California Department of Education
Assessment Development & Administration Division

California Assessment of Student Performance and Progress

California Science Test Field Test Technical Report

2017–18 Administration

Final Submitted June 20, 2019

Educational Testing Service

Contract No. CN150012

Table of Contents

List of Tables

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

Term	Definition
2PL-IRT	two-parameter item response theory model
3D	three dimensional
AD	Assessment Development
AERA	American Educational Research Association
AI	artificial intelligence
AIR	American Institutes for Research
AIS	average item score
APA	American Psychological Association
ASL	American Sign Language
CA NGSS	California Next Generation Science Standards
CAA	California Alternate Assessment
CAASPP	California Assessment of Student Performance and Progress
CALPADS	California Longitudinal Pupil Achievement Data System
CalTAC	California Technical Assistance Center
CAST	California Science Test
CCC	crosscutting concepts
CCR	California Code of Regulations
CCSS	Common Core State Standards
CDE	California Department of Education
CR	constructed response
CRSC	CR Scoring Systems and Capabilities
DCIs	disciplinary core ideas
DIF	differential item functioning
EC	Education Code
ECD	Evidence-Centered Design
ECV	explained common variance
ELA	English language arts/literacy
eSKM	Enterprise Score Key Management System
ETS	Educational Testing Service
FIA	final item analysis
GPCM	generalized partial credit model
IMS	Instructional Management Systems
KSA	knowledge, skills, and abilities
LEA	local educational agency
MC	multiple choice
MH DIF	Mantel-Haenszel Differential Item Functioning
MIRT	multidimensional item response theory
MSE	mean square error
MST	multistage adaptive test

Term	Definition
NCME	National Council on Measurement in Education
ONE	Online Network for Evaluation
OTI	Office of Testing Integrity
PAR	Psychometric Analysis and Research
PE	performance expectation
PIA	preliminary item analysis
PT	performance task
QA	quality assurance
QC	quality control
QTI	Question and Test Interoperability
QWK	quadratic weighted kappa
SBE	State Board of Education
SD	standard deviation
SEPs	science and engineering practices
SFTP	secure file transfer protocol
SMD	standardized mean difference
STAIRS	Security and Test Administration Incident Reporting System
SVM	Support Vector Machine
TDS	test delivery system
TEI	technology-enhanced items
TIF	test information function
TOMS	Test Operations Management System
UAT	User Acceptance Testing
USC	United States Code

Chapter 1: Introduction

1.1. Background

In October 2013, Assembly Bill 484 established the California Assessment of Student Performance and Progress (CAASPP) as the new student assessment system that replaced the Standardized Testing and Reporting program. The primary purpose of the CAASPP System of assessments is to assist teachers, administrators, and students and their parents or guardians by promoting high-quality teaching and learning through the use of a variety of item types and assessment approaches. These tests provide the foundation for the state's school accountability system.

California adopted the California Next Generation Science Standards (CA NGSS) in September 2013. The California Science Test (CAST) is an online assessment aligned with the CA NGSS. It was administered as a pilot for the first time during the 2016–17 CAASPP administration, followed by a field test administration for the 2017–18 CAASPP administration. The assessment is for students in grades five, eight, and high school. For the CAST field test, all students in grade twelve were tested, and students in grades ten and eleven had the opportunity to test, as discussed in subsection *1.4 Intended Population*.

During the 2017–18 administration, the overall CAASPP System had the following components:

- Smarter Balanced assessments and tools:
  - Summative Assessments—Online assessments for English language arts/literacy (ELA) and mathematics in grades three through eight and grade eleven
  - Interim Assessments—Optional resources developed for grades three through eight and grade eleven designed to inform and promote teaching and learning by providing information that can be used to monitor student progress toward mastery of the Common Core State Standards (CCSS) and that may be administered to students at any grade level
  - Digital Library—Tools, lesson plans, and practices designed to help teachers utilize formative assessment processes for improved teaching and learning in all grades
- California Alternate Assessments (CAAs) for ELA and mathematics in grades three through eight and grade eleven for students with significant cognitive disabilities
- Science assessments in grades five, eight, and high school (grades ten, eleven, or twelve; these are the CAST and the CAA for Science)
- A primary language assessment, the Standards-based Tests in Spanish for Reading/Language Arts, in grades two through eleven (optional for eligible Spanish-speaking English learners)
- A new primary language assessment, the California Spanish Assessment, delivered in field test forms at selected local educational agencies (LEAs), to students in grades three through eight and high school who are Spanish-speaking English learners or students seeking a measure that recognizes their Spanish reading, writing, and listening skills

More background information about the CAASPP System can be found on the CAASPP Description – *CalEdFacts* web page at http://www.cde.ca.gov/ta/tg/ai/cefcaaspp.asp.

1.2. Purpose of the Field Test

The purpose of the CAST field test was to mirror the upcoming operational test as closely as possible, so as to

- provide information on the performance of newly developed CA NGSS–aligned items and item types—in particular, the technology-enhanced items (TEIs) that involve the use of dynamic stimuli and other types of new media (e.g., animations of scientific phenomena, virtual engineering challenges, simulated experiments);
- provide information on the functionality of items with regard to science content rendered by the test delivery system (TDS), with special attention paid to the custom interaction items; and
- provide data for research studies to inform future test design and score reporting decisions.

The CAST field test was intended to assess item performance and not student performance. Therefore, forms are not equated and only preliminary indicators are reported for the CAST field test (refer to *Chapter 7: Reporting*).

1.3. Field Test Content

The test administered at each grade or grade span comprised three segments, A, B, and C, with the content of each assigned randomly to students without regard to their level of performance. Both discrete items and performance tasks (PTs) were included in the tests.

The test delivery system at each grade or grade span randomly assigned students any two of four different item blocks in Segment A, with each containing 16–20 discrete items. Each test also contained two of eight different PTs in Segment B, with each task presenting four to six items. Finally, either one of eight different Segment C blocks was selected (each with 13–14 discrete items), or a third PT that was different from the first two PTs was selected.

The numbers of items reported for Segment B are spread out over eight performance tasks in each grade or grade span.

The PTs were designed to provide students with an opportunity to demonstrate their ability to apply knowledge and higher-order thinking skills to explore and analyze a complex, real-world scenario. The discrete items included traditional multiple-choice items, constructed-response (CR) items and innovative TEIs (refer to subsection *3.3 Item Development*).

A fixed braille form was available for students with visual impairments (refer to subsection *2.3 Test Administration*). The braille form was composed of two Segment A blocks totaling 32–37 items, two Segment B performance tasks of 10–12 items, and one Segment C block of 13 items.

Table 1.1 lists the total number of unique items per segment for each form. On the grade five form, 13 discrete items were repeated among several Segment C blocks (one item repeated three times). On the high school form, one discrete item was repeated between two Segment C blocks.

Table 1.1 Number of Unique Items Assessed on the CAST Field Test

Segment	Grade 5	Grade 8	High School
A (four sections)	75	64	72
B (eight PTs)	44	43	46
C (eight sections)	90	104	106

1.4. Intended Population

The CAST field test was a census test administered to approximately 1.4 million students in the general population. The intended population was all students in grades five, eight, and twelve, as well as high school students in grades ten and eleven who were assigned by their LEA (refer to subsection *4.1 Student Participation Requirement* for more details about the high school grade assignments).

Students eligible for alternate assessments took the CAA for Science in grades five, eight, twelve, as well as high school students in grades ten and eleven who were designated by their LEA. Analyses of the results of the CAA for Science are reported separately.

1.5. Testing Window and Times

The CAST field test was administered during a testing window selected by the LEA, with the first possible date of administration being April 2, 2018, and the last possible date being July 16, 2018 (*California Code of Regulations,* Title 5*,* Education, Division 1, Chapter 2, Subchapter 3.75, Article 2*,* Section 855[a][2]).

Like other CAASPP assessments, the CAST field test was untimed for students. A student could take the CAST field test within the LEA's testing window over as many days as required to meet a student's needs (5 *CCR,* Section 855[a][3]). The average time it took a student to complete the test was roughly two hours.

1.6. Preparation for LEAs

To ensure the 2017–18 test administration was a successful experience for CAST test administrators and students, Educational Testing Service (ETS) provided onsite test administration workshops in various locations throughout California in January and February 2018 and produced webcasts and videos with detailed information on CAASPP test administration procedures. In addition, ETS provided a number of test administration resources to schools and LEAs. These resources included detailed information on topics such as technology readiness, test administration, test security, accommodations, TDS, and other general testing rules.

1.7. Groups and Organizations Involved with the CAST

1.7.1. State Board of Education (SBE)

The SBE is the state agency that establishes educational policy for kindergarten through grade twelve in the areas of standards, instructional materials, assessment, and accountability. The SBE adopts textbooks for kindergarten through grade eight, adopts regulations to implement legislation, and has the authority to grant waivers of the *Education Code*.

In addition to adopting the rules and regulations for itself, its appointees and California's public schools, the SBE also is the state educational agency responsible for overseeing

California's compliance with the Every Student Succeeds Act and the state's Public School Accountability Act, which measures the academic performance and progress of schools on a variety of academic metrics (CDE, 2017).

1.7.2. California Department of Education (CDE)

The CDE oversees California's public school system, which is responsible for the education of more than 6,200,000 children and young adults in more than 10,450 schools.[1] California aims to provide a world-class education for all students, from early childhood to adulthood. The CDE serves the state by innovating and collaborating with educators, school staff, parents/guardians, and community partners which together, as a team, prepares students to live, work, and thrive in a highly connected world.

Within the CDE, it is the Performance, Planning, & Technology Branch that oversees programs promoting innovation and improving student achievement. Programs include oversight of statewide assessments and the collection and reporting of educational data (CDE, 2018a). Within the Performance, Planning & Technology Branch, the Assessment Development & Administration Division manages the development and administration for all statewide assessments.

1.7.3. California Educators

A variety of California educators, including teachers and school administrators, who were selected based on their qualifications, experiences, demographics, and geographic locations in regard to population types, were invited to participate in the entire CAST assessment development process. These California educators participated in tasks that included defining the purpose and scope of the assessment, assessment design, item development, and scoring the constructed-response items.

1.7.4. Contractors

1.7.4.1 Educational Testing Service (ETS)

The CDE and the SBE contract with ETS to develop, administer, and report the CAST, although for the 2017–18 administration, only preliminary indicators were reported. As the prime contractor, ETS has the overall responsibility of working with the CDE to implement and maintain an effective assessment system and to coordinate the work of ETS with its subcontractors. Activities directly conducted by ETS include but are not limited to the following:

- Providing management of the program activities
- Supporting and training counties, LEAs, and direct funded charter schools
- Providing tiered help desk support to LEAs
- Hosting and maintaining a website with resources for LEA CAASPP coordinators
- Developing, hosting, and providing support for the Test Operations Management System (TOMS)
- Developing all CAST test items
- Scoring CR items

[1] Retrieved from the CDE Fingertip Facts on Education in California – *CalEdFacts* web page at https://bit.ly/31gJtRz

- Constructing, producing, and controlling the quality of CAASPP test forms and related test materials
- Processing student test assignments
- Producing and distributing student score reports
- Completing all psychometric procedures
- Developing a summary score reporting website that can be viewed by the public

1.7.4.2 American Institutes for Research (AIR)

ETS also monitors and manages the work of AIR, ETS' subcontractor for the CAASPP System of online assessments. Activities AIR conducts include

- Providing the AIR proprietary TDS, including the Student Testing Interface, Test Administrator Interface, secure browser, and training tests;
- Hosting and providing support for its TDS and the Online Reporting System, a component of the overall CAASPP Assessment Delivery System;
- Scoring machine-scorable items; and
- Providing Level 3 technology help desk support to LEAs.

1.8. Systems Overview and Functionality

1.8.1. Test Operations Management System (TOMS)

TOMS is the password-protected, web-based system that LEAs use to manage all aspects of CAASPP testing. TOMS serves various functions, which, for the CAST pilot, included but were not limited to the following:

- Managing test administration windows
- Assigning and managing CAST online user roles
- Managing student test assignments and accessibility resources
- Providing a platform for authorized user access to secure materials such as user information and access to the *Security and Test Administration Reporting System* form and the Appeals module

TOMS receives student enrollment data and LEA/school hierarchy data from the California Longitudinal Pupil Achievement Data System (CALPADS) via a daily feed. CALPADS is "a longitudinal data system used to maintain individual-level data including student demographics, course data, discipline, assessments, staff assignments, and other data for state and federal reporting."[2] LEA staff involved in the administration of the CAST assessments, such as LEA coordinators, test site coordinators, test administrators, and test examiners are assigned varying levels of access to TOMS. For example, only an LEA coordinator has permission to set up the LEA's test administration window; a test administrator cannot download student reports. A description of user roles is more extensively explained in the *2017–18 Online Test Administration Manual* (CDE, 2018b).

[2] From the CDE California Longitudinal Pupil Achievement Data System (CALPADS) web page at http://www.cde.ca.gov/ds/sp/cl/.

1.8.2. Test Delivery System (TDS)

TDS is the means by which the statewide online assessments are delivered to students. Components of TDS include

- Test Administrator Interface, the web browser–based application that allows test administrators to activate student tests and monitor student testing;
- Student Testing Interface, on which students take the test using the secure browser; and
- Secure browser, the online application through which the Student Testing Interface may be accessed. The secure browser prevents students from accessing other applications during testing.

1.8.3. Online Reporting System (ORS)

LEAs use the ORS to view participation results from the CAASPP assessments. The primary purpose of the ORS is for LEAs to access completion data to determine which students need to complete testing or start testing.

1.8.4. Training Tests

The publicly available training tests are provided to prepare students for the summative assessment. These tests, available for grades five and eight and high school, simulate the experience of the CAST online assessments. The training tests align with performance expectations, but do not produce scores. An accompanying scoring guide is available that describes related scoring considerations (CDE, 2019). Students may access them using a web browser.

The purposes of the training test are to

- allow students and administrators to quickly become familiar with the user interface and components of TDS and the process of starting and completing a testing session, and
- introduce students and administrators to new grade-specific items similar to those on the operational test, which included discrete items and performance tasks.

1.8.5. Constructed Response (CR) Scoring Systems for Educational Testing Service (ETS)

CR items from the TDS are routed to ETS' CR scoring systems. CR items are scored by certified raters. More information regarding scoring of CR items is available in *Chapter 5: Scoring*.

For the CAST field test, targeted efforts were made to hire qualified raters from existing CAASPP rater pools and California science teachers. The hired human raters were provided in-depth training and were certified before starting the scoring process. Human raters were organized under a scoring leader and were provided CAST scoring materials such as anchor sets, scoring rubrics, validity samples, qualifying sets, and condition codes for unscorable responses within the interface. The quality control processes for CR scoring are explained further in *Chapter 8: Quality Control*.

The CR items can also be rated by artificial intelligence (AI) scoring engines (e.g., the *c-rater*™ system). The use of such engines often requires models be built with reliable human rating data. AI scoring was not used to score responses from the field test. Instead, a careful data collection design was used to provide data to build the AI scoring engine for future use. The details of the CR sampling plan that supported the AI model building is provided in subsection *5.1.1 Sampling Process*.

The *c-rater* engine is ETS' system for the automatic, analytic content scoring of short free-text responses that range from a phrase to several sentences in length. The technology is designed to score items that elicit specific information: a correct answer. It works by identifying the content of a response, regardless of the grammatical or stylistic form chosen by the author.

ETS' process requires test designers to define the required content but does not ask them to predict every aspect of the form of student language. The *c-rater* engine can filter out potential nonscorable responses (e.g., responses in a language other than English, nonattempt responses such as "I don't know," etc.). This is applied both during the artificial intelligence (AI)-scoring model building step to ensure AI-scoring models are built on reliable data, and also when the AI-scoring model is deployed to ensure that such responses are filtered and scored correctly. Any response that is entirely non-English that the *c-rater* engine detects will be given a specific advisory designation and handled following the policy established with the CDE (e.g., give these responses a condition code and send them to a human rater, give them a condition code and do not score, etc.). For the 2017–18 CAST administration, ETS built models for 57 field test items using the field test data.

1.9. Limitations of the Assessment

Because of the innovative item types being used to assess these new standards, only a limited number of accessibility features were available for CAST items. ETS continues to conduct research and collaborate with experts to inform further refinements for the available resources as needed.

Another unique challenge of the CAST field test was the alignment of the tests with the CA NGSS, because California recently adopted the CA NGSS; these standards are distinctly different from the previous California science standards. Additionally, it was challenging to align the field tests to the assessment model. Because the purpose of the field tests was to evaluate the items rather than students' knowledge, skills, and abilities, the field tests were not a full representation of the assessment model of the CA NGSS. Refer to subsection *3.3 Item Development* for the model.

Results for the CAST were reported using preliminary indicators, which are descriptive statements with corresponding threshold scores. Preliminary indicators are general, rather than precise, indications of student content knowledge. Their purpose is to help LEAs during the transition period. Therefore, caution should be used when interpreting the preliminary indicator results.

1.10. Overview of the Technical Report

This technical report addresses the characteristics of the CAST administered in spring 2018 and contains nine additional chapters as follows:

- Chapter 2 presents an overview of processes involved in the CAST field test, including descriptions of item development, test administration, and psychometric analyses.
- Chapter 3 discusses the detailed procedures of item development for the CAST field test.

- Chapter 4 describes the details of administering the CAST field test, as well as the procedures followed by ETS to ensure test security.
- Chapter 5 summarizes the scoring approaches and types of scores that are reported for the CAST field test.
- Chapter 6 summarizes the statistical procedures and results for 2017–18, including classical item analyses, test completion rates and analyses, and differential item functioning analyses.
- Chapter 7 summarizes the types of scores and score reports for the CAST field test.
- Chapter 8 discusses the various procedures used to ensure the quality of the CAST field test.
- Chapter 9 describes the development and administration of the survey questionnaires for test administrators and students and the results of analyses of their responses.
- Chapter 10 discusses the various procedures used to gather information to improve the CAST as well as strategies to implement possible improvements.

References

California Code of Regulations, Title 5*,* Education, Division 1, Chapter 2, Subchapter 3.75, Article 2, § 855. Retrieved from https://bit.ly/3mCk1OU

California Department of Education. (2017, October). *State Board of Education responsibilities.* Retrieved from https://bit.ly/3bv2OQT

California Department of Education. (2018b). *CAASPP online test administration manual, 2017–18 administration.* Sacramento, CA: California Department of Education. Retrieved from https://bit.ly/3GBmwZH

California Department of Education. (2018a, August). *Organization.* Retrieved from https://bit.ly/3BHKUFc

California Department of Education. (2019). *California Science Test training items scoring guide, grade 5.* Sacramento, CA: California Department of Education. Retrieved from https://bit.ly/3mB2Ahj

Chapter 2: An Overview of the Field Test Processes

This chapter presents an overview of processes Educational Testing Service (ETS) implemented to develop items for use in the California Science Test (CAST) field test, including a description of the item types developed, item development specifications, form assembly, field test administration, item scoring, and psychometric analyses and reporting. These processes include those that are entirely internal to ETS and those that are undertaken in coordination with the California Department of Education (CDE), the American Institutes for Research (AIR), or both the CDE and AIR.

2.1. Item Development

CAST item development incorporates innovations and best practices from national science assessments. For the field test, items with featured simulations were developed that integrated the dimensions of the performance expectations (PEs) while maintaining appropriateness for the test-taking audience. California science teachers assisted in creating these items, and California teacher committees were instrumental in determining both the proper integration of the PE dimensions as well as grade-level appropriateness.

2.1.1. Design Guidelines

ETS content specialists referred to design patterns and task templates as part of the incipient Evidence-Centered Design (ECD) documentation created by ETS researchers and based on current educational research to properly frame the construct measured in each item (Mislevy, Almond, & Lukas, 2003). As such, all items developed and used in the 2017–18 CAST field test administration are appropriate for the grade level and aligned with the California Next Generation Science Standards (CA NGSS).

2.1.1.1 Design Patterns

The design patterns were developed to define and further unpack each of the eight science-focused science and engineering practices (SEPs) and to identify characteristics of the practice. The engineering focus in two SEPs was defined and unpacked in design patterns apart from the science focus of the same SEPs. The SEP was used as an entry point for item development, both because it represents a fundamental difference between previous science standards and the CA NGSS and because the practice is less familiar to item developers.

During the development of the design patterns, it was determined that each SEP could be further unpacked into several subpractices and that each subpractice could include a set of associated focal knowledge, skills, and abilities.

2.1.1.2 Task Templates

Each task template was developed to focus on the subpractice level and included task features to create items. However, during the field-test development cycle, it became clear that it was necessary to identify the breakdown across the disciplinary core ideas and to integrate that with the task templates at a PE level. This level of detail is now being developed by ETS research and assessment development experts and is included in the Item Specifications.

2.1.2. Content Guidelines

Throughout the item writing process, ETS developers adhered to ETS' foundational guidelines for quality item writing. These guidelines formed the basis for training of item writers and the rigorous review process that is implemented for every item. Additionally, task models and the CA NGSS PEs were used to guide the writing of items for the field test. Refer to subsection *3.3 Item Development* for the guidelines of item writing, including the item specifications.

ETS trained California science teachers to develop items for the CAST field test during an item writing workshop in April 2016 (refer to subsections *3.3.5 Item Writer Training* and *3.3.4 Selection of Item Writers*). California science teachers were instructed to produce items that spanned a variety of SEPs and science domains (i.e., Life Sciences, Physical Sciences, Earth and Space Sciences, and Engineering, Technology and Applications of Science) to provide as wide an array of items as possible for the field test forms construction.

2.1.3. Item Types Guidelines

Because the item writers had limited experience in writing innovative items in general and were not familiar with the CA NGSS specifically, a limited number of item types were assigned, including some technology-enhanced item (TEI) types and constructed-response (CR) items. A key factor in determining the assignment of PEs to each item writer was the teaching experience and focus of expertise that the item writer possessed. ETS also generated item sets—performance tasks—internally to measure more complex skills in a particular domain.

The CAST field test was designed to assess the CA NGSS using discrete items, single and multipoint items, and performance tasks. There were a variety of item types, including traditional multiple-choice (MC), CR items, some familiar TEI types, as well as some new TEI types that utilized simulations and animations. Refer to subsection *3.3 Item Development* for more details on item volumes developed; refer to subsection *3.2 Item Types and Features* for the types of items used in the CAST field test.

2.2. Test Assembly

The 2017–18 field test design was based on the SBE-approved high-level test design for an operational assessment, which requires that all students in the tested grades participate in three segments of the test: Segment A, Segment B, and Segment C. ETS designed the general field test forms, to be taken in approximately two hours. ETS used historical timing data from previous assessments that had the same item types to estimate the amount of time needed to complete MC, CR, and TEI types. Subsection *3.7* provides additional details about test assembly.

2.3. Test Administration

It was of the utmost priority to administer the CAST field test in a secure, confidential, standardized, consistent, and appropriate manner. Additional information about the administration of the CAST field test can be found in *Chapter 4: Test Administration*.

2.3.1. Test Security and Confidentiality

All tests within the CAASPP System are secure. For the CAST field test, every person with access to test materials maintained the security and confidentiality of the tests. ETS' internal Code of Ethics requires that all test information, including tangible materials (e.g., test questions and test results), confidential files, processes, and activities are kept secure. To ensure security for all tests that ETS develops or handles, ETS maintains an Office of Testing Integrity (OTI). A detailed description of the OTI and its mission is presented in subsection *4.6.1 ETS' Office of Testing Integrity (OTI)*.

In the pursuit of enforcing secure practices, ETS strives to safeguard the various processes involved in a test development and administration cycle. The practices related to each of the following security processes are discussed in detail in subsection *4.6 Test Security and Confidentiality*.

2.3.2. Accessibility Resources

ETS offered a number of accessibility features available for the CAST field test.

2.3.2.1 Universal Tools

Universal tools are accessibility features of the assessment that are available to all students based on student preference and selection (Smarter Balanced, 2016, p. 6).

These resources are intended for use in the operational administration pending regulatory approval by the Office of Administrative Law. They are either embedded in the test delivery system or non-embedded, meaning they are not online.

2.3.2.1.1 Embedded

- Breaks
- Calculator[3]:
  - Four-function—grade five
  - Scientific—grade eight and high school
- Digital notepad
- English glossary
- Expandable passages
- Expandable items[4]
- Highlighter
- Keyboard navigation
- Line reader
- Mark for review
- Mathematics tools (e.g., ruler, protractor)[5]
- Science charts (i.e., calendar, Periodic Table of the Elements, conversion charts)
- Science tools (e.g., analog clock, laboratory equipment)

[3] These are the same as the calculators used during administration of the Smarter Balanced Mathematics Summative Assessment.

[4] The expandable items universal tool is turned on by the test administrator in the Test Administrator Interface.

[5] These are the same as the mathematics tools used during administration of the Smarter Balanced Mathematics Summative Assessment.

- Strikethrough
- Writing tools (e.g., bold, italic, bullets, undo/redo)
- Zoom (in/out)

2.3.2.1.2 Non-embedded

- Breaks
- Scratch paper

2.3.2.2 Designated Supports

Designated supports are accessibility resources that are available for use by any student for whom the need has been indicated by an educator or a team of educators (with parent/ guardian and student input as appropriate). The CAST field test followed the Smarter Balanced recommendations for use (Smarter Balanced, 2016).

These resources are intended for use in the operational administration pending regulatory approval by the Office of Administrative Law. They are either embedded in the test delivery system or non-embedded, meaning they are not online.

2.3.2.2.1 Embedded

- Color contrast
- Masking
- Mouse pointer (size and color)
- Stacked translations (Spanish)
- Text-to-speech (items and stimuli)
- Translations (glossary)
- Turn off any universal tool(s)

2.3.2.2.2 Non-embedded

- 100s number table
- Amplification
- Calculator:
  - Four-function—grade five
  - Scientific—grade eight and high school
- Color contrast
- Color overlay
- Magnification
- Noise buffers
- Read aloud for items and stimuli
- Read aloud in Spanish
- Science charts (i.e., calendar, Periodic Table of the Elements, conversion charts)[6]
- Scribe

[6] PDFs of the science charts are available for download from the California Science Test web page on the CAASPP Portal at https://bit.ly/3jZtcXA.

- Separate setting (e.g., most beneficial time, special lighting or acoustics, adaptive furniture)
- Simplified test directions
- Translated test directions

2.3.2.3 Accommodations

Accommodations are available to students who have a documented need for the accommodations via an individualized education program or Section 504 plan. The CAST field test followed the Smarter Balanced recommendations for use (Smarter Balanced, 2016).

These resources are intended for use in the operational administration pending regulatory approval by the Office of Administrative Law. They are either embedded in the test delivery system or non-embedded, meaning they are not online.

2.3.2.3.1 Embedded

- American Sign Language (ASL) (videos)
- Braille (embosser and refreshable)
- Closed captioning
- Streamline

2.3.2.3.2 Non-embedded

- Abacus
- Alternate response options
- Print on demand
- Speech-to-text
- Word prediction

2.4. Scores

The CAST field test contained traditional MC items, TEIs, and CR items. The MC items and TEIs were machine-scored through the test delivery system (TDS). The CR items were scored by trained raters. In addition, the CAST field test data were used to build the artificial intelligence (AI) scoring models and evaluate whether these models can be used to effectively score students' responses on the CR items. *Chapter 5: Scoring* provides details on scoring samples, machine scoring in the TDS, the human scoring process, and AI scoring models.

There were no individual student scores reported for the 2017–18 CAST field test. The ETS psychometrics team prepared an aggregate data file of students' percent correct scores and the associated preliminary indicator category for LEAs.

2.5. Analyses

Psychometric analyses were conducted on the data from the CAST field test, including classical item analyses, differential item functioning (DIF) analyses, dimensionality analyses, IRT calibration, response time analyses, reliability analyses, and special research studies, including a multistage practicality study and content screen-out study. The results of these analyses support the understanding of the item performances and internal structure and provide the validity evidence for both the response processes and scoring. Refer to *Chapter 6: Analyses* for descriptions of these analyses.

2.6. Reporting

The primary purposes of the CAST field test were to evaluate the properties of the items and provide the data for research studies to inform future test designs and score reporting. The forms developed did not fully conform to the test blueprint and, therefore, were not intended to provide a precise measure of students' achievement on the CA NGSS assessment. Instead, a preliminary indicator was reported to provide a broad and early indication about an LEA's implementation of the CA NGSS. *Chapter 7: Reporting* provides more information about the preliminary indicators.

References

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence-centered design* (ETS Research Report RR-03-16). Princeton, NJ: Educational Testing Service. Retrieved from [this link is no longer active].

Smarter Balanced Assessment Consortium. (2016). *Smarter Balanced Assessment Consortium: Usability, accessibility, and accommodations guidelines*. Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from https://bit.ly/2ZKqZZx

Chapter 3: Item Development and Assembly

This chapter discusses the detailed procedures of item development and test assembly for the California Science Test (CAST) field test administration. In particular, new item types and features that differ from traditional item types are described.

3.1. Use of Evidence-Centered Design (ECD)

3.1.1. Principles

The principles and practices of ECD guided the development of all CAST items. Developed at Educational Testing Service (ETS) in 1999, ECD is a framework for designing, producing, and delivering educational assessments so that evidence collected about student performance during testing provides support for claims about what students actually know and can do. ECD is an important tool used to support assessment validity arguments as well as inferences made about student scores (Mislevy, Almond, & Lukas, 2003).

As described in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing, 2014), a coherent validity argument, including alignment evidence, is essential to supporting the appropriateness of inferences made on the basis of an assessment's results. By employing ECD during the development process, ETS built the validity argument needed to support the operational use of the CAST.

3.1.2. Incorporation into Item Development Processes

For the CAST item development process, ETS began with the existing Achieve Next Generation Science Standards (NGSS) evidence statements that provide additional detail on what students should know and be able to do and describe the NGSS performance expectations in some detail (Achieve, 2015), draft work on the task models, and draft work on task templates to outline the types of items that would elicit student output sufficient to provide evidence for the performance expectation (PE) claims.

The task-model documentation is practice-based. ETS developed one design pattern for each California NGSS (CA NGSS) science and engineering practice and began developing one to three task templates for each design pattern. Each design pattern captured the results of domain analysis by specifying knowledge, skills, and abilities (KSAs) focal to the corresponding science and engineering practice (SEP), characteristics of the SEP that differ across the three grade bands, and characteristic features of assessments that elicit evidence of the focal KSAs.

During the drafting stage, ETS further specified approaches to the task templates designed to engage students meaningfully with the SEP by specifying item characteristics, work products, and observations that can be made about student proficiency from those work products. This documentation was used during both item development and revision to ensure that the student responses elicited by the items validly reflected the integrated science understanding specified in the targeted PEs.

ECD is an inherently iterative process. Lessons learned in one stage are used to refine both test design decisions and documentation for later stages. Information documented in some artifacts that were key to the development of the CAST field-test items was later incorporated into more comprehensive documents. For example, the information contained in the design patterns described previously was, for later rounds of item development, incorporated into more robust item specifications. Item specifications for each PE assessed on the CAST include assessment targets, framed from focal KSAs, for each dimension of the PE.

Similarly, the definition of claims for the CAST is an ongoing and iterative process, one informed both by the data collected from the CAST field test and the future data collection from the operational administration in 2018–19. Comprehensive documentation of this process is captured in a white paper titled “Use of Evidence-Centered Design in CAST Item and Test Development” (ETS, 2019).

3.2. Item Types and Features

Every item assessed a CA NGSS disciplinary core idea (DCI) as well as at least one of the other two CA NGSS dimensions (i.e., SEP or crosscutting concept [CCC]). Wherever possible, a single item assessed all three dimensions. However, leading NGSS experts agreed that this was not always practical to assess all three dimensions using a single item (ETS, 2016a).

ETS used item types, individually and in combinations or sets, to measure targeted CA NGSS content. In some cases, the presentation of the content involved the use of dynamic stimuli and other types of new media—e.g., animations of scientific phenomena, real-life engineering challenges, and simulated experiments run multiple times by a student to generate data for analysis—to provide rich opportunities for students to demonstrate their scientific knowledge and skills.

