2020 No. 046

California Assessment of Student Performance and Progress (CAASPP)

California Alternate Assessment (CAA) for Science 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	Date:	June 25, 2020
Editors:	Christa Watters Sheila Schultz Andrea Sinclair		

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

This page is intentionally blank.

California Assessment of Student Performance and Progress (CAASPP)

California Alternate Assessment (CAA) for Science Alignment Study Report

Table of Contents

List of Appendices

List of Tables

List of Figures

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 Alternate Assessment (CAA) for Science and the Science Core Content Connectors (alternate achievement standards, hereafter referred to as Science Connectors; ETS, 2018a). 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, which encompasses both contractual years of the independent evaluation, 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, 2018) 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.

This is a stand–alone report on the completed CAA for Science Alignment Study, conducted in the fall of 2019. 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). The 2019–2020 CAA for Science administration was intended to be the first operational assessment. However, on March 20, 2020, all CAASPP testing was suspended due to the Coronavirus disease (COVID-19) outbreak. This suspension of testing did not allow for a sufficient and representative number of students to complete the four performance tasks. Therefore, the 2020–2021 administration will be considered the first operational year, using the 2019–2020 test form.

Overview

The CAA for Science is designed to measure performance on the Science Connectors. The Science Connectors are derived from the performance expectations (PEs) of the California Next Generation Science Standards (CA NGSS).

The CAA for Science is not a single end–of–year summative test but instead is designed to be administered following instruction throughout the school year. Four separate sessions, three operational and one field test, are administered each year, and each session consists of one embedded performance task (PT). Each PT addresses one science domain (i.e., Life Sciences, Physical Sciences, and Earth and Space

Sciences). Administration of the CAA for Science is not tied to a typical summative assessment testing window; teachers will have discretion to administer each session when they have completed instruction on that specific domain during the school year. The students' performance on the three operational PTs will be aggregated to generate an overall science score at the conclusion of the school year. The CAA for Science is administered in grades five and eight, and once in high school. The high school assessment may be administered in grade ten, eleven, or twelve. Two Science Connectors are represented in each PT, and the five items measuring each Science Connector are expected to include two low and two medium complexity test items and one high complexity test item (numbers of score points will also vary by item). Each Science Connector has a corresponding set of five test questions prefaced by a nonscorable orienting activity designed to engage students with a science concept they were previously taught.

The first step in evaluating for CAA for Science 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 CAA for Science scores. This component of the study is described in *Chapter 2: Review of CAA for Science 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: CAA for Science 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 CAA for Science 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 CAA for Science support the claims to be made about student performance on the assessment?

2. To what extent do the test forms and test items for the CAA for Science reflect the test design and test blueprint?

3. To what extent do the CAA for Science PT items link to the Science Connectors?

4. How well do the CAA for Science PT items cover the range of cognitive complexity of the Science Connectors?

Review of CAA for Science Documentation

HumRRO researchers collected and reviewed CAA for Science design and test development materials provided by California Department of Education (CDE) and Educational Testing Service (ETS) staff, as well as information about the CAA for Science shared with the public on the CDE website. HumRRO researchers evaluated the alignment of the CAA for Science 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 eleven standards were rated as at least partially covered based on the available evidence. Most of reviewed standards (82%) were rated as mostly covered. These results indicate that the CAA for Science test design and development processes and procedures adhere to the testing standards related to alignment of assessment content to academic standards.

CAA for Science Alignment Workshop and Outcomes

This CAA for Science alignment workshop was designed to collect evidence of whether the CAA for Science produces test forms that effectively measure the content and cognitive rigor reflected in the targeted content domain and the test blueprint. During the workshop, educators with experience teaching students with significant cognitive disabilities and content expertise evaluated how well the 2018–2019 field test items selected for use as operational 2019–2020 items represent the associated content standards, the Science Connectors.

Alignment Criteria Evaluated

HumRRO developed alignment criteria intended to parallel those developed for the California Science Test (CAST). CAST alignment criteria were developed by HumRRO and reviewed by CDE's CAASPP Technical Advisory Group, the National Center for Improvement in Educational Assessment (Center for Assessment), and CDE staff. The CAST alignment criteria are presented in the *CAASPP CAST Alignment Study Report*.

HumRRO developed the following modified criteria for evaluating the CAA for Science: **Link to Standards, Depth of Knowledge (DOK) Adequacy, and Range Adequacy.** For a full description of the alignment criteria and discussion of how and why the alignment criteria were created, see chapter 3. 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 criteria were not met, we would consider this to be a signal for

concern about the link between the assessment and the intended measurement construct.

Alignment Workshop Methods

HumRRO conducted the CAA for Science Alignment Study Workshop in the Sacramento area on November 5 and 6, 2019. HumRRO worked collaboratively with the CDE to recruit and select a group of 18 educators to serve on one of three CAA for Science alignment review panels (grade five, grade eight, and high school) during the two–day workshop. Due to a last–minute cancellation, the high school panel included five educators rather than six.

Across the three panels, 15 California school districts were represented. Approximately 53 percent of panelists reported currently working as teachers while the remaining 47 percent reported working in roles such as inclusion specialist, instructional specialist, or program specialist. In addition to their current professional roles, 94 percent of panelists reported having some level of experience with the NGSS. The types of experience reported ranged from participating in trainings to presenting at NGSS rollouts. Across the three panels, all responding panelists reported having experience teaching students with mild–to–moderate and/or significant disabilities and 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 into which panelists and workshop facilitators entered ratings for the test items that were reviewed. Support materials included copies of the (a) Connectors, (b) *Directions for Administration* (DFAs), (c) item content specifications, (d) detailed workshop instructions for both panelists and facilitators, (e) details on the cognitive complexity (DOK) rating categories, and (f) debriefing and evaluation forms.

ETS created three online test “forms” solely for use during the alignment workshop (grade five, eight and high school). These forms consisted of all the CAA for Science items that were ready for operational use in 2019−2020. ETS also created accounts for HumRRO researchers and workshop panelists 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 five, grade eight, 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) Connector measured by item
    b) Focal Knowledge, Skills, and Abilities (FKSAs) or Essential Understanding (EU) measured by the item
    c) Quality of the link between the item and the identified FKSA or EU
    d) Item cognitive complexity level
    e) Rating of item accessibility
    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

The HumRRO facilitator recorded the final consensus (or majority) item ratings in a spreadsheet and saved panelists' independent ratings to a USB flash drive. 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 CAA for Science test item pools. Across the three tests, panelists' ratings of the operational items provide strong support that the CAA for Science comprises items that reflect the Science Connectors at a range of complexity levels.

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

Criterion	Grade Five	Grade Eight	High School
Links to Standards	Met	Met	Met
DOK Adequacy	Met	Met	Met
Range Adequacy	Met	Met	Met

Table ES.2 summarizes the by–form alignment criteria results for the three CAA for Science tests. Similar to the item pool results, all test form versions (simplified as "form" in tables) are comprised of items that reflect the Science Connectors at a range of complexity levels.

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

Criterion	Grade Five	Grade Eight	High School
Links to Standards	100%	100%	100%
DOK Adequacy	100%	100%	50%[a]
Range Adequacy	100%	100%	100%

[a] 100% of high school form versions at least partially met the DOK Adequacy criterion.

Overall, the alignment workshop results provide strong support that the CAA for Science system produces aligned test forms. All test form versions at all grade levels at least partially met all three a priori alignment criteria. The Depth of Knowledge Adequacy criterion was not fully met for two high school test form versions; both form versions had one item more than the 41 percent acceptability threshold for Low Complexity items. Additionally, one high school form version had one item less than the 33 percent acceptability threshold for Medium Complexity items.

Conclusions

This study combined documentation review and a workshop with content experts to evaluate alignment between the California Alternate Assessment (CAA) for Science and the Science Connectors derived from the CA NGSS. Specifically, the study addressed four research questions.

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

Review of available documentation found that the test design and test blueprint for the CAA for Science support the conclusion that the testing contractor adhered to testing standards relevant to test–to–standards alignment (see Table 2.3). Review of items that were ready for operational use in 2019–2020 supports that the CAA for Science design produces aligned test forms (see table 3.23).

Research Question 2: To what extent do the test forms and test items for the CAA for Science reflect the test design and test blueprint?

Based on expert panelists' ratings, all performance tasks in all domains were linked to at least two Science Connectors. For two grade eight form versions, panelists identified three Science Connectors measured in the Life Sciences and Physical Sciences performance tasks. For all high school form versions, panelists identified three or more Science Connectors measured in the Life Sciences and Earth and Space Sciences performance tasks. This suggests that panelists did not find the high school performance tasks to be strongly focused on particular Science Connectors.

For nearly all grade five form versions, the number of items per task rated at each cognitive complexity level matched or was adjacent to the number outlined in the test blueprint. Similarly, for grade eight, most form versions had numbers of items rated at each level that matched or were adjacent to the blueprint guidelines. Discrepancies between panelists' ratings and blueprint guidelines were somewhat more pronounced for high school form versions, with some form versions rated as having higher numbers of low complexity Physical Sciences items and some form versions having higher numbers of medium and high complexity Life Sciences items. Tables depicting these comparisons are presented in Appendix C.

Research Question 3: To what extent do the CAA for Science Performance Task (PT) items link to the Science Connectors?

For all three CAA for Science tests (grade five, grade eight, and high school), all items were judged as being aligned to a Science Connector. Similarly, all performance tasks at all three grade levels measured multiple Science Connectors, Essential Understandings (EUs), and Focal Knowledge, Skills, and Abilities (FKSAs). Regardless of the version administered, every student was tested via a form that fully met the Link to Standards and Range Adequacy criteria.

