Use a systematic process to support a claim about the relative effectiveness of competing design solutions based on the strengths and weaknesses of each and how well they meet the criteria and constraints of the problem.

CAST Borderline Student Definitions High School

Borderline Level 2 Student Physical Sciences

The high school borderline Level 2 student can. . .

1. Use the periodic table to identify the relative properties of elements based on the patterns of electrons in the outermost energy level of atoms.
2. Describe forces acting on a system.
3. Use a model to identify energy at the macroscopic scale.
4. Use mathematical representation to identify the properties of waves.

Borderline Level 2 Student Life Sciences

The high school borderline Level 2 student can. . .

1. Identify the components of a model illustrating how macromolecules are used to sustain essential life processes.
2. Identify how energy and matter cycle in a living system.
3. Use concepts of statistics and probability to identify genetic variation and inheritance.
4. Identify how natural selection and changes lead to evolution.

Borderline Level 2 Student Earth and Space Sciences

The high school borderline Level 2 student can. . .

1. Use a model to identify evidence of how the universe was formed over time and the Earth's place in it.
2. Identify components of a model of how the Earth's systems change over time.
3. Identify evidence and describe a solution that reduces impacts of human activities on natural systems.

Borderline Level 2 Student Engineering, Technology, and Applications of Science

The high school borderline Level 2 student can. . .

1. Identify global challenges and real-world problems and their solutions using the engineering process.

Borderline Level 3 Student Physical Sciences

The high school borderline Level 3 student can. . .

1. Use the periodic table to predict and explain the outcome of simple, or common, chemical reactions and processes based on the valence electrons of each element and knowledge of the patterns of chemical properties.
2. Analyze mathematical representations that describe forces acting on a system.
3. Develop and use a model to describe changes in energy in a system.
4. Use mathematical representation to support a claim relating the properties of waves.

Borderline Level 3 Student Life Sciences

The high school borderline Level 3 student can. . .

1. Develop and use a model to construct an explanation for how macromolecules are used to sustain essential life processes.
2. Use a mathematical model and evidence to support explanations for how energy and matter cycle in living systems.
3. Apply concepts of statistics and probability to explain genetic variation and inheritance.
4. Construct an explanation based on evidence for how natural selection and changes lead to evolution.

Borderline Level 3 Student Earth and Space Sciences

The high school borderline Level 3 student can. . .

1. Develop a model to explain, based on evidence, how the universe was formed over time and the Earth’s place in it.
2. Develop a model, based on evidence, of how the Earth’s systems change over time.
3. Evaluate a solution that reduces impacts of human activities on natural systems through data analysis.

Borderline Level 3 Student Engineering, Technology, and Applications of Science

The high school borderline Level 3 student can. . .

1. Analyze global challenges and real-world problems and their solutions using the engineering process.

Borderline Level 4 Student Physical Sciences

The high school borderline Level 4 student can. . .

1. Explain and evaluate the chemical systems through equilibrium and mathematical representations.
2. Analyze data to predict changes in the motion of objects using mathematical relationships
3. Develop and use a model to explain changes in energy.
4. Use mathematical representation to predict and explain wave properties.

Borderline Level 4 Student Life Sciences

The High School Borderline Level 4 student can. . .

1. Utilize multiple sources of evidence to construct an explanation for how macromolecules are used to sustain essential life processes.
2. Use a mathematical model and evidence to construct or revise a complex explanation for how energy and matter cycle in living systems.
3. Apply concepts of statistics and probability to evaluate genetic variation and inheritance.
4. Construct an explanation based on evidence to evaluate how natural selection and changes lead to evolution.

Borderline Level 4 Student Earth and Space Sciences

The high school borderline Level 4 student can. . .

1. Construct an explanation by evaluating multiple sources of evidence of how the universe was formed over time and the Earth’s place in it.
2. Develop and explain a model of how the Earth’s systems change over time.
3. Compare and refine solutions that reduce impacts of human activities on natural systems through data analysis.

Borderline Level 4 Student Engineering, Technology, and Applications of Science

The high school borderline Level 4 student can. . .

1. Design or refine solutions to global challenges and real-world problems using the engineering process.

Attachment C: Evaluation Forms

Initial Evaluation of Modified and Extended Angoff Method

California Science Test

Training Evaluation Form

The purpose of this evaluation form is to obtain your feedback about the training you have received so far on the standard setting process. Your feedback will provide a basis for determining what to review before we begin the actual standard setting process.

Please indicate the degree to which you agree with each statement using the scale given (Strongly Agree, Agree, Disagree, or Strongly Disagree). Please choose only one response for each statement.

