# California Department of Education
# Assessment Development & Administration Division

caaspp
California Assessment of Student Performance and Progress

# California Assessment of Student Performance and Progress

# California Science Test 2018–2019 Technical Report

**Submitted July 31, 2020**

**Prepared for the California Department of Education by Educational Testing Service**

ETS

**Contract #CN150012**

# Table of Contents

## List of Appendices

# List of Tables

## List of Figures

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

| Term | Definition |
|---|---|
| 2PL | two-parameter logistic model |
| 3D | three dimensional |
| AERA | American Educational Research Association |
| AI | artificial intelligence |
| AIR | American Institutes for Research; now Cambium Assessment |
| AIS | average item score |
| ALTD | Assessment & Learning Technology Development |
| 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 |
| CDE | California Department of Education |
| CDS | county/district/school |
| CERS | California Educator Reporting System |
| CR | constructed response |
| CSEM | conditional standard error of measurement |
| DCIs | disciplinary core ideas |
| DIF | differential item functioning |
| DOK | depth of knowledge |
| EC | Education Code |
| ECD | Evidence-Centered Design |
| ECV | explained common variance |
| ELA | English language arts/literacy |
| eSKM | Enterprise Score Key Management |
| ETS | Educational Testing Service |
| FIA | final item analysis |
| GPCM | generalized partial credit model |
| HumRRO | Human Resources Research Organization |
| IEP | individualized education program |
| IMS | Instructional Management Systems |
| IRT | item response theory |

Table of Acronyms and Initialisms *(continuation)*

| Term | Definition |
|---|---|
| ISAAP Tool | Individual Student Assessment Accessibility Profile Tool |
| KSAs | knowledge, skills, and abilities |
| LEA | local educational agency |
| LOSS | lowest obtainable scale score |
| MC | multiple choice |
| MH DIF | Mantel-Haenszel Differential Item Functioning |
| MIRT | multidimensional item response theory |
| MST | multistage adaptive test |
| NCME | National Council on Measurement in Education |
| NR | number right |
| ONE | Online Network for Evaluation |
| ORS | Online Reporting System |
| OTI | Office of Testing Integrity |
| PAR | Psychometric Analysis and Research |
| PE | performance expectation |
| PIA | preliminary item analysis |
| PT | performance task |
| QA | quality assurance |
| QTI | Question and Test Interoperability |
| QWK | quadratic weighted kappa |
| SBE | State Board of Education |
| SD | standard deviation |
| SEM | standard error of measurement |
| SEPs | science and engineering practices |
| SFTP | secure file transfer protocol |
| SMD | standardized mean difference |
| SR | selected response |
| SRO | Scoring and Reporting Operations |
| 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

This chapter provides an overview of the California Science Test (CAST), including background information, the purpose of the test, the intended population, and organizations and systems involved.

## 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/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 CAST is an online assessment aligned with the CA NGSS. It was administered as a pilot for the first time during the 2016–2017 CAASPP administration, followed by a field test administration during the 2017–2018 CAASPP administration. The first operational CAST was administered during the 2018–2019 CAASPP administration. The assessment is for students in grades five and eight and high school.

In 2018–2019, the CAASPP System comprised the following assessments:

- 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 and that may be administered to students at any grade level
  - Digital Library—Professional development materials and instructional resources designed to help teachers use 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 and eight and high school (grades ten, eleven, or twelve), including the CAST and the CAA for Science
- The California Spanish Assessment, optional for eligible students in grades three through eight and high school and designed to measure a student's Spanish competency in reading, writing mechanics, and listening, as well as to serve as a high school measure suitable to be used in part for the California Seal of Biliteracy

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. Test Purpose

The purpose of the CAST was to assess students with federally required science assessments in grades five and eight and once in high school (i.e., grade ten, eleven, or twelve). The assessment was designed to assess the three dimensions (i.e., Science and Engineering Practices, Disciplinary Core Ideas, and Crosscutting Concepts) of the CA NGSS by using various item types, some of which involved the use of dynamic stimuli and other types of new media (e.g., animations of scientific phenomena, virtual engineering challenges, simulated experiments). As the first operational year of the assessment, this was the first time that Student Score Reports were generated (refer to *Chapter 7: Scoring and Reporting*).

## 1.3. Test Structure

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 (TDS) at each grade or grade level assigned students to two different item blocks in Segment A, with each containing 16–17 discrete items. Each test also contained two different PTs in Segment B, with each task presenting five to seven items. Finally, either one discrete item block (with 13 discrete items) or one PT (with seven to eight items) was selected in Segment C.

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 technology-enhanced items (refer to section *3.2 Item Development*).

A fixed braille form was available for students with visual impairments (refer to section *2.3 Test Administration*). The braille form was composed of two Segment A blocks totaling 32–34 items, two Segment B performance tasks of 12–13 items, and one Segment C block of 13 items.

Table 1.1 lists the total number of unique items per segment across all test forms. On the grade five form, seven discrete items were repeated among a few Segment C blocks. On the grade eight form, six discrete items were repeated among a few Segment C blocks. On the high school form, nine discrete items were repeated among a few Segment C blocks.

**Table 1.1 Number of Unique Items Assessed on the CAST**

| Segment | Grade 5 | Grade 8 | High School |
|---|---|---|---|
| A (2 blocks) | 34 | 32 | 34 |
| B (3–6 PTs) | 32 | 37 | 18 |
| C (15 discrete blocks) | 188 | 189 | 186 |
| C (5–6 PTs) | 46 | 39 | 38 |

## 1.4. Intended Population

The CAST operational assessment was administered to approximately 1.4 million students in the general population. The intended population was all students in grades five and eight as well as high school students in grade ten, eleven, or twelve who were assigned by their local educational agency (LEA) (refer to section *5.1 Student Test-Taking Requirement* for more details about the high school grade assignments).

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

## 1.5. Intended Use and Purpose of Test Scores

The results of tests within the CAASPP System are used for two primary purposes as described in *Education Code (EC)* sections 60602.5(a) and (a)(4). (Excerpted from the *EC* Section 60602 web page at http://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?lawCode=EDC&division=4.&title=2.&part=33.&chapter=5.&article=1 [outside source].)

> "60602.5(a) It is the intent of the Legislature in enacting this chapter to provide a system of assessments of pupils that has the primary purposes of assisting teachers, administrators, and pupils and their parents; improving teaching and learning; and promoting high-quality teaching and learning using a variety of assessment approaches and item types. The assessments, where applicable and valid, will produce scores that can be aggregated and disaggregated for the purpose of holding schools and local educational agencies accountable for the achievement of all their pupils in learning the California Next Generation Science Standards."
>
> "60602.5(a)(4) Provide information to pupils, parents and guardians, teachers, schools, and local educational agencies on a timely basis so that the information can be used to further the development of the pupil and to improve the educational program."

In other words, results for tests within the CAASPP System are used for two primary purposes:

1. To communicate students' progress in achieving the state's academic standards to students, parents and guardians, and teachers
2. To inform decisions that teachers and administrators make about improving the educational program

Sections 60602.5(c) and (d) provide additional information regarding use and purpose of test scores for the system of assessments:

> "60602.5(c) It is the intent of the Legislature that parents, classroom teachers, other educators, pupil representatives, institutions of higher education, business community members, and the public be involved, in an active and ongoing basis, in the design and implementation of the statewide pupil assessment system and the development of assessment instruments."
>
> "60602.5(d) It is the intent of the Legislature, insofar as is practically feasible and following the completion of annual testing, that the content, test structure, and test items in the assessments that are part of the statewide pupil assessment system become open and transparent to teachers, parents, and pupils, to assist stakeholders in working together to demonstrate improvement in pupil academic achievement. A planned change in annual

test content, format, or design should be made available to educators and the public well before the beginning of the school year in which the change will be implemented."

# 1.6. Testing Window

The CAST was administered during a testing window selected by the LEA, with the first possible date of administration being January 14, 2019, and the last possible date being July 15, 2019 (*California Code of Regulations,* Title 5 [5 *CCR*]*,* Education, Division 1, Chapter 2, Subchapter 3.75, Article 2*,* Section 855[a][2]). The testing window, which is open for a maximum of 12 weeks for each LEA, began on the day in which 66 percent of the instructional year was completed.

Like other CAASPP assessments, the CAST was untimed for students. A student could take the CAST 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.

# 1.7. Significant Developments in 2018–2019

## 1.7.1. Operational Assessment

The CAST was administered operationally for the first time during the 2018–2019 CAASPP administration.

## 1.7.2. Updated Accessibility Resources

The following changes were made to the list of CAST accessibility resources:

- Streamline was reassigned as an embedded designated support.
- "Medical device" was added as a new non-embedded designated support for all assessments.
- The Highlighter universal tool was made available in four colors.
- Scratch paper included the use of non-embedded digital graph paper.
- Hmong was added as an embedded translation glossary available as a designated support.
- A braille version was available.

# 1.8. Groups and Organizations Involved with the CAST

## 1.8.1. 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 *EC*.

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, 2020a).

## 1.8.2. California Department of Education (CDE)

The CDE oversees California's public school system, which is responsible for the education of more than 6,100,000 children and young adults in more than 10,500 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 Instruction & Measurement branch that oversees programs promoting improved student achievement. Programs include oversight of statewide assessments and the collection and reporting of educational data (CDE, 2020b).

## 1.8.3. California Educators

A variety of California educators, including teachers and school administrators—who were selected based on their qualifications, experiences, demographics, and geographic locations—were invited to participate in the various aspects of the assessment process. This included defining the purpose and scope, test design, item development, standard setting, score reporting, and scoring of the CR items of the CAST.

## 1.8.4. Contractors

### 1.8.4.1. Educational Testing Service (ETS)

The CDE and the SBE contract with ETS to develop, administer, and report the CAST. 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 items
- Scoring CR items
- Constructing, producing, and controlling the quality of CAASPP test forms and related test materials
- Processing student test assignments

[1] Retrieved from the CDE Fingertip Facts on Education in California – *CalEdFacts* web page at https://www.cde.ca.gov/ds/sd/cb/ceffingertipfacts.asp

- Producing and distributing student score reports
- Completing all psychometric procedures
- Developing a summary score reporting website that can be viewed by the public

#### 1.8.4.2. American Institutes for Research (AIR)

ETS also monitors and manages the work of AIR (now Cambium Assessment), 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 (ORS), a component of the overall CAASPP Assessment Delivery System;
- scoring machine-scorable items; and
- providing Level 3 technology help desk support to LEAs.

## 1.9. Systems Overview and Functionality

### 1.9.1. 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 for the CAST, including but 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 CAASPP Security and Test Administration Incident Reporting System/Appeals process

TOMS receives student enrollment data and LEA and 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 CAASPP coordinators, CAASPP test site coordinators, test administrators, and test examiners—are assigned varying levels of access to TOMS. For example, only an LEA CAASPP 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 explained more extensively in the *2018–19 CAASPP Online Test Administration Manual* (CDE, 2019a).

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

### 1.9.2. Test Delivery System (TDS)

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

- the Test Administrator Interface, the web browser–based application that allows test administrators to activate student tests and monitor student testing;
- the Student Testing Interface, on which students take the test using the secure browser; and
- the 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.

### 1.9.3. Online Reporting System (ORS) and California Educator Reporting System (CERS)

Currently, there are two California online reporting systems: the ORS and the CERS. Over the next two years CERS will replace ORS as the single resource where LEA staff access student results from the summative and interim CAASPP assessments as well as results from the English Language Proficiency Assessments for California.

The ORS is the system used by LEAs to view preliminary student results from the CAASPP assessments. The primary purposes of the ORS are for LEAs to access completion data to determine which students need to complete testing or start testing, and for LEAs to access preliminary score reports that can provide data for schools within the LEA. Results in the ORS are preliminary and may not be used for accountability purposes.

The CERS allows educators to view their students' assessment results using grouping and other new features. For example, educators can create customized groups from assigned student groups; for interim assessments, specific assessment items can be viewed with student responses; and a distractor analysis feature can be used to identify student strengths and needs.

### 1.9.4. Practice and Training Tests

The publicly available practice and 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 introduce students to the type of thinking needed to answer CAST items; the practice tests simulate the operational testing experience. The practice and training tests align with performance expectations but do not produce scores. Test administrators may access scoring guides that describe related scoring considerations.

The purposes of the training and practice tests are to

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

Details on practice and training tests are presented in section *5.6 Practice and Training Tests*.

### 1.9.5. 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 7: Scoring and Reporting*.

For the CAST, 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 benchmark sets, training sets, scoring rubrics, and scoring notes. The quality control processes for CR scoring are explained further in *Chapter 9: Quality Control*.

The CR items could also be rated by artificial intelligence (AI) scoring engines (e.g., the *c-rater*™ system). The use of such engines often required models be built with reliable human-rating data. For the 2017–2018 administration, a data collection design was used to provide data to build and evaluate AI models for the field test CRs. For details on AI model building and evaluation on field test CRs, refer to the *California Science Test Field Test Technical Report 2017–2018 Administration* (CDE, 2019b).

During the first operational administration in 2018–2019, AI scoring was used to score responses for those CRs with approved AI models. A careful data collection design was also used to provide data to build the AI scoring engine for future use. The details of the CR sampling plan that supported the new AI model building is provided in subsection *7.1.1.2 Sampling Process for Field Test Constructed-Response Items*.

The c-rater™ engine is ETS' system for the automated scoring of content in text-based responses. The engine uses state-of-the-art machine learning technology to score items that elicit and measure knowledge about specific content. The engine computes a large set of linguistic features from each response that relate to the content focus of the item. This broad set of features extends beyond key words to capture grammatical relationships and mitigates the impact of spelling and grammatical variation on how the model assigns scores. The c-rater™ system is ideal for evaluating expected content for subject-matter CR questions in content areas, including social studies, science, ELA, and mathematics.

ETS' process required test designers to define the required content but did not ask them to predict every aspect of the form of student language. The *c-rater* engine filters out potential, not-scorable responses (e.g., responses in a language other than English, no-attempt responses such as "I don't know," etc.). Filtering was applied both during the AI--scoring model building step to ensure AI-scoring models were built on reliable data; and when the AI-scoring model was deployed, to ensure that such responses were filtered and scored correctly.

Any response that was entirely in a language other than English as detected by *c-rater* was given a specific advisory designation and handled following the policy established with the CDE to mark these responses as not scorable and return them to the Online Network for Evaluation with an advisory code to be human-scored. If the response was in Spanish, these responses were then reviewed by Spanish biliterate raters and scored according to the rubric. If the response was not in English or Spanish, the response received a zero score.

## 1.10. Overview of the Technical Report

This technical report addresses the characteristics of the CAST administered in spring 2019 and contains nine additional chapters as follows:

- Chapter 2 presents an overview of processes involved in a CAST testing cycle. This includes item development, test administration, psychometric analyses, generation of test scores, and dissemination of score reports. It also includes information about the assignment of designated supports and accommodations.
- Chapter 3 discusses the detailed procedures of item development for the CAST to help ensure valid interpretation of test scores.
- Chapter 4 discusses the content and psychometric criteria that guide procedures of CAST test assembly.
- Chapter 5 details the processes involved in the administration of the CAST. It also describes the procedures followed by ETS to maintain test security throughout the test administration process.
- Chapter 6 summarizes the standard setting that occurred after the 2018–2019 first operational test administration.
- Chapter 7 summarizes the types of scores and score reports that are produced at the end of each administration of the CAST.
- Chapter 8 summarizes the statistical procedures and results for 2018–2019. These include:
  - classical item analyses,
  - test completion rates and analyses,
  - differential item functioning analyses, and
  - IRT analyses.
- Chapter 9 highlights the quality control processes used at various stages of development and administration of the CAST.
- Chapter 10 describes the development and administration of the survey questionnaires for students and the results of analyses of their responses.
- Chapter 11 discusses the various procedures used to gather information to improve the CAST as well as strategies to implement possible improvements.
- Chapter 12 is an addendum that addresses CAST dimensionality.

# References

*California Code of Regulations,* Title 5*,* Education, Division 1, Chapter 2, Subchapter 3.75, Article 2, Section 855. Retrieved from ~~https://govt.westlaw.com/calregs/Document/I2DB6A0BAA54F41B69BAF5553FABBE5EF?viewType=FullText&originationContext=documenttoc&transitionType=CategoryPageItem&contextData=(sc.Default)~~

California Department of Education. (2019a). *CAASPP online test administration manual, 2018–19 administration.* Sacramento, CA: California Department of Education. Retrieved from ~~http://www.caaspp.org/rsc/pdfs/CAASPP.online_tam.2018-19.pdf~~

California Department of Education. (2019b). *California Science Test Field Test Technical Report 2017–18 Administration.* (Unpublished report). Sacramento, CA: California Department of Education.

California Department of Education. (2020b, August). *Organization.* Retrieved from http://www.cde.ca.gov/re/di/or/

California Department of Education. (2020a, October). *State Board of Education responsibilities.* Retrieved from http://www.cde.ca.gov/be/ms/po/sberesponsibilities.asp

# Chapter 2: An Overview of the Operational Test Process

This chapter provides a brief description of the California Science Test (CAST) operational test process including item development, test design, test administration, scoring, reporting, and psychometric analyses. The details on each step in the process are presented in subsequent chapters.

## 2.1. Item Development

CAST item development incorporates innovations and best practices from national science assessments. For the CAST, items and associated stimuli with featured simulations were developed that integrated the dimensions of the performance expectations (PEs) while maintaining appropriateness for test takers. California science teachers assisted in creating these items, and item review meetings with California educators were instrumental in determining both the proper integration of the PE dimensions and grade-level appropriateness.

### 2.1.1. Design Guidelines

Educational Testing Service (ETS) content specialists referred to design patterns and task templates as part of the incipient Evidence-Centered Design 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 2018–2019 CAST administration were appropriate for the grade level and aligned with the California Next Generation Science Standards (CA NGSS).

### 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 item writers and for the rigorous review process that was implemented for every item. Additionally, item specifications and the CA NGSS PEs were used to guide the writing of items for the CAST. Refer to section *3.2 Item Development* for the guidelines of item writing, including the item specifications.

ETS trained California science teachers to develop items for the CAST during an item writing workshop in November 2017 (refer to subsections *3.2.4 Selection of Item Writers* and *3.2.5 Item Writer Training*). California science teachers were instructed to produce items that spanned a variety of science and engineering practices 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 CAST forms construction.

### 2.1.3. Item Types Guidelines

The CAST was designed to assess the CA NGSS using discrete items, single and multipoint items, and performance tasks (PTs). There were a variety of item types, including traditional multiple-choice (MC) items, constructed-response (CR) items, some familiar technology-enhanced item (TEI) types, as well as some new TEI types that used simulations and animations. Refer to section *3.2 Item Development* for more details on the number of items developed and to section *3.1.4 Item Types and Features* for the types of items used in the CAST.

Greater emphasis was made to fill the CAST item bank with items that have students explore phenomena using item types that best fit the construct. 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—PTs—internally to measure more complex skills in a particular domain.