For the item development process, ETS developed item types and features for the 2017–18 field test that were supported by *Instructional Management Systems (IMS) Global Question and Test Interoperability (QTI)* standards (IMS, 2016).

Table 3.1 outlines the major categories of QTI item types that were included in the CAST field test. This includes item types ranging from traditional multiple choice (MC) and constructed response (CR) (i.e., extended text) to new technology-enhanced item (TEI) types (the rest of the item types).

Table 3.1 Selected Item Types in the CAST Field Test

Feature	Description
Choice	Traditional single-select or multiple-select MC items
Extended Text	Traditional essay or other CR items, where the student provides a text response
Hot Spot	Items that present a graphic—such as an anatomical diagram or a drawing of laboratory equipment—where a student selects a part of the graphic as the response
Match	Items that present multiple pieces of evidence for a student to match to each of various alternate conclusions, and items that present a grid with row and column headings (e.g., representing alternate experimental designs to address alternate hypotheses), where a student selects table cells as the response to indicate which experimental design is appropriate to test each hypothesis

Feature	Description
Inline Choice	Items that provide multiple choices for filling in one or more blanks within a sentence or paragraph
Custom	Items where a student manipulates an object, such as a scale, a histogram, a clock, or an arrangement of laboratory materials; a collection of interactive items and custom interactive stimuli in a set with multiple-scored interactive components (e.g., simulations)

3.3. Item Development

3.3.1. Plan

The initial item development plan for the CAST field test focused on developing items that integrated at least two of the three dimensions of the CA NGSS—DCIs, SEPs, and CCCs. The plan incorporated a diverse selection of PEs to incorporate a range of SEPs, DCIs, and CCCs.

Table 3.2 shows the total number of items developed per grade to accommodate the field and training tests, as described in subsection *4.5 Training Test*.

Table 3.2 Total Number of Items Developed per Grade for the CAST Field Test

Item Type	Grade 5	Grade 8	High School
Standard discrete item types (non-CR)	177	167	177
Discrete CR	7	16	10
Custom discrete interactive items	23	20	17
Performance task items (eight tasks in each grade)	44	49	46
TOTAL	**251**	**252**	**250**

CR items included the extended text item type shown in Table 3.1. Discrete items included traditional MC items, CR items, and some familiar TEI types (e.g., match, inline choice list, zone or hot spot, etc.), as well as some new TEI types that utilized simulations and animations, which are also indicated as custom, discrete, interactive items. The performance task, which contained four to six items for the CAST field test, is designed to provide students with an opportunity to demonstrate their ability to apply knowledge and higher-order thinking skills to explore and analyze a complex, real-world scenario.

ETS developed all items for the CAST field test in accordance with the *ETS Standards for Quality and Fairness* (2014) across all phases of item and test development.

3.3.2. Process

Each CAST item was developed through a comprehensive development cycle and designed to conform to principles of quality item writing as defined by ETS. Further, each item in the CAST item bank was developed to measure a specific PE through integration of at least two of the three dimensions of the CA NGSS (i.e., DCI, CCC, and SEP). In addition, guidelines for style and for fairness—including issues related to bias and sensitivity—helped item developers and reviewers maintain consistency across the item development process.

Throughout the item writing process, ETS adhered to its foundational guidelines for quality item writing. According to these guidelines, item developers conformed to the following list of attributes for each item:

1. The question is clearly and concisely presented.
2. There is an absence of clueing in the item stem and supporting stimuli.
3. The supporting stimulus/stimuli are presented clearly and are construct-relevant.
4. There is a single correct answer (for selected-response items only).
5. Distractors are plausible, but incorrect (for selected-response only).
6. The answer key is correct.
7. The scoring rubric and annotations are accurate, precise, and complete.
8. Item format and content adhere to the principles of universal design.

3.3.3. Specifications

ETS created item specifications for the CAST field test using feedback from the California Department of Education (CDE) and California teachers with task models guiding the initial development. The item specifications are extensions of these models intended to be more specific in nature and to incorporate information and feedback gained through the development, review, and administration processes. These specifications describe the characteristics of items that consistently elicit evidence of student mastery of specified aspects of each PE. The specifications were developed in consultation with the CDE, and the CDE determined the emphasis on different aspects of each PE. The specifications include the following:

- Subpractice
- Subpractice assessment targets
- DCI assessment targets
- CCC assessment targets
- Possible phenomena or contexts
- Examples of integration of assessment targets and evidence
- Common misconceptions
- Additional assessment boundaries

In accordance with the iterative nature of ECD described previously, the item specifications used to produce the field test items will be updated annually and expanded to support subsequent rounds of item development.

3.3.4. Selection of Item Writers

Senior ETS content staff screened applications for item writers for the CAST field test, and ETS approved only those with strong content and teaching backgrounds for the item writing training program. ETS selected item writers after the training, but not all recipients of the training became an item writer.

Because some of the participants were current or former California educators, they were particularly knowledgeable about the standards assessed by the CA NGSS. All item writers met the following minimum qualifications:

- Possession of a bachelor’s degree in science or in the field of education with special focus on a particular scientific domain; an advanced degree in the relevant content was desirable

- Previous experience or training in writing items for standards-based assessments, including knowledge of the many considerations that are important when developing items for special student populations
- Previous experience or training in writing items in the grades and content areas covered by the CAST field test
- Familiarity and understanding of the CA NGSS

3.3.5. Item Writer Training

Item writer training is a vital part of establishing the validity chain for item and task development. In addition to relying on internal item writing experts for the CAST field test, ETS recruited and trained science educators with diverse science backgrounds, including California teachers, to enrich the range of ideas brought to the process and support effective teaching practices in science.

The primary goals for the training were to

1. provide teachers with knowledge, via professional development on writing items, that they can use to help develop or refine their own classroom teaching and assessments;
2. ensure that teachers who successfully completed the training were ready to develop high-quality items for the CAST field test; and
3. leverage the experiences, perspectives, and expertise of the teachers in writing items for the CAST field test.

ETS held an item writer–training workshop in November 2016 in Sacramento, California, to provide prospective item writers with professional development in several areas. A review of the general assessment development process gave trainees a sense of the total lifecycle of an item. The dimensions of the CA NGSS (i.e., DCI, CCC, and SEP) were analyzed and explored to focus on the three dimensions of the CA NGSS that items for the CAST field test were to emphasize. To achieve this three-dimensional quality and maintain validity, ETS explained how items should elicit evidence of student reasoning instead of rote recall of science content associated with the DCI. Finally, ETS shared with trainees best practices in item writing to provide clarity within the item and avoid bias or sensitivity concerns.

Given that the trainees were California educators and educational leaders, ETS also emphasized incorporation of current effective teaching practices and instructional activities. Small-group and individual work generated sample items that the ETS facilitators then used in a large-group discussion to analyze alignment to the dimensions of the PEs in question and ascertain overall item quality. The ETS team also provided post hoc feedback via email and phone calls to trained item writers on further item samples and ideas submitted ahead of contractual item submissions.

3.4. Item-Review Process

ETS placed items developed for the CAST field test through an extensive internal item-review process. This section summarizes the item-review process that confirmed the quality of CAST field test items.

Once an item was accepted for authoring, ETS employs a series of internal reviews. These reviews use established criteria to judge the quality of item content and to ensure that each item measures what it was intended to measure. These internal reviews also examine the overall quality of the test items before presentation to the CDE and item-review meetings, which are described in more detail in subsection *3.5 Content Expert Reviews*.

The ETS review process for the CAST includes the following; these are described in the next subsections.

1. Content review
2. Research review
3. Editorial review
4. Fairness review

Throughout this multistep item-review process, the lead content-area assessment specialists and development team members continually evaluate the activities and items for adherence to the rules for item development.

3.4.1. ETS Content Review

CAST items and stimuli undergo three rounds of content reviews by content-area assessment specialists with increasing levels of expertise; these rounds are called Round 1, Round 2, and Final Round. These assessment specialists verify that the items and stimuli were in compliance with the approved item specifications and with ETS' written guidelines for clarity, style, accuracy, and appropriateness for California students, as well as in compliance with the task models. Assessment specialists review each item for the following characteristics:

- Relevance of each item to the purpose of the test
- Match of each item to the task model, including Depth of Knowledge
- Match of each item to the principles of quality item writing
- Match of each item to the identified standard or standards
- Difficulty of the item
- Accuracy of the content of the item
- Readability of the item or passage
- Grade-level appropriateness of the item
- Appropriateness of any illustrations, graphs, or figures

Each item is classified with the PE that it is intended to measure. The assessment specialists check each item against its classification codes, both to evaluate the correctness of the classification and to confirm that the task posed by the item is relevant to the outcome it is intended to measure. The reviewers have the choice to accept the item and classification as written, suggest revisions, or recommend that the item be discarded. These steps occur prior to the CDE's review.

3.4.2. ETS Research Review

Internal science researchers, who also contribute to the ECD documentation, review a proportion of items with a focus on the alignment issues at the item level and provide potential refinement solutions to improving the integration of three dimensions according to the PE statements. This review process helps guide content specialists toward proper alignment to the CA NGSS standards through iterative development process of items.

3.4.3. ETS Editorial Review

After content-area assessment specialists and researchers review each item, a group of specially trained editors also review each item in preparation for consideration by the CDE and item-review meetings. The editors check items for clarity, correctness of language, appropriateness of language for the grade level assessed, adherence to the style guidelines, and conformity with accepted item-writing practices.

3.4.4. ETS Sensitivity and Fairness Review

ETS assessment specialists who are specially trained to identify and eliminate questions that contain content or wording that could be construed to be offensive to or biased against members of specific student groups—e.g., ethnic, racial, or gender—conduct the next level of review (ETS, 2014, 2016b). These trained staff members review every item before the CDE and item-review meeting reviews.

The review process promotes a general awareness of and responsiveness to the following:

- Cultural diversity
- Diversity of background, cultural tradition, and viewpoints to be found in the test-taking populations
- Changing roles and attitudes toward various groups
- Role of language in setting and changing attitudes toward various groups
- Contributions of diverse groups (including ethnic and minority groups, individuals with disabilities, and women) to the history and culture of the United States and the achievements of individuals within these groups
- Item accessibility for English learners

3.5. Content Expert Review

3.5.1. California Educators as Content Experts

In addition to the ETS internal content reviews, meetings with California educators are held at the end of the item-review process as the final content expert review that items must undergo before being placed on a field test. The California educators fill an advisory role to the CDE and ETS and provide guidance on matters related to item development for the CAST field test. These educators are responsible for reviewing all newly developed items for alignment to the CA NGSS. Meeting participants also review the items for accuracy of content, clarity of phrasing, and overall quality. In their examination of test items, participants can raise concerns related to grade appropriateness as well as gender, racial, ethnic, or socioeconomic bias.

3.5.2. Composition of Item-Review Panels

The item-review meetings for the field test items were comprised of current and former teachers, resource specialists, administrators, curricular experts, and other education professionals. Minimum qualifications to be invited to participate are

- three or more years of general teaching experience in grades kindergarten through twelve
- three or more years of teaching experience in science,
- bachelor’s or higher degree in science or education, and
- knowledge of and experience with the CA NGSS.

School administrators; local educational agency (LEA), county content, or program specialists; or university educators meet the following qualifications to be invited to participate:

- Three or more years of experience as a school administrator, LEA, county content, program specialist; or university instructor in a grade-specific area or area related to science
- Bachelor's or higher degree in a grade-specific or content area related to science
- Knowledge of and experience with the CA NGSS

Item-review meeting attendees were recruited through an online application process. Recommendations were solicited from LEAs and county offices of education as well as from the CDE and State Board of Education (SBE) staff. ETS assessment directors reviewed applications and confirmed that the applicant's qualifications met the specified criteria. Applications that met the criteria were forwarded to CDE and SBE staff for further review and agreement on item-review meeting inclusion.

3.5.3. Meetings for Review of CAST Field-Test Items

ETS content-area assessment specialists facilitated CAST field test item-review meetings. Each meeting began with a brief training session on how to review and make recommendations for revising items. ETS provided training on the following topics:

- Overview of the purpose and scope of the CAST field test
- Overview of the CAST field test design specifications
- Overview of criteria for evaluating test items
- Review and evaluation of items for fairness issues

The criteria for reviewing items included the following:

- Overall technical quality
- Alignment with the PEs
- Alignment with the construct being assessed by the standard
- Difficulty range
- Clarity
- Correctness of the answer
- Plausibility of the distractors
- Bias and sensitivity factors

ETS provided guidelines for reviewing items, which the CDE approved. A summary of the set of guidelines for reviewing items follows.

- Does the item
  - have one and only one clearly correct answer?
  - measure the achievement standard?
  - align with the construct being measured?
  - test worthwhile concepts or information?
- Is the stimulus, if any, for the item
  - required in order to answer the item?
  - likely to be interesting to students?

- clearly and correctly labeled?
- providing all the information needed to answer the item?

Once ETS staff compiled and reviewed the panel's feedback, the feedback was delivered to the CDE for further review and guidance on decisions.

3.6. Data Review

After items have been included in an operational or field test and administered to students, ETS conducts data review meetings with California teachers and the CDE after the data analysis is complete. Reviewers examined items that were flagged for item difficulty, item-total correlation, item response distribution, and differential item functioning according to predefined criteria. The ETS facilitator leads discussions about each flagged item and reviewed the content of the item to reach consensus on whether items should be accepted as is, accepted with revision, or rejected.

For items that are accepted with revision, California teachers participate in making suggested edits. As time allowed, ETS shows the statistics for items that were not flagged to determine if there are any edits that the stakeholders feel should be made prior to operational testing of the items. Refer to Table 3.3 for the results of the data review, showing the number of items accepted with and without edits and the number of items rejected outright.

Table 3.3 Data Review Results

Grade Level	Accept As Is	Accept with Edits	Reject	Total Items
5	158	0	51	209
8	172	29	10	211
High School	164	10	50	224

3.7. Test Assembly

ETS designed the general field test forms to be taken in approximately two hours. ETS used historical timing data from previous assessments that had the same item types to estimate the amount of time needed to complete MC, CR, and TEI types.

ETS designed a braille form consisting of items from the general field test further adapted to be accessible to students with visual impairments using current assistive technology while still maintaining the construct of the item on the general field test. The braille form for each grade level was comprised of two Segment A discrete item blocks, two Segment B performance tasks, and one Segment C discrete item block.

The various blocks of items that comprise each segment of the field test covered an extensive range of PEs; these PEs are shown for all three grade levels in Table 3.4.

Table 3.4 Performance Expectations Assessed on the General CAST Field Test—All Grade Levels

Grade Level	PEs Assessed	PEs Available	Percent of PEs Assessed
5	45	45	100%
8	59	59	100%
High School	68	71	96%

3.8. General Forms

The CAST 2017–18 Field Test Plan (ETS, 2018) design is based on the SBE-approved high level test design for an operational assessment, which requires that all students in the tested grades participate in three segments of the test: Segment A, Segment B, and Segment C. For Segment A, ETS assembled four field test blocks of 16 to 20 discrete items per grade in which any two blocks represented the approved blueprint for Segment A. There were some blocks that exceeded the blueprint in the number of items but met the overall points needed.

Segment B assembly included eight blocks with each block representing one performance task (PT) of four to six items each. Each PT represented the blueprint in number of items, but some PTs lacked the necessary multipoint items to meet the point values of six to seven points required in the blueprint.

Segment C assembly included eight discrete item blocks with 13 to 14 items each. In addition, the same pool of PTs used in Segment B were also used for Segment C such that a student could have received a third PT that was different than the two received in Segment B.

3.9. Forms with Accessibility Features

A subset of the general form blocks were used to provide accessible content for those students whose individualized education program indicated that one or more designated supports or accommodations be used. Items were embedded with content for text-to-speech, stacked Spanish, translation glossaries, and ASL videos.

Two of the four general form Segment A blocks were used for the designated supports and accommodations. Segment B had three PT blocks for these resources. Segment C designated one of the eight discrete blocks of 13 items as accessible for these resources, and the same PTs used for Segment B were available for Segment C.

The same Segment A, B, and C pool of items and PTs that were used for embedded designated supports and accommodations were also used for braille. However, if an item that relied heavily on visual input, whether through item type or visual stimuli, was needed to meet the blueprint, the item was either adapted or “twinned” to meet the accessible needs of the population of students with visually impairments. Adaptation may have included simplified graphics, more descriptive alternate text for images, or other change to make the item more accessible to refreshable braille devices, embossed tactile graphics, or screen readers. Twinning an item meant the item was rewritten using another item type while maintaining the same construct and storyline of the original item.

References

Achieve. (2015). *Next Generation Science Standards evidence statements.* Available from https://bit.ly/3BByezx

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

Educational Testing Service. (2019). *Use of evidence-centered design in CAST item and test development.* Manuscript submitted to the California Department of Education.

Educational Testing Service. (2018). *California Science Test 2017–18 field test plan.* Manuscript submitted to the California Department of Education.

Educational Testing Service. (2014). *ETS standards for quality and fairness*. Princeton, NJ: Educational Testing Service. Retrieved from https://bit.ly/3jZIPOK

Educational Testing Service. (2016b). *ETS guidelines for fair tests and communications*. Princeton, NJ: Educational Testing Service. Retrieved from [this link is no longer active]

Educational Testing Service. (2016a). *Proposed design for California's Next Generation Science Standards general summative assessments.* Princeton, NJ: Educational Testing Service.

Instructional Management Systems (IMS). (2016). *IMS question & test interoperability specification.* Available from https://bit.ly/3jX22Ra

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence-centered design* (ETS Research Report RR-03-16). Princeton, NJ: Educational Testing Service. Retrieved from [this link is no longer active]

Chapter 4: Test Administration

This chapter describes the details of the California Science Test (CAST) field test administration, including procedures to ensure test security and procedures to implement the test accommodations based on Standard 7.8 of the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014).

4.1. Student Participation Requirement

The CAST field test was administered to students in grades five, eight, and twelve, and students in grades ten and eleven as assigned by their local educational agency (LEA). For grades five, eight, and twelve, a census field test was conducted. In the CAST census field test, all students in California enrolled in grades five, eight, and twelve who were eligible for the general science assessments (i.e., not eligible for the California Alternate Assessment [CAA] for Science) were required to participate. Subsection *4.1.2 High School* outlines the process for grade assignment in the CAST field test for high school students.

4.1.1. Grades Five and Eight

All students enrolled in grades five and eight were registered to participate in the CAST field test. The Test Operations Management System (TOMS) assigned participant eligibility to all grades five and eight students with the exception of students with the most significant cognitive disabilities who are designated in TOMS to take the CAA for Science if their individualized education program indicates an alternate assessment.

4.1.2. High School

At the high school level, for the pilot year in 2016–17, to reduce the testing burden on schools and LEAs and to facilitate the computation of the participation rate for accountability, high schools were assigned to test their students in grades ten, eleven, or twelve. For the field test year, instead of using grade-level assignments, the operational administration rule was used to determine eligibility for testing, meaning that schools and LEAs were responsible for ensuring that students who completed or are in the process of completing their last high school science course and who are not eligible for CAA for Science participate in the field test. All students in grade twelve must have tested during this field test administration.

By implementing the high school administration rule during the field-test year, schools and LEAs had a chance to work out the logistics of tracking students for testing, and the test-taker population for the field test was more likely to resemble the operational test-taker population.

Students who participated in the pilot were not exempt from participating in the field test. Accordingly, a student’s eligibility for testing was not based solely on his or her science coursework trajectory, as it will be operationally, with the caveat that all students in grade twelve must test in the field test administration.

4.2. Participation Rates

Table 4.1 provides the composition of the test-taker population for the CAST field test for high school students. The sum of the number of schools for grades ten, eleven, and twelve did not equal the total number of unique school because each school may include multiple grades.

Table 4.1 Composition of Test-taker Population for the CAST Field Test for High School Students

Variable	Grade 10	Grade 11	Grade 12	Total
Number of Schools	104	800	2,546	2,609
Percent of Schools	4%	31%	98%	100%
Number of Students	6,384	136,663	405,008	548,055
Percent of Students	1%	25%	74%	100%

Table 4.2 presents the participation rates of each test as well as participation rates for each grade for the high school test. Note that participants are enrolled students who log on to the test.

Table 4.2 CAST Field Test Participation Rates of the Full Population

Group	Grade 5	Grade 8	HS— Grade 10	HS— Grade 11	HS— Grade 12	HS— All Grades
Number of Enrolled Students	469,247	472,094	16,628	151,135	466,600	634,363
Number of Participants	460,303	458,523	6,384	136,675	405,051	548,110
Percent of Participation	98.1	97.1	38.4	90.4	86.8	86.4

4.3. Demographic Summaries

Appendix 4.A shows the participation rates of selected demographic student groups in each test. The demographic student groups include gender, ethnicity, English-language fluency, economic status (disadvantaged or not), special education services status, and migrant status.

Demographic student groups included in the summaries in this chapter are shown in Table 4.3. The number and the percent of students for these demographic student groups are provided in appendix 4.B, starting in Table 4.B.1 through Table 4.B.5 for each grade, and in Table 4.B.6 for the high school, which combines students from grades ten, eleven, and twelve.

For the high school test, because not all students from grades ten and eleven were required to take the test, the demographic composition of the students who participated could be different from the demographic composition of the total population of students at a grade. Therefore, a column with the population’s percent was added in each of tables for high school grades (i.e., Table 4.B.3 to Table 4.B.6) to provide information on how the demographic composition for the testing population is different from the student population.

Table 4.3 Demographic Student Groups to Be Reported

Student Group	Definition
Gender	• Male • Female
Ethnicity	• American Indian or Alaska Native • Asian • Black or African American • Filipino • Hispanic or Latino • Native Hawaiian or Other Pacific Islander • White • Two or more races
English Language Fluency	• English only • Initial fluent English proficient • English learner • Reclassified fluent English proficient • To be determined • English proficiency unknown
Economic Status	• Not economically disadvantaged • Economically disadvantaged
Primary Disability Type	• No special education services • Special education services
Migrant Status	• Eligible for the Title I Part C Migrant Program • Not eligible for the Title I Part C Migrant Program

4.4. Accessibility

There was not an accessibility form for the field test as there was for the pilot. A subset of the general form blocks were used to provide accessible content. Refer to subsection *3.9 Forms with Accessibility Features* for more information about accessibility features of the CAST field test.

4.5. Training Test

The training test was designed to provide students with an opportunity to engage with California Next Generation Science Standards–aligned items, including TEIs. It also allowed students to familiarize themselves with the test settings, including universal tools, available for the field test. A *Training Items Scoring Guide* was available for test administrators to offer details about the items, student response types, correct responses, and related scoring considerations for the included sample of training items (California Department of Education [CDE], 2017a). In addition, the training test allowed educators to familiarize themselves with the organization of the CAST field test and help maintain the standardization of test administration.

Grade-specific training tests were released in March 2018 to replace the single training test with content from all grade levels that was previously available. These training tests are available through the Practice and Training Test website (CDE, 2017b). The training tests and the scoring guides for the training test are available to anyone with internet access.

4.6. Test Security and Confidentiality

For the CAST field test, every person who worked with the assessments, communicated test results, or received testing information was responsible for maintaining the security and confidentiality of the tests, including CDE staff, ETS staff, ETS subcontractors, LEA assessment coordinators, school assessment coordinators, students, parents, teachers, and others. ETS' Code of Ethics required that all test information, including tangible materials (e.g., test items), confidential files (e.g., those containing personally identifiable student information), and processes related to test administration (e.g., the configuration of secure servers) are kept secure. To ensure security for all tests that ETS develops or handles, ETS maintains an Office of Testing Integrity (OTI).

All tests within the California Assessment of Student Performance and Progress (CAASPP) System, as well as the confidentiality of student information, are protected to ensure the validity, reliability, and fairness of the results. As stated in *Standard 7.9* (AERA, APA, & NCME, 2014), "The documentation should explain the steps necessary to protect test materials and to prevent inappropriate exchange of information during the test administration session" (p. 128).

This section of the *CAST Technical Report* describes the measures intended to prevent potential test security incidents prior to testing and the actions that were taken to handle security incidents occurring during or after the testing window using the Security and Test Administration Incident Reporting System (STAIRS) process.

4.6.1. ETS' Office of Testing Integrity (OTI)

The OTI is a division of ETS that provides quality assurance services for all ETS-managed testing programs. This division resides in the ETS legal department. The Office of Professional Standards Compliance at ETS publishes and maintains the *ETS Standards for Quality and Fairness* (2014), which supports the OTI's goals and activities. The *ETS Standards for Quality and Fairness* provides guidelines to help ETS staff design, develop, and deliver technically sound, fair, and beneficial products and services and help the public and auditors evaluate those products and services.

The OTI's mission is to

- minimize any testing security violations that can impact the fairness of testing,
- minimize and investigate any security breach that threatens the validity of the interpretation of test scores, and
- report on security activities.

The OTI helps prevent misconduct on the part of students and administrators, detects potential misconduct through empirically established indicators, and resolves situations involving misconduct in a fair and balanced way that reflects the laws and professional standards governing the integrity of testing. In its pursuit of enforcing secure testing practices, the OTI strives to safeguard the various processes involved in a test development and administration cycle.

4.6.2. Procedures to Maintain Standardization of Test Security

Test security requires the accounting of all secure materials—including online summative test items and student data—before, during, and after each test administration. For the CAST field test, as well as for all CAASPP assessments, the LEA CAASPP coordinator was responsible for keeping all electronic test materials secure, keeping student information confidential, and making sure the CAASPP test site coordinators and test administrators were properly trained regarding security policies and procedures.

The CAASPP test site coordinator was responsible for mitigating test security incidents at the test site and for reporting incidents to the LEA CAASPP coordinator.

The test administrator was responsible for reporting testing incidents to the CAASPP test site coordinator and securely destroying printed and digital media for CAST items generated by the print-on-demand feature of the test delivery system (CDE, 2018a).

The following measures ensured the security of CAASPP System assessments administered in 2017–18:

- LEA CAASPP coordinators and test site coordinators must have signed and submitted a “CAASPP Test Security Agreement for LEA CAASPP coordinators and CAASPP test site coordinators” form to the California Technical Assistance Center (CalTAC) before ETS granted the coordinators access to TOMS. (*California Code of Regulations*, Title 5 [5 *CCR*], Education, Division 1, Chapter 2, Subchapter 3.75, Article 1, Section 859[a])
- Anyone having access to the testing materials must have signed and submitted a “Test Security Affidavit for Test Examiners, Test Administrators, Proctors, Translators, Scribes, and Any Other Person Having Access to CAASPP Tests” form to the CAASPP test site coordinator before receiving access to any testing materials. (5 *CCR*, Section 859[c])

In addition, it was the responsibility of every participant in the CAASPP System to report immediately any violation or suspected violation of test security or confidentiality. The CAASPP test site coordinator reported to the LEA CAASPP coordinator, and the LEA CAASPP coordinator reported to the CDE within 24 hours of the incident. (5 *CCR*, Section 859[e])

4.6.3. Security of Electronic Files Using a Firewall

A firewall software is currently used to prevent unauthorized entry to files, email, and other organization-specific information. All ETS data exchanges and internal email remain within the ETS firewall at all ETS locations, ranging from Princeton, New Jersey; to San Antonio, Texas; and to Concord and Sacramento, California.

All electronic applications that are included in TOMS remain protected by the ETS firewall software at all times. Due to the sensitive nature of the student information processed by TOMS, the firewall plays a significant role in maintaining assurance of confidentiality among the users of this information.

4.6.4. Transfer of Scores via Secure Data Exchange

Due to the confidential nature of test results, ETS currently uses secure file transfer protocol (SFTP) and encryption for all data file transfers; test data is never sent via email. SFTP is a method for reliable and exclusive routing of files. Files reside on a password-protected server that only authorized users can access. ETS shares an SFTP server with the CDE.

On that site, ETS posts Microsoft Word and Excel files, Adobe Acrobat PDFs, or other document files for the CDE to review; the CDE returns reviewed materials in the same manner. Files are deleted upon retrieval.

The SFTP server is used as a conduit for the transfer of files; secure test data is only temporarily stored on the shared SFTP server. Industry-standard secure protocols are used to transfer test content and student data from the ETS internal data center to any external systems.

ETS enters information about the files posted to the SFTP server in a web form on a SharePoint website; a CDE staff member monitors this log throughout the day to check the status of deliverables and downloads and deletes the file from the SFTP server when its status shows it has been posted.

4.6.5. Data Management in the Secure Database

ETS currently maintains a secure database to house all student demographic data and assessment results. Information associated with each student has a database relationship to the LEA, school, and grade codes as the data is collected during operational testing. Only individuals with the appropriate credentials can access the data. ETS builds all interfaces with the most stringent security considerations, including interfaces with data encryption for databases that store test items and student data. ETS applies best and up-to-date security practices, including system-to-system authentication and authorization, in all solution designs.

All stored test content and student data is encrypted. Industry-standard secure protocols are used to transfer test content and student data from the ETS internal data center to any external systems. ETS complies with the Family Educational Rights and Privacy Act (20 *United States Code [USC]* § 1232g; 34 *Code of Federal Regulations* Part 99) and the Children’s Online Privacy Protection Act (15 USC §§ 6501-6506, P.L. No. 105–277, 112 Stat. 2681–1728).

In TOMS, staff at LEAs and test sites have different levels of access appropriate to the role assigned to them.

4.6.6. Statistical Analysis on Secure Servers

During CAASPP testing, ETS information technology staff retrieves data files from the American Institutes for Research and loads those files into a database. The ETS Data Quality Services staff extracts the data from the database and performs quality control procedures (e.g., the values of all variables are as expected) before passing files to the ETS statistical analysis group (refer to subsection *8.6 Quality Control of Psychometric Processes* for data validation processes undertaken by ETS Data Quality Services). The statistical analysis staff stores the files on secure servers. All staff involved with the data adheres to the ETS Code of Ethics and the ETS Information Protection Policies to prevent any unauthorized access to the data.

4.6.7. Student Confidentiality

To meet requirements of the Every Student Succeeds Act as well as state requirements, LEAs must collect demographic data about students’ ethnicity, disabilities, parent/guardian education, and so forth during the school year. ETS takes every precaution to prevent any of this information from becoming public or being used for anything other than for testing and score reporting purposes. These procedures are applied to all documents in which student demographic data appears, such as technical reports.

4.6.8. Security and Test Administration Incident Reporting System (STAIRS) Process

Test security incidents, such as improprieties, irregularities, and breaches, are prohibited behaviors that give a student an unfair advantage or compromise the secure administration of the tests, which, in turn, compromise the reliability and validity of test results (CDE, 2018b). Whether intentional or unintentional, failure by staff or students to comply with security rules constitutes a test security incident. Test security incidents have impacts on scoring and affect students' performance on the test.

For the CAST field test, LEA CAASPP coordinators and CAASPP test site coordinators verified that all test security and summative administration incidents were documented by filling out the secure STAIRS form for reporting, which contained selectable options to guide coordinators in their submittal. After the form was submitted, an email containing a case number and next steps was sent to the submitter (and to the LEA CAASPP coordinator, if the form was submitted by the CAASPP test site coordinator). Coordinators could not file an appeal without the case number that is created by submitting the *CAASPP STAIRS* form. The *CAASPP STAIRS* form provided the LEA CAASPP coordinator, the CDE, and the CalTAC with the opportunity to interact and communicate regarding the STAIRS process. (CDE, 2018b)

Any incidents were then resolved when the LEA CAASPP coordinator or CAASPP test site coordinator either filed an appeal to reset, re-open, invalidate, restore, or grant a grace period extension to a student's test, or by following other instructions in a system-generated email in response to the *CAASPP STAIRS* form submittal.

The following types of STAIRS reports were also forwarded to the CDE:

- Student cheating
- Security breach (where either a student or an adult exposed secure materials)
- Accidental access to a summative assessment
- Incorrect Statewide Student Identifier used (i.e., intentionally switched)
- Restoring a test that had been reset
- Student unable to review previous answers (i.e., 20-minute pause rule)

The CDE reviewed appeals requests. Appeals could not be requested without a STAIRS case number (CDE, 2018b).

