2018 No. 087

California Assessment of Student Performance and Progress (CAASPP)

2018 Independent Evaluation 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: Michele M. Hardoin
Arthur Thacker
Rebecca Norman Dvorak
D.E. (Sunny) Becker

Editor: Christa Watters

Date: December 28, 2018

California Assessment of Student Performance and Progress (CAASPP)

2018 Independent Evaluation Report

Table of Contents

Executive Summary ES-1

Chapter 1: Overview of the 2018–20 Evaluation Plan 1-1
- Background 1-1
- Initial Development 1-2
- Technical Advisory Group Review 1-3
- Overall Goals and Timeline 1-3
- Study Designs 1-5

Chapter 2: California Science Test Alignment Study 2-7
- Background 2-7
- Research Questions 2-10
- CAST Alignment Acceptability Criteria 2-10
- Data Collection Methods and Measurement Instruments 2-11
- Data Analyses 2-15
- Schedule of Milestones 2-21

Chapter 3: Impact on Instruction and Student Learning Case Study 3-23
- Background 3-23
- Selection of Impact Case Study Participants 3-24
- Data Collection 3-27
- Data Analysis Summary 3-29
- Reporting Results 3-35
- Schedule of Milestones 3-35

Chapter 4: California Alternate Assessment for Science Alignment Study 4-37
- Overview 4-37
- Research Questions 4-38
- Data Collection Methods and Measurement Instruments 4-39
- Data Analyses 4-42
- Draft Schedule of Study Milestones 4-43

References 45

Glossary of Acronyms 47

Appendix A: Theory of Action for CAASPP and the Smarter Balanced Assessment System A-1

Appendix B: CAST Alignment Study Workshop Data Collection Plan B-1

Appendix C: Impact Case Study Definition of “Full Implementers” C-1

Appendix D: Items for Impact Case Study Eligibility Survey of Prescreened LEA CAASPP Coordinators D-1

Appendix E: Criteria for Selection of Impact Case Study Participants from Eligible LEAs E-1

Appendix F: POC Polling Information and Example Questions for Impact Case Study....................F-1

Appendix G: Impact Case Study Data Analysis Plan G-1

Appendix H: Detailed Descriptions of Figures with Images H-1

List of Tables

Table 1.1 Overall Goals for Each 2018–20 Evaluation Study....................1-4

Table 1.2 Schedule of Planned Evaluation Activities for 2018–20 1-5

Table 2.1. CAST Domain Claims 2-8

Table 2.2. Rating Scale for Evaluating Strength of Evidence from Test Developer....2-12

Table 2.3 CAST Alignment Study Research Questions and Main Data Sources 2-15

Table 2.4. Schedule of CAST Alignment Study Activities....................2-21

Table 3.1 Cross Reference of Data Sources and Draft Impact Case Study Research Questions 3-31

Table 3.2 Schedule of Impact Case Study Activities 3-35

Table 4.1. CAA for Science Alignment Study Research Questions and Main Data Sources....................4-42

Table 4.2. Draft Schedule of CAA for Science Alignment Study Activities 4-43

Table B1. Number of CAST 2019 Operational Items to be Rated B-1

Table B2. Mock-up of Panelists’ Coding Spreadsheet B-3

Table B3. Sample Standards Coding Information for Panelists....................B-4

Table B4. Tentative Agenda for CAST Alignment Study Workshop B-5

Table E1. Theoretical Distribution of Target Characteristics of Impact Case Study LEAs in 2018–19 School Year E-1

List of Figures

Figure 2.1. PE Distribution for Segment A of the CAST Grade 5 Assessment 18

Figure 2.2. Example of a Wright Map comparing examinee ability and item difficulty distributions. 19

Figure 2.3. Sample graph illustrating number of items (horizontal axis) addressing each domain (vertical axis) by form. 20

This page is intentionally blank.

Executive Summary

The California Department of Education (CDE) awarded a contract in the summer of 2018 to the Human Resources Research Organization (HumRRO) for the independent evaluation of the CAASPP System. The evaluation is defined in California *Education Code* (*EC*) Section 60649, which states that evaluation activities "may include, but not necessarily be limited to, a variety of internal and external studies such as validity studies, alignment studies, and studies evaluating test fairness, testing accommodations, testing policies, and reporting procedures, and consequential validity studies specific to pupil populations such as English learners (ELs) and pupils with disabilities." The law requires development of a three-year evaluation plan of independent evaluation activities, and it prohibits duplication of studies conducted as part of federal peer review or by assessment contractors.

HumRRO served as the first independent evaluator of the CAASPP System from July 2015 through June 2018. During those years, HumRRO conducted and reported on five research studies. The scope of the current independent 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 System. HumRRO will produce annual evaluation reports that summarize all work completed during the previous year, stand-alone reports for individual research studies, and a comprehensive final report. This is the first report in this series, the 2018 annual evaluation report. Given the restricted elapsed time between contract award and the submission of this report, its contents are necessarily limited. It describes the development and review of the plans for the research studies, in cooperation with CDE staff and the CAASPP Technical Advisory Group (TAG). This report presents the final CAASPP 2018–20 Evaluation Plan in its entirety, as well as several draft data collection instruments developed for the first year's studies.

The 2018–20 Evaluation Plan consists of the CAASPP System's theory of action (CDE, 2018a), presented in Appendix A, and detailed plans for each evaluation study, including a timeline for major milestones. Timing for implementation of each study is based on CDE priorities and the expected dates of operational administration of assessments; however, schedules and the detailed study designs will be reviewed each year and may be revised.

- California Science Test (CAST) Alignment Study, to be conducted during the 2018–19 school year, is presented in chapter 2.

- Impact on Instruction and Student Learning Case Study (hereafter, Impact Case Study), a two-year effort to be conducted during the 2018–19 and 2019–20 school years, is presented in chapter 3.

- California Alternate Assessment (CAA) for Science Alignment Study, scheduled to be conducted during the 2019–20 school year, is presented in chapter 4.

This page is intentionally blank.

Chapter 1: Overview of the 2018–20 Evaluation Plan

Background

The California Assessment of Student Performance and Progress (CAASPP) System continues to be an ambitious and important effort by the state of California to provide feedback on individual student achievement and monitor progress in implementing effective instruction aligned with the Common Core State Standards (CCSS) for English language arts/literacy (ELA) and mathematics and the California Next Generation Science Standards (CA NGSS). This system of assessments provides the compass to guide individual student learning through implementation and improvement of curricula. The CAASPP System represents a substantial financial investment by the state as well as a significant investment of educator and student time to participate in the various assessments.

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

The CDE specified in its Request for Proposals (RFP) that the 2018–20 independent evaluation focus on the following CAASPP System components:

- Smarter Balanced Summative Assessments (ELA and mathematics), required for grades three through eight and grade eleven, comprised of a computer-adaptive test (CAT) and a performance task (PT).
- Smarter Balanced Interim Assessments (ELA and mathematics), optional assessments designed for grades three through eight and grade eleven, available for use by educators from kindergarten through grade twelve to monitor student performance throughout the school year.
- Smarter Balanced formative assessment measurement tools and resources, available in the Smarter Balanced Digital Library.
- California Science Test (CAST), aligned with the NGSS for California Public Schools, Kindergarten through Grade Twelve (CA NGSS), required for students in grades five, eight and once in high school. The first operational administration of this new assessment will be in the spring of 2019.

- California Alternate Assessment for Science (CAA for Science), serving students with the most significant cognitive disabilities in grades three through eight and grade eleven. The first operational administration of this new assessment is planned for the 2019–20 school year.

The CDE awarded the contract for the 2018–20 independent evaluation of the CAASPP System to the Human Resources Research Organization (HumRRO). HumRRO served as the first CAASPP System evaluator from 2015–18. Copies of our annual and comprehensive final reports are available on the California Department of Education (CDE) Web page (https://www.cde.ca.gov/ta/tg/ca/caaspprptstudies.asp). The current contract calls for annual evaluation reports that summarize all work completed during the previous year, stand-alone reports for individual research studies, and a comprehensive final report. The annual reports prepared for this contract will include all data analyses pursuant to *EC* Section 60649.

Given the compressed timeline for development and submission of the 2018 report, as specified in the evaluation contract, the contents of this report are necessarily limited. This report begins with an overview of the development of the plans for three research studies, which HumRRO carried out in cooperation with CDE staff and the CAASPP Technical Advisory Group (TAG). The remainder of the report presents the final CAASPP 2018–20 Evaluation Plan in its entirety, as well as several draft data collection instruments developed for the first year's studies.

An ongoing evaluation is important to ensure that California gets the intended return on its investment in the CAASPP System. The evaluation can provide evidence to demonstrate the validity of intended interpretations of test scores used as measures of student learning relative to targeted content standards, and it can offer recommendations for potentially improving alignment between what an assessment measures and what it's intended to measure. The evaluation can also provide insight into how CAASPP results are used to improve instruction at the student, classroom, school, local educational agency (LEA), and statewide levels.

Initial Development

In response to the CDE's RFP, HumRRO developed a draft 2018–20 Evaluation Plan that proposed our approach to each of the three required studies:

- California Science Test (CAST) Alignment Study
- Impact on Instruction and Student Learning Study, a two-year effort
- California Alternate Assessment (CAA) for Science Alignment Study

The design of the studies was aligned with the theory of action for the CAASPP System, as articulated by the CDE (see Appendix A). The theory of action indicates components of the system should work together to accurately assess student achievement relative to grade level curriculum standards and provide information to educators to improve instruction, thereby improving student achievement. For each study, HumRRO's proposed Evaluation Plan included the rationale for the study, the research questions to be answered, an overview of the methods planned and data to be collected, and proposed data analyses.

The CAASPP Evaluation project orientation meeting was held in July 2018. During this meeting, as well as during two teleconference calls held prior to the meeting, HumRRO and CDE staff discussed the proposed approaches to the studies. We also discussed possible study variations that, as a whole, could accomplish the goals of the evaluation within the time frames and resources available. The schedule for implementation of each study was considered, both to meet CDE priorities and to coordinate with the timeline for operational administration of the CAST and CAA for Science assessments. Contributing to the discussions were the Director of the Assessment Development and Administration Division and CDE staff representatives from the CAASPP Lead Office; the Psychometrics, Evaluation, and Data Office; the Science Office; the ELA and Mathematics Office; the Interim Assessments, Digital Library, and Systems Office; and the Fiscal Support Office. Educational Testing Service (ETS) staff responsible for the CAASPP components being studied by the evaluation also contributed to discussions at the project orientation meeting.

Based on outcomes of meetings and discussions with the CDE, HumRRO refined the proposed evaluation plan and submitted the official first draft of the 2018–20 Evaluation Plan to the CDE on August 15, 2018, for formal review and comment.

Technical Advisory Group Review

HumRRO's 2018–20 Evaluation Plan was included as an agenda item for the September 2018 TAG meeting. Prior to the meeting, TAG members and ETS staff were sent the first draft of the 2018 annual evaluation report, which included the 2018–20 Evaluation Plan. During the meeting, HumRRO presented in detail the research questions, methods, and planned analyses for the CAST Alignment Study and the Impact Case Study. TAG members, whose role is to advise California's testing programs, discussed and critiqued the proposed study designs, pointing out aspects needing clarification and suggesting possible revisions to consider. Additionally, for the Impact Case Study, TAG members shared their concerns about the feasibility of HumRRO's planned approach for collaboration with educators, suggesting additional possible incentives to consider for study participants. HumRRO also presented, in brief, the CAA for Science Alignment Study, which will be developed to a greater level of detail when the assessment is closer to the operational phase. Subsequent to the TAG meeting, several TAG members and ETS provided written feedback on the study designs for HumRRO's and CDE's consideration. HumRRO addressed the input received from the TAG and ETS in the final draft of the 2018 annual evaluation report.

Overall Goals and Timeline

The studies included in the 2018–20 Evaluation Plan will provide information about how well specific parts of the CAASPP System as delivered are meeting the intended goals of the program as expressed in the theory of action for the CAASPP System.

Table 1.1 gives an overview of the goals of each independent evaluation study and indicates the year in which each study will be conducted.

Table 1.1 Overall Goals for Each 2018–20 Evaluation Study

Study Title and Year Conducted	Goals
CAST Alignment Study, 2018–19	• Evaluate the degree of alignment between the CAST test items and test forms with the California Next Generation Science Standards (CA NGSS). • The CAST Alignment Study Report should guide future item development and provide validity evidence suitable for submission for federal peer review under the Every Student Succeeds Act (ESSA).
Impact Case Study, 2018–19 and 2019–20	• Collaborate with and gather extensive qualitative data from a small sample of schools and LEAs (case studies), purposefully selected based on their use of CAASPP components and resources. The small sample will aim to broadly represent the diversity of the state with respect to geographic location, academic achievement, and size (student enrollment), as well as student population characteristics (i.e., socioeconomic disadvantage and English learner status). • Investigate the context and various approaches used by the small sample of schools and LEAs to implement and integrate the components of the CAASPP System to inform instruction and improve student learning. • The two reports for the Impact Case Study will each describe in detail one school year's findings of the studied LEAs' and schools' use of CAASPP components and their impacts on instruction and student learning. The report will document in detail the local context for each case study. A separate sample of LEAs and schools will be investigated each school year of the study.
CAA for Science Alignment Study, 2019–20	• Evaluate the degree of alignment between the CAA for Science test items and test forms with the Core Content Connectors, which are based on the CA NGSS and were developed to form the basis for test development. • The CAA for Science Alignment Study report should guide future item development and provide validity evidence suitable for submission for federal peer review under ESSA.

Table 1.2 presents a summary list of key activities and milestones for implementing the 2018–20 Evaluation Plan.

Table 1.2 Schedule of Planned Evaluation Activities for 2018–20

Activity	Time Frame
Orientation Meeting with CDE staff: In-person meeting to review all tasks and project timeline and to address questions and concerns	July 2018
Management Meetings with CDE Staff: Biweekly calls to discuss progress, plans, and issues	July 2018–December 2020
State Board of Education (SBE) Meetings: Meet with SBE staff and provide presentations at Board meetings.	As requested, up to two times annually, July 2018–December 2020
Technical Advisory Group (TAG) Meetings: Meet with and provide presentations, including detailed designs, review of progress on studies, preliminary findings from studies, and Evaluation Plan updates.	Three times annually, July 2018–December 2020
CAASPP Contractor Annual Planning Meeting: Attend meeting to learn of planned updates to the system, concerns, processes, scope, and schedule.	Annually, July 2018–June 2020
Conduct the CAST Alignment Study and deliver a stand-alone study report.	July 2018–June 2020
Conduct the Impact Case Study and deliver two stand-alone study reports.	Annually, July 2019–December 2020
Conduct the CAA for Science Alignment Study and deliver a stand-alone study report.	July 2019–June 2020
Develop and deliver annual report.	Annually, July 2018–December 2020
Develop and deliver comprehensive final report.	July–December 2020
Maintain comprehensive plan and schedule for project activities and deliverables.	July 2018–December 2020
Submit monthly written progress reports to describe evaluation progress, plans, and issues.	July 2018–December 2020

Study Designs

The remaining chapters of this report describe in detail the three research studies listed in Table 1.1. Each chapter presents research questions, methods for data collection and analysis, academic literature related to the methods, descriptions of measurement instruments to be developed for data collection, and a schedule of major activities.