Research Question 4: How well do the CAA for Science PT items cover the range of cognitive complexity of the Science Connectors?

For all three grade level CAA for Science tests, items were rated at each of the three levels of cognitive complexity. The number of items rated at each level of cognitive complexity fell within appropriate ranges for the item pools of all three grade level tests.

For grade five and grade eight, all test form versions included appropriate numbers of items from each cognitive complexity level. Two of the four high school test form versions had one item more than the acceptability threshold that was rated at Low Complexity. One high school test form version also had one item less than the acceptability threshold that was rated at Medium Complexity.

This page is intentionally blank.

Chapter 1: Introduction

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.

The California Alternate Assessment (CAA) for Science alignment study aims to provide validity evidence for this test as a measure of science achievement for the population of students for which it was designed—students with severe cognitive disabilities. This study focuses on links between the Science Core Content Connectors (alternate achievement standards, hereafter referred to as Science Connectors), and the test forms and test items developed to assess them. The Science Connectors are derived from the performance expectations (PEs) of the California Next Generation Science Standards (CA NGSS), which also define the science construct(s) to be measured.

The CAA for Science is not a single end–of–year summative test but instead is designed to be administered following instruction throughout the school year. Four separate sessions, three operational and one field test, are administered each year, and each session consists of one embedded performance task (PT). Each PT addresses one science domain (i.e., Life Sciences, Physical Sciences, and Earth and Space Sciences). Administration of the CAA for Science is not tied to a typical summative assessment testing window; teachers will have discretion to administer each session when they have completed instruction on that specific domain during the school year. The students' performance on the three operational PTs will be aggregated to generate an overall science score at the conclusion of the school year. The CAA for Science is to be administered in grades five and eight, and once in high school. The high school assessment may be administered in grade ten, eleven, or twelve. Two Science Connectors are represented in each PT, and the PT is expected to include two low and two medium complexity test items and one high complexity test item (numbers of score points will also vary by item). Each Science Connector has a corresponding set of five test questions prefaced by a nonscorable orienting activity designed to engage students with a science concept they were previously taught.

As illustrated in Figure 1.1, the Science Connectors are disaggregated into discrete Focal Knowledge, Skills, and Abilities (FKSAs) and Essential Understandings (EUs), which are basic concepts. Test questions are written to assess the FKSAs and EUs. There are one to six FKSAs and one EU for each Science Connector. Each EU, but not all FKSAs for a Science Connector, will be assessed in a single embedded PT.

Figure 1.1. CAA for Science standards continuum. (See Appendix D for alt text.)

Because there are 20, 24, and 28 Science Connectors for grades five, eight, and high school, respectively, the full breadth of the Science Connectors cannot be represented by three PTs that measure only two Science Connectors each. The CAA for Science is expected to rotate Science Connectors from year to year, building to fuller representation of the content over time. All content from the 72 identified Science Connectors will be assessed across a five–year span.

An alignment study for an assessment with this structure must approach evidence gathering in two ways. First, it must demonstrate that the aggregation of the three sessions provides an adequate representation of the science content specified by the Science Connectors. This alignment task supports the overall score and is the key evidence required by the Every Student Succeeds Act (ESSA) under federal peer review guidance. There is only one claim for the alternate assessment for science, and that claim indicates students should demonstrate performance "across the domains." Additionally, each session should adequately represent its tested domain, even if student–level scores are not produced at the domain level. Because teachers administer the assessment one–on–one, uneven or inadequate representation could lead to unwanted instructional or curricular changes over time. To avoid such consequences, test administrators should have confidence the assessment is a fair representation of the domain. While the sessions would not be expected to generate entirely reliable score estimates, each domain–level session should represent the intended domain. Data were collected to demonstrate the extent to which the Science Connectors and associated content domains are adequately represented.

We note that any student–level results represent a sampling of the Science Connectors. The CDE will "cover" all the Science Connectors across five years. Adequate representation, as described above, means that the assessments cover the two Science Connectors per performance task they are intended to cover, and that the PTs as a group represent the intended content domains.

The research questions and methodology for this alignment study address the structure and design of the CAA for Science and the ensuing results. The detailed design of the CAA for Science Alignment Study is included in the 2018–20 CAASPP Evaluation Plan, which was presented in the publicly available *https://www.cde.ca.gov/ta/tg/ca/documents/caaspp18evalrpt.pdf*.

Research Questions

Activities conducted for the CAA for Science Alignment Study provide information to answer the following research questions:

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

2. To what extent do the test forms and test items for the CAA for Science reflect the test design and test blueprint?

3. To what extent do the CAA for Science PT items link to the Science Connectors?

4. How well do the CAA for Science PT items cover the range of cognitive complexity of the Science Connectors?

This page is intentionally blank.

Chapter 2: Review of CAA for Science Alignment Documentation

Introduction

In preparation for the alignment of the California Alternate Assessment (CAA) for Science to the Science Connectors (which are derived from the CA NGSS), HumRRO evaluated how closely the CAA for Science test alignment documentation adheres to *The Standards for Educational and Psychological Testing* (AERA, APA & NCME, 2014; hereafter referred to as the *Testing Standards*). CAA for Science is a computer–based assessment administered in grades five and eight and once in high school to students with the most significant cognitive disabilities. It was field tested in the 2018–2019 school year and was to be administered operationally for the first time in 2019–2020. CAA for Science has one overall claim, that "Students can demonstrate performance associated with the expectations described by the Science Connectors linked to the CA NGSS across the domains of Earth and Space Sciences; Life Sciences; Physical Sciences; and Engineering, Technology, and Application of Science."

Method

Our evaluation of the test design and development documentation was informed by industry best practices as outlined in the *Testing Standards*. First, HumRRO researchers identified standards from the *Testing Standards* that are directly relevant to how alignment is considered during test development. Next, we 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. Researchers compared and discussed their independent ratings to reach a final consensus rating for each standard.

Document Review

We worked in cooperation with the California Department of Education (CDE) and Educational Testing Service (ETS) to obtain documentation related to California Alternate Assessment (CAA) for Science alignment. We also searched the California Assessment of Student Performance and Progress (CAASPP) website to identify additional relevant information. A list of documents we received is presented in Appendix A.

We developed a rating scale to evaluate the degree to which the evidence for the assessment supports adherence to the *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 was found in the materials.[a]
2	Little evidence of the Standard was 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 was 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.

[a] Materials include all documents and data provided, any emails or phone calls with CDE/ETS staff, as well as information we found online.

Results

Ratings for Testing Standards

The results presented in Table 2.2 represent the analysis of our review of assessment planning and item development processes. Table 2.2 provides an overall rating for each relevant testing standard based on our review of all available information.

Table 2.2 Ratings on the Selected Testing Standards for CAA for Science 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.	• CAA for Science Item Review Meeting Slides • CAASPP Item Acceptance Criteria for Item Review Committee • Depths of Knowledge • Universal Design for Item Development • California Next Generation Science Standards Core Content Connectors for Alternate Assessments Report	4

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.	• CAA for Science Item Writer Template • CAA for Science Item Writing Workshop Checklist and Guide • CAA for Science Blueprint • CAA for Science Item Metadata • CAA for Science Form Planners • Cognitive Complexity Definitions • CAA for Science Prioritized Connectors Memorandum • Development Plan for the California Next Generation Science Standards Alternate Core Content Connectors • California Next Generation Science Standards Core Content Connectors for Alternate Assessments Report	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.	• CAA for Science Blueprint • CA NGSS Core Content Connectors for Alternate Assessments • CAA Prioritized Connectors Memorandum • Development Plan for the California Next Generation Science Standards Alternate Core Content Connectors • California Next Generation Science Standards Core Content Connectors for Alternate Assessments Report	3
Standard 2.3. For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant indices of reliability/precision should be reported.	• CAASPP 2019–2020 Student Score Report Mockup – CAA for Science	5

Table 2.2 (cont.)

Standard	Supporting Documentation	Standard Rating
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.	• CAA for Science Item Review Meeting Slides • CAA for Science Item Writing Workshop Guidelines • CAA for Science Item Writing Workshop Checklist and Guide • CAASPP Item Acceptance Criteria for Item Review Committee • Depths of Knowledge • Universal Design for Item Development • CAA for Science Test Specifications	5
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.	• *Directions for Administration* (DFA) • CAA for Science Administration Planning Guides • Alternate Assessment IEP Team Guidance (https://www.cde.ca.gov/ta/tg/ca/caaiepteamrev.asp) • Test Examiner Survey Extract • CAASPP Matrix One • CAA ELA, Mathematics, and Science Test Examiner Tutorial	5

Table 2.2 (cont.)