4.6.8.1 Impropriety

A testing impropriety is an unusual circumstance that has a low impact on the individual or group of students who are testing and has a low risk of potentially affecting student performance on the test, test security, or test validity. An impropriety can be corrected and contained at a local level. An impropriety should have been reported to the LEA CAASPP coordinator and CAASPP test site coordinator immediately. The coordinator reported the incident within 24 hours, using the online *CAASPP STAIRS* form.

4.6.8.2 Irregularity

A testing irregularity is an unusual circumstance that impacts an individual or a group of students who are testing and may potentially affect student performance on the test, or impact test security or test validity. These circumstances can be corrected and contained at the local level and submitted in the online Appeals System for resolution. An irregularity should have been reported to the LEA CAASPP coordinator and CAASPP test site coordinator immediately. The coordinator reported the irregularity within 24 hours, using the online *CAASPP STAIRS* form.

4.6.8.3 Breach

A testing breach is an event that poses a threat to the validity of the test and requires immediate attention and escalation to CalTAC (for social media breaches) or the CDE (for all other breaches) via telephone. Examples may have included such situations as a release of secure materials or a security or system risk. These circumstances have external implications for the CDE and may result in a CDE decision to remove the test item(s) from the available secure item bank. A breach incident should have been reported to the LEA CAASPP coordinator immediately.

4.6.9. Appeals

For test security incidents reported in STAIRS that result in a need to reset, reopen, invalidate, or restore individual online student assessments, the CDE must approve the request. In most instances, an appeal was submitted to address a test security breach or irregularity. The LEA CAASPP coordinator or CAASPP test site coordinator may submit appeals in TOMS. All submitted appeals are available for retrieval and review by the appropriate credentialed users within a given organization. However, the view of appeals was restricted according to the user role as established in TOMS (CDE, 2018b).

Table 4.4 describes types of appeals available during the 2017–18 CAASPP administration.

Table 4.4 Types of Appeals

Type of Appeal	Description
Reset	Resetting a student’s summative assessment removes that assessment from the system and enables the student to start a new assessment from the beginning.
Invalidation	Invalidated summative tests will be scored and scores will be provided on the Student Score Report with a note that an irregularity occurred. (Note that for the 2017–18 administration, results of the CAST field test were not reported except as Preliminary Indicators.) The student(s) will be counted as participating in the calculation of the school’s participation rate for accountability purposes.
Re-open	Reopening a summative test allows a student to access an assessment that has already been submitted.
Restore	Restoring a summative test returns a test from the Reset status to its prior status. This action could only be performed on tests that have been previously reset.
Grace Period Extension	Permitting a Grace Period Extension allows the student to review previously answered questions upon logging back on to the assessment after expiration of the pause rule.

References

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

California Department of Education. (2017a). *2017–18 California Science Test training items scoring guide.* Sacramento, CA: California Department of Education. Retrieved from https://bit.ly/3bzpoYF

California Department of Education. (2018a). *CAASPP Smarter Balanced online test administration manual, 2017–18 administration.* Sacramento, CA: California Department of Education. Retrieved from https://bit.ly/3GBmwZH

California Department of Education. (2017b). *CAST grades five, eight, and high school training test*. Retrieved from [this link is no longer active]

California Department of Education. (2018b). *Security incidents and appeals procedure guide, 2017–18 administration.* Sacramento, CA: California Department of Education. Retrieved from https://bit.ly/2ZMeTPA

Educational Testing Service. (2014). *ETS standards for quality and fairness*. Princeton, NJ: Educational Testing Service. Retrieved from https://bit.ly/3jZIPOK

Appendix 4.A: Participation Rates

Notes:

- This set of tables show the percent of participants of selected demographic student groups in each test.
- The total numbers of registered students are derived from the version 4 of the production data file (“P4”).
- A student is considered a participant if he or she was enrolled during the active testing window and logged on to the test.
- High school grades are ten, eleven, and twelve.

Table 4.A.1 CAST Field Test Participation Rates for Grade Five by Student Group

Group	Number of Eligible Students	Number of Participants	Percent of Participation
All students	469,247	460,303	98.1
Male	240,381	235,428	97.9
Female	228,866	224,875	98.3
English learner	90,830	89,288	98.3
English only	269,847	263,488	97.6
Reclassified fluent English proficient	89,703	89,131	99.4
Initial fluent English proficient	17,841	17,673	99.1
To be determined	323	242	74.9
English proficiency unknown	703	481	68.4
Economically disadvantaged	292,516	287,774	98.4
Not economically disadvantaged	176,731	172,529	97.6
American Indian or Alaska Native	2,408	2,322	96.4
Asian	42,815	42,372	99.0
Native Hawaiian or Other Pacific Islander	2,229	2,168	97.3
Filipino	9,160	9,074	99.1
Hispanic or Latino	257,974	254,411	98.6
Black or African American	25,431	24,680	97.1
White	107,588	104,305	97.0
Two or more races	18,232	17,795	97.6
Special education services	57,800	55,265	95.6
No special education services	411,447	405,038	98.4
Migrant	4,017	3,959	98.6
Nonmigrant	465,230	456,344	98.1

Table 4.A.2 CAST Field Test Participation Rates for Grade Eight by Student Group

Group	Number of Eligible Students	Number of Participants	Percent of Participation
All students	472,094	458,523	97.1
Male	241,502	234,304	97.0
Female	230,592	224,219	97.2
English learner	55,950	53,957	96.4
English only	258,868	249,617	96.4
Reclassified fluent English proficient	134,622	132,886	98.7
Initial fluent English proficient	21,738	21,398	98.4
To be determined	306	232	75.8
English proficiency unknown	610	433	71.0
Economically disadvantaged	284,018	276,490	97.4
Not economically disadvantaged	188,076	182,033	96.8
American Indian or Alaska Native	2,558	2,383	93.2
Asian	44,694	44,140	98.8
Native Hawaiian or Other Pacific Islander	2,364	2,297	97.2
Filipino	11,464	11,320	98.7
Hispanic or Latino	253,673	247,625	97.6
Black or African American	25,938	24,818	95.7
White	112,365	107,783	95.9
Two or more races	15,992	15,379	96.2
Special education services	52,786	49,459	93.7
No special education services	419,308	409,064	97.6
Migrant	3,722	3,651	98.1
Nonmigrant	468,372	454,872	97.1

Table 4.A.3 CAST Field Test Participation Rates for Grade Ten by Student Group

Group	Number of Eligible Students	Number of Participants	Percent of Participation
All students	16,628	6,384	38.4
Male	8,523	3,400	39.9
Female	8,105	2,984	36.8
English learner	1,499	828	55.2
English only	9,308	3,602	38.7
Reclassified fluent English proficient	5,370	1,790	33.3
Initial fluent English proficient	444	160	36.0
To be determined	5	2	40.0
English proficiency unknown	2	2	100.0
Economically disadvantaged	11,835	4,548	38.4
Not economically disadvantaged	4,793	1,836	38.3
American Indian or Alaska Native	155	72	46.5
Asian	508	240	47.2
Native Hawaiian or Other Pacific Islander	90	65	72.2
Filipino	358	222	62.0
Hispanic or Latino	10,353	3,553	34.3
Black or African American	1,132	517	45.7
White	3,595	1,478	41.1
Two or more races	322	190	59.0
Special education services	1,989	936	47.1
No special education services	14,639	5,448	37.2
Migrant	303	62	20.5
Nonmigrant	16,325	6,322	38.7

Table 4.A.4 CAST Field Test Participation Rates for Grade Eleven by Student Group

Group	Number of Eligible Students	Number of Participants	Percent of Participation
All students	151,135	136,675	90.4
Male	76,518	69,045	90.2
Female	74,617	67,630	90.6
English learner	13,834	11,857	85.7
English only	75,405	68,091	90.3
Reclassified fluent English proficient	53,336	48,933	91.7
Initial fluent English proficient	8,397	7,682	91.5
To be determined	74	52	70.3
English proficiency unknown	89	60	67.4
Economically disadvantaged	92,170	82,531	89.5
Not economically disadvantaged	58,965	54,144	91.8
American Indian or Alaska Native	827	718	86.8
Asian	13,236	12,460	94.1
Native Hawaiian or Other Pacific Islander	632	581	91.9
Filipino	4,035	3,830	94.9
Hispanic or Latino	86,349	77,655	89.9
Black or African American	7,572	6,608	87.3
White	33,649	30,506	90.7
Two or more races	3,569	3,242	90.8
Special education services	14,497	12,343	85.1
No special education services	136,638	124,332	91.0
Migrant	879	716	81.5
Nonmigrant	150,256	135,959	90.5

Table 4.A.5 CAST Field Test Participation Rates for Grade Twelve by Student Group

Group	Number of Eligible Students	Number of Participants	Percent of Participation
All students	466,600	405,051	86.8
Male	237,796	205,248	86.3
Female	228,804	199,803	87.3
English learner	39,216	30,119	76.8
English only	249,629	214,986	86.1
Reclassified fluent English proficient	142,398	128,391	90.2
Initial fluent English proficient	34,373	31,184	90.7
To be determined	151	91	60.3
English proficiency unknown	833	280	33.6
Economically disadvantaged	266,638	229,473	86.1
Not economically disadvantaged	199,962	175,578	87.8
American Indian or Alaska Native	2,656	2,120	79.8
Asian	45,868	42,071	91.7
Native Hawaiian or Other Pacific Islander	2,425	2,044	84.3
Filipino	13,544	12,540	92.6
Hispanic or Latino	244,029	211,389	86.6
Black or African American	27,687	22,222	80.3
White	112,795	98,078	87.0
Two or more races	13,447	11,424	85.0
Special education services	46,310	35,049	75.7
No special education services	420,290	370,002	88.0
Migrant	2,932	2,675	91.2
Nonmigrant	463,668	402,376	86.8

Table 4.A.6 CAST Field Test Participation Rates for High School (All Grades Tested) by Student Group

Group	Number of Eligible Students	Number of Participants	Percent of Participation
All students	634,363	548,110	86.4
Male	322,837	277,693	86.0
Female	311,526	270,417	86.8
English learner	54,549	42,804	78.5
English only	334,342	286,679	85.7
Reclassified fluent English proficient	201,104	179,114	89.1
Initial fluent English proficient	43,214	39,026	90.3
To be determined	230	145	63.0
English proficiency unknown	924	342	37.0
Economically disadvantaged	370,643	316,552	85.4
Not economically disadvantaged	263,720	231,558	87.8
American Indian or Alaska Native	3,638	2,910	80.0
Asian	59,612	54,771	91.9
Native Hawaiian or Other Pacific Islander	3,147	2,690	85.5
Filipino	17,937	16,592	92.5
Hispanic or Latino	340,731	292,597	85.9
Black or African American	36,391	29,347	80.6
White	150,039	130,062	86.7
Two or more races	17,338	14,856	85.7
Special education services	62,796	48,328	77.0
No special education services	571,567	499,782	87.4
Migrant	4,114	3,453	83.9
Nonmigrant	630,249	544,657	86.4

Appendix 4.B: Demographic Summary

Notes:

- This set of tables are presented separately for grades five, eight, and high school, which shows categories for grades ten, eleven, and twelve.
- All students in California enrolled in grades five, eight, and twelve who were eligible for the general science assessments were required to participate in the field test.
- Students in grades ten and eleven who completed or were in the process of completing their last high school science course and who were not eligible for CAA for Science participated in the field test, per their LEA’s discretion.
- The percentages of student groups may not sum to 100 due to rounding. In addition, the percentages of students within a race and ethnic category may not sum to 100 due to missing racial or ethnic data for some students.
- For grades ten, eleven, and twelve, population percent is calculated based on all students in that grade level, regardless of whether or not they have taken the CAST.
- Because some students from the high school registration file were not included in the California Longitudinal Pupil Achievement Data System, some percentages do not sum to 100.

Table 4.B.1 Demographic Summary for Grade Five

Group	Number of Valid Scores	Percent of Valid Scores
All students	460,303	100.0
Male	235,428	51.1
Female	224,875	48.9
English learner	89,288	19.4
English only	263,488	57.2
Reclassified fluent English proficient	89,131	19.4
Initial fluent English proficient	17,673	3.8
To be determined	242	0.1
English proficiency unknown	481	0.1
Economically disadvantaged	287,774	62.5
Not economically disadvantaged	172,529	37.5
American Indian or Alaska Native	2,322	0.5
Asian	42,372	9.2
Native Hawaiian or Other Pacific Islander	2,168	0.5
Filipino	9,074	2.0
Hispanic or Latino	254,411	55.3
Black or African American	24,680	5.4
White	104,305	22.7
Two or more races	17,795	3.9
Special education services	55,265	12.0
No special education services	405,038	88.0
Migrant	3,959	0.9
Nonmigrant	456,344	99.1

Table 4.B.2 Demographic Summary for Grade Eight

Group	Number of Valid Scores	Percent of Valid Scores
All students	458,523	100.0
Male	234,304	51.1
Female	224,219	48.9
English learner	53,957	11.8
English only	249,617	54.4
Reclassified fluent English proficient	132,886	29.0
Initial fluent English proficient	21,398	4.7
To be determined	232	0.1
English proficiency unknown	433	0.1
Economically disadvantaged	276,490	60.3
Not economically disadvantaged	182,033	39.7
American Indian or Alaska Native	2,383	0.5
Asian	44,140	9.6
Native Hawaiian or Other Pacific Islander	2,297	0.5
Filipino	11,320	2.5
Hispanic or Latino	247,625	54.0
Black or African American	24,818	5.4
White	107,783	23.5
Two or more races	15,379	3.4
Special education services	49,459	10.8
No special education services	409,064	89.2
Migrant	3,651	0.8
Nonmigrant	454,872	99.2

Table 4.B.3 Demographic Summary for Grade Ten

Group	Number of Valid Scores	Percent of Valid Scores	Population Percent
All students	6,384	100.0	100.0
Male	3,400	53.3	51.0
Female	2,984	46.7	48.5
English learner	828	13.0	9.2
English only	3,602	56.4	56.0
Reclassified fluent English proficient	1,790	28.0	31.7
Initial fluent English proficient	160	2.5	2.6
To be determined	2	0.0	0.1
English proficiency unknown	2	0.0	0.5
Economically disadvantaged	4,548	71.2	70.8
Not economically disadvantaged	1,836	28.8	28.7
American Indian or Alaska Native	72	1.1	0.9
Asian	240	3.8	3.0
Native Hawaiian or Other Pacific Islander	65	1.0	0.5
Filipino	222	3.5	2.0
Hispanic or Latino	3,553	55.7	62.0
Black or African American	517	8.1	6.8
White	1,478	23.2	21.5
Two or more races	190	3.0	2.0
Special education services	936	14.7	11.9
No special education services	5,448	85.3	87.6
Migrant	62	1.0	1.9
Nonmigrant	6,322	99.0	97.7

Table 4.B.4 Demographic Summary for Grade Eleven

Group	Number of Valid Scores	Percent of Valid Scores	Population Percent
All students	136,675	100.0	100.0
Male	69,045	50.5	50.2
Female	67,630	49.5	49.1
English learner	11,857	8.7	9.3
English only	68,091	49.8	49.3
Reclassified fluent English proficient	48,933	35.8	35.1
Initial fluent English proficient	7,682	5.6	5.5
To be determined	52	0.0	0.1
English proficiency unknown	60	0.0	0.7
Economically disadvantaged	82,531	60.4	61.1
Not economically disadvantaged	54,144	39.6	38.2
American Indian or Alaska Native	718	0.5	0.5
Asian	12,460	9.1	8.6
Native Hawaiian or Other Pacific Islander	581	0.4	0.4
Filipino	3,830	2.8	2.7
Hispanic or Latino	77,655	56.8	57.0
Black or African American	6,608	4.8	5.0
White	30,506	22.3	22.0
Two or more races	3,242	2.4	2.3
Special education services	12,343	9.0	9.6
No special education services	124,332	91.0	89.7
Migrant	716	0.5	0.5
Nonmigrant	135,959	99.5	98.8

Table 4.B.5 Demographic Summary for Grade Twelve

Group	Number of Valid Scores	Percent of Valid Scores	Population Percent
All students	405,051	100.0	100.0
Male	205,248	50.7	50.1
Female	199,803	49.3	47.6
English learner	30,119	7.4	8.8
English only	214,986	53.1	52.5
Reclassified fluent English proficient	128,391	31.7	29.1
Initial fluent English proficient	31,184	7.7	7.1
To be determined	91	0.0	0.0
English proficiency unknown	280	0.1	2.5
Economically disadvantaged	229,473	56.7	55.6
Not economically disadvantaged	175,578	43.3	42.1
American Indian or Alaska Native	2,120	0.5	0.5
Asian	42,071	10.4	9.7
Native Hawaiian or Other Pacific Islander	2,044	0.5	0.5
Filipino	12,540	3.1	2.9
Hispanic or Latino	211,389	52.2	50.7
Black or African American	22,222	5.5	5.8
White	98,078	24.2	23.8
Two or more races	11,424	2.8	2.8
Special education services	35,049	8.7	12.8
No special education services	370,002	91.3	84.9
Migrant	2,675	0.7	0.6
Nonmigrant	402,376	99.3	97.1

Table 4.B.6 Demographic Summary for High School (All Grades Tested)

Group	Number of Valid Scores	Percent of Valid Scores	Population Percent
All students	548,110	100.0	100.0
Male	277,693	50.7	50.2
Female	270,417	49.3	47.9
English learner	42,804	7.8	8.9
English only	286,679	52.3	51.9
Reclassified fluent English proficient	179,114	32.7	30.6
Initial fluent English proficient	39,026	7.1	6.6
To be determined	145	0.0	0.0
English proficiency unknown	342	0.1	2.0
Economically disadvantaged	316,552	57.8	57.2
Not economically disadvantaged	231,558	42.2	40.9
American Indian or Alaska Native	2,910	0.5	0.6
Asian	54,771	10.0	9.3
Native Hawaiian or Other Pacific Islander	2,690	0.5	0.5
Filipino	16,592	3.0	2.8
Hispanic or Latino	292,597	53.4	52.5
Black or African American	29,347	5.4	5.7
White	130,062	23.7	23.4
Two or more races	14,856	2.7	2.7
Special education services	48,328	8.8	12.1
No special education services	499,782	91.2	86.1
Migrant	3,453	0.6	0.6
Nonmigrant	544,657	99.4	97.5

Chapter 5: Scoring

This chapter summarizes the types of scoring approaches that were used for each type of item in the California Science Test (CAST) field test forms, including machine scoring, human scoring, and the process for building artificial intelligence (AI) scoring models. CAST field test assessments included traditional multiple-choice (MC) items, technology-enhanced items (TEIs), and constructed-response (CR) items. The traditional MC items and the TEIs were machine-scored, while the CR items were human-scored. AI scoring models were built for future AI scoring.

5.1. Human Scoring for Constructed-Response Items

5.1.1. Sampling Process

The CAST field test at each tested grade level included selected-response (SR) and CR items. The SR items including the MC and TEIs are machine-scorable; thus, all students' SR items in the field test were scored. Not all responses to the CR items were scored in the field test. Instead, a random sample of responses was drawn for each CR item to be scored.

The CAST field test is a census administration, with all eligible students in each tested grade taking the field test. Sufficient samples of student responses for item analyses, key psychometric studies (including item calibration and testing the dimensionality), and construction of AI scoring models for potential use in future operational administrations can be obtained without having to score every tested student's response to every administered CR item.

A two-batch sampling design for each tested grade was used to maximize the utility of the scored student responses by providing the necessary samples for each of the intended uses of the CR scores while minimizing the number of responses that needs to be scored.

The first batch involved drawing a separate sample of students for each block of items administered in the field test and scoring all CRs for the selected students for a given block. For instance, if block A1 had two CRs, then the selected students for this block would have both CRs human scored. The second batch involved sampling students who were administered particular Segment A-B forms (i.e., unique combinations of two A blocks and two B blocks, such as A1, A2, B1, B2).

Both batches provided data for item calibration, the screener study (refer to subsection *6.8.2 Content Screen-Out Study*), and for building AI scoring models (refer to subsection *5.3 Artificial Intelligence (AI) Scoring Model Building*). However, due to scheduling constraints, only the first batch was used for item analysis. The second batch was used to support the dimensionality study (refer to subsection *6.4 Test Dimensionality Analyses*).

Sampling for batch one was conducted first to meet the data review schedule. The second batch sample was drawn two weeks later to allow for more students to have taken each A-B form of interest: Given that there were 168 unique combinations of A-B forms, it took longer to meet the desired sample sizes for the second batch sample. The sampling process for each batch is described in the next subsections.

5.1.1.1 Batch-One Sampling

The batch-one sample involved first creating the sampling frame and then drawing the samples at each tested grade level for each block. The sampling frame was established when the available set of tested students roughly matched the overall testing population by demographics (within ± 5 percent of the population compositions found in the California Longitudinal Pupil Achievement Data System [CALPADS][7]) to ensure that the sampling frames were sufficiently representative of the population. In a few instances, the available tested students' demographic composition differed by more than ± 5 percentage points from the population composition, but these instances were expected due to the assignment rules of the blocks. For instance, of the eight discrete-item C blocks, only C1 in each grade was assigned to students who needed certain accommodations; thus it was expected that the percentage of students in Special Education programs receiving C1 would be higher than the general population, and the percentage of Special Education students would be lower for those students receiving blocks C2 to C8.

5.1.1.1.1. Exclusion Rules

Before drawing students for the batch-one samples, certain exclusion rules were applied to remove cases that were more likely to confound, rather than inform, the results. For each block within each grade level, students who were classified as unmotivated were removed. Such students were identified by the following two rules:

1. Students who only responded to less than 25 percent of their administered test items in the block of interest are considered unmotivated. A nonresponse is considered any item coded as "Omit" or "Not Seen."
2. Students who completed their assigned items in less than the minimum testing time for the block are considered unmotivated. This is determined by the first percentile of average MC item time from the CAST pilot test in 2016–17 for each grade level, to provide conservative estimates of the minimum time a student who is not motivated would take to complete an item. These minimum times per item are 11 seconds for grade five, 10 seconds for grade eight, and 5 seconds for high school.

In addition, students who had left CR responses blank (i.e., items coded as Omit or Not Seen) within a block were removed from consideration, as they would not provide any useful data for AI model building and only a minimal number of responses were being scored. Generally, less than one percent of students were removed for these reasons for the sampling frames for each block within each grade.

5.1.1.1.2. Selection of the Random Sample

For each of the blocks in grades five and eight with at least one CR item, 4,000 students were randomly sampled. For high school blocks, the original plan was to obtain 2,000 students per high school grade level (ten, eleven, and twelve) per block, resulting in a total of 6,000 students per block (versus only 4,000 per block in grades five and eight). However, due to the small number of grade ten students participating in the field test, not all blocks contained 2,000 students per block. For those items, all students from grade ten were included and no random sample was drawn. As a result, not all CR items in the high school assessment have 6,000 responses scored.

[7] For grades five, eight, and twelve, all eligible students were assigned to take the CAST. However, only a subset of local educational agencies (LEAs) tested students in grades ten and eleven. CALPADS includes data on all students in a grade, not just those registered to take the CAST. Thus, to obtain the population-level statistics for grades ten and eleven, the test registration data in the Test Operations Management System was used to subset the CALPADS data to only those students who had registered for the CAST.

The demographic compositions for the selected samples were then checked against the population demographics. If the difference is within +5/-5 range, the sample was accepted. However, because of the need to balance competing priorities to include students from Early Adopter LEAs and have roughly representative samples of the population, as well as to account for the unbalanced sample for the blocks with the accessibility features, this requirement was not always satisfied. In such cases, some modifications were made as described in the next subsection.

5.1.1.1.3. Sample Priorities

There were two priorities in drawing the batch-one samples: (1) to include students from Early Adopter LEAs; and (2) to have roughly representative samples of the population.

The sampling procedure prioritized selecting students from LEAs that were known early adopters of the California Next Generation Science Standards (CA NGSS). Students from such LEAs were more likely to have been exposed to instruction aligned with the CA NGSS, an important consideration for valid score results.

It was also important to have a representative sample of the population so that subsequent results using the CR items can be generalized to all students (e.g., AI models were appropriate for all student groups). However, these two priorities were in conflict because the students in the Early Adopter LEAs were demographically distinct from the general population; they tended to have higher proportions of economically disadvantaged students and students in minority racial or ethnic groups. Accordingly, if the samples were only composed of students from Early Adopter LEAs, then they would not be representative of the full population.

To balance these two goals, when possible, half of each sample was drawn from students in the Early Adopter LEAs and the other half from all other LEAs. In some cases, less than half of the required sample size were students in Early Adopter LEAs. In those cases, all of the Early Adopter LEA students were selected and the rest of the required sample size, which would be more than half, was drawn from the non–Early Adopter LEAs.

In addition, as previously mentioned, only certain blocks—A1, A2, B1, B2, B3, and C1—had designated supports and accommodations available, including text-to-speech, American Sign Language, stacked Spanish translations, and translation glossaries. Accordingly, certain student groups that were more likely to need these supports or accommodations were overassigned to these blocks and underassigned to other blocks, resulting in deviations from the population demographic composition. The most affected student group was students in Special Education programs. Thus, for all batch-one samples, stratified samples by Special Education status were drawn to ensure the samples were representative of the population. In some cases, samples were further stratified by English learners (versus students who are not English learners), as this group was also affected, but to a lesser degree, to these systematic assignments of blocks.

5.1.1.2 Batch-Two Sampling

The primary purpose of the batch-two sampling was to produce datasets for use in the dimensionality study, which would benefit from having completely scored data for all students—both SR and CRs scored. The dimensionality study focused on examining the structure of the assessment for student scores, which are only based on segments A and B. Thus, batch-two sampling involved sampling by prespecified A-B forms—no C blocks were considered for batch-two sampling.

Three forms were chosen at each grade to allow for the dimensionality study to be replicated three times within a grade level. They are summarized in Table 5.1.

Table 5.1 Forms by Grade Level for Batch-Two Sampling

Form	Grade 5	Grade 8	High School
1	A2A3B4B7	A3A4B2B6	A1A3B4B5
2	A1A4B6B7	A2A4B3B5	A3A4B6B7
3	A1A2B2B3	A1A2B1B2	A1A2B1B2

Form 3 in each grade was a fully accessible form in that it contained all designated supports and accommodations. The total count of students administered this form was higher than the other forms as it was one of only a few forms to which certain student groups who needed the designated supports and accommodations were routed.

5.1.1.2.1. Exclusion Rules

Just as with the batch-one sample, the sampling frames were first created by filtering out students for each form who were unmotivated or had any blank CR responses within the form of interest. Unmotivated students were identified similarly as for batch one—those students who responded to less than 25 percent of the administered items with the form of interest or completed their assigned items in less than the minimum testing time for their form. These exclusion rules filtered out 0 to 0.30 percent of students per grade per form.

5.1.1.2.2. Selection of the Random Sample

For grades five and eight, 2,000 students were drawn for each of the three forms. For high school, the intent was to draw 2,000 students per tested high school grade level: grades ten, eleven, and twelve. However, there were fewer than 100 grade ten students administered forms 1 and 2 each and fewer than 200 grade ten students administered form 3, given the low volumes overall for grade ten during the field test. Such low numbers are insufficient to support the dimensionality analysis and thus, grade ten was excluded from batch-two sampling.

For grade eleven, there were fewer than 2,000 students available for Forms 1 and 2. Consequently, all students who were administered these forms were included. Table 5.2 provides the sample size for each of the forms included in batch-two sampling. As done in batch-one sampling, stratified sampling was used for forms that couldn't reach a reasonable representative sample of the population.

Table 5.2 Sample Size for Forms in Batch-Two Sampling

Grade	Form	Blocks	Sample Size
Grade 5	1	A2-A3-B4-B7	2,000
Grade 5	2	A1-A4-B6-B7	1,946
Grade 5	3	A1-A2-B2-B3	1,987
Grade 8	1	A3-A4-B2-B6	2,000
Grade 8	2	A2-A4-B3-B5	2,000
Grade 8	3	A1-A2-B1-B2	1,983
High School	1	A1-A3-B4-B5	1,338 (Grade 11) + 2,000 (Grade 12)
High School	2	A3-A4-B6-B7	1,376 (Grade 11) + 2,000 (Grade 12)
High School	3	A1-A2-B1-B2	2,000 (Grade 11) + 2,000 (Grade 12)

5.1.2. Scoring Rubric Development

Educational Testing Service's (ETS') Assessment Development (AD) group developed 24 performance tasks—eight per grade span—to be included in the 2017–18 CAST field test administration for measuring more complex skills. During item development, draft scoring metrics (rubrics) were created with the point scale and descriptions. ETS included these rubrics with the associated items in the internal and external review processes described in subsection *3.4 Item-Review Process*. Rubrics were edited as needed on the basis of feedback from the CDE and California teachers during the item review and range-finding processes. Exemplar responses of each score point were provided for scoring guidance as benchmarks.

5.1.3. Range Finding

Soon after receiving a large volume of CR responses from California schools, ETS began the range-finding process by randomly selecting a wide variety of student response samples. The goal was to ensure sufficient responses at each score point on the rubric to create sets of responses for training and certifying (qualifying) raters (scorers) and for monitoring raters during the scoring process. Another part of the range-finding process included annotating responses to provide further guidance on why a response received a certain rating. The following steps describe how the range finding process was implemented.

1. ETS AD staff used the rubric (scoring guide) for each item to randomly select and score responses to represent each score point on an item's rubric. The number of responses selected varied by prompt and was based on the number of points and the prompts that were preselected for certifying and training raters. Scored samples were needed for the various purposes summarized in Table 5.3.

Table 5.3 CAST Field Test Sample Selection for Human Scoring Procedures

Sample Type	Purpose	Number of Sets and Samples in Sets	Configuration of Sets
Certification	Certification samples for verifying scoring accuracy of potential raters and Scoring Leaders	• Two sets of 10 samples per set for one high school 2-point prompt • Mixed score points	Three to five samples for each score point per set

Sample Type	Purpose	Number of Sets and Samples in Sets	Configuration of Sets
Training	Training samples with annotations for rater training and scoring practice	• Two sets of seven samples per grade • One prompt for the high school training set; mixed score points • For grades five and eight, one training set for a composite item and one training set for a noncomposite item; mixed score points	Two to three samples for each score point per set
Benchmarks	Benchmark samples with annotations that represent exemplar responses at each score point on the rubric	One set of 4 to 12 samples per unique prompt per grade (60 unique prompts total)	Two to three samples for each score point
Calibration	Calibration samples for evaluating rater scoring performance on specific prompts	• Two sets of five samples per set for one prompt per grade • Mixed score points	One to three samples for each score point per set
Validity	Validity samples inserted into rater's scoring queue to monitor the quality of scoring	One set of 20 samples per prompt; mixed score points	Six to 12 samples for each score point

2. Responses were scored by two independent, experienced raters using the Online Network for Evaluation (ONE) system. ETS AD staff also wrote annotations, or short notes, with each score point to explain why a response earned a particular rating. Annotations helped raters make explicit connections between the scoring guide and responses, and thus informed their careful and accurate scoring of responses. ETS provided the CDE with the independent ratings, scored samples, annotations, and recommendations for which responses would go in the different scoring materials (i.e., certification, benchmark, training, calibration, and validity, as summarized in Table 5.3).

3. CDE and ETS content experts reviewed the samples, scores, and rationale for all set designations to agree upon the scores and samples to use for specific sets. The annotations for the samples also were reviewed and refined as needed.

4. ETS obtained feedback on the rubrics, benchmarks, and training samples from a total of seven teachers. The teachers were recruited from the existing California Assessment of Student Performance and Progress (CAASPP) rater pool based on their background in teaching science and experience with CR scoring. ETS compiled written and verbal feedback from the teachers and provided it to the CDE.
5. The CDE reviewed the teacher feedback and made final decisions about prompts, rubrics, and scoring materials.
6. ETS created all final sample sets in the ONE system and used these samples as part of a system of training and controls for verifying the quality and consistency of pilot scoring.