- Chapter 2 presents the California Science Test (CAST) Alignment Study.
- Chapter 3 presents the Impact on Instruction and Student Learning Study.
- Chapter 4 presents the California Alternate Assessment (CAA) for Science Alignment Study.

This page is intentionally blank.

Chapter 2: California Science Test Alignment Study

Background

This alignment study is one means to gather evidence to demonstrate the validity of intended interpretations and uses of California Science Test (CAST) scores. HumRRO's approach to the study is conceptualized to verify that CAST uses a reasoned approach to sampling the content within the California Next Generation Science Standards (CA NGSS), in this way indicating whether the CAST effectively measures what it is intended to measure.

The CA NGSS provides a framework for science education and includes assessable statements, called Performance Expectations (PEs), of what students should know and be able to do. The following three major components, also referred to as dimensions, are combined to produce the PEs:

- Disciplinary Core Ideas (DCIs) are the key ideas in science that have broad importance within or across multiple science or engineering disciplines. As students progress through grade levels, they experience how these core ideas build on each other.
- Science and Engineering Practices (SEPs) describe what scientists do to investigate the natural world and what engineers do to design and build systems. The practices better explain and extend what is meant by "inquiry" in science and the range of cognitive, social, and physical practices that it requires. Students engage in practices to build, deepen, and apply their knowledge of core ideas and crosscutting concepts.
- Crosscutting Concepts (CCCs) help students explore connections across the four domains of science, including Physical Sciences, Life Sciences, Earth and Space Sciences, and Engineering. When these concepts, such as "cause and effect," are made explicit for students, they can help students develop a coherent and scientifically-based view of the world around them.

Because the CA NGSS are designed as interactions among statements about content, the alignment method must allow for multiple standards to align with a single complex item or performance task.

The CAST is a computer-based assessment that will be administered to students in grades five, eight, and once in high school. The CAST was field tested in spring 2018 and will be administered operationally for the first time in January–July of 2019. The 2018–19 administration will be fixed-form (non-adaptive). The CAST includes three science domains (Physical Sciences, Life Sciences, and Earth and Space Sciences) and one engineering domain (Engineering, Technology, and Application of Science). Items written to assess PEs associated with Engineering, Technology, and Application of Science will be assigned to one of the three science domains, depending upon the context of their stimulus. California's Environmental Principles and Concepts will also be used as context for items, as appropriate to the three science domains. The High Level Test Design for the CAST includes the following three segments:

- Segment A: a set of selected response and short constructed response items taken by all students.
- Segment B: a set of performance tasks taken by all students.
- Segment C: a set of items comparable to Segment A or B, matrixed across test forms, each taken by a sample of students.

The 2018–19 administration will not fully implement this test design, because Segment C will consist only of field test items without item statistics. Starting in 2018–19, results from the first two segments will be used to report individual student scores. For the 2019–20 and subsequent administrations, the matrix portion of the test will provide school- and LEA-level information about student achievement on a broader sample of content than would be possible otherwise. Segment C will not be used for individual score reporting.

Table 2.1 presents claims as listed in the CAST blueprint (ETS, 2017). Bold font has been added to highlight domain-level text. CAST has four claims, one overall and three separate science domain claims.

Table 2.1. CAST Domain Claims

Domains	Claim
3D Overall	Students can demonstrate performances associated with the expectations of the California Next Generation Science Standards, through the integration of Science and Engineering Practices, Disciplinary Core Ideas, and Crosscutting Concepts **across the domains of Physical Sciences, Life Sciences, Earth and Space Sciences, and Engineering, Technology, and Application of Science**.
3D Physical Sciences	Students can demonstrate performances associated with the expectations **in the disciplinary area of Physical Sciences** within the California Next Generation Science Standards, through the integration of Science and Engineering Practices, Disciplinary Core Ideas, and Crosscutting Concepts.
3D Life Sciences	Students can demonstrate performances associated with the expectations **in the disciplinary area of Life Sciences** within the California Next Generation Science Standards, through the integration of Science and Engineering Practices, Disciplinary Core Ideas, and Crosscutting Concepts.
3D Earth and Space Sciences	Students can demonstrate performances associated with the expectations **in the disciplinary area of Earth and Space Sciences** within the California Next Generation Science Standards, through the integration of Science and Engineering Practices, Disciplinary Core Ideas, and Crosscutting Concepts.

The structure of CAST requires that alignment should be considered in at least two ways. First, individual students' scores should be sufficiently valid and reliable to support their intended interpretations, and this can only happen if CAST uses a reasoned approach to sampling the content within the CA NGSS in the first two test segments. One level of alignment reporting must therefore be based only on Segments A and B. For the remainder of the study design, this will be referred to as "student-level alignment."

In years subsequent to the first operational administration, CAST results at the school and LEA levels will incorporate Segment C, in addition to Segments A and B. HumRRO expects test reporting to be less specific at the student-level than at the school or LEA level. Inferences made at the school or LEA level will reference the full set of operationally administered test items, including those in Segment C. Alignment reporting should take the full operational item pool into account. We refer to this as "overall alignment." At the time of this study, Segment C will not be part of the operational CAST. Items to populate Segment C will be field tested in 2019 for operational administration in 2020. This means that these items will not be finalized or available for review by panelists as part of this study. The contributions of Segment C will be estimated based on item metadata (item content coding and complexity indications) only.

These two ways of considering alignment match the CAST test blueprint design explicitly. The blueprint indicates, "For scoring and reporting purposes, each of the three science domains will constitute one third of the test (items written to assess PEs associated with Engineering, Technology, and Application of Science will be assigned to one of the three science domains, depending upon the context of their stimulus)." It continues, "For the segments contributing to individual student scores (Segment A and Segment B), it is not possible to assess all PEs in a single testing year. As a result, PEs assessed in Segment A and Segment B will be rotated from year to year so that all PEs can be assessed in the segments contributing to individual scores over the course of a three-year period." We will use student-level alignment results to evaluate the CAST for this purpose.

The blueprint describes the use of Segment C in years subsequent to the first operational administration: "For the segment contributing only to group scores (Segment C), matrix sampling (the administration of a number of different versions across the state) will allow for assessment of all PEs annually at a state-wide level." We will use overall alignment results to evaluate the CAST for this purpose. These results should also serve as a reasonable proxy for the "three-year" combined alignment of the student-level results. As described above, Segment C will be evaluated based only on item metadata, and only for content representation. These items will not be available for independent panel review at the time of the study.

Research Questions

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

1. To what extent do the test design and intended distributions for science domains and dimensions for the CAST support the claims to be made about student performance on the assessment?
2. To what extent do the intended distributions for science domains and dimensions for the CAST represent an appropriate sampling of the content as set forth in the CA NGSS?
3. To what extent do the CAST test forms and test items reflect the test design and intended distributions for science domains and dimensions?
4. To what extent do CAST tasks and items integrate at least two dimensions (i.e., disciplinary core idea, crosscutting concept, and/or science and engineering practice)?
5. To what extent do CAST test forms show balance across the disciplinary areas (physical sciences, life sciences, and earth and space sciences)?
6. Does the CAST include items that are of sufficiently high cognitive complexity to address the CA NGSS?
7. What is the distribution of item difficulties within test forms and the distribution of student ability?

CAST Alignment Acceptability Criteria

Traditional alignment studies have generic acceptability criteria — such as six items per standard or at least half of the items at or above the Depth of Knowledge (DOK) of the matched standard—and are not appropriate for the three-dimensional nature of the CA NGSS.

To evaluate the evidence collected to answer the research questions, we will develop acceptability criteria specific to CAST results. Draft acceptability criteria will be developed for research questions 1 through 6, based on a thorough review of ETS's reporting strategy and documentation (especially mock score reports and other test information) as well as literature on evidence-centered design (and other Principled approaches to Assessment Design, Development, and Implementation (PADDI)), federal peer review guidance, and validation. HumRRO is retaining nationally recognized NGSS experts to review the draft criteria and reach consensus on any revisions. The final criteria will be used to determine the appropriateness of the evidence statements and interpretations of results described by ETS.

Data Collection Methods and Measurement Instruments

Extant data to be collected for this study include:

- *CAST test developer (ETS) documentation of how alignment was considered during test development.* This should include item and form development guidance, including 2018–19 Block Builders and Form Planners; test blueprint; 2018–19 item specifications and statistical specifications; High Level Test Design; item tryout and review procedures; procedures for reviewing and addressing item tryout information; validity and reliability evidence for the anticipated test form; guiding documents that illustrate the overall goals and philosophy underlying the assessment, such as the final version of ETS's *Use of Evidence-Centered Design in the Development of the California Science Test*. When validity or reliability evidence is not yet available, the testing contractor should provide clear plans for data collection and analyses, as well as any criteria that will be used to judge the appropriateness of the assessments' alignment.

- *CAST item metadata.* Metadata to include item parameters, *p*-values, depth of knowledge, cognitive complexity (DOK or similar, if available), and coding to the CA NGSS for all Segment A, B, and C items. Data will come from the 2019 administration (operational for Segments A and B, field test for Segment C). For Segments A and B, some metadata (i.e., coding to the CA NGSS) will be shared with panelists (see Appendix B). Note that analyses of metadata from Segment C will be limited to content designations and cognitive complexity only.

- *CAST items from Segments A and B that will be operationally administered in 2019.* Access to all test items must be in the same format as the items that will be viewed by students; however, to facilitate item ratings by panelists, items will be presented by domain (i.e., Life Sciences, Physical Sciences, Earth and Space Sciences). We assume there will be multiple operational forms (combinations of Segment A, which is the same for all students, and different versions of Segment B) and that the number of unique items (Segments A + B) will total no more than 70 items for each tested grade level. See Appendix B for additional details.

- *Student-level file of CAST scores for the overall test and for any reported sub-scores.* Student-level files do not need to include item-level scores. We do not need to identify individual students, schools, or districts for these analyses and thus will request that no personally identifiable information (PII) be provided.

- *Information on how CAST results are reported at the student- and school or LEA levels.* This information should include procedures for assigning individual students to performance levels, performance level descriptors, and any reporting categories used in aggregate level reports.

Data to be generated during this study include:

- *Independent ratings of strength of evidence of alignment considerations in test developer (ETS) documentation*. A customized rating form will be developed, guided by the Standards for Educational and Psychological Testing (APA, AERA, NCME, 2014) and federal peer review guidance. Three researchers will use the form to evaluate the documentation.

- *Expert panelists' ratings of CAST items.* Three panels (elementary, middle, and high school panels), each comprised of six educators familiar with the CA NGSS, will evaluate CAST items during an in-person workshop. Ratings of what standards the items assess, in terms of PEs and for the three CA NGSS dimensions (DCI, CCC, SEP), as well as ratings of the items' cognitive complexity requirements will be collected.

Evaluation of test developer documentation

HumRRO will create a customized rating form to capture reviewers' evaluation of test developer (ETS) documentation of how alignment was considered in test development. The form's design will be guided by the *Standards for Educational and Psychological Testing (hereafter referred to as SEPT)* (APA, AERA, NCME, 2014), which describe requirements for developing, reviewing, and trying out test items (mostly Chapter 4) as well as requirements for documenting the processes used (Chapters 4 and 6). Evaluation forms and essential questions will also be guided by Achieve's Criteria for Procuring and Evaluating High-Quality and Aligned Summative Assessments (Achieve, 2018). Additional components will be added to the form as necessary to ensure clear parallels to the federal peer review guidance. Testing standards and rating components will be selected to support the claims structure established for the CAST. The draft rating form will be submitted to the CDE Contract Monitor for review and approval. Table 2.2 presents the scale that will be used for the review.

Table 2.2. Rating Scale for Evaluating Strength of Evidence from Test Developer

Rating Level	Description
1	No evidence of the SEPT standard found in the documentation
2	Little evidence of the SEPT standard found in the documentation; less than half of the SEPT standard covered in the documentation and/or evidence of key aspects of the SEPT standard could not be found.
3	Some evidence of the SEPT standard found in the documentation; more than half of the SEPT standard covered in the documentation, including some key aspects of the SEPT standard.
4	Evidence in the documentation mostly covers the SEPT standard; the standard is largely covered in the documentation, including key aspects of the SEPT standard.

A rating will be made for each identified “relevant” SEPT standard. “Relevant” in this case refers to SEPT standards that specifically refer to alignment or SEPT standards that reference evidence for which alignment data would be considered a part (e.g., item review processes). In some cases, an individual SEPT standard may require consideration of multiple features. For example, Standard 4.8 from *Standards for Educational and Psychological Testing* states:

> “Standard 4.8. The test review process should include empirical analyses and/or the use of expert judges to review items and scoring criteria. When expert judges are used, their qualifications, relevant experiences, and demographic characteristics should be documented, along with the instructions and training in the item review process that the judges receive.”

The example is a compound testing standard, and it may be desirable to divide it into several component rating dimensions such as characteristics of expert judges and quality of training, among others. For each identified SEPT standard or rating component, three senior HumRRO researchers, all experienced in third-party evaluation of assessment systems, will independently assign a rating based on the evidence collected. After an initial round of independent ratings, the three researchers will meet to discuss any areas of disagreement (non-adjacent ratings) and to identify any gaps in the documentation received. If necessary, HumRRO will follow up with the testing contractor regarding questions or to request additional documentation. Once all identified SEPT standards have been independently rated, the researchers will compare and discuss their ratings and reach a final consensus rating for each SEPT standard or rating component.

Raters will record their ratings as well as the document(s) containing the evidence on which they based each rating. A single rated SEPT standard may reference multiple documents. Raters will provide a written justification for each rating, noting strengths and areas where evidence is potentially missing/undocumented or incomplete. This information will be provided for each rated SEPT standard or rating component. For any criterion receiving a “1” rating, HumRRO will ensure that no evidence of the standard was found because no evidence exists, rather than having that rating result from a logistical or communication error in which the testing contractor did not provide the necessary materials. For consensus ratings of “1,” HumRRO will make an additional request to the testing contractor to verify evidence for that criterion is truly missing. If additional evidence is provided, HumRRO will revise ratings as appropriate. The ratings will be arranged by claim (see Table 2.1) to facilitate validation.

CAST Alignment Workshop: Evaluating Items for Alignment with CA NGSS

HumRRO will work collaboratively with the CDE Contract Monitor to recruit and select a total of 18 educators to serve on three CAST alignment review panels (one elementary, one middle, and one high school panel). Panelists must be very familiar with the CA NGSS and have at least three years of experience as California educators. The CDE provided HumRRO with contact information for two sources for panelists: (a) the science subcommittee of the Curriculum and Instruction Steering Committee (CISC), comprised of 13 CA NGSS-knowledgeable educators for all grade levels, and (b) the CA NGSS Early Implementers, comprised of 26 educators for levels kindergarten through grade eight. Panelists may be individuals on these lists, or individuals, including teachers, who are recommended by people on these lists. The high school panel must include at least two biology educators, one chemistry educator, and one physics educator. At least one educator knowledgeable about the earth and human activity standards will be recruited for each panel; this educator may also serve in one of the previously mentioned roles (e.g., physics educator) or may be a dedicated earth science educator. Each panel should also include educators with experience teaching English learners (ELs) and/or students with disabilities (SWD) who take the CAST.