## 2.2. Test Assembly

The 2018–2019 CAST 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. The first two segments comprise the operational assessment; Segment C is for field-testing future items.

ETS designed the general CAST forms to be taken in approximately two hours and 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. *Chapter 4: Test Assembly* provides details about test assembly.

### 2.2.1. Test Blueprint

Blueprints represent a set of constraints and specifications to which each test form must conform. The CAST has three main subcontent areas or domains: Life Sciences, Physical Sciences, and Earth and Space Sciences. It consists of both discrete items and PTs. The blueprints of the assessment are shown in table 4.1.

For the 2018–2019 operational assessment, each student took 8 to 12 discrete items from each domain that were worth 12 to 18 points total, and two PTs that belong to different domains and were worth 12 to 14 points total.

## 2.3. Test Administration

The CAST is administered online using the secure browser and test delivery system, ensuring a secure, confidential, standardized, consistent, and appropriate administration for students. Additional information about the administration of the CAST can be found in *Chapter 5: Test Administration*.

### 2.3.1. Test Security and Confidentiality

All tests within the California Assessment of Student Performance and Progress (CAASPP) System are secure. For the CAST, 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 *5.7.1 ETS' Office of Testing Integrity (OTI)* in *Chapter 5: Test Administration*.

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 section *5.7 Test Security and Confidentiality*.

### 2.3.2. Procedures to Maintain Standardization

ETS takes all necessary measures to ensure the standardization of CAST administration. The measures for standardization include, but are not limited to, the aspects described in these subsections.

#### 2.3.2.1. Test Administrators

The CAST grade and grade-level assessments are administered in conjunction with the other assessments that compose the CAASPP System. ETS employs processes to ensure the standardization of an administration cycle; these processes are discussed in more detail in subsection *5.3 Procedures to Maintain Standardization*.

Staff at local educational agencies (LEAs) involved in CAST administration include LEA CAASPP coordinators, CAASPP test site coordinators, and test administrators. The responsibilities of each of the staff members are described in the *CAASPP Online Test Administration Manual* (California Department of Education [CDE], 2019a).

#### 2.3.2.2. Test Directions

Several series of instructions regarding the CAASPP administration are compiled in detailed manuals and provided to the LEA staff. Such documents include, but are not limited to, the following:

- ***CAASPP Online Test Administration Manual—***This is a manual that provides test administration procedures and guidelines for LEA CAASPP coordinators, and CAASPP test site coordinators, as well as the script and directions for administration to be followed exactly by test administrators during a testing session (CDE, 2019a). (Refer to *5.3.4.2 CAASPP Online Test Administration Manual* in chapter 5 for more information.)
- ***Test Operations Management System (TOMS) Pre-Administration Guide for CAASPP Testing—***This is a manual that provides instructions for TOMS allowing LEA staff, including LEA CAASPP coordinators and CAASPP test site coordinators, to perform a number of tasks including setting up test administrations, adding and managing users, assigning tests, and configuring online student test settings (CDE, 2018). (Refer to *5.3.4.3 TOMS Pre-Administration Guide for CAASPP Testing* in chapter 5 for more information.)

## 2.4. Universal Tools, Designated Supports, and Accommodations

All public school students participate in the CAASPP System of assessments, including students with disabilities and ELs. Additional resources are sometimes needed for these students. The CDE provides a full range of assessment resources for all students, including those who are ELs and students with disabilities. There are four different categories of student accessibility resources in the California assessment accessibility system, including universal tools, designated supports, accommodations, and unlisted resources that are permitted for use in CAASPP online assessments. These are listed in the CDE web

document, *Matrix One: Universal Tools, Designated Supports, and Accommodations for the CAASPP System* (CDE, 2019b).[3]

**Universal tools** are available to all students. These resources may be turned on and off when embedded as part of the technology platform for the online CAST assessments on the basis of student preference and selection.

**Designated supports** are available to all students when determined as needed by an educator or team of educators, with parent/guardian and student input as appropriate, or when specified in the student's individualized education program (IEP) or Section 504 plan.

**Accommodations** must be permitted for the online CAST assessments for all eligible students when specified in the student's IEP or Section 504 plan.

**Unlisted resources** are non-embedded and made available if specified in the eligible student's IEP or Section 504 plan and only on approval by the CDE.

Assignment of designated supports and accommodations to individual students based on student need is made in TOMS by the LEA CAASPP coordinator or CAASPP test site coordinator, either through individual assignment through the student's profile in TOMS; by uploading of settings for multiple students that were either selected and entered into a macro-enabled template called the Individual Student Assessment Accessibility Profile (ISAAP) Tool that created an upload file; or entered into a template without macros. These designated supports and accommodations were delivered to the student through the test delivery system at the time of testing. Refer to section *1.9 Systems Overview and Functionality* in *Chapter 1: Introduction* for more details regarding this system.

Appendix 2.A presents counts and percentages of students assigned designated supports, accommodations, or unlisted resources for the 2018–2019 CAST administration. Table 2.A.1 presents data for grades five and eight tests and table 2.A.2, for grades ten, eleven, and twelve individually and aggregated as high school. The tables in appendix 2.A were created using student demographic data that was in version 3 of the production data file ("P3") that was updated on December 17, 2019.

### 2.4.1. Resources for Selection of Accessibility Resources

The full list of the universal tools, designated supports, and accommodations that are used in all CAASPP online assessments, including the CAST, is documented in Matrix One (CDE, 2019b). Part 1 of Matrix One lists the embedded universal tools, designated supports, and accommodations available for CAST online testing. Parts 2 and 3 of Matrix One include the non-embedded universal tools, designated supports, accommodations, and unlisted resources that are available. School-level personnel, IEP teams, and Section 504 teams use Matrix One when deciding how best to support the student's test-taking experience.

The Smarter Balanced Assessment Consortium's *Usability, Accessibility, and Accommodations Guidelines* ("*Guidelines*") (Smarter Balanced, 2019b) aids in the selection of universal tools, designated supports, and accommodations deemed necessary for individual students to take the CAST.[4] The *Guidelines* apply to all students and promote an

[3] This technical report is based on the version of Matrix One that was available during the 2018–2019 CAST administration.

[4] This technical report is based on the version of the *Usability, Accessibility, and Accommodations Guidelines* that was available during the 2018–2019 CAST administration.

individualized approach to the implementation of assessment practices. The *Guidelines* are intended to provide recommendations regarding universal tools, designated supports, and accommodations. Another manual, the *Smarter Balanced Usability, Accessibility, and Accommodations Implementation Guide* (Smarter Balanced, 2014), provides suggestions for implementation of these resources.

In addition to assigning accessibility resources individually and via file upload in TOMS, LEAs had the option of using the ISAAP Tool, which was adapted to include the CAST, to assign resources to students, to facilitate selection of the accessibility resources that match student access needs for the CAST. The CAASPP ISAAP Tool was used by LEAs in conjunction with the *Guidelines* as well as with state regulations and policies (such as Matrix One) related to assessment accessibility as a part of the ISAAP process. LEA personnel, including IEP and Section 504 plan teams, used the CAASPP 2018–2019 ISAAP Tool to facilitate the selection of designated supports and accommodations for students.

### 2.4.2. Delivery of Accessibility Resources

Universal tools, designated supports, and accommodations can be delivered as either embedded or non-embedded resources. Embedded resources are digitally delivered features or settings available as part of the technology platform for the online CAST. Examples of embedded resources include the braille language resource, color contrast, and closed captioning.

Non-embedded resources are available, when provided by the LEA, for both online and paper-pencil CAASPP assessments. These resources are not part of the technology platform for the computer-administered CAASPP tests. Examples of non-embedded resources include magnification, noise buffers, and the use of a scribe.

Refer to section *5.5 Universal Tools, Designated Supports, and Accommodations for Students with Disabilities* for a detailed description of the accessibility resources available to students taking the CAST.

### 2.4.3. Unlisted Resources

An unlisted resource is an instructional resource that a student regularly uses in daily instruction, assessment, or both, that has not been previously identified as a universal tool, designated support, or accommodation. Matrix One includes an inventory of unlisted resources that have already been identified and are preapproved (CDE, 2019b). During the 2018–2019 CAST administration, an LEA CAASPP coordinator or CAASPP test site coordinator had the option to submit a web form in TOMS to request such a resource for an eligible student. The resource was specified in the eligible student's IEP or Section 504 plan and only was assigned with the CDE's approval.

For an unlisted resource to be approved, it must not change the construct of what is being tested. If it did, test results for a student using an unlisted resource that was approved but changed the construct of what was being tested was considered valid for accountability purposes. The student received a score with a footnote that the test was administered under conditions that resulted in a score that may not be an accurate representation of the student's achievement.

## 2.5. Standard Setting

After the 2018–2019 first operational test administration, standard setting was conducted and achievement levels were established. Student performance on the reporting scale was designated into one of four achievement levels:

1. Level 1—Standard Not Met
2. Level 2—Standard Nearly Met
3. Level 3—Standard Met
4. Level 4—Standard Exceeded

For information regarding achievement levels, refer to *Chapter 6: Standard Setting* for a description of the process used to set achievement-level standards.

## 2.6. Scores

Individual student scores were reported for the 2018–2019 CAST for the first time. Student performance on the reporting scale was designated into one of the four achievement levels listed in the previous section.

For information regarding score specifications and score reports, refer to *Chapter 7: Scoring and Reporting*.

### 2.6.1. Score Reporting

TOMS is a secure website hosted by ETS that permits LEA users to manage aspects of CAASPP test administration such as test assignment and the assignment of test settings. It also provides a secure means for LEA CAASPP coordinators to download Student Score Reports as PDF files.

CAST scores could also be viewed through the Online Reporting System (ORS), a secure website that provides authorized users with interactive and cumulative online reports for the CAST at the student, school, and LEA levels. The ORS provides three types of score reports: an individual student score report, a school report, and an LEA report. Refer to subsection *7.4.1 Online Reporting* for details about TOMS and the ORS and subsection *7.4.3 Types of Score Reports* for the content of each type of score report.

Note that the California Educator Reporting System was not available during 2018–2019 testing and is not discussed.

### 2.6.2. Aggregation Procedures

To provide meaningful results to the stakeholders, CAST scores for a given grade are aggregated at the school, LEA or direct funded charter school, county, and state levels. State-level results are available on the Test Results for California's Assessments web page at http://caaspp.cde.ca.gov/. The aggregated scores are presented for all students or selected demographic student groups.

Aggregate scores are generated by combining student scores. They can be created by combining results at the state, LEA or direct funded charter school, or school level; combining for all students; or by combining results for students who represent selected demographic student groups.

Aggregate results by demographic variables are described in section *7.3 Student Test Scores* of this report. In table 7.C.1 through table 7.C.6 in appendix 7.C, students are grouped by demographic characteristics, including gender, ethnicity, English language

fluency, special education service status, and economic status, as well as crosstab analysis for ethnicity and economic status. The tables show the numbers of students with valid scores in each group, scale score means and standard deviations, and the percentage in each achievement level. To protect student privacy, statistics are presented in the tables as "N/A" when the number of students in the sample is fewer than 11.

Table 5.3 in section *5.2 Demographic Summaries* provides definitions for the demographic student groups included in the tables.

## 2.7. Analyses

Psychometric analyses were conducted on the data from the CAST, including classical item analyses, differential item functioning (DIF) analyses, dimensionality analyses, IRT calibration, response time analyses, reliability analyses, validity analysis, and special research studies (the multistage practicality study and the content screen-out study). The results of these analyses support the understanding of the item performances and the internal structure of the test and provide the validity evidence for both the response processes and scoring. Refer to section *8.8 Validity Evidence* in *Chapter 8: Analyses* for descriptions of these analyses.

# References

California Department of Education. (2019b). Matrix one: Universal tools, designated supports, and accommodations for the California Assessment of Student Performance and Progress for 2018–19. Sacramento, CA: California Department of Education. Retrieved from https://www.cde.ca.gov/ta/tg/ai/caasppmatrix1.asp

California Department of Education. (2019a). CAASPP online test administration manual, 2018–19 administration. Sacramento, CA: California Department of Education. Retrieved from http://www.caaspp.org/rsc/pdfs/CAASPP.online_tam.2018-19.pdf

California Department of Education. (2018). TOMS pre-administration guide for CAASPP testing. Sacramento, CA: California Department of Education. Retrieved from http://www.caaspp.org/rsc/pdfs/CAASPP.TOMS-pre-admin-guide.2018-19.pdf

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 http://www.ets.org/Media/Research/pdf/RR-03-16.pdf

Smarter Balanced Assessment Consortium. (2014). *Smarter Balanced Assessment Consortium: Usability, accessibility, and accommodations implementation guide.* Los Angeles, CA: Smarter Balanced Assessment Consortium and National Center on Educational Outcomes. Retrieved from https://portal.smarterbalanced.org/library/en/usability-accessibility-and-accommodations-implementation-guide.pdf

Smarter Balanced Assessment Consortium. (2019). *Smarter Balanced Assessment Consortium: Usability, accessibility, and accommodations guidelines*. Los Angeles, CA: Smarter Balanced Assessment Consortium. Retrieved from https://portal.smarterbalanced.org/library/en/usability-accessibility-and-accommodations-guidelines.pdf

# Chapter 3: Item Development

This chapter discusses the detailed procedures of item development for the 2018–2019 California Science Test (CAST) operational 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, 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. Theory of Action Model

One of the stakeholder priorities presented to the California State Board of Education (SBE) in March 2016 was the "focus on providing information to support the *improvement of teaching and learning*" (California Department of Education [CDE], 2016).

The principles of ECD pervade all aspects of the CAST, including item development, so that the CAST is able to gather evidence of student proficiency that can be used to support improvement in how science is taught. Because CAST items are aligned to a wide variety of performance expectations (PEs), the CAST allows students to demonstrate a wide array of scientific knowledge and skills. In this way, the CAST supports instruction that "encourages students to build the knowledge and skills needed for college and careers" (CDE, 2019).

As part of ECD, ETS continually analyzes student performance data and feedback from external stakeholders to improve alignment to the standards and provide students with grade-level appropriate phenomena. ETS incorporates feedback from external stakeholders in the CAST development process to allow for continual improvement of the assessment and to impact instructional strategies. These support the validity argument for the CAST and support the claims enumerated in the CAST blueprint.

The logic model provides the sequence of how the CAST was conceptualized, starting with the components that led to the design and development of the assessment, the anticipated actions by stakeholders, and outcomes (both intended for the intermediate and long-term futures; and potentially unintended).

#### 3.1.2.1. Components

The CAST was aligned to the California Next Generation Science Standards (CA NGSS) to use rigorous three-dimensional standards that emphasize continual building of knowledge and skills, as well as to

- assess both a breadth and depth of the CA NGSS via items aligned to the three dimensions of the standards, and
- include performance tasks that encourage test takers to thoroughly explore a phenomenon through the dimensions of the CA NGSS, in addition to discrete items.

Accessibility resources are available to students to ensure equity of the assessment.

#### 3.1.2.2. Actions

**Stakeholders** are an essential part of ECD. Stakeholder groups involved in the logic model are students, educators, and those developing the assessment.

***Students*** engage in performance tasks that give the opportunity to experience authentic science.

***Educators*** will

- participate in item writing workshops to improve understanding in the multidimensionality of the CA NGSS,
- access item content specifications to develop properly aligned instruction, and
- access released training and practice tests to better understand CAST content.

***Assessment developers*** refine development practices and ways to obtain and apply external feedback after reviewing item field-test performance data.

#### 3.1.2.3. Intermediate Outcomes

After administration, the results from the CAST can show how students have begun to make sense of phenomena using the knowledge and skills learned through aligned instruction.

Results from the CAST

- provide educators, students, parents, and guardians with information about the student's progress; and
- help advise schools and local educational agencies (LEAs) on strengths and weaknesses in their instructional programs to provide better alignment to the CA NGSS.

#### 3.1.2.4. Long-term Outcomes

Both the student and educator stakeholder groups can benefit from a science assessment developed using ECD.

***Students*** develop the ability to provide mechanistic reasoning about phenomena in the natural and designed world around them.

***Educators*** continue to better align instruction with the three dimensions of the CA NGSS to promote greater science proficiency.

#### 3.1.2.5. Unintended Outcomes

As with any new endeavor, there may be unintended outcomes that will be overcome.

For example, in early administrations of the CAST, teachers and students will have had limited access to teaching and learning of the CA NGSS. Therefore, score users may initially misinterpret CAST results due to limited access to the CA NGSS.

This will be resolved as student exposure to the CA NGSS increases and educators have put effective strategies in place for presenting the concepts associated with the CA NGSS.

### 3.1.3. 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 PEs in some detail (Achieve, 2015). ETS drafted work on the task models and drafted work on task templates to outline the types of items that would elicit student output sufficient to provide evidence for the PE claims.

The task-model documentation is practice-based. ETS developed one design pattern for each CA NGSS science and engineering practice (SEP) 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 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. The approaches were 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. Detailed information on item specifications is presented in subsection *3.2.3 Specifications*.

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 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 administration and the data collection from the operational administration in 2018–2019. 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).

### 3.1.4. Item Types and Features

Every CAST 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 2018–2019 CAST 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. This includes item types ranging from traditional multiple-choice (MC) items and constructed-response (CR) items (i.e., extended text) to new technology-enhanced item (TEI) types (the remainder of the item types).

**Table 3.1 Selected Item Types in the CAST**

| 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 |
| **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) |

# 3.2. Item Development

## 3.2.1. Plan

The initial item development plan for the CAST 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 CAST.

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

| Item Type | Grade 5 | Grade 8 | High School |
|---|---|---|---|
| Standard discrete item types (non-CR) | 177 | 179 | 176 |
| Discrete CR | 11 | 10 | 10 |
| Performance task items (eight tasks in each grade) | 46 | 39 | 38 |
| **TOTAL** | **234** | **228** | **224** |

The standard discrete item types from table 3.2 include traditional MC items, familiar TEI types (e.g., match, inline choice list, etc.), and new TEI types with simulations and animations, which are also indicated as custom interactive discrete items.

The performance task, which contained four to six items for the CAST, was 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 in accordance with the *ETS Standards for Quality and Fairness* (2014) across all phases of item and test development.

## 3.2.2. 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 or stimuli is presented clearly and is 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.

### 3.2.3. Specifications

ETS created item specifications for the CAST using feedback from the 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:

- Science and Engineering 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 CAST items will be updated periodically to support subsequent rounds of item development.

### 3.2.4. Selection of Item Writers

Senior ETS content staff screened applications for CAST item writers, 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 shared the following 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
- Familiarity and understanding of the CA NGSS

### 3.2.5. 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, 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; and
3. leverage the experiences, perspectives, and expertise of the teachers in writing items for the CAST.

ETS held an item writer–training workshop in November 2017 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 life cycle 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 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.

## 3.3. Item Review Process

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

Once an item was accepted for authoring, ETS employed a series of internal reviews. These reviews used 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 examined the overall quality of the test items before presentation to the CDE and item review meetings, which are described in more detail in section *3.4 Content Expert Review*.