Standard	Supporting Documentation	Standard Rating
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.	• 2016 Science Framework for California Public Schools (https://www.cde.ca.gov/ci/sc /cf/ casciencefframework2016.asp) • CAA for Science Blueprint • CAA for Science Statistical Specifications for Assessmen Development • CAA for Science Item Review Meeting Slides • CAA for Science Item Writing Workshop Guidelines • CAA for Science Item Writing Workshop Checklist and Guide • CAASPP Item Acceptance Criteria for Item Review Committee • Depths of Knowledge • Universal Design for Item Development • CAA for Science Prioritized Connectors Memorandum • Development Plan for the California Next Generation Science Standards Alternate Core Content Connectors • California Next Generation Science Standards Core Content Connectors for Alternate Assessments Report	4

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).	• 2016 Science Framework for California Public Schools • Alternate Assessment IEP Team Guidance (https://www.cde.ca.gov/ta/tg/ca/caaiepteamrev.asp) • CA NGSS Core Content Connectors for Alternate Assessments • CAA for Science Blueprint • CAA for Science Statistical Specifications for Assessment Development • CAA for Science Item Review Meeting Slides • 2019–2020 CAA for Science Administration Planning Guide • Alternate Assessment IEP Team Guidance	4
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.	• CAA Prioritized Connectors Memorandum • Development Plan for the California Next Generation Science Standards Alternate Core Content Connectors • California Next Generation Science Standards Core Content Connectors for Alternate Assessments Report	3

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.	• CAA for Science Blueprint • Proposed Design for California's Next Generation Science Standards General Summative Assessments • CAA for Science Item Review Meeting Slides	4
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.	• CA NGSS Core Content Connectors for Alternate Assessments • CAA for Science Blueprint • California Alternate Assessment General Item Specifications • CAA for Science Planning Guides • CAA for Science 2019–2020 Form Planners • CAA for Science 2020 Administration Options (*Visio–CAAS 2020 Admin Options V5.pdf*) • CAA for Science Cognitive Complexity Definitions	4

Rationales for Ratings for Testing Standards

Next, we discuss the rationales for our ratings in Table 2.2 and explain to what extent the Standard was met. We also provide suggestions for further strengthening adherence 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.

Development and prioritization of the Science Connectors included review by experts external to the testing program. The report outlined the qualifications and experience required to serve as a reviewer.

Item review training and support materials provide no information on the expertise of reviewers. Item review training slides provide some description of the consensus process, but there is no information about any adjudication that happened when consensus could not be reached. Levels of rater agreement should be reported when independent ratings are a component of the review process.

Standard setting is scheduled for summer 2020, but no information about how expert judges will be selected is available at this time.

All items are multiple choice (MC) or multi–select (MS), so there are no raters involved in scoring.

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.

The template used for generating test items required item writers to enter the elements of the content domain (Performance Expectation [PE]/Science Connectors, Focal Knowledge, Skills and Abilities [FKSAs]/Essential Understandings [EUs]) that the item was intended to measure. The CAA for Science Cognitive Complexity Definitions document guides item development for the CAA for Science. The document sets parameters for stimuli and individualization, item types, item max points, vocabulary, and readability. The complexity level guidelines provide guidance on the type of cognitive processes, length of descriptions, and number/complexity of options for each of the three levels of possible item complexity.

The CAA for Science Blueprint (ETS, 2018b) describes that each of the three science domains will be assessed equally by using one performance task for each; each PT relates to two Science Connectors and includes 10 items (4 easy, 4 moderate, 2 difficult). Over five years, all Science Connectors will be sampled. The blueprint also states that "The Science Connectors were developed in a multistage process, beginning in fall 2015 and involving California educators, Educational Testing Service assessment experts, and edCount, a firm that provides consultation on the quality of assessment systems. The goal was to represent the CA NGSS with appropriate levels of challenge and rigor for the targeted population of students."

The CAA Prioritized Connectors Memorandum process identified tandem teams of independent stakeholders who selected which Science Connectors they believed applied to daily life functions and skills of the target population and how accessible the content of these Science Connectors was to these students. The stakeholders then prioritized the elementary and middle school Science Connectors based on how they aligned with the high school Science Connectors. The individuals who participated in the process of identifying Science Connectors that provide appropriate levels of challenge and rigor to students with significant cognitive disabilities were chosen for their special education expertise and their experience working with students with significant cognitive disabilities. The participants were selected in accordance with the qualifications cited in the *Development Plan for the California Next Generation Science Standards Alternate Core Content Connectors* and the *California Next Generation Science Standards Core Content Connectors for Alternate Assessments Report.*

Administration Planning Guides note that for Science Connectors with more than one Focal Knowledge, Skills, and Abilities (FKSA), assessment of all FKSAs will occur over multiple years. The Overview of the 2019–2020 CAA for Science Administration and the CAA for Science 2020 Administration Options further illustrate how the content will be covered.

Each form planner documents which item aligns with which Connector, FKSA, or Essential Understanding (EU). Similarly, metadata files present the content alignment of each item, as intended by item writers (and presumably agreed upon by reviewers).

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.

Because the CAA for Science is machine–scored, human scoring processes are not part of the argument for validity. Thus, that portion of this Standard is not relevant to the CAA for Science.

The CA NGSS incorporate considerable evidence showing how students develop from naive to sophisticated understanding of science concepts. When identifying the content on which CAA for Science scores would be based, test developers made considerable effort to identify a subset of content that it would be feasible to assess for this student population while also maintaining fidelity to the full range of content.

The Science Connectors for Alternate Assessments Report provides some theoretical support for the development of the Science Connectors through literature review on the conceptual model of learning and understanding among students with significant cognitive disabilities. Empirical evidence to support that the test measures the cognitive operations of test takers (e.g., Focal knowledge, skills, and abilities) is needed.

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

The Student Score Report mockup presents a performance level and an overall scale score, along with a state average scale score for comparison. It does not appear that error bands will be presented around reported scale scores; however, the final report format has not yet been determined.

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.

Item writing guidelines outline topics that should be avoided due to their potential to be bothersome to students.

The Item Writer Checklist and Guide reminds item writers to use "simple, direct, and unambiguous language" and to "not use idiomatic phrases." It also includes lists of *Potential Sources of Bias and Sensitivity* and *Guiding Questions to Use to Check for Bias and/or Sensitivity.*

Training and support materials for item reviewers include *Guidelines for Language,* such as checking that items do not use words, phrases, names, or terms that may be culturally insensitive or unfamiliar to people of any given culture. Training slides include focus on Universal Design principles (including "subject matter is clearly defined so that all irrelevant cognitive, sensory, emotional, and physical barriers are removed") and bias and sensitivity issues.

The CAA for Science Test Specifications include guidelines for removing items that demonstrate differential item functioning (DIF).

The item complexity guidelines in the document specify the appropriate length of sentences and options, as well as the types of cognitive activities appropriate to the level.

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.

The *Directions for Administration* (DFAs) provide alternative text for students with visual impairment and options for individualization. Each DFA states: "Take advantage of options for individualization if offered in this DFA and remember that test examiners can always use accommodations and resources to best meet a student's individual needs, as documented in the student's IEP. Please note that all items may be individualized based upon the student's IEP". For every DFA, there is an alternative text for a student with visual impairment.

The DFAs and Planning Guides note to test administrators that "all items may be individualized based upon the student's IEP." The Examiner's Tutorial provides detailed instructions about how to identify appropriate accessibility supports, including how to request unlisted resources. The tutorial reinforces that only approved supports should be used during test administration.

The Alternate Assessment IEP Team Guidance available on the CDE website states that "Through individualization, test examiners can use materials that the student is most comfortable using to access the science concept. Individualization does not change the standard being assessed." The Examiner's Tutorial provides examples of appropriate individualization, as outlined in the DFA.

The Test Examiner Survey asks assessment administrators if they provided individualization and if students took advantage of the individualization offered.

CAASPP Matrix One provides a detailed list of embedded and non–embedded accommodations that are and are not allowed for different tests, including the CAA.

The Test Examiner's Tutorial includes instructions for accessibility resources including how to request unlisted resources for approval and discusses individualization in some detail.

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.

Documentation on the development and prioritization of the Science Connectors demonstrates that the test was designed to assess content that is appropriate for the intended population.

The Science Framework provides an overview of the intended uses of science assessment scores in California. It includes information about the test development process at a high level (pilot, followed by field test, followed by operational), how long test events should last, how they will be delivered (via computer), at which grade levels students will be tested, and score use (accountability reporting at both the state and local levels).

The CAA for Science blueprint indicates that the development committee used the California NGSS to develop the Science Connectors, which were adapted to the needs of the test population from the California NGSS. The CAA Prioritized Connectors Memorandum outlines in brief the process taken to ensure that the Science Connectors were (a) considered important by stakeholders and ETS and (b) relevant and accessible to the target population. The blueprint also identifies the content that will be covered over a five–year period.

The CDE website (https://www.cde.ca.gov/ta/tg/ca/caascience.asp) states that, "the purpose of the CAA for Science is to measure what students know and can do in science. These measures help identify and address gaps in knowledge or skills early so students can receive the support they need." Additional evidence is needed to support that the test scores can be used to identify gaps in knowledge and skills.

The CAA for Science Test Specifications includes some description of the test design (number of performance tasks and associated points), as well as item specifications such as ranges for item difficulty, point biserial correlation, and DIF.

Item development and review documentation includes evidence of considering content validity, fairness, and potential sources of construct irrelevant variance.

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 CAA for Science Test Specifications are focused on statistical specifications. However, the 2016 Science Framework for California Public Schools provides information about the purpose and uses of the test (it also describes how scores from different types of assessments can be used, which is essentially a rationale), and the construct(s) measured. The Alternate Assessment IEP Team Guidance lists eligibility criteria for students to participate in the CAAs.

The Science Connectors show the relationship between the larger NGSS domain and the content domain tested by the CAA. However, more documentation about how the Science Connectors were determined would be helpful.

The CAA for Science blueprint states the purpose of the assessment is “to assess whether a student can demonstrate performance associated with expectations outlined in the Science Connectors.” The blueprint identifies the examinee population as "students with the most significant cognitive disabilities." The blueprint and test specifications provide information about what content is included and how it is weighted (points per performance task), which informs score interpretation.

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.

Although the CAA for Science test specifications were not reviewed by external experts, the content of the test as outlined in the Science Connectors was evaluated by external

experts for its appropriateness for the intended student population. The report outlines the qualifications that were required of experts who participated in the review, though it does not include a description of the final set of reviewers. The report also summarizes the purpose, process, and results of the review, though an explicit rationale to support the selection of the final set of Science Connectors is not included. In addition, an external assessment expert reviewed the process used to prioritize the Science Connectors.