Statement	Strongly Agree	Agree	Disagree	Strongly Disagree
I understand the purpose of this workshop.	[Response]	[Response]	[Response]	[Response]
The large-group facilitator explained things clearly.	[Response]	[Response]	[Response]	[Response]
The panel facilitator explained things clearly.	[Response]	[Response]	[Response]	[Response]
I understand the purpose of the Achievement Level Descriptors (ALDs) in this process.	[Response]	[Response]	[Response]	[Response]
I understand what is meant by the borderline student.	[Response]	[Response]	[Response]	[Response]
I understand what the ordered item booklet is.	[Response]	[Response]	[Response]	[Response]
The training in the standard setting methods seems adequate to give me the information I need to complete my assignment.	[Response]	[Response]	[Response]	[Response]
The training in the Bookmark method seems adequate to give me the information I need to complete my assignment.	[Response]	[Response]	[Response]	[Response]
I understand the steps I am to follow to make my standard setting judgments.	[Response]	[Response]	[Response]	[Response]
I understand how to enter my judgments using the survey software.	[Response]	[Response]	[Response]	[Response]
I am ready to complete my standard setting judgments.	[Response]	[Response]	[Response]	[Response]

If you checked "Disagree" or "Strongly Disagree" for any of the statements on the previous page, please indicate what additional information or explanations you need.

Have you participated in a standard setting workshop before today?

Final Evaluation

California Science Test

Standard Setting Final Evaluation

The purpose of the final evaluation form is to obtain your feedback about the standard setting process overall. Your feedback will provide a basis for evaluating the training, methods, and materials in the standard setting process. Your responses will be anonymous; no individuals will be identified.

Gender
Female Male Non-binary

Race
American Indian/Alaskan Native Asian Black or African American

Filipino Hispanic or Latino Pacific Islander

White not Hispanic Other

Grade(s) you currently teach. (Check all that apply.)
3 4 5 6 7 8 9–12

Other

What subjects do you currently teach? (Check all that apply.)
All Subjects Math Science Social Studies English Other

How many years of experience do you have teaching science?
1–3 4–6 7–10 More than 10

How many years of experience do you have working with the CA NGSS?
Not Applicable 1–2 3–4 More than 4

Does your experience include students from these populations? (Check all that apply.)
General education English learners Special education

1. How useful was each of the following materials or procedures in completing the standard setting process?

Statement	Not at All Useful	Somewhat Useful	Very Useful
Completing the prework assignment	[Response]	[Response]	[Response]
Taking the test before making judgments	[Response]	[Response]	[Response]
Defining the borderline students	[Response]	[Response]	[Response]
Practicing the procedure	[Response]	[Response]	[Response]
Group discussions	[Response]	[Response]	[Response]
Impact information (percent of students in each achievement level)	[Response]	[Response]	[Response]

2. How influential was each of the following in making your judgments?

Statement	Not at All Influential	Somewhat Influential	Very Influential
Achievement level descriptors	[Response]	[Response]	[Response]
Borderline student definitions	[Response]	[Response]	[Response]
My perception of the difficulty of the items and tasks	[Response]	[Response]	[Response]
My experiences with the students	[Response]	[Response]	[Response]
Group discussions	[Response]	[Response]	[Response]
Judgments and rationales of other panelists	[Response]	[Response]	[Response]
Percent of students in each achievement level	[Response]	[Response]	[Response]
My sense of what students need to know to be proficient	[Response]	[Response]	[Response]

3. How appropriate was the *amount of time* you were given to complete the different components of the process?

Statement	Too Little Time	About Right	Too Much Time
Training in the procedure (Angoff)	[Response]	[Response]	[Response]
Training in the procedure (Extended Angoff)	[Response]	[Response]	[Response]
Test familiarization	[Response]	[Response]	[Response]
Group discussion	[Response]	[Response]	[Response]

4. CAST threshold scores

Do you believe that the final recommended threshold score for entering each of the achievement levels is too low, about right, or too high?

Achievement Level	Too Low	About Right	Too High
Level 2	[Response]	[Response]	[Response]
Level 3	[Response]	[Response]	[Response]
Level 4	[Response]	[Response]	[Response]

Do you support the final recommendations of the panel?

Yes No

Attachment D: Nondisclosure Agreement Form

California Department of Education

Confidentiality Agreement

Test security for California Assessment of Student Performance and Progress (CAASPP) is of the utmost importance, and it is the California Department of Education's obligation to ensure the security of all test materials. The nature and content of any test, test item, proposed or draft test item, or other secure assessment material, including but not limited to the specific language or the subject of test items or proposed or draft test items and any art such as drawings, graphs, tables and sketches, must not be divulged.