5.1.4. Rater Recruitment and Certification Processes

Several weeks prior to the start of CR scoring, ETS recruited a pool of eligible CAST raters from invited California science teachers as well as from the current CAASPP Smarter Balanced pool of eligible raters from California. All CAST raters were required to have a bachelor’s degree to be eligible to attempt certification. The scoring pool consisted of California educators; the remaining pool of raters represented a variety of backgrounds in business, education, and other fields. Approximately 500 raters were used for the 2017–18 CAST field test, scoring 384,867 responses across the three grade levels.

Certification served as an initial screening to ensure that ETS’ CR Scoring Systems and Capabilities (CRSC) team had a sufficient number of qualified raters in place to meet the demands of scoring. One 2-point prompt (e.g., a response that can earn 0, 1, or 2 points) selected from among the high school prompts was used for certification. Training samples were provided for the rater to review and practice rating before attempting certification. If a rater passed certification on the high school prompt, he or she was eligible to calibrate on the grade-specific prompts once scoring began.

Raters were required to achieve an 80 percent exact match to the CDE-approved rating for the responses on at least one of the certification sets to be eligible for calibration on a specific grade-level test prompt. If raters did not pass either certification set, they were excused from scoring 2017–18 CAST field-test items.

5.1.5. Rater and Scoring Leader Training

ETS selected scoring leaders to oversee a group of raters during the scoring process. Scoring leaders are experienced raters who have demonstrated high scoring accuracy from previous scoring projects at ETS and are invited to act as a scoring leader on a project. For the 2017–18 CAST field-test administration, the scoring leader backread (read behind), guided, and retrained raters as needed. Scoring leaders monitored the small group of raters on a shift, usually up to 10 raters, to assist CRSC with scoring quality.

5.1.5.1 Training for Scoring Leaders

ETS assessment specialists conducted virtual training sessions for scoring leaders by means of conference calls using online conferencing tools. The purpose of the training was to discuss the duties of scoring leaders and to provide specific grade-level guidance on particular prompts. The training included guidance on using condition codes that are applied to nonscorable responses such as blank (B), insufficient (I), or those in a language other than English (L); communication with raters; how to monitor raters; and other information necessary for their role during scoring.

5.1.5.2 Training for Raters

Training for raters occurred within the ONE system. Raters were provided ONE system training documents as well as program-specific information that they could refer to at any time. Prior to attempting calibration, raters were given a window of time to review all training materials in the system and practice scoring using the prescored training sets. After raters completed a training set, they were provided with annotations for each response as a rationale for the rating assigned.

The scoring training provided for each potential rater was designed using CDE-approved materials developed by ETS and followed the three-step progression noted.

5.1.5.2.1. Step One: Review the scoring guide and benchmarks.

Training for scoring began with an overview of the scoring guide, or rubric, and benchmarks. In the ONE system, the rubric was accessed through a tab called [**Scoring Guide**]. The benchmarks, also called anchors, were accessed in ONE through the [**Benchmarks**] tab. The benchmarks had annotations associated with them to call the rater's attention to specific content in the sample responses.

5.1.5.2.2. Step Two: Score training sets.

After orientation to the scoring guide and the benchmark function, raters progressed through an online content training in the ONE system in which they reviewed several sets of sample responses, assigned scores, and received feedback on their scores based on the CDE-approved rating for each response and applicable supporting annotation. Training sets, also called feedback sets, are samples of responses that provided the rater annotations after each sample was completed. The feedback sets for the 2017–18 CAST field-test administration contained a mixed set of sample responses for each score point on the rubric as well as feedback in the form of annotations after a rater submitted a score. When raters completed the feedback sets, they could attempt calibration.

5.1.5.2.3. Step Three: Set calibration.

Calibration is a system-supported control to ensure raters meet a specified standard of accuracy when scoring a series of prescored responses. Raters calibrated before they were allowed to score, meaning they scored a certain percentage of responses accurately from a set of responses called a calibration set. The passing percentage was determined by the program and number of responses in a set.

In general, calibration can be put in place at the beginning of a four- or eight-hour scoring shift prior to starting a new grade or new prompt or at specified intervals during a scoring window. Typically, raters are allowed two chances to calibrate successfully. If raters meet the standard on the first attempt, they proceed directly to scoring responses. If raters are unsuccessful, they may review training sets and attempt to calibrate again with a new calibration set. If they are unsuccessful after both attempts, they are dismissed from that scoring shift.

Calibration can be used as a means to control rater and group drift, which are changes in behavior that affect scoring accuracy between test administrations. Calibration can be used throughout a scoring season (e.g., January through July) to check scoring accuracy on a prescored set of responses. In the case of the 2017–18 CAST field test, calibration was set at once per grade during a seven-day period. In comparison, because the scoring window for the 2016–17 pilot was less than a week, calibration was conducted prior to the start of scoring.

For the 2017–18 CAST field-test administration, raters were permitted to score any prompt for a grade if they passed calibration on their first prompt with a 90 percent exact match for items that are scored 0 or 1 point or an 80 percent match for items that are scored 0, 1, or 2 points.

5.1.5.3 Scoring Rules and Processes

ETS implemented the following scoring rules and processes for CAST pilot scoring:

- ETS psychometric staff provided a sampling plan that includes the responses selected to be scored. Refer to subsection *5.1.1 Sampling Process* for the sampling plan.
- The sampling plan was uploaded to ONE to activate the responses for scoring.
- Approximately 15 percent of responses were double-scored to facilitate the building of AI scoring models. Raters were not aware when a second scoring was occurring and so did not have access to the first score.
- Raters did not have access to condition codes and were instructed during training to defer any nonscorable responses to their scoring leader for scoring. The condition codes were:
  - Blank (B): The response area was completely blank.
  - Insufficient (I): The response had no meaningful response or even a guess at a possible answer (e.g., random keystrokes, opinions of the test).
  - Nonscorable Language (L): The language of the response was not English.
- Scoring leaders were trained to apply condition codes to nonscorable responses.
- Raters were instructed to apply zero (0) scores when there was an attempt to answer the question but the information was incorrect so could not earn the minimum score. If the rater was unsure, he or she deferred responses to the scoring leader.

5.1.6. Scoring Monitoring and Quality Management

In addition to the calibration function described previously, raters were monitored closely for the quality of their scoring throughout the scoring window. During a scoring shift, scoring leaders read behind raters at a rate of 10 percent or more of the responses scored by each individual rater to determine if raters were applying the scoring guide and benchmarks accurately and consistently. When necessary, the scoring leader redirected the rater by referencing the rubric, benchmarks, or both the rubric and benchmarks to explain why a response should have received a different score. When a rater was scoring inconsistently, the backreading proportion may be more than 10 percent.

Prescored responses from validity sets were also inserted into the rater's queue for every 10 responses scored. These were inserted in random positions and not fixed so a rater was unaware which response was a validity response. The ETS CR Performance Measures and Analytics group, in conjunction with AD, reviewed the statistics on the validity responses daily to determine if raters needed retraining.

The ONE system offers a comprehensive set of tools that the scoring leaders and scoring management staff used to monitor the progress and accuracy of individual raters and raters in aggregate. Reports produced to show rater productivity and performance presented how many responses a rater scored during a shift and how two raters scored the same response (i.e., interrater reliability).

5.1.7. Interrater Reliability

The ONE system captures interrater reliability by monitoring data for responses that are double-scored. For the CAST field test, 13 percent of grade five, 14 percent of grade eight, and 17 percent of high school responses were double-scored for studies to be made for possible AI scoring. The interrater reliability reports included the number and percent of exact matches for each rater and the number and percent of adjacent and discrepant scores. Scoring management reviewed the interrater reliability statistics for each prompt to determine if there were any issues that needed to be addressed during scoring. The interrater reliability statistics are included in subsection *6.7.5 Interrater Reliability Analyses*.

5.1.8. Validity Responses and Sets

High interrater reliability is an important goal, and the analysis of related data helps to identify errant scoring. However, validity responses and sets are the most important tools in ensuring scoring accuracy.

Unlike interrater data, which show a comparison of one rater versus another, validity data indicate the rater’s ongoing ability to match CDE-approved scores when scoring prescored validity responses that are indistinguishable from live responses.

ETS utilized sample responses approved during the range-finding process to create an initial set of 20 validity responses per prompt to represent all points across the score scale. ETS estimated 20 validity responses per grade and prompt would be sufficient for the scoring window.

Review of incorrectly scored validity responses was an ongoing process that alerted scoring leaders to specific needs for monitoring and retraining. Routine procedures included focused backreading that could lead to one-on-one retraining sessions between scoring leaders and individual raters. Additionally, scoring leaders and ETS AD staff worked together to identify any trends in errant scoring patterns to determine if a broader retraining effort would be beneficial, such as the creation of an additional training set to reanchor, or refocus, the group in the accurate application of a particular aspect of the scoring guide.

ETS AD and CRSC staff reviewed raters’ scoring patterns and make judgment calls on whether to dismiss a rater. Raters who were unable to maintain an adequate standard of accuracy after retraining were disqualified from scoring the item. When a rater was dismissed, ETS scoring leadership reviewed the rater’s scoring patterns to determine if all scores assigned by the rater during the time period in question should be nullified and the responses routed for rescoring.

Features such as backreading, interrater reliability reporting functions, and validity response insertion and reporting functions allowed scoring leaders to quickly identify inaccurate scoring patterns and take appropriate corrective actions.

5.2. Machine Scoring for Selected Response Items

CAST field test assessments included machine-scorable, traditional MC items and TE items that were scored by the test delivery system (TDS). In the TDS, responses to the test forms were compared with the answer keys or scoring rubrics embedded in the TDS to determine the score points. A real-time, quality-monitoring component was built into the TDS. After a test was administered to a student, the TDS passed the resulting data to the Quality Assurance system to ensure a score from the machine-scoring system was scored accurately. The details of the quality control are provided in subsection 8.5 Quality Control of Scoring and Reporting.

5.3. Artificial Intelligence (AI) Scoring Model Building

5.3.1. Data Collection

After the field test, ETS collected a sample of students' responses to 57 CR items with human score(s) assigned, as described in subsection 5.1.1 Sampling Process. The number of responses to be double-scored was set at 800 for all items. For items in grades five and eight, the percentage of double scoring is 20 percent. For items on the high school assessment, the total number of responses to be scored varied across items, from 4,000–6,000; and the percent of double scoring varied from 13.1 percent to 18.4 percent. ETS also collected student demographic information such as gender, ethnicity, and economic status to use for student group analysis during the model-evaluation stage.

5.3.2. Model Training

At ETS, the steps to build AI scoring models for scoring text-based responses involved the automatic extraction and modeling of linguistic features. Natural language processing techniques were used to extract construct-relevant linguistic features from a set of human scored responses. Using the linguistic features extracted from the data, statistical models were built to predict the scores that human raters would assign to that response. Statistical modeling methods included, for example, multiple linear regression and support vector machines.[8] Each model was built using 10-fold cross-validation method, which randomly splits the entire dataset for an item in 10 subsets, and nine instances of the data are used to train the model while the tenth instance is used to test the predictive ability of the model. The subsets are rotated so that the final model for each item uses the entire dataset for training and testing.

Each model then went through an evaluation stage with multiple statistical criteria, such as Pearson's *r* and quadratic-weighted kappa, using the predictions from each testing instance. The evaluation was performed at the overall data set level as well as student group–level and reported in the next subsection.

Figure 5.1 provides a cycle chart illustrating the primary steps in the model-building and evaluation processes. First, three human-scored responses with scores of 1, 1, and 3 are funneled to natural language processing tools to extract linguistic features. An arrow points to the next step, statistical modeling. Here, the model-building process ends. The resulting model from the previous steps is sent to model evaluation.

[8] A Support Vector Machine (SVM) performs classification by finding the hyperplane that maximizes the margin between two classes. The vectors (cases) that define the hyperplane are the support vectors (Vapnick, 1995). The Support Vector Regression is an extension of SVMs and uses the same principles as the SVM for classification, with only a few minor differences (Drucker, Burgess, et al., 1996).

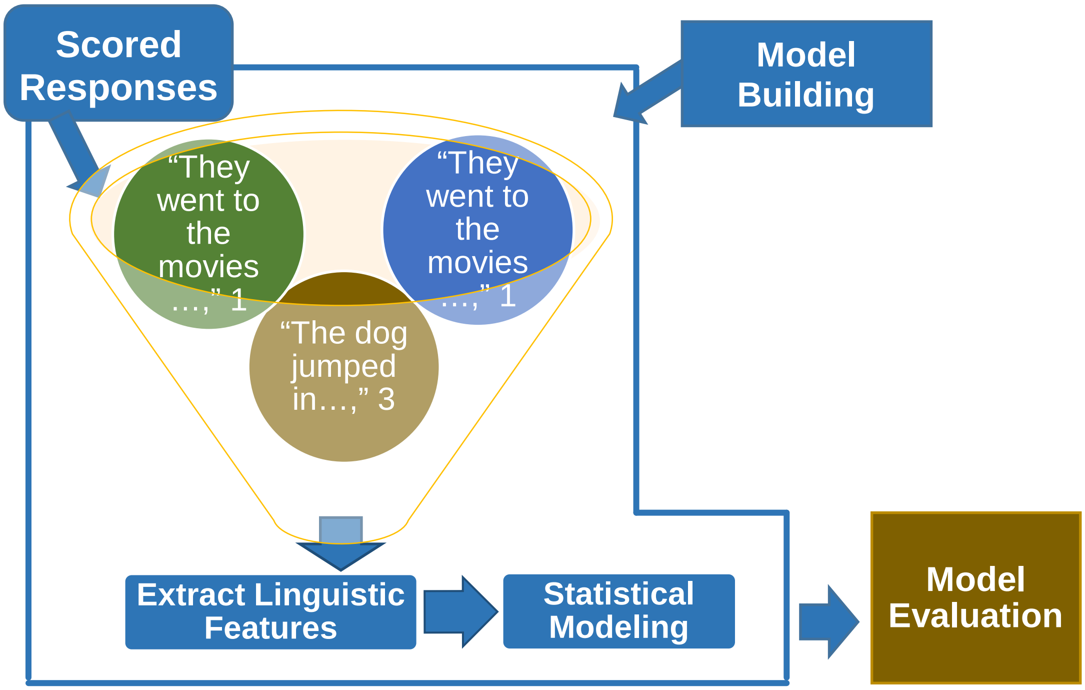

Figure 5.1 Model Building and Evaluation Process

5.3.3. Model Evaluation

One of the important factors in building AI scoring models with good performance was the use of data with reliable human scores. A commonly used indicator for evaluating human scoring reliability is to use more than one rater on a large enough sample of responses and evaluate the extent to which they agree with each other. The agreement rates between two human raters for the samples for the 57 field test CR items are shown in appendix 6.F.

The quality of the AI scores was evaluated by comparing the statistical metrics produced by the Human-Human and Human-AI scoring procedures. The statistical metrics used for the comparison of Human-Human and Human-AI were the mean, standard deviation (SD), Pearson Correlation *r*, quadratic-weighted kappa (QWK), and mean square error (MSE). Table 5.4 presents the statistical metrics for the Human-Human and Human-Machine scores. Additional information about Human-Human scoring can be found in appendix 6.F.

Table 5.4 Human–Human and Human–AI Rater Agreement

Grade Level	Prompt ID	AI Mean	AI SD	H1-H2 Pearson Correlation	H-AI Pearson Correlation	H-AI QWK	H1-H2 MSE	H-AI MSE
5	VH667949	0.48	0.50	0.76	0.75	0.75	0.12	0.13
5	VH668026	0.58	0.49	0.77	0.76	0.76	0.11	0.12
5	VH695572	0.27	0.45	0.51	0.62	0.61	0.22	0.16
5	VH709025	0.57	0.62	0.91	0.51	0.50	0.09	0.44
5	VH709052	0.29	0.45	0.76	0.79	0.79	0.10	0.09

Grade Level	Prompt ID	AI Mean	AI SD	H1-H2 Pearson Correlation	H-AI Pearson Correlation	H-AI QWK	H1-H2 MSE	H-AI MSE
5	VH731235	0.41	0.49	0.57	0.62	0.62	0.20	0.19
5	VH733167	0.57	0.53	0.59	0.63	0.63	0.27	0.23
5	VH737471	0.69	0.46	0.55	0.60	0.60	0.21	0.18
5	VH810103	0.18	0.38	0.71	0.78	0.78	0.09	0.07
5	VH810308	0.17	0.37	0.72	0.76	0.75	0.09	0.08
5	VH810523	0.22	0.42	0.69	0.68	0.68	0.12	0.12
5	VH810549	0.08	0.27	0.47	0.52	0.49	0.15	0.11
5	VH810950	0.11	0.32	0.72	0.76	0.76	0.05	0.05
5	VH811101	0.19	0.39	0.53	0.56	0.56	0.17	0.15
5	VH813229	1.15	0.87	0.91	0.87	0.86	0.15	0.21
8	VH695226	1.47	0.72	0.84	0.80	0.80	0.19	0.23
8	VH699333	0.41	0.49	0.59	0.57	0.57	0.20	0.21
8	VH702216	0.71	0.71	0.82	0.72	0.71	0.23	0.33
8	VH702611	1.27	0.84	0.95	0.90	0.89	0.08	0.16
8	VH728143	0.94	0.86	0.92	0.84	0.83	0.15	0.27
8	VH730085	0.49	0.65	0.73	0.72	0.71	0.31	0.29
8	VH734423	1.12	0.87	0.90	0.86	0.86	0.17	0.22
8	VH738505	0.66	0.48	0.80	0.82	0.82	0.09	0.08
8	VH738912	0.89	0.88	0.94	0.88	0.88	0.11	0.20
8	VH803445	0.60	0.64	0.85	0.82	0.82	0.15	0.17
8	VH803496	0.23	0.42	0.55	0.58	0.58	0.18	0.16
8	VH803535	0.64	0.79	0.87	0.87	0.87	0.17	0.17
8	VH803647	0.67	0.47	0.72	0.77	0.77	0.14	0.11
8	VH804554	0.93	0.56	0.70	0.67	0.66	0.26	0.26
8	VH805907	0.19	0.39	0.59	0.64	0.64	0.15	0.12
8	VH807320	0.57	0.53	0.71	0.68	0.68	0.21	0.21
8	VH809423	0.86	0.86	0.92	0.86	0.86	0.14	0.22
8	VH809632	0.24	0.43	0.59	0.53	0.53	0.17	0.19
8	VH814728	1.10	0.78	0.84	0.61	0.60	0.30	0.60
8	VH826960	0.51	0.66	0.86	0.77	0.77	0.15	0.23
8	VH810601	0.17	0.37	0.76	0.77	0.77	0.07	0.06
8	VH811932	0.19	0.42	0.66	0.60	0.58	0.18	0.18
8	VH811273	0.08	0.31	0.73	0.78	0.74	0.10	0.07
HS	VH651810	0.35	0.48	0.86	0.88	0.88	0.06	0.05
HS	VH651815	0.43	0.51	0.74	0.78	0.78	0.16	0.12
HS	VH696269	0.64	0.69	0.77	0.77	0.77	0.25	0.24

Grade Level	Prompt ID	AI Mean	AI SD	H1-H2 Pearson Correlation	H-AI Pearson Correlation	H-AI QWK	H1-H2 MSE	H-AI MSE
HS	VH702164	0.21	0.42	0.62	0.63	0.60	0.21	0.19
HS	VH730945	0.49	0.57	0.82	0.77	0.76	0.17	0.20
HS	VH804572	0.38	0.49	0.84	0.85	0.85	0.08	0.07
HS	VH804586	0.10	0.32	0.67	0.74	0.70	0.12	0.09
HS	VH804610	0.21	0.41	0.75	0.75	0.75	0.09	0.08
HS	VH805894	0.39	0.49	0.93	0.92	0.92	0.03	0.04
HS	VH807293	0.34	0.51	0.62	0.56	0.55	0.30	0.30
HS	VH807384	0.36	0.51	0.58	0.63	0.62	0.30	0.24
HS	VH808368	0.24	0.43	0.71	0.73	0.73	0.11	0.10
HS	VH807168	0.04	0.20	0.28	0.33	0.24	0.20	0.15
HS	VH807248	0.03	0.17	0.41	0.36	0.32	0.09	0.07
HS	VH805924	0.25	0.44	0.45	0.45	0.43	0.36	0.30
HS	VH736248	0.02	0.14	0.45	0.47	0.38	0.07	0.06
HS	VH808361	0.57	0.50	0.69	0.78	0.78	0.15	0.11
HS	VH710712	0.32	0.48	0.50	0.57	0.55	0.33	0.25
HS	VH807280	0.18	0.41	0.66	0.62	0.61	0.18	0.17

References

Drucker, Harris, Burges, Christopher J. C., Kaufman, Linda, Smola, Alexander J., & Vapnik, Vladimir N. (1997). Support vector regression machines. In *Advances in neural information processing systems 9*, NIPS 1996 (pp. 155–161). Cambridge, MA: MIT Press.

Vapnik, Vladimir N. (1995). *The nature of statistical learning theory.* New York, NY: Springer-Verlag.

Chapter 6: Analyses

This chapter summarizes the results of the item- and test-level analyses on samples from the 2017–18 California Science Test (CAST) field test administration. Analyses include the following:

- Classical item analyses
- Differential item functioning (DIF) analyses
- Test dimensionality analyses
- Item calibration
- Response time analyses
- Reliability
- Research studies

6.1. Samples Used for the Analyses

Two item analyses were run for the CAST field test: the preliminary item analyses (PIA) and the final item analyses (FIA).

PIA identifies potentially problematic items for further evaluation and is run as soon as a sufficient volume of data is collected, to obtain stable estimates. In CAST, the PIA was planned to run when the volume reached at least 2,000 responses per item. For the constructed-response (CR) items, only a sample of responses were scored (refer to subsection *5.1.1 Sampling Process* for details). The analysis sample for the CR items includes only the students who were selected to have their responses scored. The PIA for the CR items was run after the sampled responses were all scored. The sample size for the CR items is approximately 4,000 per item. For the machine scorable items, all students were subject for inclusion. The actual sample size used for PIA analyses was at least 5,000.

The FIA is conducted after the administration was completed. All student responses that met the inclusion rule are included in the analyses. The inclusion rules used in CAST field test item analyses and item calibration include the following:

- Students who logged on the test and answered at least one item were included in the item analysis and item calibration.
- At the item level, items with responses or scores labeled as “omit” were included and treated as “incorrect” for item analyses and calibration.
- At the item level, missing responses due to “not reached” or “missing CR scores by design” were excluded from item analyses and calibration. “Not reached” is the result of a student who started the test but never completed it during the testing window.
- For score reporting, missing responses for the machine-scorable items due to “omit” and “not reached” were treated as “incorrect.” CR items are not included in the score reporting for the field test.

6.2. Classical Item Analyses

Items scored as one (correct) or zero (incorrect) are referred to as dichotomous items. Items scored from zero to some number of points greater than one are called polytomous items. The classical item analysis includes the computation of item-by-item proportion-correct indices (*p*-values) and the item-total correlation indices for both dichotomous and polytomous items. In addition, the omit rate of items, distractor analysis, and the distributions of score categories for the polytomous items are also included in the classical item analyses results. Lastly, the associated flagging rules of these statistics are used to identify items that are not performing as expected.

6.2.1. Classical Item Difficulty Indices (*p*-value and Average Item Score)

For dichotomous items, item difficulty is indicated by its *p*-value, which is the proportion of students who answer the item correctly. The range of *p*-values is from 0.00 to 1.00. Items with high *p*-values are easier items; those with low *p*-values are more difficult. Dichotomous items are flagged for review if they have *p*-values above 0.95 (i.e., too easy) or below 0.20 (i.e., too difficult).

The formula for the *p*-value for a dichotomous item is:

$$p\text{–}value_{dich} = \frac{\sum X_{ic}}{N_i} \tag{6.1}$$

where,

X_{ic} is the number of students who answered item i correctly, and

N_i is the total number of students who were presented with item i.

For polytomous items, the difficulty is indicated by the average item score (AIS). The AIS can range from 0.00 to the maximum total possible points for an item. Desired AIS values for polytomous items generally fall within the range of 20 percent to 95 percent of the maximum obtainable item score; items with values outside this range are flagged for review. To facilitate the interpretation, the AIS values for polytomous items are often expressed as the proportion of the maximum possible score, which are equivalent to the *p*-values of dichotomous items.

For polytomous items, the *p*-value is defined as:

$$p\text{–}value_{poly} = \frac{\sum_j X_{ij}}{N_i \times Max(X_i)} \tag{6.2}$$

where,

X_{ij} is the score assigned for a given polytomous item i and student j,

N_i is the total number of students who were presented with item i, and

$Max\ (X_i)$ is the maximum possible score for item i.

6.2.2. Item-Total Correlations

The item-total correlation statistic describes the relationship between students’ performance on a specific item and students’ performance on the total assessment. It is calculated as the correlation coefficient between the item score and total score—specifically, the polyserial correlation is used as the index of item-total correlation for both polytomous and dichotomous items. Statistically, it is calculated as the correlation between an observed continuous variable and an unobserved continuous variable hypothesized to underlie the variable with ordered categories (Olsson, Drasgow, & Dorans, 1982).

Typically, the PIA is run by form; the total number of raw score points for the form is used as the criterion score in calculating the item-total correlations. Due to the block design of the field test, there are 1,680 different combination of blocks (i.e., forms). Because it will take extensive time to accumulate enough volume by form to run the PIA for the field test, the PIA was, instead, run by block for segment A and C blocks and by a pair of blocks for performance tasks (PTs) for all machine-scorable items.

The criterion score for calculating the item-total correlation was the total number of raw score points of all machine-scorable items in segment A and C blocks, because the CR items had not been scored at that time; or a pair of blocks for items appearing in Segment B PTs. The reason to use a pair of blocks instead of a single block for PTs was because each PT block includes only four to six items, so the total score might not be stable enough to be used as a criterion score.

For the CR items in segments A and C, the PIA was run by block with the criterion scores as the total number of raw scores points from all items (i.e., machine-scorable items as well as the CR items). For the CR items in Segment B, the PIA could not be run by a pair of blocks because CR sampling was conducted by single block only, making it difficult to obtain a sample with sufficient size that has all CR items in the pair of PTs scored. For these items, a raw score that uses the machine-scorable items in one of the Segment A blocks and the Segment B block was used as the criterion score.

Theoretically, the polyserial correlation ranges from -1.0 (for a perfect negative relationship) to 1.0 (for a perfect positive relationship) and is estimated as:

$$r_{polyreg} = \frac{\widehat{\beta} s_{tot}}{\sqrt{\widehat{\beta}^2 s_{tot}^2 + 1}} \tag{6.3}$$

where,

β is the item parameter to be estimated from the data, with the estimate denoted as $\hat{\beta}$, using maximum likelihood estimation; it is a regression coefficient (slope) for predicting the continuous version of an item score onto the continuous version of the total score, and

s_{tot} is the standard deviation of the criterion (the students' theta scores).

There are as many regressions as the number of boundaries between scores with all sharing a common slope, *β*. For a polytomous item, there are *m*-1 regressions, where *m* is the number of score points on the item. Beta (*β*) is the slope for all *m*-1 regressions.

Desired values for this correlation are positive and larger than 0.20. A relatively high item-total correlation coefficient value is desired, as it indicates that students with higher total raw scores on the overall test tend to perform better on the item than students with lower total raw scores. An item with a negative item-total correlation typically signifies a problem with the item, as that indicates that (1) the higher-ability students on the overall test tend to respond incorrectly to the item, if dichotomous, or are assigned a low score for the item (if polytomous); or (2) the lower-ability students on the overall test are responding correctly to the item, if dichotomous, or are assigned a high score for that item (if polytomous).

6.2.3. Distribution of Item Scores

For polytomous items, examination of the distribution of scores helps to show how well the items performed. If no students were given the highest possible score, the item may not be functioning as expected. The item may be confusing, poorly worded, or just unexpectedly difficult; the scoring rubric may be flawed; or students may not have had an opportunity to learn the content. If the rubric for an item allowed for partial credit but nearly all students received either full credit or no credit, the rubric may be inappropriate for the item. Items with a low percentage (i.e., less than three percent) of students obtaining any score point are flagged for review.

6.2.4. Omission

An item is considered "omitted" if it was seen but not answered (i.e., it was left blank). Because the CAST requires students to provide answers to all items on a page before they can move on to the next page, the possibility of an omission would be very small.

6.2.5. Distractor Analyses

6.2.5.1 The Proportion of Students Choosing Each Distractor

For the CAST, distractor analyses were conducted on selected response items (i.e., items that were not CRs). The statistics for each item included the proportion of students selecting each distractor (incorrect response), computed for the group of all students in the analysis sample, and also computed separately for the highest-performing 20 percent of students. Items were flagged for review if more high-performing students chose any distractor rather than the key. Such a result indicates that the item may have multiple correct answers or have the wrong key (i.e., the item is miskeyed).

6.2.5.2 Distractor-Total Correlation

For selected-response items, the distractor-total correlation describes the relationship between selecting a distractor for a specific item and performance on the total test. The polyserial correlation is calculated for the distractors, like the item-total correlation previously described, except that the regressions are implemented on the distractors rather than the keys. Items with distractor-total correlations above 0.00 (i.e., are positive) are flagged for review, as these items may have multiple correct answers, be miskeyed, or have other content issues.

6.2.6. Summary of Classical Item Analyses Flagging Criteria

In summary, items are flagged for review if the item analysis yields any of the following results:

- **Difficulty flags** indicate extreme values of the proportion-correct (for dichotomous items) or the proportion of the possible maximum points earned (for polytomous items).
  - A value less than 0.2 suggests that the item might be too difficult.
  - A value greater than 0.95 suggests that the item might be too easy.
- A **discrimination flag** indicates that the item does not discriminate effectively between high- and low-ability students. Items with a polyserial correlation less than 0.20 are flagged.
- An **omit flag** is set if the nonresponse rates greater than five percent for both dichotomous and polytomous items.

- A **distractor flag** is used for any distractors having positive correlation with the criterion score.
- A **miskey flag** is used for selected response items when more of the high-ability examinee group—the top 20 percent of examinees on the total test—choose any distractor rather than choosing the response keyed as correct.
- The **underrepresented score point flag** is used for any item that has less than three percent of the students at any score level.

Educational Testing Service's (ETS') Psychometric Analysis and Research staff and Assessment Development staff carefully reviewed each of the flagged items at the end of the item analyses and summarized the results for the California Department of Education (CDE).

6.2.7. Classical Item Analysis Results Summary

The overall item difficulty distributions are presented in Table 6.1. Item difficulty distributions by item type are shown in Table 6.A.1; item difficulty distributions by content domain are presented in Table 6.A.2. Because item analyses were not run by grade in high school for CR items, the total number of items for high school with all grades combined is more than the total number of items by grade in these tables (i.e., 221 versus 202).

Table 6.1 Item Difficulty Distributions

Grade	0≤p<0.2	0.2≤p<0.4	0.4≤p<0.6	0.6≤p<0.8	0.8≤p≤1.0	Total Number of Items
Grade 5	30	70	77	28	4	209
Grade 8	28	90	70	21	2	211
High School—Grade 10	67	86	40	8	1	202
High School—Grade 11	48	83	52	16	3	202
High School—Grade 12	53	81	53	12	3	202
High School—All Grades	59	94	53	12	3	221

Overall item-total correlation distributions are presented in Table 6.2. Item-total correlation distributions by item type are shown in Table 6.B.1; item-total correlation distributions by content domain are presented in Table 6.B.2.

