2020 No. 040

California Assessment of Student Performance and Progress (CAASPP)

California Science Test (CAST) Alignment Study Report

Prepared for: California Department of Education
Assessment Development and Administration Division
1430 N Street, Suite 4401
Sacramento, CA 95814-5901

Prepared under: CN180100

Authors: Emily Dickinson
Arthur Thacker
Michele Hardoin

Editors: Lauress Wise
Christa Watters

Date: April 6, 2020

Headquarters: 66 Canal Center Plaza, Suite 700, Alexandria, VA 22314 | Phone: 703.549.3611 | https://www.humrro.org

California Assessment of Student Performance and Progress (CAASPP)

California Science Test (CAST) Alignment Study Report

Table of Contents

List of Appendices

List of Tables

This page is intentionally blank.

Executive Summary

Pursuant to California *Education Code* (*EC*) Section 60649, the Human Resources Research Organization (HumRRO) is continuing its independent evaluation of the California Assessment of Student Performance and Progress (CAASPP) System. The scope of the current evaluation is to conduct three research studies from July 2018 through December 2020 and provide objective technical advice and consultation on activities related to the implementation of specific components of the CAASPP. This report summarizes a study of the alignment between the California Science Test (CAST) and the California Next Generation Science Standards (CA NGSS). Alignment studies are required as part of the federal assessment peer review process, provide validity evidence that the assessment is measuring the intended content, and inform future assessment item development.

The 2018–20 CAASPP Evaluation Plan is presented in HumRRO's *2018 CAASPP Independent Evaluation Report*, which is publicly available online (https://www.cde.ca.gov/ta/tg/ca/documents/caaspp18evalrpt.pdf). The report consists of the CAASPP System's theory of action (CDE, 2018a) and detailed plans for each evaluation study. The plan also includes a timeline for major study milestones; the timeline is based on California Department of Education (CDE) priorities and the anticipated dates of operational administration of assessments.

The CAST became operational in 2018–19. This is a stand-alone report on the completed 2019 CAST Alignment Study. A preliminary report on the progress of the study was presented in HumRRO's *CAASPP 2019 Independent Evaluation Report* (https://www.cde.ca.gov/ta/tg/ca/documents/caaspp19evalrpt.pdf).

Overview

The CAST is designed to measure performance on CA NGSS. Within the CA NGSS, performance expectations (PEs) are assessable statements of what students should know and be able to do. The following three major components, also referred to as dimensions, are combined to operationalize the PEs:

1. Disciplinary Core Ideas (DCIs) are the key ideas in science that have broad importance within or across multiple science or engineering disciplines. These core ideas build on each other as students progress through grade levels. The DCIs are grouped into the following domains: Physical Sciences; Life Sciences; Earth and Space Sciences; and Engineering, Technology, and the Application of Science (hereafter, Engineering).

2. Crosscutting Concepts (CCCs) help students explore connections across the four domains of science mentioned above in item 1. When these concepts, such as "cause and effect," are made explicit for students, they can help students develop a coherent and scientifically based view of the world around them.

3. Science and Engineering Practices (SEPs) describe what scientists do to investigate the natural world and what engineers do to design and build systems. The practices better explain and extend what is meant by "inquiry" in science and the range of cognitive, social, and physical practices that it requires. Students engage in practices to build, deepen, and apply their knowledge of core ideas and crosscutting concepts.

Evaluating alignment for the CAST represents a significant challenge because of the nature of the content, the organization of the content standards, and the test design. The three major components of the CA NGSS (DCIs, CCCs, and SEPs) are integrated into the three assessed science disciplines (earth and space sciences, life sciences, and physical sciences). The test is designed such that students' knowledge is expected to be integrated and to accumulate to create a deep understanding of science content. Developing tests and test items that adequately sample such complex and integrated content is especially challenging. When an item measures a single standard or concept, the alignment process is relatively straightforward. However, test development and alignment become more complex when standards are designed as interactions among statements about content.

The CAST is a computer-based, fixed-form (non-adaptive) assessment administered to students in grades five, eight, and once in high school (i.e., grades 10, 11, or 12). The CAST was field-tested in spring 2018 and administered operationally for the first time in January–July of 2019. The 2019 assessment included three segments, two of which contributed to an individual student's score. The third segment was used for field testing purposes only. This alignment study focused on "student-level alignment," analyzing items from the two operational segments used to compute student-level scores in order to collect evidence that individual students' scores should be sufficiently valid and reliable to support their intended interpretations. Minor changes were made to the CAST test design and blueprint in 2020 (adding one performance task and a small reduction in the number of selected response items), but those changes do not impact the conclusions drawn in this report.

The first step in evaluating for CAST alignment was to investigate the nature of the assessment itself: how the standards guided the development of the test items (and how the standards and items should therefore relate to one another) and the interpretations to be made from CAST scores. This component of the study is described in *Chapter 2: Review of CAST Documentation*. HumRRO then modified traditional alignment methods to account for the test structure and design, a process in keeping with best practices in test validation that facilitates using alignment study results in an overall validity argument. This component of the study is described in *Chapter 3: CAST Alignment Workshop and Outcomes*.

Research Questions

Evidence of the alignment between assessments and standards is a requirement under the U.S. Department of Education's assessment peer review process. Alignment evidence supports that students' test scores can be used to make valid inferences

about student performance on the content being tested. The CDE identified several research questions to guide the alignment evidence collected. Activities conducted for the CAST Alignment Study were designed to provide information to answer the following research questions:

1. To what extent do the test design and test blueprint for the CAST support the claims to be made about student performance on the assessment?

2. To what extent does the test blueprint for the CAST represent an appropriate sampling of the content as set forth in the CA NGSS?

3. To what extent do the test forms and test items for the CAST reflect the test design and test blueprint?

4. To what extent do CAST tasks and items integrate disciplinary core ideas, crosscutting concepts, and/or science and engineering practices?

5. To what extent do test forms show balance across the science domains used for CAST scoring and reporting purposes (earth and space sciences, life sciences, and physical sciences)?

6. Do the CAST items range from low to high cognitive complexity (i.e., depth of knowledge or DOK) and provide a sufficient number of items across the range of cognitive complexity?

7. How well does CAST fit the population being tested, in terms of the distribution of item difficulties within test forms and the distribution of student ability?

Review of CAST Documentation

HumRRO researchers collected and reviewed CAST design and test development materials provided by California Department of Education (CDE) and Educational Testing Service (ETS) staff, as well as information about the CAST shared with the public on the CDE website. HumRRO researchers evaluated the alignment of the CAST test design and development documentation to the *Standards for Educational and Psychological Testing* (AERA, APA & NCME, 2014; hereafter referred to as the *Testing Standards*).

First, HumRRO researchers identified specific standards from the *Testing Standards* that are directly relevant to how alignment is considered during test development. Next, researchers identified and collected the types of documentation needed to provide evidence that these standards were met. Finally, two HumRRO researchers independently reviewed the documentation and rated the extent to which each standard was met. These independent ratings were compared and discussed to reach a final consensus rating for each standard.

HumRRO developed and applied the following five-point rating scale to evaluate the degree to which the evidence for the assessment supports alignment to each standard:

1. No evidence of the Standard found in the Materials.
2. Little evidence of the Standard found in the materials; less than half of the Standard was covered in the materials and/or evidence of key aspects of the Standard could not be found.
3. Some evidence of the Standard found in the materials; approximately half of the Standard was covered in the materials, including some key aspects of the Standard.
4. Evidence in the materials mostly covered the Standard.
5. Evidence in the materials fully covered all aspects of the Standard.

From the *Testing Standards*, the following eleven standards were identified for review:

- Standard 1.9. When a validation rests in part on the opinions or decisions of expert judges, observers, or raters, procedures for selecting such experts and for eliciting judgments or ratings should be fully described. The qualifications and experience of the judges should be presented. The description of procedures should include any training and instructions provided, should indicate whether participants reached their decisions independently, and should report the level of agreement reached. If participants interacted with one another or exchanged information, the procedures through which they may have influenced one another should be set forth.

- Standard 1.11. When the rationale for test score interpretation for a given use rests in part on the appropriateness of test content, the procedures followed in specifying and generating test content should be described and justified with reference to the intended population to be tested and the construct the test is intended to measure or the domain it is intended to represent. If the definition of the content sampled incorporates criteria such as importance, frequency, or criticality, these criteria should also be clearly explained and justified.

- Standard 1.12. If the rationale for score interpretation for a given use depends on premises about the psychological processes or cognitive operations of test takers, then theoretical or empirical evidence in support of those premises should be provided. When statements about the processes employed by observers or scorers are part of the argument for validity, similar information should be provided.

- Standard 2.3. For each total score, sub-score, or combination of scores that is to be interpreted, estimates of relevant indices of reliability/precision should be reported.

- Standard 3.2. Test developers are responsible for developing tests that measure the intended construct and for minimizing the potential for tests' being affected by construct-irrelevant characteristics, such as linguistic, communicative, cognitive, cultural, physical, or other characteristics.

- Standard 3.9. Test developers and/or test users are responsible for developing and providing test accommodations, when appropriate and feasible, to remove construct-irrelevant barriers that otherwise would interfere with examinees' ability to demonstrate their standing on the target constructs.

- Standard 4.0. Tests and testing programs should be designed and developed in a way that supports the validity of interpretations of the test scores for their intended uses. Test developers and publishers should document steps taken during the design and development process to provide evidence of fairness, reliability, and validity for intended uses for individuals in the intended examinee population.

- Standard 4.1. Test specifications should describe the purpose(s) of the test, the definition of the construct or domain measured, the intended examinee population, and interpretations for intended uses. The specifications should include a rationale supporting the interpretations and uses of test results for the intended purpose(s).

- Standard 4.6. When appropriate to documenting the validity of test score interpretations for intended uses, relevant experts external to the testing program should review the test specifications to evaluate their appropriateness for intended uses of the test scores and fairness for intended test takers. The purpose of the review, the process by which the review is conducted, and the results of the review should be documented. The qualifications, relevant experiences, and demographic characteristics of expert judges should also be documented.

- Standard 4.12. Test developers should document the extent to which the content domain of a test represents the domain defined in the test specifications.

- Standard 12.4. When a test is used as an indicator of achievement in an instructional domain or with respect to specified content standards, evidence of the extent to which the test samples the range of knowledge and elicits the processes reflected in the target domain should be provided. Both the tested and the target domains should be described in sufficient detail for their relationship to be evaluated. The analyses should make explicit those aspects of the target domain that the test represents, as well as those aspects that the test fails to represent.

All of the eleven identified standards were rated as fully covered based on the available evidence. These results indicate that the CAST test design and development processes and procedures closely adhere to the testing standards related to alignment of assessment content to academic standards.

CAST Alignment Workshop and Outcomes

This CAST alignment workshop was designed to collect evidence of whether the CAST produces test forms that effectively measure the content and cognitive rigor reflected in the targeted content domain and the test blueprints. During the workshop, educators with content expertise evaluated how well the 2019 test items represent the associated content standards, the California Next Generation Science Standards (CA NGSS).

Alignment Criteria Evaluated

Alignment criteria were developed by HumRRO and reviewed by staff from the National Center for Improvement in Educational Assessment (Center for Assessment). These criteria were developed based on the documentation provided by CDE and ETS (the testing contractor), and they represent several aspects of the overall alignment of the CAST to the CA NGSS. Failure to meet any single criterion does not indicate that the test is invalid or flawed in some way, only that that aspect of the assessment may need to be addressed through future item development or by other means.

Alignment criteria are grounded in the Webb alignment method (1997, 1999, 2002). The Webb method includes four major indicators to evaluate alignment. These indicators rely on statistical analyses to assess how well items on the assessment, regardless of item type and point value, match the state's content standards. The four alignment indicators are categorical concurrence, depth-of-knowledge consistency, range-of-knowledge correspondence, and balance-of-knowledge representation.

HumRRO drew from Webb's concepts (e.g., depth-of-knowledge) and the principles of Webb alignment criteria as the basis for developing alignment criteria specific to the CAST. Webb's criteria provided categories for creating alignment criteria more suited to three-dimensional assessments and content standards. For a full discussion of how and why the alignment criteria were created, see chapter 3. HumRRO developed the following modified criteria for evaluating the CAST: **Link to Standards, DOK Adequacy, Range Adequacy,** and **Balance-of-Knowledge Correspondence (Revised for Science),** or simply **Balance**. To address the multidimensional nature of the CAST, we added a fifth criterion, **Multidimensional Adequacy.**

Alignment Workshop Methods

HumRRO conducted the CAST Alignment Study Workshop in the Sacramento area on February 28 and March 1, 2019. HumRRO worked collaboratively with the CDE to recruit and select a group of 18 educators to serve on one of three CAST alignment review panels (grade five, grade eight, and high school) during the two-day workshop.

Across the three panels, 14 California school districts were represented. Approximately 50 percent of panelists reported being a current teacher (including lead teacher), and the remaining 50 percent reported working in roles such as coordinator, specialist, program director, or superintendent. In addition to their current professional roles, all panelists reported having some level of experience with the CA NGSS. The types of experience

reported ranged from teaching the new standards to students to providing CA NGSS-related training to other educators. Across the three panel groups, all panelists who provided responses reported experience teaching students from diverse socioeconomic and cultural backgrounds as well as experience teaching English learners.

HumRRO developed several data collection tools (see Appendix B) and adapted other materials to support the data collection process. Data collection tools included electronic spreadsheets for panelists and workshop facilitators to enter test item ratings. Support materials included copies of the CA NGSS and appendices (both paper and electronic), copies of the CAST item specifications, detailed workshop instructions for both panelists and facilitators, details on the cognitive complexity (DOK) rating categories and debriefing and evaluation forms. ETS created three online test forms for the alignment workshop (grade 5, 8 and high school) consisting of all the operational 2019 CAST items. ETS also created accounts for HumRRO researchers to securely access the items using the CAASPP Interim Assessment Viewing System (IAVS).

Alignment panelists received two rounds of training at the outset of the alignment workshop. First, the full group of panelists received general training that provided some background on alignment and a high-level description of the alignment process. Following the general training session, panelists moved into grade-level panel groups (grade 5, grade 8, and high school) and received more detailed training on the data collection (rating) processes and procedures.

After the panel-specific training presentation by the HumRRO facilitator, each panel engaged in a calibration activity using the first few (1–3) items. Panelists accessed the items electronically and made their independent ratings. Panelists discussed their independent ratings and engaged in consensus discussion to come to agreement on the final item ratings of record. Once panelists had a clear understanding of the rating process and a common understanding of the rating categories, they moved on to rating the remaining operational items.

Item ratings were generated via the following steps:

1. Panelists reviewed test items independently and assigned ratings of:
    a. PE measured by item
    b. DCI measured by item (up to two DCIs, primary and secondary)
    c. CCC measured by item (up to two CCCs, primary and secondary)
    d. SEP measured by item (up to two SEPs, primary and secondary)
    e. Item Depth of Knowledge (DOK)
    f. Comments to clarify ratings or to provide feedback on quality of item or associated phenomenon
2. Panelists discussed their independent ratings.
3. HumRRO facilitator shared item metadata provided by ETS.

4. Panelists came to consensus (or majority) ratings.
5. HumRRO facilitator recorded consensus/majority ratings.

The HumRRO facilitator recorded the final consensus (or majority) item ratings in a spreadsheet. Panelists then completed a debriefing form and a process evaluation survey before being released from the workshop. The debriefing form was designed to give panelists the opportunity to provide their individual, qualitative perspective on the quality of alignment. The evaluation survey elicited feedback about the quality of the workshop processes and procedures (see chapter 3 for more detail on workshop processes and procedures).

Alignment Workshop Results

Table ES.1 summarizes the alignment criteria results for the three summative assessment science test item pools. Across the three tests, panelists' ratings of the operational items provide strong support that the CAST is composed of multidimensional items that reflect a range of the CA NGSS. The ratings also support that the items generally reflect appropriate levels of cognitive complexity and a balance among the CA NGSS dimensions.

Table ES.1 Summary of Item Pool Results by Criterion and Grade Level

Criterion	Grade 5	Grade 8	High School
Links to Standards	Met	Met	Met
DOK Adequacy	Met	Partially met	Partially met
Range Adequacy	Met	Met	Met
Balance of Knowledge	Met	Partially met	Met
Multidimensional Adequacy	Met	Met	Met

Table ES.2 summarizes the test form alignment criteria results for the three summative assessment science tests. Similar to the item pool results, all test forms are composed of multidimensional items that reflect a range of the CA NGSS. Grade eight and high school test forms were evaluated as not fully reflecting an appropriate range of cognitive complexity levels, notably due to slightly more than 10 percent of items rated at DOK Level 1. Not all grade five and grade eight test forms were evaluated as fully reflecting an appropriate balance among the CA NGSS dimensions, though all calculated balance index values were within three points of the threshold value.

Table ES.2 Percentage of Grade Level Forms Fully Meeting Each Criterion

Criterion	Grade 5	Grade 8	High School
Links to Standards	100	100	100
DOK Adequacy	100	0[a]	0[a]
Range Adequacy	100	100	100
Balance of Knowledge	60[b]	93[b]	100
Multidimensional Adequacy	100	100	100

[a] 100 percent of grade eight and high school forms at least partially met the DOK Adequacy criterion.
[b] 100 percent of grade five and eight forms at least partially met the Balance-of-Knowledge criterion.

Overall, the alignment workshop results provide strong support that the CAST is designed to produce aligned test forms. All test forms at all grade levels at least partially met all five *a priori* alignment criteria that were evaluated. Alignment criteria that were not fully met for all test forms include Depth of Knowledge Adequacy and Balance of Knowledge.

Forms that did not meet the Depth of Knowledge Adequacy criterion contained slightly more Level 1 DOK items than the 10 percent maximum outlined in the criterion (see chapter 3 for an explanation of the alignment criteria applied). Note, also, that for each form, the number of Level 3 DOK items exceeded the ten percent minimum outlined. Failure to meet our proposed alignment criteria is often mitigated by demonstrating that test forms do meet goals outlined in test blueprints, which are reflective of the test's design and goals. At the time of this study, the CAST blueprints did not contain guidelines regarding the distribution of DOK levels. We recommend that such guidelines be added to the blueprint, along with a rationale for the range of items at each DOK level. Such a rationale might include, for example, that performance tasks are designed to lead students through simple to complex sense-making of the science phenomenon under investigation.

All forms that did not meet the Balance of Knowledge criterion were within three points of the minimum balance index threshold. This is likely the reflection of a single or very small number of items being aligned to one dimension over another. The CA NGSS dimensions are designed to be integrated; the categories of each tend to overlap. It is not uncommon for experts to disagree with one another on the specific SEP and CCC codes that should be assigned to a test item. Although no formal confidence intervals around the minimum balance index have been established (in prior alignment research or in this study), the proximity of the calculated index values to the threshold suggest all test forms demonstrated a reasonable level of balance among the SEP and CCC categories.

Conclusions

This study combined documentation review and item ratings by content experts to evaluate the alignment between the California Science Test (CAST) and the California Next Generation Science Standards (CA NGSS). Here we present the conclusions reached for each of the seven research questions posed at the beginning of the study:

Research Question 1: To what extent do the test design and test blueprints for the CAST support the claims to be made about student performance on the assessment?

Review of available documentation found that the test design and test blueprints for the CAST support the conclusion that the testing contractor adhered to testing standards relevant to test-to-standards alignment (see table 2.2). Review of operational test forms from the 2018–19 administration support that the CAST design produces aligned test forms (see table 3.35).

Research Question 2: To what extent does the test blueprint for the CAST represent an appropriate sampling of the content as set forth in the CA NGSS?

The CAST is designed such that its content at each grade level will rotate across years, each year sampling different content from the CA NGSS. The rotation is designed to allow CAST to address the full breadth of the CA NGSS over a three-year span. Table ES.3 compares the number of PEs that should be tested each year in order to meet the test blueprint with the number of PEs tested via the item pool in Year 1, based on expert panelists' ratings. The PEs assessed via the 2018–19 item pool are sufficient to support that the CAST is on track to address the full breadth of the CA NGSS after two additional operational administrations.

Table ES.3 Comparison of PE Needs Per Administration and PEs Tested in Year 1

CAST Item Pool Grade Level	Physical Sciences PEs Needed Per Year	Physical Sciences PEs Tested in Year 1	Life Sciences PEs Needed Per Year	Life Sciences PEs Tested in Year 1	Earth & Space Sciences PEs Needed Per Year	Earth & Space Sciences PEs Tested in Year 1
Grade 5	5–6	11	4	10	4–5	9
Grade 8	6–7	13	7	14	5	10
High School	8	10	8	12	6–7	9

Research Question 3: To what extent do the CAST test forms and test items reflect the test design and test blueprints?

Based on expert panelists' ratings, the number of items linked to each content domain, science and engineering practice, and crosscutting concept align with the guidelines presented in the CAST blueprints. In only a small number of instances did the number of items rated as aligned to a particular dimension fall slightly outside of the ranges specified in the blueprint. Tables depicting these comparisons are presented in Appendix C.

Research Question 4: To what extent do CAST tasks and items integrate more than one disciplinary core idea, crosscutting concept, and/or science and engineering practice?

Expert reviewers found that most of the CAST items, across the grade levels, measure a performance expectation by integrating a disciplinary core idea, crosscutting concept, and/or science and engineering practice (and are therefore multidimensional). Table ES.4 summarizes the percentage of items on each test form that were rated as multidimensional. Across the grade levels, the majority of items were rated as multidimensional, and more than half of items on any test form were rated as integrating all three dimensions.

Table ES.4 Summary of Multidimensional Items by Grade Level

Grade Level	Range of Percentages of Items Aligned to Two or More Dimensions	Range of Percentages of Items Aligned to All Three Dimensions
Grade 5	91–93	64–80
Grade 8	91–98	88–95
High School	98–100	84–86

Research Question 5: To what extent do CAST test forms show balance across the disciplinary areas used for scoring and reporting purposes (earth and space sciences, life sciences, and physical sciences)?

CAST forms across the grade levels reflect reasonable balance across the disciplinary areas used for scoring and reporting purposes (earth and space sciences, life sciences, and physical sciences), as well as across the CA NGSS science and engineering practices and crosscutting concepts. This was determined by calculating Webb's balance index for each. This index takes into consideration (a) the number of content domains, SEPs, and CCCs measured by the items and (b) the proportion of items measuring each domain, SEP, or CCC. For most forms across the grade levels, an *a priori*-defined minimum index was met. For a smaller number of forms, this index was missed by only three points on a 100-point scale.

Research Question 6: Do the CAST items range from low to high cognitive complexity and provide a sufficient number of items across the range of cognitive complexity?

Expert reviewers indicated that CAST items vary in cognitive complexity, with slightly more than the *a priori* limit of 10 percent at Level 1 DOK and also more than the *a priori* minimum of 10 percent at Level 3 DOK.

Research Question 7: How well does CAST fit the population being tested, in terms of the distribution of item difficulties within test forms and the distribution of student ability?

Item-person maps, or Wright Maps, illustrate the correspondence between test takers' ability and the difficulty of the test items. Ideally, test items will be at an appropriate level of difficulty to measure the test takers' ability level, ensuring that the test provides information about test performance that is meaningful and useful. For example, test scores on a test in which most items are too difficult for most test takers would result in an underestimation of true achievement levels. Item-person maps for each grade level were produced by ETS. HumRRO conducted additional item mapping analyses, classifying items into achievement levels based on the score associated with having a 50 percent probability of responding correctly to an item (or receiving full points for a multi-point items). This classification represents the achievement level at which each item is providing the most information about student performance. Item-person maps and item-achievement level classification results are presented in Appendix D.

In the evaluation of this operational administration, the item-person maps in Appendix D generally depict item difficulty being aligned with students' ability. For all three grades, the distribution of item difficulties generally lines up with the distribution of student ability levels. For high school, the item difficulty distribution relative to the student ability distribution has a slightly more upward shift compared to the other two grades. This indicates that the high school test has fewer items that are at a difficulty level that is comparable to students on the lower end of the ability distribution. Across grade levels and forms, item-achievement level classifications indicate that the largest percentage of items tended to be classified at Achievement Level 2, with some exceptions. In grade eight and high school, there were some forms in which a slightly higher percentage of items were rated at Achievement Level 4. This is in part due to multipoint items being classified based on the probability of earning full points (i.e., the ability level associated with having a 50% probability of getting the full two points on a two-point test item). Classifying items based on the probability of earning at least partial points (i.e., the ability level associated with having a 50% probability of getting at least one point on a two-point test item) would likely result in fewer items classified at Achievement Level 4.

Classifying items into achievement levels provides insight into how well a test form can differentiate among different levels of student performance. This is done by calculating the probability of answering each item correctly at each student ability level. Items are then classified into achievement levels based on the student ability level associated with having a 50% probability of answering the item correctly. During standard setting, CAST

achievement levels were set such that the largest percentage of students are expected to be classified at Achievement Level 2 based on the 2018–19 spring operational test administration. Thus, it makes sense that a large proportion of items would be targeting students at this level. But test forms also contained items targeting the higher achievement levels, and, to a lesser extent, Level 1 Achievement, thus providing information about student performance at all levels. It is important to note that California educators are still developing strategies for teaching the CA NGSS in the classroom. As students have more opportunities to learn the CA NGSS, the correspondence between student ability and item difficulty is expected to shift.

Recommendations

The study results were generally very positive and do not indicate that any major changes in test development or forms construction processes and procedures are needed. We do offer one recommendation for improving the CAST blueprints:

1. **Add recommended cognitive complexity distributions to the CAST blueprints, along with a rationale for the targets set for each level.**

Establishing guidelines for cognitive complexity in the CAST blueprints will enhance item development and test form construction by clearly stating the proportions of items at each cognitive complexity level that each test form should include. This information will be helpful in ongoing evaluations of the adequacy of the item pool for building multiple test forms and for verifying that forms contain items from an appropriate range of cognitive complexity levels. These guidelines should include a rationale for each cognitive complexity level, noting why some levels are emphasized over others and how this design reflects the intent of the CA NGSS as well as the interpretation and use of CAST scores.

This page is intentionally blank.

Chapter 1: Introduction

Background

The California Assessment of Student Performance and Progress (CAASPP) System, launched in 2014, was intended to assist teachers, administrators, students, and parents by promoting high-quality teaching and learning using a variety of assessment approaches and item types. The statewide student assessments monitor progress in implementing effective instruction aligned with the Common Core State Standards (CCSS) for English language arts/literacy (ELA) and mathematics and the California Next Generation Science Standards (CA NGSS). The Smarter Balanced ELA and mathematics tests have been operational since 2014, and the California Alternate Assessments in ELA and mathematics have been operational since 2016. The California Science Test (CAST) became operational in spring 2019, and the California Alternate Assessment in Science (CAA Science) became operational during the 2019–20 school year. The CAASPP System also includes an optional Spanish reading language arts test, the California Spanish Assessment (CSA), which became operational in 2019. These assessments aim to shift the focus away from accountability toward a comprehensive plan for promoting high-quality teaching and learning for all students, including students with disabilities (SWDs) and English learners (ELs). The CAASPP System represents a substantial financial investment by the state as well as a significant investment of educator and student time.