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 CAA for Science blueprint demonstrates how each science domain will be measured via the associated performance tasks. Each performance task is intended to measure two Science Connectors via 10 items at varying levels of cognitive complexity. The blueprint also describes how the content to be tested (as identified through the process of identifying Science Connectors) will be covered over a 5–year period.

The California Science Test (CAST) Evidence–Centered Design White Paper and the Proposed Design for California's Next Generation Science Standards General Summative Assessments provide background about the content domain. The latter discusses "CA NGSS Assessments," which include both CAST and CAA. The CAA is aligned to the CA NGSS through the process of identifying Science Connectors based on the Performance Expectations (PE) outlined in the CA NGSS.

The Item Review Meeting training slides indicate that CDE and ETS convened a panel of experts to review CAA items. Experts were tasked with judging whether items (a) aligned with Science Connectors, (b) were written clearly/concisely, (c) met assigned complexity levels, (d) followed Universal Design (UD) principles, (e) were engaging and appropriate for the population, and (f) were free from bias/offensive material.

No evidence about scoring criteria is available at the time of this report, as standard setting for the CAA for Science was postponed to Summer 2021 due to COVID–19.

More empirical evidence to support that the test represents the content domain, such as confirmatory factor analysis (CFA), would be helpful.

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 CA NGSS Core Content Connectors for Alternate Assessments Report outlines the Science Connectors and Focal Knowledge, Skills, and Abilities/Essential Understandings [FKSAs/EUs]) that are to be tested (i.e., target domain).

The CAA for Science blueprint identifies the content that will be covered over a five–year period (tested domain). The blueprint also outlines the number of items at each cognitive complexity level within each tested science domain. The blueprint references both the full CA NGSS and the Science Connectors.

The Item Specifications provide detail about how test items are designed to reflect the different levels of cognitive complexity. The Cognitive Complexity Definitions document includes a description of the types of cognitive activities required by each level of cognitive complexity.

The CAA for Science Planning Guides indicate which Science Connectors and, and within each Science Connector, which FKSAs are being assessed for a given year. These documents, in concert with other documents listing all the Science Connectors/FKSAs, indicate what is and is not being tested.

Each CAA for Science Form Planner documents which item aligns with which Science Connector, FKSA, or EU. The CAA for Science 2020 Administration Options (*Visio–CAAS 2020 Admin Options V5.pdf*) document provides a visual of what each grade–band's form versions consist of, breaking down the task and test form version, as well as the anchor and field test Connector measured on each test form version.

Evidence is needed of the extent to which the test elicits the processes reflected in the target domains (e.g., Focal knowledge, skills, and abilities). Item-level and test-level data may provide this evidence as the assessment becomes operational. Prior to operational data analyses, this evidence is often gathered through cognitive labs or similar studies.

Summary and Discussion

Table 2.3 summarizes the results of the review across the selected standards. All eleven standards were rated as at least partially covered based on the available evidence. Most of reviewed standards (82%) were rated as mostly covered. It is important to note that these ratings are based on evidence that is available during the first operational year of the CAA for Science. Collecting evidence of test validity is an ongoing process; evidence noted as missing during this review may become available as the test is administered in subsequent years.

These results indicate that the CAA for Science test design and development processes and procedures largely 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 form versions are adequately aligned to the Science Connectors.

Table 2.3 Summary of Document Review Results

Number of Standards Rated	Percent Fully Covered	Percent Mostly Covered	Percent Partially Covered
11	36	46	18

This page is intentionally blank.

Chapter 3: CAA for Science Alignment Workshop and Outcomes

Introduction

HumRRO conducted a workshop to examine the content alignment of the California Alternate Assessment (CAA) for Science. This alignment study provides evidence regarding the extent to which the CAA for Science produces test forms that effectively measure what is intended. It does so by evaluating how well the 2019 test items sample the construct represented by the associated content standards, the Science Connectors. 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 CAA for Science items to the Science Connectors. 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 three major alignment criteria.

CAA for Science Alignment Criteria

HumRRO developed alignment criteria intended to parallel those developed for the California Science Test (CAST). CAST alignment criteria were developed by HumRRO and reviewed by CDE's CAASPP Technical Advisory Group, the National Center for Improvement in Educational Assessment (Center for Assessment), and CDE staff. The CAST alignment criteria are presented in the *CAASPP CAST Alignment Study Report*.

Table 3.1 summarizes the criteria we used to evaluate alignment of the CAA items to the Science Connectors. 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 criteria were not met, we would consider this to be a signal for concern about the link between the assessment and the intended measurement construct.

Table 3.1 CAA Alignment Criteria

Criteria	Description
Link to Standards	HumRRO calculates the percentage of items panelists rate as directly and clearly matched to a Science Connector. The criterion is defined as fully met if 90% of items are matched to a Science Connector.
DOK Adequacy	HumRRO calculates the percentage of items panelists rate as reflecting each of three DOK levels (Low, Medium, and High; see Appendix B for definitions) is calculated. The criterion is considered fully met if 25−41% of items are rated at Low Complexity, 33−50% of items are rated at Medium Complexity, and 17−33% of items are rated at High Complexity.
Range Adequacy	HumRRO calculates the percentage of items panelists rate as directly and clearly matched to a Focal Knowledge, Skills, and Abilities (FKSA) or Essential Understanding (EU). The criterion is fully met if each performance task is aligned to at least two Science Connectors and at least two EUs and one FKSA.

Method

We evaluated the alignment criteria based on item ratings and professional judgments collected during the 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 18 educators to serve on three CAA alignment review panels (six educators each for grade five, grade eight, and high school). Due to a last–minute cancellation, the high school panel included five educators rather than six. The three panels represented a total of 15 California school districts.

Approximately 53 percent of panelists reported currently working as teachers while the remaining 47 percent reported working in roles such as inclusion specialist, instructional specialist, or program specialist. In addition to their current professional roles, 94 percent of panelists reported having some level of experience with the NGSS. The types of experience reported ranged from participating in trainings to presenting at NGSS rollouts. Across the three panels, all responding panelists reported having experience teaching students with mild–to–moderate and/or significant disabilities and 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 Demographics of CAA for Science Alignment Panelists

Panel	# of Panelists	# of Districts	% Female/ % Male	% Hispanic/ % Non–Hispanic	Years of Experience Mean (SD)
Grade Five	6	6	83/17	17/83	15.50 (9.14)
Grade Eight	6	6	67/33	33/67	15.00 (9.13)
High School	5	5	60/40	60/40	11.60 (6.35)

Workshop Logistics

HumRRO conducted a two–day CAA for Science Alignment Study Workshop in the Sacramento area on November 5–6, 2019. During the workshop, panels of educators evaluated how well each CAA for Science item assessed the Science Connectors. 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 blueprint), information about the California approach to classifying item cognitive complexity, and item metadata.

HumRRO developed several data collection tools and adapted other existing materials to support the data collection process. Data collection tools included electronic spreadsheets into which panelists and workshop facilitators entered ratings for the test items that were reviewed. Support materials included copies of the (a) Science Connectors, (b) *Directions for Administration* (DFAs), (c) item specifications, (d) detailed workshop instructions for both panelists and facilitators, (e) details on the cognitive complexity (DOK) rating categories, and (f) debriefing and evaluation forms. Examples of workshop materials are presented as appendices to this report.

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 panels and received detailed training on the data collection processes and procedures. Those processes and procedures are described in the following section.

Workshop Processes and Procedures

During the workshop, each panelist had a workstation that contained two laptops and a binder containing alignment materials and supporting documentation; electronic

versions of materials were also provided. Operational test items were accessed on one laptop via an online secure platform set up by ETS. Electronic rating forms were saved onto panelists' other laptops. Panelists were given access to paper and electronic copies of the Science Connectors, DFAs, and item specifications. They were also given paper copies of detailed alignment process steps, descriptions of the cognitive complexity (DOK) categories, and descriptions of item accessibility.

After the panel–specific training presentation by the HumRRO facilitator, each panel engaged in a calibration activity that involved the first 1–3 items. Panelists accessed the items electronically and independently rated each item. Panelists discussed their independent ratings and engaged in consensus discussion until they reached agreement on the final item ratings. Once panelists had a clear understanding of the rating process and a common understanding of the rating categories, they proceeded to independently complete ratings for the remaining operational items.

The panelists rated only a small group of operational items at a time. For each group of items, panelists first independently rated each item. Then panelists discussed their ratings for an item; reviewed the item metadata; engaged in more discussion; then reached their final consensus/majority rating for the item before moving on to the next. Once we recorded consensus/majority ratings for one group of items, the panel moved on to the next group and repeated this process. Panelists generated item ratings via the following steps:

1. Panelists reviewed test items independently and assigned ratings of:
    a. Connector measured by item
    b. Focal Knowledge, Skills, and Abilities (FKSA) or Essential Understanding (EU) measured by the item
    c. Quality of the link between the item and the identified FKSA or EU
    d. Item cognitive complexity level
    e. Rating of item accessibility
    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.

After all panelists completed their independent ratings, the HumRRO facilitator managed the group discussion and encouraged the panelists to share their ratings. The facilitator polled the panelists about their ratings and asked them to share their rationale when independent ratings differed. Panelists were trained to retain their independent ratings unless they realized they had made a coding error or the group's discussion revealed an error in their thinking about an item and/or the Science Connectors.

Following an initial discussion, the HumRRO facilitator projected the item metadata for panelists to review, discuss, and use to reach consensus on the final rating for each item. If panelists could not reach true consensus, the facilitator recorded the rating that reflected 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. The debriefing form was designed to give panelists the opportunity to provide their individual, qualitative perspective on the overall alignment of the CAA for Science test. The evaluation survey elicited feedback about the quality of the workshop processes and procedures. Table 3.3 summarizes the results of the workshop evaluation survey.