By signing below, you acknowledge and agree that the CAASPP test materials are highly secure and that the unauthorized disclosure of any test materials associated with CAASPP could result in substantial monetary and nonmonetary costs to the State to replace the test and materials. You agree that your access to CAASPP test items, proposed or draft test items, or any other test materials is only for the purpose of review as charged by your role as a member of this panel. You agree not to reproduce the tests or any questions within them, directly or indirectly, and not to reveal the nature or content of the test or test items to any other person other than those participating in this meeting.

I understand that the CAASPP California Science Test (CAST) operational items and scoring guides for the Standard Setting Meeting are classified as confidential. I understand that these materials cannot be discussed outside of the meeting, posted publicly, sold, or reproduced. The materials included in this workshop contain information copyrighted by the Regents of the University of California, the California Department of Education, and/or independent publishers.

____
Signature

____
Print Name

____
Affiliation/Organization

____
Date

Appendix 2: Grades Five and Eight Results with Two Items Dropped from Each Grade

Table 2.1. Mean Raw Score Threshold Scores at the End of Each Round: Grade Five

Level	Round 1	Round 2	Round 3
Level 2	24	19	19
Level 3	48	38	38
Level 4	64	54	53

Table 2.2. Mean Raw Score Threshold Scores at the End of Each Round: Grade Eight

Level	Round 1	Round 2	Round 3
Level 2	21	15	16
Level 3	54	40	39
Level 4	72	56	55

Table 2.3. Raw Score Standard Errors of Judgement (SEJs) by Round: Grade Five

Level	Round 1	Round 2	Round 3
Level 2	1.30	1.01	0.99
Level 3	1.46	1.35	1.15
Level 4	1.03	1.56	1.40

Table 2.4. Raw Score SEJs by Round: Grade Eight

Level	Round 1	Round 2	Round 3
Level 2	1.92	1.10	0.93
Level 3	1.72	1.02	0.72
Level 4	0.71	1.78	1.43

Table 2.5. Projected Distribution of 2018–19 Students Based on Round 2 Recommendations: Grade Five

Achievement Level	Threshold Total Scale Score	Percentage
Level 1	-	18.10
Level 2	177	41.60
Level 3	206	28.90
Level 4	230	11.40

Table 2.6. Projected Distribution of 2018–19 Students Based on Round 2 Recommendations: Grade Eight

Achievement Level	Threshold Total Scale Score	Percentage
Level 1	-	7.80
Level 2	169	55.40
Level 3	209	27.50
Level 4	232	9.30

Table 2.7. Projected Distribution of 2018–19 Students Based on Round 3 Recommendations: Grade Five

Achievement Level	Threshold Total Scale Score	Percentage
Level 1	-	18.10
Level 2	177	41.60
Level 3	206	28.00
Level 4	229	12.30

Table 2.8. Projected Distribution of 2018–19 Students Based on Round 3 Recommendations: Grade Eight

Achievement Level	Threshold Total Scale Score	Percentage
Level 1	-	10.20
Level 2	171	50.20
Level 3	207	29.50
Level 4	231	10.10

Table 2.9. Projected Percentage of 2018–19 Students at or Above the Recommended Threshold Score, +/-1 CSEM and +/-2 CSEM for Grade Five

Threshold	Level 2 Scale Score	Level 2 Percent at or Above	Level 3 Scale Score	Level 3 Percent at or Above	Level 4 Scale Score	Level 4 Percent at or Above
-2 CSEM	167	95.3	194	56.8	219	22.7
-1 CSEM	172	89.6	200	48.7	224	17.4
Panel Recommended	177	81.9	206	40.3	229	12.3
+1 CSEM	182	74.7	212	32.1	234	7.6
+2 CSEM	187	67.0	218	24.4	239	4.1

Table 2.10. Projected Percentage of 2018–19 Students at or Above Recommended Threshold Score, +/-1 CSEM and +/-2 CSEM for Grade Eight

Threshold	Level 2 Scale Score	Level 2 Percent at or Above	Level 3 Scale Score	Level 3 Percent at or Above	Level 4 Scale Score	Level 4 Percent at or Above
-2 CSEM	161	98.7	195	55.0	221	22.0
-1 CSEM	166	95.8	201	47.2	226	16.0
Panel Recommended	171	89.8	207	39.6	231	10.1
+1 CSEM	176	82.8	213	31.6	236	5.8
+2 CSEM	181	75.5	219	24.4	241	2.3