California *Education Code* (*EC*) Section 60649 requires the independent evaluation of the CAASPP System, stating that "evaluation activities may include a variety of internal and external studies such as validity studies, alignment studies, and studies evaluating test fairness, testing accommodations, testing policies, and reporting procedures, and consequential validity studies specific to pupil populations such as English learners and pupils with disabilities." The law requires development of a plan to assess independent evaluation activities, and it prohibits duplication of studies conducted as part of a federal peer review process or by California Department of Education (CDE) assessment contractors.

The Human Resources Research Organization (HumRRO) served as the first CAASPP System evaluator from 2015–18. Copies of our annual and comprehensive final reports are publicly available online (https://www.cde.ca.gov/ta/tg/ca/caaspprptstudies.asp).

The CDE awarded the contract for the 2018–20 independent evaluation of the CAASPP System to HumRRO in July 2018. The current contract calls for annual evaluation reports that summarize all work completed during the previous year, stand-alone reports for individual research studies, and a comprehensive final report. Within a few months of the award, HumRRO submitted to the CDE the first required annual evaluation report (Hardoin, Thacker, Dvorak, Becker, 2018). That report's core contents included the 2018–20 Evaluation Plan, which described the design of three research studies approved by the CDE and scheduled within the contract period. HumRRO recently submitted to the CDE the second annual evaluation report (Hardoin, Dvorak, Thacker,

Paulsen, Gribben, Handy, 2019). That report described activities conducted and results obtained to date from the 2018–19 studies. The present report is a stand-alone report for the CAST Alignment Study. A Comprehensive Final Evaluation Report will be delivered in 2020 and will include evaluation findings from each of the three annual reports (2018, 2019, and 2020).

HumRRO approaches alignment studies as one means to gather evidence to demonstrate the validity of intended interpretations and uses of the assessment scores. Alignment studies can tell us how well a set of test items fully samples the construct represented by the associated content standards. That is, alignment studies indicate whether a test effectively measures what it is intended to measure.

For the CAST, evaluating alignment represents a significant challenge because of the nature of the content and the content standards. The California Next Generation Science Standards (CA NGSS) provide a framework for science education. Within the CA NGSS, performance expectations (PEs) are assessable statements of what students should know and be able to do. The following three major components, also referred to as dimensions, are combined to operationalize the PEs:

1. Disciplinary Core Ideas (DCIs) are the key ideas in science that have broad importance within or across multiple science or engineering disciplines. These core ideas build on each other as students progress through grade levels. The DCIs are grouped into the following domains: Physical Sciences, Life Sciences, Earth and Space Sciences, and Engineering, Technology, and the Application of Science (hereafter, Engineering).

2. Crosscutting Concepts (CCCs) help students explore connections across the four domains of science mentioned above in item 1. When these concepts, such as "cause and effect," are made explicit for students, they can help students develop a coherent and scientifically based view of the world around them.

3. Science and Engineering Practices (SEPs) describe what scientists do to investigate the natural world and what engineers do to design and build systems. The practices better explain and extend what is meant by "inquiry" in science and the range of cognitive, social, and physical practices that it requires. Students engage in practices to build, deepen, and apply their knowledge of core ideas and crosscutting concepts.

The three major components of the CA NGSS (DCIs, CCCs, and SEPs) are integrated into the three science disciplines (earth and space sciences, life sciences, and physical sciences). In the CAST test design, each of these three disciplines assesses engineering, technology, and application of science. The design of the test is further complicated by the premise that students' knowledge is expected to be integrated and to accumulate to create a deep understanding of science content. Students are expected to apply their knowledge and generalize across the three major components. Developing tests and test items that adequately sample such complex and integrated content is especially challenging. When an item measures a single standard or concept,

the alignment process is relatively straightforward. However, test development and alignment become more complex when standards are designed as interactions among statements about content.

The first step in evaluating for CAST alignment was to investigate the nature of the assessment itself: how the standards guided the development of the test items (and how the standards and items should therefore relate to one another) and the interpretations to be made from CAST scores. HumRRO then modified traditional alignment methods to account for the test structure and design, a process in keeping with best practices in test validation that facilitates using alignment study results in an overall validity argument. This process also supports federal peer review goals.

The CAST is a computer-based, fixed-form (non-adaptive) assessment administered to students in grades five, eight, and once in high school (i.e., grades 10, 11, or 12). The CAST was field-tested in spring 2018 and administered operationally for the first time in January–July of 2019. The 2019 assessment included the following three segments:

- Segment A: a set of selected response and short constructed-response items (two blocks (A1 and A2) were administered operationally in 2019). Each tested student was administered blocks A1 and A2.

- Segment B: a set of two performance tasks (PT) (five performance tasks were available for Segment B in grade five, six in grade eight, and three in high school; two were selected per test form.). Each tested student was administered two segment B performance tasks.

- Segment C: a set of items comparable to Segment A or B, highly matrixed across test forms, each taken by a smaller sample of students than Segments A or B. Segment C included only field test items (discrete and PT), not operational items.

For the 2019 CAST, results from the first two segments were used to report individual student scores. Segment C was not used for individual score reporting but for field test purposes only. The high-level test design planned for a portion of Segment C to include operational items that would provide school- and LEA-level information about student achievement on a broader sample of content than would be possible otherwise. At the time of this alignment investigation, only Segments A and B were administered operationally. All results in this report are based on Segments A and B only.

Because students who took the test in 2019 could potentially have been administered any combination of Segments A and B, a student testing event (or test form) was defined in this evaluation as any possible combination of Segments A and B. This means that there were 10 possible forms for grade five, 15 for grade eight, and 3 for high school. Alignment analyses will be conducted for each potential form and the results will be summarized across forms. Table 1.1 presents the possible combinations of A and B Segments (all students administered both A1 and A2, plus all possible combinations of the available B segments).

Table 1.1 Combinations of A and B Segments

Grade Level	A Segment	B Segment
5	A1A2	B1B2 B1B3 B1B4 B1B5 B2B3 B2B4 B2B5 B3B4 B3B5 B4B5
8	A1A2	B1B2 B1B3 B1B4 B1B5 B1B6 B2B3 B2B4 B2B5 B2B6 B3B4 B3B5 B3B6 B4B5 B4B6 B5B6
High School	A1A2	B1B2 B1B3 B2B3

It should be noted that the operational test design for 2018–19 does not represent the final design. Changes to the CAST design and blueprint were approved by the California State Board of Education (SBE) in January 2020. Specifically, the number of stand-alone items in Segments A and C were reduced to allow for a third performance task in a third science domain to be added to Segment B without extending student testing time. The screener, originally planned to be used as the student transitioned from Segment A to Segment B, was eliminated. Segment C will be used for field testing purposes only and will not be used for group reporting, as initially planned. These changes do not impact the conclusions drawn from this study. The increase in performance tasks improves content representation to offset the loss of content representation from the stand-alone items. The screener was not part of this study, and group reporting was not addressed in this report.

The content of the CAST will also rotate across years, each year sampling different content from the CA NGSS. The rotation is designed to allow CAST to address the full breadth of the CA NGSS in a three-year span. This alignment study was conducted during the first operational year of testing, so it will not be possible to evaluate how well CAST addresses the breadth of the content standards over three years. HumRRO will be able to use the initial year's data, however, to estimate whether one administration can address roughly one third of the intended PEs. If so, the three-year rotation is feasible as a sampling plan for addressing the full breadth of the CA NGSS.

The CAST blueprint indicates, "For scoring and reporting purposes, each of the three science domains will constitute one third of the test (items written to assess PEs associated with Engineering, Technology, and Application of Science will be assigned to one of the three science domains, depending upon the context of their stimulus)." It continues, "For the segments contributing to individual student scores (Segment A and Segment B), it is not possible to assess all PEs in a single testing year. For example, there are 14 PEs assessed in grade five, each of which would require multiple items to fully assess. As a result, PEs assessed in Segment A and Segment B will be rotated from year to year so that all PEs can be assessed in the segments contributing to individual scores over the course of a three-year period." HumRRO will use student-level alignment results to evaluate the CAST to ensure that the PEs contributing to individual student scores on the 2018–19 CAST are adequate to support coverage of the full set of PEs over three administrations.

Research Questions

Activities conducted for the CAST Alignment Study were designed to provide information to answer the following research questions:

1. To what extent do the test design and test blueprints for the CAST support the claims to be made about student performance on the assessment?

2. To what extent does the test blueprint for the CAST represent an appropriate sampling of the content as set forth in the CA NGSS?

3. To what extent do the test forms and test items for the CAST reflect the test design and test blueprints?

4. To what extent do CAST tasks and items integrate more than one disciplinary core idea, crosscutting concept, and/or science and engineering practice?

5. To what extent do test forms show balance across the disciplinary areas of the CAST used for scoring and reporting purposes (earth and space sciences, life sciences, and physical sciences)?

6. Do the CAST items range from low to high cognitive complexity and provide a sufficient number of items across the range of cognitive complexity?

7. How well does CAST fit the population being tested, in terms of the distribution of item difficulties within test forms and the distribution of student ability?

Organization and Contents of the CAST Alignment Study Report

The remaining chapters and appendices of this report describe the CAST Alignment Study activities, findings, and conclusions.

- Chapter 2, *Review of CAST Documentation*, presents the methods, rating scale, and data analysis activities HumRRO conducted to evaluate the alignment of development documentation of the CAST to relevant *Standards for Educational and Psychological Testing* (AERA, APA & NCME, 2014), hereafter referred to as *Testing Standards*. The chapter identifies the list of CAST documents reviewed for each test standard and describes the rationale for HumRRO's alignment rating. The chapter concludes with a summary of HumRRO's evaluation of CAST documentation.

- Chapter 3, *CAST Alignment Workshop and Outcomes*, presents HumRRO's method for evaluating the alignment of the pool of CAST items and the 2019 CAST forms to the CA NGSS and CAST blueprint. The chapter presents HumRRO's five alignment criteria; describes the alignment workshop data collection activities, including panelist training and item rating procedures; and presents results of data analysis. The results section provides outcomes by grade level (i.e., grades 5, 8, and high school) for each alignment criterion. The chapter concludes with an overall summary of HumRRO's evaluation of the alignment of CAST grade-level item pools and test forms.

- Chapter 4, *Conclusions*, presents HumRRO's response to each of the seven research questions in the alignment study. HumRRO's responses were informed by results of the CAST documentation review and the CAST item ratings by content experts. The chapter concludes with an overall summary of HumRRO's evaluation of the alignment of CAST grade-level item pools and test forms.

- Appendix A, *CAST Documentation Reviewed by HumRRO,* lists the file names of all documents reviewed for the study. Documents are grouped by these topics of focus: (a) CA NGSS standards, core concepts, and performance expectations; (b) test design; (c) item development and information; (d) DOK information; (e) test fairness, accessibility, and accommodations; (f) item scoring; (g) field test; and (h) teacher training.

- Appendix B, *Alignment Workshop Materials*, includes documents provided to content experts participating in the workshop. Materials include the workshop agenda, panelist item rating instructions, debriefing questions, and evaluation of alignment workshop training and procedures.

- Appendix C, *Test Form-Blueprint Comparison*, presents tables comparing results of panelists' item ratings with test blueprint ranges for domains and SEPs, by grade level. Test blueprint ranges are for Segment A items only.

- Appendix D, *Item-Person Maps and Item-to-Achievement Level Classifications*, presents ETS's item-person maps, which display for each grade level the comparison between CAST item difficulty and student performance (scale scores), and HumRRO's item-to-achievement level classifications, which summarize the achievement levels at which CAST items provide the most information about student performance.

This page is intentionally blank.

Chapter 2: Review of CAST Documentation

Introduction

To begin the alignment study and build knowledge of the California Science Test (CAST), HumRRO researchers collected and reviewed CAST design and test development materials provided by California Department of Education (CDE) and Educational Testing Service (ETS) staff, as well as information about the CAST shared with the public on the CDE website.

HumRRO researchers completed the first major task of the alignment study by conducting an evaluation of the alignment of CAST test design and development documentation to the *Standards for Educational and Psychological Testing* (AERA, APA & NCME, 2014; hereafter referred to as the *Testing Standards*). This chapter presents the methods and outcome of the evaluation of CAST documentation. The review of CAST materials also informed HumRRO's plans and preparation for the second major task of the study, the alignment workshop.

Method

HumRRO's evaluation of the test design and development documentation was informed by industry best practices as outlined in the *Testing Standards*. First, HumRRO researchers identified specific standards from the *Testing Standards* that are directly relevant to how alignment is considered during test development. Next, researchers identified and collected the types of documentation needed to provide evidence that these standards were met. Finally, two HumRRO researchers independently reviewed the documentation and rated the extent to which each standard was met. These independent ratings were compared and discussed to reach a final consensus rating for each standard.

Documents Collected

HumRRO worked in cooperation with the CDE and ETS to obtain documentation related to the design and development of the CAST. We also searched CDE CAASPP website pages to identify additional relevant information. Appendix A lists the full complement of documents HumRRO collected and reviewed. The documents generally focus on the following areas: CA NGSS standards, core concepts, and performance expectations; test design; item development and information; DOK (i.e., cognitive complexity) information; test fairness, accessibility, and accommodations; item scoring; field test; and teacher training.

Rating Scale

HumRRO developed a rating scale to evaluate the degree to which the evidence for the assessment supports adherence to these testing standards. The rating scale ranged

from 1 to 5, with higher scores indicating stronger evidence for compliance with the standard (See table 2.1).

Table 2.1 Rating Scale for Evaluating Strength of Evidence for Testing Standards

Rating Level	Description
1	No evidence of the Standard found in the materials.[a]
2	Little evidence of the Standard found in the materials; less than half of the Standard was covered in the materials and/or evidence of key aspects of the Standard could not be found.
3	Some evidence of the Standard found in the materials; approximately half of the Standard covered in the materials, including some key aspects of the Standard.
4	Evidence in the materials mostly covered the Standard.
5	Evidence in the materials fully covered all aspects of the Standard.

[a] Materials include all documents and data provided, any emails or phone calls with CDE and/or ETS staff, as well as information available on the CDE website.

Results

Ratings for Testing Standards

The results in table 2.2 represent the outcomes of HumRRO's review of assessment planning and item development processes. The leftmost column in table 2.2 presents the evaluated testing standards. The center column lists the names of the files provided by ETS as supporting documentation for the processes and procedures related to each evaluated testing standard. Finally, the rightmost column provides an overall rating for each testing standard based on our review of this supporting documentation.

Table 2.2 Ratings on the Testing Standards for CAST Alignment

Standard	Supporting Documentation	Standard Rating
Standard 1.9. When a validation rests in part on the opinions or decisions of expert judges, observers, or raters, procedures for selecting such experts and for eliciting judgments or ratings should be fully described. The qualifications and experience of the judges should be presented. The description of procedures should include any training and instructions provided, should indicate whether participants reached their decisions independently, and should report the level of agreement reached. If participants interacted with one another or exchanged information, the procedures through which they may have influenced one another should be set forth.	• VH651815 Benchmark and Annotation Examples Table • VH651815_Fossil Map_Scoring Notes • HS_DRAFT_Content Training for Raters and Scoring Leaders_021519 • ETS Machine Scoring Introduction • CAST Constructed Response Scoring Overview • 234-2018C_FOR REVIEW_CAST Technical Report._022119	5

Table 2.2 (cont.)

Standard	Supporting Documentation	Standard Rating
Standard 1.11. When the rationale for test score interpretation for a given use rests in part on the appropriateness of test content, the procedures followed in specifying and generating test content should be described and justified with reference to the intended population to be tested and the construct the test is intended to measure or the domain it is intended to represent. If the definition of the content sampled incorporates criteria such as importance, frequency, or criticality, these criteria should also be clearly explained and justified.	• castblueprint • 239-2016 FOR ARCHIVE – CA NGSS GEN Assessment Design_022416: intended uses of results • DRAFT CAST Evidence-Centered Design White Paper 070218: Item and test development • CAST Academy Item Specs grade 5, 8, and HS • 110317-01_FOR ETS_CAST_IWW_Nov 2017_110717_FINAL • CAASPP Item Acceptance Criteria for IRC 021618_v3 • CAST OIW Part 3 Gr5_Final • CAST OIW Part 3 HS_Final • CAST OIW Part 3 MS_Final • CAST OIW Part 4_PT_FINAL • Item Authoring Template • Item_Review_040218 • PT_WritingTemplate • Form Planners_all grades (7) folder • 239-2016 FOR ARCHIVE – CA NGSS GEN Assessment Design_022416 • 167-2019C-v3_FOR ARCHIVE CAST_Test_Specs_092518: Item selection guidelines for operational forms	5

Table 2.2 (cont.)

Standard	Supporting Documentation	Standard Rating
Standard 1.12. If the rationale for score interpretation for a given use depends on premises about the psychological processes or cognitive operations of test takers, then theoretical or empirical evidence in support of those premises should be provided. When statements about the processes employed by observers or scorers are part of the argument for validity, similar information should be provided.	• 239-2016 FOR ARCHIVE – CA NGSS GEN Assessment Design_022416: intended uses of results • DRAFT CAST Evidence-Centered Design White Paper 070218: Item and test development • CAST Academy Item Specs grade 5, 8, and HS • 110317-01_FOR ETS_CAST_IWW_Nov 2017_110717_FINAL • CAST ECD White Paper-2nd submission to CDE 6-29-2018 • CAASPP Item Acceptance Criteria for IRC 021618_v3 • CAST OIW Part 3 Gr5_Final • CAST OIW Part 3 HS_Final • CAST OIW Part 3 MS_Final • CAST OIW Part 4_PT_FINAL • Item Authoring Template • Item_Review_040218 • PT_WritingTemplate • Form Planners_all grades (7) folder • 239-2016 FOR ARCHIVE – CA NGSS GEN Assessment Design_022416 • 167-2019C-v3_FOR ARCHIVE CAST_Test_Specs_092518: Item selection guidelines for operational forms • 15_Facilitators-Guide • 12_CAST-Academy-Slides-Handout • VH651815 Benchmark and Annotation Examples Table • VH651815_Fossil Map_Scoring Notes • HS_DRAFT_Content Training for Raters and Scoring Leaders_021519 • ETS Machine Scoring Introduction • CAST Constructed Response Overview • Assigning of DOK_v2 • CriteriaforStatewidesummativescience assessments_03192018 • DOK-Science • Webbs_DOK_Guide	5

Table 2.2 (cont.)

Standard	Supporting Documentation	Standard Rating
Standard 2.3. For each total score, sub-score, or combination of scores that is to be interpreted, estimates of relevant indices of reliability/precision should be reported.	• DRAFT CAST Evidence-Centered Design White Paper 070218 • 234-2018C_FOR REVIEW_CAST Technical Report._022119	5
Standard 3.2. Test developers are responsible for developing tests that measure the intended construct and for minimizing the potential for tests being affected by construct-irrelevant characteristics, such as linguistic, communicative, cognitive, cultural, physical, or other characteristics.	• DRAFT CAST Evidence-Centered Design White Paper 070218 • ATF ETS Accessibility Handbook Content Development: Accessibility of Test Content • CAST Editorial & Graphics Style Guide V4 • CAST_Review discrete Items Process Map_D013019 • ETS Guidelines for Fair Tests and Communications • ETS Standards for Quality and Fairness • Fairness Review Book for Assessment Specialists • Fairness Training PowerPoint for Assessment Specialists • Universal Design Training for Assessment Specialists • 234-2018C_FOR REVIEW_CAST Technical Report._022119 • 102717-06-v3_FOR ARCHIVE_CAST Item Type Specifications_030718	5

Table 2.2 (cont.)

Standard	Supporting Documentation	Standard Rating
Standard 3.9. Test developers and/or test users are responsible for developing and providing test accommodations, when appropriate and feasible, to remove construct-irrelevant barriers that otherwise would interfere with examinees' ability to demonstrate their standing on the target constructs.	• DRAFT CAST Evidence-Centered Design White Paper 070218: Accessibility • 239-2016 FOR ARCHIVE – CA NGSS GEN Assessment Design_022416: Appendix C • castaccesssupt (accessibility supports for operational testing) • Sciencebentobox0918 • Applying Principle of Digital Accessibility training for Assessment Specialists • ATF ETS Accessibility Handbook Content Development • Universal Design Training for Assessment Specialists • 102717-06-v3_FOR ARCHIVE_CAST Item Type Specifications_030718 • CAST Accessibility Supports, Operational Testing	5
Standard 4.0. Tests and testing programs should be designed and developed in a way that supports the validity of interpretations of the test scores for their intended uses. Test developers and publishers should document steps taken during the design and development process to provide evidence of fairness, reliability, and validity for intended uses for individuals in the intended examinee population.	• DRAFT CAST Evidence-Centered Design White Paper 070218 • For-Kbacher_060618-01-v2 FOR ETS_CAST_Field Test Data Review PowerPoint_Final • 239-2016 FOR ARCHIVE – CA NGSS GEN Assessment Design_022416 • CAST Academy Item Specs grade 5, 8, and HS • 060618-02-v2 FOR ARCHIVE_CAST Field Test Data Review Reference Sheet 061	5

Table 2.2 (cont.)

Standard	Supporting Documentation	Standard Rating
Standard 4.1. Test specifications should describe the purpose(s) of the test, the definition of the construct or domain measured, the intended examinee population, and interpretations for intended uses. The specifications should include a rationale supporting the interpretations and uses of test results for the intended purpose(s).	• castblueprint • DRAFT CAST Evidence-Centered Design White Paper 070218 • 239-2016 FOR ARCHIVE – CA NGSS GEN Assessment Design_022416	5
Standard 4.6. When appropriate to documenting the validity of test score interpretations for intended uses, relevant experts external to the testing program should review the test specifications to evaluate their appropriateness for intended uses of the test scores and fairness for intended test takers. The purpose of the review, the process by which the review is conducted, and the results of the review should be documented. The qualifications, relevant experiences, and demographic characteristics of expert judges should also be documented.	• DRAFT CAST Evidence-Centered Design White Paper 070218: consultation with external advisors	5

Table 2.2 (cont.)

Standard	Supporting Documentation	Standard Rating
Standard 4.12. Test developers should document the extent to which the content domain of a test represents the domain defined in the test specifications.	• 239-2016 FOR ARCHIVE – CA NGSS GEN Assessment Design_022416 • DRAFT CAST Evidence-Centered Design White Paper 070218	5
Standard 12.4. When a test is used as an indicator of achievement in an instructional domain or with respect to specified content standards, evidence of the extent to which the test samples the range of knowledge and elicits the processes reflected in the target domain should be provided. Both the tested and the target domains should be described in sufficient detail for their relationship to be evaluated. The analyses should make explicit those aspects of the target domain that the test represents, as well as those aspects that the test fails to represent.	• castblueprint • 239-2016 FOR ARCHIVE – CA NGSS GEN Assessment Design_022416: Plan for Sampling CA NGSS Content • CAST_2019 OP_Gr5_Non accessible BB_for CDE_v11 • CAST_2019 OP_Gr8_Non accessible BB_for CDE_v9 • CAST_2019 OP_HS_Non accessible BB_for CDE_v9 • Updated DOK CAST_2019 OP • 060118-02-v3_FOR ARCHIVE_CAST ECD White Paper • 168-2018-v2_FOR_ARCHIVE_IDP_032118 • CAST_Gr5_2020 NID_IDP_v01 • CAST_Gr8_2020 NID_IDP_v01 • CAST_HS_2020 NID_IDP_v01	5

Rationales for Ratings for Testing Standards

This section presents the rationales for HumRRO's ratings in table 2.2 and explains to what extent each relevant testing standard was met based on evidence from the test development documentation. HumRRO also provides suggestions for further strengthening compliance with the testing standards.

Standard 1.9. When a validation rests in part on the opinions or decisions of expert judges, observers, or raters, procedures for selecting such experts and for eliciting judgments or ratings should be fully described. The qualifications and experience of the judges should be presented. The description of procedures should include any training and instructions provided, should indicate whether participants reached their decisions independently, and should report the level of agreement reached. If participants interacted with one another or exchanged information, the procedures through which they may have influenced one another should be set forth.

"CAST Constructed Response Scoring Overview" describes the qualifications necessary for raters to be hired. The document provides description of training for each level of the rating hierarchy (Chief Scoring Leaders, Scoring Leaders, and Raters). An overview training presentation, "HS_Draft_Content Training for Raters and Scoring Leaders", and an individual item's scoring guide, scoring notes, and benchmarks illustrate the details necessary for effective scoring of CAST constructed response items. "234-2018C_FOR REVIEW_CAST Technical Report_022119" provides description of the level of agreement measures used and the empirical results of those measures.

Standard 1.11. When the rationale for test score interpretation for a given use rests in part on the appropriateness of test content, the procedures followed in specifying and generating test content should be described and justified with reference to the intended population to be tested and the construct the test is intended to measure or the domain it is intended to represent. If the definition of the content sampled incorporates criteria such as importance, frequency, or criticality, these criteria should also be clearly explained and justified.

This standard is covered extensively by the supporting documents. The test structure is described in the blueprint and the test specifications documents. The rationale for test design and score interpretation, and the procedures followed in specifying and generating test content are described and justified in the Evidence Centered Design (ECD) White Paper. Detailed information about how the items were developed is provided in the item authoring template and the PDF files of item review workshops for different grades and item types.

Standard 1.12. If the rationale for score interpretation for a given use depends on premises about the psychological processes or cognitive operations of test takers, then theoretical or empirical evidence in support of those premises should be provided. When statements about the processes employed by

observers or scorers are part of the argument for validity, similar information should be provided.

The documents "239-2016 FOR ARCHIVE – CA NGSS GEN Assessment Design_022416" and ECD White Paper provide a description of human and machine item scoring. The document "239-2016 FOR ARCHIVE – CA NGSS GEN Assessment Design_022416" provides a description of technology enhanced item types and the supports allowed for students to increase assessment accessibility. Detailed information about how the items were developed is provided in the item authoring template and the PDF files of item review workshops for different grades. To further support evidence for this standard, additional information about machine scoring procedures and metrics, and training protocols for teachers for hand scoring items is needed. The additional information should also include specifications about how training is evaluated and determined to be effective.

The documents "110317-01_FOR ETS_CAST_IWW_Nov 2017_110717_FINAL", "CAST ECD White Paper-2nd submission to CDE 6-29-2018", OIW Grade specific documents, Form Planner documents, and VH651815 series all illustrate how CAST items are designed to require multidimensional cognitive operations. The VH651815 series of documents and "HS_DRAFT_Content Training for Raters and Scoring Leaders_021519" demonstrate how raters should approach rating multidimensional items.

Standard 2.3. For each total score, sub-score, or combination of scores that is to be interpreted, estimates of relevant indices of reliability/precision should be reported.

The document "DRAFT CAST Evidence-Centered Design White Paper 070218" provides information about how scores should be interpreted. Also, item files provide information about item parameter SEM. Reliability estimates were reported in the field test technical report "234-2018C_FOR REVIEW_CAST Technical Report. _022119". Reliability and SEM estimates were reported for total scores and domains (Physical Sciences, Life Sciences, Earth and Space Sciences) by form for each grade. Interrater reliability measures were also reported for each of the constructed response items.

Standard 3.2. Test developers are responsible for developing tests that measure the intended construct and for minimizing the potential for tests' being affected by construct-irrelevant characteristics, such as linguistic, communicative, cognitive, cultural, physical, or other characteristics.