HumRRO will work collaboratively with the CDE Contract Monitor to determine specific dates in February 2019 and the specific location for the workshop. HumRRO will secure the meeting space and arrange for meals during workshop days, arrange lodging and travel for panelists, and provide all necessary equipment for the workshop, including two laptop computers per panelist, one laptop to view the assessment items and the other to document their ratings. Panelists will be required to sign nondisclosure agreements as a condition of participation.

HumRRO will conduct a two-day workshop, during which panels of educators will evaluate how well each item assesses the CA NGSS. The educators will make ratings regarding what standards items assess, accounting for the three-dimensional nature of the standards. They will rate each item according to its cognitive complexity requirements. They will discuss discrepant ratings and attempt to reach consensus or near-consensus when they disagree about important ratings. The data produced during the panel workshop will be used to evaluate alignment of CAST to the CA NGSS.

Panelists will be trained to rate CAST items using a grid system specifically designed to capture the three-dimensional nature of the content standards. Panelists will also assign a cognitive complexity rating to each item. HumRRO expects to use the cognitive complexity rating system used by the CDE and ETS during test development, in order to allow for direct comparisons of the ratings. A detailed description of the rating processes and examples of the rating forms are included in Appendix B.

Data Analyses

Table 2.3 summarizes the data that will be analyzed to answer each research question.

Table 2.3 CAST Alignment Study Research Questions and Main Data Sources

Research question, abbreviated form	Data to be analyzed
1. To what extent do the test design and intended distributions for science domains and dimensions for the CAST support the claims to be made about student performance on the assessment?	Independent rating scale data, structured by selected SEPT *standards* and peer review guidance
2. To what extent do the intended distributions for science domains and dimensions for the CAST represent an appropriate sampling of the content as set forth in the CA NGSS?	Ratings of CAST documentation from ETS; application of acceptability criteria established to support the appropriateness of reporting based on the CAST design
3. To what extent do the CAST test forms and test items reflect the test design and intended distributions for science domains and dimensions?	Depictions of CAST test content based on ETS metadata compared to test design documents and test blueprint
4. To what extent do CAST tasks and items integrate at least two dimensions?	Expert panelists' item ratings on three CA NGSS dimensions; depictions of items by dimensions in various graphical representations
5. To what extent do CAST test forms show balance across the disciplinary areas?	Expert panelists' item ratings on three CA NGSS dimensions; depictions of items by domain in various graphical representations
6. Does the CAST include items that are of sufficiently high cognitive complexity to address the CA NGSS?	Expert panelists' cognitive complexity ratings by dimension and domain, represented numerically and graphically
7. What is the distribution of item difficulties within test forms and the distribution of student ability?	CAST metadata from ETS and de-identified student score file

Panelist alignment ratings

Panelists will generate independent rating data for analyses, including indications of the cognitive complexity of items and the specific three-dimensional standards the items assess. When ratings are numeric and easily coded, statistics for ratings will be computed. When ratings are more complex, such as indications of three-dimensional standards, variance will be displayed in graphics.

Panelists' original data will be retained to allow us to determine the extent to which each panelist revised their original ratings. HumRRO will compute agreement statistics based on the panelists original data (panelists compared to other panelists), and on the final consensus data (panelist consensus data compared to contractor's metadata). Item data from panelists will be considered final after this step.

As a reminder, at the time of the study, Segment C will not be operational, but items will be field tested to support Segment C in 2020. HumRRO will compare panelists' data to the testing contractor's metadata related to item-to-standard coding for Segments A and B to gauge how closely panelists' data match the metadata. HumRRO will conduct analyses of Segment C based on the existing contractor-supplied metadata (content designations and cognitive complexity only).

The report of workshop results will include an assessment of the level of agreement among raters within each panel and overall agreement with the test developer's metadata. Any areas of significant disagreement will be reported to the test developers for consideration and comment. Overall alignment results (including Segment C) will be computed after the 2019 CAST administration based on item metadata only.

Regarding analysis of cognitive complexity ratings, traditionally, cognitive complexity has been evaluated in alignment studies by comparing the complexity of the test items to that of specific statements about student knowledge, skills, and abilities in the test content standards. One could simply match the item to its standard and directly compare to see if the item's cognitive complexity was lower, equal to, or higher than the standard's cognitive complexity. Rather than comparing panelists' cognitive complexity ratings to a specific standards statement, graphs will be generated to depict the distribution of cognitive complexity for the alignment results at the student level and overall. This approach supports the assertion that CAST should have items that represent a range of cognitive complexity and that range should be reasonably evenly dispersed by dimension and by content domain. Results will be depicted as the proportions of items at each cognitive complexity level by content domain and the proportion of items at each cognitive complexity by dimension. HumRRO will not combine content and dimension (e.g., Life Sciences SEP items) for these comparisons, as the numbers of items represented in each category would be too small to sensibly interpret. An attenuated range of cognitive complexity ratings is expected (toward higher cognitive complexity) for CAST items because of the nature of the CA NGSS. This portion of the review is designed to indicate any areas where complexity is unevenly represented by groups of items.

The data will also be parsed by dimension and domain, and represent the cognitive complexity distribution data in graphs for each. Cognitive complexity ratings assigned by panelists can also be graphed alongside the item cognitive complexity provided by the test developer in the metadata. The resulting set of graphs will allow for examination of the distribution of cognitive complexity by test purpose (student-level or overall), by domain, and by dimension, and all cognitive complexity data from panelists will be juxtaposed with the vendor's data. Lastly, the data will be provided in tables (mean cognitive complexity level, variance, and range) to facilitate comparisons to acceptability criteria.

The final results from the panelists' item-level indications will then be compared to the intended test blueprint. The panelists will have provided us with which DCIs, CCCs, and SEPs are addressed by items on each test. These can be directly compared to ETS test blueprint spreadsheets. Figure 2.1 shows the blueprint table for Grade 5 (ETS, 2017). By comparing the panelists' ratings to the blueprint, we can determine if the CAST has met the specifications related to items per DCI strand and domain. HumRRO will also be able to report the intersections among SEPs and PEs.

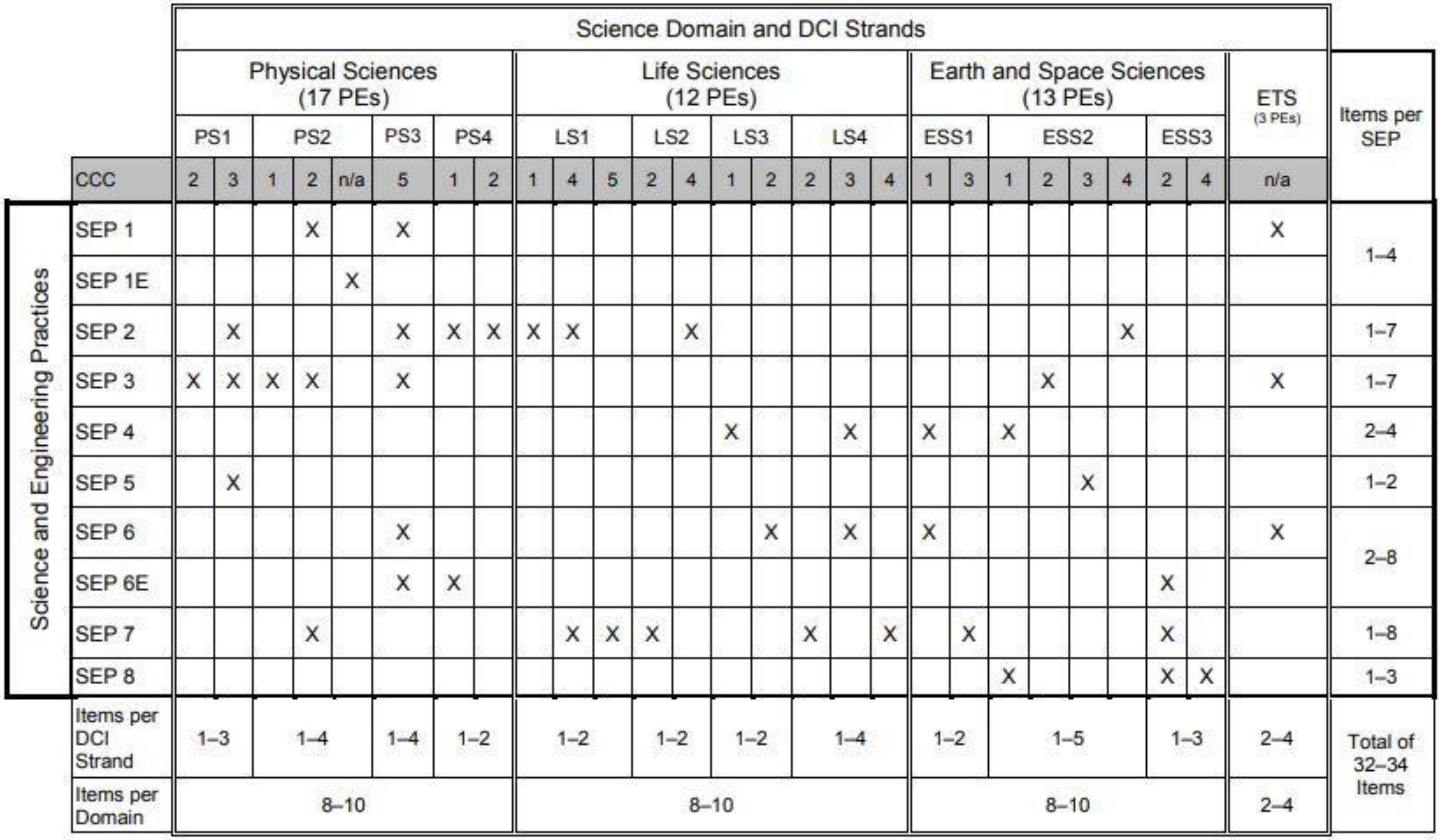
Test item difficulty and student ability

Test item difficulty can be compared to students' performance using a Wright Map (Callingham & Bond, 2006; Wilson & Draney, 2002). Figure 2.2 provides a sample Wright Map that has persons on the left side of the map and items on right. The "persons" side of the map is simply a histogram (frequency representation) depicting the overall score of all the test takers on the theta scale. The more Xs in each column on the left side of the map, the more people had that specific score. Higher-ability persons are located toward the top of the map.

The right side of the map has a histogram of item difficulties. The histogram is in roughly the shape of the normal curve. This indicates there are more items in the middle of the range of difficulty than on the easier or more difficult ends of the theta scale. Higher-difficulty items are toward the top of the map. In the example, there are a few items that are very easy for this population of students but no items with a difficulty parameter of more than 2 on the theta scale.

A Wright Map is one way to depict how items and persons are dispersed across the difficulty range of an assessment. If persons were clustered in areas where there were few or no items, this would be an indication the test was not functioning as intended. The test items might be too easy or too difficult for the tested population. The Wright Map can also tell us if items are clustered in a specific area of the scale. This can also be problematic, as it can limit the reliability of person measures that fall outside the area where most of the items are located.

Interpretation of the Wright Maps for this initial examination of the CAST may be challenging due to students' opportunity-to-learn. The CA NGSS represent a significant shift in the way science is expected to be taught and learned, requiring both deep knowledge of science content and the ability to use that content in potentially unfamiliar contexts. It may be some time before most California teachers change their instructional practices to account for this change in the science standards. Students may not perform very well on early iterations of the CAST. If there is a mismatch between test item difficulty and student ability, it will not be possible to parse opportunity-to-learn issues from concerns regarding item difficulty. It is nonetheless important to examine the match between test difficulty and student ability estimates. Monitoring changes in relative item



		Science Domain and DCI Strands																											
		Physical Sciences (17 PEs)								Life Sciences (12 PEs)										Earth and Space Sciences (13 PEs)								ETS (3 PEs)	Items per SEP
		PS1		PS2			PS3	PS4		LS1			LS2		LS3		LS4			ESS1		ESS2				ESS3			
	CCC	2	3	1	2	n/a	5	1	2	1	4	5	2	4	1	2	2	3	4	1	3	1	2	3	4	2	4	n/a	
Science and Engineering Practices	SEP 1				X		X																					X	1–4
	SEP 1E					X																							
	SEP 2		X				X	X	X	X	X			X											X				1–7
	SEP 3	X	X	X	X		X																X					X	1–7
	SEP 4														X			X		X		X							2–4
	SEP 5		X																					X					1–2
	SEP 6						X									X		X		X								X	2–8
	SEP 6E						X	X																		X			
	SEP 7				X						X	X	X				X		X		X					X			1–8
	SEP 8																					X				X	X		1–3
	Items per DCI Strand	1–3		1–4			1–4	1–2		1–2			1–2		1–2		1–4			1–2		1–5				1–3		2–4	Total of 32–34 Items
	Items per Domain	8–10								8–10										8–10								2–4	

∗ For scoring and reporting purposes, items written to assess PEs associated with Engineering, Technology, and Application of Science will be assigned to one of the three science domains, depending upon the context of their stimulus.

(See Appendix H for a detailed description of the figure.)

Figure 2.1. PE Distribution for Segment A of the CAST Grade 5 Assessment

```
Higher ability --------------- PERSONS -+- ITEMS-----------------Higher Difficulty
    3                                 +                                 3
                                      |
                                      |
                                      |
        More able Students            |          More difficult items
                                      |
                                      |
                                      |
                                      |
                                   X  |
    2                              X  +                                 2
                                   X  |  X
                                  XX T|
                                 XXX  |
                                 XXX  |  X
                             XXXXXXX  |
                              XXXXXX  |T                Level 4
-----------------------------------------------------------------------------
                             XXXXXXX  |  XX
                    XXXXXXXXXXXXXXXX S|  XXX
                       XXXXXXXXXXXXX  |  XX
    1               XXXXXXXXXXXXXXXX  +  XX                             1
                      XXXXXXXXXXXXXX  |  XXXX
                         XXXXXXXXXXX  |  XXXXXX
               XXXXXXXXXXXXXXXXXXXXX M|S XXXXX
                           XXXXXXXXX  |  XXXXX
                          XXXXXXXXXX  |  XXXXXX
                   XXXXXXXXXXXXXXXXX  |  XXXXXXXX
    Level 3          XXXXXXXXXXXXXXX  |  XXXXXXXXX      Level 3
    -------------------------------------------------------------------------
                             XXXXXXX S|  XXXXXX
                           XXXXXXXXX  |  XXXXXXXXX
    0                        XXXXXXX  +M XXXXXXX                        0
                        XXXXXXXXXXXX  |  XXXXXXX
                                 XXX  |  XXXXXXXXXXX
                                  XX  |  XXXXXXX
                                   X T|  XXXXX
                                      |  XXXXX
                                   X  |  XXXXXXXXXX     Level 2
-----------------------------------------------------------------------------
                                      |S XXX
       Less able Students          X  |  XXXX
                                      |  XX
   -1                                 +  XXX                           -1
                                      |  XX
                                      |  XXX
                                      |
                                      |T
                                      |
                                      |
                                      |
                                      |  X
                                      |  X
   -2                                 +                                -2
                                      |
                                      |  X
                                      |
                                      |  X
                                      |
                                      |
                                      |            Less difficult items
                                      |
                                      |
   -3                                 +                                -3
```