The ETS review process for the CAST includes the following; these tasks 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 at ETS continually evaluated the activities and items for adherence to the rules for item development.

### 3.3.1. ETS Content Review

CAST items and stimuli underwent 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 verified that the items and stimuli complied with the approved item specifications and with ETS' written guidelines for clarity, style, accuracy, and appropriateness for California students, as well as complied with the task models. Assessment specialists reviewed 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 was classified with the PE that it was intended to measure. The assessment specialists checked each item against its classification codes, both to evaluate the correctness of the classification and to confirm that the task posed by the item was relevant to the outcome it was intended to measure. The reviewers had the choice to accept the item and classification as written, suggest revisions, or recommend that the item be discarded. These steps occurred prior to the CDE's review.

### 3.3.2. ETS Research Review

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

### 3.3.3. ETS Editorial Review

After content-area assessment specialists and researchers reviewed each item, a group of specially trained editors also reviewed each item in preparation for consideration by the CDE and item review meeting panelists. The editors checked 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.

### 3.3.4. 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—conducted the next level of review (ETS, 2014, 2016b). These trained staff members reviewed every item before the CDE and item review panelist 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

# 3.4. Content Expert Review

## 3.4.1. California Educators as Content Experts

In addition to the ETS internal content reviews, meetings with California educators were held at the end of the item review process as the final content-expert review that items must undergo before being placed on the CAST. The California educators filled an advisory role to the CDE and ETS and provided guidance on matters related to item development for the CAST.

These educators were responsible for reviewing all newly developed items for alignment to the CA NGSS. Meeting participants also reviewed the items for accuracy of content, clarity of phrasing, and overall quality. In their examination of test items, participants could raise concerns related to grade appropriateness as well as gender, racial, ethnic, or socioeconomic bias.

## 3.4.2. Composition of Item Review Panels

The panelists for the item review meetings for CAST items include current and former teachers, resource specialists, administrators, curricular experts, and other education professionals. Minimum qualifications to be invited to participate were

- three or more years of general teaching experience in kindergarten through grade 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; LEA, county content, or program specialists; or university educators met the following qualifications to be invited to participate:

- Three or more years of experience as a school administrator, LEA, county content, or 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

Every effort was made to ensure that groups of item reviewers included a wide representation of genders, geographic regions, and ethnic groups in California. Table 3.3 shows the educational qualifications, present occupation, and credentials of the 21 individuals who participated in the CAST item review.

**Table 3.3 CAST Item Reviewer Qualifications**

| Qualification Type | Qualification | Total |
|---|---|---|
| **Occupation** | Special Education Teacher | 0 |
| **Occupation** | Educational Specialist | 8 |
| **Occupation** | General Education Teacher | 13 |
| **Highest Degree Earned** | Bachelor's Degree | 10 |
| **Highest Degree Earned** | Master's Degree | 10 |
| **Highest Degree Earned** | Doctorate | 1 |
| **K–12 Teaching Credential** | Elementary Teaching (multiple subjects) | 3 |
| **K–12 Teaching Credential** | Secondary Teaching (single subject) | 16 |
| **K–12 Teaching Credential** | Special Education | 0 |
| **K–12 Teaching Credential** | Reading Specialist | 0 |
| **K–12 Teaching Credential** | English Learner (CLAD, BCLAD) | 0 |
| **K–12 Teaching Credential** | Administrative | 0 |
| **K–12 Teaching Credential** | Other | 2 |

Item reviewers were recruited through an online application process. Recommendations were solicited from LEAs and county offices of education as well as from the CDE and SBE staff. ETS assessment directors reviewed applications and confirmed that an applicant's qualifications met the specified criteria. Applicants who met the criteria had their information forwarded to CDE and SBE staff for further review and agreement before invitations to participate were distributed.

### 3.4.3. Meetings for Review of Operational CAST Items

ETS content-area assessment specialists facilitated CAST 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
- Overview of the CAST 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. The set of guidelines for reviewing items is summarized next.

- 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 on whether to field-test the items.

## 3.5. Data Review Meeting

After items were included in an operational or field test and administered to students, ETS prepared the items and the associated statistics for review by the CDE and California educators.

ETS conducted an introductory training to highlight any new issues and serve as a statistical refresher. Reviewers then made decisions about which items should be included in the item bank for future assembly. If an item was considered problematic and not to be included in the item bank, it would be either removed from the bank or revised and once again follow the steps in the item development process, including field testing. ETS psychometric and content staff were available to reviewers throughout this process.

Content staff facilitated the meeting, confirming that all educators weighed in on each flagged item to confirm there were no concerns, from a content perspective, as it pertained to the flag. ETS psychometricians provided training on the item statistics and responded to questions about the item statistics during the item discussion. The data review meeting participants reviewed the content and statistics of each item and then made a recommendation to accept or reject an item.

Content staff recorded each participant's recommendations and comments regarding the flagged items. The feedback was referenced when working with the CDE to reconcile educator feedback and to make a final decision on whether or not to include the item in the operational pool.

Refer to table 3.4 for the results of the item data review, showing the number of items accepted with and without edits and the number of items rejected outright. The rejection rates were 7 percent, 9 percent, and 5 percent for grade five, grade eight, and high school, respectively.

**Table 3.4 Item Data Review Results**

| Grade or Grade Level | Accept As Is | Reject | Total Items | Rejection Rate |
|---|---|---|---|---|
| Grade 5 | 97 | 7 | 104 | 7% |
| Grade 8 | 113 | 11 | 124 | 9% |
| High school | 136 | 7 | 143 | 5% |

# References


Achieve. (2015). *Next Generation Science Standards evidence statements.* Available from https://www.nextgenscience.org/evidence-statements/

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. (2016). *California's Next Generation Science Standards (CA NGSS) assessment plan.* Presented to the California State Board of Education in March 2016. Sacramento, CA: California Department of Education. Retrieved from https://www.cde.ca.gov/be/ag/ag/yr16/documents/mar16item02slides.pdf

California Department of Education. (2019) *California Science Test assessment fact sheet.* Sacramento, CA: California Department of Education. Retrieved from ~~https://www.cde.ca.gov/ta/TG/ca/documents/castfactsheet.pdf~~

Educational Testing Service. (2014). *ETS standards for quality and fairness*. Princeton, NJ: Educational Testing Service. Retrieved from https://www.ets.org/s/about/pdf/standards.pdf

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

Educational Testing Service. (2016b). *ETS guidelines for fair tests and communications*. Princeton, NJ: Educational Testing Service. Retrieved from ~~https://www.ets.org/s/about/pdf/ets_guidelines_for_fair_tests_and_communications.pdf~~

Educational Testing Service. (2019). *Use of evidence-centered design in CAST item and test development.* [Unpublished report.] Princeton, NJ: Educational Testing Service.

Instructional Management Systems (IMS). (2016). *IMS question & test interoperability specification.* Available from https://www.imsglobal.org/question/

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 http://www.ets.org/Media/Research/pdf/RR-03-16.pdf

# Chapter 4: Test Assembly

Educational Testing Service (ETS) assembled the 2018–2019 California Science Test (CAST) operational test forms in consultation with California Department of Education (CDE) content experts after the data review meeting was completed. Only items accepted by the data review committee or the CDE were used to build the operational forms. The forms were built to meet the test blueprint and other test content specifications, as well as the psychometric specifications. This chapter discusses the detailed procedures of the test assembly.

## 4.1. Test Design

The CAST design is based on the State Board of Education (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. Segments A and B contribute to individual student score reporting. Segment C is used for field-testing purposes, where the items do not count toward students' score reporting.

In the 2018–2019 design, Segment A contained two blocks of discrete items, each with 16 to 17 items, with a total of 42 to 44 points equally distributed between the two blocks. All students received the same two Segment A blocks. The delivery order of the two Segment A blocks was random, with either one being given first to the student.

Segment B included two performance tasks (PTs) from the pool. Each PT consisted of four to seven items worth six to seven points total. A PT may contain embedded field test items to meet the test blueprint. Each PT was identified as applying to one of the three science domains through analysis of the items. Where a PT had items that were part of more than one science domain, a determination of the primary science domain for the PT was based upon the content of the PT. Each student received two PTs from two different (primary) domains out of the three main content domains—Life Sciences, Physical Sciences, and Earth and Space Sciences. The PTs were presented after the discrete blocks and were delivered randomly, with either one delivered first.

Segment C contained 15 discrete blocks and up to six PTs in the pool. Each student received either one discrete block or one PT. Each student was administered one discrete block with 13 items or one PT with seven to eight items.

## 4.2. Test Blueprints and Other Content Specifications

### 4.2.1. Test Blueprints

Table 4.1 shows the CAST blueprint approved by the California SBE in November 2017 that was used to build the 2018–2019 operational forms. For details on test blueprint for the CAST, please refer to the *California Science Test Blueprint* (CDE, 2017).

### Table 4.1 CAST Blueprint for Segments Contributing to Individual Scores

| Science Content Domain and Disciplinary Core Idea (DCI)** | Segment A: Discrete Items by DCI—Grade 5 | Segment A: Discrete Items by DCI—Grade 8 | Segment A: Discrete Items by DCI—HS | Segment B: Performance Tasks (PTs) |
|---|---|---|---|---|
| **Physical Sciences (PS)1:** Matter and Its Interactions | 1–3 | 1–5 | 2–7 | 0–1 PTs for all PS DCI strands |
| **PS2:** Motion and Stability: Forces and Interactions | 1–4 | 1–4 | 1–5 | 0–1 PTs for all PS DCI strands |
| **PS3:** Energy | 1–4 | 1–4 | 1–4 | 0–1 PTs for all PS DCI strands |
| **PS4:** Waves and Their Applications in Technologies for Information Transfer | 1–2 | 1–2 | 1–4 | 0–1 PTs for all PS DCI strands |
| **ETS1:** Engineering Design | * | * | * | 0–1 PTs for all ETS (PS) DCI strands |
| **Total for PS:** | • **8–12 items**<br>• **12–18 points** | • **8–12 items**<br>• **12–18 points** | • **8–12 items**<br>• **12–18 points** | • **4–6 items per PT**<br>• **6–7 points total** |
| **Life Sciences (LS)1:** From Molecules to Organisms: Structures and Processes | 1–2 | 1–6 | 1–6 | 0–1 PTs for all LS DCI strands |
| **LS2:** Ecosystems: Interactions, Energy and Dynamics | 1–2 | 1–4 | 1–7 | 0–1 PTs for all LS DCI strands |
| **LS3:** Heredity: Inheritance and Variation of Traits | 1–2 | 1–2 | 1–2 | 0–1 PTs for all LS DCI strands |
| **LS4:** Biological Evolution: Unity and Diversity | 1–4 | 1–5 | 1–5 | 0–1 PTs for all LS DCI strands |
| **ETS1:** Engineering Design | * | * | * | 0–1 PTs for all ETS (LS) DCI strands |
| **Total for LS:** | • **8–12 items**<br>• **12–18 points** | • **8–12 items**<br>• **12–18 points** | • **8–12 items**<br>• **12–18 points** | • **4–6 items per PT**<br>• **6–7 points total** |

Table 4.1 *(continuation)*

| Science Content Domain and Disciplinary Core Idea (DCI)** | Segment A: Discrete Items by DCI—Grade 5 | Segment A: Discrete Items by DCI—Grade 8 | Segment A: Discrete Items by DCI—HS | Segment B: Performance Tasks (PTs) |
|---|---|---|---|---|
| **Earth and Space Sciences (ESS)1:** Earth's Place in the Universe | 1–2 | 1–3 | 1–5 | 0–1 PTs for all ESS DCI strands |
| **ESS2:** Earth's Systems | 1–5 | 1–5 | 1–6 | 0–1 PTs for all ESS DCI strands |
| **ESS3:** Earth and Human Activity | 1–3 | 1–4 | 1–5 | 0–1 PTs for all ESS DCI strands |
| **ETS1:** Engineering Design | * | * | * | 0–1 PTs for all ETS (ESS) DCI strands |
| **Total for ESS:** | • **8–12 items**<br>• **12–18 points** | • **8–12 items**<br>• **12–18 points** | • **8–12 items**<br>• **12–18 points** | • **4–6 items per PT**<br>• **6–7 points total** |
| **TOTAL:** | • **32–34 items**<br>• **42–44 points** | • **32–34 items**<br>• **42–44 points** | • **32–34 items**<br>• **42–44 points** | • **2 PTs**<br>• **8–12 items**<br>• **12–14 points** |

* Across the three science content domains, a student will receive two to four items assessing Engineering, Technology, and the Applications of Science. The item(s) may be discrete or part of a PT.

** The CAST Item Specifications provide greater detail on the assessment targets by performance expectation.

The test blueprint also specifies the performance expectation (PE) distribution for Segment A items by the disciplinary core ideas (DCIs) (within each content domain), science and engineering practices (SEPs) and crosscutting concepts (CCCs). These tables are included in appendix 4.A, in figure 4.A.1 through figure 4.A.3.

Segment A is designed to assess a student's mastery of a breadth of PEs of the California Next Generation Science Standards in the grade bands tested (grade five, grade eight, and high school—either grade ten, eleven, or twelve).

The tables display an "X" for the intersections of SEPs, DCIs, and CCCs articulated in the PEs. These intersections represent opportunities to develop items that can be used to assemble Segment A. While each individual item reflects the intersection of a SEP, DCI, and CCC, the tables also indicate the proposed distribution of Segment A items by DCI, SEP, and CCC. Segment A had 8 to 10 items in each of the three science domains: Physical Sciences, Life Sciences, and Earth and Space Sciences. Two to four items were in the Engineering, Technology, and Applications of Science PEs, but for scoring and reporting purposes, items written to those PEs were assigned to one of the three science domains depending on the context of their stimulus.

### 4.2.2. Other Content Specifications

The two Segment A blocks were built to be parallel in terms of other content and statistical specifications.

Segment A was limited to an item type distribution of 40 to 50 percent multiple-choice items and 40 to 50 percent technology-enhanced items per form. Each Segment A block had one to two constructed-response (CR) items. Most PTs only contained one CR item. Cognitive complexity as measured by depth of knowledge (DOK) had a distribution for Segment A where a minimum of 5 percent of the items on the form had DOK 1, 45 to 60 percent measured at DOK 2, and 30 to 45 percent measured at DOK 3 or 4.

The sequence of items used in the operational form were as close as possible to their sequence when they were field-tested.

## 4.3. Psychometric Criteria

The item statistics such as the *p*-value (item difficulty; refer to subsection *8.2.1 Classical Item Difficulty Indices (p-value)* for more details on this statistic); and item total polyserial correlation (item discrimination; refer to subsection *8.2.2 Item-Total Score Correlations* for more details on this statistic) obtained from the field test administration were used to inform the item selection for the operational forms. At the form level, the distribution of *p*-values ranged from 0.2 to 0.95, with the center around 0.5.

The following psychometric criteria were applied in the form assembly:

- The *p*-value is between 0.2 and 0.95. A *p*-value less than 0.2 suggests that the item might be too difficult; a *p*-value greater than 0.95 suggests that the item might be too easy. Items that were too easy or too difficult were not used as they provided little information on evaluating students' abilities.
- The item-total polyserial correlation is at least 0.2. Most items selected had polyserial correlations higher than 0.3.
- Items with C-DIF should not be used unless it is necessary for content coverage (refer to subsection *8.3.3 Classification* for more details on the differential item

functioning [DIF] classification). All C-DIF items were reviewed by a DIF panel that included members of the focal groups that were affected and who confirmed the items were not biased before the items could be selected for use. The panelists did not have a vested interest in the outcome of the decision.

# 4.4. Test Production Process

## 4.4.1. Selection of Items

From the eligible item pool, test developers selected items that, as a whole

- met the coverage specifications of the test blueprints (subsection *4.2.1 Test Blueprints*),
- met item selection criteria developed by the ETS psychometrics team (section *4.3 Psychometric Criteria*),
- represented a wide variety of item types, and
- provided a wide variety of item context.

## 4.4.2. Verification of Statistics

ETS test developers sent the proposed assessment to the ETS psychometrics team for approval. The proposed assessment was reviewed to ensure that all statistical guidelines were met for both individual items and the assembled segments.

## 4.4.3. Test Forms

The numbers of unique operational forms are

- 7 forms for grade five,
- 12 forms for grade eight, and
- 3 forms for high school (grade ten, eleven, or twelve).

By including field test blocks, the total numbers of forms are

- 147 forms for grade five,
- 240 forms for grade eight, and
- 60 forms for high school (grade ten, eleven, or twelve).

All the forms mentioned previously were used for the 2018–2019 administration. The variation in the number of forms across grades is partially due to the number of operational PTs available for each grade level.

## 4.4.4. Content Review of Forms

After psychometric approval, the proposed assessment underwent two additional content reviews and one editorial review. The form content reviewers are test developers who work on assessment programs other than the CAST and who are able to bring a fresh perspective to the review. They were given the appropriate materials and documentation to complete the following tasks:

- Verification of item keys
- Identification of possible clueing across the items
- Verification that individual items meet the standard
- Verification of coverage of the standards
- Identification of any possible grammatical or production errors

### 4.4.5. CDE Review of Forms

Following the ETS content review, all proposed assessments were sent to the CDE for review to ensure the proposed assessments met the CAST blueprint requirements and to check for possible clueing between items or statistical issues. The CDE was provided with block builders that catalogued information and collected CDE comments on the assembled segments.

Comments from the CDE were resolved during a virtual meeting with the ETS test development team.

### 4.4.6. Configuration of the Test Delivery System (TDS)

Once all the test reviews were completed and concerns, if any, had been resolved, the official ordered item sequence of the proposed forms was sent to the American Institutes for Research (AIR) (now Cambium Assessment) for configuration of the TDS.

Each item underwent an extensive platform review on different operating systems such as Windows, Linux, and iOS to ensure that the item looked consistent across all platforms.

The platform review was conducted by a team at AIR consisting of a team leader and several team members. The team leader presented the item as it was approved in ETS and AIR item banks. Each team member was assigned a different platform—hardware device and operating system—and reviewed the item to see that it rendered as expected. This platform review meeting ensured that all items were presented consistently to all students regardless of testing device or operating system for standardization of the test administration.

Prior to operational deployment, the testing system and content were deployed to a staging server where they were subject to user acceptance testing (UAT) by both ETS and AIR staff. The TDS UAT served as both a software evaluation and a content approval.

Following the UAT by ETS and AIR staff, separate UAT cycles were conducted by the CDE. The UAT review provided the CDE with an opportunity to interact with the exact test that would be administered to the students. The CDE had to approve the CAST UAT before the test could be released for administration to students.

### 4.4.7. Test Form Delivery

Students were randomly assigned blocks of items for each segment during the 2018–2019 administration.

## 4.5. Performance Expectation (PE) Coverage

The various blocks of items that comprise each segment of the CAST covered an extensive range of PEs; these PEs at the operational-item-pool level are shown for all three grade levels in table 4.2.