The technical report, "234-2018C_FOR REVIEW_CAST Technical Report. _022119" provides a high-level description of the editorial, sensitivity, and fairness review CAST items underwent. Further detail of the depth and breadth of these reviews is included in multiple ETS documents including the "CAST Editorial & Graphics Style Guide V4", "CAST_Review discrete Items Process Map_D013019", "ETS Guidelines for Fair Tests and Communications", "ETS Standards for Quality and Fairness", "Fairness Review Book for Assessment Specialists", "Fairness Training PowerPoint for Assessment

Specialists", and "Universal Design Training for Assessment Specialists". These documents and trainings review each of the possible construct-irrelevant characteristics outlined in the standard. Differential Item Functioning (DIF) analyses were run for each of the items across gender, race, ELL status, special education, and socio-economic categories. Also, proficiency category distributions were compared across different demographic groups.

Standard 3.9. Test developers and/or test users are responsible for developing and providing test accommodations, when appropriate and feasible, to remove construct-irrelevant barriers that otherwise would interfere with examinees' ability to demonstrate their standing on the target constructs.

This standard overlaps with Standard 3.2. The document "DRAFT CAST Evidence-Centered Design White Paper 070218" provides information about the intended construct being measured and describes accommodations that may be provided to the students. The document "239-2016 FOR ARCHIVE – CA NGSS GEN Assessment Design_022416: Appendix C" also outlines student accommodations; the accommodations for CAST are documented in greater detail in "CAST Accessibility Supports, Operational Testing," examining how each accommodation responds to a particular student need. ETS also provides (a) an "ETS Accessibility Handbook" with guidance for developing fully accessible tests and (b) "Applying Principles of Digital Accessibility training" that describes ETS approaches to task design for item accessibility.

Standard 4.0. Tests and testing programs should be designed and developed in a way that supports the validity of interpretations of the test scores for their intended uses. Test developers and publishers should document steps taken during the design and development process to provide evidence of fairness, reliability, and validity for intended uses for individuals in the intended examinee population.

The documents "DRAFT CAST Evidence-Centered Design White Paper 070218," "239-2016 FOR ARCHIVE – CA NGSS GEN Assessment Design_022416," "CAST Academy Item Specs grade 5, 8, and HS," and "060618-02-v2 FOR ARCHIVE_CAST Field Test Data Review Reference Sheet 061" provide detailed descriptions of the test development process. The document "For-Kbacher_060618-01-v2 FOR ETS_CAST_Field Test Data Review PowerPoint_Final" also describes how the items functioned on a pilot test for the general population of students. Document "239-2016 FOR ARCHIVE – CA NGSS GEN Assessment Design_022416" describes the assessment design, outlining intended uses of the assessment results, including that the test will not be used for high-stakes purposes at the individual student level. It also describes plans to incorporate best practices such as evidence-centered design (ECD) principles and practices in item development. The CAST Item Specifications detail how the CA NGSS guided item development, and "060618-02-v2 FOR ARCHIVE_CAST Field Test Data Review Reference Sheet 061" outlines the rules for flagging potentially problematic items, including flags for items that may reflect student characteristics other than ability.

Standard 4.1. Test specifications should describe the purpose(s) of the test, the definition of the construct or domain measured, the intended examinee population, and interpretations for intended uses. The specifications should include a rationale supporting the interpretations and uses of test results for the intended purpose(s).

The documents "castblueprint," "DRAFT CAST Evidence-Centered Design White Paper 070218," and "239-2016 FOR ARCHIVE – CA NGSS GEN Assessment Design_022416" describe the purpose of the test, the intended population, and interpretation of intended uses.

Standard 4.6. When appropriate to documenting the validity of test score interpretations for intended uses, relevant experts external to the testing program should review the test specifications to evaluate their appropriateness for intended uses of the test scores and fairness for intended test takers. The purpose of the review, the process by which the review is conducted, and the results of the review should be documented. The qualifications, relevant experiences, and demographic characteristics of expert judges should also be documented.

The section "Consultation with internal advisors" of the document "DRAFT CAST Evidence-Centered Design White Paper 070218" provides information about content experts' review of the assessment. The document also contains the biographical sketches of the reviewers.

Standard 4.12. Test developers should document the extent to which the content domain of a test represents the domain defined in the test specifications.

The documents "239-2016 FOR ARCHIVE – CA NGSS GEN Assessment Design_022416" and "DRAFT CAST Evidence-Centered Design White Paper 070218" provide detailed information about which content domains are represented on the test.

Standard 12.4. When a test is used as an indicator of achievement in an instructional domain or with respect to specified content standards, evidence of the extent to which the test samples the range of knowledge and elicits the processes reflected in the target domain should be provided. Both the tested and the target domains should be described in sufficient detail for their relationship to be evaluated. The analyses should make explicit those aspects of the target domain that the test represents, as well as those aspects that the test fails to represent.

The documents "castblueprint," "CAST_2019 OP_Gr5_Non accessible BB_for CDE_v11," "CAST_2019 OP_Gr8_Non accessible BB_for CDE_v9," "CAST_2019 OP_HS_Non accessible BB_for CDE_v9," "Updated DOK CAST_2019 OP," and "239-2016 FOR ARCHIVE – CA NGSS GEN Assessment Design_022416: Plan for Sampling CA NGSS Content" provide information about the target domains represented by the assessment. The document " 060118-02-v3_FOR ARCHIVE_CAST ECD White Paper"

includes a description of the segment-based design of CAST forms, noting while individual forms may not cover the breadth of the CA NGSS, coverage (at the school/corporation level, not the student level) will be achieved over the course of multiple administrations. The multiple "IDP"-named files document progress toward developing an item pool that adequately reflects the breadth and the depth of the NGSS.

Summary and Discussion

All of the eleven standards were rated as fully covered based on the available evidence. These results indicate that the CAST test design and development processes and procedures closely adhere to the testing standards related to alignment of assessment content to academic standards. Chapter 3 of this report describes the alignment workshop convened to document the extent to which test forms are adequately aligned to the CA NGSS.

Chapter 3: CAST Alignment Workshop and Outcomes

Introduction

This alignment study provides evidence of whether the CAST is designed to produce test forms that effectively measure the content and cognitive rigor reflected in the targeted content domain and the test blueprints. It does so by evaluating how well the 2019 test items fully sample the construct represented by the associated content standards, the California Next Generation Science Standards (CA NGSS). The first section of this chapter presents the alignment criteria HumRRO used for the evaluation. The next sections describe the methods HumRRO used to complete the second major task for the study: collection and analysis of item-level ratings from content experts on the alignment of CAST items to the CA NGSS. The chapter describes the recruitment and demographics of the panels of content experts and the workshop data collection procedures. The chapter concludes with the results of HumRRO's analysis of panelists' ratings. For each grade level, results are organized by the five major alignment criteria.

CAST Alignment Criteria

Alignment criteria were developed by HumRRO and reviewed by staff from the National Center for Improvement in Educational Assessment (Center for Assessment). The reviewers were highly experienced in both alignment methodologies and the CA NGSS. Reviewers made several comments that helped to clarify how the criteria would be communicated and operationalized for the study. The criteria were presented to California's CAASPP Technical Advisory Group (TAG) and finalized prior to the alignment workshop.

It is important to remember that no assessment is perfectly aligned. These criteria were developed based on the documentation provided by CDE and ETS, and they represent several aspects of the overall alignment of the CAST to the CA NGSS. Failure to meet any single criterion does not indicate that the test is invalid or flawed in some way, only that that aspect of the assessment may need to be addressed through future item development or by other means. An alignment study should be formative in nature and provide the state and the testing company with actionable results to make the assessment more closely mirror the CA NGSS.

The Webb alignment method (1997, 1999, 2002) was originally designed to align content standards with large-scale assessments. Dr. Norman Webb researched and refined this method over time. His approach is often cited and has been reviewed by the Council of Chief State School Officers (CCSSO).[1] The Webb method includes four major indicators to evaluate alignment. These indicators rely on statistical analyses to assess how well items on the assessment, regardless of item type and point value, match the state's content standards. The four alignment indicators are categorical

[1] For background information on alignment, see https://ccsso.org/sites/default/files/2018-07/TILSA%20Evaluating%20Alignment%20in%20Large-Scale%20Standards-Based%20Assessment%20Systems.pdf.

concurrence, depth-of-knowledge consistency, range-of-knowledge correspondence, and balance-of-knowledge representation. While it was not appropriate to implement Webb's methodology for this study, mainly because of the multidimensionality of the content standards and the way the content is sampled across years, we did use Webb's criteria to help guide our methodology and the development of criteria for judging the alignment of the CAST. Below, we briefly describe Webb's criteria, and then describe the similar criteria developed for use with CAST.

Webb's ***Categorical Concurrence*** is a basic measure of alignment between content standards and test items. This term refers to the proportion of overlap between the content stated in the standards document and that assessed by items on the test. Webb's criterion is based on the minimum number of items required to achieve acceptable reliability for reporting. HumRRO prefers to directly examine the reliability of the science assessments, which will be available in the forthcoming technical report[2] for the CAST. Reliability of scores should be evaluated for overall science scores and sub-scores at the student level and for any aggregate scores or sub-scores computed for schools, districts, or the state.

Webb's categorical concurrence criterion is derived by determining if there are at least six items per reporting category on the assessment. California will produce an overall student score and sub-scores at the domain level (e.g., life, physical, and earth and space sciences). So, at the most basic level, California could meet Webb's criteria if at least six items per domain were included on the assessments. This would not be a robust criterion for determining the sufficiency of items for generating reliable student scores.

The CA NGSS are written as performance expectations (PEs) through which students can demonstrate understanding of the content. These PEs were developed based on the DCI, SEP, and CCC the students are expected to have learned at each grade level. The PEs incorporate DCI, SEP and CCC. Test items might directly address the PE, or they might address the supporting DCI, SEP, or CCC. Ideally, an item would be linked to both a PE and some number of DCI, SEP, or CCC, but that may not always be possible given the relatively discrete nature of selected-response test items. It may be necessary to address all aspects of a standard through multiple test items.

For this criterion, the results section of this chapter reports the proportion of items that panelists matched to the PEs for science. The proportions also indicate the number of items not judged to relate to any PE. To be judged acceptable, at least 50 percent of the test items must be directly matched to a PE. We use 50 percent match to PE as one component of this criterion because some items are expected to be matched only to DCI, SEP, or CCC. Ideally, all items would match at least one PE, DCI, SEP, or CCC. However, it is possible for an assessment to have acceptable alignment with one or two weak items (as judged by panelists). To be judged acceptable for the second component of this criterion, at least 90 percent of items must be matched to either a PE, DCI, SEP, or CCC. To be judged acceptable, the test form must meet both components. We will refer to this criterion as **Link to Standards**.

[2] The technical report will be authored by ETS.

Webb's ***Depth-of-Knowledge*** (DOK) ***Consistency*** statistic measures the type of cognitive processing required by items compared to the cognitive processing required by the matched content standards. For example, is a student expected to simply identify or recall basic facts, to use reason to manipulate information, or to strategize how to best solve a complex problem? In another instance, a student may be asked to identify the planets of our solar system among several answer choices. This task would be rated less complex (have a lower DOK) than comparing the composition of the planets in preparation for landing unmanned probes.

The purpose of using DOK as a measure of alignment is to determine whether a test item and its corresponding standard are written at the same level of cognitive complexity. In Webb's method, panelists make two separate judgments about cognitive complexity, one rating for the standard and one rating for the item. These two judgments are compared to determine whether the item is written at the same level as the standard to which it is linked. Webb (1997) refers to this comparison as *Depth-of-Knowledge consistency*.

Webb's DOK consistency category is nearly impossible to implement when the standards are multidimensional. Doing so would require panelists to determine the DOK for each potential combination of standard and dimension. For science, it is also the case that the test standards can be interpreted in multiple ways and each combination of standard and dimension would represent a range of cognitive complexities depending on the specific knowledge, skills, and abilities that were being addressed. So, even if we could generate the number of DOK ratings required by the science standards, our ratings would likely be vague, unreliable, and inflated (Webb's rule is to assign the higher DOK level if the standard is ambiguous). Therefore, no attempt will be made to match item DOK with standard DOK for this study.

It is still, however, important to determine if CAST test items reflect the level of cognitive complexity indicated by the CA NGSS. Looking at the standards more globally, HumRRO found they focus on requiring students to use their science knowledge and skills to investigate potentially unfamiliar phenomena. Focusing on science in this way means that students are expected to engage in more complex reasoning than simply recalling science terms or generating simple answers using familiar algorithms. HumRRO therefore reasoned that California's science assessment should include few, if any, low-complexity items. Webb uses a four-point scale for DOK, with level one being low. For an assessment based on the CA NGSS, HumRRO would expect no more than 10 percent of items to be rated at level one. HumRRO selected the 10 percent threshold to reflect the CA NGSS focus on complex multidimensional science content. Level one items are expected to make up a minimal part of the assessment, if they are included at all. Webb's scale also includes a level four rating, which is seldom used for summative tests. This level of cognitive processing requires deep engagement of the students with the content, in multiple ways, typically over an extended period of time. This level is similar to producing a thesis or generating an extensive investigation of some scientific phenomenon a student would observe, collect data about, and generate a report to describe. HumRRO does not expect CAST to include level four items but does expect the assessments to be primarily a mix of DOK level two and three items. HumRRO also expects more level two items than level three items. Level three items require more input

or time for students to respond, and it would not be practical to include primarily level three items on a summative assessment. In other states, notably Colorado, science standards are presented with a DOK range included. The range indicates the level of items and the level of instruction that are expected when addressing the content of the standards. In most cases in Colorado standards, the DOK range for standards is 1–3, with the mode clearly at level 2.

For this criterion, the results section of this chapter reports the proportion of items panelists rated at each DOK level. Guidelines for an appropriate distribution of item DOK levels are not included in the CAST blueprints. As such, to be judged acceptable, no more than 10 percent of items may be rated at level 1 and no less than 10 percent of items must be rated at level 3. If there are more than 10 percent of items at level 1 or fewer than 10 percent of items at level 3, the DOK level of the items as a group would be judged too low to adequately represent the California science standards. We will refer to this criterion as **DOK Adequacy**.

Webb's ***Range-of-Knowledge Correspondence*** examines the extent to which the test items reflect the full range of knowledge, skills, and abilities contained in the standards document. Where categorical concurrence notes whether a sufficient number of items on the test covers each general content topic (reporting category), the range-of-knowledge correspondence measure indicates the number of specific content objectives within each broader topic that are assessed by the test items.

Webb's range-of-knowledge correspondence criterion requires that at least 50 percent of the standards from each reporting category are addressed on the assessment. As stated above, California intends to report students' overall science scores and domain level sub-scores, but not finer-grained sub-scores (e.g., physics, chemistry, ecology, cross-cutting concepts, science and engineering principles). Meeting Webb's range-of-knowledge criterion would thus require that at least half of the full range of standards for science be represented on the tests. Given the three-dimensional nature of the standards, this criterion is not practical. The number of potential combinations of domains and dimensions represent too many standards to address in any single testing event. Even assessing at the PE level, if one were to address every PE on a single assessment, the number of required items would be impractical. HumRRO believes it is necessary, therefore, to sample the standards for assessing students. The standards emphasize students making meaning from information gathered from new or unfamiliar phenomena. They are expected to have a deep understanding of SEPs and CCCs, and that knowledge is expected to provide tools to use across DCIs in all content domains. We will focus on SEPs and CCCs for this criterion rather than on trying to address the full breadth of the science standards.

Because students are expected to use their knowledge of SEPs and CCCs across multiple standards and content domains, we would expect these dimensions to be high priorities on California's science assessments. HumRRO also expects there to be few, if any, items on the tests that measure only an SEP or a CCC, and that these concepts are measured in context with DCIs from legitimate scientific phenomena. Items are coded to indicate if they measure an SEP or CCC, or both.

For this criterion, the results section of this chapter reports the proportion of items that panelists rate as measuring SEPs and CCCs. To be judged acceptable, at least 50 percent of the eight SEPs and seven CCCs must be directly measured by items on the tests. Hence, the assessments should contain items that address at least four SEPs and four CCCs to meet this criterion. We will refer to this criterion as **Range Adequacy**.

Webb's ***Balance-of-Knowledge Representation*** focuses on content coverage in yet more detail. In this case, the number of items matched to the content objective does matter. The balance of representation criterion determines whether the assessment measures the content objectives equitably within each content topic using only those content objectives identified by panelists as measured by the test item. Based on Webb's (1997) method, items should be distributed evenly across the objectives per content topic for good balance. The balance-of-knowledge representation is determined by calculating an index, or score, for each content topic. Each topic should meet or surpass a minimum index level to demonstrate adequate balance.

It would not be possible to compute a single interpretable balance-of-knowledge representation index for a three-dimensional assessment. The interaction of the dimensions and domains would yield too many objectives (and too many PEs) to include on a summative test form. It does, however, make sense to consider that each content domain should be represented rather evenly, or purposefully, on an assessment. It might also be sensible to declare that the three dimensions should be represented rather evenly, or purposefully, on an assessment. Acceptability for the CAST test will be determined using the same metric as Webb uses for balance-of-knowledge correspondence with the notable exception that it will be computed twice: once for domain, and again for dimension.

For this criterion, the results section of this chapter reports the number and proportion of items panelists matched to each domain and dimension. HumRRO uses the same index for acceptability as Webb uses for traditional assessments (0.70; the balance index formula can be found in http://wat.wceruw.org/Training%20Manual%202.1%20Draft%20091205.doc. For CAST, however, balance criteria must be met by domain and by dimension for the assessment to be considered adequately aligned. HumRRO will refer to this criterion as **Balance-of-Knowledge Correspondence (Revised for Science),** or simply as **Balance**.

Finally, the CAST items are written to be multidimensional. They are intended to measure more than isolated science content knowledge and are expected to address CCC and SEP in addition to DCI and/or specific PE. For this criterion, the results section of this chapter reports the proportion of items panelist rated as related to multiple science concepts across DCI, CCC, and SEP. To be judged acceptable, at least 90 percent of items must address more than one dimension. HumRRO will refer to this criterion as **Multidimensional Adequacy**.

Table 3.1 summarizes the criteria used to evaluate the alignment of CAST items to the CA NGSS. Failure to meet a single criterion would not indicate that the test is insufficiently aligned to generate meaningful scores, but that attention to that aspect of the test should be addressed through future item development. If several of the criteria

were not met, it would signal that HumRRO should be concerned with the link between the assessment and the intended measurement construct.

Table 3.1 CAST-to-CA NGSS Alignment Criteria

Criteria	Description
Link to Standards	The percentage of items that panelists rate as directly and clearly matched to a PE, DCI, SEP, and/or CCC is calculated. The criterion is met if 50 percent or more of the items are matched to a specific PE and at least 90 percent of items are matched to at least one PE, DCI, SEP, or CCC.
DOK Adequacy	The percentage of items rated by panelists as reflecting each of Webb's DOK levels (Recall, Skill/Concept, Strategic Thinking) is calculated. The criterion is met if fewer than 10 percent of items are rated as DOK level 1 (Recall) and more than 10 percent of items are rated at DOK level 3 (Strategic Thinking).
Range Adequacy	The percentage of SEPs and/or CCCs that panelists rate as directly and clearly matched to one or more items is calculated. The criterion is met if at least 50 percent of CCCs and 50 percent of SEPs are aligned to test items (at least 4 CCCs and 4 SEPs).
Balance-of-Knowledge Correspondence (Revised for Science)	The number of items that panelists rate as directly and clearly matched to a content domain (e.g., Life Sciences), SEP, and/or CCC is calculated. Webb's balance-of-knowledge correspondence index is computed separately for each of these CA NGSS dimensions based on the total number of items that were matched to any content domain, SEP, and/or CCC and the proportion of those items that were matched to each specific content domain, SEP, and CCC. The criterion is met if the calculated balance index is 70 percent or higher for domains and dimensions.
Multidimensional Adequacy	The percentage of items that panelists rate as directly and clearly matched to at least one DCI, SEP, and/or CCC is calculated. The criterion is met if at least 90 percent of items are aligned to more than one dimension.

Methods

The evaluation of the alignment criteria is based on item ratings and professional judgments collected during an alignment workshop. This section describes the workshop participants (henceforth referred to as "alignment panelists" or "panelists"), workshop materials, training, and workshop processes and procedures.

Alignment Panelists

HumRRO worked collaboratively with the CDE to recruit and select a group of 18 educators to serve on three CAST alignment review panels (grade five, grade eight, and high school). Across the three panels, 14 California school districts were represented.

Approximately 50 percent of panelists reported being a current teacher (including lead teacher), and the remaining 50 percent reported working in roles such as coordinator, specialist, program director, or superintendent. In addition to their current professional roles, all panelists reported having some level of experience with the CA NGSS. The types of experience reported ranged from teaching the CA NGSS to providing CA NGSS-related training to other educators. Across the three panel groups, all panelists who provided responses reported experience teaching students from diverse socioeconomic and cultural backgrounds as well as experience teaching English learners. Table 3.2 summarizes the demographics of the alignment panelists.

Table 3.2 CAST Alignment Panelists' Demographics

Panel	# of Panelists	# of Districts	% Female/ %Male	% Hispanic/ %Non-Hispanic	Years of Experience Mean (SD)
Grade 5	6	5	67%/33%	17%/83%	20.17 (6.91)
Grade 8	6	4	100%/0%	17%/83%	15.67 (8.85)
High School	6	5	50%/50%	17%/83%	16.10 (9.36)

Workshop Logistics

HumRRO conducted a two-day CAST Alignment Study Workshop in the Sacramento area on February 28–March 1, 2019. During the workshop, panels of educators evaluated how well each CAST item assessed the CA NGSS. Prior to entering the workshop, panelists were required to sign nondisclosure agreements as a condition of participation.

Workshop Materials

CDE and ETS provided HumRRO with documents and data to facilitate the development of materials for the alignment workshop. These included test design documentation (e.g., item specifications, test blueprints) and information about the California approach to classifying item cognitive complexity (depth of knowledge [DOK]

using Webb's four levels) for the operational 2019 CAST items. ETS created three online test forms for the alignment workshop (grade 5, 8 and high school) consisting of all the operational 2019 CAST items. ETS also created accounts for HumRRO researchers to securely access the items using the CAASPP Interim Assessment Viewing System (IAVS).

HumRRO developed several data collection tools and adapted other materials to support the data collection process. Data collection tools included electronic spreadsheets for panelists and workshop facilitators to enter test item ratings. Support materials included copies of the CA NGSS and appendices (both paper and electronic), copies of the CAST item specifications, detailed workshop instructions for both panelists and facilitators, details on the cognitive complexity (DOK) rating categories, and debriefing and evaluation forms. Example workshop materials are presented in Appendix B.

Training

Alignment panelists received two rounds of training at the outset of the alignment workshop. First, the full group of panelists received general training that provided some background on alignment and a high-level description of the alignment process. Following the general training session, panelists moved into grade-level panel groups and received more detailed training on the data collection (rating) processes and procedures. Those processes and procedures are described in more detail in the following section.

Workshop Processes and Procedures

During the workshop, each panelist had a workstation with two laptops and a three-ring binder containing grade level alignment materials (CA NGSS and associated appendices and CAST item specs). Electronic versions of the materials in the binder were also saved to a folder on panelists' laptops (to facilitate electronic searching and to ensure accessibility of the materials for all panelists). Panelists accessed operational test items via the online secure platform set up by ETS. Electronic rating forms were saved onto panelists' laptops. Panelists also received paper copies of the training presentation, detailed alignment process steps, item specifications, and descriptions of the cognitive complexity (DOK) categories.

After the panel-specific training presentation by the HumRRO facilitator, each panel engaged in a calibration activity using the first few (1–3) items. Panelists accessed the items electronically and made their independent ratings. Panelists discussed their independent ratings and engaged in consensus discussion to come to agreement on the final item ratings of record. Once panelists had a clear understanding of the rating process and a common understanding of the rating categories, they moved on to rating the remaining operational items.

The panelists rated a small group of operational items at a time. For each group of items, panelists first made their ratings for each item independently. Then panelists discussed their ratings for an item, looked at the item metadata, engaged in more

discussion, then reached their final consensus/majority rating for that item before moving on to the next. When consensus could not be reached, the facilitator recorded the majority rating. Once consensus/majority ratings were recorded for that group of items, the panel moved on to the next group and repeated this process.

In summary, item ratings were generated via the following steps:

1. Panelists reviewed test items independently and assigned ratings of:
    a. PE measured by item
    b. DCI measured by item (up to two DCI, primary and secondary)
    c. CCC measured by item (up to two CCC, primary and secondary)
    d. SEP measured by item (up to two SEP, primary and secondary)
    e. Item Depth of Knowledge (DOK)
    f. Comments to clarify ratings or to provide feedback on quality of item or associated phenomenon
2. Panelists discussed their independent ratings.
3. HumRRO facilitator shared item metadata.
4. Panelists came to consensus (or majority) ratings.
5. HumRRO facilitator recorded consensus/majority ratings.

Once all panelists had completed their independent ratings for a group of items, the HumRRO facilitator managed the group discussion and encouraged all panelists to share their ratings. Typically, the facilitator polled the group about each rating, and asked for panelists to provide a rationale when independent ratings differed among them. Panelists were trained to retain their independent ratings unless they realized that they had made a coding error, or if group discussion revealed to them an error in their thinking about an item and/or the CA NGSS. After the initial discussion, the HumRRO facilitator projected the ETS item metadata from CAST item development for the group to review and discuss and reach consensus on the final ratings for each item. If the group could not reach true consensus, the facilitator recorded the rating of the majority of panelists.

The HumRRO facilitator recorded the final consensus (or majority) item ratings in a spreadsheet. Once all consensus statements were recorded, panelists completed a debriefing form and a process evaluation survey before being released from the workshop. The debriefing form was designed to give panelists the opportunity to provide their individual, qualitative perspective on the quality of alignment. The evaluation survey elicited feedback about the quality of the workshop processes and procedures. Table 3.3 summarizes the workshop quality results.

Table 3.3 CAST Alignment Evaluation Survey Results

Evaluative Statement	% Strongly Disagree	% Disagree	% Somewhat Disagree	% Somewhat Agree	% Agree	% Strongly Agree
The training presentation in the large group provided useful information about the CAST and HumRRO's alignment method.	0.0	0.0	5.9	5.9	47.1	41.2
After the additional training in my small group, I felt prepared to review and rate test items.	0.0	0.0	0.0	5.9	52.9	41.2
HumRRO staff seemed knowledgeable of the CAST and alignment steps.	0.0	0.0	0.0	0.0	17.6	82.4
The Panelist Instruction document was clear, understandable, and useful in performing the alignment steps.	0.0	0.0	0.0	5.9	35.3	58.8
The Excel file was understandable and relatively easy to use to enter item ratings.	0.0	0.0	0.0	5.9	41.2	52.9
The process for reaching consensus ratings was conducted fairly.	0.0	0.0	0.0	0.0	17.6	82.4

Results

This section summarizes the data and information collected during the alignment workshop. The results are presented for each grade level, the item pool, and by test form. Results are presented as percentages that have been rounded to the nearest whole number, for ease of interpretation. For each grade level, all operational 2019 CAST test items were evaluated on the five alignment criteria: (1) Link to Standards, (2) DOK Adequacy, (3) Range Adequacy, (4) Balance-of-Knowledge Correspondence (Revised for Science), and (5) Multidimensional Adequacy.