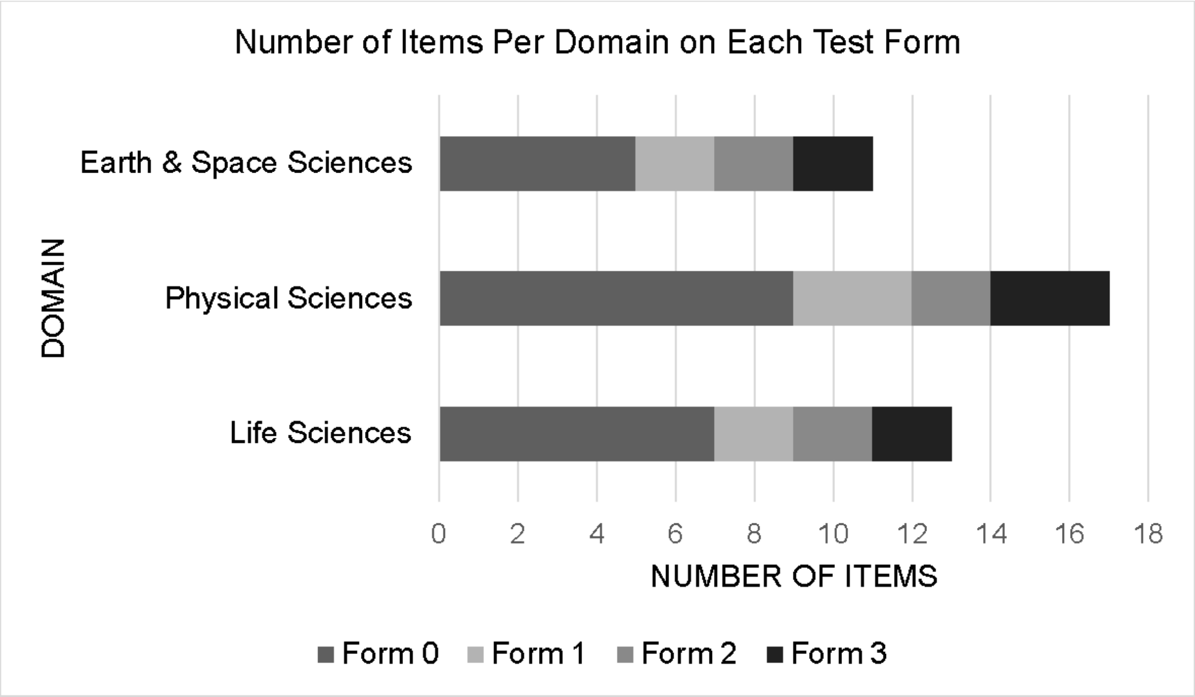
(See Appendix H for a detailed description of the figure.)

Figure 2.2. Example of a Wright Map comparing examinee ability and item difficulty distributions.

difficulties across administrations of CAST could signal shifts in teaching pedagogy and improvements in students' opportunity-to-learn.

Using metadata from the test developer and the de-identified student score file, Wright Maps will be created to depict the distribution of items for the student-level alignment evaluation. Wright Maps can also be produced for each dimension and content domain based on the coding of the items by the test vendor and/or by the workshop panelists.

Alternatively, the items can also be represented in graphics similar to Wright Maps that depict the range of cognitive complexity by dimension and by domain. As an example, Figure 2.3 presents a graph for a fictitious science test with three content domains and four test forms. Form 0 is common, administered to all students, while forms 1–3 are spiraled throughout the population of test takers. The graph shows how many items represent each domain by form, and gives a clear indication of how the matrix design functions to increase the representation of the content domains when considered across forms. The graphic also clearly demonstrates areas where content is distributed unevenly, both overall and by form. These kinds of graphs are often easier to interpret than Wright Maps and will be included in the stand-alone alignment study report (as appendices) to further assist readers' in understanding the results.



(See Appendix H for a detailed description of the figure.)

Figure 2.3. Sample graph illustrating number of items (horizontal axis) addressing each domain (vertical axis) by form.

Schedule of Milestones

Table 2.4 provides a timeline for specific study milestones.

Table 2.4. Schedule of CAST Alignment Study Activities

Description of Study Activities or Deliverable	Time Frame
Collect CAST documentation, secure test material, and metadata from ETS.	July 2018–January 2019
Recruit expert panelists for alignment workshop.	August–December 2018
Develop forms and conduct rating activity to evaluate CAST contractor documentation.	October–December 2018
Develop alignment acceptability criteria for CA NGSS assessment.	September–October 2018
Develop, produce, and quality check all workshop materials, including rating forms.	December 2018–January 2019
Conduct alignment workshop with expert panels.	February 2019
Complete analysis of alignment data.	September 2019
Develop detailed stand-alone report.	September 2019–January 2020
Submit draft stand-alone report.	January 2020
Submit final stand-alone report.	June 1, 2020

This page is intentionally blank.

Chapter 3: Impact on Instruction and Student Learning Case Study

Background

The Impact on Instruction and Student Learning Case Study (hereafter, Impact Case Study) will use a case study approach to deeply investigate and produce a richly detailed summary of the CAASPP System's impact in a modest number of LEAs and schools that use a variety of components of the CAASPP System. By using a case study design, HumRRO can intensively explore a manageable number of schools and LEAs to elicit concrete examples of how and why specific CAASPP System components (described later) are used, their impact on instruction and student learning, as well as the perceived benefits, strengths, and challenges of using them. The Impact Case Study will also provide narrative descriptions of structures that support or are barriers to the successful implementation of CAASPP components and their integration into an effective educational ecosystem. The LEAs and schools included in the Impact Case Study can be considered early adopters of specific components of the full suite of CAASPP System components, and their experiences will be useful to LEAs and schools that may not be implementing some or all of the available CAASPP components. The Impact Case Study will also inform the CDE about particular challenges and suggest where potential future improvements to the system can be made.

Creswell (1998) described a case study as an appropriate research approach when one is interested in the in-depth study of a "case" bounded in time or place. Patton (2015) noted that a "case" can be many different things, depending on the focus and field of study. Moss and Haertel (2016) use the label "Small N or Comparative Case Studies" (CCS) for studies with "more than one case, but typically fewer than fifty, purposively chosen so as to illuminate the question or phenomenon of interest. Typically, cases are chosen so as to contrast with respect to some set of key features. In CCS, within-case analyses are supplemented by cross-case comparisons, which help to support generalization."

For this study, a case is defined as an LEA that has fully implemented the CAASPP System for a minimum duration of two school years (the study year and the year prior). To conduct a case study, one should gather a large amount of data to provide an in-depth picture of the "case" (Creswell, 1998). Like other forms of qualitative research, case studies tend to rely on use of inductive reasoning, meaning rather than beginning with specific hypotheses, we may start with general research questions and develop ideas and themes from the data itself (Creswell & Plano-Clark 2007). Consistent with these approaches, our study methods will rely on inductive reasoning guided by a set of research questions. We plan multiple types of data collection, as described further in this section, to provide an in-depth look at the implementation of CAASPP for a selection of LEAs and a sample of their schools. For example, we will include qualitative interviews that "sacrifice uniformity of questioning to achieve fuller development of information… and because the fuller responses obtained by the qualitative study can not be easily categorized, their analysis will rely less on counting and correlating and more on interpretation, summary, and integration" (Weiss, 1994).

Selection of Impact Case Study Participants

Implementing a sampling plan to select and recruit LEA and school participants will be the first activity of the Impact Case Study.

Sampling Plan

The sampling plan is a multi-step process to secure six eligible LEAs for the 2018–19 school year case studies:

1. With CDE input, develop an operational definition of full implementation of the CAASPP System. See Appendix C for the definition of *full CAASPP implementer*.

2. Obtain Interim Assessment Block (IAB) and Interim Comprehensive Assessment (ICA) usage data—by LEA, school, and test—from the 2017–18 school year from ETS. Usage data will indicate manner of test administration (standardized or nonstandardized) and whether hand scoring is involved.

3. Obtain Digital Library login information from the 2017–18 school year from CDE.

4. Prescreen LEAs eligible for recruitment by applying the CDE-approved operational definition of full implementer (e.g., eliminate LEAs that did not administer any IABs in the 2017–18 school year) to all LEAs in the state. At this stage, the CDE may ask HumRRO to eliminate some LEAs based on their involvement in other studies or due to other valid concerns.

5. Develop a very brief Web-based Eligibility Survey, informed by the full implementer criteria, to further screen LEAs for recruitment (e.g., ask about use of CDE CAASPP resources, the IA Reporting System, Digital Library; degree of staff turnover). See Appendix D for survey questions.

6. Administer the online Eligibility Survey to CAASPP Coordinators for the prescreened LEAs.

7. Apply the CDE-approved full implementer criteria and other screening criteria (e.g., preference for including users of the IA Reporting System and excluding LEAs that experienced recent high turnover) to survey responses to identify the sampling frame for eligible LEAs.

8. Stratify the eligible pool of LEAs, using recent CAASPP summative assessment data (ELA and mathematics) and demographic characteristics, and apply the criteria for selecting LEAs to be recruited. See Appendix E for criteria for selecting LEAs.

Recruitment of Impact Case Study Participants

The CDE will play an important role in helping HumRRO gain participation of targeted LEAs for the study. LEAs who join the study must commit to full engagement throughout the school year and demonstrate a willingness to complete the assigned tasks. Each

LEA must arrange for staff from two or three schools to participate. If available, we will seek one high school, one middle school, and one elementary school to participate, though not all LEAs may include eligible schools at each level. HumRRO will include one direct-funded charter LEA, which may have only a single school.

HumRRO will develop a Memorandum of Understanding (MOU) to be agreed to by each LEA, defining expectations of the LEA, its participating schools, and HumRRO. The MOU will identify a Study Point of Contact (POC) for the LEA and for each school, and will include a statement that, with permission, allows sharing of specific information across participating LEAs.

When recruiting LEAs, the time commitment for POCs as well as for teachers and other LEA and school staff will be estimated. We anticipate each participating LEA and school POC will provide 8 hours of assistance in a typical month, and no more than 12 hours during a single month, to conduct this evaluation. To support the administrative tasks of POCs, HumRRO will offer each POC a fixed honorarium of $100 per month of data collection, for a period of nine months. HumRRO will work with the CDE to explore other possible benefits to offer LEAs as appreciation for participation in the study, such as waived fees for CAASPP professional development opportunities. POC activities include:

- Assist with logistics for HumRRO site visits (e.g., arrange for focus groups and interviews with LEA and school leaders who use CAASPP components)
- Participate in interviews or focus groups (one in-person, one virtual)
- Conduct POC polling phone calls or written responses (monthly)
- Provide information on use of interim assessments and digital library of formative tools
- If agreed to by the LEA, provide interim assessment data and climate survey data to HumRRO – either through access to the online system, or downloaded and de-identified. If provided as downloaded data, this information will be shared through a secure FTP site.

In addition to LEA and school POCs, the data collection effort will also require participation from teachers at each selected school. Teacher activities include:

- Participate in one in-person focus group (i.e., a selection of teachers who use CAASPP components)
- Provide information to the school POC in response to monthly polling questions (may not require a response from all teachers every month)

For the 2019–20 school year, selection of a new set of six LEAs will be based on study needs or gaps identified during the first year, in consultation with the CDE. We leave open the possibility that one or more LEAs participating in the 2018–19 evaluation will continue on for an additional year, if there are justifications for doing so and the LEA(s) are willing.

Research Questions

The CAASPP Theory of Action (CDE, 2018a) was used as a guide to define the research questions for this two-year case study, and it will also guide development of data collection instruments, analysis of the obtained data, and reporting of results. Answers to each research question will be informed by teachers, school leaders, and LEA staff, with the intent that answers to questions may differ depending on the level. Similarly, several questions address both classroom instruction and student learning; however, the impacts on each aspect will be examined separately. In addition, we will seek information about unanticipated barriers to implementation.

The following draft research questions address the three Smarter Balanced components included in the CAASPP System (i.e., summative assessments, interim assessments, and the Digital Library) and the CAASPP Theory of Action; changes in instruction; and changes in learning. HumRRO plans to collect narrative descriptions of examples from LEA and school educators as evidence to answer many of the research questions. The new science summative assessments will not be administered operationally until 2018–19; therefore, we have excluded CAST and CAA for Science from the research questions but may possibly include them in year two of the study.

Questions related to the full suite of Smarter Balanced components (Summative assessments, Interim assessments, and Digital Library of formative tools) included in the CAASPP System:

1. What are the characteristics and contexts of sampled schools/LEAs that have implemented the full suite of Smarter Balanced components?

2. How does implementation of Smarter Balanced components for ELA/literacy and mathematics vary across schools/LEAs? What instructions and supports are provided to educators for implementing the components?

3. What aspects of Smarter Balanced components are perceived as most beneficial for improving classroom instruction and student learning across schools/LEAs?

4. What changes to the Smarter Balanced components and supporting resourcesdo LEA and school staff believe would improve support for their use of CAASPP components to improve classroom instruction and student learning?

5. How do educators/schools/LEAs use and integrate results from the summative, interim, and formative assessment resources for each content domain (ELA/literacy and mathematics) with each other and with other measures to enhance classroom instruction and student learning? What challenges are faced and how are they overcome?

6. How do students from schools that use the full suite of Smarter Balanced components perceive classroom opportunities to learn about summative assessment item types and topics for each content domain (ELA/literacy and mathematics)?

Questions related to Smarter Balanced Summative Assessments Only:

7. How do educators/schools/LEAs use summative assessment data—including, but not limited to, information about student proficiency levels and progress towards college- and career-readiness—in ELA/literacy and mathematics to inform classroom instruction and make decisions?

Questions related to Smarter Balanced Interim Assessments and Digital Library of Formative Tools Only:

8. What interim assessments are used for ELA/literacy and mathematics for schools/LEAs that have implemented the full CAASPP System, and at what grade levels and frequency?

9. What decision-making processes are used by educators/schools/LEAs to determine what ELA/literacy and mathematics interim assessments to use, who should administer them, and how frequently they should be administered?

10. To what extent have educators/schools/LEAs incorporated ELA/literacy and mathematics IABs into their classes? What, if any, classroom assessments have been replaced in the process? Why, and what are the implications?

11. How do educators/schools/LEAs use information from ELA/literacy and mathematics interim assessments to track individual student progress and/or inform classroom instruction?

12. How is information on student/school/LEA performance on ELA/literacy and mathematics interim assessments used at the school/LEA level to determine the effectiveness of practices and curricular materials for teaching the targeted standards (i.e., CCSS)?

13. How is the Smarter Balanced Digital Library of formative tools used to improve classroom instruction (e.g., share information with students to help them monitor their own performance; better align instruction, curricula, and assessments)?

Data Collection

The following data will be gathered from extant sources:

- *Statewide assessment data*. Records of summative assessment administration results, counts of interim assessments administered in each content domain (including interim assessments that require hand scoring).

- *Demographic records*. Data with LEA characteristics, including student population, number of schools, student demographics, and historical achievement on summative assessments.

- *School Climate data*. School climate data (e.g., perceived student, teacher, and parental engagement; and academic expectations and rigor) will be collected from each LEA's three participating schools, if available.

- *Assessment data from LEAs and schools*. Data on interim assessment use and scores, by classroom and teacher, for all interim assessment administrations in each LEA's three participating schools. Expect to collect at three points from each school POC during the year.

- *Smarter Balanced Digital Library data*. Data on account user login by LEA, school, and individual.

Data to be generated during this study include:

- *Data from Eligibility Survey*. LEA CAASPP coordinators' responses to online survey to screen for LEAs that are full implementers of the Smarter Balanced components of the CAASPP System.