Table 6.2 Item-Total Correlation Distributions

Grade	r<0	0≤r<0.2	0.2≤r<0.3	0.3≤r<0.4	0.4≤r<0.5	r≥0.5	Total Number of Items
Grade 5	1	3	5	16	25	159	209
Grade 8	0	5	7	15	34	150	211
High School—Grade 10[9]	0	13	15	31	54	88	201
High School—Grade 11	0	10	7	26	39	120	202
High School—Grade 12	0	8	4	26	40	124	202
High School—All Grades	0	8	5	28	41	139	221

6.3. Differential Item Functioning (DIF) Analyses

In examining the DIF between groups, the reference group is often designated as the group that is assumed to have an advantage, while the focal group refers to the group anticipated to be disadvantaged by the test.

DIF analyses were conducted for 2017–18 CAST items with sufficient sample sizes. The sample size requirements for the DIF analyses were 100 in the smaller of either the focal group or the reference group and 400 in the combined focal and reference groups. These sample size requirements are based on standard operating procedures with respect for DIF analyses at ETS.

If an item performs differentially across identifiable student groups—for example, gender or ethnicity—when students are matched on ability, the item may be measuring something else other than the intended construct (i.e., possible evidence of bias). It is important, however, to recognize that item performance differences flagged for DIF might be related to actual differences in relevant knowledge or skills between student groups (i.e., impact) or statistical Type I error, which might falsely find DIF in an item. As a result, DIF analysis is used mainly as a statistical tool to identify *potential* item bias. Subsequent reviews by content experts and bias and sensitivity experts are required to determine the source and meaning of performance differences.

6.3.1. Dichotomous Items

The Mantel-Haenszel (MH) DIF statistic was calculated for dichotomous items (Mantel & Haenszel, 1959; Holland & Thayer, 1985). For this method, students are classified to relevant student groups of interest (e.g., gender or ethnicity). Students at each total score level in the focal group (e.g., females) are compared with examinees at each total score level in the reference group (e.g., males). The common odds ratio—that is, the proportion of correct response over the proportion of incorrect response—is estimated across all levels of matched student ability using the formula in equation 6.9 (Dorans & Holland, 1993). The

[9] There was one item for which none of the students in grade ten received credit. Therefore, no item-total correlation could be calculated for that item.

resulting estimate is interpreted as the relative probability of success on a particular item for members of two groups when matched on ability.

$$\alpha_{MH} = \frac{\left(\sum_m R_{rm} \frac{W_{fm}}{N_{tm}}\right)}{\left(\sum_m R_{fm} \frac{W_{rm}}{N_{tm}}\right)} \tag{6.4}$$

where,

m indexes the score categories,

R_{rm} is the number of students in the reference group at score level m who answer the item correctly,

W_{fm} is the number of students in the focal group at score level m who answer the item incorrectly,

N_{tm} is the total number of students at score level m,

R_{fm} is the number of students in the focal group at score level m who answer the item correctly, and

W_{rm} is the number of students in the reference group at score level m who answer the item incorrectly.

To facilitate the interpretation of MH results, the common odds ratio is frequently transformed to the delta scale using the following formula (Holland & Thayer, 1985):

$$MH\ D\text{-}DIF = \text{-}2.35 \ln\left[\alpha_{MH}\right] \tag{6.5}$$

Positive values indicate DIF in favor of the focal group (i.e., positive DIF items are differentially easier for the focal group), whereas negative values indicate DIF in favor of the reference group (i.e., negative DIF items are differentially easier for the reference group).

6.3.2. DIF Procedure for Polytomous Items

The standardization DIF (Dorans & Schmitt, 1993; Zwick, Thayer, & Mazzeo, 1997; Dorans, 2013) in conjunction with the Mantel chi-square statistic (Mantel, 1963; Mantel & Haenszel, 1959) is calculated for polytomous items. The standardized mean difference (SMD) compares the item means of the two groups after adjusting for differences in the distribution of students across all items and is calculated using the following formula:

$$SMD = \frac{\sum_{m=1}^{M} N_{fm} \times E_f(Y \mid X = m)}{\sum\limits_{m=1}^{M} N_{fm}} - \frac{\sum_{m=1}^{M} N_{fm} \times E_r(Y \mid X = m)}{\sum\limits_{m=1}^{M} N_{fm}} = \frac{\sum_{m=1}^{M} D_m}{\sum_{m=1}^{M} N_{fm}} \tag{6.6}$$

where,

X is the criterion score (total raw score),

Y is the item score,

M is the number of score levels on X,

D is the difference in the distribution of students at score level m,

N_{rm} is the number of students in the reference group at score level m,

N_{fm} is the number of students in the focal group at score level m,

E_r is the expected item score for the reference group, and

E_f is the expected item score for the focal group.

A positive SMD value means that, conditional on the criterion score, the focal group has a higher mean item score than the reference group (i.e., the item is differentially easier for the focal group). In contrast, a negative SMD value means that, conditional upon the criterion score, the focal group has a lower mean item score than the reference group (i.e., the item is differentially harder for the focal group).

6.3.3. Classification

Based on the DIF statistics and significance tests, items are classified into three categories and assigned values of A, B, or C (Holland & Wainer, 1993). Category A items contain negligible DIF, Category B items exhibit slight to moderate DIF, and Category C items possess moderate to large DIF values.

The flagging criteria for dichotomous items are presented in Table 6.3; the flagging criteria for polytomous items are provided in Table 6.4.

Table 6.3 DIF Categories for Dichotomous Items

DIF Category	Criteria
A (negligible)	• Absolute value of MH D-DIF is not significantly different from zero, or is less than one. • Positive values are classified as "A+" and negative values as "A-."
B (moderate)	• Absolute value of MH D-DIF is significantly different from zero but not from one, and is at least one; OR • Absolute value of MH D-DIF is significantly different from one, but is less than 1.5. • Positive values are classified as "B+" and negative values as "B-."
C (large)	• Absolute value of MH D-DIF is significantly different from one, and is at least 1.5. • Positive values are classified as "C+" and negative values as "C-."

Table 6.4 DIF Categories for Polytomous Items

DIF Category	Criteria
A (negligible)	Mantel Chi-square *p*-value > 0.05 or |SMD/SD| ≤ 0.17
B (moderate)	Mantel Chi-square *p*-value < 0.05 and 0.17< |SMD/SD| ≤ 0.25
C (large)	Mantel Chi-square *p*-value < 0.05 and |SMD/SD| > 0.25

Note: SMD = standardized mean difference; SD = total group standard deviation of item score

DIF analyses were conducted on each test for designated comparison groups. Groups were defined on the basis of demographic variables, including gender, race or ethnicity, and primary disabilities, if the number of students in the group was sufficient. These comparison groups are specified in Table 6.5.

Table 6.5 Student Groups for DIF Comparison

DIF Type	Focal Group	Reference Group
Gender	Female	Male
Ethnicity	American Indian or Alaska Native	White
Ethnicity	Asian	White
Ethnicity	Black or African American	White
Ethnicity	Hispanic or Latino	White
English fluency	English learner	English only
Disability	Special education services	No special education services
Economic status	Economically disadvantaged	Not economically disadvantaged

6.3.4. Items Exhibiting Significant DIF

Summarized DIF results are given in Table 6.6, Table 6.7, and Table 6.8 for grades five and eight and high school respectively. Items showing C-level DIF and are considered as biased by the DIF review panel will be deactivated for future use.

Test developers are instructed to avoid selecting other C-level items that disadvantage a focal group (negative C-DIF) for future test forms unless their inclusion is deemed essential to meeting test-content specifications. If the sample size requirement for conducting DIF analyses was not met, that item was categorized in “insufficient counts.”

Table 6.6 Number of Items Flagged by DIF Category for Grade Five

Focal Group–Reference Group	DIF Category A	DIF Category B-	DIF Category B+	DIF Category C-	DIF Category C+	Insufficient Counts
Female–Male	204	3	2	0	0	0
Asian–White	207	0	2	0	0	0
Black–White	197	3	6	3	0	0
Hispanic–White	202	3	3	1	0	0
American Indian or Alaska Native–White	179	7	7	0	1	15
English learner–English only	196	4	5	3	1	0
Special education services–No special education services	194	9	2	2	2	0
Economically disadvantaged–Not economically disadvantaged	204	2	2	1	0	0

Table 6.7 Number of Items Flagged by DIF Category for Grade Eight

Focal Group–Reference Group	DIF Category A	DIF Category B-	DIF Category B+	DIF Category C-	DIF Category C+	Insufficient Counts
Female–Male	202	2	7	0	0	0
Asian–White	205	1	5	0	0	0
Black–White	202	6	3	0	0	0
Hispanic–White	207	3	1	0	0	0
American Indian or Alaska Native–White	177	6	5	0	0	23
English learner–English only	185	13	7	6	0	0
Special education services–No special education services	192	11	4	4	0	0
Economically disadvantaged–Not economically disadvantaged	208	3	0	0	0	0

Table 6.8 Number of Items Flagged by DIF Category for High School

Focal Group–Reference Group	DIF Category A	DIF Category B-	DIF Category B+	DIF Category C-	DIF Category C+	Insufficient Counts
Female–Male	213	4	2	2	0	0
Asian–White	206	1	11	0	3	0
Black–White	203	11	2	5	0	0
Hispanic–White	213	8	0	0	0	0
American Indian or Alaska Native–White	196	5	0	1	0	19
English learner–English only	175	13	17	14	2	0
Special education services–No special education services	196	11	5	9	0	0
Economically disadvantaged–Not economically disadvantaged	217	4	0	0	0	0

6.4. Test Dimensionality Analyses

The CA NGSS—the standards on which the grade-level CAST assessments are based—are referred to as three dimensional (3D) because of the interrelationships of the disciplinary core ideas (DCIs), science and engineering practices (SEPs), and crosscutting concepts (CCCs). The CAST is designed to reflect a commitment to the 3D approach in both the writing of the test items, all of which are aligned with at least two of the three dimensions, and in the assembly of test forms.

There are a number of questions that need to be addressed for reporting reliable student scores that afford valid inferences about students' mastery of the CA NGSS. For example:

- Does the test measure primarily a single dominant trait (e.g., science) or does it clearly distinguish the more specific traits defined by the DCIs, SEPs, and CCCs?
- Do the PTs measure something different than the discrete items?
- Do the technology enhanced items measure anything different from the traditional item types (e.g., multiple-choice or CR items)?

These questions can be addressed by a test dimensionality study and the answers to these questions directly impact how the test items should be calibrated and how the scores should be reported. The purpose of this dimensionality study is to examine the dimensional structure of the CAST to provide evidence on whether the CAST measures a single integrated science construct or several related knowledge subdomains.

The methodology and results of this study are briefly reviewed. Refer to the *Report on the Psychometric Studies with California Science Test Field Test Data* (ETS, 2019) for additional information*.*

6.4.1. Form Selection

The CAST field test used a block design, where each segment included multiple item blocks and each student was randomly assigned a portion of the blocks. For each grade-level assessment, this created 168 combinations of segment A and B blocks. Instead of conducting the analyses for each one of these 168 forms, three forms were carefully selected to be evaluated using the following guidelines:

- Due to the limitation of the field test item pool, not all blocks fully met the test blueprint requirement. The combination of the blocks should best conform to the test blueprint.
- In the field test, students were randomly assigned two PTs that could be from the same or different content domains. The two PTs selected in each form for this study should be from different content domains to best mimic the operational test.
- In the field test, every segment has one or two accessible blocks for students requesting accommodations. A form with accessible blocks should be selected for each grade so students needing accommodations can be included in the study.

6.4.2. Analysis Sample

For each form in grades five and eight, a random sample of 2,000 students was selected to have their responses to all CR items in the form scored. For high school, the original plan was to score 2,000 students for each grade. However, because there were fewer than 100 students from grade ten at the form level, random samples of 2,000 students taking each from only grades eleven and twelve were selected to have their responses to all CR items in the form scored. Refer to subsection on 5.1.1 Sampling Process section for more details.

6.4.3. Methodology

Two different models within the multidimensional item response theory (MIRT) framework were used to evaluate the test dimensionality in this study: a bifactor model and an MIRT model with correlated factors. Refer to the *Report on the Psychometric Studies with California Science Test Field Test Data* (ETS, 2019) for details about the specifications of these two types of models.

The multidimensional study examines five distinct ways of assigning items to substantive categories or dimensions:

1. Content domain classification (e.g., Life Sciences)
2. Each item's SEP classification
3. CCC classification
4. Item type or format (i.e., multiple-choice items vs. technology enhanced items)
5. A division of the discrete items from those assigned to PTs

Evaluating the dimensionality of a test is a subjective judgment that weighs different sources of empirical evidence. To determine whether the CAST is multidimensional or essentially unidimensional, the following evidence is considered:

- Item loadings on the general factor and on the group specific factor: If most items have high loadings on the general factor and low loadings on the group-specific factor, it suggests that a unidimensional model is sufficient for the data.
- The variance explained by the general factor and by the group specific factor: The following indices (Rodriguez, Reise, & Haviland, 2016) were used:
  - OmegaH and OmegaHS: OmegaH estimates the proportion of variance in total scores that can be attributed to a single general factor. OmegaHS reflects the reliability of a subscale score after controlling for the variance due to the general factor. High values of OmegaHS indicate that, after controlling for the variance due to the general factor, there is still a larger amount of the variance that can be explained by the group-specific variance, which could be an indicator of multidimensionality.
  - Explained common variance (ECV) (Sijtsma, 2009; Ten Berge & Socan, 2004): ECV is the ratio of the variance explained by the general factor divided by the variance explained by the general and the group factor. A high ECV value is evidence of an essentially unidimensional model.

6.4.4. Results

Results for all forms in all grade-level assessments are consistent and suggest there is no clear multidimensionality in the classifications evaluated; a unidimensional IRT model is safe and effective in calibrating the items and reporting students' scores. For the full details on the results, refer to the *Report on the Psychometric Studies with California Science Test Field Test Data* (ETS, 2019).

6.5. Item Calibration

IRT is a mathematical model that characterizes the probability of a given response as a function of a test-taker’s true ability. IRT can be used to calibrate items (i.e., fit the mathematical model), link item parameters to a given ability metric, scale or equate test scores across different forms or test administrations, evaluate item performance, build an item bank, and assemble test forms.

This subsection describes how IRT models are used in CAST field test for calibrating items. Note that no scale scores and achievement levels are reported for the CAST field test (refer to *Chapter 7: Reporting*). Therefore, the purpose of the IRT calibration for the CAST field test is to provide item parameters that are on the same scale to facilitate the research studies. For this purpose, only items that were not rejected by the data review and the CDE were included in the calibration.

6.5.1. Item Response Models

On the basis of the results from the test dimensionality study, a unidimensional model was used to calibrate the CAST field test items. The two-parameter item response theory model (2PL-IRT) was used to calibrate the dichotomous items and the generalized partial credit model (GPCM) (Muraki, 1992) was used to calibrate the polytomous items. The 2PL-IRT model is a special case of the GPCM when the maximum number of score points for the item is 1. FlexMIRT® (Cai, 2016), a multilevel and multiple-group IRT software package (version 3.5.1), is used for the calibration.

The mathematical form of the GPCM is the following:

$$P_{ih}(\theta_j) = \begin{cases} \dfrac{\exp(\sum\limits_{v=1}^{h} a_i(\theta_j - b_i + d_{iv}))}{1 + \sum\limits_{c=1}^{n_i} \exp(\sum\limits_{v=1}^{c} a_i(\theta_j - b_i + d_{iv}))}, & \text{if score } h = 1, 2, ..., n_i \\[2em] \dfrac{1}{1 + \sum\limits_{c=1}^{n_i} \exp(\sum\limits_{v=1}^{c} a_i(\theta_j - b_i + d_{iv}))}, & \text{if score } h = 0 \end{cases} \quad , \qquad (6.7)$$

where,

$P_{ih}\left(\theta_j\right)$ is the probability of student with proficiency θ_j obtaining score h on item i,

n_i is the maximum number of score points for item i,

a_i is the discrimination parameter for item i,

b_i is the location parameter for item i, and

d_{iv} is the category parameter for item i on score v, and

$d_{i1} = 0$.

When $n_i = 1$, equation 6.7 becomes an expression of the two-parameter logistic model for the dichotomous items.

6.5.2. Data Preparation

Items flagged at the PIA were reviewed by the data review committee and the CDE (refer to subsection *3.6 Data Review*). Items that were rejected by the data review committee and the CDE were excluded from the calibration.

The sample used in the item calibration includes all students who have participated in the CAST field test, with the exception of those who were considered “unmotivated” based on the guidelines defined in subsection *5.1.1.1.1 Exclusion Rules* for the batch-one sample and *5.1.1.2.1 Exclusion Rules* for the batch-two sample*.*

Similar to the classical item analyses, “omit” items were treated as incorrect. The “not-administered” items and CR items that were administered but not scored were treated as not presented.

The calibration for the high school assessment was conducted using a multigroup analyses, where the mean and variance of the ability estimates are set to 0 and 1 for grade eleven and freely estimated for grades ten and twelve. The item parameters—the item discrimination, the location, and the categories parameters—are set to equal across three grades.

The FlexMIRT output was evaluated to examine whether every execution of FlexMIRT converged. The item parameter estimates were examined for reasonableness. Items with unreasonably large parameter values or standard errors were noted and removed from the subsequent Multistage Adaptive Test (MST) Practicality Study in subsection *6.8.1*.

6.5.3. Summary of IRT parameters

The overall summary of the IRT *a*-parameter estimates is shown in Table 6.9. The number of items in each of the *a*-value intervals is shown for each grade, as well as the minimum, maximum, mean, and standard deviation (SD) values. The summaries of the IRT *a*-parameter estimates for each grade assessment are presented in appendix 6.C, in Table 6.C.1 through Table 6.C.3 by item type and Table 6.C.4 through Table 6.C.6 by content domain.

Table 6.9 Item Discrimination Parameter Distribution by Grade

IRT-a Range	Grade 5	Grade 8	High School
a<0	0	4	2
0≤a<0.2	0	6	13
0.2≤a<0.4	14	17	16
0.4≤a<0.6	13	22	26
0.6≤a<0.8	20	30	28
0.8≤a<1.0	31	32	31
1.0≤a<1.2	36	39	17
1.2≤a<1.4	20	24	20
1.4≤a<1.6	11	10	8
1.6≤a<1.8	8	8	6

IRT-a Range	Grade 5	Grade 8	High School
1.8≤a<2.0	1	4	3
a≥2.0	3	3	0
Minimum	0.20	-0.20	-0.18
Maximum	2.29	2.42	1.98
Mean	1.00	0.91	0.82
SD	0.41	0.47	0.45
Number of Items	**157**	**199**	**170**

Similar information for the IRT *b*-parameter estimates is shown in Table 6.10 for the number of items in each of the *b*-parameter intervals and the summary statistics such as the minimum, maximum, mean, and SD values. The summaries, broken down by item type and content domain, are presented in Table 6.D.1 through Table 6.D.6 in appendix 6.D.

Table 6.10 Item Difficulty Parameter Distribution by Grade

IRT-b Range	Grade 5	Grade 8	High School
b < −3.5	0	3	4
−3.5 ≤ b < −3.0	0	1	0
−3.0 ≤ b < −2.5	1	0	0
−2.5 ≤ b < −2.0	0	2	0
−2.0 ≤ b < −1.5	2	2	2
−1.5 ≤ b < −1.0	8	2	4
−1.0 ≤ b < −0.5	20	14	9
−0.5 ≤ b < 0	33	34	21
0 ≤ b < 0.5	29	31	26
0.5 ≤ b < 1.0	23	40	27
1.0 ≤ b < 1.5	22	20	19
1.5 ≤ b < 2.0	8	19	20
2.0 ≤ b < 2.5	5	8	9
2.5 ≤ b < 3.0	1	5	7
3.0 ≤ b < 3.5	1	4	10
b ≥ 3.5	4	14	12
Min	-2.90	-22.80	-22.58
Max	9.75	118.47	15.29
Mean	0.42	1.34	0.95
SD	1.35	8.73	3.01
Number of Items	**157**	**199**	**170**

6.6. Response Time Analyses

The CAST includes three segments: Segment A, Segment B, and Segment C. Each student received two blocks with 16–20 items each in Segment A, two performance tasks (PTs) with four to six items each in Segment B, and either one PT of four to six items or one block of 13 or 14 discrete items in Segment C. The test is untimed at the administration.

The estimated time for students to complete the test was 60 minutes for Segment A, 40 minutes for Segment B, and 20 minutes for Segment C. The time[10] it took students to complete a test was recorded and analyzed. Summaries of the times students spent by test segment, item type, and for the total test are given in Table 6.E.1, Table 6.E.2, and Table 6.E.3, respectively. Because the testing time for a discrete block is typically longer than that for a PT, the testing time for the total test in Table 6.E.3 was broken down for students who received a third PT in Segment C (i.e., two A blocks + three PTs) and those who received a discrete block in Segment C (i.e., two A blocks + two PTs + one C block).

6.7. Reliability Analyses

Two types of reliabilities are reported in this chapter: the reliability of the test scores and the reliability of the CR scoring.

Reliability is the extent to which differences in test scores reflect true differences in the knowledge, ability, or skill being tested rather than fluctuations due to measurement error. Thus, reliability is the consistency of the scores across conditions that do not differ systematically and only contain random measurement errors. In statistical terms, the variance in the distributions of test scores—essentially, the differences among individuals—is due partly to real differences in the knowledge, skill, or ability being tested (true variance) and due partly to measurement errors inherent in the measurement process (error variance). The reliability coefficient is an estimate of the proportion of the total variance that is true variance.

Reliability coefficients range from 0 to 1. The higher the reliability coefficient for a set of scores, the more likely individuals are to obtain very similar scores upon repeated testing occasions (assuming there is no memory or practice effect) if the students do not change in their level of the knowledge or skills measured by the test.

Reliability of the CR scoring is the extent to which two different raters given consistent scores on the same response. In this report, the interrater reliability analyses include the percent of exact and adjacent agreement between the two raters, and the quadratic-weighted kappa coefficient.

6.7.1. Internal Consistency Reliability

There are several different ways of estimating reliability of the test scores. One type of reliability estimate reported here is an internal-consistency estimate, which is derived from analysis of the consistency of the performance of individuals across items within a test.

[10] The timing data is based on capturing the amount of time spent on answering the item(s) on each page.

6.7.2. Cronbach’s Alpha Coefficient

In classical test theory, the reliability coefficient can be defined as the squared correlation between the observed score and the true score, which is equal to the correlation between parallel observed scores (Lord and Novick, 1968, p. 61). In applied settings, the requirement of repeated administrations is impractical, and methodologies estimating reliability from relationships among student performances on items within a single test form are often used. Coefficient alpha (Cronbach, 1951) is among the most common of these methodologies.

Cronbach’s alpha is defined as

$$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{i=1}^{K} s_{X_i}^2}{s_X^2}\right), \tag{6.8}$$

where,

K is the number of items in the test,

$s_{X_i}^2$ is the observed variance of item i in the test, and

s_X^2 is the observed variance of the total test score.

Because CAST field test forms have mixed item types (PT and non-PT), it is more appropriate to report stratified alpha (Feldt & Brennan, 1989). Stratified alpha is a reliability estimate computed by dividing the test into parts (strata), computing coefficient alpha separately for each part, and using the results to estimate a reliability coefficient for the total score. Stratified alpha is used here because different parts of the test consist of different item types and may measure different skills. The formula for the stratified alpha is:

$$\rho_{strata} = 1 - \frac{\sum_j \sigma_{X_j}^2 (1-\alpha_j)}{\sigma_X^2}, \tag{6.9}$$

where,

$\sigma_{X_j}^2$ is the variance for strata j of the test,

σ_X^2 is the total variance of the test, and

α_j is the Cronbach’s alpha for strata j of the test.

Estimates of stratified alpha are computed by substituting sample estimates for the parameters in the formula.

6.7.3. Standard Error of Measurement (SEM)

The SEM provides a measure of score instability on a different metric. The SEM is the square root of the error variance in the scores, i.e., the standard deviation of the distribution of the differences between students’ observed scores and their true scores. The SEM is calculated by:

$$SEM = s_X \sqrt{1-\alpha} \tag{6.10}$$

where,

α is the reliability estimated in equation 6.8, and

s_X is the standard deviation of the total score.

The SEM is useful in determining the confidence interval that likely captures a student's true score. A student's true score can be thought of as the mean of observed scores a student would earn over an infinite number of independent administrations of the test. Approximately 95 percent of the students will have scores within the range of their true scores: -1.96 SEMs to their true scores +1.96 SEMs (Crocker & Algina, 1986). For example, if a student's observed score on a given test equals 345 points, and the SEM equals 5, one can be 95 percent confident that the student's true score lies between 335 and 355 points (i.e., 345 ± 10).

6.7.4. Results for the Fixed Forms

As described in subsection *6.4 Test Dimensionality Analyses*, three forms were chosen for the dimensionality study. These forms were the ones closest to the blueprint and have all the CR responses scored. Reliabilities reported in this subsection are restricted to these three forms.

Table 6.11 through Table 6.13 provide the reliability estimates and the SEMs for the three grade-level assessments. The correlations between the observed content domain scores and the reliability and the SEM, for the total test and then by the content domain, are provided for each assessment.

Table 6.11 Test Reliability of Total Score and Subscores for Grade Five

Form	Domain	Number of Items	Total Score	Physical Sciences	Life Sciences	Earth and Space Sciences	Reliability	SEM
1	Total Score	50	1.00	0.93	0.87	0.89	0.88	3.19
1	Physical Sciences	18	0.93	1.00	0.70	0.74	0.78	1.98
1	Life Sciences	17	0.87	0.70	1.00	0.67	0.67	1.72
1	Earth and Space Sciences	15	0.89	0.74	0.67	1.00	0.69	1.78
2	Total Score	49	1.00	0.92	0.88	0.93	0.89	3.18
2	Physical Sciences	16	0.92	1.00	0.72	0.76	0.70	2.01
2	Life Sciences	14	0.88	0.72	1.00	0.73	0.70	1.52
2	Earth and Space Sciences	19	0.93	0.76	0.73	1.00	0.76	1.93
3	Total Score	49	1.00	0.93	0.88	0.89	0.89	3.18
3	Physical Sciences	18	0.93	1.00	0.72	0.73	0.78	2.13
3	Life Sciences	14	0.88	0.72	1.00	0.70	0.72	1.56
3	Earth and Space Sciences	17	0.89	0.73	0.70	1.00	0.71	1.77

Table 6.12 Test Reliability of Total Score and Subscores for Grade Eight

Form	Domain	Number of Items	Total Score	Physical Sciences	Life Sciences	Earth and Space Sciences	Reliability	SEM
1	Total Score	44	1.00	0.89	0.90	0.85	0.85	3.38
1	Physical Sciences	15	0.89	1.00	0.68	0.63	0.66	2.11
1	Life Sciences	16	0.90	0.68	1.00	0.67	0.72	1.86
1	Earth and Space Sciences	13	0.85	0.63	0.67	1.00	0.60	1.86
2	Total Score	44	1.00	0.93	0.88	0.85	0.86	3.17
2	Physical Sciences	18	0.93	1.00	0.73	0.67	0.74	2.03
2	Life Sciences	10	0.88	0.73	1.00	0.65	0.67	1.61
2	Earth and Space Sciences	16	0.85	0.67	0.65	1.00	0.58	1.81
3	Total Score	44	1.00	0.92	0.91	0.82	0.87	3.27
3	Physical Sciences	15	0.92	1.00	0.74	0.64	0.73	2.07
3	Life Sciences	17	0.91	0.74	1.00	0.65	0.72	1.93
3	Earth and Space Sciences	12	0.82	0.64	0.65	1.00	0.51	1.63

Table 6.13 Test Reliability of Total Score and Subscores for High School

Form	Domain	Number of Items	Total Score	Physical Sciences	Life Sciences	Earth and Space Sciences	Reliability	SEM
1	Total Score	49	1.00	0.84	0.91	0.87	0.86	3.13
1	Physical Sciences	16	0.84	1.00	0.65	0.60	0.60	1.72
1	Life Sciences	17	0.91	0.65	1.00	0.69	0.72	1.93
1	Earth and Space Sciences	16	0.87	0.60	0.69	1.00	0.70	1.73
2	Total Score	48	1.00	0.87	0.90	0.89	0.88	3.03
2	Physical Sciences	16	0.87	1.00	0.68	0.65	0.69	1.71
2	Life Sciences	16	0.90	0.68	1.00	0.69	0.71	1.79
2	Earth and Space Sciences	16	0.89	0.65	0.69	1.00	0.72	1.74
3	Total Score	44	1.00	0.70	0.89	0.89	0.81	3.10
3	Physical Sciences	12	0.70	1.00	0.47	0.48	0.34	1.50
3	Life Sciences	15	0.89	0.47	1.00	0.65	0.63	2.00
3	Earth and Space Sciences	17	0.89	0.48	0.65	1.00	0.69	1.80

6.7.5. Interrater Reliability Analyses

To monitor the consistency of ratings assigned to students' responses by human raters, approximately 15 percent of the human-scored CR responses received a second rating ("backreading"); the responses in this subsample were randomly selected and scored by two raters. The two sets of ratings are used to compute statistics describing the consistency (reliability) of the human ratings. This interrater consistency is described in two ways:

1. Percentage agreement between two human raters,
2. Quadratic-weighted kappa coefficient.

6.7.6. Percentage Agreement

Percentage agreement between two raters includes the percentage of exact score agreement, the percentage of adjacent score agreement, and the percentage of exact plus adjacent score agreement. Adjacent score agreement means agreement between scores that differ by just one point. The fewer the item score points, the fewer degrees of freedom on which two raters can vary, and the higher the percentage of agreement.

6.7.7. Quadratic-weighted Kappa

Quadratic-weighted kappa is also used because kappa does not take into account the degree of disagreement between raters. It is a generalization of the simple kappa coefficient using weights to quantify the relative difference between categories. The range of the quadratic-weighted kappa is from 0.0 to 1.0, with perfect agreement being equal to 1.0.

For a human-scored item with m categories, one can construct an $m \times m$ rating table with scores provided by two raters, A and B. Suppose m is the maximum obtainable score for each item, n_{st} is the number of responses for which rater A's score = s, and rater B's score = t, n_{s+} is the number of responses for which rater A's score = s, n_{+t} is the number of responses for which rater B = t, and n_{++} is the number of all responses from either rater A or rater B. The weighted kappa coefficient is defined as:

$$\kappa_{st} = \frac{\left(\sum_{s=0}^{m}\sum_{t=0}^{m} w_{st}\frac{n_{st}}{n_{++}}\right) - \left(\sum_{s=0}^{m}\sum_{t=0}^{m} w_{st}\frac{n_{s+}n_{+t}}{n_{++}^{2}}\right)}{1-\left(\sum_{s=0}^{m}\sum_{t=0}^{m} w_{st}\frac{n_{s+}n_{+t}}{n_{++}^{2}}\right)} \tag{6.11}$$

For quadratic-weighted kappa, the weights are:

$$w_{st} = 1-\frac{(s-t)^2}{m^2} \tag{6.12}$$

6.7.8. Results

Table 6.F.1 through Table 6.F.3 present the results of the interrater analyses and descriptive statistics of the ratings by the two raters on CR items, including the following:

- Number of score points in each item
- Number of raters for each round of rating (total count)
- Mean of the item score for nominal rater 1 and rater 2
- Standard deviation of the item score for nominal rater 1 and rater 2

- Percent of exact agreement
- Quadratic-weighted kappa

Quadratic-weighted kappa statistics provide evidence of the degree to which a student's score is consistent from one rater to another. Research has shown the values of quadratic-weighted kappa greater than 0.70 indicate excellent agreement (Williamson, Xi, & Breyer, 2012).

Given the criteria mentioned, the results of these four items in Table 6.F.1 show 34 out of 57 items with quadratic-weighted kappa higher than 0.7. The interrater agreement is also high, with the percent of exact agreement ranging from 69.07 percent to 96.78 percent. Scoring performance statistics were monitored carefully at the pool level throughout scoring. Where low agreement was identified for an item, raters were directed to retrain prior to continuing to score. Scoring leaders also closely monitored and mentored raters throughout scoring when any issues with an individual were identified.