Grade Five

This section summarizes results for the grade five science assessment. For each alignment criterion, the first table presents results for the grade five item pool and the second table presents the results for the grade five test forms.

Criterion 1: Link to Standards

This criterion is evaluated based on the percentage of items that panelists rate as directly and clearly matched to a PE, DCI, SEP, and/or CCC. The criterion is considered Acceptable if 50 percent or more of the items in the item pool or on a test form are matched to a specific PE and at least 90 percent of items are matched to at least one PE, DCI, SEP, or CCC.

Table 3.4 shows that 97 percent of the 60 grade five CAST items were matched to a specific PE by panelists, and that 98 percent of items were matched to at least one PE, DCI, SEP, or CCC. Based on this, Criterion 1, Link to Standards, is met for the grade five items.

Table 3.4 Grade Five Item Pool Results for Criterion 1: Link to Standards (n items= 60)

Sub-criterion	Percentage	Acceptable?
Items matched to a specific PE	97	Yes
Items matched to at least one PE, DCI, SEP, or CCC	98	Yes

Table 3.5 presents the results from a by-form analysis of the same ratings. Across forms, 95–98 percent of items on the form were rated as measuring a specific PE, and 98 percent of items were matched to at least one PE, DCI, SEP, or CCC. Criterion 1 is met for all of the grade five test forms.

Table 3.5 Grade Five Test Form Results for Criterion 1: Link to Standards

Sub-criterion	Range of Percentages	Number of Forms Meeting Criterion
Items matched to a specific PE	95–98	10 of 10
Items matched to at least one PE, DCI, SEP, or CCC	98–98	10 of 10

Criterion 2: DOK Adequacy

This criterion is evaluated based on the percentage of items rated by panelists as reflecting each of Webb's DOK levels (Recall, Skill/Concept, Strategic Thinking). The criterion is considered Acceptable if fewer than 10 percent of items in the item pool or on a test form are rated as DOK level 1 (Recall) and more than 10 percent of items are rated at DOK level 3 (Strategic Thinking).

Table 3.6 shows that less than 10 percent of grade five CAST items were rated at DOK Level 1 and more than 10 percent of grade five CAST items were rated at DOK Level 3. Criterion 2, DOK Adequacy, is met for the grade five items.

Table 3.6 Grade Five Item Pool Results for Criterion 2: DOK Adequacy (n items= 60)

DOK level	Percentage	Acceptable?
Level 1 - Recall	2	Yes
Level 2 - Skill/Concept	55	Yes
Level 3 - Strategic Thinking	43	Yes

Table 3.7 presents the results from a by-form analysis of the same ratings. Across the 10 grade five science forms, less than 10 percent of items were rated DOK Level 1 and more than 10 percent of items were rated at DOK Level 3. Criterion 2, DOK Adequacy, is met for all the grade five test forms.

Table 3.7 Grade Five Test Form Results for Criterion 2: DOK Adequacy

DOK level	Range of Percentages	Number of Forms Meeting Criterion
Level 1 – Recall	2–2	10 of 10
Level 2 - Skill/Concept	43–59	10 of 10
Level 3 - Strategic Thinking	39–55	10 of 10

Criterion 3: Range Adequacy

This criterion is evaluated based on the percentage of SEPs and/or CCCs that panelists rate as directly and clearly matched to one or more items. The criterion is considered Acceptable if at least 50 percent of CCCs and 50 percent of SEPs are aligned to test items (at least 4 CCCs and 4 SEPs).

Table 3.8 shows that more than half of the SEPs and CCCs were represented by the pool of grade five items. The CCC that was not matched to any item was *Stability and change*. Criterion 3, Range Adequacy, is met for the grade five items.

Table 3.8 Grade Five Item Pool Results for Criterion 3: Range Adequacy

CA NGSS Dimension	Percentage	Acceptable?
SEP (n=8)	100	Yes
CCC (n=7)	86	Yes

Table 3.9 presents the results from a by-form analysis of the same ratings. Across the forms, more than half of the SEPs and CCCs were matched to items. Criterion 3, Range Adequacy, is met for all of the grade five test forms.

Table 3.9 Grade Five Test Form Results for Criterion 3: Range Adequacy

CA NGSS Dimension	Range of Percentages	Number of Forms Meeting Criterion
SEP (n=8)	88–100	10 of 10
CCC (n=7)	86–86	10 of 10

Criterion 4: Balance-of-Knowledge Correspondence (Revised for Science)

This criterion is evaluated based on number of items that panelists rate as directly and clearly matched to a content domain (e.g., Life Sciences), SEP, and/or CCC. HumRRO computed Webb's balance-of-knowledge correspondence index separately for each of these CA NGSS dimensions based on the total number of items that were matched to any content domain, SEP, and/or CCC and the proportion of those items that were matched to each specific content domain, SEP, and CCC. The criterion is considered Acceptable if the calculated balance index is 70 or higher.

Table 3.10 shows that the balance indexes were 81, 77, and 78 for Content Domain, SEP, and CCC, respectively. This indicates an acceptable level of balance among the science domains, practices, and concepts that were matched to items. Criterion 4, Balance-of-Knowledge Correspondence, is met for the grade five items.

Table 3.10 Grade Five Item Pool Results for Criterion 4: Balance-of-Knowledge Correspondence

CA NGSS Dimension	Balance Index	Acceptable?
Content Domain	81	Yes
SEP	77	Yes
CCC	78	Yes

Table 3.11 presents the results from a by-form analysis of the same ratings. For all 10 grade five science test forms, there is an acceptable level of balance among the content domains and SEP that were matched to items. For all but four (4) test forms, there was an acceptable level of balance among the CCCs that were matched to items. For those forms that did not demonstrate acceptable balance, there was a large proportion of items matched to *Cause and effect: mechanism and explanation*. Criterion 4 Balance-of-Knowledge Correspondence is met for the majority of grade five test forms.

Table 3.11 Grade Five Item Pool Results for Criterion 4: Balance-of-Knowledge Correspondence

CA NGSS Dimension	Range of Balance Indexes	Number of Forms Meeting Criterion
Content Domain	81–91	10 of 10
SEP	78–87	10 of 10
CCC	67–76	6 of 10

Comparing panelist's ratings to the item metadata provides some useful context for these results. Generally, panelists ratings tended to agree with the metadata regarding item alignment to the *Cause and effect: mechanism and explanation* CCC. For five of the 60 items, panelists rated the items as aligned to the Cause and effect: mechanism and explanation CCC, whereas the metadata aligned them to the *Patterns* CCC (1 item), the *Scale, proportions, and quantity* CCC (2 items) or no CCC (2 items). The forms for which the minimum balance index was not met all contained the same B segment (B3). This segment contains two items aligned to the *Scale, proportions, and quantity* CCC, which panelists thought were better aligned to the *Cause and effect: mechanism and explanation* CCC.

Criterion 5: Multidimensional Adequacy

This criterion is evaluated based on the percentage of items that panelists rate as directly and clearly matched to at least one DCI, SEP, and/or CCC. The criterion is considered Acceptable if at least 90 percent of items in the item pool or on a test form are aligned to more than one dimension.

Table 3.12 show that 93 percent of the 60 grade five CAST items were rated as measuring at least two CA NGSS dimensions. Seventy-two percent (72%) of the grade 5 items were rated as measuring a DCI, SEP, and CCC. Criterion 5, Multidimensional Adequacy, is met for the grade five items.

Table 3.12 Grade Five Item Pool Results for Criterion 5: Multidimensional Adequacy (n items= 60)

Criterion	Percentage	Acceptable?
Items are aligned to more than one dimension	93	Yes

Table 3.13 presents the results from a by-form analysis of the same ratings. For all 10 forms, over 90 percent of items were rated as measuring two or more CA NGSS dimensions. There was a single item, which appeared on all test forms, that panelists found to be not aligned to any part of the standards. Three items were rated as measuring only one dimension, and all forms contained either 2 or 3 of these items. Criterion 5, Multidimensional Adequacy, is met for all grade five test forms.

Table 3.13 Grade Five Test Form Results for Criterion 5: Multidimensional Adequacy

Criterion	Range of Percentages	Number of Forms Meeting Criterion
Items are aligned to more than one dimension	91–93	10 of 10

Grade Eight

This section summarizes results for the grade eight science assessment. For each alignment criterion, the first table presents results for the grade eight item pool and the second table presents the results for the grade eight test forms.

Criterion 1: Link to Standards

Table 3.14 shows that 98 percent of the 62 grade eight CAST items were matched to a specific PE by panelists, and 98 percent of items were matched to at least one PE, DCI, SEP, or CCC. Based on this, Criterion 1, Link to Standards, is met for the grade eight items.

Table 3.14 Grade Eight Item Pool Results for Criterion 1: Link to Standards (n items= 62)

Sub-criterion	Percentage	Acceptable?
Items matched to a specific PE	98	Yes
Items matched to at least one PE, DCI, SEP, or CCC	98	Yes

Table 3.15 presents the results from a by-form analysis of the same ratings. Across forms, 91–98 percent of items on the form were rated as measuring a specific PE, and were matched to at least one PE, DCI, SEP, or CCC. Criterion 1, Link to Standards, is met for all of the grade eight test forms.

Table 3.15 Grade Eight Test Form Results for Criterion 1: Link to Standards

Sub-criterion	Range of Percentages	Number of Forms Meeting Criterion
Items matched to a specific PE	91–98	15 of 15
Items matched to at least one PE, DCI, SEP, or CCC	91–98	15 of 15

Criterion 2: DOK Adequacy

Table 3.16 shows that just slightly over 10 percent (11%) of grade eight CAST items were rated at DOK Level 1 and more than 10 percent of grade eight CAST items were rated at DOK Level 3. Criterion 2, DOK Adequacy, is partially met for the grade eight items.

Table 3.16 Grade Eight Item Pool Results for Criterion 2: DOK Adequacy (n items= 62)

DOK level	Percentage	Acceptable?
Level 1 – Recall	11	No
Level 2 - Skill/Concept	54	Yes
Level 3 - Strategic Thinking	34	Yes

Note. One item did not receive a DOK rating because panelists felt there was not enough information in the item stem to get a correct answer.

Table 3.17 presents the results from a by-form analysis of the same ratings. All grade eight science forms had 12–18 percent of items rated DOK Level 1, slightly above the 10 percent threshold. All grade eight test forms had more than 10 percent of items rated at DOK Level 3. Criterion 2, DOK Adequacy, is partially met for the grade eight test forms.

Table 3.17 Grade Eight Test Form Results for Criterion 2: DOK Adequacy

DOK level	Range of Percentages	Number of Forms Meeting Criterion
Level 1 - Recall	12–18	0 of 15
Level 2 - Skill/Concept	46–61	15 of 15
Level 3 - Strategic Thinking	21–40	15 of 15

Note. One item did not receive a DOK rating because panelists felt there was not enough information in the item stem to determine a correct answer.

Comparing panelists' ratings to the item metadata provides some useful context for these results. Panelists rated five segment A items (15.6%) at Level 1 DOK, compared to one segment A item (3%) identified as Level 1 DOK in the item metadata. All items that panelists rated at Level 1 DOK were either identified as Level 2 DOK in the metadata or did not have a DOK rating stored in the metadata. No items from segments B1–B4 were rated as Level 1 DOK by panelists. One item in segment B5 (16%) and one item in segment B6 (25%) were rated at Level 1 DOK by panelists. The segment B5 item did not have a DOK level stored in the metadata, and the segment B6 item was identified as Level 2 DOK in the metadata. The level of agreement between the panelists' ratings and the item metadata will be further discussed in the *Discussion* section of this report.

Criterion 3: Range Adequacy

Table 3.18 shows that 88 percent of SEPs and all CCCs were represented by the pool of grade eight items. The SEPs that are not matched to any item are *Asking questions (for science)* and *Designing solutions (for engineering)*. Criterion 3, Range Adequacy, is met for the grade eight items.

Table 3.18 Grade Eight Item Pool Results for Criterion 3: Range Adequacy

CA NGSS Dimension	Percentage	Acceptable?
SEP (n=8)	88	Yes
CCC (n=7)	100	Yes

Table 3.19 presents the results from an analysis of the same ratings by test form. Across the forms, more than half of SEP and CCC were matched to items. Criterion 3, Range Adequacy, is met for all the grade eight test forms.

Table 3.19 Grade Eight Test Form Results for Criterion 3: Range Adequacy

CA NGSS Dimension	Range of Percentages	Number of Forms Meeting Criterion
SEP (n=8)	88–88	15 of 15
CCC (n=7)	100–100	15 of 15

Criterion 4: Balance-of-Knowledge Correspondence (Revised for Science)

Table 3.20 shows that the balance indexes were 85, 64, and 78 for Content Domain, SEP, and CCC, respectively. This is acceptable balance among the science domains and concepts, but less than an acceptable level of balance among the practices that

were matched to items. Criterion 4, Balance-of-Knowledge Correspondence, is partially met for the grade eight items.

Table 3.20 Grade Eight Item Pool Results for Criterion 4: Balance-of-Knowledge Correspondence

CA NGSS Dimension	Balance Index	Acceptable?
Content Domain	85	Yes
SEP	64	No
CCC	78	Yes

Table 3.21 presents results from a by-form analysis of the same ratings. For all 15 grade eight science test forms, HumRRO found an acceptable level of balance among the science domains. For 14 of the 15 forms we found an acceptable level of balance among the SEPs and the CCCs that were matched to items. For those forms that did not demonstrate acceptable balance, we found a large proportion of items matched to *Cause and effect: mechanism and explanation*. Criterion 4, Balance-of-Knowledge Correspondence, is met for the majority of grade eight test forms.

Table 3.21 Grade Eight Test Form Results for Criterion 4: Balance-of-Knowledge Correspondence

CA NGSS Dimension	Range of Balance Indexes	Number of Forms Meeting Criterion
Content Domain	77–84	15 of 15
SEP	68–74	14 of 15
CCC	69–86	14 of 15

Table 3.21 demonstrates that although the item pool is not balanced across the SEPs and CCCs, there were sufficient items for creating balanced forms, with a small number of exceptions. The test form with segments B1 and B3 had a relatively large number of alignments to the *Constructing explanations* SEP, but this was reflected in both the panelists ratings as well as the item metadata. The test form with segments B1 and B2 had a relatively large number of alignments to the *Cause and effect* CCC, and similarly this was exhibited in both ratings and metadata. For both forms, the balance index was very close to the minimum threshold of 70.

Criterion 5: Multidimensional Adequacy

Table 3.22 shows that 98 percent of the 62 grade eight CAST items were rated as measuring at least two CA NGSS dimensions. Approximately 97 percent of the grade eight items were rated as measuring a DCI, SEP, and CCC. Criterion 5, Multidimensional Adequacy, is met for the grade eight items.

Table 3.22 Grade Eight Item Pool Results for Criterion 5: Multidimensional Adequacy (n items= 62)

Criterion	Percentage	Acceptable?
Items are aligned to more than one dimension	98	Yes

Table 3.23 presents the results from a by-form analysis of the same ratings. For all 15 forms, 91–98 percent of items were rated as measuring two or more CA NGSS dimensions. Criterion 5, Multidimensional Adequacy, is met for all grade eight test forms.

Table 3.23 Grade Eight Test Form Results for Criterion 5: Multidimensional Adequacy

CA NGSS Dimension	Range of Percentages	Number of Forms Meeting Criterion
Items are aligned to more than one dimension	91–98	15 of 15

High School

This section summarizes results for the high school assessment. For each alignment criterion, the first table presents results for the high school item pool and the second table presents the results for the high school test forms

Criterion 1: Link to Standards

Table 3.24 shows that 98 percent of the 50 high school CAST items were matched to a specific PE by panelists, and that 98 percent of items were matched to at least one PE, DCI, SEP, or CCC. Based on these findings, Criterion 1, Link to Standards, is met for the high school items.

Table 3.24 High School Item Pool Results for Criterion 1: Link to Standards (n items= 50)

Sub-criterion	Percentage	Acceptable?
Items matched to a specific PE	98	Yes
Items matched to at least one PE, DCI, SEP, or CCC	98	Yes

Table 3.25 presents the results of the same ratings by form. Across forms, 98–100 percent of items on the forms were rated as measuring a specific PE, and 98–100 percent of items were matched to at least one PE, DCI, SEP, or CCC. Criterion 1, Link to Standards, is met for all of the high school test forms

Table 3.25 High School Test Form Results for Criterion 1: Link to Standards

Sub-criterion	Range of Percentages	Number of Forms Meeting Criterion
Items matched to a specific PE	98–100	3 of 3
Items matched to at least one PE, DCI, SEP, or CCC	98–100	3 of 3

Criterion 2: DOK Adequacy

Table 3.26 shows that more than 10 percent (16%) of high school CAST items were rated at DOK Level 1 and more than 10 percent of high school CAST items were rated at DOK Level 3. Criterion 2, DOK Adequacy, is partially met for the high school items.

Table 3.26 High School Item Pool Results for Criterion 2: DOK Adequacy (n items= 50)

DOK level	Percentage	Acceptable?
Level 1- Recall	16	No
Level 2- Skill/Concept	50	Yes
Level 3- Strategic Thinking	34	Yes

Table 3.27 presents the results of the same ratings by test. None of the three high school science forms had less than 10 percent of items rated as DOK Level 1. All high school test forms had more than 10 percent of items rated at DOK Level 3. Criterion 2, DOK Adequacy, is partially met for the three high school test forms.

Table 3.27 High School Test Form Results for Criterion 2: DOK Adequacy

DOK level	Range of Percentages	Number of Forms Meeting Criterion
Level 1- Recall	12–18	0 of 3
Level 2- Skill/Concept	49–52	3 of 3
Level 3- Strategic Thinking	33–38	3 of 3

Comparing panelists' ratings to the item metadata provides some useful context for these results. Panelists rated four segment A items (11.8%) at Level 1 DOK, compared to one segment A item (2.9%) at Level 1 DOK in the item metadata. All segment A items that panelists rated at Level 1 DOK were identified as Level 2 DOK in the metadata. Four items (25%) from blocks B1–B3 were rated as Level 1 DOK by panelists, compared to two items (12.5%) identified as Level 1 DOK in the metadata. Three of these four items rated at Level 1 DOK are from block B3. One of these three block B3 items was also identified as Level 1 DOK in the metadata, whereas the other

two were identified as Level 2 DOK in the metadata. The level of agreement between the panelists' ratings and the item metadata will be presented in the *Discussion* section of this chapter.

Criterion 3: Range Adequacy

Table 3.28 shows that all SEPs and all CCCs were represented by the pool of high school items. Criterion 3, Range Adequacy, is met for the high school items.

Table 3.28 High School Item Pool Results for Criterion 3: Range Adequacy

CA NGSS Dimension	Percentage	Acceptable?
SEP (n=8)	100	Yes
CCC (n=7)	100	Yes

Table 3.29 presents the by-form analysis of the same ratings. Across the forms, more than half of the SEPs and CCCs were matched to items. Criterion 3, Range Adequacy, is met for all of the high school test forms.

Table 3.29 High School Test Form Results for Criterion 3: Range Adequacy

CA NGSS Dimension	Range of Percentages	Number of Forms Meeting Criterion
SEP (n=8)	100–100	3 of 3
CCC (n=7)	100–100	3 of 3

Criterion 4: Balance-of-Knowledge Correspondence (Revised for Science)

Table 3.30 shows that the balance indexes were 83, 76, and 76 for Domain, SEP and CCC, respectively. This indicates an acceptable level of balance among the science domains, practices, and concepts that were matched to items. Criterion 4, Balance-of-Knowledge Correspondence, is met for the high school items.

Table 3.30 High School Item Pool Results for Criterion 4: Balance-of-Knowledge Correspondence

CA NGSS Dimension	Balance Index	Acceptable?
Content Domain	83	Yes
SEP	76	Yes
CCC	76	Yes

Table 3.31 presents the results of an analysis by test form. For all three of the science test forms, there is an acceptable level of balance among the science domains, SEP, and CCC. Criterion 4, Balance-of-Knowledge Correspondence, is met for the three high school test forms.

Table 3.31 High School Test Form Results for Criterion 4: Balance-of-Knowledge Correspondence

CA NGSS Dimension	Range of Balance Indexes	Number of Forms Meeting Criterion
Content Domain	82–82	3 of 3
SEP	75–77	3 of 3
CCC	76–78	3 of 3

Criterion 5: Multidimensional Adequacy

Table 3.32 shows that 96 percent of the 50 high school CAST items were rated as measuring at least two CA NGSS dimensions. Eighty-four percent (84%) of the high school items were rated as measuring a DCI, SEP, and CCC. Criterion 5, Multidimensional Adequacy, is met for the high school items.

Table 3.32 High School Item Pool Results for Criterion 5: Multidimensional Adequacy (n items= 50)

Criterion	Percentage	Acceptable?
Items are aligned to more than one dimension	96	Yes

Table 3.33 presents the results from an analysis of the same ratings by test form. For all three forms, at least 98 percent of items were rated as measuring two or more CA NGSS dimensions. Criterion 5, Multidimensional Adequacy, is met for all high school test forms.

Table 3.33 High School Test Form Results for Criterion 5: Multidimensional Adequacy

CA NGSS Dimension	Range of Percentages	Number of Forms Meeting Criterion
Items are aligned to more than one dimension	98–100	3 of 3

Summary and Discussion

Summary Results

Table 3.34 summarizes the alignment criteria results for the three summative assessment science test item pools. Across the three tests, panelists' ratings of the operational items provide strong support that the CAST is composed of multidimensional items that reflect a range of the CA NGSS. The ratings also support that the items generally reflect appropriate levels of cognitive complexity and a balance among the CA NGSS dimensions.

Table 3.34 Summary of Item Pool Results by Criterion and Grade Level

Criterion	Grade 5	Grade 8	High School
Links to Standards	Met	Met	Met
DOK Adequacy	Met	Partially met	Partially met
Range Adequacy	Met	Met	Met
Balance of Knowledge	Met	Partially met	Met
Multidimensional Adequacy	Met	Met	Met

Table 3.35 summarizes the test form alignment criteria results for the three summative assessment science tests. Similar to the item pool results, all test forms are composed of multidimensional items that reflect a range of the CA NGSS. Grade eight and high school test forms were evaluated as not fully reflecting an appropriate range of cognitive complexity levels, notably due to slightly more than 10 percent of items rated at DOK Level 1. Not all grade five and grade eight test forms were evaluated as fully reflecting an appropriate balance among the CA NGSS dimensions, though all calculated balance index values were within three points of the threshold value.

Table 3.35 Percentage of Grade Level Forms Fully Meeting Each Criterion

Criterion	Grade 5	Grade 8	High School
Links to Standards	100	100	100
DOK Adequacy	100	0[a]	0[a]
Range Adequacy	100	100	100
Balance of Knowledge	60[b]	93[b]	100
Multidimensional Adequacy	100	100	100

[a] 100 percent of grade eight and high school forms at least partially met the DOK Adequacy criterion.
[b] 100 percent of grade five and eight forms at least partially met the Balance-of-Knowledge criterion.

Discussion

Overall, the alignment workshop results provide strong support that the CAST design produces aligned test forms. All test forms at all grade levels at least partially met all five *a priori* alignment criteria that were evaluated. Alignment criteria that were not fully met for all test forms include Depth of Knowledge Adequacy and Balance of Knowledge.

Forms that did not meet the Depth of Knowledge Adequacy criterion contained slightly more Level 1 DOK items than the 10 percent maximum outlined in the criterion. Note, also, that for each form, the number of Level 3 DOK items exceeded the ten percent minimum outlined. Failure to meet our proposed alignment criteria is often mitigated by demonstrating that test forms do meet goals outlined in test blueprints, which are reflective of the test's design and goals. At the time of this study, the CAST blueprints did not contain guidelines regarding the distribution of DOK levels. We recommend that such guidelines be added to the blueprint, along with a rationale for the range of items at each DOK level. Such a rationale may include, for example, that performance tasks are designed to lead students through simple to complex sense-making of the science phenomenon under investigation.

All forms that did not meet the Balance of Knowledge criterion were within three points of the minimum balance index threshold. This is likely the reflection of a single or very small number of items being aligned to one dimension over another. The CA NGSS dimensions are designed to be integrated; the categories of each tend to overlap. It is not uncommon for experts to disagree with one another on the specific SEP and CCC codes that should be assigned to a test item. Although no formal confidence intervals around the minimum balance index have been established (in prior alignment research or in this study), the proximity of the calculated index values to the threshold suggest all test forms demonstrated a reasonable level of balance among the SEP and CCC categories.

This raises the issue of the level of agreement between the alignment workshop panelists and item developers. Table 3.36 presents the percentage of agreement between the final consensus ratings and the item metadata. To calculate these values, the final consensus ratings for each item were compared to the item metadata from ETS. If the consensus rating and metadata matched, then agreement was noted. The values in table 3.36 reflect the percentage of items for which there was agreement between the consensus rating and metadata for each dimension/DOK. It is important to note the final consensus ratings were recorded after the panel had viewed and discussed the item metadata, and so levels of agreement reflect the panels' ratings after taking the metadata into consideration. The highest level of disagreement was observed on the high school DOK ratings. The panel disagreed with the DOK level reported in the item metadata for slightly more than half of the high school items. When there was disagreement, high school panelists tended to rate the items at the lower adjacent DOK level compared to the metadata.

Table 3.36 Percentage of Agreement with Item Metadata

CAST Item Pool Grade Level	PE	DCI[a]	SEP	CCC	DOK
Grade 5	92	85	60	70	66
Grade 8	97	95	90	94	66
High School	92	78	72	80	46

[a] DCI values are based on agreement at the sub-practice level.

This page is intentionally blank.

Chapter 4: Conclusions

This study combined documentation review and item ratings by content experts to evaluate the alignment between the California Science Test (CAST) and the California Next Generation Science Standards (CA NGSS). Specifically, the study addressed seven research questions. This chapter presents the response to each research question, based on the study results.

Research Question 1: To what extent do the test design and test blueprints for the CAST support the claims to be made about student performance on the assessment?

Review of available documentation found that the test design and test blueprints for the CAST support the conclusion that the testing contractor adhered to testing standards relevant to test-to-standards alignment (see table 2.2). Review of operational test forms from the 2018–19 administration support that the CAST design produces aligned test forms (see table 3.35).

Research Question 2: To what extent does the test blueprint for the CAST represent an appropriate sampling of the content as set forth in the CA NGSS?

The CAST is designed such that its content at each grade level will rotate across years, each year sampling different content from the CA NGSS. The rotation is designed to allow CAST to address the full breadth of the CA NGSS over a three-year span. Table 4.1 compares the number of PEs that should be tested each year in order to meet the test blueprint with the number of PEs tested via the item pool in Year 1, based on expert panelists' ratings. The PEs assessed via the 2018–19 item pool were sufficient to support that the CAST is on track to address the full breadth of the CA NGSS after two additional operational administrations.

Table 4.1 Comparison of PE Needs Per Administration and PEs Tested in Year 1

CAST Item Pool Grade Level	Physical Sciences PEs Needed Per Year	Physical Sciences PEs Tested in Year 1	Life Sciences PEs Needed Per Year	Life Sciences PEs Tested in Year 1	Earth & Space Sciences PEs Needed Per Year	Earth & Space Sciences PEs Tested in Year 1
Grade 5	5–6	11	4	10	4–5	9
Grade 8	6–7	13	7	14	5	10
High School	8	10	8	12	6–7	9

Research Question 3: To what extent do the CAST test forms and test items reflect the test design and test blueprints?