- *Data from in-person visits to LEAs and schools*. Detailed notes and audio recordings of LEA leaders', school leaders', and teachers' responses to interview and focus group questions about the use of Smarter Balanced components of the CAASPP System. Artifacts such as professional learning community (PLC) meeting schedules, teacher lesson plans that incorporate formative tools, school calendars, handouts from student assessment data review meetings, and professional development materials. HumRRO will submit draft interview and focus group protocols (topic guides) to the CDE for review in advance of the first LEA site visit. HumRRO also provided the draft protocol to the CAASPP TAG members in advance of the September TAG meeting for review and feedback from CAASPP TAG members.

- *Data from monthly phone or e-mail polling of LEA and school POCs, who will collect LEA leader, school leader, and teacher responses*. Narrative responses to or discussion of one to three questions per month related to use of Smarter Balanced components. Questions may target particular types of respondents (e.g., elementary teachers, middle school principals, high school math teachers, or LEA curriculum and instruction staff) or may be the same for all types. We will submit a courtesy copy of questions to the CDE at least one week in advance but will not require approval before polling. This will allow the CDE to monitor the types of questions being asked and have an opportunity to suggest revisions and/or additional questions, but will not unnecessarily burden the CDE. See Appendix F for examples.

- *Data from end of school year Web-based focus groups with LEA and school POCs*. Detailed notes and audio recordings of LEA and school POC groups' responses to focus group questions, including generic and LEA-specific questions (e.g., use of different types of available resources, such as Digital Library Connections, and why the resource was used and if the resource was effective).

- *Data from Student Focus Groups led by school POCs*. We will ask each school POC to conduct one focus group during the year with a sample of students in their school to gain their perspectives on the various aspects of the CAASPP system from the student level. The focus group should include students from more than one grade, and students of mixed ability levels. HumRRO will provide a focus group protocol specific to the school type (elementary, middle, or high).

- *Data from Summative Assessment Student Questionnaires*. A small number of multiple-choice or multiple-select questions will be administered to all students following the 2020 ELA/literacy and mathematics Smarter Balanced Summative Assessments, as timing precludes administration in 2019. HumRRO will provide the questions and response options.

Data Analysis Summary

The questions included on all protocols will be aligned with the research questions listed earlier. The focus group and interview questions will be primarily open ended. As appropriate, questions will include associated probes to ensure an accurate and consistent understanding of, and thorough answers to, the questions. Survey questions will be primarily multiple choice or scaled response items. See Appendix G for detailed descriptions of analyses planned for each type of data collected for the Impact Case Study. A summary of the data analyses planned is as follows:

- *Qualitative analysis of interviews and focus group data and collected artifacts.* Analyses will involve a cycle of iterative steps: gathering data, examining data, comparing prior data to new data, writing up field notes before conducting more interviews and focus groups, and making plans to gather new data through revisions to the protocols. We will analyze qualitative data by systematically and progressively narrowing the patterns and themes that emerge. In addition to identifying common themes across LEAs and schools, unique examples and descriptions will be selected to address each research question.

- *Analysis of POC Polling Data.* First, the POC polling data will be analyzed as each set of monthly responses is received from LEA and School POCs. We will code responses by LEA identification, respondent group (i.e., elementary school, middle school, high school, LEA), disposition (e.g., positive, negative, neutral), and theme. Individual responses may be coded multiple times; for example, if a single narrative response contains both positive and negative elements. In preparation for end-of-school-year focus group protocol development, each LEA’s cumulative POC polling data will be analyzed. In addition to coding for commonalities, we will select rich, descriptive examples of uses of Smarter Balanced components that address our research questions. In preparation for annual reports, previously analyzed themes from the POC polling data will be analyzed by respondent group, across LEAs. Overall trends as well as trends disaggregated by school characteristic (e.g., high or low achievement, school size, high or low percentages of disadvantaged students) will be evaluated.

- *Analysis of Interim Assessment Data.* Descriptive statistics will characterize the extent of use, type of use (e.g., content area, grade level, standardized versus non-standardized administration, IABs versus ICAs), timing of use, and level and trends in student scores. Analyses will include a static summary of all use to date, as well as patterns of use during the school year. Results will be summarized by LEA and overall to facilitate the development of POC polling questions and end-of-year focus groups.
- *Analysis of Use of the Digital Library*. We will use descriptive statistics to summarize the nature and frequency of use of different types of resources included in the Digital Library. We will use qualitative analyses to identify and describe reasons educators seek available resources, and the effectiveness of these resources.
- *Review of School Climate Survey Data:* If available for schools selected in this study, HumRRO will collect extant climate survey data for each year beginning with the 2016–17 school year. We will use this information to help describe our sample of LEAs and schools to better understand underlying characteristics (e.g., perceived student, teacher, and parental engagement; and academic expectations and rigor) that may impact or interact with choices made regarding the use of Smarter Balanced components.

Table 3.1 provides a cross reference between data to be analyzed and the draft research questions for the Impact Case Study (RQ 1–13). A “Yes” or “No” indicates for each data source whether it is applicable to each research question. This evaluation will primarily consist of qualitative analyses to address each question. We will incorporate quantitative information from ETS interim assessment usage data, CDE Digital Library login information, school climate surveys (if available), and any POC polls collected over the evaluation to help us describe the LEAs and charter in our sample. Each data collection activity will be designed to address one or more research questions.

Table 3.1 Cross Reference of Data Sources and Draft Impact Case Study Research Questions

Research Question (abbreviated)	Eligibility Survey	Initial Site Visits	Ongoing POC Polling	End-of-School-Year Focus Groups	Interim & Formative Data	Climate Survey Data	Student Focus Groups with School POCs	Student Question-naires (2020)
1. What are the characteristics and contexts of sampled schools/LEAs that have implemented the full suite of Smarter Balanced components?	Yes	Yes	No	No	No	Yes	No	No
2. How does implementation of Smarter Balanced components for ELA/literacy and mathematics vary across schools/LEAs? What instructions and supports are provided to educators for implementing the components?	No	Yes	Yes	Yes	Yes	No	No	No
3. What aspects of Smarter Balanced components are perceived as most beneficial for improving classroom instruction and student learning across schools/LEAs?	No	Yes	Yes	Yes	No	No	Yes	No
4. What changes to the Smarter Balanced components and supporting resources do LEA and school staff believe would improve support for their use of CAASPP components to improve classroom instruction and student learning?	No	No	Yes	Yes	No	No	Yes	No

Research Question (abbreviated)	Eligibility Survey	Initial Site Visits	Ongoing POC Polling	End-of-School-Year Focus Groups	Interim & Formative Data	Climate Survey Data	Student Focus Groups with School POCs	Student Question-naires (2020)
5. How do educators/schools/LEAs use and integrate results from the summative, interim, and formative assessment resources for each content domain (ELA/literacy and mathematics) with each other and with other measures to enhance classroom instruction and student learning? What challenges are faced and how are they overcome?	No	Yes	Yes	Yes	Yes	No	No	No
6. How do students from schools that use the full suite of Smarter Balanced components perceive classroom opportunities to learn about summative assessment item types and topics for each content domain (ELA/literacy and mathematics)?	No	No	No	No	No	No	Yes	Yes
7. How do educators/schools/LEAs use summative assessment data—including, but not limited to, information about student proficiency levels and progress towards college- and career-readiness—in ELA/literacy and mathematics to inform classroom instruction and make decisions?	No	Yes	Yes	Yes	No	No	No	No

Research Question (abbreviated)	Eligibility Survey	Initial Site Visits	Ongoing POC Polling	End-of-School-Year Focus Groups	Interim & Formative Data	Climate Survey Data	Student Focus Groups with School POCs	Student Question-naires (2020)
8. What interim assessments are used for ELA/literacy and mathematics for schools/LEAs that have implemented the full CAASPP System, and at what grade levels and frequency?	No	No	No	No	Yes	No	No	No
9. What decision-making processes are used by educators/schools/LEAs to determine what ELA/literacy and mathematics interim assessments to use, who should administer them, and how frequently they should be administered?	No	Yes	Yes	Yes	Yes	No	No	No
10. To what extent have educators/schools/LEAs incorporated ELA/literacy and mathematics IABs into their classes? What, if any, classroom assessments have been replaced in the process? Why, and what are the implications?	No	Yes	Yes	Yes	Yes	No	No	No
11. How do educators/schools/LEAs use information from ELA/literacy and mathematics interim assessments to track individual student progress and/or inform classroom instruction?	No	Yes	Yes	Yes	Yes	No	No	No

Research Question (abbreviated)	Eligibility Survey	Initial Site Visits	Ongoing POC Polling	End-of-School-Year Focus Groups	Interim & Formative Data	Climate Survey Data	Student Focus Groups with School POCs	Student Question-naires (2020)
12. How is information on student/ school/LEA performance on ELA/literacy and mathematics interim assessments used at the school/LEA level to determine the effectiveness of practices and curricular materials for teaching the targeted standards (i.e., CCSS)?	No	Yes	Yes	Yes	Yes	No	No	No
13. How is the Smarter Balanced Digital Library of formative tools used to improve classroom instruction (e.g., share information with students to help them monitor their own performance; better align instruction, curricula, and assessments)?	No	Yes	Yes	Yes	Yes	No	Yes	No

Reporting Results

The following reports will be developed for this study:

- *Annual evaluation reports*. The Impact Case Study report will be included as a chapter within the corresponding annual Independent Evaluation reports and will describe the work completed during the prior school year as it pertains to each research question.
- *Annual stand-alone Impact Case Study reports*. Following each evaluation year, we will produce an Impact Case Study Report to provide a self-contained account of the study for the six LEAs studied that year. It will fully describe sample selection, data protocols and methods, analyses, findings, and recommendations. Improvements and other changes to the CAASPP System will be clearly identified.
- *Comprehensive final report*. The evaluation will be fully described in the Comprehensive Final Report. Details will include the research questions, study design, data collection procedures, analyses, findings, contextual factors, limitations, challenges, recommendations for the CDE regarding improvement of the CAASPP System, and implications of the findings for future implementation. Data collection instruments and protocols will be included in appendices.

Schedule of Milestones

Table 3.2 provides a preliminary timeline for specific study milestones.

Table 3.2 Schedule of Impact Case Study Activities

Description of Study Activities or Deliverable	Evaluation Year	Time Frame
Administer Eligibility Survey, recruit LEAs and associated schools, and establish POCs	2018–19 and 2019–20	July–October
In-person visits to LEAs and schools	2018–19 and 2019–20	October–November
Polling of LEA and school POCs	2018–19 and 2019–20	November–May, Monthly
End-of-school-year focus groups	2018–19 and 2019–20	April–May
Student Questionnaire in summative assessments for ELA and Mathematics	2019–20	February–June
Focus Group with 2019 LEA and School POCs	2019–20	April–May
Tracking interim assessments and use of formative resources	2018–19 and 2019–20	November, February, and May
Summative test results	2018–19 and 2019–20	August–September
Develop and submit stand-alone Impact Case Study Report	2018–19 and 2019–20	September

This page is intentionally blank.

Chapter 4: California Alternate Assessment for Science Alignment Study

Overview

The alignment study for the CAA for Science assessment will be very similar in approach to that of the alignment study for CAST. The study will aim to provide validity evidence for the CAA for Science as a measure of science achievement for the population of students for which it was designed—students with severe cognitive disabilities. This study will focus on links between the Core Content Connectors, which are based on the CA NGSS and were developed to define the science construct(s) to be measured, and the test forms and test items that were developed to assess them.

The CAA for Science is intended to function similarly to an “end-of-instruction” rather than an “end-of-year” summative assessment. The test will be given as three separate domain sessions, one for life sciences, one for physical sciences, and one for earth and space sciences. Administration is not tied to an administration window, as for a typical summative assessment, and 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 sessions will be aggregated to generate an overall science score. The CAA for Science is administered in grades five, eight, and once in high school. The high school assessment may be administered in grade ten, eleven, or twelve. The CAA for Science is designed such that each session is represented by one complex embedded performance task. Two Core Content Connectors are represented in each task, and the task is expected to have a mix of low, medium, and high complexity test items (or score points). Obviously, the two connectors cannot represent the full breadth of the Core Content Connectors available. There are 20, 24, and 28 Core Content Connectors for grades five, eight, and high school, respectively. The CAA for Science is expected to rotate connectors from year to year to build to fuller representation of the content over time. This assessment is not expected to guide instruction based on a single administration.

Alignment studies for an assessment with this structure must approach evidence gathering in two ways. First, it must demonstrate the aggregation of the three sessions provides an adequate representation of the science content specified by the Core Content Connectors. This alignment task supports the overall score and is the key evidence required by 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 this 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 will be collected to demonstrate the science Core Content Connectors are adequately represented, and those same data will be used to ensure the content domains are evenly represented.

Typical alternate assessment alignment studies, such as those guided by the Links for Academic Learning method (Flowers, Wakeman, Browder, & Karvonen, 2007), focus as much on evaluating the link between the regular education standards and the alternate standards as on the assessment itself. This study will not reevaluate the standards-to-alternate standards link but will favor a comprehensive examination of how the CAA for Science tasks and forms support the claim to be made about student performance. Documentation surrounding creation of the Core Content Connectors will be referenced, but links between them and the CA NGSS will not be reevaluated.

Typical alternate assessment alignment studies also tend to reevaluate the accessibility of the test items for various disability categories (e.g., vision impaired, hearing impaired, autism) as well as the communication level of the tested students. While accessibility is vitally important to an alternate assessment, it is not an activity conducted as part of an alignment study. The focus during this study is on alignment rather than accessibility; however, studies that demonstrate accessibility will be referenced, rather than our repeating evaluations of accessibility.

The research questions and methodology for this alignment study were designed specifically to address the structure and design of the CAA for Science and the results to be reported.

Research Questions

Activities conducted for the CAA for Science Alignment Study are 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 assessment tasks link to the CA NGSS Core Content Connectors?

4. How well do the CAA for Science assessment tasks cover the range of cognitive complexity of the CA NGSS Core Content Connectors?

To evaluate the evidence collected to answer the research questions, acceptability criteria will be developed specific to CAA for Science results. Similar to the process described for the CAST Alignment Study, the CAA for Science acceptability criteria will be developed prior to data collection and analyses to guard against inadvertently adjusting the acceptability criteria to better fit the findings. Draft criteria, along with supporting documentation, will be submitted for review to a small group of experts external to HumRRO. This group will review the draft criteria, comment on their adequacy, suggest adjustments or revisions, and reach consensus on the final set of acceptability criteria for the CAA for Science Alignment Study. These criteria will be used to determine the appropriateness of the evidence statements and interpretations of results described by the test developer. The criteria will be submitted for review by the CDE and TAG. A member of the CAASPP TAG with experience and expertise assessing special populations will critique our approach to ensure the nature of the CAA for Science is considered in all alignment judgments.

Data Collection Methods and Measurement Instruments

Extant data to be collected for this study include:

- *CAA for Science test developer documentation of how alignment was considered during test development.* This should include item and form development guidance, test blueprint and item specifications; item tryout and review procedures, procedures for reviewing and addressing item tryout information; validity and reliability evidence for the anticipated test form; and guiding documents that illustrate the overall goals and philosophy underlying the assessment such as a theory of action, interpretive argument, or other similar documents.

- *CAA for Science item metadata.* Metadata are to include item parameters, *p*-values, cognitive complexity (DOK or similar, if available), and coding to the CA NGSS. Data may come from field test or operational administrations.