**Table 4.2 Performance Expectations Assessed on the CAST—All Grade Levels**

| Grade or Grade Level | PEs Assessed | PEs Available | Percent of PEs Assessed |
|---|---|---|---|
| Grade 5 | 45 | 45 | 100% |
| Grade 8 | 59 | 59 | 100% |
| High school | 69 | 71 | 97% |

# 4.6. Special Forms

## 4.6.1. Braille Form

ETS designed a braille form for students with visual impairment. 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.

The items appeared in the same or similar positions on the braille form as they did in Segment A of the form for the general population. The braille form Segment B included two PTs from two different science content domains. These two PTs were selected from the eight PTs in Segment B of the general population form. The braille form Segment C included one PT and one discrete block. The PT was one of the six field test PTs from the general form. The discrete block included six items from one of the 15 discrete field test blocks from the general form.

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 accessibility needs of the population of students with visual impairment. Adaptation may have included simplified graphics, more descriptive alternative text for images, or other changes to make the item more accessible to refreshable braille devices, embossed tactile graphics, or screen readers. Adaptation did not change the item type. Twinning an item meant the item was rewritten using another item type while maintaining the same construct and storyline of the original item. Whether items were adapted or twinned, overall cognitive complexity was maintained as closely as possible with the original parent item.

## 4.6.2. Forms with Accessibility Features Other Than Braille

A subset of the general form blocks was used to provide accessible content for those students who had one or more designated supports or accommodations assigned, as determined by an educator, individualized education program, or Section 504 plan. Items were embedded with content for text-to-speech, stacked Spanish, translation glossaries, and ASL videos. Refer to *5.5 Universal Tools, Designated Supports, and Accommodations for Students with Disabilities* for a list of designated supports and accommodations available during the 2018–2019 CAST administration.

Both Segment A blocks found on the general form included designated supports and accommodations. Segment B had three PT blocks for that used these resources. Segment C designated one of the 15 discrete blocks and two PTs in different science domains as accessible for these resources. Students received either the discrete block or a PT for Segment C.

# Reference

California Department of Education. (2017). *California Science Test Blueprint.* Sacramento, CA: California Department of Education. Retrieved From https://www.cde.ca.gov/ta/tg/ca/documents/castblueprint.pdf

# Chapter 5: Test Administration

This chapter describes the details of California Science Test (CAST) operational test administration, as well as the procedures followed by Educational Testing Service (ETS) to ensure test security.

## 5.1. Student Test-Taking Requirement

The CAST was administered to students in grades five and eight as well as high school students in grade ten, eleven, or twelve who were assigned by their local educational agency (LEA). The CAST is a science assessment for the general student population (i.e., those students who are not otherwise eligible for the California Alternate Assessment [CAA] for Science). Subsection *5.1.2 High School* outlines the process for grade assignment in the CAST for high school students.

### 5.1.1. Grades Five and Eight

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

### 5.1.2. High School

At the high school level, schools and LEAs were responsible for assigning students in grade ten or eleven. Guidelines were provided by the California Department of Education (CDE) suggesting that students who completed or were in the process of completing their last high school science course and who were not eligible for the CAA for Science take the operational test. All grade twelve students were assigned to take the CAST.

Table 5.1 provides the composition of the test-taker population for the CAST for high school students. The sum of the number of schools for grades ten, eleven, and twelve does not equal the total number of unique schools because each school may include multiple grades.

A total of 557,251 students in grades ten, eleven, and twelve from 2,686 California schools took the CAST during the 2018–2019 administration. The test-taker population was comprised of 4 percent of grade ten, 46 percent of grade eleven, and 50 percent of grade twelve students.

**Table 5.1 Composition of Test-Taker Population for the CAST for High School Students**

| Variable | Grade 10 | Grade 11 | Grade 12 | Total |
|---|---|---|---|---|
| Number of Schools | 247 | 1,551 | 2,418 | 2,686 |
| Percent of Schools | 9% | 58% | 90% | 100% |
| Number of Students | 23,352 | 257,271 | 276,628 | 557,251 |
| Percent of Students | 4% | 46% | 50% | 100% |

Table 5.2 presents the test-taking rates for all grades. Note that test takers are students who enrolled and logged on to the test. Among the students who enrolled, the percentages of those who took the CAST ranged from 82.7 percent to 98.4 percent across grade levels. The highest test-taking rate was in grade five, with 98.4 percent. The lowest test-taking rate was found in grade ten, with 82.7 percent. Of the high school, grade eleven held the highest test taking rate of 94 percent.

**Table 5.2 CAST Test-Taking 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 | 463,976 | 474,832 | 28,247 | 273,743 | 323,195 | 625,185 |
| Number of Test Takers | 456,604 | 463,151 | 23,352 | 257,271 | 276,628 | 557,251 |
| Percent of Test Takers | 98.4 | 97.5 | 82.7 | 94.0 | 85.6 | 89.1 |

## 5.2. Demographic Summaries

Table 5.A.1 through table 5.A.6 in appendix 5.A show the test-taking rates of selected demographic student groups for each test. The demographic student groups include gender, ethnicity, English-language fluency, economic status (disadvantaged or not), special education services status, migrant status, parent military status, and homeless status.

Demographic student groups included in the summaries in this chapter are shown in table 5.3. The number and the percent of students for these demographic student groups are provided in appendix 5.B, starting in table 5.B.1 through table 5.B.5 for each grade, and in table 5.B.6 for high school.

**Table 5.3 Demographic Student Groups to Be Reported**

| Category | Student Groups |
|---|---|
| Gender | • Male<br>• Female |
| Ethnicity | • American Indian or Alaska Native<br>• Asian<br>• Black or African American<br>• Filipino<br>• Hispanic or Latino<br>• Native Hawaiian or Other Pacific Islander<br>• White<br>• Two or more races |
| English Language Fluency | • English only<br>• Initial fluent English proficient<br>• English learner<br>• Reclassified fluent English proficient<br>• To be determined<br>• English proficiency unknown |

Table 5.3 *(continuation)*

| Category | Student Groups |
|---|---|
| Economic Status | • Not economically disadvantaged<br>• Economically disadvantaged |
| Primary Disability Type | • No special education services<br>• Special education services |
| Migrant Status | • Eligible for the Title I Part C Migrant Program<br>• Not eligible for the Title I Part C Migrant Program |
| Military | • Eligible based on parent's most recent active military status<br>• Not eligible based on parent's most recent active military status |
| Homeless | • Designated as homeless in the California Longitudinal Pupil Achievement Data System (CALPADS)<br>• Not designated as homeless in CALPADS |

# 5.3. Procedures to Maintain Standardization

The test administration procedures are designed so that the tests are administered in a standardized manner. ETS takes all necessary measures to ensure the standardization of test administration, as described in this section.

## 5.3.1. LEA CAASPP Coordinator

An LEA CAASPP coordinator was designated by the district superintendent at the beginning of the 2018–2019 school year. LEAs include public school districts, statewide benefit charter schools, State Board of Education–authorized charter schools, county office of education programs, and direct funded charter schools.

LEA CAASPP coordinators are responsible for ensuring the proper and consistent administration of the CAASPP assessments. In addition to the responsibilities set forth in 5 *CCR* Section 857, their responsibilities include

- adding CAASPP test site coordinators and test administrators into TOMS;
- training CAASPP test site coordinators and test administrators regarding the state and CAASPP assessment administration as well as security policies and procedures;
- reporting test security incidents (including testing irregularities) to the CDE;
- overseeing test administration activities;
- printing out checklists for CAASPP test site coordinators and test administrators to review in preparation for administering the summative assessments;
- distributing and collecting scorable and nonscorable materials for students who take paper-pencil tests;
- filing a report of a testing incident in STAIRS; and
- requesting an Appeal (if indicated by TOMS prompts while reporting an incident using the STAIRS/Appeal process).

### 5.3.2. CAASPP Test Site Coordinator

A CAASPP test site coordinator is trained by the LEA CAASPP coordinator for each test site (5 *CCR* Section 857[f]). A test site coordinator must be an employee of the LEA and must sign a security agreement (5 *CCR* Section 859[a]).

A test site coordinator is responsible for identifying test administrators and ensuring that they have signed CAASPP Test Security Affidavits (5 *CCR* Section 859[d]). CAASPP test site coordinators' duties may include

- adding test administrators into TOMS;
- entering test settings for students;
- creating testing schedules and procedures for a school consistent with state and LEA policies;
- working with technology staff to ensure secure browsers are installed and any technical issues are resolved;
- monitoring testing progress during the testing window and ensuring all students take the test, as appropriate;
- coordinating and verifying the correction of student data errors in the California Longitudinal Pupil Achievement Data System;
- ensuring a student's test session is rescheduled, if necessary;
- addressing testing problems;
- reporting security incidents;
- overseeing administration activities at a school site;
- filing a report of a testing incident in STAIRS; and
- requesting an Appeal (if indicated by TOMS prompts while reporting an incident using the STAIRS/Appeal process).

### 5.3.3. Test Administrators

Test administrators are identified by CAASPP test site coordinators as individuals who will administer the CAST.

A test administrator must sign a security affidavit (5 *CCR* Section 850[ae]). A test administrator's duties may include

- ensuring the physical conditions of the testing room meet the criteria for a secure test environment;
- administering the CAASPP assessments, including the CAST;
- reporting all test security incidents to the test site coordinator and LEA CAASPP coordinator in a manner consistent with state and LEA policies;
- viewing student information prior to testing to ensure that the correct student receives the proper test with appropriate resources and reporting potential data errors to test site coordinators and LEA CAASPP coordinators;

- monitoring student progress throughout the test session using the Test Administrator Interface; and
- fully complying with all directions provided in the directions for administration CAASPP (CDE, 2019a).

## 5.3.4. Instructions for Test Administrators

### 5.3.4.1. Test Administrator Directions for Administration

The directions for administration of the CAST used by test administrators to administer the CAST to students are included in the *CAASPP Online Test Administration Manual* (CDE, 2019a). Test administrators must follow all directions and guidelines and read, word-for-word, the instructions to students in the "SAY" boxes to ensure standardization of test administration. Additionally, the *CAASPP Online Test Administration Manual* provides information to test administrators regarding the systems involved in testing, including sections on the TDS, so they may become familiar with the testing application used by their students (CDE, 2019a).

### 5.3.4.2. *CAASPP Online Test Administration Manual*

The *CAASPP Online Test Administration Manual* (CDE, 2019a) contains information and instructions on overall procedures and guidelines for all LEA and test site staff involved in the administration of online assessments. Sections include the following topics:

- Roles and responsibilities of those involved with CAASPP testing
- Test administration resources
- Test security
- Administration preparation and planning
- General test administration
- Test administration directions and scripts for test administrators
- Overview of the student testing application
- Instructions for steps to take before, during, and after testing

Appendices include definitions of common terms, descriptions of different aspects of the test and systems associated with the test, and checklists of activities for LEA CAASPP coordinators, CAASPP test site coordinators, and test administrators.

### 5.3.4.3. *TOMS Pre-Administration Guide for CAASPP Testing*

TOMS is a web-based application that allows LEA CAASPP coordinators to set up test administrations, add and manage users, submit online student test settings, and order paper-pencil tests. TOMS modules include the following (CDE, 2018a):

- **Test Administration Setup—**This module allows LEAs to determine and calculate dates for the LEA's 2018–2019 administration of the CAST.
- **Adding and Managing Users—**This module allows LEA CAASPP coordinators to add CAASPP test site coordinators and test administrators to TOMS so that the designated user can administer, monitor, and manage the CAASPP Smarter Balanced assessments.
- **Student Test Assignment—**This module allows LEA CAASPP coordinators to designate students to take the CAST in grade ten or eleven. (Students in grades five and eight are assigned automatically to take the CAST.)

- **Online Student Test Settings—**This module allows LEA CAASPP coordinators and CAASPP test site coordinators to configure online test settings so students receive the assigned accessibility resources for the online assessments.

#### 5.3.4.4. Other System Manuals

Other manuals were created to assist LEA CAASPP coordinators and others with the technological components of the CAASPP System and are listed next.

- ***Technical Specifications and Configuration Guide for CAASPP Online Testing—***This manual provides information, tools, and recommended configuration details to help technology staff prepare computers and install the secure browser to be used for the online CAASPP assessments (CDE, 2018b).
- ***Security Incidents and Appeals Procedure Guide—***This manual provides information on how to report a testing incident and submit an Appeal to reset, reopen, invalidate, or restore individual online student assessments (CDE, 2019b).
- ***Accessibility Guide for CAASPP Online Testing—***This manual provides descriptions of the accessibility features for online tests as well as information about supported hardware and software requirements for administering tests to students using accessibility resources, including those with a braille accommodation using Job Access With Speech (JAWS®) (software) or a braille embosser (hardware) (CDE, 2019c).

## 5.4. LEA Training

ETS established and implemented a training plan for LEA assessment staff on all aspects of the assessment program. The CDE and ETS, in collaboration with the CDE Senior Assessment Fellows and other stakeholders as needed, determined the audience, topics, frequency, and mode (in-person, webcast, videos, modules, etc.) of the training, including such elements as format, participants, and logistics.

ETS conducted eight in-person pretest workshops and a pretest webcast for the 2018–2019 administration.

Following approval by the CDE, the ancillary materials were posted for each webcast on the CAASPP website at http://www.caaspp.org/training/caaspp/ so the LEAs could download the training materials.

### 5.4.1. In-person Training

ETS also provided a series of in-person trainings. Beginning in January 2019, the first in-person trainings provided were the pretest CAASPP workshops, which focused on training LEA CAASPP coordinators on how to prepare for administering the CAASPP online assessments. Training was also provided to focus on interpreting and using results. Eight in-person post-test workshops and one webcast were offered in May and June 2019. The post-test workshop and webcast were titled “2018–19 CAASPP Results Are In—Now What?” An additional, stand-alone webcast, “CAASPP Principles of Scoring and Reporting Webcast,” was presented on July 24, 2019.

### 5.4.2. Webcasts

ETS provided a series of live webcasts throughout the school year that were archived and made available for training LEA and test site staff as well as test administrators. Webcast viewers were provided with a method of electronically submitting questions to the presenters during the webcast. The webcasts were recorded and archived for on-demand viewing on the CAASPP Summative Assessments Videos and Archived Webcasts web page at http://www.caaspp.org/training/caaspp/. CAASPP webcasts were available to everyone and required neither preregistration nor a logon account.

### 5.4.3. Videos and Narrated PowerPoint Presentations

To supplement the live webcasts and in-person workshops, ETS also produced short "how to" videos and narrated PowerPoint presentations that were available on the CAASPP Summative Assessments Videos and Archived Webcasts web page. In total, 20 recorded webcasts and tutorials were available.

## 5.5. Universal Tools, Designated Supports, and Accommodations for Students with Disabilities

The purpose of universal tools, designated supports, and accommodations in testing is to allow *all* students the opportunity to demonstrate what they know and what they are able to do, rather than giving students who use these resources an advantage over other students or artificially inflating their scores. Universal tools, designated supports, and accommodations minimize or remove barriers that could otherwise prevent students from demonstrating their knowledge, skills, and achievement in a specific content area.

### 5.5.1. Identification

All public school students participate in the CAASPP System, including students with disabilities and English learners. The Smarter Balanced Assessment Consortium's *Usability, Accessibility, and Accommodations Guidelines* (Smarter Balanced, 2019) and the CDE's Matrix One (CDE, 2019d) are intended for school-level personnel and individualized education program (IEP) and Section 504 plan teams to select and administer the appropriate universal tools, designated supports, and accommodations as deemed necessary for individual students. The CAST follows the Smarter Balanced recommendations for use (Smarter Balanced, 2019).

The *Guidelines* apply to all students and promote an individualized approach to the implementation of assessment practices. Another web document, the *Smarter Balanced Resources and Practices Comparison Crosswalk* (Smarter Balanced, 2018), connects the assessment resources described in the *Guidelines* with associated classroom practices.

Another manual, the *Smarter Balanced Usability, Accessibility, and Accommodations Implementation Guide* (Smarter Balanced, 2014), provides suggestions for implementation of these resources. Test administrators are given the opportunity to participate in the CAST practice and training tests so that students have the opportunity to familiarize themselves with a designated support or accommodation prior to testing.

### 5.5.2. Assignment

Once the student's IEP or Section 504 plan team decided which accessibility resource(s) the student should use, LEA CAASPP coordinators and CAASPP test site coordinators used TOMS to assign designated supports and accommodations to students prior to the start of a test session.

There are three ways the student's accessibility resource(s) could be assigned:

1. Using the Individual Student Assessment Accessibility Profile Tool to identify the accessibility resource(s) and then uploading the spreadsheet it creates into TOMS (This process is discussed in more detail in subsection *2.4.1 Resources for Selection of Accessibility Resources*.)
2. Using the Online Student Test Settings template to enter students' assignments and then uploading the spreadsheet into TOMS
3. Entering assignments for each student individually in TOMS

If a student's IEP or Section 504 plan team identified and designated a resource not identified in Matrix One, the LEA CAASPP coordinator or CAASPP test site coordinator needed to submit a request for an unlisted resource to be approved by the CDE. The CDE then determined whether the requested unlisted resource changed the construct being measured after all testing has been completed.

## 5.5.3. Available Resources

### 5.5.3.1. Universal Tools

Universal tools are accessibility features of the assessment that are available to all students based on student preference and selection (Smarter Balanced, 2019).

These resources were used in the operational administration. They were either embedded in the test delivery system or non-embedded, meaning they were not online.

#### *5.5.3.1.1. Embedded*

- Breaks
- Calculator[5]:
  - Four-function—grade five
  - Scientific—grade eight and high school
- Digital notepad
- English glossary
- Expandable items[6]
- Expandable passages
- Highlighter
- Keyboard navigation
- Line reader
- Mark for review
- Mathematics tools (e.g., ruler, protractor)[7]
- Science charts (i.e., calendar, Periodic Table of the Elements, conversion charts)

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

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

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

- Science tools (e.g., analog clock, laboratory equipment)
- Strikethrough
- Writing tools (e.g., bold, italic, bullets, undo/redo)
- Zoom (in/out)

#### *5.5.3.1.2. Non-embedded*

- Breaks
- Scratch paper

### 5.5.3.2. 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).

These resources were used in the operational administration. They were either embedded in the test delivery system or non-embedded, meaning they were not online.

#### *5.5.3.2.1. Embedded*

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

#### *5.5.3.2.2. Non-embedded*

- 100s number table
- Amplification
- Calculator:
  - Four-function—grade five
  - Scientific—grade eight and high school
- Color contrast
- Color overlay
- Magnification
- Medical device
- Noise buffers
- Read aloud for items and stimuli
- Read aloud in Spanish

- Science charts (i.e., calendar, Periodic Table of the Elements, conversion charts)[8]
- Scribe
- Separate setting (e.g., most beneficial time, special lighting or acoustics, adaptive furniture)
- Simplified test directions
- Translated test directions

#### 5.5.3.3. Accommodations

Accommodations are available to students who have a documented need for the accommodations via an IEP or Section 504 plan.