Table 3.3 CAA for Science Alignment Evaluation Survey Results

Evaluative Statement	% Strongly Dis–agree	% Dis–agree	% Some–what Dis–agree	% Some–what Agree	% Agree	% Strongly Agree
The training presentation in the large group provided useful information about the CAA for Science and HumRRO's alignment method.	0.0	0.0	0.0	6.3	43.8	50.0
After the additional training in my small group, I felt prepared to review and rate test items.	0.0	0.0	0.0	6.3	18.8	75.0
HumRRO staff seemed knowledgeable of the CAA for Science and alignment steps.	0.0	0.0	0.0	0.0	12.5	87.5
The Panelist Instruction document was clear, understandable, and useful in performing the alignment steps.	0.0	0.0	0.0	0.0	18.8	81.3
The Excel file was understandable and relatively easy to use to enter item ratings.	0.0	0.0	0.0	6.3	25.0	68.8
The process for reaching consensus ratings was conducted fairly.	0.0	0.0	0.0	0.0	18.8	81.3

Results

This section summarizes the data/information collected during the alignment workshop. The results are presented for each grade level separately, and separately for the item pool and by test form version.

Grade Five

The grade five science operational test items were evaluated on three alignment criteria: (1) Link to Standards, (2) DOK Adequacy, and (3) Range Adequacy.

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 Connector. The criterion is considered Acceptable if at least 90 percent of items are matched to a Connector.

Table 3.4 shows that panelists matched 100 percent of the 60 grade five CAA for Science items to a Connector. 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[a]

Sub–criterion	Percentage	Acceptable?
Items matched to a Connector	100	Yes

[a] n = 60 items

Table 3.5 presents the results from a by–form analysis of the items rated as directly and clearly matched to a Connector. Across the four grade five form versions, panelists rated 100 percent of items as measuring a Connector. Criterion 1 is met for all grade five test form versions.

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 Connector	100–100	4 of 4

Criterion 2: DOK Adequacy

This criterion is evaluated based on the percentage of items rated by panelists as reflecting each of the cognitive complexity levels (Low, Medium, High; see Appendix B for definitions). The criterion is considered Acceptable if 25−41 percent of items are rated at Low Complexity, 33−50 percent of items are rated at Medium Complexity, and 17−33 percent of items are rated at High Complexity.

Table 3.6 shows that 35 percent of grade five CAA for Science items were rated at Level 1, 38 percent were rated at Level 2, and 27 percent were rated at 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[a]

DOK level	Percentage	Acceptable?
Level 1– Low	35	Yes
Level 2– Medium	38	Yes
Level 3– High	27	Yes

[a] n = 60

Table 3.7 presents the results from a by–form analysis of the same ratings. Across the four grade five science form versions, 37–40 percent of items were rated Level 1, 37–40 percent were rated Level 2, and 23 percent were rated at Level 3. Criterion 2 DOK Adequacy is met for all grade five test form versions.

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– Low	37–40	4 of 4
Level 2– Medium	37–40	4 of 4
Level 3– High	23–23	4 of 4

Criterion 3: Range Adequacy

This criterion is evaluated based on the Focal Knowledge, Skills, and Abilities (FKSAs) and Essential Understanding (EU) that panelists rated as directly and clearly matched to test items. The criterion is considered Acceptable if the ten items composing each performance task (PT) are aligned to at least two Science Connectors and at least two EUs and one FKSA.

Table 3.8 shows that the ten items composing each of the grade five PTs were aligned to at least two Science Connectors and at least two EUs and one FKSA. Criterion 3 Range Adequacy is met for the grade five item pool.

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

Sub–criterion	Percentage	Acceptable?
PTs aligned to at least two Science Connectors and at least two EUs and one FKSA	100	Yes

Table 3.9 presents the results from a by–form analysis of the same ratings. Across the form versions, the items composing each grade five PTs were aligned to at least two Science Connectors and at least two EUs and one FKSA. Criterion 3 Range Adequacy is met for all grade five test form versions.

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

Sub–criterion	Range of Percentages	Number of Forms Meeting Criterion
PTs aligned to at least two Science Connectors and at least two EUs and one FKSA	100–100	4 of 4

Grade Eight

This section summarizes results for the grade eight science assessment. The grade eight science operational test items were evaluated on the same alignment criteria described in the *grade five* section.

Criterion 1: Link to Standards

Table 3.10 shows that panelists matched 100 percent of the 59 grade eight CAA for Science items to a Connector. Based on this, Criterion 1, Link to Standards, is met for the grade eight items.

Table 3.10 Grade Eight Item Pool Results for Criterion 1: Link to Standards[a]

Sub–criterion	%	Acceptable?
Items matched to a Connector	100	Yes

[a] n = 59

Table 3.11 presents the results from a by–form analysis of the same ratings. Across the four form versions, 100 percent of items on the form were rated as measuring a Connector. Criterion 1, Link to Standards, is met for all grade eight test form versions.

Table 3.11 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 Connector	100–100	4 of 4

Criterion 2: DOK Adequacy

Table 3.12 shows that of the 59 grade eight CAA for Science items, 32 percent were rated at Level 1, 36 percent were rated at Level 2, and 32 percent were rated at Level 3. Criterion 2, DOK Adequacy, is met for the grade eight items.

Table 3.12 Grade Eight Item Pool Results for Criterion 2: DOK Adequacy[a]

DOK level	Percentage	Acceptable?
Level 1– Low	32	Yes
Level 2– Medium	36	Yes
Level 3– High	32	Yes

[a] n = 59

Table 3.13 presents the results from a by–form analysis of the same ratings. Across the grade eight science form versions, panelists rated 30–37 percent of items at Level 1, 33–43 percent at Level 2, and 27–30 percent at Level 3. Criterion 2, DOK Adequacy, is met for all grade eight test form versions.

Table 3.13 Grade Eight Test Form Results for Criterion 2: DOK Adequacy

DOK level	Range of Percentages	Number of Forms Meeting Criterion
Level 1– Low	30–37	4 of 4
Level 2– Medium	33–43	4 of 4
Level 3– High	27–30	4 of 4

Criterion 3: Range Adequacy

Table 3.14 shows that the ten items composing each grade eight PTs were aligned to at least two Science Connectors and at least two Essential Understandings (EUs) and one Focal Knowledge, Skills, and Ability (FKSAs), Criterion 3, Range Adequacy, is met for the grade eight item pool.

Table 3.14 Grade Eight Item Pool Results for Criterion 3: Range Adequacy

Sub–criterion	Percentage	Acceptable?
PTs aligned to at least two Science Connectors and at least two EUs and one FKSA	100	Yes

Table 3.15 presents the results from a by–form analysis of the same ratings. Across the form versions, the ten items composing PTs were aligned to at least two Science Connectors and at least two EUs and one FKSA. Criterion 3, Range Adequacy, is met for all grade eight test form versions.

Table 3.15 Grade Eight Test Form Results for Criterion 3: Range Adequacy

Sub–criterion	Range of Percentages	Number of Forms Meeting Criterion
PTs aligned to at least two Science Connectors and at least two EUs and one FKSA	100–100	4 of 4

High School

This section summarizes results for the high school assessment. The high school operational test items were evaluated on the same alignment criteria described in the G*rade Five* and G*rade Eight* sections.

Criterion 1: Link to Standards

Table 3.16 shows that panelists matched 100 percent of the 59 high school CAA for Science items to a Connector. Based on this, Criterion 1, Link to Standards, is met for the high school items.

Table 3.16 High School Item Pool Results for Criterion 1: Link to Standards[a]

Sub–criterion	Percentage	Acceptable?
Items matched to a Connector	100	Yes

[a] n = 59

Table 3.17 presents the results from a by–form analysis of the same ratings. Across form versions, 100 percent of items on the form were rated as measuring a Connector. Criterion 1, Link to Standards, is met for all high school test form versions.

Table 3.17 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 Connector	100–100	4 of 4

Criterion 2: DOK Adequacy

Table 3.18 shows that 37 percent of high school CAA for Science items were rated at Level 1, 41 percent were rated at Level 2, and 22 percent were rated at Level 3. Criterion 2, DOK Adequacy, is met for the high school items.

Table 3.18 High School Item Pool Results for Criterion 2: DOK Adequacy[a]

DOK level	Percentage	Acceptable?
Level 1– Low	37	Yes
Level 2– Medium	41	Yes
Level 3– High	22	Yes

[a] n = 59

Table 3.19 presents the results from a by–form analysis of the same ratings. On two of the four form versions, 43 percent of the items were rated as Low complexity, which was one item more than the acceptability threshold. For one high school form version, 30% of the items were rated at Level 2, which was one item less than the acceptability

threshold. All high school test form versions had 17–27 percent of the items rated at Level 3. Criterion 2, DOK Adequacy, is fully met for two form versions, and is partially met for the two other high school test form versions.

Table 3.19 High School Test Form Results for Criterion 2: DOK Adequacy

DOK level	Range of Percentages	Number of Forms Meeting Criterion
Level 1– Low	37–43	2 of 4
Level 2– Medium	30–47	3 of 4
Level 3– High	17–27	4 of 4

Criterion 3: Range Adequacy

Table 3.20 shows that the ten items composing each high school PT aligned to at least two Science Connectors and at least two Essential Understandings (EUs) and one Focal Knowledge, Skill, and Ability (FKSAs). Criterion 3, Range Adequacy, is met for the high school items.