6.8. Research Studies

6.8.1. Multistage Adaptive Test (MST) Practicality Study

6.8.1.1 Description

Adaptive tests can provide more precise estimates of student ability, with improvement most notable at extreme ability levels (van der Linden, 2005). They do so by tailoring the difficulty of the test to the performance level of the student. Because CAST Segment A is comprised largely of discrete items and would appear to be a good candidate for adaptation, this study evaluated, for each grade level, whether an adaptive Segment A will improve measurement of student ability over a linear form, and whether there is enough improvement to offset the complexity and risk inherent in all adaptive testing.

Given the size of the item bank, a multistage test (MST) instead of an item-level computer adaptive test is more likely to succeed. As such, the field test MST study is only the first opportunity to evaluate the increased efficiency of an MST for Segment A.

The CAST item pool will be expanded from the 2017–18 field test to the 2018–19 first operational administration. ETS will replicate the study in the summer of 2019 with the expanded item pool to make final decisions on (1) whether the item pool supports an MST and, if so, (2) the number of stage-two difficulty levels and their corresponding decision thresholds for operational implementation in 2019–20, which is the earliest possible implementation of an adaptive Segment A.

This study used the field-test data to investigate MST of two stages with one router block at the first stage and two or three levels of difficulty at the second stage. The goal of this study was to

- inform MST design decisions (e.g., number of levels of difficulty in stage two);
- determine the extent to which MST improves measurement of student ability in comparison to a linear Segment A form;
- establish procedures to support efficient replication of the study in 2018–19 with an expanded item pool; and
- establish MST assembly procedures to support (potential) operational implementation in 2019–20.

The methodology and results are briefly reviewed in this subsection. Refer to the *Report on the Psychometric Studies with California Science Test Field-Test Data* (ETS, 2019) for the full details.

6.8.1.2 Methodology

The MST practicality study was conducted using the item parameters and student distributions estimated from the field test data. Items used had good classical item analysis results—well-fit by their item response theory model—and item parameters that fell within appropriate ranges.

In this subsection, MST panels are described as follows:

- **MST 1-2 design:** MST panel with two levels of difficulty in the second stage
- **MST 1-3 design:** MST panel with three levels of difficulty in the second stage

In both cases, the panels were assembled to conform to both the content rules in the blueprint and statistical specifications. The content rules include, for instance, the number of items and points per domain.

In terms of statistical specifications, the router block should be comprised of items that have a wide distribution of item difficulties of moderate item discriminations, while each of the stage-two difficulty blocks should contain items that measure well across a narrower and targeted range of performance. An easy block was assembled to target performance levels about 0.75 SDs below the average student score. The hard block targeted a performance centered about 0.75 SDs above the average student score. Both the router block and the medium block (when assembled for MST 1-3 design) targeted the center of the performance range.

The first threshold, t_1, was set at where the information functions for the easy and medium difficulty second-stage blocks cross. Doing so achieves the goal of routing each student along the path that is likely to be most informative. Similarly, the upper threshold, t_2, was set at the point where the medium and hard second-stage information functions cross. (A similar procedure was used if there were only two levels of difficulty at the second stage: The intersection of the two curves was the single threshold.)

The performance of the MST 1-2 and MST 1-3 designs were evaluated against a linear form (a combination of router and medium difficulty blocks) with respect to the following criteria:

- **Test information functions (TIFs):** The information functions for the second stage levels; expected to be reasonably distinctive for the panel to be meaningful
- **Relative efficiency:** Ratio of the information function for the two designs being compared at every true ability level
- **Measurement precision:** Measured by the conditional standard error of measurement, i.e., the standard deviation of the estimated ability distribution at every true ability level
- **Conditional bias:** The difference between the expected value of the estimated ability distribution and the true ability level
- **Routing rates:** The percent of students being routed to each path (i.e., easy, medium, or difficult second stage level)

- **Average item overlap rates:** The average of proportion of items in common across tests administered to any two students
- **Measured student group achievement gap**

6.8.1.3 Results

The mean difficulty parameters for the entire pool were all positive, indicating the pools were difficult in general. Based on the calibrated item pools for grades five and eight and high school, test forms using MST 1-2 and MST 1-3 designs were assembled using automated test assembly methods (van der Linden, 2005). The assembly took into consideration content constraints (e.g., the test blueprint requirements) and the statistical specification targets, finding the set of items that satisfied both the content and statistical constraints.

In this study, the target statistical specification was set so the router and medium second-stage blocks had TIFs that were centered at the average student score; the easy second-stage block centered at 0.75 standard deviation below the average student score; and the hard second-stage block centered at 0.75 standard deviation above the average student score.

In general, the test forms satisfied content blueprint requirements and statistical requirements. However, because of a lack of sufficient number of polytomous items in the item pool, the test length at the second stage for grade five and high school was increased to 20 items to meet the requirement in the blueprint regarding the total score points. No attempts were made to limit the extent to which any item appeared in two or more of the second-stage blocks. Such sharing was necessary given the small size of the item pool and the rigor of the content requirements. The ETS content team then conducted a thorough review of the assembled forms to ensure that assembled forms did not have item pairs that were not to appear together in the same form, or if item content clued other item content.

For the MST 1-3 forms, the information functions for grade eight were well-behaved, providing substantial information near the points where they were targeted. The information functions of medium and difficult blocks for grade five did not provide as much information as the easy and router blocks. In addition, a high percentage of shared items were observed between blocks at the second stage. The information functions corresponding to the easy and medium blocks for high school did not provide as much information as the difficult and router blocks. Finally, a high percentage of common items were also observed between blocks at the second stage. The routing thresholds were set for all grades and the routing rates were all reasonable.

For all three grades, MST designs outperformed the linear form in terms of conditional SEM. In terms of relative efficiency, for moderate ability levels, the linear form was as or slightly more precise than the MST. For low or high proficiency levels, the linear form would need to have as many as 16 additional items to match the precision of the MST for grade five and high school and would need to have as many as 23 additional items to match the precision of the MST for grade eight. All three designs provided similar estimates of student group achievement gaps.

The performance of MST 1-2 and MST 1-3 designs were comparable in all the evaluation criteria used. The comparability of MST 1-3 and MST 1-2 designs can be attributed to the large number of shared second-stage items, perhaps indicating limited size of the item pool.

The MST 1-2 design is recommended based on the field test data, given it increased measurement precision compared to the linear form. The medium block of MST 1-3 did not provide greater improvement in terms of overall measurement in relation to the MST 1-2 design.

6.8.1.4 Limitations of the Study

The MST practicality study was conducted based on item parameters and empirical student ability distributions estimated from field-test data. However, there are a few limitations with the field test data that would impact the generalizability of the results.

First, not all local educational agencies have fully implemented the CA NGSS into curricular and classroom practices. Students' future familiarity with the CA NGSS and the CAST could cause differences between field-test performance and performance on the operational assessment.

In addition, students' motivation to take the field test may have impacted performance.

Finally, several characteristics of the field test item pool also limited the generalizability of the results. For example, the number of items is limited, especially the polytomous items for grade five and high school. In addition, items tended to be difficult for all grades, especially for high school.

Given these limitations, these findings should be revisited in spring 2019, when the item pool expands. Study results in 2019–20 are expected to be similar or better, given an expanded item pool.

6.8.2. Content Screen-Out Study

6.8.2.1 Description

Students receive two PTs in CAST Segment B, where the context of each PT has a primary domain—one of the three main science content domains of Life Sciences, Earth and Space Sciences, or Physical Sciences—and, in some cases, a secondary domain. For the field test administration, students were randomly assigned any two PTs from the pool of eight available PTs in each tested grade. In operational administrations, students must receive two PTs in two different domains.

There are a number of ways in which PT assignment could take place. For instance, random assignment could be used, but then certain students may be advantaged or disadvantaged if students are found to perform better in contexts with which they are interested and experienced and they happen to be assigned PTs in the domains in which they are most or least familiar. Alternatively, performance in Segment A, which is comprised of 32–34 items, roughly spread evenly across the three domains, could be used to screen out PTs in the domain in which the student demonstrated the weakest performance, so as not to disadvantage any student in their assignment of PTs in Segment B. For instance, students who performed conspicuously poorly on Life Sciences items in Segment A would be assigned PTs in the other two domains, whereas students who performed similarly across all three domains in Segment A would be randomly assigned two PTs from any two domains.

Such screening out would be helpful to inform selection of PTs only if (1) student performance tended to differ by science domain and (2) student performance in Segment A was predictive of performance in Segment B. This study investigates these conditions using the 2017–18 field test data.

The methodology and results of this study are briefly reviewed. Refer to the *California Science Test (CAST): A Report on the Psychometric Studies with California Science Test Field Test Data* (ETS, 2019) for the full details.

6.8.2.2 Methodology

This study used grade eight field-test data for students who had at least two of the three PTs of interest across their Segment B and Segment C sections of the CAST. There were 112,125 students who met this criteria, with about 16.5 percent—or 18,526—of them assigned to all three PTs. Each PT had one CR, but given the limited CR scoring (refer to subsection *5.1.1 Sampling Process*), only about seven percent of students had at least one CR scored across their PTs.

Because of the loss of items in Segment A following CDE and educator data reviews, students' Segment A scores were based on 22 to 31 items, with an average of 25 items. Data loss was not even across the three content domains: Students had 7 to 11 Life Sciences items in Segment A with an average of 9 items, 4 to 9 Earth and Space Sciences items with an average of 6 items, and 9 to 12 Physical Sciences items with an average of 10 items. The low numbers of Earth and Space Sciences items made the Segment A Earth and Space Sciences score less reliable than that of the other two domains.

The analysis involved first computing comparable scores and subscores for students for the Segment A overall score, Segment A domain subscores, Segment B overall score[11], and individual PT subscores. Such scores were computed by taking the inverse of the test characteristic curve formed by the items associated with each score or subscore. Correlations and disattenutated correlations were then computed to describe the association among the scores, particularly between the Segment A domain scores and the individual PT scores.

Then, an alignment index was computed as a measure of the alignment between students' performance on Segment A and their assigned PTs in Segment B. Details of this index are provided in the *Report on the Psychometric Studies with California Science Test Field Test Data* (ETS, 2019).

Finally, three linear models were run to determine the best predictors of the overall Segment B score. The first model predicted the overall B score only with the overall Segment A score, the second model added the individual Segment A domain scores, and the third model added the alignment index.

Operational implementation of a screener mechanism is advisable if Segment A domain scores added substantially to the prediction of Segment B scores and if the alignment index also added substantially.

6.8.2.3 Results

Evidence that content screening would be necessary or prudent would include differential performance across domains in Segment A and strong relationships between Segment A domain subscores and the corresponding PTs in the same domains. The Segment A domain scores were moderately correlated once measurement error was taken into account (.69 to .76), suggesting that, generally, students perform comparably across the domains with only some differential performance.

[11] The Segment B overall score includes PTs assigned in Segment B and Segment C.

Note, however, that true score correlations are better approximated using multidimensional IRT, as was done in the dimensionality study using three forms of segments A and B, than with disattenuated correlations, particularly those based on rough approximations for the score reliabilities. For all three forms, the MIRT estimated true-score correlations among content domains were greater than .9, indicating little differential performance by content domain. Regardless of which domain each PT represented, they all correlated highest with the Physical Sciences Segment A subscore instead of their corresponding domain score. Moreover, the PT scores were even more highly correlated with the overall Segment A score, suggesting that performance in Segment B is better informed by overall Segment A performance than performance on Segment A items in the same domain as the PT. These results, however, should be interpreted with caution, given the small numbers of items per score.

The results of the linear models further suggested a screener may not be useful as all three models had R^2 values within 0.0016 of each other, ranging from .3399 to .3415. Model two, with both the overall Segment A score and the domain A subscores, did fit significantly better than model one with only the Segment A score. However, the very small difference in R^2 values indicates there is no practical difference between the models. Model three had the same R^2 value as model two to the fourth decimal place, suggesting that the alignment index did not help explain any additional variation in the overall Segment B score.

Some follow-up analyses were conducted to further probe the potential utility of a screener. First, students were divided into groups by their weakest domain in Segment A. Then, students' scores in Segment B were compared for each of the three possible pairings of the three PTs (Life Sciences and Earth and Space Sciences PTs, Life Sciences and Physical Sciences PTs, and Earth and Space Sciences and Physical Sciences PTs). If students who were not assigned any PTs in the domain in which they were weakest in Segment A performed better on average than students who were assigned a PT in their weakest domain, that would serve as evidence in favor of a screener. Table 6.14 shows the results of this analysis.

The results were not definitive. In fact, students with the weakest performances in Physical Sciences in Segment A and who were not assigned a Physical Sciences PT in Segment B performed significantly worse than those who were assigned a Physical Sciences PT in Segment B. These results were strongly tied to having only one PT per domain. The particular Physical Sciences PT in the study was easier than the other two PTs. Thus, students who were assigned it generally performed better than those who were not, regardless of in which domain they were weakest in Segment A.

Table 6.14. Average Total B Scores for Students by Weakest Domain

Weakest Domain	PT	Assigned weakest?	Average B Score	Significantly lower than score for those not assigned the weakest domain?
Life Sciences (LS)	B1 (LS)-B2 (ESS)	Yes	-0.41	Yes
Life Sciences	B1 (LS)-B3 (PS)	Yes	-0.20	No
Life Sciences	B2 (ESS)-B3 (PS)	No	-0.23	NA
Earth and Space Sciences (ESS)	B2 (ESS)-B1 (LS)	Yes	-0.26	Yes
Earth and Space Sciences	B2 (ESS)-B3 (PS)	Yes	-0.07	No
Earth and Space Sciences	B1 (LS)-B3 (PS)	No	-0.05	NA
Physical Sciences (PS)	B3 (PS)-B2 (ESS)	Yes	-0.15	No
Physical Sciences	B3 (PS)-B1 (LS)	Yes	-0.19	No
Physical Sciences	B2 (ESS)-B1 (LS)	No	-0.35	NA

The CAST field test screener study provided weak evidence that a screener from Segment A to Segment B would be useful in limiting the extent that students are advantaged or disadvantaged by the domain-specific PTs they are assigned. The regression results also suggest that even if necessary, it may be difficult or impossible to properly apply a screener. This is likely because students with true performance profiles across content domains are difficult to distinguish from those whose performance differs across domains only due to measurement error.

6.8.2.4 Limitations of the Study

This data suffers from several limitations, as detailed in the *Report on the Psychometric Studies with California Science Test Field Test Data* (ETS, 2019), that require interpreting these results with caution. Of primary concern is that the study can be conducted only for grade eight because after the educator data review and CDE item review, only grade eight had at least one intact PT per domain. In fact, grade eight had exactly three intact PTs, one for each domain, which is another major limitation as it is not clear if results are representative of these three PTs only or generalizable to the behavior of the PTs in general.

However, this study will be replicated in 2019, with the first year of operational data, to inform a final decision on the implementation of a screener for CAST testing. Accordingly, the primary purpose of the screen-out study using the 2017–18 field test data is to establish study protocols to facilitate replication of the study in 2019, when there will be tight timelines to make and implement a final decision.

References

Cai, L. (2016). flexMIRT® R 3.5.1: Flexible multilevel and multidimensional item response theory analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group, LLC.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20 (1): 37–46.

Crocker, L. M., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart and Winston.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Dorans, N. J. (2013). ETS contributions to the quantitative assessment of item, test, and score fairness. *ETS Research Report Series*, i–38.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Dorans, N. J., & Schmitt, A. P. (1993). Constructed response and differential item functioning: A pragmatic approach. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 135–65). Hillsdale, NH: Lawrence Erlbaum Associates, Inc.

Educational Testing Service. (2019). *Report on the psychometric studies with California Science Test field test data.* (Draft manuscript).

Feldt, L. S., & Brennan, R. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). New York: Macmillan.

Holland, P. W., & Thayer, D. T. (1985). *An alternative definition of the ETS delta scale of item difficulty* (Research Report 85–43). Princeton, NJ: Educational Testing Service.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.

Lord, F.M. and Novick, M.R. (1968) Statistical theories of mental test scores. Addison-Wesley, Menlo Park.

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association*, *58*, 690–700.

Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analyses of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719–48.

Muraki, Eiji. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Olsson, U., Drasgow, F., & Dorans, N. J. (1982). The polyserial correlation coefficient. *Psychometrika, 47*, 337–347.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, *21*(2), 137–150.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*(1), 145–154.

Ten Berge, J. M., & Socan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika*, *69*(4), 613–625.

van der Linden, W. J. (2005). *Linear models for optimal test design.* New York, NY: Springer.

Williamson, D.M., Xi, X., & Breyer, F.J. (2012), A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, 31: 2–13.

Zwick, R., Thayer, D. T., & Mazzeo, J. (1997). Descriptive and inferential procedures for assessing differential item functioning in polytomous items. *Applied Measurement in Education, 10*(4), 321–344.

Appendix 6.A: Item Difficulty Distribution

Notes:

- Item analyses were not run by grade in high school for CR items. Therefore, all counts for CR items in high school grades ten through twelve are shown as “NA.” As a result, the total numbers of items for high school are not the same as the ones for each grade in high school because item analyses were not run by grade in high school for CR items.
- Item types are as follows:
  - MC = Multiple-choice item
  - CR = Constructed-response item
  - TEI = Technology-enhanced item
  - Composite = Composite item (an item type that includes multiple parts)

Table 6.A.1 Item Difficulty Distributions by Item Type

Grade	Item Type	0≤p<0.2	0.2≤p<0.4	0.4≤p<0.6	0.6≤p<0.8	0.8≤p≤1.0	Total Number of Items
Grade 5	MC	8	32	46	18	2	106
Grade 5	CR	4	5	3	1	0	13
Grade 5	TEI	17	30	26	7	2	82
Grade 5	Composite	1	3	2	2	0	8
Grade 8	MC	4	45	38	9	2	98
Grade 8	CR	3	9	5	3	0	20
Grade 8	TEI	19	31	20	7	0	77
Grade 8	Composite	2	5	7	2	0	16
High School—Grade 10	MC	17	62	27	1	0	107
High School—Grade 10	CR	NA	NA	NA	NA	NA	NA
High School—Grade 10	TEI	44	23	13	7	1	88
High School—Grade 10	Composite	6	1	0	0	0	7

Grade	Item Type	0≤p<0.2	0.2≤p<0.4	0.4≤p<0.6	0.6≤p<0.8	0.8≤p≤1.0	Total Number of Items
High School—Grade 11	MC	11	50	37	9	0	107
High School—Grade 11	CR	NA	NA	NA	NA	NA	NA
High School—Grade 11	TEI	33	30	15	7	3	88
High School—Grade 11	Composite	4	3	0	0	0	7
High School—Grade 12	MC	14	50	37	6	0	107
High School—Grade 12	CR	NA	NA	NA	NA	NA	NA
High School—Grade 12	TEI	35	28	16	6	3	88
High School—Grade 12	Composite	4	3	0	0	0	7
High School—All Grades	MC	12	52	37	6	0	107
High School—All Grades	CR	9	9	1	0	0	19
High School—All Grades	TEI	34	30	15	6	3	88
High School—All Grades	Composite	4	3	0	0	0	7

Table 6.A.2 Item Difficulty Distributions by Content Domain

Grade	Content Domain	0≤p<0.2	0.2≤p<0.4	0.4≤p<0.6	0.6≤p<0.8	0.8≤p≤1.0	Total Number of Items
Grade 5	PS	15	35	42	15	0	107
Grade 5	LS	5	10	12	9	2	38
Grade 5	ESS	10	25	23	4	2	64
Grade 8	PS	4	35	26	6	1	72
Grade 8	LS	12	30	27	3	0	72
Grade 8	ESS	12	25	17	12	1	67
High School—Grade 10	PS	30	31	10	1	0	72
High School—Grade 10	LS	23	33	14	3	1	74
High School—Grade 10	ESS	14	22	16	4	0	56
High School—Grade 11	PS	23	32	11	6	0	72
High School—Grade 11	LS	15	33	20	3	3	74
High School—Grade 11	ESS	10	18	21	7	0	56
High School—Grade 12	PS	26	31	11	4	0	72
High School—Grade 12	LS	16	32	21	2	3	74
High School—Grade 12	ESS	11	18	21	6	0	56
High School—All Grades	PS	26	39	10	4	0	79
High School—All Grades	LS	19	33	21	2	3	78
High School—All Grades	ESS	14	22	22	6	0	64

Appendix 6.B: Item-Total Correlation

Notes:

- Out of all grade ten students, none received credit on one technology-enhanced item (TEI) in the Physical Sciences domain. As a result, the total number of TEIs for grade ten students is one fewer than that for grades eleven and twelve students (in Table 6.B.1) and the total number of Physical Sciences items for grade ten students is one fewer than for grades eleven and twelve students (in Table 6.B.2).
- Item analyses were not run by grade in high school for CR items. Therefore, all counts for CR items in high school grades ten through twelve are shown as “NA.” As a result, the total numbers of items for high school are not the same as the ones for each grade in high school because item analyses were not run by grade in high school for CR items.

Table 6.B.1 Item-Total Correlation Distributions by Item Type

Grade	Item Type	r<0	0≤r<0.2	0.2≤r<0.3	0.3≤r<0.4	0.4≤r<0.5	r≥0.5	Total Number of Items
Grade 5	MC	1	2	2	13	19	69	106
Grade 5	CR	0	0	0	0	0	13	13
Grade 5	TEI	0	1	3	3	6	69	82
Grade 5	Composite	0	0	0	0	0	8	8
Grade 8	MC	0	2	5	8	20	63	98
Grade 8	CR	0	0	0	0	1	19	20
Grade 8	TEI	0	3	2	7	11	54	77
Grade 8	Composite	0	0	0	0	2	14	16
High School—Grade 10	MC	0	8	9	20	37	33	107
High School—Grade 10	CR	NA	NA	NA	NA	NA	NA	NA
High School—Grade 10	TEI	0	5	6	11	14	51	87
High School—Grade 10	Composite	0	0	0	0	3	4	7
High School—Grade 11	MC	0	5	6	18	25	53	107
High School—Grade 11	CR	NA	NA	NA	NA	NA	NA	NA
High School—Grade 11	TEI	0	5	1	8	14	60	88
High School—Grade 11	Composite	0	0	0	0	0	7	7

Grade	Item Type	r<0	0≤r<0.2	0.2≤r<0.3	0.3≤r<0.4	0.4≤r<0.5	r≥0.5	Total Number of Items
High School—Grade 12	MC	0	3	4	18	26	56	107
High School—Grade 12	CR	NA	NA	NA	NA	NA	NA	NA
High School—Grade 12	TEI	0	5	0	8	14	61	88
High School—Grade 12	Composite	0	0	0	0	0	7	7
High School—All Grades	MC	0	3	5	18	25	56	107
High School—All Grades	CR	0	0	0	1	3	15	19
High School—All Grades	TEI	0	5	0	9	13	61	88
High School—All Grades	Composite	0	0	0	0	0	7	7

Table 6.B.2 Item-Total Correlation Distributions by Content Domain

Grade	Content Domain	r<0	0≤r<0.2	0.2≤r<0.3	0.3≤r<0.4	0.4≤r<0.5	r≥0.5	Total Number of Items
Grade 5	PS	1	2	2	7	13	82	107
Grade 5	LS	0	0	0	4	5	29	38
Grade 5	ESS	0	1	3	5	7	48	64
Grade 8	PS	0	2	4	4	14	48	72
Grade 8	LS	0	0	1	5	11	55	72
Grade 8	ESS	0	3	2	6	9	47	67
High School—Grade 10	PS	0	6	8	10	18	29	71
High School—Grade 10	LS	0	4	5	15	20	30	74
High School—Grade 10	ESS	0	3	2	6	16	29	56
High School—Grade 11	PS	0	4	5	10	10	43	72
High School—Grade 11	LS	0	2	2	11	15	44	74
High School—Grade 11	ESS	0	4	0	5	14	33	56
High School—Grade 12	PS	0	2	4	10	14	42	72
High School—Grade 12	LS	0	2	0	11	15	46	74
High School—Grade 12	ESS	0	4	0	5	11	36	56
High School—All Grades	PS	0	2	4	11	15	47	79
High School—All Grades	LS	0	2	1	12	14	49	78
High School—All Grades	ESS	0	4	0	5	12	43	64

Appendix 6.C: Item Discrimination Parameter Distribution

Notes:

- MC = Multiple-choice item
- CR = Constructed-response item
- TEI = Technology-enhanced item
- Composite = Composite item (an item type that includes multiple parts)

Table 6.C.1 Item Discrimination Parameter Distribution by Item Type for Grade Five

IRT-a Range	MC	CR	TEI	Composite	Number of Items
a < 0	0	0	0	0	0
0 ≤ a < 0.2	0	0	0	0	0
0.2 ≤ a < 0.4	10	0	4	0	14
0.4 ≤ a < 0.6	12	0	1	0	13
0.6 ≤ a < 0.8	13	0	6	1	20
0.8 ≤ a < 1.0	12	1	17	1	31
1.0 ≤ a < 1.2	16	4	15	1	36
1.2 ≤ a < 1.4	10	2	6	2	20
1.4 ≤ a < 1.6	7	1	3	0	11
1.6 ≤ a < 1.8	3	3	2	0	8
1.8 ≤ a < 2.0	0	0	1	0	1
a ≥ 2.0	2	0	1	0	3
Minimum	0.20	0.96	0.25	0.63	NA
Maximum	2.29	1.77	2.09	1.35	NA
Mean	0.93	1.33	1.02	1.02	NA
SD	0.44	0.28	0.36	0.31	NA
Number of Items	**85**	**11**	**56**	**5**	**157**

Table 6.C.2 Item Discrimination Parameter Distribution by Item Type for Grade Eight

IRT-a Range	MC	CR	TEI	Composite	Number of Items
a < 0	3	0	1	0	4
0 ≤ a < 0.2	4	0	2	0	6
0.2 ≤ a < 0.4	11	0	6	0	17
0.4 ≤ a < 0.6	14	3	5	0	22
0.6 ≤ a < 0.8	12	3	14	1	30
0.8 ≤ a < 1.0	15	0	10	7	32
1.0 ≤ a < 1.2	14	8	15	2	39
1.2 ≤ a < 1.4	14	2	7	1	24
1.4 ≤ a < 1.6	5	2	2	1	10
1.6 ≤ a < 1.8	4	1	3	0	8
1.8 ≤ a < 2.0	1	0	3	0	4
a ≥ 2.0	0	0	3	0	3
Minimum	-0.20	0.44	-0.19	0.62	NA
Maximum	1.84	1.72	2.42	1.43	NA
Mean	0.84	1.02	0.98	0.98	NA
SD	0.45	0.37	0.52	0.22	NA
Number of Items	**97**	**19**	**71**	**12**	**199**

Table 6.C.3 Item Discrimination Parameter Distribution by Item Type for High School

IRT-a Range	MC	CR	TEI	Composite	Number of Items
a < 0	0	0	2	0	2
0 ≤ a < 0.2	11	0	2	0	13
0.2 ≤ a < 0.4	12	0	4	0	16
0.4 ≤ a < 0.6	18	1	7	0	26
0.6 ≤ a < 0.8	23	1	3	1	28
0.8 ≤ a < 1.0	14	2	14	1	31
1.0 ≤ a < 1.2	10	0	7	0	17
1.2 ≤ a < 1.4	4	2	14	0	20
1.4 ≤ a < 1.6	1	2	5	0	8
1.6 ≤ a < 1.8	5	0	1	0	6
1.8 ≤ a < 2.0	0	1	2	0	3
a ≥ 2.0	0	0	0	0	0
Minimum	0.02	0.51	-0.18	0.61	NA
Maximum	1.76	1.98	1.94	0.88	NA
Mean	0.70	1.17	0.96	0.75	NA
SD	0.41	0.46	0.47	0.19	NA
Number of Items	**98**	**9**	**61**	**2**	**170**

Table 6.C.4 Item Discrimination Parameter Distribution by Content Domain for Grade Five

IRT-a Range	Life Sciences	Physical Sciences	Earth and Space Sciences	Number of Items
a < 0	0	0	0	0
0 ≤ a < 0.2	0	0	0	0
0.2 ≤ a < 0.4	2	7	5	14
0.4 ≤ a < 0.6	4	8	1	13
0.6 ≤ a < 0.8	3	12	5	20
0.8 ≤ a < 1.0	5	19	7	31
1.0 ≤ a < 1.2	3	20	13	36
1.2 ≤ a < 1.4	1	9	10	20
1.4 ≤ a < 1.6	5	3	3	11
1.6 ≤ a < 1.8	4	3	1	8
1.8 ≤ a < 2.0	0	0	1	1
a ≥ 2.0	1	2	0	3
Minimum	0.34	0.20	0.22	NA
Maximum	2.20	2.29	1.90	NA
Mean	1.10	0.95	1.01	NA
SD	0.49	0.39	0.39	NA
Number of Items	**28**	**83**	**46**	**157**

Table 6.C.5 Item Discrimination Parameter Distribution by Content Domain for Grade Eight

IRT-a Range	Life Sciences	Physical Sciences	Earth and Space Sciences	Number of Items
a < 0	0	3	1	4
0 ≤ a < 0.2	2	3	1	6
0.2 ≤ a < 0.4	5	5	7	17
0.4 ≤ a < 0.6	6	11	5	22
0.6 ≤ a < 0.8	11	9	10	30
0.8 ≤ a < 1.0	11	13	8	32
1.0 ≤ a < 1.2	18	13	8	39
1.2 ≤ a < 1.4	9	4	11	24
1.4 ≤ a < 1.6	3	3	4	10
1.6 ≤ a < 1.8	1	3	4	8
1.8 ≤ a < 2.0	2	1	1	4
a ≥ 2.0	0	2	1	3
Minimum	0.11	-0.19	-0.20	NA
Maximum	1.88	2.42	2.40	NA
Mean	0.93	0.85	0.96	NA
SD	0.39	0.50	0.50	NA
Number of Items	**68**	**70**	**61**	**199**

Table 6.C.6 Item Discrimination Parameter Distribution by Content Domain for High School

IRT-a Range	Life Sciences	Physical Sciences	Earth and Space Sciences	Number of Items
a < 0	1	0	1	2
0 ≤ a < 0.2	7	5	1	13
0.2 ≤ a < 0.4	5	6	5	16
0.4 ≤ a < 0.6	9	8	9	26
0.6 ≤ a < 0.8	12	9	7	28
0.8 ≤ a < 1.0	10	11	10	31
1.0 ≤ a < 1.2	7	8	2	17
1.2 ≤ a < 1.4	6	4	10	20
1.4 ≤ a < 1.6	3	2	3	8
1.6 ≤ a < 1.8	1	3	2	6
1.8 ≤ a < 2.0	1	2	0	3
a ≥ 2.0	0	0	0	0
Minimum	-0.13	0.07	-0.18	NA
Maximum	1.94	1.98	1.74	NA
Mean	0.77	0.83	0.85	NA
SD	0.44	0.48	0.44	NA
Number of Items	**62**	**58**	**50**	**170**

Appendix 6.D: Item Difficulty Parameter Distribution

Notes:

- MC = Multiple-choice item
- CR = Constructed-response item
- TEI = Technology-enhanced item
- Composite = Composite item (an item type that includes multiple parts)

Table 6.D.1 Item Difficulty Parameter Distribution by Item Type for Grade Five

IRT-b Range	MC	CR	TEI	Composite	Number of Items
b < −3.5	0	0	0	0	0
−3.5 ≤ b < −3.0	0	0	0	0	0
−3.0 ≤ b < −2.5	0	0	1	0	1
−2.5 ≤ b < −2.0	0	0	0	0	0
−2.0 ≤ b < −1.5	1	0	1	0	2
−1.5 ≤ b < −1.0	5	0	3	0	8
−1.0 ≤ b < −0.5	13	1	5	1	20
−0.5 ≤ b < 0	19	1	11	2	33
0 ≤ b < 0.5	19	2	7	1	29
0.5 ≤ b < 1.0	9	2	11	1	23
1.0 ≤ b < 1.5	8	5	9	0	22
1.5 ≤ b < 2.0	5	0	3	0	8
2.0 ≤ b < 2.5	3	0	2	0	5
2.5 ≤ b < 3.0	0	0	1	0	1
3.0 ≤ b < 3.5	1	0	0	0	1
b ≥ 3.5	2	0	2	0	4
Minimum	-1.73	-0.56	-2.90	-0.76	NA
Maximum	4.40	1.25	9.75	0.81	NA
Mean	0.29	0.65	0.61	-0.15	NA
SD	1.10	0.61	1.76	0.60	NA
Number of Items	**85**	**11**	**56**	**5**	**157**