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

##### *5.5.3.3.1. Embedded*

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

##### *5.5.3.3.2. Non-embedded*

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

## 5.6. Practice and Training Tests

Practice and training tests are available publicly for the CAST. The practice tests simulate the operational testing experience, and the training tests introduce students to the type of thinking needed to answer CAST items. Practice and training tests are available through the Practice and Training Test website linked on the Online Practice and Training Tests Portal web page at http://www.caaspp.org/practice-and-training/.

Grade-level-specific practice tests were released in November 2018. The practice test is designed to be a more authentic representation of the summative CAST. It contains the same types of items and content as the summative CAST, as it was built using the blueprint. Similar to the summative CAST, the practice test should take about two hours to be administered. *Practice Items Scoring Guides* for each grade level (CDE, 2019e; CDE, 2019f; CDE, 2019g) were available in January 2019.

The grade-level-specific training tests can be taken by students in all tested grades. All unique item types available on the operational test are covered in the training tests.

[8] PDFs of the science charts are available for download from the California Science Test web page on the CAASPP website at http://www.caaspp.org/administration/about/science/.

# 5.7. Test Security and Confidentiality

For the CAST, 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/guardians, 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* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 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.

## 5.7.1. ETS's 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* (ETS, 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.

## 5.7.2. 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, 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, 2019a).

The following measures ensured the security of CAASPP System assessments administered in 2018–2019:

- 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 electronically 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])

## 5.7.3. 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.

Refer to section *1.9 Systems Overview and Functionality* in *Chapter 1: Introduction* for more information on TOMS.

### 5.7.4. 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.

### 5.7.5. 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.

### 5.7.6. Statistical Analysis on Secure Servers

During CAASPP testing, ETS information technology staff retrieves data files from the American Institutes for Research (now Cambium Assessment) 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 section *9.6 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 adhere to the ETS Code of Ethics and the ETS Information Protection Policies to prevent any unauthorized access to the data.

## 5.7.7. 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.

## 5.7.8. Student Test Results

### 5.7.8.1. Types of Results

The following deliverables are produced for reporting of the CAST:

- Individual Student Score Reports (printed and electronic)
- Internet reports—available on a public web reporting site—aggregated by content area and state, county, LEA, or test site

### 5.7.8.2. Security of Results Files

ETS takes measures to protect files and reports that show students' scores and achievement levels. ETS is committed to safeguarding all secure information in its possession from unauthorized access, disclosure, modification, or destruction. ETS has strict information security policies in place to protect the confidentiality of both student and client data. ETS staff access to production databases is limited to personnel with a business need to access the data. User IDs for production systems must be person-specific or for systems use only.

ETS has implemented network controls for routers, gateways, switches, firewalls, network tier management, and network connectivity. Routers, gateways, and switches represent points of access between networks. However, these do not contain mass storage or represent points of vulnerability, particularly for unauthorized access or denial of service.

ETS has many facilities, policies, and procedures to protect computer files. Software and procedures such as firewalls, intrusion detection, and virus control are in place to provide for physical security, data security, and disaster recovery. ETS is certified in the BS 25999-2 standard for business continuity and conducts disaster recovery exercises annually. ETS routinely backs up all data to either disks through deduplication or to tapes, all of which are stored off site.

Access to the ETS Computer Processing Center is controlled by employee and visitor identification badges. The Center is secured by doors that can only be unlocked by the badges of personnel who have functional responsibilities within its secure perimeter. Authorized personnel accompany visitors to the ETS Computer Processing Center at all times. Extensive smoke detection and alarm systems, as well as a preaction fire-control system, are installed in the Center.

### 5.7.8.3. Security of Individual Results

ETS protects individual students' results on both electronic files and paper reports during the following events:

- Scoring
- Transfer of scores by means of secure data exchange
- Reporting

- Analysis and reporting of erasure marks
- Posting of aggregate data
- Storage

In addition to protecting the confidentiality of testing materials, ETS' Code of Ethics further prohibits ETS employees from financial misuse, conflicts of interest, and unauthorized appropriation of ETS property and resources. Specific rules are also given to ETS employees and their immediate families who may take a test developed by ETS (e.g., a CAASPP assessment). The ETS OTI verifies that these standards are followed throughout ETS. This verification is conducted, in part, by periodic on-site security audits of departments, with follow-up reports containing recommendations for improvement.

### 5.7.9. 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, 2019b). 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.

LEA CAASPP coordinators and CAASPP test site coordinators must ensure that all test security and summative administration incidents are documented by following the prompts in TOMS that guided coordinators in their submittal. An Appeal is a request to reset, restore, reopen, invalidate, or grant a grace period extension to a student's test. If an Appeal to a student's test was warranted, TOMS provided additional prompts to file the Appeal.

After a case 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 CalTAC with the opportunity to interact and communicate regarding the STAIRS process (CDE, 2019b).

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 cases 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)

Appeals requests were reviewed by the CDE or CalTAC. When a request to submit an Appeal was approved, the coordinator received a system-generated email with the Appeal type that was approved (CDE, 2019b).

#### 5.7.9.1. 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.

#### 5.7.9.2. 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.

#### 5.7.9.3. 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.

### 5.7.10. 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, 2019b).

Table 5.4 describes types of appeals available during the 2018–2019 CAASPP administration.

**Table 5.4 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. The student(s) will be counted as participating in the calculation of the school's participation rate for accountability purposes. The score will be counted as "not proficient" for aggregation into the CAASPP results. |

Table 5.4 *(continuation)*

| Type of Appeal | Description |
|---|---|
| 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.<br><br>A grace period extension will only be granted in cases where there was a disruption to a test session, such as a technical difficulty, fire drill, schoolwide power outage, earthquake, or other act beyond the control of the test administrator. |

# 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. (2018b). *Technical specifications and configuration guide for CAASPP online testing*. Sacramento, CA: California Department of Education. Retrieved from ~~http://www.caaspp.org/rsc/pdfs/CAASPP.tech-specs-and-config-guide.2018-19.pdf~~

California Department of Education. (2018a). *TOMS pre-administration guide for CAASPP testing*. Sacramento, CA: California Department of Education. Retrieved from ~~http://www.caaspp.org/rsc/pdfs/CAASPP.TOMS-pre-admin-guide.2018-19.pdf~~

California Department of Education. (2019c). *Accessibility guide for CAASPP online testing*. Sacramento, CA: California Department of Education. Retrieved from ~~http://www.caaspp.org/rsc/pdfs/CAASPP.accessibility-guide.2018-19.pdf~~

California Department of Education. (2019d). Matrix one: Universal tools, designated supports, and accommodations for the California Assessment of Student Performance and Progress for 2018–19. Sacramento, CA: California Department of Education. Retrieved from https://www.cde.ca.gov/ta/tg/ai/caasppmatrix1.asp

California Department of Education. (2019a). *CAASPP online test administration manual, 2018–19 administration.* Sacramento, CA: California Department of Education. Retrieved from ~~http://www.caaspp.org/rsc/pdfs/CAASPP.online_tam.2018-19.pdf~~

California Department of Education. (2019b). *Security incidents and appeals procedure guide, 2018–19 administration.* Sacramento, CA: California Department of Education. Retrieved from ~~http://www.caaspp.org/rsc/pdfs/CAASPP.stairs-appeals-guide.~~2018-19.pdf

California Department of Education. (2019e). California Science Test Practice Items Scoring Guide Grade Five. Sacramento, CA: California Department of Education. Retrieved from ~~http://www.caaspp.org/rsc/resources/CAST.practice-scoring-guide-gr5.2018-19.pdf~~

California Department of Education. (2019f). California Science Test Practice Items Scoring Guide Grade Eight. Sacramento, CA: California Department of Education. Retrieved from ~~http://www.caaspp.org/rsc/resources/CAST.practice-scoring-guide-gr8.2018-19.pdf~~

California Department of Education. (2019g). California Science Test Practice Items Scoring Guide High School. Sacramento, CA: California Department of Education. Retrieved from ~~http://www.caaspp.org/rsc/resources/CAST.practice-scoring-guide-hs.2018-19.pdf~~

Educational Testing Service. (2014). *ETS Standards for Quality and Fairness*. Princeton, NJ: Educational Testing Service. Retrieved from https://www.ets.org/s/about/pdf/standards.pdf

Smarter Balanced Assessment Consortium. (2014). *Smarter Balanced Assessment Consortium: Usability, accessibility, and accommodations implementation guide.* Los Angeles: Smarter Balanced Assessment Consortium. Retrieved from ~~https://portal.smarterbalanced.org/library/en/usability-accessibility-and-accommodations-implementation-guide.pdf~~

Smarter Balanced Assessment Consortium. (2018). *Smarter Balanced Resources and Practices Comparison Crosswalk.* Los Angeles: Smarter Balanced Assessment Consortium. Retrieved from ~~https://portal.smarterbalanced.org/library/en/uaag-resources-and-practices-comparison-crosswalk.pdf~~

Smarter Balanced Assessment Consortium. (2019). *Smarter Balanced Assessment Consortium: Usability, accessibility, and accommodations guidelines.* Los Angeles: Smarter Balanced Assessment Consortium. Retrieved from https://portal.smarterbalanced.org/library/en/usability-accessibility-and-accommodations-guidelines.pdf

# Chapter 6: Standard Setting

This chapter summarizes the standard setting process through which California Science Test (CAST) threshold scores and achievement levels were established. Included are a background of the development of the CAST, an overview of the standard setting methodology, a summary of the standard setting procedure, the description of the achievement level descriptors (ALDs) (California Department of Education [CDE], 2019), the panel recommendations, and the results. The detailed standard setting information for the CAST is described in the *Standard Setting Technical Report for the California Science Test* (Educational Testing Service [ETS], 2019).

## 6.1. Background

The CAST is an online assessment aligned with the California Next Generation Science Standards (CA NGSS) (CDE, 2019). The first operational administration of the CAST occurred during the 2018–2019 CAASPP administration. Standard setting was required so that threshold scores and achievement levels were available for the fall 2019 release of results in the CAST score reports. Achievement levels for the CAST are as follows:

- Level 4—Standard Exceeded
- Level 3—Standard Met
- Level 2—Standard Nearly Met
- Level 1—Standard Not Met

Figure 6.1 presents the score reporting hierarchy for the CAST, which was approved in November 2019 by the California State Board of Education (SBE), along with the General (Policy) ALDs describing the expectations at each performance level across grade five, grade eight, and high school.

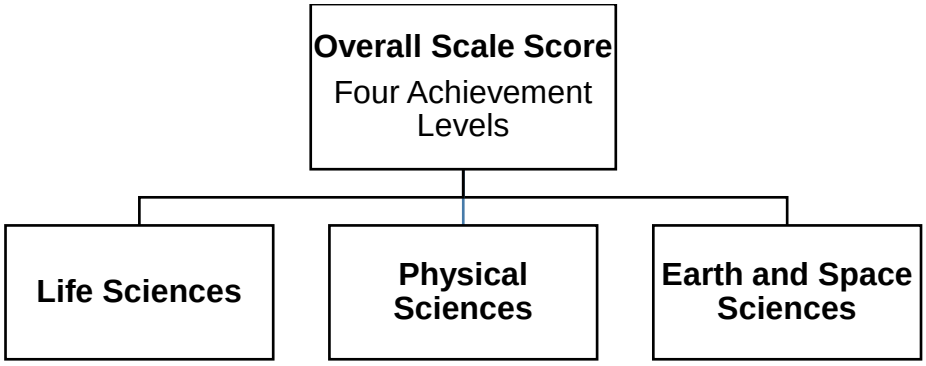


**Figure 6.1 CAST score reporting hierarchy**

To develop threshold-score recommendations aligned to the score-reporting hierarchy, ETS conducted standard setting workshops in Sacramento, California, for the CAST for grade five, grade eight, and high school on August 6–9, 2019. The overall score was considered in the standard setting process.

## 6.2. Achievement Level Descriptors (ALDs)

The CAST General ALDs are descriptors that convey the expectation at each achievement level across grade five, grade eight, and high school (CDE, 2018). They were provided to the panelists for prereading prior to the standard setting workshop.

Using the General ALDs, as well as the blueprint, a panel of educators familiar with both students taking the CAST and the CA NGSS were recruited to develop CAST Range ALDs, which include more detailed expectations for each domain and are grade specific. The panel reviewed the SBE-approved general ALDs and a draft of CAST Range ALDs. The result of the panel meeting was a revised draft; these Range ALDs were reviewed and approved by the CDE following the educator panel review, on May 24, 2019. Panelists referenced the SBE-approved general ALDs and the Range ALDs as part of the standard setting process, to develop definitions of the expectations at the threshold of the performance levels, called the borderline student definitions.

## 6.3. Standard Setting Methodology

Standard setting refers to a class of methodologies, the result of which is that one or more achievement threshold scores are applied to a scale score to report achievement levels. The purpose of the standard setting process for the CAST was to collect recommendations from California educators for the placement of the threshold scores for review by the CDE, with final determination and approval by the SBE.

For the overall score, the standard setting panel recommended threshold scores to indicate the score that must be earned for a student to reach the beginning (i.e., threshold) of three of the four achievement levels (levels 2 through 4); final recommendations were made for the overall score.

Specifically, the Modified Angoff and the Extended Angoff methods were implemented to collect panelists' judgments on 1-point and 2-point items, respectively.

### 6.3.1. Modified and Extended Angoff Methods

The Modified Angoff method (Angoff, 1971; Brandon, 2004; Hambleton & Pitoniak, 2006) is a probability-based standard setting method. For 1-point items, each panelist judged the item on the likelihood that the borderline student would answer the item correctly.

The Extended Angoff method (Cizek & Bunch, 2007; Hambleton & Plake, 1995) was used for the 2-point items. For these items, the task was to decide on the assigned score value that would most likely be earned by the borderline student for each constructed-response item; panelists reviewed the scoring rules or scoring rubric for each of these items. In standard setting, one of the critical components is to include a standard setting panel of experts who can provide appropriate consideration and judgments. The panel begins by becoming familiar with the test and considering the content assessed and the relative difficulty of the items. The test-familiarization stage also allows the panelists to experience the test in a manner that is similar to an operational test administration, which allows the panelists to get a sense of the test taker's experience. After independently reviewing the assessment, the panelists discuss the content measured and the relative difficulty of the items.

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

During this review, each panelist independently considers the level of knowledge and skills required to respond to the item as well as the features of a response that would earn a particular score as defined by the scoring rubric. Each panelist decides on the score most

likely to be earned by each borderline student from the possible values a student can earn. Panelists are reminded to refer to the knowledge and skills of the borderline student definition and the scoring rules and not to expect the three levels to match to the three possible scores.

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

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

## 6.4. Standard Setting Procedures

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

### 6.4.1. Panelists

Prior to the standard setting, panelists were recruited to include a diverse, representative group of California educators who have experience in the science education of students in grades five or eight or high school who take the CAST as well as familiarity with the CA NGSS.

Panelists were assigned to one of three panels of educators; each panel focused on one grade-level CAST: grade five, grade eight, or high school. The number of panelists from this population of educators was 15 for grades five and eight and 16 for high school.

Because standard setting is based on expert judgment—informed by student performance data—it is important that panelists collectively reflect the diversity of the students and the educators working with students who take the assessment. Special efforts were made to assemble panels that were representative of the geographic and socioeconomic diversity of California in general and the CAST student and educator population. Panels included a sample across genders, ethnic and racial backgrounds, and geographical regions in California. A majority of the educators indicated they had more than 10 years' experience teaching science. All panels were primarily comprised of teachers currently teaching science. More detailed information about panelists' demographic characteristics and background is provided in S*tandard Setting Technical Report for the California Science Test* (ETS, 2019).

## 6.4.2. Materials

At the standard setting workshop, panelists received training materials and a set of operational materials. Materials and data were based on the pool of items administered in 2018–2019. For each CAST assessment, the following materials were provided to each panelist:

- CAST general and range ALDs
- Test familiarization materials
  - Paper assessment for entering answers and notes
  - Answer key with scoring rules where appropriate
  - Rubrics
- Judgment materials
  - Survey forms on tablets, one per panelist
- Practice and training materials
- Impact data based on the 2018–2019 administration of the CAST
- Evaluation forms
  - Training evaluation form
  - Final evaluation form
- Workshop agenda

## 6.4.3. Process

The workshop process included a general session, where all panelists were provided an overview of the purpose of the meeting, their role and the roles of facilitators and observers, and an explanation of the approach used in the standard setting for the CAST. Educators were then guided to grade-specific panel rooms, where they completed the training and judgment process (Brandon, 2004; Cizek & Bunch, 2007).

### 6.4.3.1. Training

Training was provided on the following topics:

- Test familiarization
- CAST general and range ALDs for the panelists' assigned grade and the CA NGSS
- Development of borderline student definitions
- Standard setting judgment process for the modified and extended Angoff methods
  - Training and practice prior to the first round of judgments
- Interpretation of feedback data for the 1-point and 2-point items

### 6.4.3.2. Judgments and Feedback

Feedback was provided to the panelists based on Round 1 judgments. Panelists reviewed item-level judgment information for each item; this allowed panelists to identify where their judgment was closer to other panelists or more diverse. Panelists discussed judgments and rationales and made notes. Additionally, the mean, minimum, maximum, and range of the overall score based on panel judgments (from low to high) were projected in the room. After discussing all of this feedback, panelists were asked to make an independent Round 2 judgment on the items for all levels.

Round 2 results were again displayed and discussed. After the panelists discussed the data, the student performance data showing the impact (or consequence) data for the

Round 2 judgments was presented. This performance data was based on 2018–2019 CAST student performance. The feedback showed what percentage of students would fall into each level based on these decisions. After the room-level discussions, panelists were invited to continue with table-level discussions as needed.

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

Panelists were reminded at this time that there would be additional reviews by the CDE and final approval by the SBE.

## 6.5. Results of the Standard Setting

Results from the CAST standard setting after Round 3 included recommended thresholds for each test (grade five, grade eight, and high school). The *Standard Setting Technical Report for the California Science Test* (ETS, 2019) presents details about the following results from the standard setting workshops:

- The mean threshold score recommendations for each test at the end of each round
- Standard errors of judgment, scale scores, and conditional standard errors of measurement at or above the recommended threshold scores

Table 6.1 through table 6.3 show the projected percentage of students statewide who would be placed at each achievement level based on the results of the 2018–2019 operational test administration of the CAST in each of the administered grade levels. The threshold score is the minimum score (on standard setting scale score metric) needed to achieve an achievement level.

Scales provided in these tables were presented and used in the standard setting process and are more user-friendly than scores in the theta metric. However, it should be noted that the scores presented are ***not*** the reported scale scores for the CAST. The scale used in the tables was created, based on the 2018–2019 operational test data, for standard setting (prior to the approval of the official scale for the CAST) and was used as a tool for the standard setting process.