Based on expert panelists' ratings, the number of items linked to each content domain, science and engineering practice, and crosscutting concept align with the guidelines presented in the CAST blueprints. In only a small number of instances did the number of items rated as aligned to a particular dimension fall slightly outside of the ranges specified in the blueprint. Tables depicting these comparisons are presented in Appendix C.

Research Question 4: To what extent do CAST tasks and items integrate more than one disciplinary core idea, crosscutting concept, and/or science and engineering practice?

Expert reviewers found that most of the CAST items, across the grade levels, measure a performance expectation by integrating a disciplinary core idea, crosscutting concept, and/or science and engineering practice. Table 4.2 summarizes the percentage of items on each test form that was rated as multidimensional. Across the grade levels, the majority of items were rated as multidimensional, and more than half of items on any test form were rated as integrating all three dimensions.

Table 4.2 Summary of Multidimensional Items by Grade Level

Grade Level	Range of Percentages of Items Aligned to Two or More Dimensions	Range of Percentages of Items Aligned to All Three Dimensions
Grade 5	91–93	64–80
Grade 8	91–98	88–95
High School	98–100	84–86

Research Question 5: To what extent do CAST test forms show balance across the disciplinary areas used for scoring and reporting purposes (earth and space sciences, life sciences, and physical sciences)?

CAST forms across the grade levels reflect reasonable balance across the disciplinary areas used for scoring and reporting purposes (earth and space sciences, life sciences, and physical sciences), as well as across the CA NGSS science and engineering practices and crosscutting concepts. This was determined by calculating Webb's balance index for each. This index takes into consideration (a) the number of content domains, SEPs, and CCCs measured by the items and (b) the proportion of items measuring each domain, SEP, or CCC. For most forms across the grade levels, an *a priori*-defined minimum index was met (see tables 3.11, 3.21, and 3.31). For a smaller number of forms, this index was missed by only three points on a 100-point scale.

Research Question 6: Do the CAST items range from low to high cognitive complexity and provide a sufficient number of items across the range of cognitive complexity?

Expert reviewers indicated that CAST items vary in cognitive complexity, with slightly more than the *a priori* limit of 10 percent at Level 1 DOK and also more than the *a priori* minimum of 10 percent at Level 3 DOK. (see tables 3.7, 3.17, and 3.27).

Research Question 7: How well does CAST fit the population being tested, in terms of the distribution of item difficulties within test forms and the distribution of student ability?

Item-person maps, or Wright maps, illustrate the correspondence between test takers' ability and the difficulty of the test items. Ideally, test items will be at an appropriate level of difficulty to measure the test takers' ability level, ensuring that the test provides information about test performance that is meaningful and useful. For example, test scores on a test in which most items are too difficult for most test takers would result in an underestimation of true achievement levels. Item-person maps for each grade level were produced by ETS. HumRRO conducted additional item mapping analyses, classifying items into achievement levels based on the score associated with having a 50 percent probability of responding correctly to an item (or receiving full points for a multi-point items). This classification represents the achievement level at which each item is providing the most information about student performance. Item-person maps and item-achievement level classification results are presented in Appendix D.

In the evaluation of this operational administration, the item-person maps in Appendix D generally depict item difficulty being aligned with students' ability levels. For all three grades, the distribution of item difficulties generally lines up with the distribution of student ability levels. For high school, the item difficulty distribution relative to the student ability distribution has a slightly more upward shift compared to the other two grades. This indicates that the high school test has fewer items that are at a difficulty level that is comparable to students on the lower end of the ability distribution. Across grade levels and forms, item-achievement level classifications indicate that the largest percentage of items tended to be classified at Achievement Level 2, with some exceptions. In grade eight and high school, there were some forms in which a slightly higher percentage of items were rated at Achievement Level 4. This is in part due to multipoint items being classified based on the probability of earning full points (i.e., the ability level associated with having a 50 percent probability of getting the full two points on a two-point test item). Classifying items based on the probability of earning at least partial points (i.e., the ability level associated with having a 50 percent probability of getting at least one point on a two-point test item) would likely result in fewer items classified at Achievement Level 4.

Classifying items into achievement levels provides insight into how well a test form can differentiate among different levels of student performance. This is done by calculating the probability of answering each item correctly at each student ability level. Items are then classified into achievement levels based on the student ability level associated with having a 50 percent probability of answering the item correctly. During standard setting,

CAST achievement levels were set such that the largest percentage of students are expected to be classified at Achievement Level 2 based on the 2018–19 spring operational test administration. Thus, it makes sense that a large proportion of items would be targeting students at this level. But test forms also contained items targeting the higher achievement levels, and Level 1 Achievement to a lesser extent, thus providing information about student performance at all levels. It is important to note that California educators are still developing strategies for teaching the CA NGSS in the classroom. As students have more opportunities to learn the CA NGSS, the correspondence between student ability and item difficulty is expected to shift.

Recommendations

The study results were generally very positive and do not indicate that any major changes in test development or forms construction processes and procedures are needed. We do offer one recommendation for improving the CAST blueprints:

1. **Add recommended cognitive complexity distributions to the CAST blueprints, along with a rationale for the targets set for each level.**

Establishing guidelines for cognitive complexity in the CAST blueprints will enhance item development and forms construction by clearly stating the proportions of items at each cognitive complexity level that should be included on each test form. This information will be helpful in ongoing evaluations of the adequacy of the item pool for building multiple test forms and for verifying that forms contain items from an appropriate range of cognitive complexity levels. These guidelines should include a rationale for each cognitive complexity level, noting why some levels are emphasized over others and how this design reflects the intent of the CA NGSS as well as the interpretation and use of CAST scores.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Hardoin, M. M., Norman Dvorak, R., Thacker, A. A., Paulsen, J., Gribben, M., & Handy, K. (2019). California Assessment of Student Performance and Progress (CAASPP): 2019 independent evaluation report (2019 No. 102). In S. Schultz, L. Wise, & C. Watters (Eds.). Alexandria, VA: Human Resources Research Organization.

Hardoin, M. M., Thacker, A. A., Norman Dvorak, R., & Becker, D. E. (2018). California Assessment of Student Performance and Progress (CAASPP): 2018 independent evaluation report (2018 No. 087). In C. Watters (Ed.). Alexandria, VA: Human Resources Research Organization.

Webb, N. L. (1997). Criteria for alignment of expectations and assessments in mathematics and science in education (Research Monograph No. 6). Madison, WI: National Institute for Science Education.

Webb, N. L. (1999). Alignment of science and mathematics standards and assessments in four states (Research Monograph No. 18). Madison, WI: University of Wisconsin–Madison, National Institute for Science Education.

Webb, N. L. (2002) Alignment study in language arts, mathematics, science, and social studies of state standards and assessments for four states. Washington, D.C.: Council of Chief State School Officers, December, 2002.

This page is intentionally blank.

Glossary of Acronyms

Acronym	Glossary
CA NGSS	California Next Generation Science Standards
CAASPP	California Assessment of Student Performance and Progress
CAST	California Science Test
CCC	Crosscutting Concept
DCI	Disciplinary Core Idea
PE	Performance Expectation
SEP	Science and Engineering Practice

This page is intentionally blank.

Appendix A: CAST Documentation Reviewed by HumRRO

Table A.1. CAST Documents Reviewed

Document Focus	Document File Name
NGSS Standards, Core Concepts, and Performance Expectations	• 05_Grade-5-Performance-Expectations • 06_Middle-School-Performance-Expectations • 07_High-School-Performance-Expectations • 04_Dimensions-of-the-CA-NGSS • 11_Appendix-1-from-the-California-Science-Framework
Test Design	• Castblueprint • Form Planners_all grades (7) folder • CriteriaforStatewidesummativescienceassessments_03192018 • CAST Editorial & Graphics Style Guide V4 • CAST ECD White Paper-2nd submission to CDE 6-29-2018 • DRAFT CAST Evidence-Centered Design White Paper 070218 • Gr 8 Reference sheet • High School Reference sheet • 060118-02-v3_FOR ARCHIVE_CAST ECD White Paper • 167-2019C-v3_FOR ARCHIVE CAST_Test_Specs_092518 • 239-2016 FOR ARCHIVE – CA NGSS GEN AssessmentDesign_022416 • CAST Phenomenon Memo • SBE CAASPP update Aug 2018 • SBE memo NGSS Imp feb2018
Item Development and Information	• CAASPP Item Acceptance Criteria for IRC 021618_v3 • Item Authoring Template • CAST_Review discrete Items Process Map_D013019 • CAST OIW Part 3 Gr5_Final • CAST OIW Part 3 MS_Final • CAST OIW Part 3 HS_Final • CAST OIW Part 4_PT_Final • Item_Review_040218 • CAST Academy Item Specs grade 5, 8, and HS • 110317-01_FOR ETS_CAST_IWW_Nov 2017_110717_FINAL • PT_WritingTemplate • CAST_Gr5_2020 NID_IDP_v01 • CAST_Gr8_2020 NID_IDP_v01 • CAST_HS_2020 NID_IDP_v01 • 102717-06-v3_FOR ARCHIVE_CAST Item Type Specifications_030718

Table A.1. (cont.)

Document Focus	Document File Name
Item Development and Information (cont.)	• 168-2018-v2_FOR_ARCHIVE_IDP_032118 • CAST_2019 OP_Gr5_Non accessible BB_for CDE_v1 • CAST_2019 OP_Gr8_Non accessible BB_for CDE_v9 • CAST_2019 OP_HS_Non accessible BB_for CDE_v9
DOK Information	• DOK-Science • Webbs_DOK_Guide • Updated DOK CAST_2019 OP • Training of TD'ers on Assigning of DOK values • Assigning of DOK_v2
Test Fairness, Accessibility, and Accommodations	• Sciencebentobox0918 • ATF ETS Accessibility Handbook Content Development • castaccesssupt • Applying Principle of Digital Accessibility training for Assessment Specialists • Universal Design Training for Assessment Specialists • ETS Guidelines for Fair Tests and Communications • Fairness Review Book for Assessment Specialists • ETS Standards for Quality and Fairness • Fairness Training PowerPoint for Assessment Specialists • 060618-02-v2 FOR ARCHIVE_CAST Field Test Data Review Reference Sheet 061
Item Scoring	• HS_DRAFT_Content Training for Raters and Scoring Leaders_021519 • VH651815 Benchmark and Annotation Examples Table • VH651815_Fossil Map_Scoring Notes • VH651815_Fossil Map_Rubric • CAST Constructed Response Scoring Overview • ETS Machine Scoring Introduction
Field Test	• 234-2018C_FOR REVIEW_CAST Technical Report._022119 • For-[CDE Staff]_060618-01-v2 FOR ETS_CAST_Field Test Data • Review PowerPoint_Final
Teacher Training	• 12_CAST-Academy-Slides-Handout • 15_Facilitators-Guide • 18_Instructional-Shifts-Handout • How to Read NGSS - Final 4-19-13

Note: Documents reviewed may not have been cited for supporting an evaluated testing standard.

Appendix B: Alignment Workshop Materials

List of Materials

Workshop Agenda

Panelist Instructions (including sample Performance Expectation from CA NGSS and sample Item Specifications)

Example Item Rating Form

Depth-of-Knowledge Help Sheet

Debriefing Survey

Workshop Evaluation Survey

Workshop Agenda

California Science Test (CAST) Alignment Study
February 28 and March 1, 2019
Sacramento Marriott Rancho Cordova
Rancho Cordova, CA

Day 1 - Thursday, February 28

8:00 – 8:30 a.m.	Panelists sign in and sign CAASPP Confidentiality Agreement
8:30 – 10:30 a.m.	Welcome, introductions, logistics, and general training
10:30 – 10:45 a.m.	Break - Report to Grade 5, Grade 8, and High School Panel Rooms
10:45 – 11:30 a.m.	Panel Introductions and Training on Item Viewing
11:30 – 12:00 noon	Review Panelist Instructions and Rating Processes
12:00 – 1:00 p.m.	Buffet Lunch (*staggered release of each Panel*)
1:00 – 2:00 p.m.	Begin iterative alignment rating process: • Independent rating • Discussion and consensus building • Group review of metadata • Final independent and consensus ratings
2:00 – 2:45 p.m.	Continue iterative alignment rating process
2:45 – 3:00 p.m.	Break
3:00 – 4:30 p.m.	Continue iterative alignment rating process

Day 2 - Friday, March 1

8:30 – 10:00 a.m.	If needed: Review and Correct Rating Spreadsheets; Continue iterative alignment rating process
10:00 – 10:15 a.m.	Break
10:15 – 12:00 noon	Continue iterative alignment rating process
12:00 – 1:00 p.m.	Buffet Lunch (*staggered release of each Panel*)
1:00 – 2:30 p.m.	Continue iterative alignment rating process
2:30 – 2:45 p.m.	Break
2:45 – 4:15 p.m.	Complete iterative alignment rating process
4:15 – 4:30 p.m.	Debrief, workshop evaluation, and adjourn

Panelist Instructions

1	CA NGSS	Print and electronic copy
2	CAST Item Specifications	Print and electronic copy
3	CAST Rating Spreadsheet	Excel (panelists and facilitator)
4	CAST DOK rating guide	Print copy
5	CA NGSS Appendix F and G	Electronic, with print copy of first page for reference
6	CAST Items	Accessed via computer link
7	Panelist Instructions	Print copy
8	Debriefing/Evaluation Form	Print copy
9	Demographic Questionnaire	Print copy

Panelists *NOT* allowed cell phones or open email at table

Prior to alignment steps:

1. Introductions
2. Review all of the materials that panelists should have
    a. Laptops for recording ratings in Excel and accessing CAST items
    b. Panelist Instructions
    c. CA NGSS
    d. Item Specifications
    e. Depth of Knowledge (DOK) Levels for Science
3. Additional documents will be handed out as needed
    a. Demographic Questionnaire
    b. Debriefing/Evaluation form

1 Rate CAST Items

Train Task:

1. Panelists will review several CAST items and will assign each item's DOK level and enter the standard information that best matches with what the item measures.
2. Access CAST_ItemRating Spreadsheet Excel file:
    a. Locate the file on the desktop, double click to open.
    b. Panelists Save As file name and add **underscore and their 3 initials** to the file name (e.g., CAST ItemRating_*panelgroup*_**ymn**). Make sure no one in the group has the same initials.
    c. Autosave (under File, Options) should already be set to 1 minute, **but hit save often**.

3. Review rating categories on Excel form and talk about how to enter data on **first worksheet tab.**
    a. Panelists will only need to review items on the first tab. The other tabs are for internal use only. If any issue occurs with auto-fill for cells, facilitator can address by revising the second tab—hopefully will not be needed.
    b. Columns A through D are filled with information about each CAST item. Column A (hidden) provides the ETS unique item identifier. This will not be used by panelists, but is provided as information for the facilitator in case any items seem to be out of sequence. Column B provides the sequence number. This number will be used by the panelists to make sure everyone is talking about the same item. Panelists should make sure they are viewing the same item as the item listed on the Excel file that they are rating. Column C provides item type (for reference—does not play into alignment). Column D provides the testing contractor's identification of the Domain. This should facilitate finding the correct PE. Items could be mis-identified or may address multiple domains (e.g. life sciences and physical sciences).
    c. Column E asks panelists to identify the PE and type in the associated code from the CA NGSS. For example, **5-LS1-1.** The first number indicates the grade level**,** then the domain, followed by numbers indicating specific PEs within this grade/domain. Panelists should be very familiar with these codes and the CA NGSS document. They can use any version of the CA NGSS they like (if they brought their own).
    d. Columns F, G, and H are for panelists to identify the relevant DCI the item measures. Panelists should use both the CA NGSS and Item Specifications for this task. The specific numbered and lettered sub-groupings for DCIs only exist in the Item Specifications. Panelists should select the DCI from the drop-down menu on their spreadsheet (F). Then, they will need to select the number (G) and letter (H) for the more specific sub-category for that DCI.
    e. Columns I, J, and K are for any identified secondary DCI. The process is the same for completing it as for columns F, G, and H. Panelists—we do not expect there to be secondary DCI for most items.
    f. Columns L and M are for panelists to identify a primary CCC and, if necessary, a secondary CCC. Panelists—we do not expect most items to measure a secondary CCC. Panelists should select the CCC from the drop-down menu on their spreadsheet. Panelists—there may not be a CCC for all items.
    g. Columns N and O are for panelists to identify a primary SEP and, if necessary, a secondary SEP. Panelists—we do not expect most items to measure a secondary SEP. Panelists should select the SEP from the drop-down menu on their spreadsheet. Panelists—there may not be an SEP for all items.
    h. **All secondary fields are for when items measure multiple things. It is not appropriate to identify two DCIs, CCCs, or SEPs because there is**

vagueness between them and panelists are undecided on which is most appropriate.

i. Column P is for panelists to provide the DOK level that best represents the cognitive demand of the item. The verbs are a clue as to the level, but do not rely on that approach. Use your resources—DOK for science, Hess document. Column Q panelists enter any comments of notes regarding the quality of the item or the phenomenon the item references. Panelists should take notes on their own, discuss them, and the facilitator should capture the main agreed upon points in the consensus spreadsheet.

Conduct Task:

1. Panelists rate the first item independently, all indicated fields. Next, panelists discuss their ratings. Focus on why there is disagreement, if any, and what the most appropriate selections should be. Do not spend time discussing items where everyone agrees. Be sure you are comfortable with the Dimensions. Review the meta-data. Discuss any discrepancies between panelists' decisions and the meta-data. Remember that item writers do not select CCC or SEP, but meta-data reflects the CCC or SEP that go with the PE in the item specifications. Should be the same—but may not be. Settle on consensus ratings. Repeat at least 3 times, one item at a time. Panelists should not change ratings after discussion and review unless they are **certain** they want to (due to a coding error or someone convincing them that there is a better match). No changes after seeing the meta-data. We will capture consensus ratings among the panelists, reflecting the inclusion and consideration of the meta-data, but we want to be able to gauge the differences between initial panelists' ratings and final consensus ratings.
2. Panelists should rate all remaining CAST items independently in sets of 3-8 items before discussing and settling on consensus. Repeat the process above for each set of items.
3. Panelists should work independently; however, they may have the occasional discussion about any item(s) that is causing someone difficulty.
4. You should complete between 15-20 items on Day 1—closer to 20 is encouraged. That will leave about 35-45 for Day 2. There is a practice effect and no additional training on Day 2, so this should be ample time. The facilitator will monitor discussion time and encourage quicker consensus as needed (majority if necessary).

2 CAST Debrief

Conduct Task:

1. Participate in a discussion of "whole test." Use guiding questions
    a. Was there anything that surprised you about the CAST items?
    b. Were there major omissions (not specific DCIs—we know we can't test all of them in a given year)?
    c. Are there ways you would like to see the CAST improved? Be specific.
    d. Other issues that panelists feel are important.

2. Complete CAST Debriefing Form.
3. Panelist's responses will be confidential and anonymous at the individual level.
4. The front of the document asks about the alignment of the CAST in general terms.
5. The back of the document asks how well HumRRO did training and conducting the workshop.

Science Major Dimensions

Science performance expectations are built around the following three major dimensions.

Science and Engineering Practices...describe the major practices scientists employ as they investigate and build models and theories about the world and what engineers use as they design and build systems.

They include:

1. Ask questions (for science)
2. Define problems (for engineering)
3. Develop and use models
4. Plan and conduct investigations
5. Analyze and interpret data
6. Use mathematical and computational thinking
7. Construct explanations (for science)
8. Design solutions (for engineering)
9. Engage in scientific argument from evidence
10. Obtain, evaluate, and communicate information

Disciplinary Core Ideas…represent a set of science and engineering ideas for K-12 science education that have broad importance across multiple sciences or engineering disciplines; provide a key tool for understanding or investigating more complex ideas and solving problems; relate to the interests and life experiences of students; are teachable and learnable over multiple grades at increasing levels of sophistication.

They include:

1. Physical Science
2. Life Science
3. Earth and Space Science

Crosscutting Concepts…represent common threads or themes that span across science disciplines (biology, chemistry, physics, environmental science, Earth/space science) and have value to both scientists and engineers because they identify universal properties and processes found in all disciplines.

They include:

1. Patterns
2. Cause and Effect: Mechanism and Explanations
3. Scale, Proportion, and Quantity
4. Systems and System Models
5. Energy and Matter: Flows, cycles, and conservation
6. Structure and Function
7. Stability and Change

See sample Performance Expectation for grade 5 next page

Next Generation Science Standards for California Public Schools, Kindergarten through Grade Twelve

Grade Five
Standards Arranged by Disciplinary Core Ideas

California Department of Education

Clarification statements were created by the writers of CA NGSS to supply examples or additional clarification to the performance expectations and assessment boundary statements.

*The performance expectations marked with an asterisk integrate traditional science content with engineering through a Practice or Disciplinary Core Idea.

**California clarification statements, marked with double asterisks, were incorporated by the California Science Expert Review Panel

The section entitled "Disciplinary Core Ideas" is reproduced verbatim from *A Framework for K–12 Science Education: Practices, Cross-Cutting Concepts, and Core Ideas*. Revised March 2015.

5-LS1 From Molecules to Organisms: Structures and Processes

Students who demonstrate understanding can:

Support an argument that plants get the materials they need for growth chiefly from air and water.

[Clarification Statement: Emphasis is on the idea that plant matter comes mostly from air and water, not from the soil.]

The performance expectations above were developed using the following elements from the NRC document *A Framework for K–12 Science Education*:

Science and Engineering Practices	Disciplinary Core Ideas	Crosscutting Concepts
Engaging in Argument from Evidence Engaging in argument from evidence in 3–5 builds on K–2 experiences and progresses to critiquing the scientific explanations or solutions proposed by peers by citing relevant evidence about the natural and designed world(s). • Support an argument with evidence, data, or a model. (5-LS1-1)	**LS1.C: Organization for Matter and Energy Flow in Organisms** Plants acquire their material for growth chiefly from air and water. (5-LS1-1)	**Energy and Matter** • Matter is transported into, out of, and within systems. (5-LS1-1).

Connections to other DCIs in fifth grade: 5.PS1.A (5-LS1-1)

Articulation of DCIs across grade-bands: K.LS1.C (5-LS1-1); 2.LS2.A (5-LS1-1); MS.LS1.C (5-LS1-1)

California Common Core State Standards Connections:

ELA/Literacy

RI.5.1	Quote accurately from a text when explaining what the text says explicitly and when drawing inferences from the text. (5-LS1-1)
RI.5.9	Integrate information from several texts on the same topic in order to write or speak about the subject knowledgeably. (5-LS1-1)
W.5.1.a–d	Write opinion pieces on topics or texts, supporting a point of view with reasons and information. (5-LS1-1)

Mathematics

MP.2	Reason abstractly and quantitatively. (5-LS1-1)
MP.4	Model with mathematics. (5-LS1-1)
MP.5	Use appropriate tools strategically. (5-LS1-1)
5.MD.1	Convert among different-sized standard measurement units within a given measurement system (e.g., convert 5 cm to 0.05 m), and use these conversions in solving multi-step, real world problems. (5-LS1-1)

Item Specifications for PE 3-LS3-2 Heredity: Inheritance and Variation of Traits

Students who demonstrate understanding can:

Use evidence to support the explanation that traits can be influenced by the environment.

[Clarification Statement: Examples of the environment affecting a trait could include normally tall plants grown with insufficient water are stunted; and, a pet dog that is given too much food and little exercise may become overweight.]

Science and Engineering Practices	Disciplinary Core Ideas	Crosscutting Concepts
Constructing Explanations and Designing Solutions Constructing explanations and designing solutions in 3–5 builds on K–2 experiences and progresses to the use of evidence in constructing explanations that specify variables that describe and predict phenomena and in designing multiple solutions to design problems. • Use evidence (e.g., observations, patterns) to support an explanation.	**LS3.A: Inheritance of Traits** 3. Other characteristics result from individuals' interactions with the environment, which can range from diet to learning. Many characteristics involve both inheritance and environment. **LS3.B: Variation of Traits** 3. The environment also affects the traits that an organism develops.	**Cause and Effect** • Cause and effect relationships are routinely identified and used to explain change.

Assessment Targets

Assessment targets describe the focal knowledge, skills, and abilities for a given three-dimensional Performance Expectation. Please refer to the Introduction for a complete description of assessment targets.

Science and Engineering Subpractice(s)

Please refer to appendix A for a complete list of Science and Engineering Practices (SEP) subpractices. Note that the list in this section is not exhaustive.

6.1 Ability to construct explanations of phenomena

6.2 Ability to evaluate explanations of phenomena

Science and Engineering Subpractice Assessment Targets

Please refer to appendix A for a complete list of SEP subpractice assessment targets. Note that the list in this section is not exhaustive.

6.1.1 Ability to construct quantitative and/or qualitative explanations of observed relationships

6.1.2 Ability to apply scientific concepts, principles, theories, and big ideas to construct an explanation of a real-world phenomenon

6.1.3 Ability to use models and representations in scientific explanations

6.2.2 Ability to use data to support or refute an explanatory account of a phenomenon

Disciplinary Core Idea Assessment Targets

LS3.A.3a Describe that traits can be influenced by the environment

LS3.A.3b Describe that inherited traits vary between organisms of the same type

LS3.A.3c Describe that some traits result from the combination of inherited information and environmental influence

LS3.A.3d Describe environmental factors that can influence traits

LS3.B.3a Describe that the environment can affect the traits an organism develops

LS3.B.3b Describe that traits can be variable due to environmental conditions

LS3.B.3c Use reasoning to connect evidence and support an explanation about environmental influence on inherited traits in organisms

Crosscutting Concept Assessment Target(s)

CCC2 Identify and test cause and effect relationships to explain change

Examples of Integration of Assessment Targets and Evidence

Note that the list in this section is not exhaustive.

Task provides data comparing appearance of a trait under different conditions:

- Makes a quantitative and/or qualitative conclusion regarding the relationships between dependent and independent variables (6.1.1, LS3.A.3, and CCC2)

- Describes how the evidence allows for the distinction between causal and correlational relationships (6.1.1, LS3.A.3, and CCC2)

Task provides data on different plant heights in the same species of plant with different amounts of a particular variable:

- Student correctly uses scientific concepts, principles, theories, and big ideas to explain how the evidence supports a conclusion about environmental influence on traits (6.1.2, LS3.B.3, and CCC2)

Task provides a model about how the environment can influence a trait:

- Uses scientific models to construct an explanation of a phenomenon (6.1.3, LS3.B.3, and CCC2)

- Uses models to represent their explanation (6.1.3, LS3.B.3, and CCC2)

Task provides data to describe the impact of the environment on a particular trait under different conditions:

- Uses data to support an explanatory account of a phenomena (6.2.2, LS3.A.3a, LS3.B.3, and CCC2)

- Uses data to refute an explanatory account of a phenomena (6.2.2, LS3.A.3a, LS3.B.3, and CCC2)

Environmental Principles and Concepts

- EP2: The long-term functioning and health of terrestrial, freshwater, coastal, and marine ecosystems are influenced by their relationships with human societies.

Possible Phenomena or Contexts

Note that the list in this section is not exhaustive.