Table 6.D.2 Item Difficulty Parameter Distribution by Item Type for Grade Eight

IRT-*b* Range	MC	CR	TEI	Composite	Number of Items
b < −3.5	3	0	0	0	3
−3.5 ≤ b < −3.0	0	0	1	0	1
−3.0 ≤ b < −2.5	0	0	0	0	0
−2.5 ≤ b < −2.0	0	0	2	0	2
−2.0 ≤ b < −1.5	2	0	0	0	2
−1.5 ≤ b < −1.0	0	0	1	1	2
−1.0 ≤ b < −0.5	8	2	3	1	14
−0.5 ≤ b < 0	17	3	12	2	34
0 ≤ b < 0.5	17	3	7	4	31
0.5 ≤ b < 1.0	17	4	17	2	40
1.0 ≤ b < 1.5	8	5	6	1	20
1.5 ≤ b < 2.0	10	1	7	1	19
2.0 ≤ b < 2.5	5	0	3	0	8
2.5 ≤ b < 3.0	1	0	4	0	5
3.0 ≤ b < 3.5	2	0	2	0	4
b ≥ 3.5	7	1	6	0	14
Minimum	-22.80	-0.83	-3.46	-1.14	NA
Maximum	118.47	3.55	10.69	1.64	NA
Mean	1.68	0.66	1.25	0.27	NA
SD	12.35	1.02	2.38	0.77	NA
Number of Items	**97**	**19**	**71**	**12**	**199**

Table 6.D.3 Item Difficulty Parameter Distribution by Item Type for High School

IRT-*b* Range	MC	CR	TEI	Composite	Number of Items
b < −3.5	1	0	3	0	4
−3.5 ≤ b < −3.0	0	0	0	0	0
−3.0 ≤ b < −2.5	0	0	0	0	0
−2.5 ≤ b < −2.0	0	0	0	0	0
−2.0 ≤ b < −1.5	0	0	2	0	2
−1.5 ≤ b < −1.0	3	0	1	0	4
−1.0 ≤ b < −0.5	4	0	5	0	9
−0.5 ≤ b < 0	16	0	5	0	21
0 ≤ b < 0.5	17	2	7	0	26
0.5 ≤ b < 1.0	13	3	11	0	27
1.0 ≤ b < 1.5	10	3	5	1	19
1.5 ≤ b < 2.0	11	1	8	0	20
2.0 ≤ b < 2.5	5	0	3	1	9
2.5 ≤ b < 3.0	4	0	3	0	7
3.0 ≤ b < 3.5	5	0	5	0	10
b ≥ 3.5	9	0	3	0	12
Minimum	-22.58	0.31	-13.22	1.23	NA
Maximum	15.29	1.53	5.64	2.13	NA
Mean	1.12	0.93	0.64	1.68	NA
SD	3.30	0.42	2.77	0.64	NA
Number of Items	**98**	**9**	**61**	**2**	**170**

Table 6.D.4 Item Difficulty Parameter Distribution by Content Domain for Grade Five

IRT-*b* Range	Life Sciences	Physical Sciences	Earth and Space Sciences	Number of Items
b < −3.5	0	0	0	0
−3.5 ≤ b < −3.0	0	0	0	0
−3.0 ≤ b < −2.5	0	1	0	1
−2.5 ≤ b < −2.0	0	0	0	0
−2.0 ≤ b < −1.5	0	0	2	2
−1.5 ≤ b < −1.0	5	3	0	8
−1.0 ≤ b < −0.5	4	13	3	20
−0.5 ≤ b < 0	7	16	10	33
0 ≤ b < 0.5	4	13	12	29
0.5 ≤ b < 1.0	2	14	7	23
1.0 ≤ b < 1.5	2	14	6	22
1.5 ≤ b < 2.0	2	3	3	8
2.0 ≤ b < 2.5	2	3	0	5
2.5 ≤ b < 3.0	0	1	0	1
3.0 ≤ b < 3.5	0	0	1	1
b ≥ 3.5	0	2	2	4
Minimum	-1.34	-2.90	-1.73	NA
Maximum	2.25	4.40	9.75	NA
Mean	0.05	0.39	0.68	NA
SD	1.04	1.11	1.80	NA
Number of Items	**28**	**83**	**46**	**157**

Table 6.D.5 Item Difficulty Parameter Distribution by Content Domain for Grade Eight

IRT-*b* Range	Life Sciences	Physical Sciences	Earth and Space Sciences	Number of Items
b < −3.5	0	2	1	3
−3.5 ≤ b < −3.0	0	1	0	1
−3.0 ≤ b < −2.5	0	0	0	0
−2.5 ≤ b < −2.0	1	0	1	2
−2.0 ≤ b < −1.5	0	1	1	2
−1.5 ≤ b < −1.0	1	1	0	2
−1.0 ≤ b < −0.5	1	6	7	14
−0.5 ≤ b < 0	14	9	11	34
0 ≤ b < 0.5	10	13	8	31
0.5 ≤ b < 1.0	17	13	10	40
1.0 ≤ b < 1.5	7	8	5	20
1.5 ≤ b < 2.0	5	7	7	19
2.0 ≤ b < 2.5	4	1	3	8
2.5 ≤ b < 3.0	1	2	2	5
3.0 ≤ b < 3.5	0	3	1	4
b ≥ 3.5	7	3	4	14
Minimum	-2.13	-22.80	-5.21	NA
Maximum	9.83	118.47	10.69	NA
Mean	1.19	1.92	0.86	NA
SD	2.10	14.49	2.16	NA
Number of Items	**68**	**70**	**61**	**199**

Table 6.D.6 Item Difficulty Parameter Distribution by Content Domain for High School

IRT-*b* Range	Life Sciences	Physical Sciences	Earth and Space Sciences	Number of Items
b < −3.5	3	0	1	4
−3.5 ≤ b < −3.0	0	0	0	0
−3.0 ≤ b < −2.5	0	0	0	0
−2.5 ≤ b < −2.0	0	0	0	0
−2.0 ≤ b < −1.5	2	0	0	2
−1.5 ≤ b < −1.0	0	1	3	4
−1.0 ≤ b < −0.5	2	3	4	9
−0.5 ≤ b < 0	8	1	12	21
0 ≤ b < 0.5	10	7	9	26
0.5 ≤ b < 1.0	8	14	5	27
1.0 ≤ b < 1.5	4	9	6	19
1.5 ≤ b < 2.0	7	9	4	20
2.0 ≤ b < 2.5	6	1	2	9
2.5 ≤ b < 3.0	6	0	1	7
3.0 ≤ b < 3.5	3	5	2	10
b ≥ 3.5	3	8	1	12
Minimum	-22.58	-1.02	-13.22	NA
Maximum	9.80	15.29	3.57	NA
Mean	0.68	1.79	0.30	NA
SD	3.80	2.40	2.27	NA
Number of Items	**62**	**58**	**50**	**171**

Appendix 6.E: Response Time Analyses

Notes:

- Response time analyses were based on students who logged on the test and whose total testing time at the test level did not equal to zero.
- In Table 6.E.1, PT refers to performance task. According to the test design, half of the students received a PT and the other half of the students received a discrete item block in Segment C. Segment C (PT) provides a summary of the testing time to complete Segment C if students received a PT. Segment C (Discrete) provides a summary of the testing time to complete Segment C if students received a discrete item block.
- Because response time was recorded at the page level, items that were on a page with multiple items were excluded in the analysis in Table 6.E.2.
- The following abbreviations apply in Table 6.E.2:
  - MC = Multiple-choice item
  - CR = Constructed-response item
  - TEI = Technology-enhanced item
  - Composite = Composite item (an item type that includes multiple parts)
- The criterion for students to be included in Table 6.E.3 is that they have no “Not Seen” items.

Table 6.E.1 Average Testing Time (in Minutes) by Segment

Grade	Segment	N	Mean	SD	Min	Max	1 Percentile	10 Percentile	25 Percentile	50 Percentile	75 Percentile	90 Percentile	99 Percentile
Grade 5	Segment A	460,271	90.2	48.2	0.1	1301.2	20.8	43.8	58.6	79.9	109.8	147.8	260.3
Grade 5	Segment B	457,498	22.0	14.2	0.0	490.5	3.3	8.2	12.7	19.1	27.7	38.5	71.5
Grade 5	Segment C (PT)	228,177	10.4	8.4	0.0	508.5	1.0	3.1	5.2	8.5	13.2	19.4	40.9
Grade 5	Segment C (Discrete)	229,483	22.9	15.2	0.0	614.4	2.2	7.9	13.5	20.0	28.6	40.4	77.0

Grade	Segment	N	Mean	SD	Min	Max	1 Percentile	10 Percentile	25 Percentile	50 Percentile	75 Percentile	90 Percentile	99 Percentile
Grade 8	Segment A	458,460	76.8	37.3	0.0	824.5	13.8	37.9	52.2	70.5	94.4	122.7	199.1
Grade 8	Segment B	451,229	15.5	11.2	0.0	440.9	1.5	4.4	8.0	13.5	20.1	27.9	53.9
Grade 8	Segment C (PT)	224,999	8.1	7.6	0.0	277.4	0.5	1.9	3.7	6.6	10.2	15.1	36.5
Grade 8	Segment C (Discrete)	227,260	21.0	13.3	0.0	321.5	2.0	6.5	12.2	19.0	26.9	36.7	65.7
HS—Grade 10	Segment A	6,384	51.9	25.5	2.3	238.9	7.5	22.7	34.7	48.8	65.1	83.7	130.0
HS—Grade 10	Segment B	6,315	10.5	8.1	0.1	109.1	1.3	3.2	5.2	8.6	13.3	19.3	40.7
HS—Grade 10	Segment C (PT)	3,197	5.2	5.8	0.0	91.4	0.4	1.2	2.1	3.8	6.3	10.2	28.1
HS—Grade 10	Segment C (Discrete)	3,122	12.3	9.4	0.0	95.5	1.1	2.7	5.4	10.6	17.1	23.8	44.1
HS—Grade 11	Segment A	136,658	52.4	26.6	0.0	456.8	6.0	22.1	35.0	49.4	65.7	84.5	136.5
HS—Grade 11	Segment B	134,935	10.8	8.3	0.0	243.7	1.1	3.0	5.4	9.0	13.8	19.9	40.7
HS—Grade 11	Segment C (PT)	67,331	5.4	6.0	0.0	173.8	0.4	1.2	2.2	4.0	6.6	10.4	29.6
HS—Grade 11	Segment C (Discrete)	67,763	13.1	9.5	0.0	155.4	1.0	2.8	6.1	11.6	17.6	24.6	44.1
HS—Grade 12	Segment A	404,923	43.1	22.0	0.1	531.0	5.1	16.8	28.2	41.0	55.1	70.2	108.6
HS—Grade 12	Segment B	400,134	8.4	6.3	0.0	291.1	1.0	2.4	4.1	7.0	10.9	15.6	30.2
HS—Grade 12	Segment C (PT)	199,403	4.1	4.2	0.0	172.5	0.3	0.9	1.7	3.1	5.2	7.9	20.0
HS—Grade 12	Segment C (Discrete)	201,129	10.5	7.8	0.0	152.2	0.9	2.2	4.6	9.2	14.6	20.3	35.1
HS—All Grades	Segment A	547,965	45.5	23.6	0.0	531.0	5.3	17.9	29.7	43.0	57.9	74.3	118.6
HS—All Grades	Segment B	541,384	9.0	7.0	0.0	291.1	1.0	2.5	4.4	7.5	11.7	16.8	33.6
HS—All Grades	Segment C (PT)	269,931	4.4	4.8	0.0	173.8	0.4	1.0	1.8	3.3	5.5	8.6	22.7
HS—All Grades	Segment C (Discrete)	272,014	11.2	8.3	0.0	155.4	0.9	2.3	4.9	9.8	15.4	21.5	38.2

Table 6.E.2 Average Testing Time (in Minutes) by Item Type

Grade	Item Type	N	Mean	SD	Min	Max	1 Percentile	10 Percentile	25 Percentile	50 Percentile	75 Percentile	90 Percentile	99 Percentile
Grade 5	MC	9,720,070	1.7	2.1	0.0	455.9	0.1	0.4	0.7	1.2	2.0	3.3	9.9
Grade 5	CR	1,376,800	7.4	7.9	0.0	600.8	0.3	1.6	2.8	5.0	9.1	15.7	38.5
Grade 5	TEI	9,153,445	1.9	2.3	0.0	585.7	0.1	0.4	0.7	1.3	2.2	3.7	10.6
Grade 5	Composite	1,583,476	4.4	5.1	0.0	244.6	0.0	1.1	1.8	2.9	5.1	8.9	24.9
Grade 8	MC	7,803,834	1.5	1.8	0.0	114.9	0.0	0.2	0.6	1.1	1.8	3.0	8.6
Grade 8	CR	1,756,231	5.7	5.3	0.0	280.9	0.1	1.4	2.6	4.4	7.2	11.3	25.7
Grade 8	TEI	7,042,908	1.8	2.0	0.0	199.2	0.1	0.3	0.7	1.3	2.2	3.5	9.5
Grade 8	Composite	2,600,436	3.2	3.1	0.0	193.5	0.1	0.7	1.4	2.4	3.9	6.2	15.2
HS—Grade 10	MC	137,533	1.1	1.3	0.00	42.80	0.0	0.1	0.3	0.8	1.4	2.2	5.7
HS—Grade 10	CR	16,854	2.7	2.9	0.00	43.00	0.1	0.4	1.0	1.9	3.5	5.7	13.7
HS—Grade 10	TEI	86,508	1.2	1.5	0.00	61.20	0.1	0.2	0.4	0.8	1.6	2.6	7.0
HS—Grade 10	Composite	15,774	2.5	2.5	0.00	37.60	0.2	0.6	1.0	1.7	3.2	5.4	12.2
HS—Grade 11	MC	2,942,925	1.1	1.4	0.00	119.50	0.0	0.1	0.3	0.8	1.4	2.2	6.0
HS—Grade 11	CR	361,786	2.8	2.9	0.00	106.60	0.1	0.4	1.0	2.1	3.6	5.8	13.5
HS—Grade 11	TEI	1,851,593	1.3	1.6	0.00	186.10	0.1	0.2	0.4	0.9	1.6	2.6	7.3
HS—Grade 11	Composite	338,491	2.4	2.5	0.00	94.60	0.2	0.6	0.9	1.6	3.1	5.1	11.5
HS—Grade 12	MC	8,726,884	0.9	1.0	0.00	126.00	0.0	0.1	0.2	0.7	1.2	1.8	4.5
HS—Grade 12	CR	1,072,151	2.2	2.2	0.00	132.10	0.1	0.3	0.8	1.6	2.9	4.6	10.3
HS—Grade 12	TEI	5,485,614	1.0	1.2	0.00	148.00	0.1	0.1	0.3	0.7	1.3	2.2	5.6
HS—Grade 12	Composite	1,005,017	2.1	2.0	0.00	93.80	0.2	0.5	0.9	1.4	2.6	4.4	9.3
HS—All Grades	MC	11,807,342	0.9	1.1	0.0	126.0	0.0	0.1	0.3	0.7	1.2	1.9	4.9
HS—All Grades	CR	1,450,791	2.4	2.5	0.0	132.1	0.1	0.3	0.8	1.7	3.1	4.9	11.2
HS—All Grades	TEI	7,423,715	1.1	1.3	0.0	186.1	0.1	0.2	0.4	0.8	1.4	2.3	6.0
HS—All Grades	Composite	1,359,282	2.2	2.2	0.0	94.6	0.2	0.5	0.9	1.5	2.7	4.6	10.0

Table 6.E.3 Average Testing Time (in Minutes) for the Total Test

Grade	Segment	N	Mean	SD	Min	Max	1 Percentile	10 Percentile	25 Percentile	50 Percentile	75 Percentile	90 Percentile	99 Percentile
Grade 5	2 A blocks + 2 PTs +1C block	228,237	132.1	65.6	4.3	1698.0	31.7	66.4	89.0	119.7	160.3	210.6	357.5
Grade 5	2 A blocks + 3 PTs	228,583	125.5	62.8	3.7	1123.7	31.1	62.8	84.0	113.4	152.6	201.5	342.0
Grade 8	2 A blocks + 2 PTs +1C block	224,074	110.6	51.9	4.3	1249.2	21.4	55.3	76.7	103.0	135.1	173.5	279.6
Grade 8	2 A blocks + 3 PTs	225,486	103.1	48.5	3.6	1396.7	20.2	51.6	71.3	95.7	126.1	162.0	260.5
HS—Grade 10	2 A blocks + 2 PTs +1C block	3,088	72.5	36.0	3.9	366.1	13.5	32.5	48.5	67.3	91.4	117.9	180.7
HS—Grade 10	2 A blocks + 3 PTs	3,197	69.8	33.9	4.1	314.3	11.5	31.4	46.7	65.5	86.5	112.3	174.6
HS—Grade 11	2 A blocks + 2 PTs +1C block	67,020	74.5	37.7	3.5	521.7	10.0	31.5	49.7	70.3	93.0	119.8	193.4
HS—Grade 11	2 A blocks + 3 PTs	67,416	70.1	34.9	3.1	491.3	9.8	30.5	47.2	66.3	87.4	112.2	181.8
HS—Grade 12	2 A blocks + 2 PTs +1C block	199,004	60.3	30.5	2.2	618.1	8.4	24.3	39.5	57.6	76.8	97.3	151.9
HS—Grade 12	2 A blocks + 3 PTs	199,795	57.1	28.6	1.6	686.0	7.9	23.4	37.7	54.5	72.6	91.8	143.7
HS—All Grades	2 A blocks + 2 PTs +1C block	269,112	64.0	33.1	2.2	618.1	8.7	25.7	41.6	60.5	81.1	103.7	167.9
HS—All Grades	2 A blocks + 3 PTs	270,408	60.5	30.9	1.6	686.0	8.2	24.8	39.7	57.3	76.5	97.7	157.2

Appendix 6.F: Interrater Reliability for CR Scoring

Table 6.F.1 Interrater Reliability for Constructed-Response Items for Grade Five

Prompt ID	Score Points	Number of Raters for Rating 1	Number of Raters for Rating 2	H1 Mean	H1 STD	H2 Mean	H2 STD	H1-H2 Percent Exact Agreement	H1-H2 Percent Adjacent Agreement	H1-H2 QWK
VH667949	1	6003	800	0.47	0.50	0.48	0.50	87.88	100.00	0.76
VH668026	1	6003	799	0.56	0.50	0.58	0.49	88.61	100.00	0.77
VH695572	1	8003	800	0.32	0.47	0.34	0.47	78.13	100.00	0.50
VH709025	2	8003	800	0.67	0.69	0.68	0.70	91.25	100.00	0.91
VH709052	1	6003	800	0.30	0.46	0.29	0.45	90.00	100.00	0.76
VH731235	1	6000	800	0.40	0.49	0.39	0.49	79.63	100.00	0.57
VH733167	2	8000	800	0.57	0.57	0.58	0.58	74.13	99.50	0.59
VH737471	1	6002	800	0.64	0.48	0.65	0.48	79.50	100.00	0.55
VH810103	1	4002	800	0.20	0.40	0.19	0.39	90.88	100.00	0.71
VH810308	1	6000	800	0.20	0.40	0.20	0.40	91.13	100.00	0.72
VH810523	1	6000	800	0.27	0.45	0.26	0.44	88.00	100.00	0.69
VH810549	1	4001	799	0.15	0.36	0.18	0.39	85.11	100.00	0.47
VH810950	1	8003	799	0.10	0.30	0.11	0.31	94.74	100.00	0.72
VH811101	1	4001	800	0.24	0.42	0.23	0.42	83.13	100.00	0.53
VH813229	2	6002	800	1.14	0.90	1.14	0.90	85.75	99.88	0.91

Table 6.F.2 Interrater Reliability for Constructed-Response Items for Grade Eight

Prompt ID	Score Points	Number of Raters for Rating 1	Number of Raters for Rating 2	H1 Mean	H1 STD	H2 Mean	H2 STD	H1-H2 Percent Exact Agreement	H1-H2 Percent Adjacent Agreement	H1-H2 QWK
VH695226	2	8001	800	1.42	0.79	1.43	0.79	82.88	99.25	0.84
VH699333	1	3999	799	0.41	0.49	0.43	0.49	80.23	100.00	0.59
VH702216	2	3999	800	0.70	0.80	0.70	0.80	77.25	100.00	0.82
VH702611	2	4003	799	1.24	0.91	1.25	0.91	92.24	100.00	0.95
VH728143	2	6005	800	0.96	0.94	0.96	0.94	86.50	99.50	0.92
VH730085	2	8004	800	0.53	0.75	0.53	0.76	77.38	97.13	0.73
VH734423	2	6003	799	1.11	0.90	1.12	0.90	83.98	99.75	0.90
VH738505	1	8003	800	0.66	0.48	0.64	0.48	90.88	100.00	0.80
VH738912	2	6003	799	0.94	0.94	0.94	0.95	90.49	99.50	0.94
VH803445	2	6005	799	0.60	0.70	0.58	0.69	85.86	99.87	0.85
VH803496	1	6000	800	0.26	0.44	0.27	0.44	82.50	100.00	0.55
VH803535	2	8004	800	0.68	0.81	0.67	0.82	84.75	99.50	0.87
VH803647	1	6002	800	0.61	0.49	0.60	0.49	86.50	100.00	0.72
VH804554	1	4003	800	0.94	0.67	0.91	0.66	74.25	99.88	0.70
VH805907	1	4000	800	0.24	0.42	0.26	0.44	84.63	100.00	0.59
VH807320	2	4000	798	0.60	0.61	0.61	0.62	78.57	100.00	0.71
VH809423	2	4001	799	0.88	0.91	0.89	0.91	86.23	100.00	0.92
VH809632	1	4003	800	0.29	0.45	0.28	0.45	83.38	100.00	0.59
VH814728	3	6002	800	1.13	0.95	1.09	0.94	70.26	100.00	0.83
VH826960	2	6003	800	0.50	0.74	0.51	0.74	88.38	98.88	0.86
VH810601	1	4002	799	0.17	0.37	0.17	0.37	93.25	100.00	0.76
VH811932	2	4004	797	0.23	0.51	0.25	0.51	82.48	100.00	0.66
VH811273	2	8002	800	0.13	0.42	0.14	0.42	90.75	99.88	0.73

Table 6.F.3 Interrater Reliability for Constructed Response Items for High School

Prompt ID	Score Points	Number of Raters for Rating 1	Number of Raters for Rating 2	H1 Mean	H1 STD	H2 Mean	H2 STD	H1-H2 Percent Exact Agreement	H1-H2 Percent Adjacent Agreement	H1-H2 QWK
VH651810	1	10112	1531	0.37	0.48	0.36	0.48	93.53	100.00	0.86
VH651815	2	10003	1501	0.43	0.54	0.45	0.55	84.88	99.80	0.74
VH696269	2	12681	1189	0.62	0.74	0.63	0.74	77.38	99.16	0.77
VH702164	2	4368	1566	0.30	0.53	0.30	0.53	79.69	99.68	0.62
VH730945	2	10001	1500	0.52	0.70	0.54	0.70	84.40	99.47	0.82
VH804572	1	10001	1500	0.38	0.48	0.39	0.49	92.20	100.00	0.83
VH804586	2	4378	1574	0.15	0.43	0.15	0.42	89.19	99.67	0.67
VH804610	1	9262	1500	0.21	0.41	0.21	0.41	91.46	100.00	0.75
VH805894	1	9142	1499	0.39	0.49	0.39	0.49	96.78	100.00	0.93
VH807293	2	9139	1499	0.43	0.63	0.42	0.63	70.23	100.00	0.62
VH807384	2	5917	1800	0.41	0.60	0.43	0.59	73.00	98.93	0.58
VH808368	2	13338	1199	0.26	0.44	0.26	0.44	88.76	100.00	0.71
VH807168	1	9241	1499	0.17	0.38	0.16	0.37	79.99	100.00	0.28
VH807248	1	5777	1800	0.08	0.27	0.08	0.27	91.39	100.00	0.41
VH805924	2	9248	1500	0.37	0.56	0.38	0.59	69.64	98.08	0.45
VH736248	1	4356	1552	0.07	0.26	0.06	0.25	92.95	100.00	0.45
VH808361	1	9379	1500	0.55	0.50	0.55	0.50	84.86	100.00	0.69
VH710712	2	12683	1189	0.39	0.58	0.40	0.57	69.07	99.47	0.50
VH807280	2	4367	1566	0.23	0.52	0.22	0.51	84.13	99.33	0.66

Chapter 7: Reporting

The primary purpose of the California Science Test (CAST) field test was to provide the data used to evaluate the properties of the items and for research studies to inform future test designs and score reporting.

Because the forms developed did not fully conform to the test blueprint, the results are not intended to represent a precise measure of students' achievement of the science assessment based on the California Next Generation Science Standards [CA NGSS]). Instead, a preliminary indicator was reported to provide a broad and early indication about a local educational agency's (LEA's) implementation of the CA NGSS to meet the requirement for reporting results to parents or guardians and the public. The preliminary indicators include a percent of maximum points earned—referred to as "percent correct" in this chapter—and an indicator category. These indicators have not been equated to adjust for form difficulty, thus making them more useful for gauging group, rather than individual student, performance.

7.1. Percent Correct Scores

The percent correct scores calculated for the field test only include the machine-scorable items in Segment A. The CR items are not included because, as described in subsection *5.1.1 Sampling Process*, only a sample of the responses to each CR prompt was scored to support the item analyses, item response theory calibration, and test dimensionality study. As a result, not all students have scores on every CR item taken. The performance tasks (PTs) in Segment B are not included because, compared with the discrete items in Segment A, the PTs are relatively new to the students, who might still be taking the time to familiarize themselves with these types of items. Because of this lack of familiarity, results might not truly reflect what students do or do not know about the measured knowledge or skills. In future administrations, with the exposure granted by practice and training tests, students should become more familiar with these types of items, making it more appropriate to include students' scores on such tasks for reporting.

The percent correct is calculated using the following equation:

$$\text{Percent correct}=\frac{\text{Number of points earned for all machine scorable items in Segment A}}{\text{Maximum number of points for all machine scorable items in Segment A}} \quad (7.1)$$

If a student took the first Segment A block and exited the test, the maximum number of points for machine-scorable items in Segment A is undefined because the second Segment A block has not been assigned. For such students, the maximum number of points is defined as the average of the maximum number of points for all machine-scorable items in Segment A across all six possible combinations of Segment A blocks (i.e., A1-A2, A1-A3, A1-A4, A2-A3, A2-A4, and A3-A4).

For the high school test, during the preliminary item analyses, two items were discovered to be flawed and were removed from calculating the percent correct.

Table 7.A.1 in appendix 7.A provides the distributions of the percent correct for the total group by grade.

7.2. Preliminary Indicator Categories

The preliminary indicators are descriptive statements with corresponding threshold scores used in reporting the CAST results. Indicators are considered preliminary because they are available to parents/guardians and the public before the completion of the science assessments' development (CDE, 2018).

There are three preliminary indicator categories. Refer to Table 7.1 for a description of each. A student's preliminary indicator category provides a general indication of the student's understanding of the CA NGSS.

Table 7.1 Descriptions of the Preliminary Indicator Categories

Category	Description
3	Student performance suggests a considerable understanding of the California Next Generation Science Standards.
2	Student performance suggests a moderate understanding of the California Next Generation Science Standards.
1	Student performance suggests a limited understanding of the California Next Generation Science Standards.

Because the preliminary indicator is a general, rather than a precise, indication of student content knowledge, the indicator categories are set so the majority of students will be in the middle categories (i.e., category 2). Students who performed at or below the chance level receive an indicator category of 1. Students who performed exceedingly well—90 percent correct or above—receive an indicator category of 3.

The chance probability for each CAST grade-level assessment is calculated as the chance score across all machine-scorable items in Segment A over the maximum points possible. The chance score for each item is calculated as the maximum points of this item times the chance probability, which is defined as one divided by the number of possible options. The cut scores for the three indicator categories are shown in Table 7.2.

Table 7.2 Threshold Scores for Indicator Categories

Grade	Threshold Score Between Indicators 1 and 2	Threshold Score Between Indicators 2 and 3
Grade 5	15%	90%
Grade 8	18%	90%
High School	14%	90%

Table 7.3 provides distributions of the indicator categories for each grade for the total group of students.

Table 7.3 Percent of Students in Each Indicator Category

Grade	Indicator Category 1	Indicator Category 2	Indicator Category 3
Grade 5	4.5	95.0	0.5
Grade 8	8.8	91.1	0.1
High School—Grade 10	11.5	88.5	0.0

Grade	Indicator Category 1	Indicator Category 2	Indicator Category 3
High School—Grade 11	7.6	92.4	0.0
High School—Grade 12	9.0	91.0	0.0
High School—All Grades	8.7	91.3	0.0

The tables in appendix 7.B—Table 7.B.1 through Table 7.B.6—provide the distributions of the indicator categories by demographic variables for each grade level.