**Table 6.1 Projected Distribution of 2018–2019 Students Based on Round 3 Recommendations: Grade Five**

| Achievement Level | Threshold Score | Percentage |
|---|---|---|
| Level 1 | N/A | 18.1 |
| Level 2 | 177 | 44.4 |
| Level 3 | 207 | 25.2 |
| Level 4 | 229 | 12.3 |

### Table 6.2 Projected Distribution of 2018–2019 Students Based on Round 3 Recommendations: Grade Eight

| Achievement Level | Threshold Score | Percentage |
|---|---|---|
| Level 1 | N/A | 8.8 |
| Level 2 | 170 | 54.4 |
| Level 3 | 209 | 26.7 |
| Level 4 | 231 | 10.1 |

### Table 6.3 Projected Distribution of 2018–2019 Students Based on Round 3 Recommendations: High School

| Achievement Level | Threshold Score | Percentage |
|---|---|---|
| Level 1 | N/A | 16.5 |
| Level 2 | 174 | 55.8 |
| Level 3 | 213 | 23.3 |
| Level 4 | 238 | 4.4 |

Results presented in the *Standard-Setting Technical Report for the California Science Test* (ETS, 2019) are based on the standard setting workshop and panel-recommended threshold scores at the end of the workshop. This consisted of recommended threshold scores for each performance level for the overall scores for each grade. Following the standard setting workshop, the SBE reviewed both the panel recommendations and the State Superintendent of Public Instruction's recommendations for threshold scores. The final threshold values were established by the SBE.

# References

Angoff, W.H. (1971). Scales, norms and equivalent scores. In R.L. Thorndike (Ed.), Educational measurement (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard setting topics. *Applied Measurement in Education, 17*(1), 59–88.

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

California Department of Education. (2019). *NGSS for California public schools, K-12.* Sacramento, CA*:* California Department of Education*.* Retrieved from https://www.cde.ca.gov/pd/ca/sc/ngssstandards.asp.

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

Educational Testing Service (2019). *Standard Setting Technical Report for the California Science Test*. Princeton, NJ: Educational Testing Service.

Hambleton, R. K., & Plake, B.S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments*. Applied Measurement in Education, 8:1*, 41–55*.*

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.). *Educational Measurement* (4th ed., pp. 433–70). Westport, CT: Praeger.

# Chapter 7: Scoring and Reporting

This chapter summarizes the scoring at the item level, including the types of scoring approaches that are used for each item type in the California Science Test (CAST) operational test and the approach implemented to produce student scores. This chapter also describes scores reported at the individual student level and various reports that are generated.

## 7.1. Scoring for Constructed-Response Items

The CAST at each tested grade level included selected-response (SR) and constructed-response (CR) items. The SR items are machine-scored and the CR items are scored by either human scoring or the artificial intelligence (AI) scoring engine.

All the CR items used in the operational segments (i.e., segments A and B) were field-tested during the 2017–2018 administration in which AI models were built and evaluated. Some CR items have AI models approved for operational use.

Out of the 27 CR items used in segments A and B of the 2018–2019 operational forms across three grades, 14 of them were human-scored and 13 were AI-scored. A random sample of approximately 9,600 responses for each CR prompt were double-scored by human raters. Items were double-scored to provide a measure of interrater reliability for quality control of the CR item scoring.

For all the field test CR prompts, not all students' responses were scored by human raters. Instead, a random sample of responses drawn from each field test CR item were scored and those first human ratings were used to support the item analyses, item response theory (IRT) analyses, and construction of AI scoring models for potential use in future operational administrations. A portion of the sampled responses was randomly selected and double-scored (second human ratings) for validating the first human ratings.

### 7.1.1. Sampling Process

There were two CR item sampling processes for the 2018–2019 administration: one for operational items (for double scoring only) and the other for field test items (for both first and second human ratings). The simple random sampling and, in some cases, stratified sampling methods for demographic student groups were used.

#### 7.1.1.1. Sampling Process for Double Scoring of Operational Constructed-Response Items

For the 2018–2019 administration, the CAST program double-scored approximately 9,600 students for each operational CR item for the purposes of reporting interrater agreement statistics and quality control of CR item scoring. The double scoring also supported the evaluation of new AI models.

The sampling for double scoring of operational CR items was conducted randomly, by item, at the time of scoring in the Educational Testing Service (ETS) Online Network for Evaluation (ONE). The sample was a representation of population demographics, including gender, ethnicity, special education services status, English language fluency, and economic status.

#### 7.1.1.2. Sampling Process for Field Test Constructed-Response Items

Each student taking the CAST was expected to receive five blocks: two discrete blocks in Segment A, two performance tasks (PTs) in Segment B, and one discrete block or one PT in Segment C. Field test CR items were sampled by block, meaning if a student was sampled for a block, that student's responses to all CR items in that block were scored.

For CR prompts in grades five and eight assessments, 1,800 responses per prompt were sampled. For high school, because students may test in grade ten, eleven, or twelve, the percentage of students from each grade taking the test could be drastically different. To account for the uncertainty in sampling composition by grades, 2,500 responses per prompt were sampled for high school. Out of these sampled responses—for grades five and eight (1,800 for each prompt) and for high school (2,500 for each prompt)—approximately 800 responses per prompt were randomly selected and double-scored.

##### *7.1.1.2.1. Creating the Sampling Frame*

The sampling 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 a 5 percent difference of the population compositions found in the California Longitudinal Pupil Achievement Data System—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 fifteen 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 C15.

Certain exclusion rules were applied when creating the sampling frame to remove cases that were more likely to confound, rather than inform, the results. For example, for each block within each grade level, students who completed their assigned items in less than the minimum testing time for the block were considered unmotivated. The first percentile of average item time for multiple-choice (MC) items from the CAST field test in 2017–2018 for each grade level multiplied by the total number of items in a block was used to provide conservative estimates of the minimum time a student who was not motivated would take to complete a block. These minimum times per item were 3 seconds for grade five, 2.4 seconds for grade eight, and 1.8 seconds for high school.

##### *7.1.1.2.2. Selection of Random Samples*

Simple random sampling was used to draw samples. The demographic compositions for the selected samples were then checked against the population demographics. If the difference was less than 5 percent, the sample was accepted. However, as previously mentioned, only certain blocks 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 over-assigned to these blocks and under-assigned to other blocks, resulting in deviations from the population demographic composition. The most affected student group was students in special education programs. Thus, for these blocks, 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.

## 7.1.2. Human Scoring

### 7.1.2.1. Scoring Rubric Development

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 section *3.4 Content Expert Review*. Rubrics were edited as needed on the basis of feedback from the California Department of Education (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.

### 7.1.2.2. Range Finding

Range finding is the process of identifying student responses that will be used as anchor (benchmark) samples to help ensure that CR items are scored consistently and reliably.

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 Assessment & Learning Technology Development (ALTD) 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 needed for various purposes are summarized in table 7.1.

**Table 7.1 CAST 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 | Two to four samples for each score point represented per set |
| Training | Training samples with annotations for rater training and scoring practice | One set of 10 samples per grade for each prompt | Two to four samples for each score point per set |

Table 7.1 *(continuation)*

| Sample Type | Purpose | Number of Sets and Samples in Sets | Configuration of Sets |
|---|---|---|---|
| Benchmarks | Benchmark samples with annotations that represent exemplar responses at each score point on the rubric | One set of six to eight 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 | 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 | Four to ten samples for each score point |

2. Responses were scored by two independent, experienced raters using the ONE system. ETS ALTD 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 7.1).
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.

#### 7.1.2.3. Rater Recruitment and Certification process

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 included California educators as well as other raters representing a variety of backgrounds in business, education, and other fields.

Table 7.2 shows the characteristics of the CAST raters. Among the 459 raters with teaching experience in science, 61 currently worked in a K–12 school in California, 5 were fluent in Spanish, and 393 had experience teaching in a K–12 school. Approximately 1,500 raters were used for the 2018–2019 CAST, scoring 9,522,138 operational responses as well as 98,104 field test responses across the three grade levels.

**Table 7.2 Summary of Characteristics of Human Raters Scoring the CAST**

| Characteristic | N |
|---|---|
| Experience teaching in Science | 459 |
| Fluent in Spanish | 37 |
| Experience teaching in a kindergarten (K)–12 school | 272 |
| Currently works in a K–12 school in California | 230 |
| Others—Not meeting any of the previous criteria | 500 |
| **Total raters scoring in 2018–2019** | **1,498** |

Certification served as an initial screening to ensure that ETS' Scoring and Reporting Operations (SRO) team had a sufficient number of qualified raters in place to meet the demands of scoring. One 2-point prompt (i.e., 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, the rater 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 the 2018–2019 CAST items.

#### 7.1.2.4. Rater and Scoring Leader Training

ETS selected scoring leaders to oversee a group of raters during the scoring process. Scoring leaders were experienced raters who had demonstrated high scoring accuracy from previous scoring projects at ETS and were invited to act as a scoring leader on a project. For the 2018–2019 CAST 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 SRO with scoring quality.

#### 7.1.2.5. 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 communication with raters, how to monitor raters, and other information necessary for their role during scoring.

#### 7.1.2.6. 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.

##### *7.1.2.6.1. 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.

##### *7.1.2.6.2. 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 2018–2019 CAST 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.

##### *7.1.2.6.3. 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. Raters typically 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 2018–2019 CAST, calibration was set at once per item during a three-day period.

For the 2018–2019 CAST 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.

#### 7.1.2.7. Scoring Rules and Processes

ETS implemented the following scoring rules and processes for CAST operational and field test scoring:

- Operational responses were scored via both human and AI scoring.
  - Human scoring was comprised of approximately 57 percent of all responses, with AI scoring accounting for the remaining 43 percent.
  - Approximately 9,600 responses per item were double-scored as part of continuous quality management. Raters were not aware when a second scoring was occurring and did not have access to the first score.
- Field test responses were only scored via human scoring.
  - Approximately 40 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.
- For the 2018–2019 CAST scoring, the use of condition codes was retired. Raters were instructed to apply zero (0) scores when there was an attempt to answer the question but the information was incorrect. If the rater was unsure, the rater deferred responses to the scoring leader.
- For field test items only, ETS psychometric staff provided a sampling plan that included the responses selected to be scored. Refer to subsection *7.1.1.2 Sampling Process for Field Test Constructed-Response Items* for the sampling plan. The sampling plan was uploaded to ONE to activate the responses for scoring.

#### 7.1.2.8. 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 might 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 ALTD, 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).

#### 7.1.2.9. Interrater Reliability for Operational Items

The ONE system captured interrater reliability by monitoring data for responses that are double-scored. Approximately 9,600 CAST responses per item were double scored. The statistics included the percentage agreement between the two raters, kappa, and the quadratic-weighted kappa (QWK). For detailed descriptions of these statistics, refer to subsection *8.7.2.1 Interrater Agreement* in *Chapter 8: Analyses*. 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 shown in table 7.3 through table 7.8. Table 7.3 through table 7.5 include the operational items that were human-scored. Table 7.6 through table 7.8 include the operational items that were AI-scored, where Rater 1 scores in these tables refer to the AI scores.

These tables show that the percentage of students for whom the human raters were in exact agreement ranged from 77.65 percent to 97.10 percent for 1-point items and 63.99 percent to 91.95 percent for 2-point items across all grade levels. The percentage of students for whom the human and AI raters were in exact agreement ranged from 78.75 percent to 95.63 percent for 1-point items and 74.77 percent to 86.90 percent for 2-point items. The exception was one 2-point item in high school where the exact agreement was 48.12 percent.

These tables also show that the QWK ranged from 0.51 to 0.94 for human-scored items and 0.56 to 0.90 for AI-scored items, which indicates a moderate-to-high level of agreement between two raters.

**Table 7.3 Interrater Reliability and Descriptive Statistics for the Ratings by Two Raters in Human-Scoring of Operational Items for Grade Five**

| Prompt | Item ID | Score Points | Rater 1 N | Rater 2 N | Kappa | QWK | Percent Exact | Percent Adjacent | Percent Exact + Adjacent | Rater 1 Item Score Mean | Rater 1 Item Score SD | Rater 2 Item Score Mean | Rater 2 Item Score SD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | VH709025 | 2 | 9,598 | 9,598 | 0.87 | 0.91 | 91.95 | 7.97 | 99.92 | 0.67 | 0.70 | 0.66 | 0.70 |
| 2 | VH733167 | 2 | 9,599 | 9,599 | 0.44 | 0.62 | 63.99 | 33.39 | 97.37 | 0.79 | 0.76 | 0.78 | 0.76 |
| 3 | VH737471 | 1 | 9,599 | 9,599 | 0.54 | 0.54 | 77.65 | 22.35 | 100.00 | 0.59 | 0.49 | 0.59 | 0.49 |
| 4 | VH811101 | 1 | 9,598 | 9,598 | 0.51 | 0.51 | 80.46 | 19.54 | 100.00 | 0.28 | 0.45 | 0.27 | 0.44 |
| 5 | VH813229 | 2 | 9,597 | 9,597 | 0.79 | 0.91 | 86.77 | 12.90 | 99.67 | 1.14 | 0.89 | 1.14 | 0.89 |
| N/A | **AVERAGE** | N/A | 9,598 | 9,598 | 0.63 | 0.70 | 80.16 | 19.23 | 99.39 | 0.69 | 0.66 | 0.69 | 0.66 |

### Table 7.4 Interrater Reliability and Descriptive Statistics for the Ratings by Two Raters in Human-Scoring of Operational Items for Grade Eight

| Prompt | Item ID | Score Points | Rater 1 N | Rater 2 N | Kappa | QWK | Percent Exact | Percent Adjacent | Percent Exact + Adjacent | Rater 1 Item Score Mean | Rater 1 Item Score SD | Rater 2 Item Score Mean | Rater 2 Item Score SD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | VH738912 | 2 | 9,599 | 9,599 | 0.83 | 0.94 | 90.09 | 9.72 | 99.81 | 1.01 | 0.94 | 1.01 | 0.94 |
| 2 | VH803496 | 1 | 9,597 | 9,597 | 0.66 | 0.66 | 86.44 | 13.56 | 100.00 | 0.28 | 0.45 | 0.28 | 0.45 |
| 3 | VH804554 | 2 | 9,600 | 9,600 | 0.57 | 0.71 | 73.69 | 25.63 | 99.31 | 0.72 | 0.69 | 0.73 | 0.70 |
| 4 | VH809423 | 2 | 9,600 | 9,600 | 0.73 | 0.88 | 82.51 | 17.36 | 99.88 | 1.00 | 0.88 | 1.00 | 0.88 |
| 5 | VH809632 | 1 | 9,599 | 9,599 | 0.59 | 0.59 | 84.20 | 15.80 | 100.00 | 0.26 | 0.44 | 0.26 | 0.44 |
| 6 | VH811932 | 2 | 9,597 | 9,597 | 0.64 | 0.77 | 81.28 | 16.64 | 97.92 | 0.50 | 0.73 | 0.50 | 0.73 |
| N/A | **AVERAGE** | N/A | 9,599 | 9,599 | 0.67 | 0.76 | 83.03 | 16.45 | 99.49 | 0.63 | 0.69 | 0.63 | 0.69 |

### Table 7.5 Interrater Reliability and Descriptive Statistics for the Ratings by Two Raters in Human-Scoring of Operational Items for High School

| Prompt | Item ID | Score Points | Rater 1 N | Rater 2 N | Kappa | QWK | Percent Exact | Percent Adjacent | Percent Exact + Adjacent | Rater 1 Item Score Mean | Rater 1 Item Score SD | Rater 2 Item Score Mean | Rater 2 Item Score SD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | VH702164 | 2 | 9,600 | 9,600 | 0.54 | 0.67 | 75.23 | 23.65 | 98.88 | 0.49 | 0.65 | 0.49 | 0.65 |
| 2 | VH804669 | 1 | 9,599 | 9,599 | 0.77 | 0.77 | 97.10 | 2.90 | 100.00 | 0.07 | 0.25 | 0.07 | 0.25 |
| 3 | VH807293 | 2 | 9,602 | 9,602 | 0.44 | 0.61 | 70.07 | 27.71 | 97.78 | 0.50 | 0.69 | 0.49 | 0.69 |
| N/A | **AVERAGE** | N/A | 9,600 | 9,600 | 0.58 | 0.68 | 80.80 | 18.08 | 98.89 | 0.35 | 0.53 | 0.35 | 0.53 |

### Table 7.6 Interrater Reliability and Descriptive Statistics for the Ratings by AI and Human Raters in AI-Scoring of Operational Items for Grade Five

| Prompt | Item ID | Score Points | Rater 1 N | Rater 2 N | Kappa | QWK | Percent Exact | Percent Adjacent | Percent Exact + Adjacent | Rater 1 Item Score Mean | Rater 1 Item Score SD | Rater 2 Item Score Mean | Rater 2 Item Score SD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | VH667949 | 1 | 9,516 | 9,516 | 0.76 | 0.76 | 87.84 | 12.16 | 100.00 | 0.49 | 0.50 | 0.47 | 0.50 |
| 2 | VH668026 | 1 | 9,524 | 9,524 | 0.58 | 0.58 | 78.75 | 21.25 | 100.00 | 0.61 | 0.49 | 0.45 | 0.50 |
| 3 | VH810103 | 1 | 9,510 | 9,510 | 0.85 | 0.85 | 95.63 | 4.37 | 100.00 | 0.18 | 0.38 | 0.17 | 0.37 |
| 4 | VH810308 | 1 | 9,510 | 9,510 | 0.74 | 0.74 | 89.98 | 10.02 | 100.00 | 0.25 | 0.44 | 0.28 | 0.45 |
| N/A | **AVERAGE** | N/A | 9,515 | 9,515 | 0.73 | 0.73 | 88.05 | 11.95 | 100.00 | 0.38 | 0.45 | 0.34 | 0.45 |

### Table 7.7 Interrater Reliability and Descriptive Statistics for the Ratings by AI and Human Raters in AI-Scoring of Operational Items for Grade Eight

| Prompt | Item ID | Score Points | Rater 1 N | Rater 2 N | Kappa | QWK | Percent Exact | Percent Adjacent | Percent Exact + Adjacent | Rater 1 Item Score Mean | Rater 1 Item Score SD | Rater 2 Item Score Mean | Rater 2 Item Score SD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | VH695226 | 2 | 9,508 | 9,508 | 0.67 | 0.83 | 82.11 | 17.33 | 99.44 | 1.46 | 0.73 | 1.44 | 0.78 |
| 2 | VH728143 | 2 | 9,536 | 9,536 | 0.61 | 0.83 | 74.77 | 24.73 | 99.50 | 0.95 | 0.86 | 0.93 | 0.92 |
| 3 | VH803535 | 2 | 9,539 | 9,539 | 0.79 | 0.90 | 86.90 | 13.00 | 99.90 | 0.70 | 0.80 | 0.70 | 0.81 |
| 4 | VH803647 | 1 | 9,538 | 9,538 | 0.76 | 0.76 | 89.33 | 10.67 | 100.00 | 0.70 | 0.46 | 0.63 | 0.48 |
| N/A | **AVERAGE** | N/A | 9,530 | 9,530 | 0.71 | 0.83 | 83.28 | 16.43 | 99.71 | 0.95 | 0.71 | 0.92 | 0.75 |

### Table 7.8 Interrater Reliability and Descriptive Statistics for the Ratings by AI and Human Raters in AI-Scoring of Operational Items for High School

| Prompt | Item ID | Score Points | Rater 1 N | Rater 2 N | Kappa | QWK | Percent Exact | Percent Adjacent | Percent Exact + Adjacent | Rater 1 Item Score Mean | Rater 1 Item Score SD | Rater 2 Item Score Mean | Rater 2 Item Score SD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | VH651810 | 1 | 9,452 | 9,452 | 0.90 | 0.90 | 95.10 | 4.90 | 100.00 | 0.43 | 0.49 | 0.44 | 0.50 |
| 2 | VH651815 | 2 | 9,523 | 9,523 | 0.71 | 0.75 | 84.44 | 15.54 | 99.98 | 0.56 | 0.54 | 0.51 | 0.57 |
| 3 | VH696269 | 2 | 9,447 | 9,447 | 0.23 | 0.56 | 48.12 | 50.99 | 99.11 | 0.73 | 0.69 | 1.14 | 0.76 |
| 4 | VH730945 | 2 | 9,522 | 9,522 | 0.61 | 0.74 | 76.89 | 22.95 | 99.83 | 0.61 | 0.59 | 0.66 | 0.73 |
| 5 | VH804572 | 1 | 9,565 | 9,565 | 0.84 | 0.84 | 91.96 | 8.04 | 100.00 | 0.49 | 0.50 | 0.46 | 0.50 |
| N/A | **AVERAGE** | N/A | 9,502 | 9,502 | 0.66 | 0.76 | 79.30 | 20.48 | 99.78 | 0.56 | 0.56 | 0.64 | 0.61 |

The CAST provides the following flagging criteria to identify operational items to be reviewed for potential elimination after scoring is completed. ETS monitored CAST activity throughout the scoring period and adjusted the training and scoring processes. ETS will continue the monitoring process and make improvements as needed.