- *CAA for Science items that will be operationally administered.* Access to all test items must be in the same format as they will be viewed by students.

- *Student-level file of CAA scores for the overall test and for any reported sub-scores.*

- *Information on how CAA results are reported at the student level.* This information should include procedures for assigning individual students to performance levels, performance level descriptors, and any reporting categories used in aggregate level reports.

Data to be generated during this study include:

- *Independent ratings of strength of evidence of alignment considerations in test developer documentation*. A customized rating form will be developed, guided by the *Standards for Educational and Psychological Testing* and federal peer review guidance. Three researchers will use the form to evaluate the documentation.

- *Expert panelists' ratings of CAA for Science items.* Three panels (elementary, middle, and high school panels), each comprised of six teachers of science, will evaluate CAA items during an in-person workshop. Ratings of what standards the items assess for the three CA NGSS dimensions (DCI, CCC, SEP), and ratings of the items' cognitive complexity requirements will be collected.

Evaluation of test developer documentation

A rating form will be created to capture reviewers' evaluation of test developer documentation of how alignment was considered in test development. The form's design will be guided by the *Standards for Educational and Psychological Testing*, which describe requirements for developing, reviewing, and trying out test items (mostly Chapter 4) as well as requirements for documenting the processes used (Chapters 4, 6, and 7). Special attention will also be given to the section on alternate assessments in Chapter 12; however, it is important to note the testing standards indicate that "Alternate assessments in education should be held to the same technical requirements that apply to regular large-scale assessments (pg. 190)." Alignment evidence for alternate assessments should be held to rigorous acceptability criteria.

Additional components will be added to the form as necessary to ensure clear parallels to the federal peer review guidance. Testing standards and rating components will be selected to support the claims structure established for the CAA for Science. The draft rating form will be submitted to the CDE Contract Monitor for review and approval.

The rating form HumRRO develops to evaluate the CAA for Science test developer's documentation is expected to be similar in format and use a similar rating scale to that presented in Table 2.2 for the CAST Alignment Study section, though additional standards relevant to alternate assessments will likely be included. Similarly, the process described for researcher evaluation of evidence will also be applied to evaluation of the test developer's documentation for the CAA for Science.

CAA for Science Alignment Workshop: Evaluating Items for Alignment with the CA NGSS Core Content Connectors

HumRRO will conduct a two-day workshop with three panels of educators. We will work collaboratively with the CDE Contract Monitor to recruit and select a total of 18 educators to serve on three CAA for Science alignment review panels (one elementary, one middle, and one high school panel). Criteria for panelists will include familiarity with the CA NGSS. The high school panel must include at least two biology teachers, one chemistry teacher, and one physics teacher. At least one educator responsible for teaching the earth and space standards will be recruited for each panel; this educator may also serve in one of the previously mentioned roles (e.g., physics teacher) or may be a dedicated earth/space science educator. Four special education teachers and two science content experts will be recruited for each panel. We will ensure at least one educator from each panel also has experience teaching ELs. The two high school panel content teachers should not have the same domain specialty (e.g., life sciences, chemistry).

HumRRO will work collaboratively with the CDE Contract Monitor to determine dates and the specific location for the workshop. HumRRO will make all logistical, equipment, and travel arrangements, as described for the CAST Alignment Study. Panelists will be required to sign nondisclosure agreements as a condition of participation. Panelists will be reimbursed per diem expenses and be paid an honorarium or substitute teacher costs will be reimbursed to their districts (according to LEA-specific costs).

The alignment workshop will be structured to include group-level activities and individual rating tasks. Whole group training (all panelists in the same room) will occur on the first day. The training will provide an overview of the task, discuss alignment concepts, review rules regarding confidentiality and data security, and orient the panelists to expectations for the remainder of the workshop. This training will include an overview of the cognitive complexity ratings used by the test developer and an overview of administration policies and procedures, including common accommodations and accessibility. In small panel groups, each alignment task will be preceded with targeted training that will be conducted by the facilitator.

The following sequence of activities is planned as a high-level, tentative agenda for the CAA for Science alignment workshop:

1. Introduce participants to the method and provide overview of alignment concepts (large group training).

2. Review test administration policies and procedures, including common accommodations and accessibility features (large group training).

3. Participants create independent ratings of CAA for Science task items (content match and cognitive complexity) (small group).

4. Facilitators help independent panelists compare their ratings with each other (outlier analyses) and with metadata and cognitive complexity data supplied by the test developer (small group).

5. Facilitators lead workshop evaluation, debrief, and dismissal (small group).

Panelists will make ratings regarding what Core Content Connectors each CAA for Science item assesses and determine if the item also assesses a SEP or CCC. The rating form will be a slightly modified version of the one presented for the CAST Alignment Study. Finally, panelists will rate the item’s cognitive complexity level using the low, medium, and high scale from the test blueprint. These data will be automatically tabulated and prepared for the next morning’s activities by the facilitator.

On the second day, panelists will review their data from Day 1 in comparison to one another and to the metadata from the test developer. They will begin by conducting outlier analyses. Each item will be reviewed by the facilitator and discrepant ratings will be discussed among the group. The facilitator will lead a consensus building discussion and panelists will be allowed to change their ratings based on this discussion. Panelists’ data will be preserved after each successive step. Once the panelists have completed outlier analyses, they will be shown the metadata from the test developer. They will note discrepancies and will again be allowed to revise their ratings based on discussion.

Panelists will note any substantive disagreement between their own ratings and the metadata. Panelists will generate a consensus statement regarding the adequacy of the CAA for Science to represent the Core Content Connectors sufficiently to generate a summary judgment of student performance in science across the domains.

Next, panelists will use the same scale as the test developer to indicate a cognitive complexity rating for each item from all CAA for Science tasks. The developer of the test blueprint recognized any CA NGSS standard and any Core Content Connector may be assessed at a variety of cognitive complexities and, to represent the content, it is important to provide access at several different levels. The blueprint requires items with cognitive complexity ratings of 1, 2, and 3 for each Core Content Connector. Judgment of how well the assessment tasks cover the range of cognitive complexity will be based on whether the full range of cognitive complexity ratings are indicated for each item.

Data Analyses

Table 4.1 summarizes the data that will be analyzed to answer each research question.

Table 4.1. CAA for Science Alignment Study Research Questions and Main Data Sources

Research question	Data to be analyzed
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?	Independent rating scale data from SEPT standards and peer review guidance.
2. To what extent do the test forms and test items for the CAA for Science reflect the test design and test blueprint?	Depictions of CAA for Science test content based on test developer's metadata compared to test design documents and test blueprint; expert panelists' item ratings.
3. To what extent do the CAA for Science assessment tasks link to the CA NGSS Core Content Connectors?	Expert panelists' item ratings.
4. How well do the CAA for Science assessment tasks cover the range of cognitive complexity of the CA NGSS Core Content Connectors?	Expert panelists' cognitive complexity ratings.

The adherence of the test items and forms to the design and blueprint will be judged in two ways. First, the test developer's metadata will be evaluated to verify it reflects the test design and blueprint in terms of the Core Content Connectors assessed, the number of points per task, and the complexity of the items within a domain session. If the item metadata include coding to indicate science dimensions, HumRRO will also verify that items are not clustered to measure any single dimension within or across domains.

The metadata provided by the test developer will be compared to data generated by the panels of educators who review test items and forms.

Draft Schedule of Study Milestones

Table 4.2 provides a preliminary timeline for specific study milestones; the timeline will be updated when further information about the operational assessment is available.

Table 4.2. Draft Schedule of CAA for Science Alignment Study Activities

Description of Study Activities or Deliverable	Time Frame
Develop alignment acceptability criteria.	July 2019
Collect CAA for Science documentation, secure test material, and metadata from ETS.	September–October 2019
Recruit panels for workshop and plan workshop logistics.	August–September 2019
Conduct alignment workshop with expert panels.	October 2019–January 2020
Analyze alignment data.	February 2020
Develop and submit draft detailed stand-alone report on CAA for Science Alignment Study.	March 2020
Submit final stand-alone report on CAA for Science Alignment Study.	June 1, 2020

The study will conclude with development and production of a CAA for Science Alignment Study Report. The report will describe the methodology in detail, include all results (e.g., alignment statistics), and make any conclusions or recommendations that are supported by the data. The report of workshop results will include an assessment of the level of agreement among raters within each panel and overall agreement with the test developer’s metadata. Any areas of significant disagreement will be reported to the test developers for consideration and comment. The report will include a nontechnical executive summary that should be accessible to most audiences. The main report will include sufficient technical details to allow for appropriate scrutiny of results and conclusions, and it will be suitable for submission for federal peer review. HumRRO will provide an electronic appendix containing item-level data that could be used to flag specific items for additional review. All electronic data will be made available to both CDE and the test developer. Only individual raters’ identification will be redacted.

This page is intentionally blank.

References

Achieve. (2018). Criteria for Procuring and Evaluating High-Quality and Aligned Summative Assessments. Retrieved from https://www.achieve.org/files/Criteria03202018.pdf

California Department of Education. (2018a). Appendix C: Theory of Action for CAASPP and the Smarter Balanced Assessment System. In *Request for Proposals (RFP) Independent Evaluation of the California Assessment of Student Performance and Progress (CAASPP) - CN170400*.

California Department of Education. (2018b). 2017–18 CAST Academy Materials. CAST Grade 5 Item Specifications. Retrieved from http://www.caaspp.org/training/caaspp/materials.html

Callingham, R., & Bond, T. (2006). Research in Mathematics Education and Rasch Measurement. Editorial in *Mathematics Education Research Journal*, 18, 2, 1-10.

Creswell, J.W. (1998). *Qualitative inquiry and research design: Choosing among five traditions*. Thousand Oaks, CA: Sage Publications.

Creswell, J.W., & Plano Clark, V.L. (2007). *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage Publications.

ETS. (2017). *California Science Test Blueprint*. Approved by the State Board of Education on November 8, 2017. (Revised 12/1/2017).

Flowers, C., Wakeman, S., Browder, D., & Karvonen, M. (2007). *Links for academic learning: An alignment protocol for alternate assessments based on alternate achievement standards.* Charlotte, NC: University of North Carolina at Charlotte. Retrieved from: http://www.naacpartners.org/LAL/documents/NAAC_AlignmentManualVer8_3.pdf

Moss, P. A., & Haertel, E. H. (2016). Engaging methodological pluralism. In D. Gitomer and C. Bell (Eds.), Handbook of research on teaching (5th ed.), (pp. 127–247). Washington, DC: AERA.

National Research Council. (2014). *Developing Assessments for the Next Generation Science Standards*. Washington, DC: The National Academies Press. https://doi.org/10.17226/18409.

Patton, M. Q. (2015). *Qualitative research & evaluation methods: Integrating theory and Practice* (4th ed.). Thousand Oaks, CA: Sage Publications.

Weiss, R. (1994). *Learning from Strangers: The Art and Method of Qualitative Interview Studies*. New York, New York: The Free Press.

Wilson, M., & Draney, K. (2002). *A technique for setting standards and maintaining them over time*. In S. Nishisato, Y. Baba, H. Bozdogan, & K. Kanefugi (Eds.), Measurement and multivariate analysis (Proceedings of the International Conference on Measurement and Multivariate Analysis, Banff, Canada, May 12-14, 2000), pp 325-332. Tokyo: Springer-Verlag.

Yarbrough, D. B., Shulha, L. M., Hopson, R. K., & Caruthers, F. A. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Thousand Oaks, CA: Sage.

Glossary of Acronyms

Acronym	Gloss
CAA	California Alternate Assessment
CAASPP	California Assessment of Student Performance and Progress
CA NGSS	NGSS for California Public Schools, Kindergarten through Grade Twelve
CAST	California Science Test
CAT	Computer-adaptive test
CCC	Crosscutting Concepts (CA NGSS)
CDE	California Department of Education
DCI	Disciplinary Core Ideas (CA NGSS)
DOK	Depth of knowledge
EC	California Education Code
EL	English learner (student)
ELA	English language arts/literacy
ESSA	Every Student Succeeds Act
IAB	Interim Assessment Block
ICA	Interim Comprehensive Assessment
LEA	Local educational agency
NGSS	Next Generation Science Standards
PE	Performance expectations (CA NGSS)
PII	Personally identifiable information
PT	Performance Task
SEP	Science and Engineering Practice (SEP)
TAG	CAASPP Technical Advisory Group

This page is intentionally blank.

Appendix A: Theory of Action for CAASPP and the Smarter Balanced Assessment System

The primary theory of action for the CAASPP program is that the components of the system work together to accurately assess student achievement relative to grade level curriculum standards, and provide information to educators to improve instruction, and thereby improve student achievement. The Smarter Balanced Assessment System has three components: summative assessments, designed for accountability purposes; interim assessments, designed to support teaching and learning throughout the year; and formative assessment processes and tools, designed to support instruction.

All of the components of the Smarter Balanced Assessment System are based on the Common Core State Standards (CCSS) which clearly specify college and career-readiness and meaningful grade-level expectations. The system is supported by consortium and state policies and practices designed to support high expectations and increased learning opportunities for students. Teachers are provided with curriculum and instructional materials and given rich professional development and other supports and resources needed to effectively teach the content embodied by the standards.

The intended purposes of the various components are outlined below.

The purposes of the Smarter Balanced summative assessments are to provide valid, reliable, and fair information about:

- Students’ ELA/literacy and mathematics achievement with respect to CCSS, measured by the ELA/literacy and mathematics summative assessments in grades 3 to 8 and high school,
- Whether students prior to grade 11 have demonstrated sufficient academic proficiency in ELA/literacy and mathematics to be on track for achieving college readiness,
- Whether grade 11 students have sufficient academic proficiency in ELA/literacy and mathematics to be ready to take credit-bearing, transferable college courses after completing their high school coursework,
- Students’ annual progress toward college- and career-readiness in ELA/literacy and mathematics, and
- Students’ ELA/literacy and mathematics proficiencies for federal accountability purposes and potentially for state and local accountability systems.

The purposes of the Smarter Balanced interim assessments are to provide valid, reliable, and fair information about:

- Student progress toward mastery of the skills in ELA/literacy and mathematics measured by the summative assessment,
- Student performance at the Claim or cluster of Assessment Targets so teachers and administrators can track student progress throughout the year and adjust instruction accordingly,
- Individual and group (e.g., school, district) performance at the Claim level in ELA/literacy and mathematics to determine whether teaching and learning are on target,
- Teacher-moderated scoring of performance events as a professional development vehicle to enhance teacher capacity to evaluate student work aligned with the standards, and
- Student progress toward the mastery of skills measured in ELA/literacy and mathematics across all students and subgroups.

The purposes of the Smarter Balanced formative assessment resources are to provide measurement tools and resources to:

- Illustrate how teachers and other educators can use assessment data to engage students in monitoring their own learning,
- Help teachers and other educators align instruction, curricula, and assessments,
- Assist teachers and other educators in using the summative and interim assessments to improve instruction at the individual and classroom levels, and
- Offer professional development and resources for how to use assessment information to improve teacher decision-making in the classroom.

Appendix B: CAST Alignment Study Workshop Data Collection Plan

This approach to alignment is conceptualized to verify that CAST uses a reasoned approach to sampling the content within the science standards. HumRRO will not compute or report proportions of potential combinations of standards addressed on the CAST (e.g., DCI x SEP x CCC). This would be inappropriate because it would not be possible to represent the full potential breadth of the standards in a single summative assessment.