Reference

California Department of Education. (2018). *Science assessments preliminary indicators FAQ, question 7.* Retrieved from https://bit.ly/3mDuGZn

Appendix 7.A: Distribution of Percent Correct Scores

Table 7.A.1 Distribution of Percent Correct Scores

Percent Correct	Grade 5 Percentage of Students	Grade 8 Percentage of Students	Grade 10 Percentage of Students	Grade 11 Percentage of Students	Grade 12 Percentage of Students	High School Percentage of Students
0	0.0	0.0	0.1	0.1	0.1	0.1
1	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0
3	0.1	0.1	0.3	0.2	0.2	0.2
4	0.0	0.0	0.0	0.0	0.0	0.0
5	0.1	0.0	0.4	0.2	0.2	0.2
6	0.1	0.2	0.5	0.4	0.5	0.5
7	0.0	0.0	0.0	0.0	0.0	0.0
8	0.4	0.0	0.7	0.4	0.5	0.5
9	0.2	0.6	1.6	1.1	1.2	1.2
10	0.1	0.0	0.7	0.4	0.5	0.5
11	1.2	0.5	1.3	1.0	1.2	1.1
12	0.0	0.9	2.1	1.3	1.6	1.6
13	0.3	0.0	1.9	1.3	1.6	1.5
14	1.8	0.9	1.6	1.0	1.2	1.1
15	0.2	1.6	5.0	3.1	3.6	3.5
16	1.6	0.0	1.1	0.7	0.8	0.8
17	1.0	1.5	1.6	1.4	1.5	1.5
18	0.6	2.4	6.8	4.6	5.2	5.1
19	2.2	0.0	0.0	0.0	0.0	0.0
20	0.7	2.1	2.4	1.6	1.7	1.7
21	0.8	3.1	6.7	5.3	5.8	5.7
22	2.5	0.0	0.0	0.0	0.0	0.0
23	1.1	2.7	4.7	3.7	3.8	3.8
24	2.6	3.5	4.2	3.9	3.9	3.9
25	0.6	0.0	0.0	0.0	0.0	0.0
26	1.9	6.9	9.2	7.7	7.7	7.7
27	2.2	0.0	0.0	0.0	0.0	0.0
28	1.2	0.0	2.9	2.0	2.1	2.1
29	1.5	7.1	6.0	5.5	5.3	5.3

Percent Correct	Grade 5 Percentage of Students	Grade 8 Percentage of Students	Grade 10 Percentage of Students	Grade 11 Percentage of Students	Grade 12 Percentage of Students	High School Percentage of Students
30	2.1	0.0	0.0	0.0	0.0	0.0
31	2.0	3.4	3.9	3.6	3.5	3.5
32	2.9	3.6	3.6	3.6	3.4	3.4
33	1.3	0.0	2.1	2.0	2.0	2.0
34	1.5	3.3	2.1	2.4	2.2	2.3
35	2.0	3.3	1.8	2.3	2.1	2.1
36	1.3	0.0	2.0	2.1	1.9	2.0
37	1.5	3.2	1.9	2.1	2.0	2.0
38	2.6	3.0	3.5	4.0	3.6	3.7
39	1.5	0.0	0.7	1.0	0.9	0.9
40	0.7	3.0	0.7	1.0	0.8	0.9
41	2.6	2.8	2.6	3.4	3.2	3.2
42	1.5	0.0	0.7	0.9	0.8	0.8
43	2.6	2.8	0.5	0.8	0.7	0.7
44	1.4	2.6	2.1	3.0	2.7	2.8
45	0.8	0.0	0.7	0.8	0.7	0.7
46	3.2	2.6	1.4	2.2	1.9	2.0
47	1.5	2.4	1.1	1.8	1.7	1.7
48	0.0	0.0	0.0	0.0	0.0	0.0
49	3.1	2.3	1.2	1.9	1.7	1.8
50	1.4	2.2	0.6	1.5	1.4	1.4
51	3.0	2.1	0.9	1.6	1.5	1.5
52	0.0	0.0	0.0	0.0	0.0	0.0
53	1.4	2.1	0.5	1.2	1.1	1.1
54	2.8	1.8	0.7	1.4	1.2	1.3
55	0.7	0.0	0.2	0.5	0.4	0.4
56	1.3	1.9	0.7	1.4	1.3	1.3
57	2.1	1.6	0.1	0.2	0.2	0.2
58	1.2	0.0	0.2	0.4	0.4	0.4
59	2.0	1.7	0.4	1.1	1.0	1.0
60	0.5	1.4	0.0	0.2	0.2	0.2
61	1.2	0.0	0.1	0.3	0.3	0.3
62	1.9	1.6	0.4	0.8	0.8	0.8

Percent Correct	Grade 5 Percentage of Students	Grade 8 Percentage of Students	Grade 10 Percentage of Students	Grade 11 Percentage of Students	Grade 12 Percentage of Students	High School Percentage of Students
63	1.0	1.2	0.1	0.4	0.4	0.4
64	1.1	0.0	0.1	0.5	0.5	0.5
65	1.1	1.5	0.0	0.2	0.2	0.2
66	0.9	1.0	0.0	0.3	0.3	0.3
67	1.1	0.0	0.1	0.4	0.4	0.4
68	1.4	1.3	0.0	0.3	0.3	0.3
69	1.4	0.8	0.0	0.4	0.3	0.3
70	0.9	0.0	0.0	0.0	0.0	0.0
71	0.7	1.6	0.0	0.2	0.3	0.3
72	0.9	0.0	0.1	0.2	0.2	0.2
73	0.8	0.0	0.0	0.0	0.0	0.0
74	1.0	1.3	0.0	0.3	0.4	0.4
75	0.4	0.0	0.0	0.0	0.0	0.0
76	1.0	0.6	0.0	0.1	0.1	0.1
77	0.6	0.3	0.0	0.1	0.1	0.1
78	0.9	0.0	0.0	0.0	0.0	0.0
79	0.5	0.5	0.0	0.1	0.1	0.1
80	0.2	0.2	0.0	0.0	0.0	0.0
81	0.8	0.0	0.0	0.0	0.0	0.0
82	0.4	0.3	0.0	0.1	0.1	0.1
83	0.3	0.1	0.0	0.0	0.0	0.0
84	0.5	0.0	0.0	0.0	0.0	0.0
85	0.2	0.2	0.0	0.0	0.0	0.0
86	0.5	0.1	0.0	0.0	0.0	0.0
87	0.2	0.0	0.0	0.0	0.0	0.0
88	0.0	0.1	0.0	0.0	0.0	0.0
89	0.4	0.0	0.0	0.0	0.0	0.0
90	0.1	0.0	0.0	0.0	0.0	0.0
91	0.0	0.1	0.0	0.0	0.0	0.0
92	0.2	0.0	0.0	0.0	0.0	0.0
93	0.0	0.0	0.0	0.0	0.0	0.0
94	0.0	0.0	0.0	0.0	0.0	0.0
95	0.1	0.0	0.0	0.0	0.0	0.0

Percent Correct	Grade 5 Percentage of Students	Grade 8 Percentage of Students	Grade 10 Percentage of Students	Grade 11 Percentage of Students	Grade 12 Percentage of Students	High School Percentage of Students
96	0.0	0.0	0.0	0.0	0.0	0.0
97	0.0	0.0	0.0	0.0	0.0	0.0
98	0.0	0.0	0.0	0.0	0.0	0.0
99	0.0	0.0	0.0	0.0	0.0	0.0
100	0.0	0.0	0.0	0.0	0.0	0.0

Appendix 7.B: Indicator Distribution

Table 7.B.1 Percent of Students in Each Indicator Category by Demographic Variables for Grade Five

Group	Indicator Category 1	Indicator Category 2	Indicator Category 3
All students	4.5	95.0	0.5
Male	5.0	94.4	0.6
Female	3.9	95.7	0.4
English learner	10.6	89.4	0.0
English only	3.6	95.8	0.6
Reclassified fluent English proficient	1.4	98.2	0.3
Initially fluent English proficient	1.1	96.7	2.3
To be determined	18.2	81.8	0.0
Economically disadvantaged	6.1	93.8	0.1
Not economically disadvantaged	1.8	97.0	1.2
American Indian or Alaska Native	5.1	94.5	0.4
Asian	1.7	96.4	2.0
Native Hawaiian or Other Pacific Islander	5.9	94.1	0.0
Filipino	1.6	97.9	0.5
Hispanic or Latino	5.7	94.2	0.1
Black or African American	8.8	91.1	0.1
White	2.2	96.9	1.0
Two or more races	2.3	96.7	1.0
Special education services	12.9	86.9	0.2
No special education services	3.3	96.1	0.6
Migrant	8.6	91.4	0.0
Nonmigrant	4.4	95.1	0.5

Table 7.B.2 Percent of Students in Each Indicator Category by Demographic Variables for Grade Eight

Group	Indicator Category 1	Indicator Category 2	Indicator Category 3
All students	8.8	91.1	0.1
Male	9.9	90.0	0.1
Female	7.7	92.3	0.0
English learner	22.5	77.5	0.0
English only	7.4	92.5	0.1
Reclassified fluent English proficient	6.6	93.3	0.0
Initially fluent English proficient	3.4	96.3	0.3
To be determined	30.2	69.8	0.0
Economically disadvantaged	11.8	88.2	0.0
Not economically disadvantaged	4.3	95.6	0.1
American Indian or Alaska Native	11.7	88.3	0.0
Asian	2.8	96.8	0.3
Native Hawaiian or Other Pacific Islander	11.7	88.3	0.0
Filipino	3.3	96.7	0.1
Hispanic or Latino	11.3	88.7	0.0
Black or African American	15.6	84.4	0.0
White	4.8	95.1	0.1
Two or more races	5.3	94.6	0.1
Special education services	22.1	77.9	0.0
No special education services	7.2	92.7	0.1
Migrant	13.5	86.5	0.0
Nonmigrant	8.8	91.2	0.1

Table 7.B.3 Percent of Students in Each Indicator Category by Demographic Variables for Grade Ten

Group	Indicator Category 1	Indicator Category 2	Indicator Category 3
All students	11.5	88.5	0.0
Male	12.6	87.4	0.0
Female	10.2	89.8	0.0
English learner	24.8	75.2	0.0
English only	10.9	89.1	0.0
Reclassified fluent English proficient	7.0	93.0	0.0
Initially fluent English proficient	5.0	95.0	0.0
To be determined	0.0	100.0	0.0
Economically disadvantaged	12.9	87.1	0.0
Not economically disadvantaged	7.9	92.1	0.0
American Indian or Alaska Native	12.5	87.5	0.0
Asian	10.0	90.0	0.0
Native Hawaiian or Other Pacific Islander	12.3	87.7	0.0
Filipino	5.0	95.0	0.0
Hispanic or Latino	12.6	87.4	0.0
Black or African American	16.1	83.9	0.0
White	8.2	91.8	0.0
Two or more races	11.1	88.9	0.0
Special education services	25.3	74.7	0.0
No special education services	9.1	90.9	0.0
Migrant	24.2	75.8	0.0
Nonmigrant	11.3	88.7	0.0

Table 7.B.4 Percent of Students in Each Indicator Category by Demographic Variables for Grade Eleven

Group	Indicator Category 1	Indicator Category 2	Indicator Category 3
All students	7.5	92.4	0.0
Male	8.8	91.1	0.0
Female	6.2	93.8	0.0
English learner	20.9	79.1	0.0
English only	6.6	93.4	0.0
Reclassified fluent English proficient	6.2	93.8	0.0
Initially fluent English proficient	4.2	95.8	0.0
To be determined	7.7	92.3	0.0
Economically disadvantaged	9.4	90.6	0.0
Not economically disadvantaged	4.8	95.2	0.0
American Indian or Alaska Native	8.6	91.4	0.0
Asian	2.8	97.1	0.1
Native Hawaiian or Other Pacific Islander	5.7	94.3	0.0
Filipino	3.8	96.2	0.0
Hispanic or Latino	9.1	90.9	0.0
Black or African American	12.9	87.1	0.0
White	5.0	94.9	0.0
Two or more races	5.2	94.8	0.0
Special education services	20.3	79.7	0.0
No special education services	6.3	93.7	0.0
Migrant	15.1	84.9	0.0
Nonmigrant	7.5	92.5	0.0

Table 7.B.5 Percent of Students in Each Indicator Category by Demographic Variables for Grade Twelve

Group	Indicator Category 1	Indicator Category 2	Indicator Category 3
All students	9.0	91.0	0.0
Male	10.1	89.9	0.0
Female	7.9	92.1	0.0
English learner	21.9	78.1	0.0
English only	8.2	91.7	0.0
Reclassified fluent English proficient	8.0	92.0	0.0
Initially fluent English proficient	5.6	94.3	0.1
To be determined	19.8	80.2	0.0
Economically disadvantaged	10.9	89.1	0.0
Not economically disadvantaged	6.4	93.5	0.1
American Indian or Alaska Native	9.8	90.2	0.0
Asian	3.8	96.0	0.1
Native Hawaiian or Other Pacific Islander	11.2	88.8	0.0
Filipino	4.8	95.2	0.0
Hispanic or Latino	10.8	89.2	0.0
Black or African American	15.5	84.5	0.0
White	6.5	93.5	0.0
Two or more races	7.0	93.0	0.0
Special education services	21.5	78.5	0.0
No special education services	7.8	92.2	0.0
Migrant	11.3	88.7	0.0
Nonmigrant	9.0	91.0	0.0

Table 7.B.6 Percent of Students in Each Indicator Category by Demographic Variables for High School

Group	Indicator Category 1	Indicator Category 2	Indicator Category 3
All students	8.7	91.3	0.0
Male	9.8	90.2	0.0
Female	7.5	92.5	0.0
English learner	21.7	78.3	0.0
English only	7.9	92.1	0.0
Reclassified fluent English proficient	7.5	92.5	0.0
Initially fluent English proficient	5.3	94.6	0.1
To be determined	15.2	84.8	0.0
Economically disadvantaged	10.6	89.4	0.0
Not economically disadvantaged	6.0	93.9	0.0
American Indian or Alaska Native	9.6	90.4	0.0
Asian	3.6	96.3	0.1
Native Hawaiian or Other Pacific Islander	10.0	90.0	0.0
Filipino	4.6	95.4	0.0
Hispanic or Latino	10.4	89.6	0.0
Black or African American	14.9	85.1	0.0
White	6.2	93.8	0.0
Two or more races	6.6	93.3	0.0
Special education services	21.3	78.7	0.0
No special education services	7.4	92.5	0.0
Migrant	12.3	87.7	0.0
Nonmigrant	8.6	91.3	0.0

Chapter 8: Quality Control

The California Department of Education (CDE) and Educational Testing Service (ETS) implemented rigorous quality control procedures throughout the test development, administration, scoring, analyses, and completion of the technical report for the California Science Test (CAST) field test. As part of this effort, ETS staff worked with its Office of Professional Standards Compliance, which publishes and maintains the *ETS Standards for Quality and Fairness* (ETS, 2014).These *Standards* support the goal of delivering technically sound, fair, and useful products and services; and assisting the public and auditors to evaluate those products and services. This chapter highlights the quality control processes used at various stages of administration.

8.1. Quality Control of Test Materials

8.1.1. Developing Test Administration Instructions

ETS staff consult with internal subject matter experts and conduct validation checks to verify that test instructions accurately match the testing processes. Copy editors and content editors review each document for spelling, grammar, accuracy, and adherence to CDE style and usage requirements as well as the CDE accessibility standards. CAST content is incorporated to fit the California Assessment of Student Performance and Progress (CAASPP) System specifications. All CAASPP documents are approved by the CDE before they can be published to the CAASPP Portal at http://www.caaspp.org/. Only nonsecure documents are posted to this website.

8.1.2. Processing Test Materials

Online tests that were submitted by students were transmitted from the American Institutes for Research (AIR) to ETS each day. Each system checked for the completeness of the student record and stopped records that were identified as having an error.

Test responses were sent for human scoring and the reader's ratings were delivered to ETS scoring systems for merging with machine-scored items, final scoring, and scoring quality checks.

8.2. Quality Control of Item Development

ETS' goal is to provide the best standards-based and innovative items for the CAST. Items developed for the CAST field test were subject to an extensive item-review process. The item writers responsible for developing CAST items and performance tasks (PTs) were trained in CAASPP and ETS policies on quality control of item content, sensitivity, and bias guidelines, as well as guidelines for accessibility to ensure that the items allow the widest possible range of students to demonstrate their content knowledge.

Once a written item is accepted for authoring—that is, once it has been entered into ETS' item bank and formatted for use in an assessment—ETS employs a series of internal and external reviews. These reviews use established criteria and specifications to judge the quality of item content and ensure that each item measures what it is intended to measure. These reviews also examine the overall quality of the test items before presentation to the CDE and item reviewers. To finish the process, a group of California educators review the items and PTs for accessibility, bias and sensitivity, and content, and make recommendations for item enhancement. The details on quality control of item development are described in subsection *3.4 Item-Review Process*.

When student response data on each item became available, ETS Psychometric Analysis and Research (PAR) staff conducted item analysis and a key check to examine whether the items performed as expected. When the CAST field test was completed and the population data was available, psychometric staff conducted a thorough item analysis and evaluated all items carefully using the statistical criteria described in subsection *6.2.6 Summary of Classical Item Analyses Flagging Criteria* to flag items that were potentially problematic due to poor item performance, content issues, item bias, or accessibility challenges. After that, a data-review process was implemented, where a group of California educators and ETS content staff reviewed the items and PTs, together with their associated statistical results, and made recommendations about item disposition.

8.3. Quality Control of Test Form Development

ETS conducted multiple levels of quality assurance checks on each assembled field test form to ensure it met the form-building specifications. Both ETS assessment development and psychometric staff reviewed and signed off on the accuracy of test forms before the forms were put into production for administration in the field test. Detailed information related to test assembly can be found in subsection *3.7 Test Assembly*.

In particular, the assembly of all test forms went through a certification process that included various checks including verifying that:

- all answers are correct,
- answers are scored correctly in the item bank,
- all items match the standard,
- all content in the item is correct,
- all items meet the statistical criteria,
- distractors are plausible,
- multiple-choice item options are parallel in structure,
- language is grade-level appropriate,
- no more than three multiple-choice items in a row have the same key,
- all art is correct,
- there are no mechanical errors in grammar, spelling, punctuation, and the like, and
- items adhere to the approved style guide.

Reviews were also conducted for functionality and sequencing during the user acceptance testing (UAT) process to ensure all items functioned as expected.

8.4. Quality Control of Test Administration

The quality of test administration for the CAST, and all assessments administered as part of the CAASPP System, was monitored and controlled through several strategies. A fully staffed support center, the California Technical Assistance Center (CalTAC), supports all local educational agencies (LEAs) in the administration of CAASPP assessments. In addition to providing guidance and answering questions, CalTAC regularly conducts outreach campaigns on particular administration topics to ensure all LEAs understand correct test administration procedures. CalTAC is guided by a core group of LEA Outreach Advocacy staff that manage communications to LEAs; provide regional and web-based trainings; and host a website, http://www.caaspp.org/, that houses a full range of manuals, videos, and other instructional and support materials.

The quality of test administration was further managed through comprehensive rules and guidelines for maintaining the security and standardization of CAASPP assessments, including the CAST field test. LEAs received training on these topics and were provided tools for reporting security incidents and resolving testing discrepancies for specific testing sessions.

The ETS Office of Testing Integrity (OTI) reinforced the quality control procedures for test administration, providing quality assurance services for all testing programs managed by ETS. The detailed procedures OTI developed and applied in quality control are described in subsection *4.6.1 ETS' Office of Testing Integrity (OTI)*.

8.5. Quality Control of Scoring and Reporting

8.5.1. Development of Scoring Specifications

A number of measures are taken to ascertain that the scoring keys are applied to the student responses as intended and the student scores are computed accurately. ETS builds and reviews the scoring system models based on the reporting specifications approved by the CDE. These specifications contain detailed scoring procedures, along with the procedures for determining whether a student has attempted a test and whether that student's response data should be included in the statistical analyses and calculations for computing summary data.

Prior to the test administration, ETS Assessment Development (AD) staff reviewed and verified the keys and scoring rubrics for each item. Then, these keys and rubrics were provided to AIR for implementing machine scoring of the selected response items. Human-scored item responses are sent electronically to the ETS Online Network for Evaluation for scoring by trained, qualified raters. In addition, the student's original response string is stored for data verification and auditing purposes. Standard quality inspections are performed on all data files, including the evaluation of each student data record for correctness and completeness. Student results are kept confidential and secure at all times.

8.5.2. Quality Control of Machine-Scoring Procedures

The American Institutes for Research (AIR), the CAASPP subcontractor, provided the test delivery system (TDS) and scored machine-scorable items. A real-time, quality-monitoring component was built into the TDS. After a test was administered to a student, the TDS passed the resulting data to the quality assurance (QA) system. QA conducts a series of data integrity checks, ensuring, for example, that the record for each test contained information for each item, keys for multiple-choice items, score points in each item, and the total number of operational items. In addition, QA also checks to ensure that the test record contains no data from items that might have been invalidated.

Data passes directly from the Quality Monitoring System to the Database of Record, which served as the repository for all test information, and from which all test information is pulled and transmitted to ETS in a predetermined results format.

8.5.3. Quality Control of Human Scoring

For human scoring, ETS employed multiple quality controls including

- raters being required to successfully pass calibration, described earlier in subsection *5.1.5.2. Training for Raters*, prior to beginning scoring at each grade level;
- scoring leaders conducting backreads during each scoring shift;

- review of statistics on validity papers; and
- review of interrater reliability statistics.

Refer to subsection *5.1 Human Scoring for Constructed-Response Items* for the topics *5.1.6 Scoring Monitoring and Quality Management*, *5.1.7 Interrater Reliability*, and *5.1.8 Validity Responses and Sets* for more specific details on these tools used for quality control of human scoring.

8.5.4. Enterprise Score Key Management System (eSKM) Processing

Prior to the start of the test administration, test-level scores are defined in a scoring model configured in ETS' Enterprise Score Key Management (eSKM) system.

After the administration starts, and after AIR completes machine scoring, item scores and responses are delivered to ETS. ETS' Centralized Repository Distribution System and Enterprise Service Bus departments collect and parse .xml files that contained student response data from AIR. The eSKM system collects and calculates individual students' overall scores (total raw scores from machine-scored items and the human-scored items) and generates student scores in the approved statistical extract format. The data extracts are sent to ETS' Data Quality Services for data validation.

Following successful validation, the student response statistical extracts are made available to the psychometric team. The eSKM system implements scoring procedures specified by the psychometric team.

8.5.5. Psychometric Processing

Prior to the administration, the ETS psychometric team verifies the score calculation is accurate by both reviewing the configuration setup and using the UAT data. When the operational data arrives, eSKM receives the individual students' item scores and item responses from AIR and calculates individual student scores for ETS' reporting systems. The psychometric team also computes individual student scores based on item scores delivered by AIR.

The scores from the two sources are then compared for internal quality control. Any differences in the scores are discussed and resolved. All scores are complied with the ETS scoring specifications and the parallel scoring process to ensure the quality and accuracy of scoring and to support the transfer of scores into the database of the student records scoring system, the Test Operations Management System.

8.6. Quality Control of Psychometric Processes

8.6.1. Development of Psychometric Specifications

The psychometric procedures for the field test were developed, reviewed, and approved prior to the receipt of student response data. The ETS psychometric team also developed specifications for each of the psychometric analyses performed. These specifications contain detailed descriptions of the analyses steps such as sample inclusion, analyses methods, and special handling of the data.

8.6.2. Quality Control for Psychometric Analyses

All psychometric analyses conducted at ETS undergo comprehensive quality checks by a team of psychometricians and data analysts. Detailed checklists and psychometric specifications are developed by members of the team for each of the statistical procedures performed on CAST results data. The classical item analyses and differential item functioning analyses are run and confirmed by independent analysts. Results are then reviewed by the psychometricians to compile a list of flagged items for ETS AD staff to review. Reviewer comments are checked by the psychometricians before the data review meetings with the CDE. The ETS AD and PAR teams worked together to evaluate and make recommendations to the CDE about any problematic items that should be removed from scoring and reporting.

References

Educational Testing Service. (2014). *ETS standards for quality and fairness*. Princeton, NJ: Educational Testing Service. Retrieved from https://bit.ly/3jZIPOK

Chapter 9: Surveys

This chapter describes the development and administration of the survey questionnaires presented to test administrators and students during the 2017–18 California Science Test (CAST) field test administration. The summary of findings and results of analyses from the survey are included.

9.1. Test Administrator Survey

The responses to the test administrator survey provided additional insight into the student test-taking experience of the CAST field test. The feedback from the survey will help in the development and administration of the CAST field tests and operational tests.

The test administrator survey was developed by Educational Testing Service (ETS) in consultation with the California Department of Education (CDE). The CDE provided guidelines in terms of the length of the survey and the number and focus of the questions. The survey questions used during the administration and the response frequencies are included in appendix 9.A.

The test administrators completed their survey via SurveyGizmo, an online survey software tool. The test administrator survey was the same regardless of the grade level monitored.

9.1.1. High School Student Survey

A student survey question was administered as a final section at the end of each field test in high school. It was available in braille for students who needed this accommodation. The question asked was, “Do you think you will be enrolling in any more science classes in high school?”

Table 9.1 provides summary results of the high school student survey. It shows that 75.3 percent of students responded “No.” Note that for the high school assessment, the student survey was self-reported and the results were different from the eligibility the LEA determines. The statistical analyses conducted on the field test data were based on the LEA eligibility standards instead of the self-reported status in the student survey.

Table 9.1 Student Survey Summary

Grade	Answer	Number of Students	Percent
10	No Response	116	1.8
10	Yes	3,952	61.9
10	No	2,316	36.3
11	No Response	2,844	2.1
11	Yes	74,560	54.6
11	No	59,271	43.4
12	No Response	8,562	2.1
12	Yes	45,234	11.2
12	No	351,255	86.7
All Grades	No Response	11,522	2.1
All Grades	Yes	123,746	22.6
All Grades	No	412,842	75.3

Appendix 9.A: Survey Results

Note that the sum of percentages in Table 9.A.1 may exceed 100 percent because a test administrator could select more than one grade as applicable to a survey question.

Table 9.A.1 Distribution of Test Administrator Responses to Question 1

Were you a test administrator for the CAST field test?	N = 4,105 test administrators
a. Yes	92%
b. No	8%

Table 9.A.2 Distribution of Test Administrator Responses to Question 2

For what grade(s) did you administer the test? Select all that apply.	N = 3,771 test administrators
a. Grade five	52%
b. Grade eight	28%
c. Grade ten	1%
d. Grade eleven	13%
e. Grade twelve	21%

Table 9.A.3 Distribution of Test Administrator Responses to Question 3

For the students to whom you administered the CAST, are you the students' science teacher?	N = 3,771 test administrators
a. Yes, I am the teacher for all of the students tested this year.	42%
b. Yes, I am the teacher for some of the students tested this year.	16%
c. No, I was not the science teacher for any of the students tested this year.	42%

Table 9.A.4 Distribution of Test Administrator Responses to Question 4

To your knowledge, did the students you tested have an opportunity to take the CAST training test prior to taking the field test?	N = 3,775 test administrators
a. All or most of the students took the training test.	16%
b. More than half of the students took the training test.	14%
c. Less than half of the students took the training test.	11%
d. Few or none of the students took the training test.	30%
e. I don't know.	29%

Table 9.A.5 Distribution of Test Administrator Responses to Question 5

Which of the following statements best describes the instructions provided in Chapter 6 “Administering the Summative Assessments to Students” of the Online Test Administration Manual, found on the caaspp.org Web site?	**N = 3,713 test administrators**
a. The instructions were very clear.	54%
b. The instructions were somewhat clear.	37%
c. The instructions were somewhat confusing.	7%
d. The instructions were very confusing.	2%

Table 9.A.6 Distribution of Test Administrator Responses to Question 6

Which of the following statements best describes the instructions provided within the test delivery system to students for the pilot?	**N =3,743 test administrators**
a. The students appeared to find the instructions very clear.	38%
b. The students appeared to find the instructions somewhat clear.	43%
c. The students appeared to find the instructions somewhat confusing.	16%
d. The students appeared to find the instructions very confusing.	3%

Table 9.A.7 Distribution of Test Administrator Responses to Question 7

Which of the following statements best describes your students’ engagement with the field test?	**N =3,750 test administrators**
a. All or most of my students appeared to be fully engaged with the field test.	15%
b. More than half of my students appeared to be fully engaged with the field test.	52%
c. Less than half of my students appeared to be fully engaged with the field test.	30%
d. Few or none of my students were engaged with the field test.	3%

Table 9.A.8 Distribution of Test Administrator Responses to Question 8

Did a majority of your students complete the CAST in one test session?	**N =3,766 test administrators**
a. Yes	45%
b. No	55%

Table 9.A.9 Distribution of Test Administrator Responses to Question 9

Typically how many test sessions did it take for your students to complete the CAST?	**N =2,051 test administrators**
a. 2 sessions	54%
b. 3 sessions	32%
c. 4 or more sessions	14%

Table 9.A.10 Distribution of Test Administrator Responses to Question 10

Which of the following statements best describes your students' interaction with the computer interface for the assessment?	N =3,768 test administrators
a. All or most of my students appeared to be able to easily navigate through the online assessment.	32%
b. More than half of my students appeared to be able to easily navigate through the online assessment.	50%
c. Less than half of my students appeared to be able to easily navigate through the online assessment.	16%
d. Few or none of my students appeared to be able to easily navigate through the online assessment.	2%

Chapter 10: Continuous Improvement

The California Science Test (CAST) field test was offered during the 2017–18 school year. Since the inception of the CAST, continuous efforts have been made to improve the grade-level assessments in various ways. This chapter summarizes the current and ongoing improvements for the CAST in the areas of test design, item development, test delivery and administration, psychometric analyses, and accessibility.

10.1. Test Design

Educational Testing Service (ETS) works in collaboration with the California Department of Education in planning, proposing, evaluating, and improving CAST test design.

The operational test form will be delivered in a similar manner as the CAST field test. In planning for the 2018–19 operational administration, ETS will follow the test blueprint and high-level test design closely to provide a testing experience much like that offered during the field test. Unlike the pilot, which was focused on testing item functionality and content, the field test was focused on preparing for the operational test.

10.2. Item Development

For the 2017–18 item development cycle, the ETS content teams used item specifications that make the alignment of all three dimensions of the California Next Generation Science Standards (CA NGSS)—disciplinary core ideas, science and engineering practices, and crosscutting concepts—clearer on CAST items. The creation and modification of the item specifications has continued with the development of the operational assessments. The newest items, when compared to those items developed previously, feature significantly more integration of the aforementioned three dimensions. When these items were shared with teacher panelists prior to field-test administration, the feedback received was positive and enthusiastic.

Work to refresh the CAST item bank will continue through subsequent development cycles with the goal of developing items of low and medium complexity, along with items of high complexity, by the 2021–22 administration.

10.3. Administration and Test Delivery

10.3.1. Survey Results

The California Assessment of Student Performance and Progress (CAASPP) program annually solicits feedback from CAASPP stakeholders through the CAASPP Post-Test Survey. Local educational agency (LEA) and test site staff, as well as test administrators and test examiners, were invited to participate in the 2017–18 CAASPP Post-Test Survey. California educators provided specific, actionable insights about the 2017–18 testing experience.

Additionally, CAST-specific surveys were conducted for both students and test administrators. More information about their results can be found in *Chapter 9: Surveys*. Those results are being used to improve the CAST.

10.3.2. Test Delivery Improvements

Test delivery changed for the field test to match operational delivery. Instead of assigning entire high school grade levels for testing by school, as was done in the pilot, all students in grade twelve were assigned to test, with students in grades ten and eleven assigned at the discretion of the LEA. This matches the policy that will be applied to the operational test. With students in grade twelve receiving a default test assignment, the high school testing population was more representative of the population that will take the operational test, making any conclusions drawn from their performance more meaningful.

The training test was also updated during this administration, giving students the opportunity to interact with all of the unique item types available on the CAST. For the 2018–19 administration, the practice and training tests will be updated as well.

10.4. Psychometric Special Studies

ETS conducted three special studies based on field-test data collected in the 2017–18 administration. The goals were to inform the operational-year score reporting and provide initial evidence and support for continuous improvement on the test designs.

The first special study was the dimensionality study. It was designed to study test dimensionality with the intent of informing what scores are to be reported and how those scores will be reported. Details of the study can be found in subsection *6.4 Test Dimensionality Analyses*.

ETS also conducted a multistage adaptive test (MST) practicality study. Its purpose was to evaluate the extent to which the MST improves measurement of student ability in comparison to a linear Segment A form and whether the improvement is enough to offset the increased complexity and risk inherent in all adaptive testing. Details of the MST study can be found in subsection *6.8.1 Multistage Adaptive Test (MST) Practicality Study*.

Finally, ETS conducted a CAST screener study to collect preliminary evidence regarding the utility of using Segment A performance to screen out performance tasks in Segment B. This study is described in subsection *6.8.2 Content Screen-Out Study*.

ETS will replicate all three studies in 2019 with the expanded item pool and testing data collected during the 2018–19 operational administration. The results from those studies will help inform the test design in future years.

10.5. Accessibility

ETS increased the number of accessibility resources available to match the upcoming operational test. The resources listed in Table 10.1 are in addition to the full suite of universal tools that were available.

Table 10.1 Additional Accessibility Resources

Category	Embedded	Non-embedded
Designated Supports	• Color Contrast • Mouse Pointer • Stacked Translations and Translated Test Directions (Spanish) • Translation Glossaries • Turn Off Any Universal Tools	• Amplification • Read-Aloud in Spanish • Science Charts • Simplified Test Directions
Accommodations	• Closed Captioning • Streamline Audio Transcript	• Alternate Response Options • Word Prediction

New accessibility resources are planned for the operational test, such as Hmong as a translated test direction language. As for the items themselves, ETS has redoubled its focus on making items accessible from the outset, reducing the need to provide extensive adaptations to make the items accessible to students with visual impairments.

In addition to student accessibility resources, a new process has been implemented to allow for a review of item content by teachers of students with visual impairments. Feedback on adapted items from teachers of the visually impaired confirms some techniques and informs other techniques used to reduce the need for item adaptation in further development cycles.