Table 3.20 High School Item Pool Results for Criterion 3: Range Adequacy

Sub–criterion	Percentage	Acceptable?
PTs aligned to at least two Science Connectors and at least two EUs and one FKSA	100	Yes

Table 3.21 presents the results from a by–form analysis of the same ratings. Across the form versions, the items composing each PT aligned to at least two Science Connectors and at least two EUs and one FKSA. Criterion 3, Range Adequacy, is met for all high school test form versions.

Table 3.21 High School Test Form Results for Criterion 3: Range Adequacy

Sub–criterion	Range of Percentages	Number of Forms Meeting Criterion
PTs aligned to at least two Science Connectors and at least two EUs and one FKSA	100–100	4 of 4

Summary and Discussion

Summary Results

Table 3.22 summarizes the alignment criteria results for the three CAA for Science test item pools. Across the three tests, panelists' ratings of the operational items provide strong support that the CAA for Science comprises items that reflect the Science Connectors at a range of complexity levels.

Table 3.22 Summary of Item Pool Results by Criterion and Grade Level

Criterion	Grade Five	Grade Eight	High School
Links to Standards	Met	Met	Met
DOK Adequacy	Met	Met	Met
Range Adequacy	Met	Met	Met

Table 3.23 summarizes the by–form alignment criteria results for the three CAA for Science tests. Similar to the item pool results, all test form versions are comprised of items that reflect the Science Connectors at a range of complexity levels. At the high school level, two of the four form versions only partially met the DOK adequacy criterion.

Table 3.23 Percent of Grade Level Forms Fully Meeting Each Criterion

Criterion	Grade Five	Grade Eight	High School
Links to Standards	100%	100%	100%
DOK Adequacy	100%	100%	50%[a]
Range Adequacy	100%	100%	100%

[a] 100% of high school form versions at least partially met the DOK Adequacy criterion.

Discussion

Overall, the alignment workshop results provide strong support that the CAA for Science system produces aligned test forms. All test form versions at all grade levels at least partially met all three *a priori* alignment criteria. The Depth of Knowledge Adequacy criterion was not fully met for two high school test form versions; both form versions had one item more than the 41 percent acceptability threshold for Low Complexity items. Additionally, one high school form version had one item less than the 33 percent acceptability threshold for Medium Complexity items.

Because criteria ratings are based on panelists' ratings of the items rather than on the intent of the item developers, it is informative to consider the level of agreement between the alignment workshop panelists and item developers. Table 3.24 presents the percent 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 provided by ETS. If the consensus rating and metadata matched, then agreement was noted. The values in Table 3.24 reflect the percent of items for which there was agreement between the consensus rating and metadata for Connector, Focal Knowledge, Skills, and Abilities/Essential Understandings (FKSAs/EUs), and cognitive complexity level. It is important to note the final consensus ratings were recorded after the panel had viewed and discussed the item metadata, so levels of agreement reflect the panels' ratings after considering the metadata. The highest level of disagreement was observed on the high school FKSA/EU rating. For 41 percent of the high school items, the panel disagreed with the FKSA/EU alignment reported in the item metadata. For approximately half of these disagreements, the panelists agreed on the Connector that the item was measuring but they disagreed on whether the item was measuring an EU or an FKSA.

Table 3.24 Percent of Agreement with Item Metadata

CAA for Science Item Pool Grade Level	Connector	FKSA/EU	Cognitive Complexity
Grade Five (n=60)	100%	88%	87%
Grade Eight (n=59)	95%	93%	81%
High School (n=59)	78%	59%	68%

This page is intentionally blank.

Chapter 4: Conclusions

This study combined documentation review and a workshop with content experts to evaluate alignment between the California Alternate Assessment (CAA) for Science and the Science Connectors derived from the CA NGSS. Specifically, the study addressed four 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 blueprint for the CAA for Science support the claims to be made about student performance on the assessment?

Review of available documentation found that the test design and test blueprint for the CAA for Science support the conclusion that the testing contractor adhered to testing standards relevant to test–to–standards alignment (see Table 2.3). Review of test form versions composed of items that were ready for operational use in 2019–2020 support that the CAA for Science design produces aligned test forms (see table 3.23).

Research Question 2: To what extent do the test forms and test items for the CAA for Science reflect the test design and test blueprint?

Based on expert panelists' ratings, all performance tasks in all domains were linked to at least two Science Connectors. For two grade eight form versions, panelists identified three Science Connectors measured in the Life Sciences and Physical Sciences performance tasks. For all high school form versions, panelists identified three or more Science Connectors measured in the Life Sciences and Earth and Space Sciences performance tasks. This suggests that panelists did not find the high school performance tasks to be strongly focused on particular Science Connectors.

For nearly all grade five form versions, the number of items per task rated at each cognitive complexity level matched or was adjacent to the number outlined in the test blueprint. Similarly, for grade eight, most form versions had numbers of items rated at each level that matched or were adjacent to the blueprint guidelines. Discrepancies between panelists' ratings and blueprint guidelines were somewhat more pronounced for high school form versions, with some form versions rated as having higher numbers of low complexity Physical Sciences items and some form versions having higher numbers of medium and high complexity Life Sciences items. Tables depicting these comparisons are presented in Appendix C.

Research Question 3: To what extent do the CAA for Science Performance Task (PT) items link to the Science Connectors?

For all three CAA for Science tests (grade five, grade eight, and high school), all items were judged as being aligned to a Science Connector. Similarly, all performance tasks at all three grade levels measured multiple Science Connectors, Essential Understandings (EUs), and Focal Knowledge, Skills, and Abilities (FKSAs). Regardless of the version administered, every student was tested via a form that fully met the Link to Standards and Range Adequacy criteria.

Research Question 4: How well do the CAA for Science PT items cover the range of cognitive complexity of the Science Connectors?

For all three grade level CAA for Science tests, items were rated at each of the three levels of cognitive complexity. The number of items rated at each level of cognitive complexity fell within appropriate ranges for the item pools of all three grade level tests.

For grade five and grade eight, all test form versions contained appropriate numbers of items from each cognitive complexity level. Two of the four high school test form versions had one item more than the acceptability threshold that was rated at Low Complexity. One high school test form version also had one item less than the acceptability threshold that was rated at Medium Complexity.

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.

California Department of Education. (2018). Appendix A: Theory of Action for CAASPP and the Smarter Balanced Assessment System. In California Assessment of Student Performance and Progress (CAASPP) 2018 independent evaluation report.

Dickinson, E. R., Thacker, A. A., & Hardoin, M. M. (2020). *California Assessment of Student Performance and Progress (CAASPP): California Science Test (CAST) alignment study report* (2020 No. 040). Alexandria, VA: Human Resources Research Organization.

ETS. (2018a). *California Next Generation Science Standards Core Content Connectors for Alternate Assessments*. Revised January 2018. Retrieved from: https://www.cde.ca.gov/ta/tg/ca/documents/ngssaltconnectors.docx

ETS. (2018b). *California Alternate Assessment for Science Blueprint*. Approved by the State Board of Education on January 18, 2018. Retrieved from: https://www.cde.ca.gov/ta/tg/ca/documents/caascienceblueprint.docx

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.

This page is intentionally blank.

Glossary of Acronyms

Acronym	Glossary
CA NGSS	California Next Generation Science Standards
CAA	California Alternate Assessment
CAASPP	California Assessment of Student Performance and Progress
CAST	California Science Test
CDE	California Department of Education
DFA	Directions for Administration
DIF	Differential Item Functioning
DOK	Depth of Knowledge
EU	Essential Understanding
ESSA	Every Student Succeeds Act
FKSAs	Focal Knowledge, Skills, and Abilities
IEP	Individualized Education Plan
LEA	Local Education Agency
PE	Performance Expectation
PT	Performance Task
UD	Universal Design

This page is intentionally blank.

Appendix A: CAA for Science Documentation Reviewed by HumRRO

Table A.1. CAA for Science Documents Reviewed

Document Focus	Document File Name
Describes how test forms are assembled.	CAA for Science Blueprint
Document produced by ETS psychometrics group listing the statistical parameters for individual items and the form as a whole.	2019–2020 Statistical Specifications
Training materials for outside item writers consisting of a slide deck and handouts.	Item Writer Workshop (IWW) Materials (3 documents)
Training and other materials for teacher reviews of items.	• Item Review Meeting (IRM) Materials • (5 documents, 1 slide deck, 1 spreadsheet) • CAA for Science May Item Review Meeting Invitee List • CAA for Science Item Review Meeting Comment Sheet
Provides the final configuration of the four test versions making up the 2019–2020 administration, with the details of each of the four PTs that constitute a "version". (Science Connectors assessed, and item set status as operational or FT).	2019–2020 Test Design (2 documents)
Contains information about the assessment priority levels of the Science Connectors and the administration years that each Connector will be (a) Field Tested and (b) Operationally Assessed.	CAA for Science 5–year Administration Plan
Excel documents that contain assessment metadata for each PT, including but not limited to item number, sequence, Connector, item type, key, and statistical information. One per grade (grades five, eight, and high school).	2019–2020 Form Planners (3 spreadsheets)
Online, self–guided training module that test examiners must complete to be certified to administer CAAs each year.	2019–2020 Test Examiner Tutorial

Table A.1. (Cont.)