In order to capture the specificity of the CAST items, and to verify that CAST represents the intended blueprint, HumRRO proposes to use the final CAST Item Specifications as our alignment guide. These specifications guided the creation of the CAST items and represent the intended CAST measurement construct. Table B1 provides the number of items to be reviewed and rated by panelists, by segment and grade.

Table B1. Number of CAST 2019 Operational Items to be Rated

Grade	Number of items Segment A	Number of items Segment B	Total number of operational items by grade
5	34 (2 blocks)	32 (5 blocks)	66
8	33 (2 blocks)	37 (6 blocks)	70
High School	34 (2 blocks)	19 (3 blocks)	53
Total	101	88	189

HumRRO will rely on panelist experts to make several item-level ratings to accomplish the alignment review. There will be a separate panel for each tested grade span. To make the task manageable, the testing contractor’s item coding metadata will be used to create subsets of items to limit the panelists. Items will be divided into those that primarily represent life sciences, physical sciences, or earth and space sciences. Because an item may address more than one of these three science domains, panelists will be given the opportunity to inform us of this, and HumRRO can then account for it in our study. The creation of item sub-sets limits the reference material needed at one time for each group of items.

Next, HumRRO will create simple unique item IDs that can be easily entered into a spreadsheet. These will likely be simple sequence numbers. Most ratings panelists make will require indicating a performance expectation, content dimensions or domains, and an item cognitive complexity rating. One complete data collection spreadsheet will be created for each grade span.

Each spreadsheet will include multiple columns. Columns will include Sub-Practice Assessment Targets, DCI Assessment Targets, and CCC Assessment Targets, as well as a column for item-level cognitive complexity ratings (such as Depth of Knowledge) and a column to rate the use of Phenomena. Table B2 displays a mock-up of the panelists' coding spreadsheet. Note that each assessment target is given a #1 and a #2 column to accommodate any item that assesses multiple targets. Panelists will also be instructed to include notes if any item does not address one or more assessment targets (the notes column does not appear in the mock-up but will be available electronically). CDE staff will review and approve the data collection instruments prior to use, as a quality control step.

Table B2. Mock-up of Panelists' Coding Spreadsheet

Item Number	Sub-Practice Assessment Target #1	Sub-Practice Assessment Target #2	DCI Assessment Target #1	DCI Assessment Target #2	CCC Assessment Target #1	CCC Assessment Target #2	Item Complexity	Phenomenon
1	(blank)	(blank)	(blank)	(blank)	(blank)	(blank)	(blank)	(blank)
2	(blank)	(blank)	(blank)	(blank)	(blank)	(blank)	(blank)	(blank)
3	(blank)	(blank)	(blank)	(blank)	(blank)	(blank)	(blank)	(blank)
4	(blank)	(blank)	(blank)	(blank)	(blank)	(blank)	(blank)	(blank)
5	(blank)	(blank)	(blank)	(blank)	(blank)	(blank)	(blank)	(blank)
6	(blank)	(blank)	(blank)	(blank)	(blank)	(blank)	(blank)	(blank)
7	(blank)	(blank)	(blank)	(blank)	(blank)	(blank)	(blank)	(blank)
8	(blank)	(blank)	(blank)	(blank)	(blank)	(blank)	(blank)	(blank)
9	(blank)	(blank)	(blank)	(blank)	(blank)	(blank)	(blank)	(blank)
10	(blank)	(blank)	(blank)	(blank)	(blank)	(blank)	(blank)	(blank)

The information panelists will code into the forms will come from a coding guide provided to panelists based on CAST item development guidance. The codes represent the standards each item was developed to measure. There are codes specific to Sub-Practice Assessment Targets, DCI Assessment Targets, and CCC Assessment Targets. A sample of the codes is listed in Table B3. The codes will be arranged by domain (Life, Physical, and Earth and Space Sciences).

Table B3. Sample Standards Coding Information for Panelists

Dimensions	Standards
Sub-Practice	2.1 Ability to Develop Models
Sub-Practice Assessment Target	2.1.1 Ability to determine the components as well as relationships among multiple components, to include or omit, a scientific event, system, or design solution
Sub-Practice Assessment Target	2.1.3 Ability to represent mechanisms, relationships, and connections to illustrate, explain, or predict a scientific event
DCI Assessment Target	PS1.A.4a. Develop a model of matter with microscopic particles as the components.
DCI Assessment Target	PS1.A.4b. Describe bulk matter as being composed of tiny particles of matter that cannot be seen.
DCI Assessment Target	PS1.A.4c. Describe the behavior of many tiny particles to explain observable phenomena involving bulk matter.
DCI Assessment Target	PS1.A.4d. Explain observable phenomena by using a model of bulk matter composed of many tiny particles.
CCC Assessment Targets	Student can: CCC3 Identify that natural objects exist from the very small to the immensely large.

For each item, panelists will indicate on the coding spreadsheet whether the item addresses each of the components of each of the standards listed. For example, for Item #1, panelists would indicate if the item addressed either of the sub-practice assessment targets (2.1.1 or 2.1.3). If so, the panelists would indicate this by placing that code in the second column beside the item number. If the item addressed a sub-practice that was different from 2.1.1 or 2.1.3, they would enter the more generic code (2.1), and provide an explanation of why the item met the more generic definition, but not the specific, in their notes. They would continue this process for DCI and CCC targets. Training would include when to indicate a code versus when not to (e.g., the item directly addresses the standard versus uses language similar to the standard). Items will have multiple codes in the spreadsheet to account for the multidimensionality of the test items.

The next column in the coding spreadsheet will capture cognitive complexity ratings. These ratings will be numeric and panelists will be trained on the selected scale prior to assigning them. HumRRO expects to use Webb's DOK rating system unless the CDE and ETS use a different system.

A fundamental principle in CA NGSS is that students must use the three dimensions to understand specific phenomena (i.e., any observable event that occurs in a natural or a designed system), and that phenomena drive science learning. The final column in the spreadsheet is for capturing panelist ratings for phenomena. Panelists would enter a 1 in this column to indicate that the phenomenon used for the test item provided an appropriate opportunity for the student to demonstrate the knowledge/skills required to provide evidence to support the intended claims and interpretations (as indicated by the content codes). Panelists would enter a 0 if no phenomenon was used for the test item or if the phenomenon used for the test item(s) did not provide an appropriate opportunity.

The workshop will be structured such that whole group training (all panelists in the same room) will occur on the morning of the first day. Each alignment task will be preceded with targeted training that will be conducted by the facilitator within each group. The whole group training will provide an overview of the task, discuss alignment concepts, review rules regarding confidentiality and data security, and orient the panelists to expectations for the remainder of the workshop. This training will include an overview of the cognitive complexity ratings used by the test developer and an overview of administration policies and procedures, including common accommodations and accessibility features. Table B4 provides a tentative agenda for the workshop.

Table B4. Tentative Agenda for CAST Alignment Study Workshop

Day	Time Period	Activity
Day 1	9–10 a.m.	Introduction to the method and overview of alignment concepts (large group training)
Day 1	10:15–11 a.m.	Review of test specifications, blueprint, and overview of item and form development processes (large group training)
Day 1	11 a.m.–4:30 p.m.	Independent ratings of CAST items (small group, content match and complexity)
Day 2	9 a.m.–12 noon	Compare independent panelists’ ratings with test developer-supplied metadata (CA NGSS and complexity coding) and indicate item-level agreement (small group)
Day 2	1–4:00 p.m.	Outlier analyses and item-level discussion; panelists make any revisions to original ratings (small group) Collect final panelist data.
Day 2	4–4:30 p.m.	Workshop evaluation, debrief, and dismissal (small group)

During the workshop, once panelists have completed their independent ratings, those ratings will be saved, and then facilitators will show them the test developer’s metadata for each item for comparison. Panelists will be directed to discuss the items that are mismatched between the test developer and most of the panelists. Panelists will be instructed to change their ratings at this stage if they believe the test developer's metadata more accurately captures the content the items measure. Panelists will then discuss their ratings compared to each other and the testing contractor to attempt to reach consensus on the most accurate alignment information for each item.

Panelists will provide final ratings following or during the consensus discussion. Their original data will be retained to allow determination of the extent to which each panelist revised their original ratings. HumRRO will compute agreement statistics based on the panelists original data (panelists compared to other panelists), and on the final consensus data (panelist consensus data compared to contractor's metadata). Item data from panelists will be considered final after this step.

The report of workshop results will include an assessment of the level of agreement among raters within each panel and overall agreement with the test developer's metadata. Any areas of significant disagreement will be reported to the test developers for consideration and comment (i.e., an electronic file will be made available that directly compares item metadata with panelists' final ratings). Overall alignment results (including Segment C) will be computed after the 2019 CAST administration based on item metadata only.

Appendix C: Impact Case Study Definition of "Full Implementers"

LEAs may use all or a selected subset of Smarter Balanced components of the CAASPP System in well-integrated and highly effective ways. For purposes of this Impact Case Study, however, HumRRO will gain the greatest amount of useful information for the CDE by focusing our case study sample on LEAs that use the Smarter Balanced components and their features extensively. To that end, HumRRO defines "full CAASPP implementers." These LEAs should have demonstrated during the 2017–18 school year at least a "modest threshold" of use of both of the optional Smarter Balanced CAASPP System components (a) IAB assessments, with or without ICAs and hand scoring, and (b) the Instructional Resources of the Digital Library, with or without use of Professional Learning resources and Playlist resources. "Modest threshold" means a sufficient amount of use beyond simply investigating system features and will be defined based on Digital Library login data and interim assessment data provided to HumRRO. Eligible LEAs need not be the heaviest users in the state. HumRRO proposes to establish an empirically based threshold of DL and IAB use after analysis of the distribution of usage in 2017–18 by test, school, and LEA.

Rationale: Given the limited sample of only six (6) LEAs, HumRRO wants to maximize useful information upon which to provide robust findings for the CDE and, where warranted, make actionable recommendations. Including LEAs that are not using both optional components in meaningful ways would result in two shortcomings to the Impact Case Study. First, the CDE is interested in the frequency of IAB use and how the results are used to inform instruction—an LEA electing not to use these would not be able to provide feedback regarding IABs. Second, if an LEA uses the Instructional Resources of the Digital Library, educators in the LEA can provide important feedback on features and content of some of the CAASPP tools most likely to be useful to inform changes to instruction and hence opportunities for student learning.

This page is intentionally blank.

Appendix D: Items for Impact Case Study Eligibility Survey of Prescreened LEA CAASPP Coordinators

1. Which of the following CAASPP training resources, developed by the California Department of Education and its vendors, did staff from your local educational agency (LEA) and/or your schools attend/use/review during the 2017–18 school year? Mark all that apply.
   - CAASPP Institute (in-person attendance or use/review of online resources)
   - Post-Test Workshop – The Results are in…Now What? (in-person attendance or use/review of online resources)
   - Summer Hand Scoring Workshop– also referred to as Interim Assessment (IA) Hand Scoring Workshop (in-person attendance or use/review of online resources)
   - Digital Library and Interim Assessment Clinic (in-person attendance or use/review of online resources)
   - *CAASPP in Action* report series, featuring LEAs sharing their successes, challenges, and lessons learned (use/review of online resource)
   - None of the above

2. How do you typically access CAASPP interim assessment results?
   - IA Reporting System only
   - Student information system (e.g., Aeries, Illuminate Education) or other local database only
   - Multiple ways (student information system, local database, and IA Reporting System)
   - Other, explain what system and why you selected it:

     ____

3. How do you provide access to CAASPP interim assessment results to classroom teachers in your LEA?
   - IA Reporting System only
   - Student information system (e.g., Aeries, Illuminate Education) or other local database only
   - Multiple ways (student information system, local database, and IA Reporting System)
   - Other, explain what system and why you selected it:

     ____

4. To what extent do teachers in your schools use the Instructional Resources in the Smarter Balanced Digital Library (DL)?
    - Never
    - Rarely
    - Sometimes
    - Often
    - Do not know

5. To what extent do teachers in your schools use the Professional Learning Resources in the DL?
    - Never
    - Rarely
    - Sometimes
    - Often
    - Do not know

6. To what extent do teachers in your schools use the Playlist Resources (such as Connections Playlists that link student performance on the IABs to specific resources in the DL)?
    - Never
    - Rarely
    - Sometimes
    - Often
    - Do not know

7. Please choose the response that best describes the participation of schools within your LEA in professional learning communities (PLCs).
    - All (or most) schools have established PLCs.
    - Some schools have established PLCs.
    - Few or no schools have established PLCs.
    - I am not aware of the use of PLCs in schools across my LEA.
    - Other and/or clarifying comments (please describe)

    ____

8. Which of the following phrases best describes the amount of teacher turnover in your schools from the 2017–18 school year to the 2018–19 school year?
    - Little/no turnover
    - Moderate turnover
    - Extensive turnover

9. Which of the following phrases best describes the amount of leadership change in your LEA and its schools from the 2017–18 school year to the 2018–19 school year?
    - Little/no changes of leadership (LEA or school)
    - Moderate changes of leadership (LEA or school)
    - Extensive changes of leadership (LEA or school)

10. Do you think a small number of educators might be interested in collaborating with HumRRO during the 2018–19 study, by sharing LEA- and school-level experiences using CAASPP System components to improve instruction and student learning?
    - Yes
    - No
    - I don’t know

This page is intentionally blank.

Appendix E: Criteria for Selection of Impact Case Study Participants from Eligible LEAs

Empirically based tiers will be defined for the LEA sample, established from the distribution of characteristics of all LEAs in the state. The size of the student population of all LEAs will be rank ordered and then divided into two groups: High (top 50%) and Low (bottom 50%). Similarly, LEAs will be classified based on their student achievement on recent summative assessments in ELA and mathematics, aggregated across grade levels and content areas. LEAs will also be divided into high and low levels of the density of EL students and density of economically disadvantaged (ED) students among the student population. HumRRO will empirically identify cut points for high versus low levels to yield relatively equal numbers of high and low LEA context indicators (demographics) within each size and achievement level grouping. HumRRO will establish minimum levels to ensure all participating LEAs offer some diversity.

After the Eligibility Survey responses are collected and analyzed, HumRRO will work with CDE to identify eligible LEAs within the tiers and classify them with respect to their degree of “full CAASPP System implementation” and other factors (e.g., use of IA Reporting System, availability of climate survey data, use of professional learning communities, etc.). To be considered for participation, an LEA will only be required to meet the definition of “full CAASPP System implementer” for ELA or mathematics.

After an eligible pool of LEAs is determined, HumRRO will select specific LEAs to recruit across these tiers. To the extent possible, purposive sampling will be used to select LEAs that represent the distribution depicted in Table E1, whereby LEAs are ordered randomly within each sampling cell.

Table E1. Theoretical Distribution of Target Characteristics of Impact Case Study LEAs in 2018–19 School Year

Case Study LEA #	Size of Student Population	Percentage of Disadvantaged Students*	Aggregated Student Achievement:
#1	Lowest 50%	Low	Lowest 50%
#2	Lowest 50%	High	Lowest 50%
#3	Lowest 50%	High	Highest 50%
#4	Highest 50%	Low	Highest 50%
#5	Highest 50%	High	Lowest 50%
#6	Highest 50%	High	Highest 50%

*An index combining percent EL and percent ED students will be generated to identify disadvantage level at each school compared to the norm across California.