- Diet and nutrient availability

- Exposure to abiotic factors (water, sunlight, chemicals, etc.)

- Activity level

- Learned responses
- Comparing two ecotypes of the same species
- Change in species composition of a community

Common Misconceptions

Note that the list in this section is not exhaustive.

- The environment cannot impact genetically determined traits.
- Organisms can consciously change their phenotypes to better survive in a given environment.

Additional Assessment Boundaries

None listed at this time.

Additional References

3-LS3-2 Evidence Statement
https://www.nextgenscience.org/sites/default/files/evidence_statement/black_white/3-LS3-2%20Evidence%20Statements%20June%202015%20asterisks.pdf

Environmental Principles and Concepts http://californiaeei.org/abouteei/epc/

California Education and the Environment Initiative http://californiaeei.org/

The *2016 Science Framework for California Public Schools Kindergarten through Grade 12*

Appendix 1: Progression of the Science and Engineering Practices, Disciplinary Core Ideas, and Crosscutting Concepts in Kindergarten through Grade 12
https://www.cde.ca.gov/ci/sc/cf/documents/scifwappendix1.pdf

Appendix 2: Connections to Environmental Principles and Concepts
https://www.cde.ca.gov/ci/sc/cf/documents/scifwappendix2.pdf

Posted by the California Department of Education, June 2019

Example Item Rating Form

ITS Item ID	Item Sequence	Item Type	Domain	Identify the Performance Expectation (PE) *(Type in PE code)*	Identify Primary Disciplinary Core Idea (DCI) Assessment Target *(Select from drop down menus)*	Identify Secondary DCI Assessment Target *(Select from drop down menus)*	Identify Primary Cross Cutting Concept (CCC) *(Select from drop down menu)*	Identify Secondary CCC *(Select from drop down menu)*	Identify Primary Science and Engineering Practice (SEP) *(Select from drop down menu)*	Identify Secondary SEP *(Select from drop down menu)*	Assign an item depth of knowledge (DOK) rating. *(Select from drop down menu)*	Comments *(Provide comments about the appropriateness of phenomena, item quality, etc.)*
10024-1690	1	MCSS - Discrete	LS									
10024-698	2	ZoneMS-Discrete	LS									
10024-1802	3	MCMS - Discrete	LS									
10024-830	4	MatchMS - Discrete	LS									
10024-538	5	MatchMS - Member	LS									

HumRRO prepopulated CAST metadata in first columns of the form:

First gray column: Unique item identification number

Second gray column: Sequential item number (order item was presented to panelists)

Third gray column: Type of item reviewed (e.g., multiple choice, matching)

Fourth gray column: Science domain the item intended to measure

Panelists entered item-level rating data in cells under blue headers of the form:

First blue column: Performance Expectation alignment

Second blue column: Primary Disciplinary Core Idea alignment

Third blue column: Secondary Disciplinary Core Idea alignment

Fourth blue column: Primary Crosscutting Concept alignment

Fifth blue column: Secondary Crosscutting Concept alignment

Sixth blue column: Primary Science and Engineering Practice alignment

Seventh blue column: Secondary Science and Engineering Practice alignment

Eighth blue column: Depth-of-Knowledge level

Ninth blue column: Panelists comments

Depth-of-Knowledge (DOK) Help Sheet

According to Norman L. Webb, Wisconsin Center for Educational Research ("Depth-of-Knowledge Levels for Four Content Areas," March 28, 2002), "interpreting and assigning Depth-of-Knowledge Levels to both objectives within standards and assessment items is an essential requirement of alignment analysis. Four levels of Depth-of-Knowledge are used for this analysis." Norman Webb's "Depth-of- Knowledge Levels for Four Content Areas" include: Language Arts (Reading, Writing), Mathematics, Science, and Social Studies.

A general definition for each of the four (Webb) Depth-of-Knowledge levels is followed by table 1, which provides further specification and examples for each of the DOK levels. Webb recommends that large-scale, on-demand assessments in reading should assess only Depth-of-Knowledge Levels 1, 2, and 3. Depth-of-Knowledge at Level 4 in science should be reserved for local assessment only.

Descriptors of DOK Levels for Science (based on Webb and Wixson, March 2002)

Level 1 Recall and Reproduction requires recall of information, such as a fact, definition, term, or a simple procedure, as well as performing a **simple** science process or procedure. Level 1 only requires students to demonstrate a rote response, use a well-known formula, follow a set procedure (like a recipe), or perform a clearly defined series of steps. A "simple" procedure is well-defined and typically involves only **one-step**. Verbs such as "identify," "recall," "recognize," "use," "calculate," and "measure" generally represent cognitive work at the recall and reproduction level. Simple word problems that can be directly translated into and solved by a formula are considered Level 1. Verbs such as "describe" and "explain" could be classified at different DOK levels, depending on the complexity of what is to be described and explained.

A student answering a Level 1 item either knows the answer or does not: that is, the answer does not need to be "figured out" or "solved." In other words, if the knowledge necessary to answer an item automatically provides the answer to the item, then the item is at Level 1. If the knowledge necessary to answer the item does not automatically provide the answer, the item is at least at Level 2.

Level 2 Skills and Concepts includes the engagement of some mental processing beyond recalling or reproducing a response. The content knowledge or process involved is **more complex** than in level 1. Items require students to make some decisions as to how to approach the question or problem. Keywords that generally distinguish a Level 2 item include "classify," "organize," "estimate," "make observations," "collect and display data," and "compare data." These actions imply **more than one step**. For example, to compare data requires first identifying characteristics of the objects or phenomenon and then grouping or ordering the objects. Level 2 activities include making observations and collecting data; classifying, organizing, and comparing data; and organizing and displaying data in tables, graphs, and charts.

Some action verbs, such as "explain," "describe," or "interpret," could be classified at different DOK levels, depending on the complexity of the action. For example, interpreting information from a simple graph, requiring reading information from the graph, is a Level 2. An item that requires interpretation from a complex graph, such as making decisions regarding features of the graph that need to be considered and how information from the graph can be aggregated, is at Level 3.

Level 3 Strategic Thinking requires deep knowledge using reasoning, planning, using evidence, and a higher level of thinking than the previous two levels. The cognitive demands at Level 3 are **complex and abstract**. The complexity results not only from the fact that there could be multiple answers, a possibility for both Levels 1 and 2, but because the multi-step task requires **more demanding reasoning**. In most instances, requiring students to explain their thinking is at Level 3; requiring a very simple explanation or a word or two should be at Level 2. An activity that has more than one possible answer and requires students to justify the response they give would most likely be a Level 3. Experimental designs in Level 3 typically involve more than one dependent variable. Other Level 3 activities include drawing conclusions from observations; citing evidence and developing a logical argument for concepts; explaining phenomena in terms of concepts; and using concepts to solve non-routine problems.

Level 4 Extended Thinking requires **high cognitive demand** and is **very complex**. Students are required to make several connections—relate ideas *within* the content area or *among* content areas—and have to select or devise one approach among many alternatives on how the situation can be solved. Many on-demand assessment instruments will not include any assessment activities that could be classified as Level 4. However, standards, goals, and objectives can be stated in such a way as to expect students to perform extended thinking. "Develop generalizations of the results obtained and the strategies used and apply them to new problem situations," is an example of a Grade 8 objective that is a Level 4. Many, but not all, performance assessments and open-ended assessment activities requiring significant thought will be at a Level 4.

Level 4 requires complex reasoning, experimental design and planning, and **probably will require an extended period of time** either for the science investigation required by an objective, or for carrying out the multiple steps of an assessment item. However, the extended time period is not a distinguishing factor if the required work is only repetitive and does not require applying significant conceptual understanding and higher-order thinking. For example, if a student has to take the water temperature from a river each day for a month and then construct a graph, this would be classified as a Level 2 activity. However, if the student conducts a river study that requires taking into consideration a number of variables, this would be a Level 4.

Table B.1 Detailed Descriptors of Depth-of-Knowledge Levels for Science

Level 1 - Recall & Reproduction

a. Recall or recognize a fact, term, definition, simple procedure (such as one step), or property
b. Demonstrate a rote response
c. Use a well-known formula
d. Represent in words or diagrams a scientific concept or relationship
e. Provide or recognize a standard scientific representation for simple phenomenon
f. Perform a routine procedure, such as measuring length
g. Perform a **simple** science process or a set procedure (like a recipe)
h. Perform a clearly defined set of steps
i. Identify, calculate, or measure

***Note*: If the knowledge necessary to answer an item automatically provides the answer, it is a Level 1.**

Level 2 - Skills & Concepts

a. Specify and explain the relationship between facts, terms, properties, or variables
b. Describe and explain examples and non-examples of science concepts
c. Select a procedure according to specified criteria and perform it
d. Formulate a routine problem given data and conditions
e. Organize, represent, and compare data
f. Make a decision as to how to approach the problem
g. Classify, organize, or estimate
h. Compare data
i. Make observations
j. Interpret information from a simple graph
k. Collect and display data

***Note*: If the knowledge necessary to answer an item does not automatically provide the answer, then the item is at least a Level 2. Most actions imply more than one step.**

Table B.1 (cont.)

Level 3 - Strategic Thinking

a. Interpret information from a complex graph (such as determining features of the graph or aggregating data in the graph)
b. Use reasoning, planning, and evidence
c. Explain thinking (beyond a simple explanation or using only a word or two to respond)
d. Justify a response
e. Identify research questions and design investigations for a scientific problem
f. Use concepts to solve non-routine problems/more than one possible answer
g. Develop a scientific model for a complex situation
h. Form conclusions from experimental or observational data
i. Complete a multi-step problem that involves planning and reasoning
j. Provide an explanation of a principle
k. Justify a response when more than one answer is possible
l. Cite evidence and develop a logical argument for concepts
m. Conduct a designed investigation
n. Research and explain a scientific concept
o. Explain phenomena in terms of concepts

***Note*: Level 3 is complex and abstract. If more than one response is possible, it is at least a Level 3 and calls for use of reasoning, justification, evidence, as support for the response.**

Level 4 - Extended Thinking

a. Select or devise approach among many alternatives to solve problem
b. Based on provided data from a complex experiment that is novel to the student, deduct the fundamental relationship between several controlled variables.
c. Conduct an investigation, from specifying a problem to designing and carrying out an experiment, to analyzing its data and forming conclusions
d. Relate ideas *within* the content area or *among* content areas
e. Develop generalizations of the results obtained and the strategies used and apply them to new problem situations

***Note*: Level 4 activities often require an extended period of time for carrying out multiple steps; however, time alone is not a distinguishing factor if skills and concepts are simply repetitive over time.**

Source: K. Hess, Center for Assessment, based on Webb, update 2005

Debriefing: Analysis of Alignment Outcomes for the California Science Test (CAST)

1. Panel: ____ Grade 5 ____Grade 8 ____High School

2. Did the items you reviewed generally represent the content in the CA NGSS that you expected to be covered? If not, what content seemed underrepresented or overrepresented?

3. Did the items generally reflect the level of cognitive complexity (DOK) you expected? If not, were item DOK levels overall lower or higher than expected?

4. Did the items you reviewed generally allow students to demonstrate performance in science? If not, please explain.

5. What is your general opinion of the alignment between the CAST items you reviewed and the CA NGSS?

   - ☐ Excellent
   - ☐ Good
   - ☐ Limited
   - ☐ Weak (please explain and provide some examples)

Comments:

Evaluation: Alignment Workshop Training and Procedures

Please indicate your agreement by marking an 'X' in the appropriate box for each statement.

1. **The training presentation in the large group provided useful information about the CAST and HumRRO's alignment method.**
   - ☐ Strongly Disagree
   - ☐ Disagree
   - ☐ Somewhat Disagree
   - ☐ Agree
   - ☐ Somewhat Agree
   - ☐ Strongly Agree

2. **After the additional training in my small group, I felt prepared to review and rate test items.**
   - ☐ Strongly Disagree
   - ☐ Disagree
   - ☐ Somewhat Disagree
   - ☐ Agree
   - ☐ Somewhat Agree
   - ☐ Strongly Agree

3. **HumRRO staff seemed knowledgeable of the CAST and alignment steps.**
   - ☐ Strongly Disagree
   - ☐ Disagree
   - ☐ Somewhat Disagree
   - ☐ Agree
   - ☐ Somewhat Agree
   - ☐ Strongly Agree

4. **The Panelist Instruction document was clear, understandable, and useful in performing the alignment steps.**
   - ☐ Strongly Disagree
   - ☐ Disagree
   - ☐ Somewhat Disagree
   - ☐ Agree
   - ☐ Somewhat Agree
   - ☐ Strongly Agree
5. **The Excel file was understandable and relatively easy to use to enter item ratings.**
   - ☐ Strongly Disagree
   - ☐ Disagree
   - ☐ Somewhat Disagree
   - ☐ Agree
   - ☐ Somewhat Agree
   - ☐ Strongly Agree
6. **The process for reaching consensus ratings was conducted fairly.**
   - ☐ Strongly Disagree
   - ☐ Disagree
   - ☐ Somewhat Disagree
   - ☐ Agree
   - ☐ Somewhat Agree
   - ☐ Strongly Agree

If you rated any statement Disagree or Strongly Disagree, suggest ideas for improvement:

If you have additional feedback, share your thoughts and comments below.

Appendix C: Test Form-Blueprint Comparison

List of Tables

This page is intentionally blank.

Table C.1 Grade 5 Comparison of Forms and Test Blueprints: Domain

Domain	Number of Items Linked by Panelists	Range from Test Blueprint
Earth and Space Sciences	10	8–10
Life Sciences	10	8–10
Physical Sciences	9	8–10
Engineering/Technology Sciences	4	2–4

Note. Range from the test blueprint refers to the number of items that the blueprint specifies should be aligned to the domain/dimension. Range from test blueprint includes Segment A items only. Segment A items are common to all test forms.

Table C.2 Grade 5 Comparison of Forms and Test Blueprints: SEP

SEP	Number of Items Linked by Panelists	Range from Test Blueprint
Asking questions (for science) or Defining problems (for engineering)	0	1–4
Developing and using models	5	1–7
Planning and carrying out investigations	4	1–7
Analyzing and interpreting data	4	2–4
Using mathematics and computational thinking	1	1–2
Constructing explanations (for science) or Designing solutions (for engineering)	6	2–8
Engaging in argument from evidence	3	1–8
Obtaining, evaluating, and communicating information	2	1–3

Note. Range from the test blueprint refers to the number of items that the blueprint specifies should be aligned to the domain/dimension. Range from test blueprint includes Segment A items only. Segment A items are common to all test forms.

Table C.3. Grade 8 Comparison of Forms and Test Blueprints: Domain

Domain	Number of Items Linked by Panelists	Range from Test Blueprint
Earth and Space Sciences	9	8–10
Life Sciences	9	8–10
Physical Sciences	11	8–10
Engineering/Technology Sciences	2	2–4

Note. Range from the test blueprint refers to the number of items that the blueprint specifies should be aligned to the domain/dimension. Range from test blueprint includes Segment A items only. Segment A items are common to all test forms.

Table C.4. Grade 8 Comparison of Forms and Test Blueprints: SEP

SEP	Number of Items Linked by Panelists	Range from Test Blueprint
Asking questions (for science) or Defining problems (for engineering)	1	1–3
Developing and using models	9	1–16
Planning and carrying out investigations	3	1–5
Analyzing and interpreting data	7	1–9
Using mathematics and computational thinking	0	1–2
Constructing explanations (for science) or Designing solutions (for engineering)	4	1–12
Engaging in argument from evidence	4	1–8
Obtaining, evaluating, and communicating information	3	1–4

Note. Range from the test blueprint refers to the number of items that the blueprint specifies should be aligned to the domain/dimension. Range from test blueprint includes Segment A items only. Segment A items are common to all test forms.

Table C.5 High School Comparison of Forms and Test Blueprints: Domain

Domain	Number of Items Linked by Panelists	Range from Test Blueprint
Earth and Space Sciences	10	8–10
Life Sciences	10	8–10
Physical Sciences	10	8–10
Engineering/Technology Sciences	4	2–4

Note. Range from the test blueprint refers to the number of items that the blueprint specifies should be aligned to the domain/dimension. Range from test blueprint includes Segment A items only. Segment A items are common to all test forms.

Table C.6 High School Comparison of Forms and Test Blueprints: SEP

SEP	Number of Items Linked by Panelists	Range from Test Blueprint
Asking questions (for science) or Defining problems (for engineering)	2	2–3
Developing and using models	8	2–6
Planning and carrying out investigations	4	2–5
Analyzing and interpreting data	3	2–5
Using mathematics and computational thinking	3	2–6
Constructing explanations (for science) or Designing solutions (for engineering)	8	2–6
Engaging in argument from evidence	2	2–6
Obtaining, evaluating, and communicating information	2	2–6

Note. Range from the test blueprint refers to the number of items that the blueprint specifies should be aligned to the domain/dimension. Range from test blueprint includes Segment A items only. Segment A items are common to all test forms.

This page is intentionally blank.

Appendix D: Item-Person Maps and Item-to-Achievement Level Classifications

List of Tables

List of Figures

This page is intentionally blank.

Item-Person Maps and Item-to-Achievement Level Classifications

Tables D.1 through D.3 are called item-person maps and present a comparison of student ability and test item difficulty. The left side of the map shows the distribution of student ability levels, or the *Theta Distribution*. The right side of the map shows the distribution of item difficulty levels. Each figure breaks across two pages, but if you were to put the two pages together, both the student ability and item difficulty distributions would take on a bell curve shape (oriented vertically).

Both student ability and item difficulty are presented on the same scale, represented by the *Value* column at the center of the map. These values are also referred to as *bins*. The students at the top of the map had the highest scores (highest ability students), while the items at the top of the map are the most difficult. The students at the bottom of the map earned the lowest scores (lowest ability students), and the items at the bottom of the map are easiest. When students and items are directly opposite each other on the map, the difficulty of the items and the ability of the students are comparable. Students and items are comparable when a student at a given ability level has about a 50 percent probability of correctly answering an item at that level of difficulty.

Figures D.1 through D.28 depict the percentage of items classified at each achievement level. Items were classified by calculating the probability of answering each item correctly at each student ability level. Items were then classified into achievement levels based on the student ability level associated with having a 50 percent probability of answering the item correctly. Achievement level cut scores were identified during a standard setting process that was separate from this study. The CAST achievement levels are Standard Not Met (Level 1), Standard Nearly Met (Level 2), Standard Met (Level 3), and Standard Exceeded (Level 4).

Table D.1 Grade Five Item-Person Map

Number of Students	Theta Distribution*	Value	Item Difficulty Distribution**	Number of Items
0	-	5.0	O####	5
0	-	4.8	-	0
0	-	4.6	-	0
0	-	4.4	-	0
0	-	4.2	O	1
812	.	4.0	O	1
376	.	3.8	-	0
292	.	3.6	#	1
977	.	3.4	#	1
466	.	3.2	O##	3
1,331	.	3.0	-	0
2,002	.X	2.8	-	0
2,071	.X	2.6	O##	3
3,651	.XX	2.4	##	2
3,566	.XX	2.2	O###	4
8,473	.XXXXX	2.0	OO##	4
7,822	.XXXXX	1.8	OOO###	6
9,009	.XXXXXX	1.6	O######	7
15,108	.XXXXXXXXXX	1.4	OOOO##########	15
15,253	.XXXXXXXXXX	1.2	OOO###	6
18,379	.XXXXXXXXXXX	1.0	OOO####	7
21,655	.XXXXXXXXXXXXX	0.8	OOOOO##############	20
29,890	.XXXXXXXXXXXXXXXXX	0.6	OOO#########	12
29,813	.XXXXXXXXXXXXXXXXX	0.4	OOOOO##########	16
25,369	.XXXXXXXXXXXXXXX	0.2	OOO###################	23
38,669	.XXXXXXXXXXXXXXXXXXXXXXX	0.0	OOOO#######################	27

*For each bin in the theta distribution column, “X” represents 1,500 students, “.” represents a value in between 1 and 1,499 students, and no students are denoted as “-”.

**For each bin in the item difficulty distribution column, “O” represents an operational item, “#” represents a field-test item, and no items are denoted as “-”.

Table D.1 (cont.)

Number of Students	Theta Distribution*	Value	Item Difficulty Distribution**	Number of Items
25,983	.XXXXXXXXXXXXXXXX	-0.2	OOOOOOOOO##########################	35
36,877	.XXXXXXXXXXXXXXXXXXXXXXX	-0.4	OOO####################	23
29,799	.XXXXXXXXXXXXXXXXXX	-0.6	OOOO############	17
31,134	.XXXXXXXXXXXXXXXXXXX	-0.8	OOOOO################	22
25,159	.XXXXXXXXXXXXXXX	-1.0	O##########	12
24,281	.XXXXXXXXXXXXXXX	-1.2	O####	5
18,209	.XXXXXXXXXXX	-1.4	OO##########	12
13,277	.XXXXXXX	-1.6	#######	7
6,585	.XXXX	-1.8	#####	5
3,973	.XX	-2.0	#	1
2,447	.X	-2.2	-	0
1,093	.	-2.4	#	1
608	.	-2.6	O	1
39	.	-2.8	##	2
168	.	-3.0	-	0
0	-	-3.2	-	0
28	.	-3.4	-	0
23	.	-3.6	-	0
0	-	-3.8	-	0
7	.	-4.0	-	0
0	-	-4.2	-	0
0	-	-4.4	-	0
0	-	-4.6	-	0
0	-	-4.8	O	1
0	-	-5.0	###	3

*For each bin in the theta distribution column, “X” represents 1,500 students, “.” represents a value in between 1 and 1,499 students, and no students are denoted as “-”.

**For each bin in the item difficulty distribution column, “O” represents an operational item, “#” represents a field-test item, and no items are denoted as “-”.

Table D.2 Grade Eight Item-Person Map

Number of Students	Theta Distribution*	Value	Item Difficulty Distribution**	Number of Items
0	-	5.0	OOO#######	10
0	-	4.8	-	0
0	-	4.6	-	0
0	-	4.4	#	1
0	-	4.2	#	1
328	.	4.0	-	0
174	.	3.8	-	0
264	.	3.6	-	0
396	.	3.4	#	1
777	.	3.2	##	2
716	.	3.0	##	2
1,460	.	2.8	#	1
2,478	.X	2.6	###	3
3,365	.XX	2.4	##	2
4,319	.XX	2.2	OO###	5
6,557	.XXXX	2.0	OO########	10
9,381	.XXXXXX	1.8	OOOOO###	8
11,460	.XXXXXXX	1.6	OOOOO#######	12
15,054	.XXXXXXXXX	1.4	OOOO#########	13
19,104	.XXXXXXXXXXX	1.2	OOO##########	18
22,034	.XXXXXXXXXXXXX	1.0	OOOOOO##########	20
22,592	.XXXXXXXXXXXXXX	0.8	OOO##########	17
26,678	.XXXXXXXXXXXXXXXX	0.6	OOOOO##################	23
28,908	.XXXXXXXXXXXXXXXXX	0.4	OOOOOO####################	27
31,560	.XXXXXXXXXXXXXXXXXXX	0.2	OOOO################	21
31,326	.XXXXXXXXXXXXXXXXXX	0.0	OOOO##############	18

*For each bin in the theta distribution column, “X” represents 1,500 students, “.” represents a value in between 1 and 1,499 students, and no students are denoted as “-”.

**For each bin in the item difficulty distribution column, “O” represents an operational item, “#” represents a field-test item, and no items are denoted as “-”.

Table D.2 (cont.)

Number of Students	Theta Distribution*	Value	Item Difficulty Distribution**	Number of Items
28,149	.XXXXXXXXXXXXXXXXX	-0.2	OO##################	20
30,704	.XXXXXXXXXXXXXXXXXXX	-0.4	OOOOOOOOO################	24
31,035	.XXXXXXXXXXXXXXXXXXX	-0.6	OOOOO###############	20
29,321	.XXXXXXXXXXXXXXXXXX	-0.8	OOO###########	14
26,624	.XXXXXXXXXXXXXXXX	-1.0	O#####	6
22,876	.XXXXXXXXXXXXXX	-1.2	OO##	4
20,942	.XXXXXXXXXXXX	-1.4	#	1
13,666	.XXXXXXXXX	-1.6	O	1
7,067	.XXXX	-1.8	O#	2
5,098	.XXX	-2.0	O#	2
2,816	.X	-2.2	-	0
1,130	.	-2.4	-	0
493	.	-2.6	#	1
426	.	-2.8	-	0
72	.	-3.0	-	0
126	.	-3.2	-	0
64	.	-3.4	-	0
0	-	-3.6	-	0
5	.	-3.8	-	0
53	.	-4.0	-	0
0	-	-4.2	-	0
0	-	-4.4	-	0
0	-	-4.6	-	0
0	-	-4.8	#	1
0	-	-5.0	#	1

*For each bin in the theta distribution column, “X” represents 1,500 students, “.” represents a value in between 1 and 1,499 students, and no students are denoted as “-”.

**For each bin in the item difficulty distribution column, “O” represents an operational item, “#” represents a field-test item, and no items are denoted as “-”.

Table D.3 High School Item-Person Map

Number of Students	Theta Distribution*	Value	Item Difficulty Distribution**	Number of Items
0	-	5.0	OOOO###########	15
0	-	4.8	-	0
0	-	4.6	-	0
0	-	4.4	#	1
0	-	4.2	##	2
805	.	4.0	##	2
154	.	3.8	#	1
513	.	3.6	O	1
550	.	3.4	-	0
861	.	3.2	OO##	4
1,131	.	3.0	#	1
1,984	.X	2.8	-	0
2,660	.X	2.6	OOO#####	8
4,713	.XXX	2.4	#######	7
5,787	.XXX	2.2	####	4
5,419	.XXX	2.0	O######	7
8,347	.XXXXX	1.8	OOO###########	15
12,643	.XXXXXXXX	1.6	O#########	10
18,117	.XXXXXXXXXXX	1.4	OOO###########	16
13,973	.XXXXXXXX	1.2	OO###########	14
19,635	.XXXXXXXXXXXX	1.0	OOO##############	18
26,761	.XXXXXXXXXXXXXXX	0.8	OO###########	15
29,860	.XXXXXXXXXXXXXXXXX	0.6	OOOOO##########	15
32,711	.XXXXXXXXXXXXXXXXXX	0.4	OOOOO####################	26
41,579	.XXXXXXXXXXXXXXXXXXXXXXXX	0.2	OOOOOOO################	23
38,566	.XXXXXXXXXXXXXXXXXXXXXXX	0.0	OOO###########	15

*For each bin in the theta distribution column, “X” represents 1,500 students, “.” represents a value in between 1 and 1,499 students, and no students are denoted as “-”.

**For each bin in the item difficulty distribution column, “O” represents an operational item, “#” represents a field-test item, and no items are denoted as “-”.

Table D.3 (cont.)