Document Focus	Document File Name
High–level explanation of the 2019–2020 CAA Science administration for LEA coordinators and test examiners, including listing of the Science Connectors, by science domain, that are assessed this year.	2019–2020 Administration Planning Guides
Information used to guide the development of items assessing the given Science Connector, including descriptions of three different item complexity levels.	CAA for Science Item Specifications (all 72)
Scripts that guide test examiners through administration of each PT.	2019–2020 *Directions for Administration*
Item data for 2019 operational items.	2019 Item Metadata
Analysis of field test data that can be used as evidence of reliability and validity.	Field Test Technical Report
Report describing the process of identifying the Science Connectors.	CAA Prioritized Connectors Memorandum Development Plan for the California Next Generation Science Standards Alternate Core Content Connectors California Next Generation Science Standards Core Content Connectors for Alternate Assessments Report
Survey that examiners complete after test administration.	Test Examiner Survey Extract
Example score reports.	CAASPP 2019–2020 Student Score Report Mockup – CAA for Science

Appendix B: Alignment Workshop Materials

This page is intentionally blank.

California Alternate Assessment (CAA) for Science Alignment Workshop
November 5–6, 2019

Embassy Suites by Hilton Sacramento Riverfront Promenade
Sacramento, CA

Agenda

Day 1 – Tuesday, November 5

8:00 – 8:30 a.m.	Panelists sign in and sign CAASPP Confidentiality Agreement (Tower Bridge Room, main floor)
8:30 – 10:30 a.m.	Welcome, introductions, logistics, and general training
10:30 – 10:45 a.m.	Break – Report to Panel Rooms • Grade 5: Sutter Board Room (second floor) • Grade 8: Tower Bridge Room (main floor) • High School: Crocker Board Room (second floor)
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 – Wednesday, November 6

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

CAA for Science Alignment Study Workshop
Panelist Instructions

1	Panelist Instructions	Print copy
2	CAA for Science Cognitive Complexity rating guide	Print copy
3	CAA for Science Accessibility rating guide	Print copy
4	Core Content Connectors for Science	Print and electronic copy
5	*Directions for Administration* (DFA)	Print and electronic copy
6	Item Specifications	Print and electronic copy
7	CAA for Science Rating Form	Excel file
8	CAA for Science Items	Accessed via computer link
9	Debriefing/Evaluation Form	Print copy
10	Demographic Questionnaire	Print copy

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

Prior to alignment ratings:

1. Introductions
2. Review all of the materials that panelists should have
    a. Laptops for recording ratings in Excel and accessing CAA for Science items
    b. Panelist Instructions
    c. Core Content Connectors for Science
    d. *Directions for Administration* (DFA)
    e. Item Specifications
    f. Cognitive complexity levels for CAA for Science
    g. Accessibility guidance for CAA for Science
3. Additional documents will be handed out as needed
    a. Demographic Questionnaire
    b. Debriefing/Evaluation form

Rate CAA for Science Items

Train Task:

1. Panelists will review several CAA for science items and will enter the core content connector ratings, cognitive complexity rating, and accessibility rating for each item.
2. Access CAA_Rating Form 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., CAA_Rating Form_*groupname*_**ymn**).
    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.
b. Columns A through D are filled with information about each CAA for Science item. Column A (hidden) provides the ETS unique item identifier. 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 content connector.
c. Column E asks panelists to identify the core content connector and type in the associated code from the Core Content Connectors for Science. An example connector code is **5–LS1–1.** The first number indicates the grade level**,** then the domain, followed by numbers indicating specific connector within this grade/domain.
d. Columns F and G are for panelists to identify the KFSA **or** EU the item measures. Panelists should refer to the Core Content Connectors for Science for the codes that correspond to the FKSA and EU (the code can be found with the PE associated with the Connector, presented in **bold**). Panelists should select the corresponding code from the drop–down menu on their rating form. If an FKSA is selected, then the EU cell (Column G) will become highlighted in black. If an EU is selected, then the FKSA cell (Column F) will become highlighted in black.
e. Columns H and I are for panelists to rate the quality of the link between the item and the identified FKSA **or** EU. If everything that the item measures is contained in the identified FKSA **or** EU, then the quality of link rating is *2– Fully linked*. If the item measures content in the identified FKSA **or** EU along with additional content, then the quality of link is *1– Partially linked*. If the item measures content that is not contained in any FKSA **or** EU, then the quality of link is *0– No link.* If an item is rated *1– Partially linked* or *0– No link*, then panelists should enter a rationale for that rating in Column I. If no rationale is entered, the cell will become highlighted in red.
f. Column J is for panelists to indicate the cognitive complexity level that best represents the cognitive demand of the item. Panelists will select the appropriate level from the drop–down menu. Panelists should refer to the CAA for Science Cognitive Complexity Rating Guide for definitions of the three levels (Low, Medium, and High).
g. Column K is for panelists to indicate if the item is accessible to most students who take the CAA for Science. Panelists will select *Yes* or *No* from the drop–down menu. Panelists should refer to the CAA for Science Accessibility Rating Guide to support this determination.

h. If an item is rated as not accessible in Column K, then panelists should enter a rationale for that rating in Column L. If no rationale is entered, the cell will become highlighted in red.
i. Column M is for panelists to enter any comments or notes regarding the quality of the item or the phenomenon the item references.

Conduct Task:

1. Panelists rate the first item independently, all indicated fields. Next, panelists discuss their ratings. The HumRRO facilitator will share the item metadata and item specifications. The group will discuss any discrepancies. The HumRRO facilitator will poll the group regarding each rating and will capture the final consensus rating. If true consensus cannot be reached, the rating of the majority of panelists will be recorded.
    a. Repeat at least 3 times, one item at a time.
    b. Panelists should not change ratings after discussion and review unless they are **certain** they made an error (e.g., coding error or misunderstanding of the standards). Do NOT change independent ratings after seeing the metadata.
2. Panelists should rate all remaining CAA for Science items independently in sets of 3–8 items before discussing and settling on consensus. The HumRRO facilitator will instruct the group on the set of items to be rated. 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.

Item Content Specifications for PE 3–LS3–1

Domain	Life Sciences
CA NGSS PE	**3–LS3–1** Analyze and interpret data to provide evidence that plants and animals have traits inherited from parents and that variation of these traits exists in a group of similar organisms.
Connector	Based on data through observation, identify similarities in the traits of a parent and the traits of an offspring and variations in similar traits in a grouping of similar organisms.
FKSA	1: Ability to identify similarities in the traits of a parent and the traits of an offspring (e.g., tall plants typically have tall offspring).
Essential Understanding	Identify variations in similar traits in a grouping of similar organisms (e.g., dogs come in many shapes and sizes, siblings look alike and different).
Point Value	1–2
Associated ETS Connector(s)	**NA**
Other Notes	**NA**

Example Item Rating Form

Item Metadata			Connector Alignment	FKSA or EU Alignment				Item Complexity	Item Accessibility		Comments
Item Sequence (Item Viewing System)	Item Type	Domain	Identify the Connector *(Select from drop down menu)*	Identify the FKSA *(Select from drop down menu)*	Identify the EU *(Select from drop down menu)*	Quality of Link to FKSA/EU *(Select from drop down menu)*	If the Quality of Link is 0 or 1, state specifically why the item content does not match a FKSA or EU	Assign an item complexity level (Low, Med, High) *(Select from drop down menu)*	Is item accessible to most students taking the CAA for Science? *(Select from drop down menu)*	If item is not accessible to most students, state specifically why	Provide comments about the appropriateness of phenomena, item quality, etc. (Optional)
1	MCSS-Member	ESS									
2	MCSS-Member	ESS									
3	MCSS-Member	ESS									
4	MCSS-Member	ESS									
5	Composite Objective-Member	ESS									
6	MCSS-Discrete	ESS									

HumRRO prepopulated CAA for Science metadata in first three columns of the form:

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

Second column: Type of item reviewed (e.g., discrete)

Third column: Science domain the item intended to measure

Panelists entered item–level rating data in cells under next nine column headers of the form:

Fourth column: Core Content Connector alignment

Fifth column: Focal Knowledge, Skills, and Abilities (FKSAs) alignment

Sixth column: Essential Understanding (EU) alignment

Seventh column: Quality of the FKSA/EU link

Eighth column: Required comment if FKSA/EU not rated as fully linked

Ninth column: Cognitive complexity level

Tenth column: Item accessibility rating

Eleventh column: Required comment if item rated as not accessible

Twelfth column: Additional panelists comments

California Alternate Assessment (CAA) for Science
Cognitive Complexity Rating Guide

- The complexity levels are generally aligned to the FKSA and EU for each Science Connector.

- The assumption is that a large number of the test takers do not have the capability to read the item text.

- Items at the low complexity level must be written using the simplest grammatical structure to assess the most basic understanding of the content described by a Science Connector. The EUs and FKSAs may progress in difficulty and complexity as the grade level increases. It is important, however, that items at the low complexity level should follow the guidelines described below, as well as the more specific guidelines in the individual item specifications for each Science Connector. These items should not be written with an assumption that their complexity should be increasing as the grade level increases.

- Complexity should be reflected in the complexity of the thought process required to respond correctly, not in the vocabulary or sentence structure of the items.

Complexity Level Guidelines

Low Complexity

- Assess the most basic elements of the knowledge and skill described by the Connector at a recall/recognize level and require no abstract reasoning or application of information.

- Will usually, but not always, be aligned to the Essential Understanding (EU)

Medium Complexity

- Assess the knowledge and skills described by the EU and FKSA at a recall, recognize, identify level, as well as, to a limited degree, the application level.

- Will usually, but not always, be aligned to the FKSA.

High Complexity

- Assess the knowledge and skills described by the FKSA at the most abstract level of the FKSA. This includes interpretation and application of information provided by data in graphs, charts, and tables as well as data provided in other graphics.

- Aligned to the FKSA.

California Alternate Assessment (CAA) for Science Accessibility Rating Guide

Item Accessibility Description

- Refers to the capacity of the student, using accommodations and other allowable individualization, to demonstrate knowledge of the content indicated by the test item.