The characteristics in Table E1 describe our ideal sample. In addition to the characteristics identified within the table, HumRRO will strive to include:

- Three LEAs from the northern half of the state and three from the southern half
- At minimum three LEAs implementing the full CAASPP system for ELA, and three for mathematics. Ideally, at least some LEAs will be full implementers for both ELA and mathematics.
- One direct-funded charter school LEA. Note that this LEA might include only a single school, rather than the three schools HumRRO plans to recruit from each of the other five LEAs (i.e., one high school, one middle school, and one elementary school).

The recruitment of LEAs will include a description of the desired characteristics of one elementary, one middle, and one high school from each LEA. HumRRO may find that an LEA that appears to be a good fit may not have the requisite types of schools. HumRRO anticipates needing to contact LEA CAASPP coordinators to confirm the identified schools meet the definition of “full implementer.”

HumRRO expects this recruitment to be an iterative process. HumRRO will begin with identification of six LEAs, but based on LEA willingness or reluctance to participate and the fit of schools within the LEA, will identify replacement LEAs until a sufficient sample is recruited.

Appendix F: POC Polling Information and Example Questions for Impact Case Study

This case study approach demands that HumRRO establish a healthy, active working relationship with our LEAs and schools and their respective Points of Contact (POCs). In addition to site visits early in the school year and focus groups near the end of the year, HumRRO will communicate with POCs regularly over the course of each school year. HumRRO will ask one to three monthly questions to better understand use of the various CAASPP components. Depending on the information sought, HumRRO will use phone calls or e-mails to collect information. In any given month, HumRRO may send different questions to elementary POCs, middle school POCs, high school POCs, and LEA POCs, which POCs will distribute to LEA leaders, school leaders, and/or teachers; or HumRRO may issue the same questions to multiple groups.

The polls to each LEA will begin after the initial site visit and run through the end of the school year. Questions will be e-mailed regularly once each month, allowing about four weeks for responses to be collected and returned to HumRRO by POCs. Questions may include but will not be limited to:

- Elementary School POC: Describe recent use of formative tools in a math classroom in your school. Include the grade level and characteristics of the class, the teacher's goals in using this tool, and the teacher's evaluation of the effectiveness of the tools. Indicate how these tools relate to the CAASPP System (e.g., from the DL).

- Middle School POC: Describe a recent Professional Learning Community (PLC) meeting that focused on some aspect of the CAASPP System. What was discussed? What benefits and/or challenges were identified? What was accomplished via this discussion?

- High School POC: Describe the use of summative assessment scores to track college and career readiness for high school students in reading and mathematics. Are formative assessment tools used to help those who are not on track? If so, which ones and how are they used?

- All School POCs: What challenges do educators at your school encounter when using interim assessments and formative assessment tools for English learners? Have teachers identified successful strategies for using these tools with this population? If so, what are they? You may want to poll educators at your school to develop your answer. (Note: Similar question for supports for students with disabilities.)

- LEA POC: Describe how your LEA has monitored interim assessment use and results over the past month, if at all. Have the results of the monitoring caused you to take any action? What benefits and/or concerns do you have?

This page is intentionally blank.

Appendix G: Impact Case Study Data Analysis Plan

Analysis of Data from Artifacts, Interviews, and Focus Groups

Qualitative data gathered from collected artifacts and from interviews and focus groups will be reviewed immediately after initial interactions with LEA participants. As additional data are collected, continual analysis will inform ongoing refinements to data collection instruments (e.g., topic guides) to progressively narrow the focus on key aspects of the participants' perspectives and experiences.

Initial analysis activities will involve a cycle of iterative steps: gathering data, examining data, comparing prior data to new data, writing up field notes before conducting more interviews and focus groups, and making plans to gather new data through revisions to the protocols. In keeping with an inductive reasoning approach to the case studies, HumRRO will avoid making premature decisions based on early analysis and interpretation of data. Consciously "pausing" will allow the team to continually reflect on procedures and integrate necessary adjustments, thus ensuring data are collected that will provide rich and meaningful information regarding the research questions.

To conduct this iterative process, researchers will reflect upon each incremental body of data they gather by considering various questions: "Why do participants act as they do?" "What else do I want to know about that participant's process?" "What new ideas have emerged in this round of data collection?" "Is this a new concept or is it the same as a previous one?" and so forth. This iterative process will lead to the collection of carefully targeted new data and the elimination of unproductive questions. Periodic reflection on data collection techniques will allow for refinement as needed to address the research questions and filter out irrelevant data.

After the fall interviews and focus groups and end-of-year focus groups are completed, a complete quantitative analysis on all interview and focus group data collected will be conducted. Because there are no predefined variables, the qualitative data will be analyzed by systematically and progressively narrowing the patterns and themes that emerge. This will entail a multistage process of organizing, categorizing, synthesizing, and analyzing the data and documenting findings to produce amalgamated themes. Additionally, variations to these themes will be represented to reflect the range of responses.

Researchers will cycle through the stages multiple times to narrow and make sense of the data. Three key steps include (a) becoming familiar with the data and identifying potential themes; (b) reviewing the data to understand each LEA context and how the participants interact with the CAASPP system; and (c) categorizing, coding, and grouping the data into themes to address research questions. HumRRO anticipates the CAASPP System's Theory of Action and the study's research questions will provide a starting point for organizing the coding categories, which will be complemented and challenged by inductively developed codes.

The Study Director will assign subsets of the interview and focus group data to analysts to categorically mark or reference units of text (e.g., words, sentences, paragraphs, and quotations) with codes or labels to indicate patterns and meaning. As HumRRO analyzes and codes the data, researchers will reduce the data to a manageable form (e.g., may use Excel features such as pivot tables to facilitate review of all responses by question, as well as filtering responses by LEA and/or grade span). Two analysts will independently scan each set of responses to ensure interrater consistency, and to consider the big picture and compile a list of themes that emerge. During this step, analysts will search for patterns (e.g., actions that participants took, perceptions about the effectiveness of a process or an action) that align and form a pattern or theme. After determining initial themes, our analysts will compile and determine the commonality across LEAs. They will consider the need to combine themes that are worded differently based on LEA-specific language. For example, LEAs may have different names for interdepartmental meetings/planning time, though these activities are highly similar. HumRRO will determine if there are specific themes that are common across most or all LEAs, and/or if there are particular themes for LEAs with common characteristics (for example, if small LEAs face challenges that are not seen in large LEAs). Additionally, important variations to these themes will be represented to reflect the range of responses.

The analysts will meet periodically with the Study Director to discuss the emerging patterns and themes, and to reconcile any major differences in what the analysts glean from the data. The Study Director and analysts also will meet at the conclusion of the coding to review results and finalize descriptions and findings to include in the final report.

Analysis of POC Polling Data

The polling data from LEA and School POCs will be analyzed in three distinct ways at different times to answer different questions. This information will be merged with data obtained through site visits for addressing research questions.

1. As each set of monthly responses is received: Written and verbal responses will be coded by LEA identification, respondent group (i.e., elementary school, middle school, high school, LEA), disposition (e.g., positive, negative, neutral), and theme. Themes will be empirically derived from a review of the responses. As the thematic coding is underway, any unclear responses will be shared with HumRRO's POC for possible follow-up. Monthly analytic results will be provided to the Study Director, with information about trends and issues to date for use when developing future polling questions.

2. In preparation for end-of-school-year focus group protocol development: HumRRO will analyze each LEA's cumulative POC polling data, to include responses over time for all four respondent groups in the LEA (i.e., LEA, elementary school, middle school, high school). In addition to a summary of the themes identified each month, patterns such as trends in disposition over time will be analyzed through simple statistics. The results of this analysis will be used to tailor LEA-specific end-of-year focus group questions to relevant issues.

3. In preparation for annual reports: Previously analyzed themes from POC polling data will be analyzed by respondent group (i.e., LEA, elementary, middle, high school), across LEAs. Overall trends as well as trends disaggregated by LEA size (based on student enrollment) and aggregated student achievement at time of recruitment will be evaluated. Differences between stratifications (e.g., high achieving LEAs value Digital Library resources more than LEAs with medium achievement) will be highlighted. Differences in these qualitative data for schools and LEAs implementing and using CAASPP in different ways will be evaluated.

Analysis of Interim Assessment Data

Data regarding interim assessment use and scores will be obtained from participating LEAs during each school year. Descriptive statistics will characterize the extent of use, type of use (e.g., content area, grade level, standardized versus non-standardized administration, IABs versus ICAs), timing of use, and level and trends in student scores. Analyses will include a static summary of all use to date, as well as patterns of use during the school year. Results will be summarized by LEA and overall to facilitate the development of POC polling questions and spring focus groups.

Analysis of Use of the Digital Library

Information on use of the Digital Library will be compiled from POCs throughout the school year through POC polling, and from login information from the CDE. Descriptive statistics will summarize the frequency of logins and the use of different types of resources. Qualitative analyses will identify common reasons for seeking available resources, and whether and how the reasons change over time within an LEA or across the full group.

This page is intentionally blank.

Appendix H:
Detailed Descriptions of Figures with Images

Figure 2.1 PE Distribution for Segment A of the CAST Grade 5 Assessment. (p. 2-18)

- The figure is a matrix illustrating the multiple dimensions of the CA NGSS and the proposed distribution of Segment A items for grade 5 by DCI, SEP, and CCC.
- The data table below is represented graphically in the figure to indicate the number of items per DCI strand. The total number of items is 32–34.

DCI Strand	PS1	PS2	PS3	PS4	LS1	LS2	LS3	LS4	ESS 1	ESS 2	ESS 3	ETS
Items per Strand	1–3	1–4	1–4	1–2	1–2	1–2	1–2	1–4	1–2	1–5	1–3	2–4

- The data table below is represented graphically in the figure to indicate the number of items per Domain. The total number of items is 32–34.

Domain	Physical Sciences (17 PEs)	Life Sciences (12 PEs)	Earth and Space Sciences (13 PEs)	ETS (3 PEs)
Items per Domain	8–10	8–10	8–10	2–4

- The data table below is represented graphically in the figure to indicate the number of items per SEP. The total number of items is 32–34.

Science and Engineering Practices	Items per SEP
SEP 1 and 1E	1–4
SEP 2	1–7
SEP 3	1–7
SEP 4	2–4
SEP 5	1–2
SEP 6 and 6E	2–8
SEP 7	1–8
SEP 8	1–3

- In the figure, an X indicates the intersections of SEPs, DCIs, and CCCs articulated in the PEs. These intersections represent opportunities to develop items that can be used to assemble Segment A. While each individual item reflects the intersection of an SEP, DCI, and CCC, the figure indicates the proposed distribution of Segment A items by DCI, SEP, and CCC.

- An X indicates that there is at least one PE at the given intersection of the three dimensions that can be sampled on a test form for Segment A.
- SEPs 1 and 6 have separate components for science and engineering (SEP 1E and SEP 6E). All other SEPs incorporate the same components for both science and engineering.
- The figure includes an X for each of the following intersections:
  - SEP 1 (3 Xs): PS2 and CCC2; PS3 and CCC5; ETS
  - SEP 1E (1 X): PS2
  - SEP 2 (8 Xs): PS1 and CCC3; PS3 and CCC5; PS4 and CCC1; PS4 and CCC2; LS1 and CCC1; LS1 and CCC4; LS2 and CCC4; ESS2 and CCC4
  - SEP3 (7 Xs): PS1 and CCC2; PS1 and CCC3; PS2 and CCC1; PS2 and CCC2; PS3 and CCC5; ESS2 and CCC2; ETS
  - SEP4 (4 Xs): LS3 and CCC1; LS4 and CCC3; ESS1 and CCC1; ESS2 and CCC1
  - SEP5 (2 Xs): PS1 and CCC3; ESS2 and CCC3
  - SEP6 (5 Xs): PS3 and CCC5; LS3 and CCC2; LS4 and CCC3; ESS1 and CCC1; ETS
  - SEP6E (3 Xs): PS3 and CCC5; PS4 and CCC1; ESS3 and CCC2
  - SEP7 (8 Xs): PS2 and CCC2; LS1 and CCC4; LS1 and CCC5; LS2 and CCC2; LS4 and CCC2; LS4 and CCC4; ESS1 and CCC3; ESS3 and CCC2
  - SEP8 (3 Xs): ESS2 and CCC1; ESS3 and CCC2; ESS3 and CCC4

Figure 2.2 Example of a Wright Map comparing examinee ability and item difficulty distributions. (p. 2–19)

- This figure provides a sample Wright Map that has persons on the left side of the map and items on right.
- The figure depicts the distribution of students’ ability levels, as indicated by their score estimates (thetas) on an assessment, on the left side of the centerline of the figure. Bars of differing heights indicate the number of students at each score level).

- Using the same scale (theta), the right side of the centerline of the figure plots the distribution of items by difficulty (again using bars to depict the number of items with difficulties at each level).
- The figure also includes horizontal bars to indicate where the cut points are for classifying students into performance categories (Level 1 through 4).
- The data table below is represented graphically in the figure.

Performance Level (High=4)	Ability Level	Number of Persons	Number of Items	Item Difficulty Parameter
4	2.2–3.0	0	0	2.2–3.0
4	2.1	1	0	2.1
4	2.0	1	0	2.0
4	1.9	1	1	1.9
4	1.8	2	0	1.8
4	1.7	3	0	1.7
4	1.6	3	1	1.6
4	1.5	7	0	1.5
4	1.4	6	0	1.4
3	1.3	7	2	1.3
3	1.2	16	3	1.2
3	1.1	13	2	1.1
3	1.0	16	2	1.0
3	0.9	14	4	0.9
3	0.8	11	6	0.8
3	0.7	21	5	0.7
3	0.6	9	5	0.6
3	0.5	10	6	0.5
3	0.4	17	8	0.4
3	0.3	15	9	0.3
2	0.2	7	5	0.2
2	0.1	9	9	0.1
2	0.0	7	7	0.0

Performance Level (High=4)	Ability Level	Number of Persons	Number of Items	Item Difficulty Parameter
2	-0.1	12	7	-0.1
2	-0.2	3	11	-0.2
2	-0.3	2	7	-0.3
2	-0.4	1	5	-0.4
2	-0.5	0	5	-0.5
2	-0.6	1	10	-0.6
1	-0.7	0	3	-0.7
1	-0.8	1	4	-0.8
1	-0.9	0	2	-0.9
1	-1.0	0	3	-1.0
1	-1.1	0	2	-1.1
1	-1.2	0	3	-1.2
1	-1.3–1.7	0	0	-1.3–1.7
1	-1.8	0	1	-1.8
1	-1.9	0	1	-1.9
1	-2.0–2.1	0	0	-2.0–2.1
1	-2.2	0	1	-2.2
1	-2.3	0	0	-2.3
1	-2.4	0	1	-2.4
1	-2.5–3.0	0	0	-2.5–3.0

Figure 2.3 Sample graph illustrating number of items (horizontal axis) addressing each domain (vertical axis) by form. (p. 2–20)

- The data table below is represented graphically in the figure.

Domain	Number of Items Form 0	Number of Items Form 1	Number of Items Form 2	Number of Items Form 3
Life Sciences	7	2	2	2
Physical Sciences	9	3	2	3
Earth & Space Sciences	5	2	2	2