Polytomous items are flagged if any of the following conditions occur:

- Exact + adjacent agreement < 0.80
- Exact agreement < 0.70
- QWK < 0.70

Dichotomous items are flagged if either of the following conditions occur:

- Exact agreement < 0.80
- QWK < 0.70

Table 7.9 shows the number of items flagged by grade and by scoring method. There were nine flagged items among the 27 operational items across grade levels. Of the nine flagged items, seven were flagged for human-human ratings and two for human-AI ratings. Flagged items were subsequently reviewed by content specialists.

**Table 7.9 Number of Operational Constructed-Response Items Flagged by Scoring Method**

| Scoring Method | Grade or Grade Level | Flagged Polytomous Items | Flagged Dichotomous | Total Flagged Items | Total Number of Scored Items | Percentage Flagged |
|---|---|---|---|---|---|---|
| Human scoring | Grade 5 | 1 | 2 | 3 | 5 | 60 |
| Human scoring | Grade 8 | 0 | 2 | 2 | 6 | 33 |
| Human scoring | High School | 2 | 0 | 2 | 3 | 67 |
| AI scoring | Grade 5 | 0 | 1 | 1 | 4 | 25 |
| AI scoring | Grade 8 | 0 | 0 | 0 | 4 | 0 |
| AI scoring | High School | 1 | 0 | 1 | 5 | 20 |

The evaluation of CR items with new, approved, AI models are presented in subsection *7.5.3 Model Evaluation*.

### 7.1.2.10. 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 used 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 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 ALTD 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 ALTD and CR Scoring Systems and Capabilities staff reviewed raters' scoring patterns and made 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.

## 7.2. Scoring for Selected-Response Items

CAST 2018–2019 assessments included machine-scorable, traditional MC items and technology-enhanced 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 quality control are provided in section *9.5 Quality Control of Scoring*.

## 7.3. Student Test Scores

ETS developed two parallel scoring systems to produce students' scores: the Enterprise Score Key Management (eSKM) scoring system, which scores and delivers individual students' scores to the ETS reporting system; and the parallel scoring system developed by ETS Psychometric Analysis and Research (PAR), which computes individual students' scores. The two scoring systems independently applied the same scoring algorithms and specifications. ETS psychometricians verified the eSKM scoring by comparing all individual student scores from PAR and resolving any discrepancies. This process was an internal quality control step that is in place to verify the accuracy of scoring. Students' scores were reported only when the two parallel systems produced identical results with acceptable tolerance.

All scores must comply 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 (TOMS).

### 7.3.1. Theta Scores

After IRT item calibration is conducted to obtain item parameter estimates, the student's theta score is then computed by the inverse test characteristic curve method (Stocking, 1996) via an iterative process. Refer to section *8.5 IRT Analyses* for more details on the IRT models and calibration. This method transforms the sum of the student's item scores into an ability estimate. That estimate is the ability level at which the sum of the expected scores on the items that the student took is equal to the sum of the raw scores that the student actually earned on those items. The range of theta scores is -4 to 4.

The same method is used to estimate a student's theta score for both the overall test and domain scores.

Individual student theta score distributions are presented in table 7.A.1 through table 7.A.6 in appendix 7.A for grades five, eight, ten, eleven, and twelve and high school, respectively.

### 7.3.2. Scale Scores

The CAST uses the IRT model to estimate students' abilities (i.e., theta scores) and then uses the IRT true score equating method to convert the theta scores to number right (NR) scores on a base form, which is comprised of 100 Rasch items with a difficulty of 0. A total theta score is converted to an NR score by the following formula:

$$NR = n\frac{e^{\theta_j}}{1+e^{\theta_j}} \tag{7.1}$$

*Refer to the Alternative Text for Equation 7.1 for a description of this equation.*

Where,

$n$ = 100.

Because all forms are equated to one base form, the NR scores account for the form difficulty differences and are comparable across forms. Because the NR scores can easily be misinterpreted as raw scores, a transformation is needed to convert them to scale scores to facilitate score interpretation. Table 7.10 shows the scaling constants for the linear transformation of an NR score to a scale score.

**Table 7.10 Scaling Constants**

| Grade or Grade Level | Slope | Intercept |
|---|---|---|
| Grade 5 | 1.0081 | 151 |
| Grade 8 | 1.0081 | 351 |
| High school | 1.0081 | 551 |

The transformation constants are derived by mapping the lowest obtainable NR score to the lowest obtainable scale score (LOSS) plus one, and the highest obtainable NR score to the highest obtainable scale score. The solutions to these linear equations are the transformation constants as shown in table 7.10.

The ranges of the reporting scale scores are 150–250, 350–450, and 550–650 for grade five, grade eight, and high school, respectively.

The CAST reports scale scores for the total test for students who have answered at least 10 items. Those who did not answer any items on the test received NS (no score) in their score

report. Those who answered one to nine items received the LOSS. The LOSS for grade five, grade eight, and high school is 150, 350, and 550, respectively.

The CAST is only considered "complete" if a student responds to at least a minimum number of operational items for the total test. Table 7.11 lists the minimum number of operational items required to fulfill the completion requirement for the CAST for the 2018–2019 administration. For example, the minimum number of item requirements for grade five is 46.

**Table 7.11 Minimum Number of Item Requirements for Test Completion**

| Grade or Grade Level | Minimum Number of Items | Life Sciences | Physical Sciences | Earth and Space Sciences |
|---|---|---|---|---|
| Grade 5 | 46 | 7 | 13 | 8 |
| Grade 8 | 43 | 7 | 8 | 7 |
| High school | 46 | 8 | 8 | 8 |

For students who answered at least 10 items but did not complete the test, a proportional adjustment on NR is used to provide an equitable score to all students. The amount of adjustment for incomplete test takers is proportional to the fraction of the test completed.

Table 7.12 shows the mean and standard deviation of both scale score and theta scores for the CAST.

**Table 7.12 Mean and Standard Deviation of Theta Scores and Scale Scores**

| Grade or Grade Level | Number of Students | Scale Score Mean | Scale Score SD | Theta Score Mean | Theta Score SD |
|---|---|---|---|---|---|
| Grade 5 | 456,221 | 201 | 22.2 | 0.0 | 1.1 |
| Grade 8 | 462,504 | 401 | 22.5 | 0.0 | 1.1 |
| High School—Grade 10 | 23,322 | 597 | 21.4 | -0.2 | 1.0 |
| High School—Grade 11 | 256,842 | 601 | 22.4 | 0.0 | 1.1 |
| High School—Grade 12 | 276,105 | 597 | 22.7 | -0.2 | 1.1 |
| High School—All Grades | 556,269 | 599 | 22.6 | -0.1 | 1.1 |

Individual student scale score distributions are presented in table 7.B.1 through table 7.B.6 in appendix 7.B for grades five, eight, ten, eleven, and twelve and high school, respectively.

### 7.3.3. Achievement Levels

After the 2018–2019 first operational test administration, a standard setting was conducted, and achievement levels were established. Student performance on the reporting scale is designated into one of four achievement levels:

- **Level 1—Standard Not Met:** Student demonstrates a minimal understanding of and ability to apply the knowledge and skills associated with the performance expectations of the California Next Generation Science Standards (CA NGSS).
- **Level 2—Standard Nearly Met:** Student demonstrates a partial understanding of and ability to apply the knowledge and skills associated with the performance expectations of the CA NGSS.

- **Level 3—Standard Met:** Student demonstrates an adequate understanding of and ability to apply the knowledge and skills associated with the performance expectations of the CA NGSS.
- **Level 4—Standard Exceeded:** Student demonstrates a thorough understanding of and ability to apply the knowledge and skills associated with the performance expectations of the CA NGSS.

The scale score ranges for achievement levels are shown in table 7.13. Percentages of students in each achievement level are in table 7.14 and their graphic representation is displayed in figure 7.1.

### Table 7.13 Scale Score Ranges for Achievement Levels

| Grade or Grade Level | Standard Not Met | Standard Nearly Met | Standard Met | Standard Exceeded |
|---|---|---|---|---|
| Grade 5 | 150–178 | 179–213 | 214–230 | 231–250 |
| Grade 8 | 350–377 | 378–414 | 415–432 | 433–450 |
| High school | 550–575 | 576–614 | 615–635 | 636–650 |

### Table 7.14 Percent of Students in Each Achievement Level for Total Scores

| Grade or Grade Level | Number of Students Tested | Percent in Achievement Level Standard Not Met | Percent in Achievement Level Standard Nearly Met | Percent in Achievement Level Standard Met | Percent in Aachievement Level Exceeded | Percent in Aachievement Level Met/Exceeded |
|---|---|---|---|---|---|---|
| Grade 5 | 456,221 | 18.9 | 49.4 | 19.7 | 12.0 | 31.7 |
| Grade 8 | 462,504 | 18.1 | 51.0 | 20.8 | 10.0 | 30.8 |
| High school—Grade 10 | 23,322 | 19.5 | 57.4 | 18.9 | 4.3 | 23.1 |
| High school—Grade 11 | 256,842 | 15.4 | 54.2 | 23.2 | 7.2 | 30.4 |
| High school—Grade 12 | 276,105 | 19.8 | 54.5 | 19.1 | 6.6 | 25.6 |
| High school—All grades | 556,269 | 17.8 | 54.5 | 20.9 | 6.8 | 27.7 |

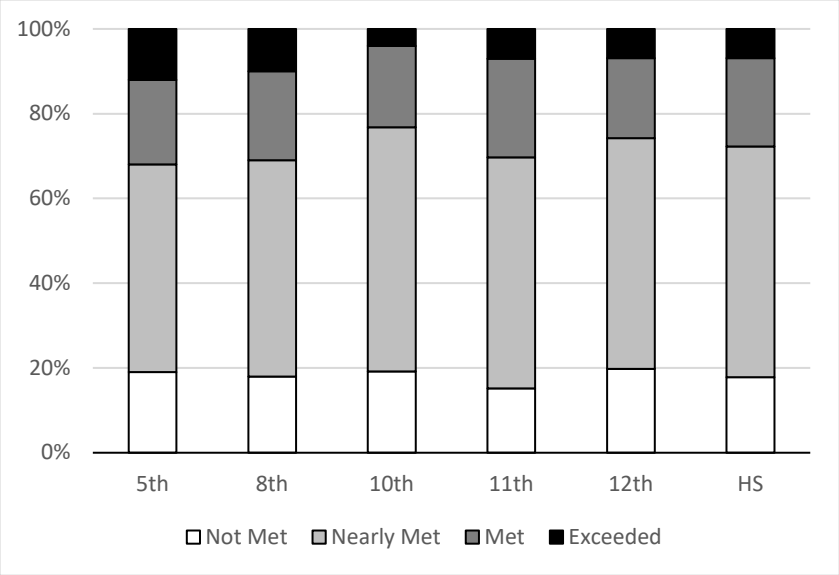


**Figure 7.1 Percentage of achievement levels**

Overall, about one third of grade five (31.7%) and grade eight (30.8%) students met or exceeded the standards, but slightly more than one quarter of high school students (27.7%) met or exceeded the standards. For high school students, the percentage was the lowest for grade ten and the highest for grade eleven.

Demographic summaries of achievement levels of total scores are presented in table 7.C.1 through table 7.C.6 in appendix 7.C for grades five, eight, ten, eleven, and twelve and high school, respectively. The description of the demographic student groups included in these tables are shown in table 5.3 in *Chapter 5: Test Administration*. The percentages of students who met or exceeded standards were similar between male and female students.

Those student groups with high percentages of students who met or exceeded standards were initial fluent English proficient, Asian, or not economically disadvantaged. In contrast, those groups with low percentages of students who met or exceeded standards were English learners, African Americans, or economically disadvantaged.

### 7.3.4. Domain Achievement Levels

Domain achievement levels are reported for students who have met the minimum number of item requirements for the domain as listed in table 7.11. Students might receive domain achievement levels for some domain(s) but not the others depending on the number of items they completed for different domains. Domain achievement levels are not reported for students who are in the LOSS condition (i.e., answered fewer than 10 items for the total test).

There are three domains for each test: Life Sciences, Physical Sciences, and Earth and Space Sciences. The detailed descriptions of the three achievement levels for each domain are shown in table 7.15.

**Table 7.15 Description of Domain Achievement Levels**

| Science Domain | Below Standard | Near Standard | Above Standard |
|---|---|---|---|
| Life Sciences | The student demonstrates minimal understanding of and ability to apply the knowledge and skills associated with the core ideas, concepts, and practices in the Life Sciences, which focus on structures and processes in living things, ecosystems, heredity, and biological evolution. | The student demonstrates some understanding of and ability to apply the knowledge and skills associated with the core ideas, concepts, and practices in the Life Sciences, which focus on structures and processes in living things, ecosystems, heredity, and biological evolution. | The student demonstrates a thorough understanding of and ability to apply the knowledge and skills associated with the core ideas, concepts, and practices in the Life Sciences, which focus on structures and processes in living things, ecosystems, heredity, and biological evolution. |
| Physical Sciences | The student demonstrates minimal understanding of and ability to apply the knowledge and skills associated with the core ideas, concepts, and practices in Physical Sciences, which focus on matter and its interactions, motion and stability, energy, and waves and their applications. | The student demonstrates some understanding of and ability to apply the knowledge and skills associated with the core ideas, concepts, and practices in the Physical Sciences, which focus on matter and its interactions, motion and stability, energy, and waves and their applications. | The student demonstrates a thorough understanding of and ability to apply the knowledge and skills associated with the core ideas, concepts, and practices in the Physical Sciences, which focus on matter and its interactions, motion and stability, energy, and waves and their applications. |

Table 7.15 *(continuation)*

| Science Domain | Below Standard | Near Standard | Above Standard |
|---|---|---|---|
| Earth and Space Sciences | The student demonstrates minimal understanding of and ability to apply the knowledge and skills associated with the core ideas, concepts, and practices in Earth and Spaces Sciences, which focus on Earth's place in the universe, Earth's systems, and Earth and human activity. | The student demonstrates some understanding of and ability to apply the knowledge and skills associated with the core ideas, concepts, and practices in the Earth and Space Sciences, which focus on Earth's place in the universe, Earth's systems, and Earth and human activity. | The student demonstrates a thorough understanding of and ability to apply the knowledge and skills associated with the core ideas, concepts, and practices in the Earth and Spaces Sciences, which focus on Earth's place in the universe, Earth's systems, and Earth and human activity. |

A student is assigned to one of the three achievement levels for a domain according to the following rules:

- Place in the Below Standard level if $\theta_d < \theta_{L3} - 1.5\,\overline{\mathrm{SEM}(\theta_d)}$
- Place in the Near Standard level if $\theta_{L3} - 1.5\,\overline{\mathrm{SEM}(\theta_d)} \leqslant \theta_d < \theta_{L3} + 1.5\,\overline{\mathrm{SEM}(\theta_d)}$
- Place in the Above Standard level if $\theta_d \geqslant \theta_{L3} + 1.5\,\overline{\mathrm{SEM}(\theta_d)}$

where

$\theta_d$ is a domain theta score,

$\overline{\mathrm{SEM}(\theta_d)}$ is the mean standard error of the domain theta scores, and

$\theta_{L3}$ is the level 3 theta threshold score of the total test.

This domain achievement level estimation method for CAST is similar to but slightly different from the Smarter Balanced approach for a claim performance level. CAST uses the mean standard error of domain theta scores, while Smarter Balanced uses the standard error of an individual student claim theta score.

Across all grade levels, the percentages of students in the Above Standard achievement level ranged from 9.2 to 12.3 for Life Sciences, from 6.1 to 14.3 for Physical Sciences, and from 6.7 to 15.2 for Earth and Space Sciences. The majority of students achieved either Near Standard or Above Standard levels.

Compared with the percentages of Below Standard, the percentages of Near Standard were higher in Life Sciences (except for grade five) and Earth and Space Sciences but lower in Physical Sciences.

Table 7.16 through table 7.18 show the percentages of domain achievement levels for Life Sciences, Physical Sciences, and Earth and Space Sciences, respectively.