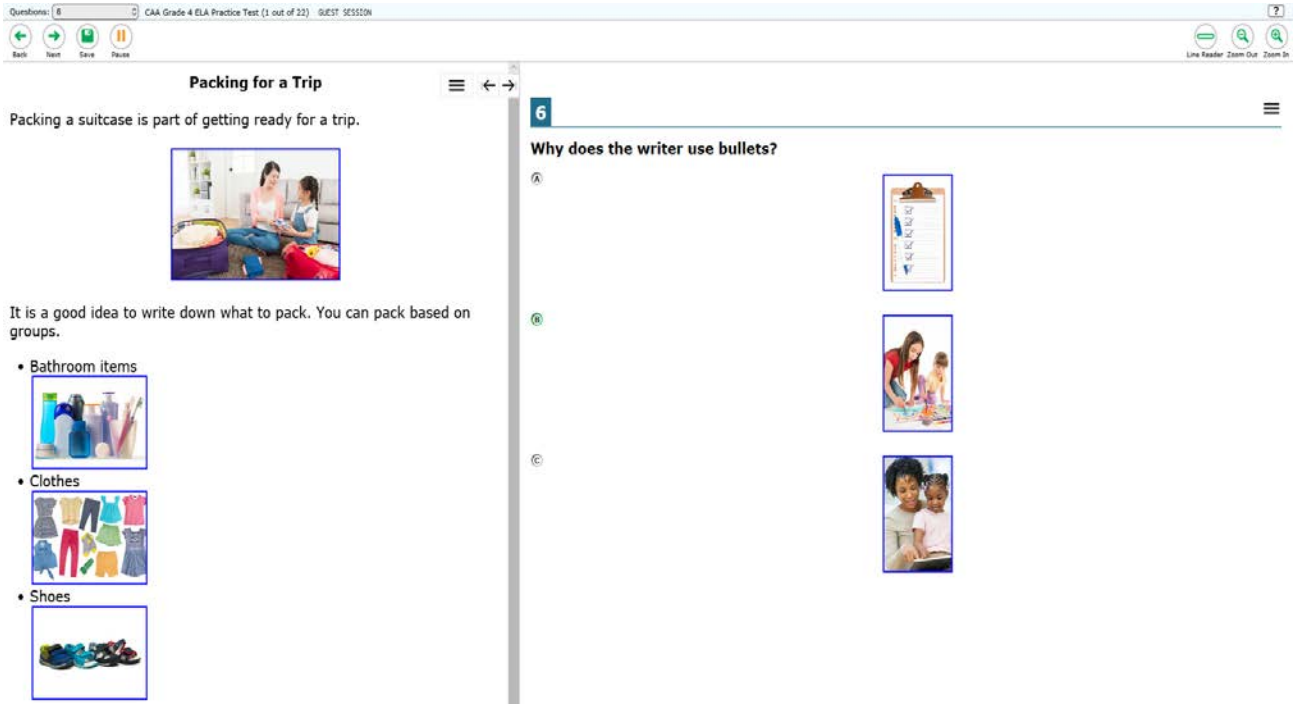
- Assume the necessary supports are provided within the limitations of the assessment requirements.

- Assume the test administrator is highly familiar with the student, the student's Individual Education Plan (IEP), and that they administer the assessment according to the instructions and training they have received.

See reverse side for an example of a "Not Accessible" Item and its Directions for Administration.

Example of a "Not Accessible" Item

The image below shows an item as it is displayed to a student taking the computer administered test. The left side of the image is the item prompt (Packing for a Trip), and the right side of the image asks a question about the prompt and presents three answer choices.



The image below is a screen shot of an excerpt of the Directions for Administration, with the script and alternative text for the specific test item above.

English Language Arts
Practice Test, No.6

Administration Script	Alternative Text for Students with Visual Impairments
SAY: **We will read about packing for a trip. Follow along as I read aloud. Then I will ask you a question.** **READ** the passage aloud and **POINT TO** the corresponding pictures as the passage is read. After the passage has been read, **SAY:** **We have finished reading. Now I will ask you a question. Why does the writer use bullets?** **To make a list [POINT], to paint a picture [POINT], or to tell a story [POINT]?**	**DESCRIBE:** *The picture shows a mom and daughter packing a suitcase.* *The picture shows several bottles of shampoo, lotion, and toothbrushes.* *The picture shows pieces of clothing.* *The picture shows different shoes.* *The picture shows a piece of paper with checkmarks on it.* *The picture shows children painting.* *The picture shows a mom telling her daughter a story.*

Debriefing: Analysis of Alignment Outcomes for the California Alternate Assessment (CAA) for Science

The questions and response options below represent the content of a paper-and-pencil survey given to panelists after they finished rating all of the CAA for Science items for their grade.

1. Panel (Grade 5, Grade 8, High School)

2. Did the items you reviewed generally represent the content in the CAA for Science standards 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 you expected? If not, were item cognitive complexity 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 CAA for Science items you reviewed and the CAA for Science standards?

   Excellent, Good, Limited, Weak (please explain and provide examples)

Comments:

Evaluation: Alignment Workshop Training and Procedures

The questions and response options below represent the content of a paper-and-pencil survey given to panelists after all rating activities were concluded.

1. **The training presentation in the large group provided useful information about the CAA for Science 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 CAA for Science 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

Table C.1 Comparison of Blueprint and Test Forms: Grade Five Science Connectors per Task and Item Complexity Levels per Task

Connectors or Items:	Number of Connectors per Task		Number of Low Complexity Items		Number of Medium Complexity Items		Number of High Complexity Items	
Domain	Blueprint	Forms	Blueprint	Forms	Blueprint	Forms	Blueprint	Forms
Earth and Space Sciences	2	2–2	4	4–5	4	3–4	2	2–2
Life Sciences	2	2–2	4	4–4	4	4–4	2	2–2
Physical Sciences	2	2–2	4	2–4	4	3–5	2	3–3

Note. Values in the Blueprint columns reflect the number designated in the test blueprint. Values in forms columns reflect the range of values across the four test forms based on panelist ratings.

Table C.2 Comparison of Blueprint and Test Forms: Grade Eight Science Connectors per Task and Item Complexity Levels per Task

Connectors or Items:	Number of Connectors per Task		Number of Low Complexity Items		Number of Medium Complexity Items		Number of High Complexity Items	
Domain	Blueprint	Forms	Blueprint	Forms	Blueprint	Forms	Blueprint	Forms
Earth and Space Sciences	2	2–3	4	2–3	4	4–5	2	3–3
Life Sciences	2	2–3	4	4–4	4	2–3	2	3–4
Physical Sciences	2	2–2	4	3–4	4	4–5	2	2–2

Note. Values in the Blueprint columns reflect the number designated in the test blueprint. Values in forms columns reflect the range of values across the four test forms based on panelist ratings.

Table C.3 Comparison of Blueprint and Test Forms: High School Science Connectors per Task and Item Complexity Levels per Task

Connectors or Items:	Number of Connectors per Task		Number of Low Complexity Items		Number of Medium Complexity Items		Number of High Complexity Items	
Domain	Blueprint	Forms	Blueprint	Forms	Blueprint	Forms	Blueprint	Forms
Earth and Space Sciences	2	3–6	4	3–3	4	3–6	2	1–4
Life Sciences	2	2–3	4	4–7	4	1–3	2	1–3
Physical Sciences	2	3–5	4	3–4	4	5–6	2	1–2

Note. Values in the Blueprint columns reflect the number designated in the test blueprint. Values in forms columns reflect the range of values across the four test forms based on panelist ratings.

This page is intentionally blank.

Appendix D: Detailed Descriptions of Figures with Image

Figure 1.1 CAA for Science standards continuum (p. 1-1).

- The CAA science standards continuum graphic depicts four connected linear boxes with the right side of each box coming to a point to indicate they flow from one to another (left to right).
- The first of four boxes is labeled CA NGSS and written below is Performance Expectation.
- The second box is labeled Science Connector and written above it is Bridge to Performance Expectation and below the box is Alternate Science Learning Goals.
- The third and fourth boxes have Assessment Targets written above both boxes. The third box is labeled FKSA and written below is Focal Knowledge, Skills, and Abilities. The fourth box is labeled EU and written below is Essential Understanding.

Example of a “Not Accessible” Item (p. B-11) .

- Top image is the item content.
- The item prompt is titled "Packing for a Trip."
- Sentence ("Packing a suitcase is part of getting ready for a trip.") is followed by a picture of a mother/daughter packing.
- Sentence ("It is a good idea to write down what to pack. You can pack based on groups.") is followed by three pictures (bathroom, pieces of clothes, shoes) presented as a bulleted list.
- Question ("Why use bullets?") is followed by answer option pictures (checklist, two children painting, and mother/daughter reading).
- Bottom image is an excerpt from the *Directions for Administration* for the English Language Arts Practice Test, No. 6.
- The first column is the Administration Script for the item.
    - SAY: We Will read about packing for a trip. Follow along as I read aloud. Then I will ask you a question.
    - READ the passage aloud and POINT TO the corresponding pictures as the passage is read. After the passage has been read,
    - SAY: We have finished reading. Now I will ask you a question. Why does the writer use bullets? To make a list [POINT], to paint a picture [POINT], or to tell a story [POINT]?

- The second column is the Alternative Text for Students with Visual Impairments.

    - DESCRIBE: The picture shows a mom and daughter packing a suitcase. The picture shows several bottles of shampoo, lotion, and toothbrushes. The picture shows pieces of clothing. The picture shows different shoes. The picture shows a piece of paper with checkmarks on it. The picture shows children painting. The picture shows a mom telling her daughter a story.