Number of Students	Theta Distribution*	Value	Item Difficulty Distribution**	Number of Items
40,686	.XXXXXXXXXXXXXXXXXXXXXXXXXX	-0.2	OOOOOOO###############	21
35,324	.XXXXXXXXXXXXXXXXXXXXXX	-0.4	OO######	8
37,773	.XXXXXXXXXXXXXXXXXXXXXXXXXX	-0.6	##################	16
37,543	.XXXXXXXXXXXXXXXXXXXXXXXXXX	-0.8	O#######	8
37,031	.XXXXXXXXXXXXXXXXXXXXXXXXXX	-1.0	OO	2
21,404	.XXXXXXXXXXXXXX	-1.2	##	2
26,085	.XXXXXXXXXXXXXXXXX	-1.4	O	1
16,242	.XXXXXXXXXX	-1.6	-	0
12,982	.XXXXXXXX	-1.8	-	0
6,677	.XXXX	-2.0	-	0
4,810	.XXX	-2.2	-	0
4,677	.XXX	-2.4	-	0
0	-	-2.6	-	0
1,684	.X	-2.8	-	0
672	.	-3.0	-	0
454	.	-3.2	-	0
551	.	-3.4	-	0
0	-	-3.6	-	0
0	-	-3.8	-	0
393	.	-4.0	-	0
0	-	-4.2	-	0
0	-	-4.4	-	0
0	-	-4.6	-	0
0	-	-4.8	##	2
0	-	-5.0	O#####	6

*For each bin in the theta distribution column, “X” represents 1,500 students, “.” represents a value in between 1 and 1,499 students, and no students are denoted as “-”.

**For each bin in the item difficulty distribution column, “O” represents an operational item, “#” represents a field-test item, and no items are denoted as “-”.

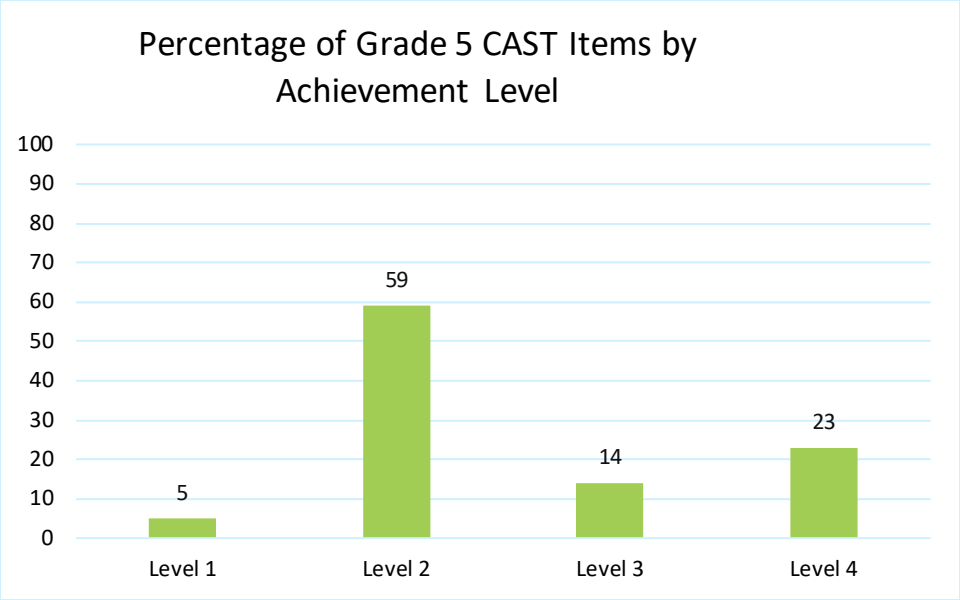

Figure D.1 Item-to-Achievement Level Classification: Grade 5 Form 1.

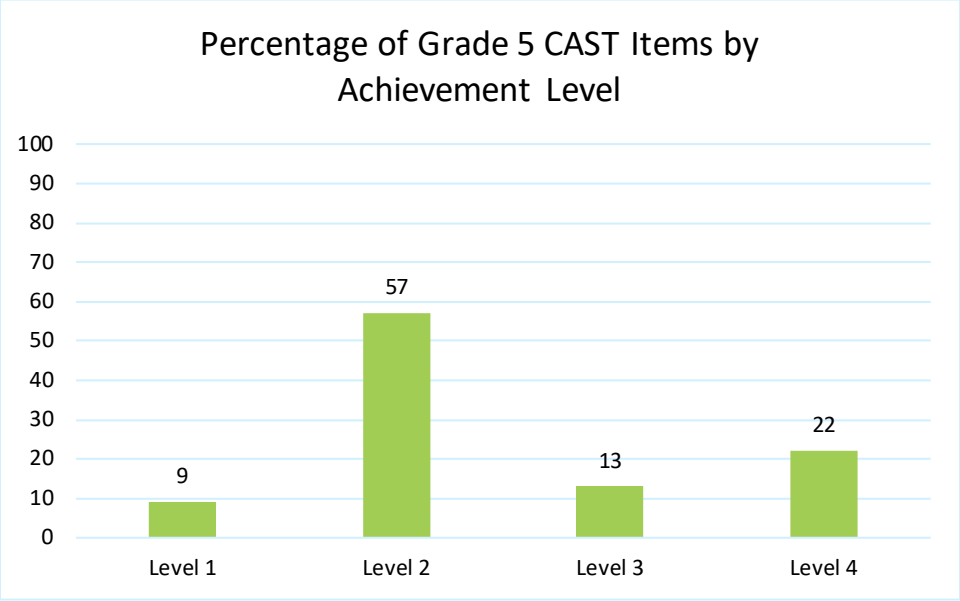

Figure D.2 Item-to-Achievement Level Classification: Grade 5 Form 2.

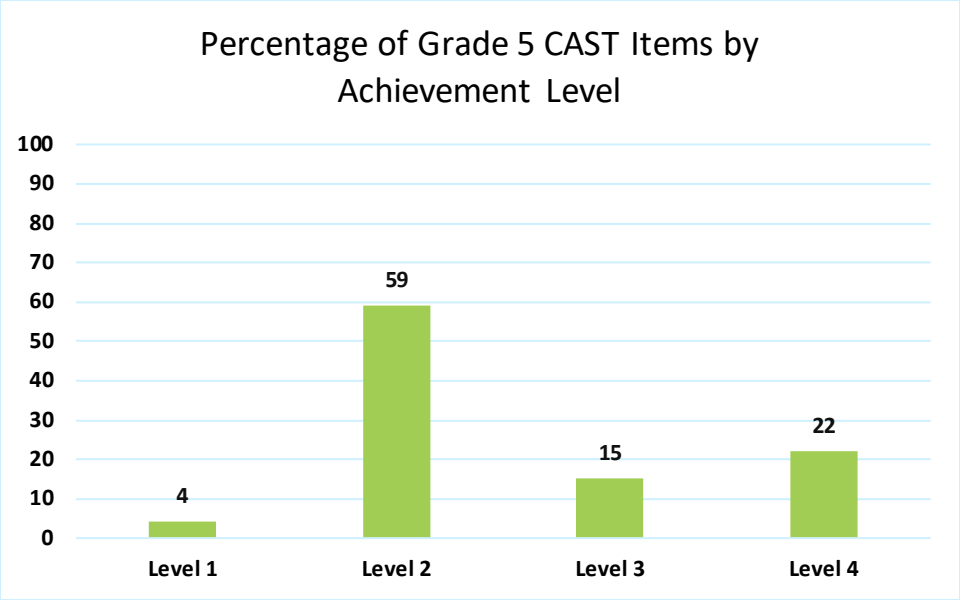

Figure D.3 Item-to-Achievement Level Classification: Grade 5 Form 3.

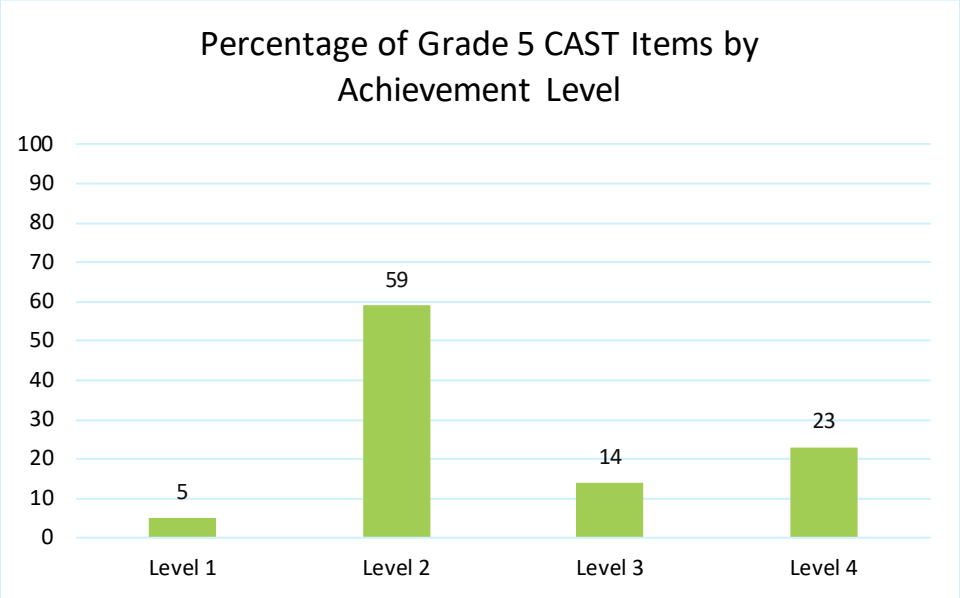

Figure D.4 Item-to-Achievement Level Classification: Grade 5 Form 4.

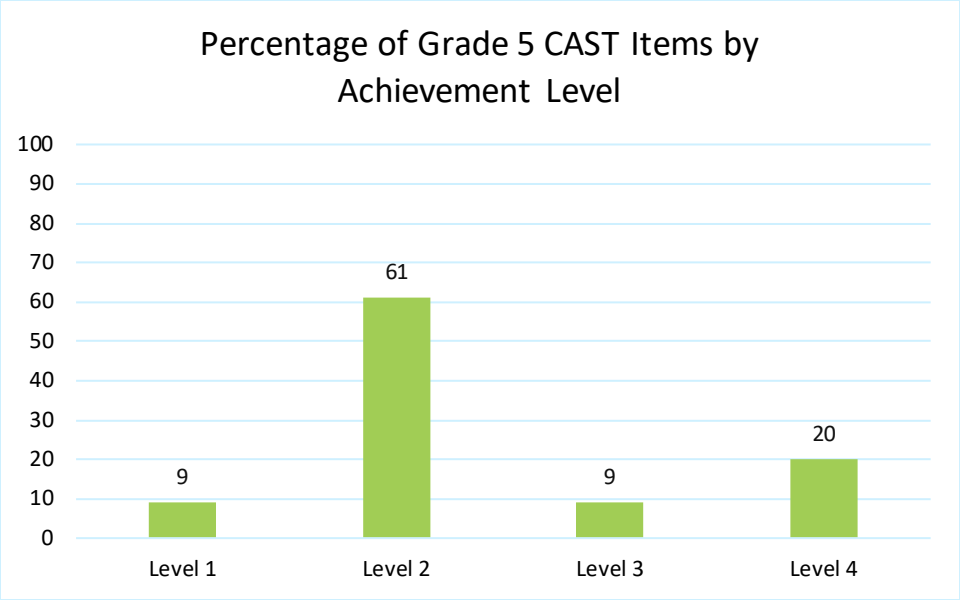

Figure D.5 Item-to-Achievement Level Classification: Grade 5 Form 5.

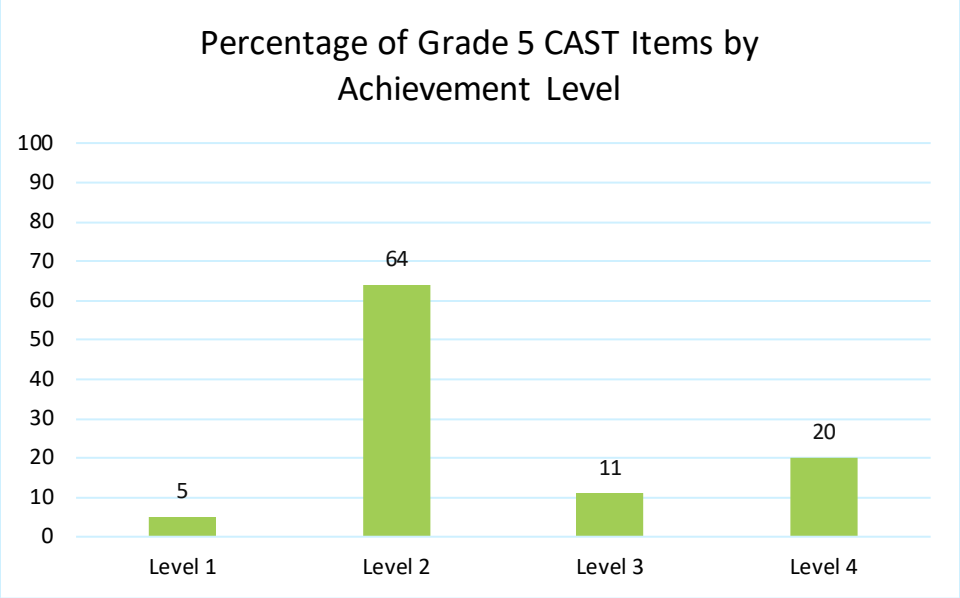

Figure D.6 Item-to-Achievement Level Classification: Grade 5 Form 6.

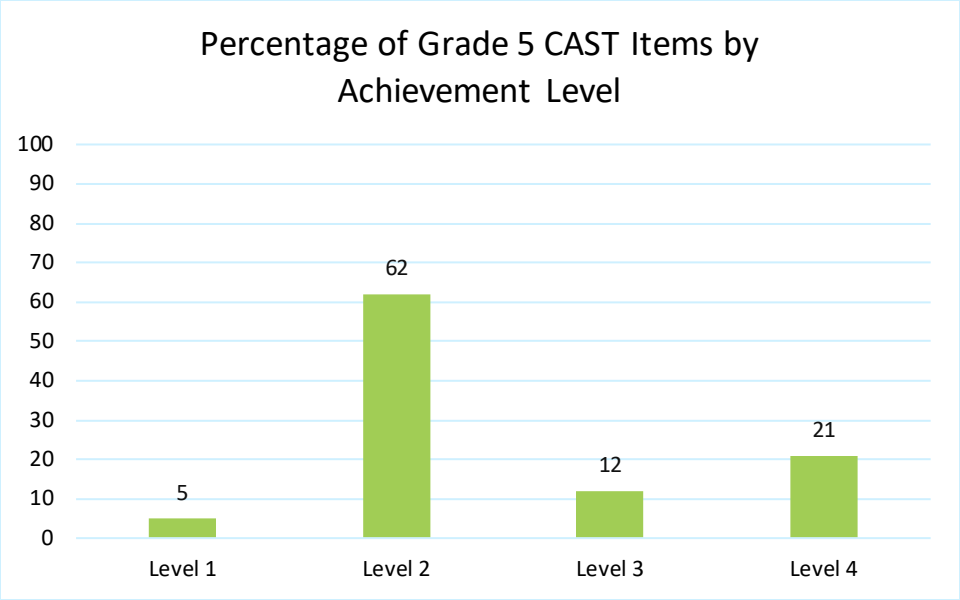

Figure D.7 Item-to-Achievement Level Classification: Grade 5 Form 7.

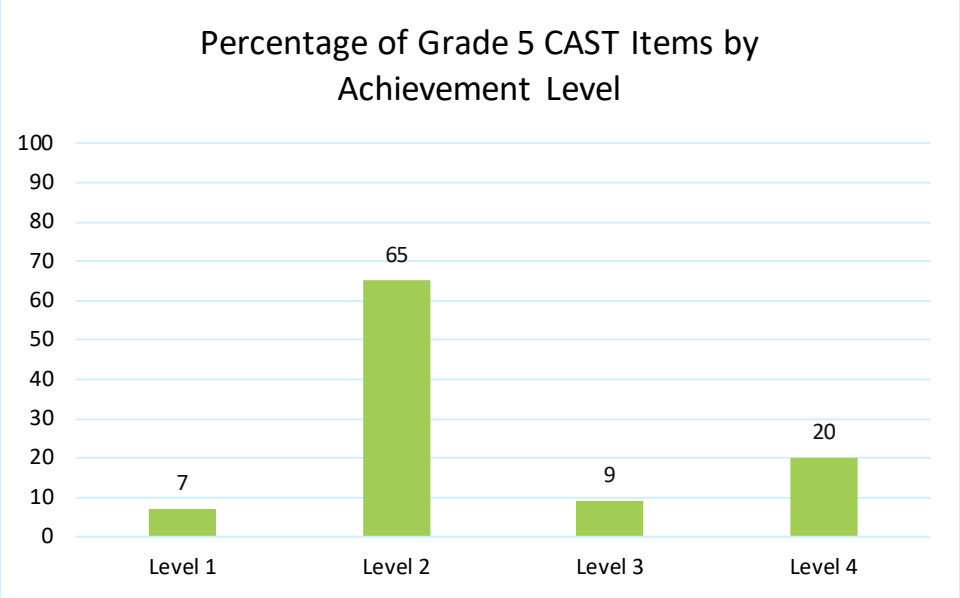

Figure D.8 Item-to-Achievement Level Classification: Grade 5 Form 8.

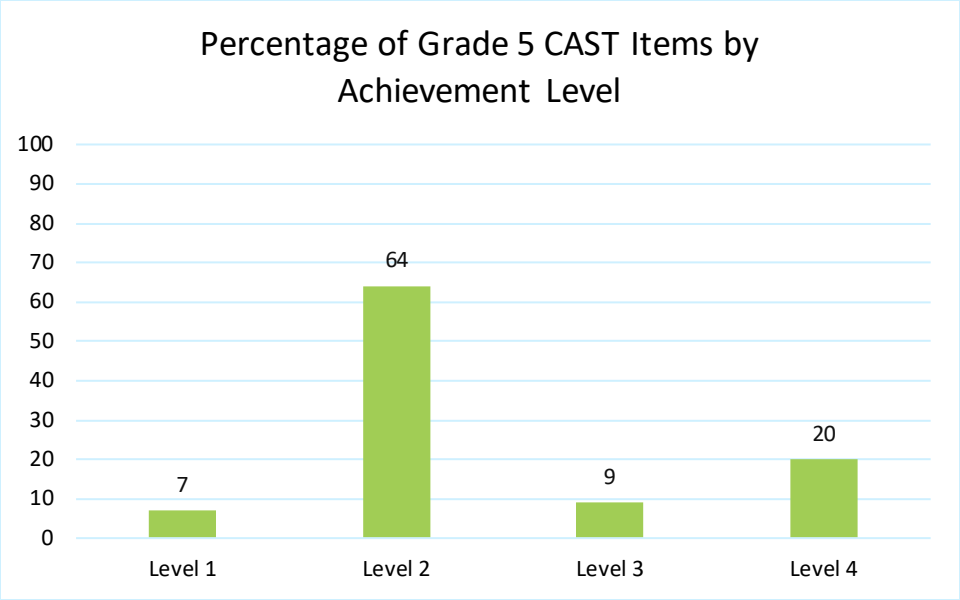

Figure D.9 Item-to-Achievement Level Classification: Grade 5 Form 9.

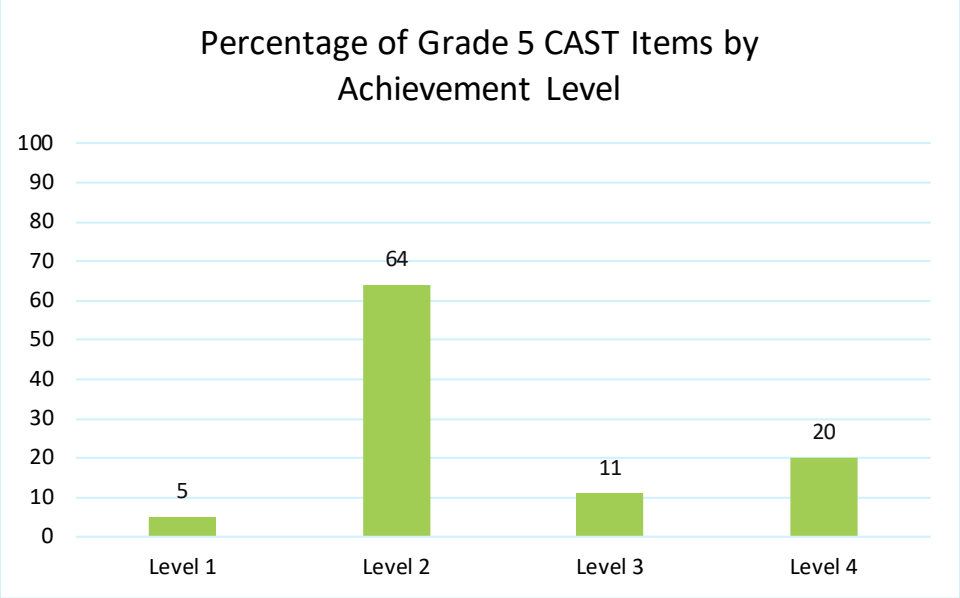

Figure D.10 Item-to-Achievement Level Classification: Grade 5 Form 10.

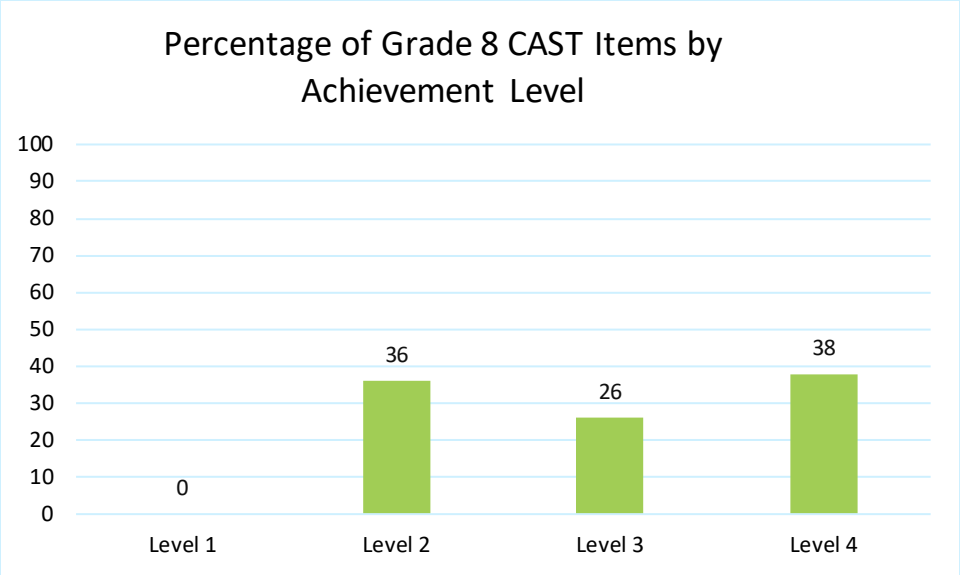

Figure D.11 Item-to-Achievement Level Classification: Grade 8 Form 1.

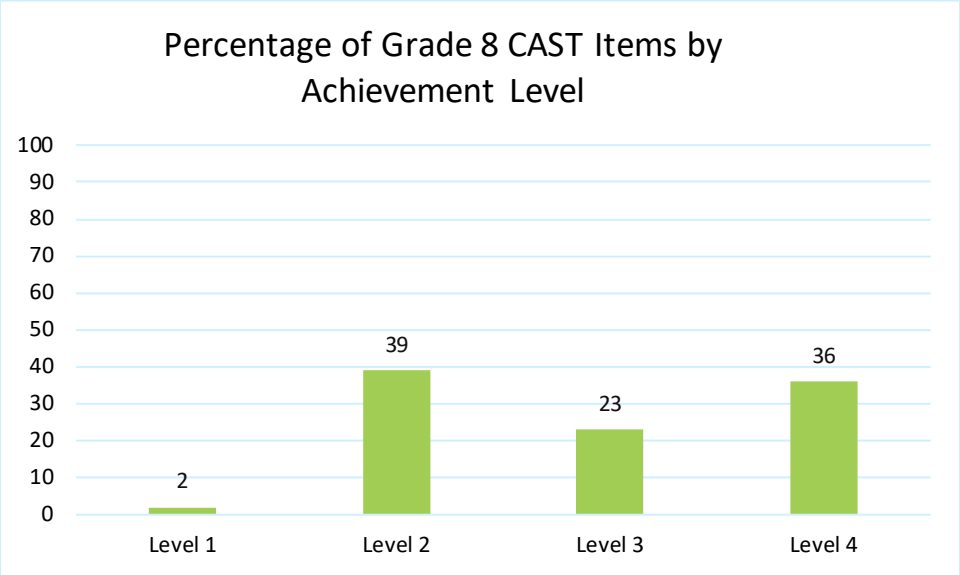

Figure D.12 Item-to-Achievement Level Classification: Grade 8 Form 2.

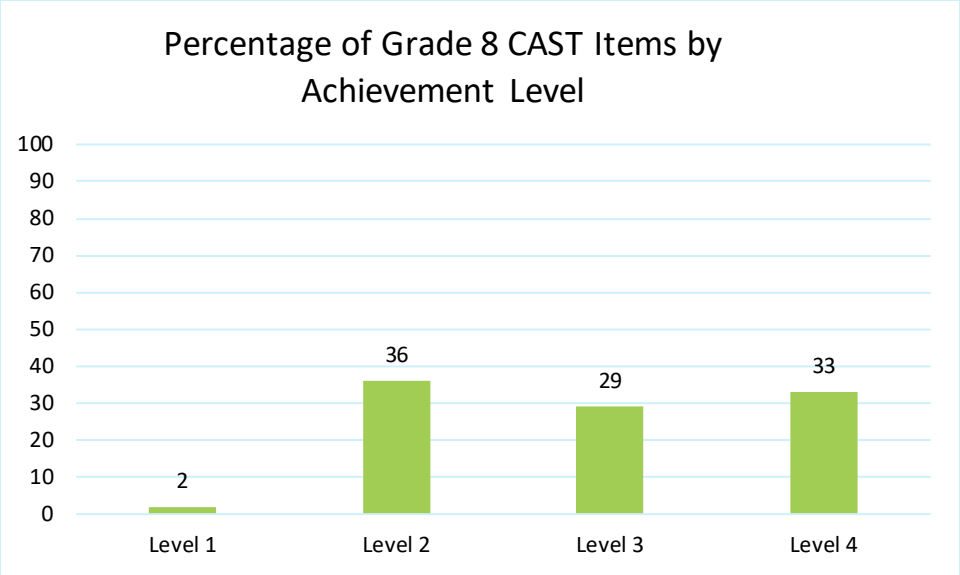

Figure D.13 Item-to-Achievement Level Classification: Grade 8 Form 3.

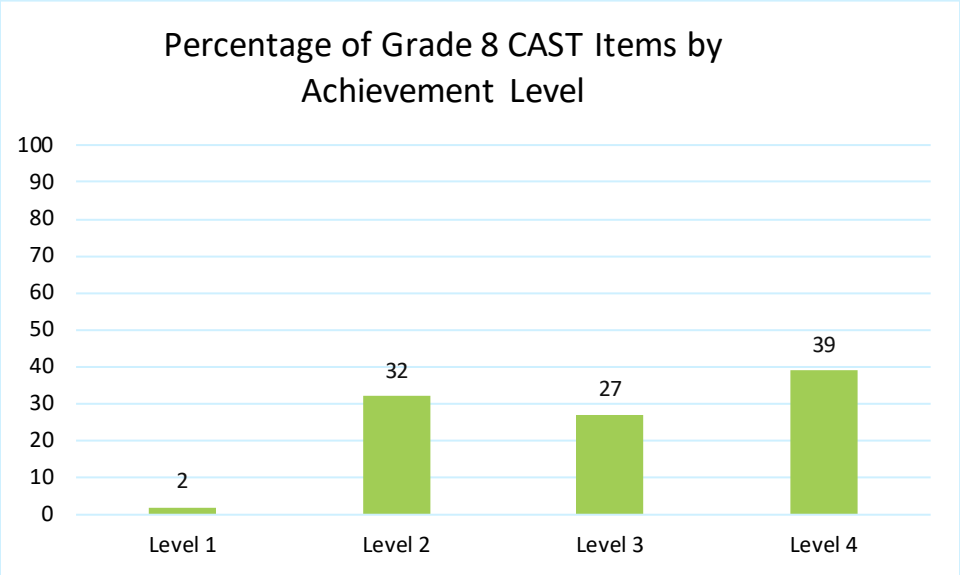

Figure D.14 Item-to-Achievement Level Classification: Grade 8 Form 4.

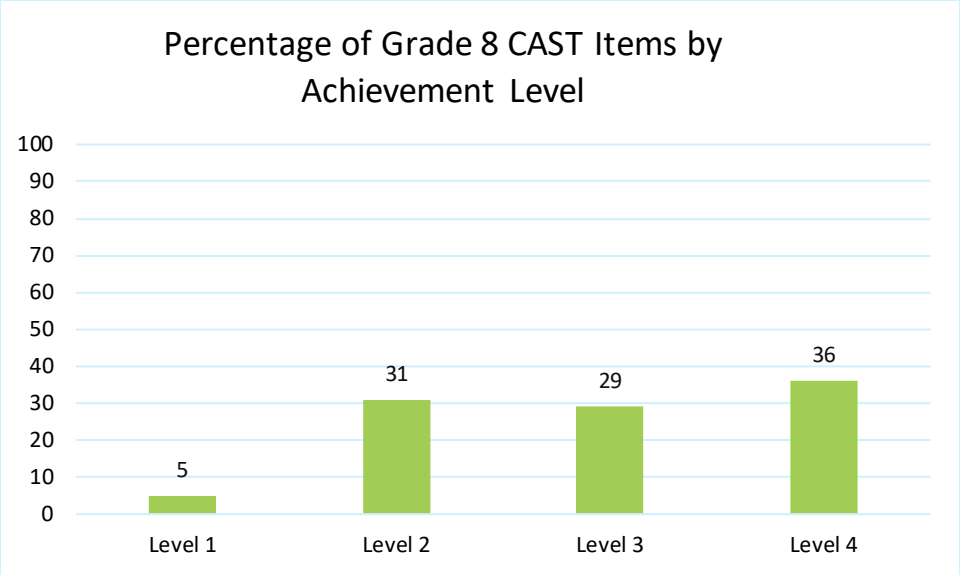

Figure D.15 Item-to-Achievement Level Classification: Grade 8 Form 5.

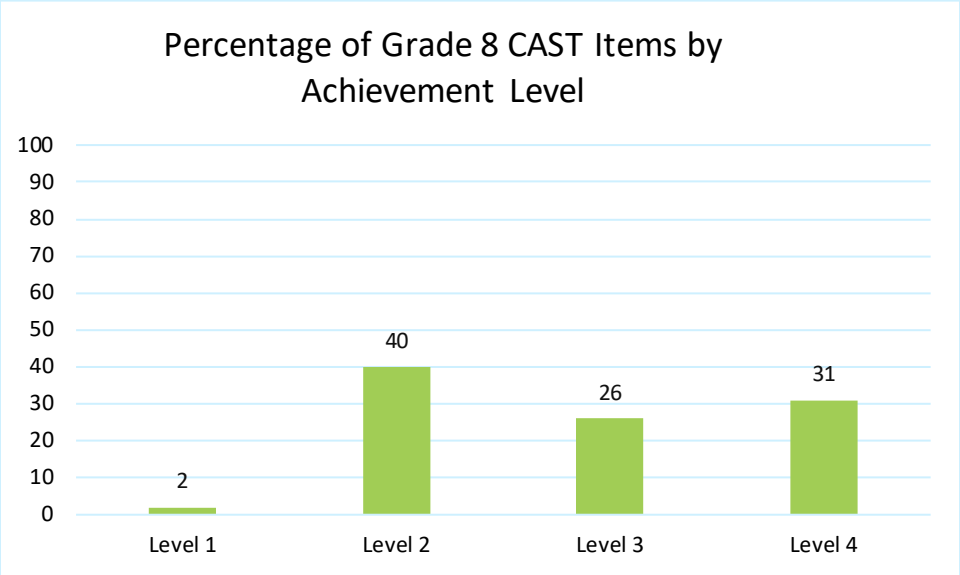

Figure D.16 Item-to-Achievement Level Classification: Grade 8 Form 6.

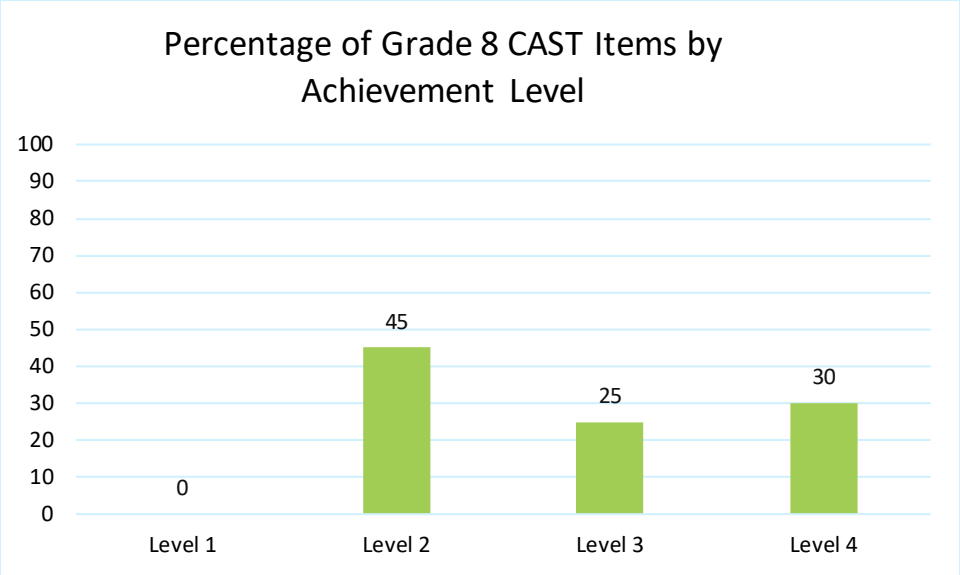

Figure D.17 Item-to-Achievement Level Classification: Grade 8 Form 7.

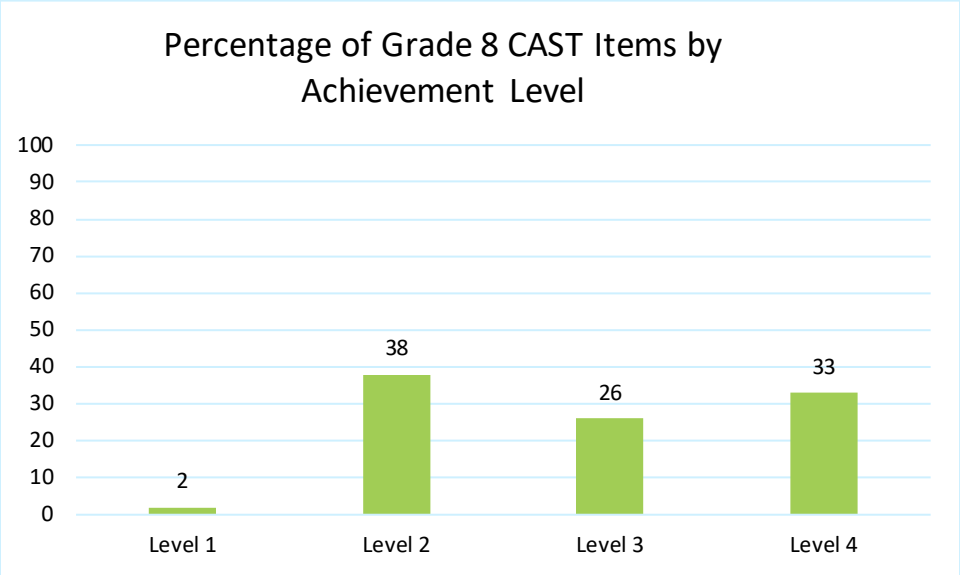

Figure D.18 Item-to-Achievement Level Classification: Grade 8 Form 8.

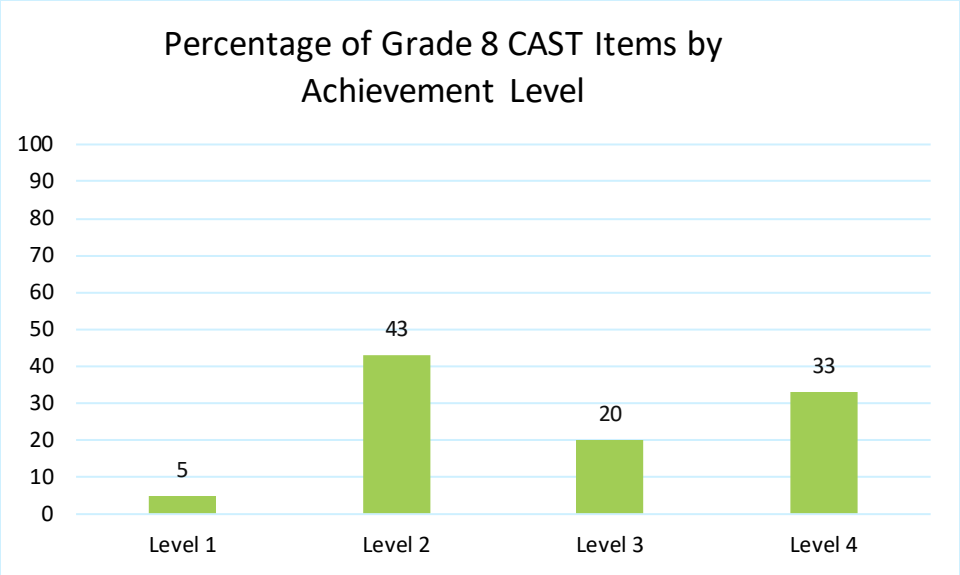

Figure D.19 Item-to-Achievement Level Classification: Grade 8 Form 9.

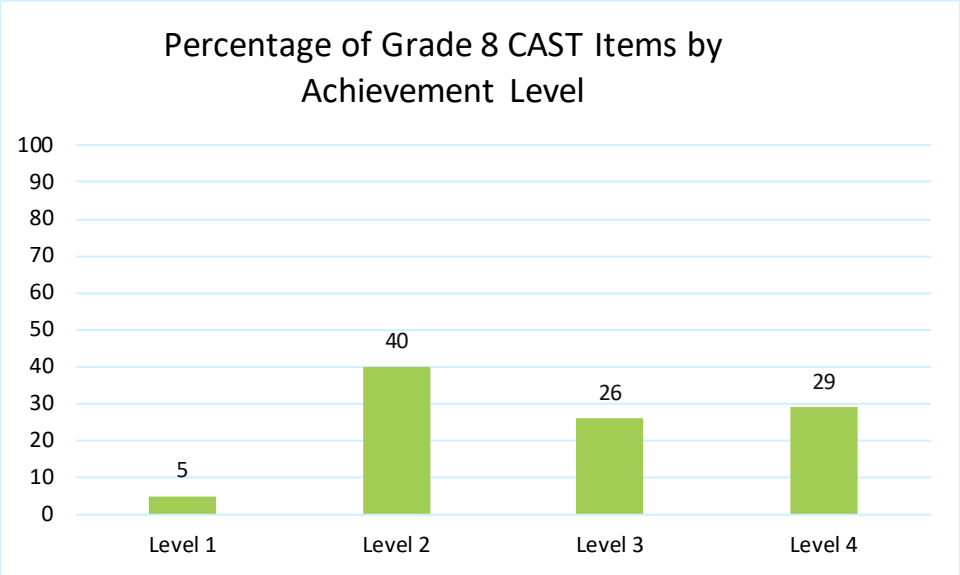

Figure D.20 Item-to-Achievement Level Classification: Grade 8 Form 10.

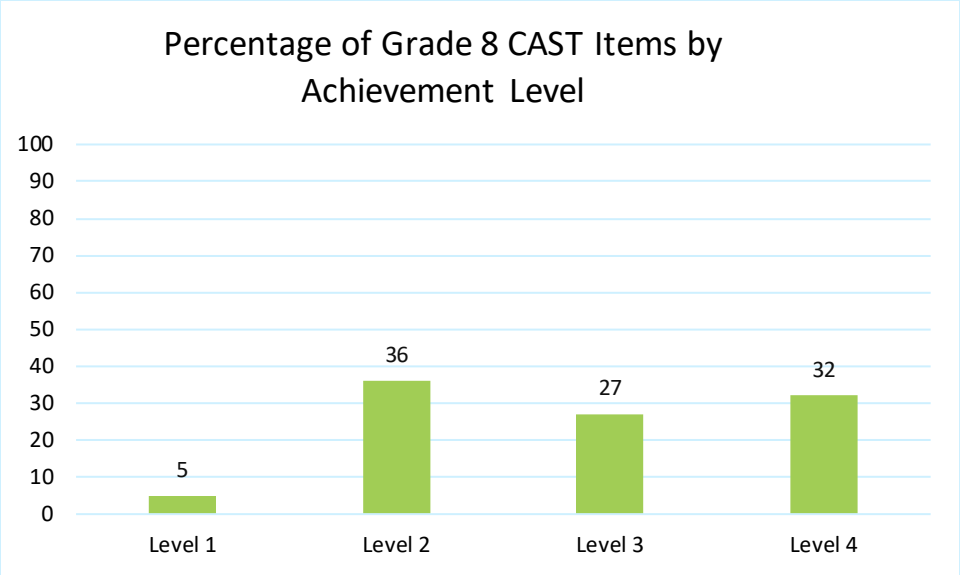

Figure D.21 Item-to-Achievement Level Classification: Grade 8 Form 11.

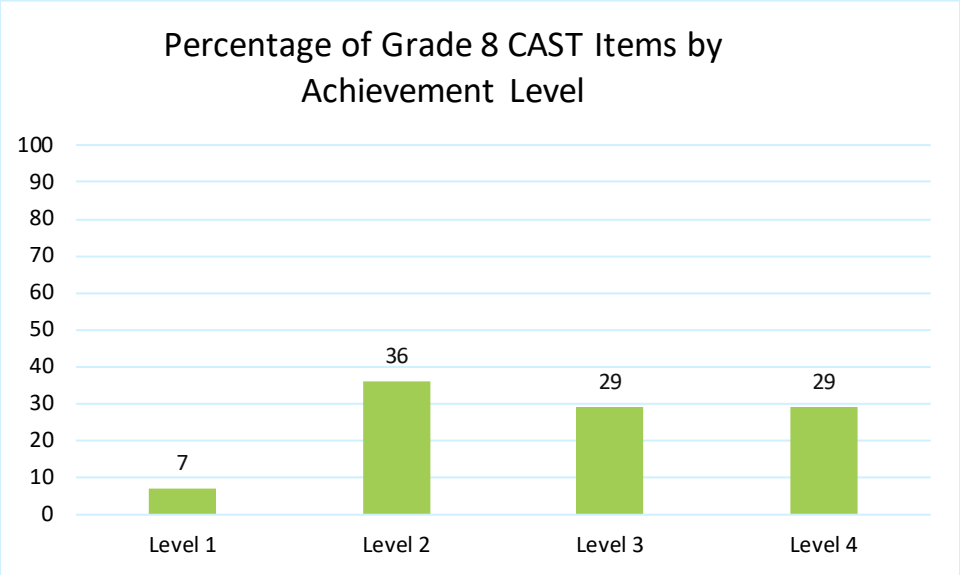

Figure D.22 Item-to-Achievement Level Classification: Grade 8 Form 12.

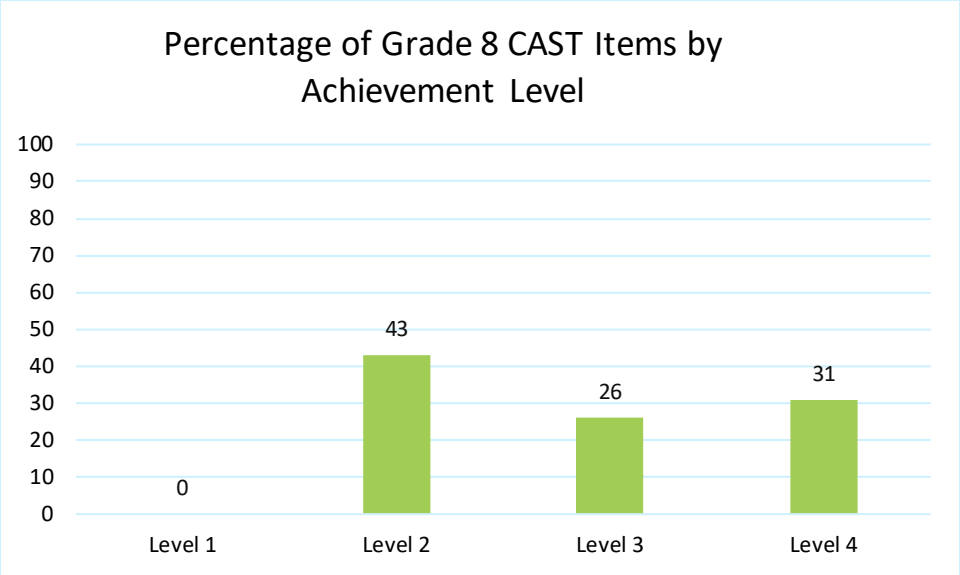

Figure D.23 Item-to-Achievement Level Classification: Grade 8 Form 13.

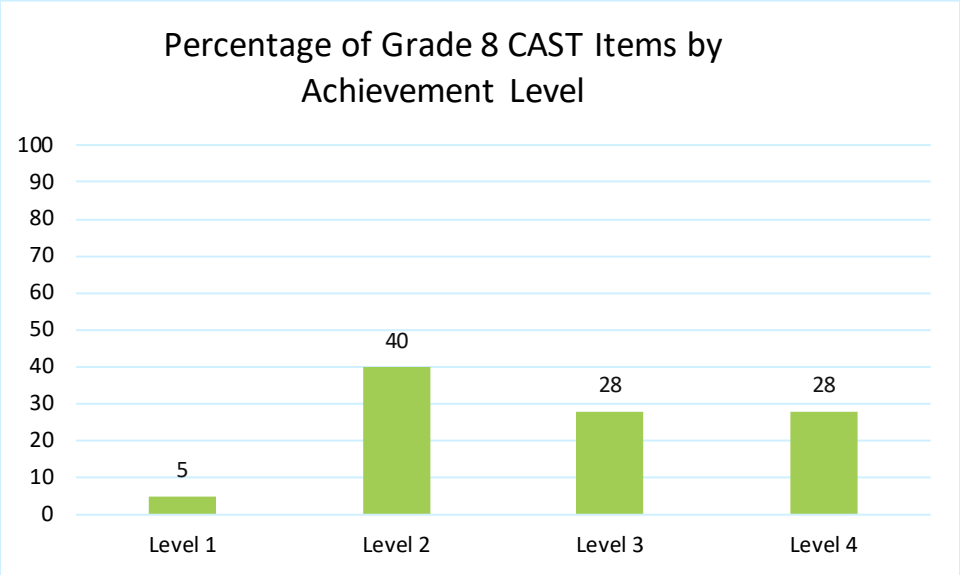

Figure D.24 Item-to-Achievement Level Classification: Grade 8 Form 14.

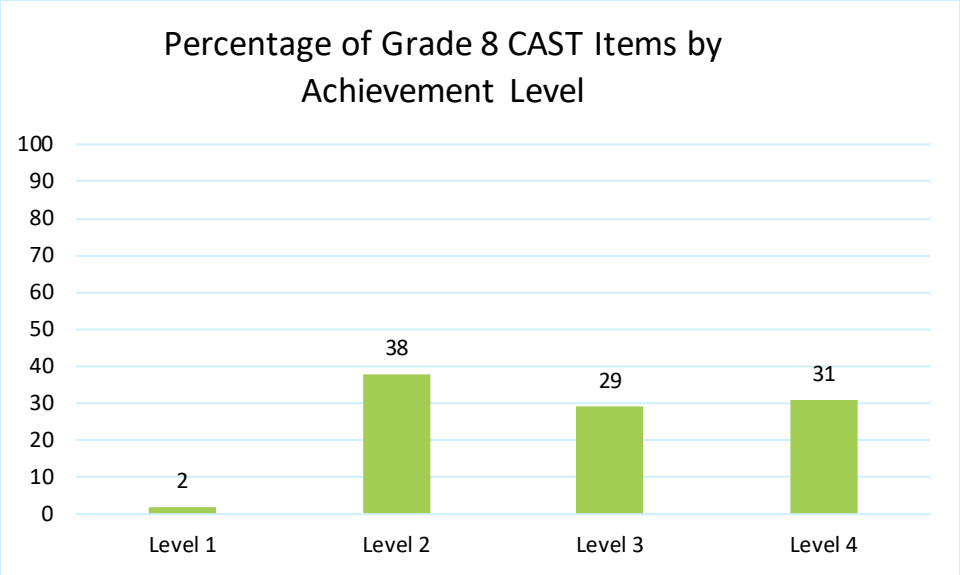

Figure D.25 Item-to-Achievement Level Classification: Grade 8 Form 15.

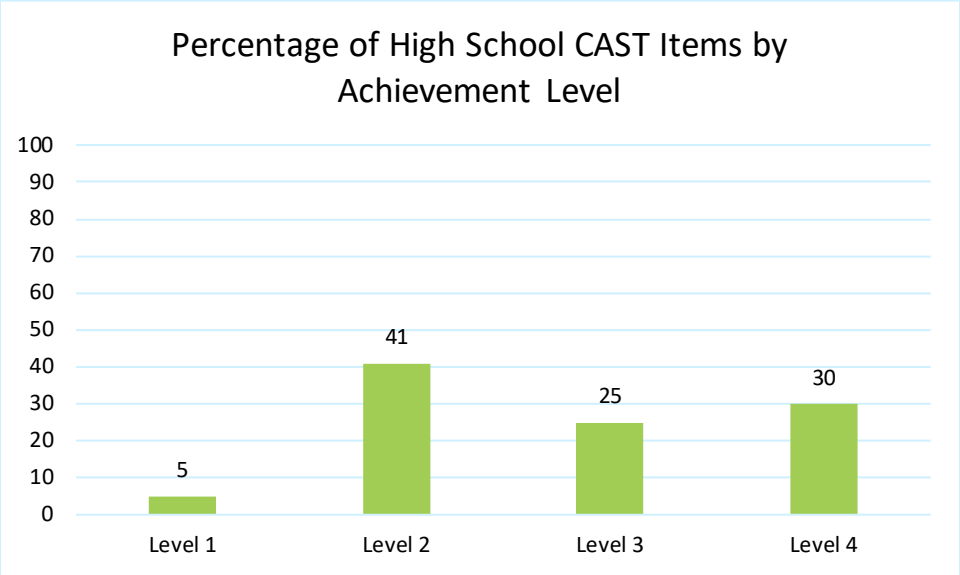

Figure D.26 Item-to-Achievement Level Classification: High School Form 1.

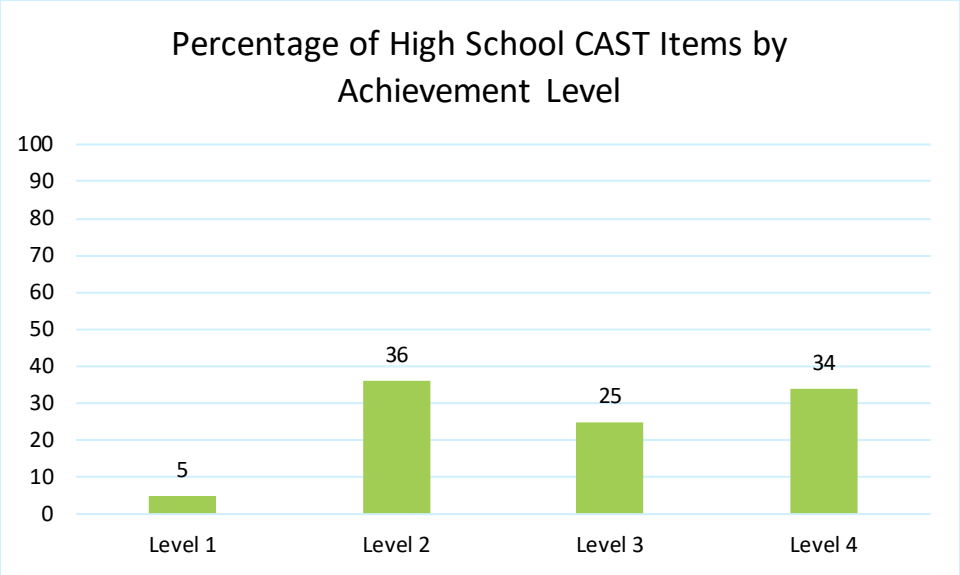

Figure D.27 Item-to-Achievement Level Classification: High School Form 2.

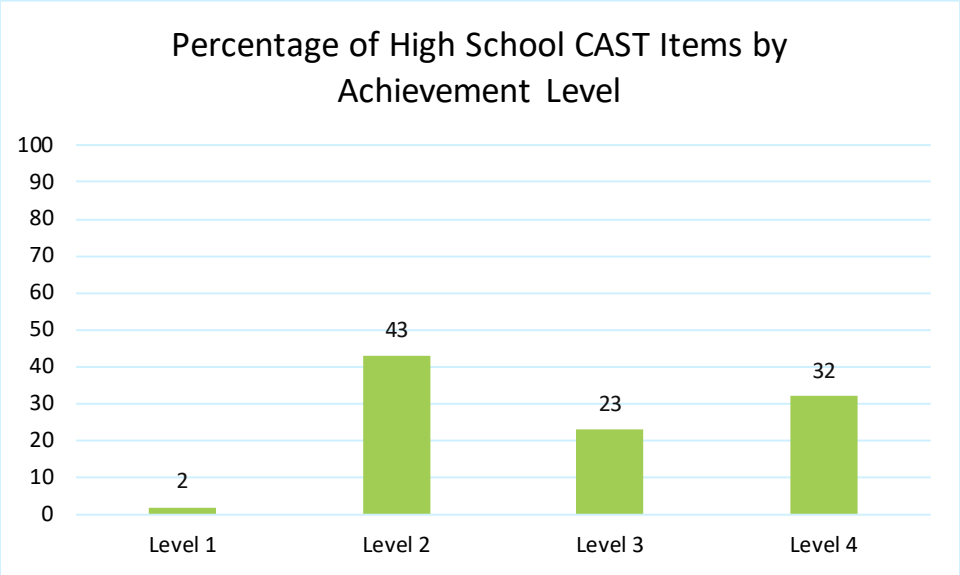

Figure D.28 Item-to-Achievement Level Classification: High School Form 3.

This page is intentionally blank.