### Table 7.16 Percent of Students in Each Achievement Level for the Life Sciences Domain

| Grade or Grade Level | Number of Students Tested | Below Standard | Near Standard | Above Standard |
|---|---|---|---|---|
| Grade 5 | 455,748 | 45.1 | 42.6 | 12.3 |
| Grade 8 | 461,580 | 39.7 | 48.8 | 11.5 |
| High school—Grade 10 | 23,247 | 38.9 | 51.9 | 9.2 |
| High school—Grade 11 | 256,300 | 33.6 | 54.1 | 12.3 |
| High school—Grade 12 | 275,174 | 39.4 | 49.5 | 11.1 |
| High school—All grades | 554,721 | 36.7 | 51.7 | 11.6 |

### Table 7.17 Percent of Students in Each Achievement Level for the Physical Sciences Domain

| Grade or Grade Level | Number of Students Tested | Below Standard | Near Standard | Above Standard |
|---|---|---|---|---|
| Grade 5 | 455,429 | 44.1 | 41.6 | 14.3 |
| Grade 8 | 461,484 | 43.9 | 43.9 | 12.2 |
| High school—Grade 10 | 23,237 | 53.4 | 40.6 | 6.1 |
| High school—Grade 11 | 256,245 | 45.9 | 43.9 | 10.2 |
| High school—Grade 12 | 275,087 | 53.0 | 38.2 | 8.8 |
| High school—All grades | 554,569 | 49.8 | 40.9 | 9.3 |

### Table 7.18 Percent of Students in Each Achievement Level for the Earth and Space Sciences Domain

| Grade or Grade Level | Number of Students Tested | Below Standard | Near Standard | Above Standard |
|---|---|---|---|---|
| Grade 5 | 455,804 | 41.6 | 46.7 | 11.7 |
| Grade 8 | 461,231 | 42.0 | 42.8 | 15.2 |
| High school—Grade 10 | 23,243 | 44.7 | 48.6 | 6.7 |
| High school—Grade 11 | 256,297 | 38.9 | 51.1 | 10.0 |
| High school—Grade 12 | 275,206 | 44.2 | 46.9 | 8.9 |
| High school—All grades | 554,746 | 41.8 | 48.9 | 9.3 |

Demographic summaries of domain achievement levels are presented in appendix 7.D for grades five, eight, ten, eleven, and twelve and high school. Table 7.D.1 through table 7.D.6 are for Life Sciences, table 7.D.7 through table 7.D.12 are for Physical Sciences, and table 7.D.13 through table 7.D.18 are for Earth and Space Sciences. The description of the demographic student groups included in these tables are shown in table 5.3 in *Chapter 5: Test Administration*.

## 7.3.5. Theta Scores Standard Errors

The conditional standard error of measurement (CSEM) is the standard deviation of the distribution of theta scores that the student would earn under different testing conditions.

In the framework of IRT, the CSEM is the reciprocal of the square root of the test information function (TIF) based on the items taken by each student. It is also the estimate of standard error for the estimate of theta. The TIF is the sum of information from each item on the test. The CSEM for a student with proficiency $\theta_j$ is

$$SE_{\theta_j} = \frac{1}{\sqrt{I(\theta_j)}} \tag{7.2}$$

*Refer to the Alternative Text for Equation 7.2 for a description of this equation.*

where,

$I(\theta_j)$ is the test information for student $j$, calculated as

$$I(\theta_j) = \sum_{i=1}^{n} I_i(\theta_j) \tag{7.3}$$

*Refer to the Alternative Text for Equation 7.3 for a description of this equation.*

and $I_i(\theta_j)$ is the item information of item $i$ for student $j$.

Item information based on the generalized partial credit model for both dichotomous and polytomous items is calculated as

$$I_i(\theta_j) = (Da_i)^2[s_{i2}(\theta_j) - s_i^2(\theta_j)] \tag{7.4}$$

*Refer to the Alternative Text for Equation 7.4 for a description of this equation.*

where,

$S_i(\theta_j)$ is the expected item score for item $i$ on a theta score $\theta_j$, calculated as

$$s_i(\theta_j) = \sum_{h=0}^{n_i} hp_{ih}(\theta_j) \tag{7.5}$$

*Refer to the Alternative Text for Equation 7.5 for a description of this equation.*

and

$$s_{i2}(\theta_j) = \sum_{h=0}^{n_i} h^2 p_{ih}(\theta_j) \tag{7.6}$$

*Refer to the Alternative Text for Equation 7.6 for a description of this equation.*

where,

$P_{ih}(\theta_j)$ is the probability of an examinee with $\theta_j$ getting score $h$ on item $i$, the computation of which is shown in equation 8.7,

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

$D$ is a scaling constant of 1.7 that makes the logistic model approximate the normal ogive model.

The CSEM is calculated based only on the answered item(s) for both complete and incomplete tests.

### 7.3.6. Scale Score Standard Errors

The conditional standard errors of theta score can be transformed onto the reporting scale. This transformation is

$$CSEM_{\text{Scale Score}} = A \bullet SE_{\theta_j} \bullet n \bullet \frac{e^{\theta_j}}{(1+e^{\theta_j})^2} \quad (7.7)$$

*Refer to the Alternative Text for Equation 7.7 for a description of this equation.*

where,

$n$=100,

$A$ is the scaling constant that equals to 1.0081 as defined in table 7.10, and

$SE_{\theta_j}$ is the standard error estimate of $\theta_j$ defined in equation 7.2.

## 7.4. Reports Produced and Scores for Each Report

The CAST provides results or score summaries that are reported for different purposes. The four major purposes are to

1. help facilitate conversations between parents/guardians and teachers about student performance,
2. serve as a tool to help parents/guardians and teachers work together to improve student learning,
3. help schools and school districts identify strengths and areas that need improvement in their educational programs, and
4. provide the public and policymakers with information about student achievement.

This section provides detailed descriptions of the uses and applications of the CAST reporting for students.

### 7.4.1. Online Reporting

TOMS is a secure website hosted by ETS that permits local educational agency (LEA) users to manage the CAST online assessments and to inform the TDS. This system uses a role-specific design to restrict access to certain tools and applications based on the user's designated role. Specific functions of TOMS include the following:

- Manage user access privileges
- Manage test administration calendars and testing windows
- Manage student test assignments
- Manage and confirm the accuracy of students' test settings (i.e., designated supports and accommodations) prior to testing
- Generate and download various reports

In addition to TOMS, there are two California online reporting systems: the Online Reporting System (ORS) and the California Educator Reporting System.

TOMS communicates with the ORS, which provides authorized users with interactive and cumulative online reports for the CAST at the student, school, and LEA levels. The ORS

provides access to two CAST reports: Score Reports, which provide preliminary score data for each administered test available in the reporting system; and the Completion Status Reports, which provide completion data in the reporting system for students taking an assessment.

Based on the CAST reporting requirements, the ORS provides the preliminary summative reports containing information outlining student knowledge and skills, as well as performance levels aligned to the assessment-specific claims. The online aggregate reports provide functionality at the student, classroom, school, and LEA levels. The online aggregate reports are available to be downloaded in PDF, Excel, and comma-separated value formats.

## 7.4.2. Special Cases

Student scores are not reported for the following cases:

- Student was absent from the test administration
- Student moved or had a medical emergency during testing
- Student's parent/guardian requested exemption from testing
- Student did not log on to test systems
- Student was invalidated in the system (not reported in aggregated reporting)

## 7.4.3. Types of Score Reports

There are three categories of CAST reports. The specific reports within each category are presented in this subsection.

### 7.4.3.1. Student Score Report

The CAST Student Score Report is the official score report for parents or guardians and includes the following metrics:

- Scale score for the CAST reported
- Achievement levels for the CAST reported (CAST achievement levels are "Standard Exceeded," "Standard Met," "Standard Nearly Met," and "Standard Not Met.")
- Domain-specific achievement levels for CAST reported (The domain-specific achievement levels are "Above Standard," "Near Standard," and "Below Standard.")

Scores for students who were assigned accommodations or designated supports are reported in the same way as for students who were not assigned accommodations or designated supports. Detailed information about accessibility resources is described in subsection *2.4 Universal Tools, Designated Supports, and Accommodations*.

In all, LEAs had four options for accessing and distributing Student Score Reports to parents/guardians:

1. Accessing electronic Student Score Reports using a locally provided parent or student portal
2. Downloading Student Score Reports from TOMS and making them available electronically using a secure local method
3. Downloading Student Score Reports from TOMS, printing them, and making them available locally
4. Purchasing paper Student Score Reports from Educational Testing Service

Further information about the Student Score Report and other reports is provided on the CDE CAASPP Student Score Report Information web page at ~~https://www.cde.ca.gov/ta/tg/ca/caasppssrinfo.asp.~~

#### 7.4.3.2. School Report

The school performance report provides group information by content area, including the school's average scale score and the percentage of students at each achievement level. This report also provides a list of students' scale scores and achievement levels.

The school scale score report is presented as a dashboard to provide group information by content area. It includes a histogram showing the distribution of students' scale scores.

These reports may be found in the ORS.

#### 7.4.3.3. District Report

The district performance report provides school-level information by content area, including the school average scale score and the percentage of students at each achievement level.

This report lists all the proficiency information for each school, including the testing status as shown in subsection *7.4.2 Special Cases*, number of students who completed testing, average scale score, and percentage of students in each achievement level.

The district scale score report is presented as a dashboard to provide cumulative information. A histogram is included to show the frequency of schools with mean scale scores in each score interval.

The CAASPP student data files for the LEA are available for the LEA CAASPP coordinator to download from TOMS. The LEA CAASPP coordinator forwards the appropriate reports to test sites.

Internet reports are accessible to the public online on the Test Results for California's Assessments website at https://caaspp-elpac.cde.ca.gov/caaspp/.

Preliminary individual student scores are also available to LEAs prior to the release of final reports via electronic reporting that can be accessed using the ORS. This application permits LEAs to view preliminary results for all tests taken.

### 7.4.4. Score Report Applications

CAST results provide parents and guardians with information about their child's progress. The results are a tool for increasing communication and collaboration between parents/guardians and teachers. The Student Score Report can be used by parents and guardians while talking with teachers about ways to improve their child's achievement of the CA NGSS.

Schools may use the CAST results to help make decisions about how best to support student achievement. CAST results, however, should never be used as the only source of information to make important decisions about a child's education.

CAST results help schools and LEAs identify strengths and weaknesses in their instructional programs. Each year, staff from schools and LEAs examine CAST results at each grade level tested. Their findings are used to help determine

- the extent to which students are learning the academic standards,
- instructional areas that can be improved,

- teaching strategies that can be developed to address needs of students, and
- decisions about how to use funds to ensure that students achieve the standards.

### 7.4.5. Criteria for Interpreting Test Scores

An LEA may use CAST results to help make decisions about student placement, promotion, retention, or other considerations related to student achievement. However, it is important to remember that a single test can provide only limited information. Other relevant information should be considered as well. Additionally, the 2018–2019 administration is the first operational administration and it is possible that the CA NGSS are not fully implemented across the state. It is advisable for parents to evaluate their child's strengths and weaknesses in the relevant topics by reviewing classroom work and progress reports in addition to the child's CAST online assessment results. It is also important to note that a student's score could vary somewhat if the student were retested.

### 7.4.6. Criteria for Interpreting Score Reports

The information presented in various reports must be interpreted with caution when making performance comparisons. When comparing scale score and achievement-level results, the user is limited to comparisons within a grade or grade band for high school. The user may compare scale scores for the same grade, within a school, between schools, or between a school and its LEA, its county, or the state.

The user can also make comparisons within the same grade across years. Caution should be taken when comparing scale scores from different grades, because the curricula are different across grade levels.

For more details on the criteria for interpreting information provided on the score reports, refer to https://startingsmarter.org/.

## 7.5. New Artificial Intelligence (AI) Model Building

After the 2018–2019 administration, AI models were built and evaluated for operational and new field test items which were human scored in 2018–2019. Approved AI models will be used for future operational scoring.

ETS built models for 47 field test items and 9 operational items that were human scored during the 2018–2019 administration. The breakdown of item counts by grade level is shown in table 7.19:

**Table 7.19 Number of Items for New AI Model Building by Grade**

| Grade or Grade Level | Number of Field Test Items | Number of Operational Items | Total |
|---|---|---|---|
| Grade 5 | 16 | 3 | 19 |
| Grade 8 | 15 | 4 | 19 |
| High school | 16 | 2 | 18 |
| **Total** | **47** | **9** | **56** |

Of the 56 AI models that were built, 31 of them were approved. The evaluation process of the 31 AI models is presented in subsection *7.5.3 Model Evaluation*.

### 7.5.1. Data Collection

After the CAST, ETS collected a sample of students' responses to 56 CR items with human score(s) assigned, as described in subsection *7.1.1 Sampling Process*.

### 7.5.2. 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.[9] Each model was built using a 10-fold cross-validation method that randomly splits the entire dataset for an item in 10 subsets. 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 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 QWK, using the predictions from each testing instance. The evaluations performed are reported in the next subsection.

[9] A Support Vector Machine (SVM) performs classification by finding the hyperpla7ne 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 7.2 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 2 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.

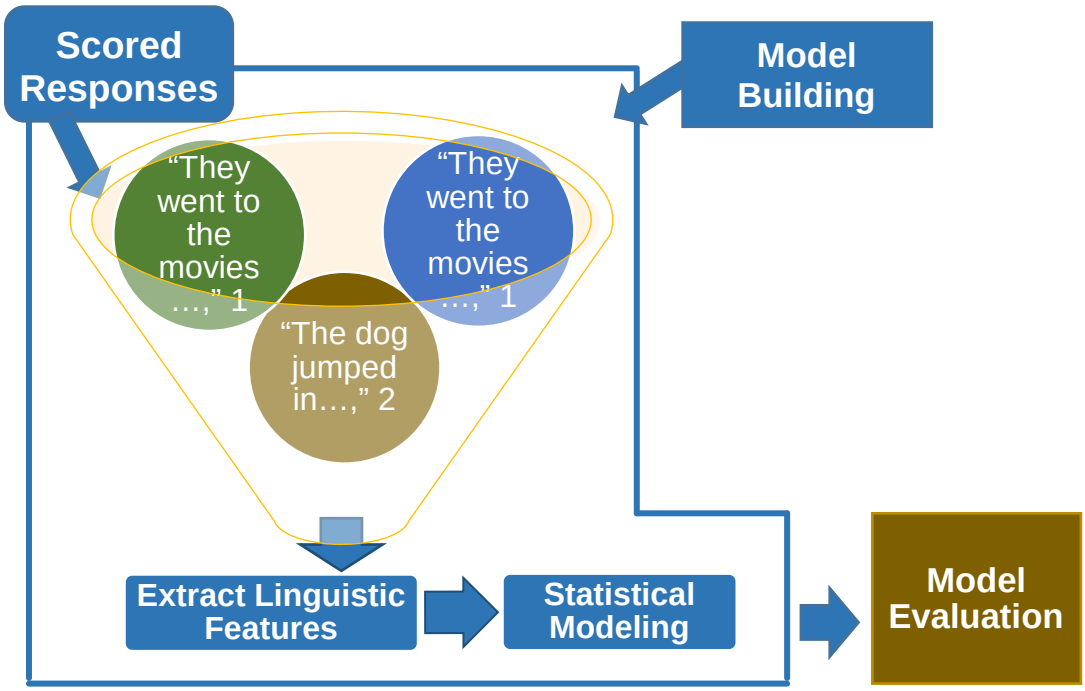


**Figure 7.2 Model building and evaluation process**

## 7.5.3. Model Evaluation

One of the important factors in building AI scoring models with good performance is the use of data with reliable human scores. A commonly used indicator for evaluating human scoring reliability is to use multiple raters on a large enough sample of responses and evaluate the extent to which they agree with each other.

Each item response had two human ratings: first and second human ratings. The second human ratings were available only on those randomly selected item responses that were double scored. The first human ratings were used to build and evaluate an AI model and the second human ratings were used to validate the first human ratings. The evaluation of an AI model includes human-human agreement, human-AI agreement, and the comparison of the two. High human-human agreement indicates that the human ratings used to build the AI model are reliable and high human-AI agreement that is similar to the human-human agreement for the item indicates that the AI model performs as expected.

# 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.

Stocking, M. L. (1996). *An alternative method for scoring adaptive tests*. Journal of Educational and Behavioral Statistics, 21, 365–389.

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

# Accessibility Information

## Alternative Text for Equation 7.1

NR is equal to n times the fraction with the numerator exponent based theta sub j and the denominator one plus exponent theta sub j.

## Alternative Text for Equation 7.2

SEM of Theta sub j equals 1 divided by the square root of I of theta sub j.

## Alternative Text for Equation 7.3

I of Theta sub j equals the sum from i equals 1 to n of I sub i of Theta sub j.

## Alternative Text for Equation 7.4

I sub i of Theta sub j equals open parenthesis D times a sub i close parenthesis squared times open bracket s sub i2 of theta sub j minus s squared sub i of theta sub j close bracket.

## Alternative Text for Equation 7.5

S sub i of Theta sub j equals the sum from h equals 0 to n sub i of h times p sub ih of theta sub j.

## Alternative Text for Equation 7.6

S sub i2 of Theta sub j equals the sum from h equals 0 to n sub i of h squared times p sub ih of theta sub j.

## Alternative Text for Equation 7.7

CSEM of sub scale score equals A times SE of theta sub j times n times the fraction with the numerator exponent based theta sub j and the denominator open parenthesis of one plus exponent theta sub j close parenthesis square.

# Chapter 8: Analyses

This chapter summarizes the results of the analyses on the data from the California Science Test (CAST) 2018–2019 operational test administration, including classical item analyses, differential item functioning (DIF) analyses, the dimensionality study, item response theory (IRT) calibration, response time analyses, reliability, validity, and two research studies (multistage adaptive test [MST] study and content screen-out study).

## 8.1. Sample Used for the Analyses

Two item analyses were run for the CAST: 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, all student responses to the operational constructed-response (CR) items and only a sample of student responses from the field test CR items were scored (refer to subsection *7.1.1.2 Sampling Process for Field Test Constructed-Response Items* for details). A two-wave approach to the PIA was taken to mitigate any scheduling risks in terms of meeting the data review timeline. The first wave involved conducting the PIA for all machine-scored items across all three segments. The second wave analysis was for all the CR items and it was conducted at a later time when the sampled responses were all scored.

The FIA was conducted near the end of the administration. Available student responses that met the inclusion rule were included in the analyses. The inclusion rules used in CAST item analyses and item calibration were as follows:

- 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” were treated as “incorrect.” Not-reached items were not included in the calculation of student scores.

Any items flagged during the PIA were sent to the data review committee (Refer to section *3.5 Data Review Meeting* for more details) for review. The California Department of Education (CDE) then made final decisions on the acceptance or rejection of the items based on the data review results. Items that were rejected by the CDE were not included in the FIA and the IRT calibration.

## 8.2. 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 well.

## 8.2.1. Classical Item Difficulty Indices (*p*-value)

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.

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

$$p\text{–}value_{dich} = \frac{\sum X_{ic}}{N_i} \tag{8.1}$$

*Refer to the Alternative Text for Equation 8.1 for a description of this equation.*

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) or *p*-value. The AIS can range from 0.00 to the maximum total possible points for an item. 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

$$p\text{–}value_{poly} = \frac{\sum_j X_{ij}}{N_i \times Max(X_i)} \tag{8.2}$$

*Refer to the Alternative Text for Equation 8.2 for a description of this equation.*

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 $i$.

Acceptable *p*-values for both dichotomous and polytomous items for CAST are between 0.20 and 0.95. Items with *p*-values outside this range were flagged for review.

## 8.2.2. Item-Total Score 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 block design of the CAST generated many different combinations of blocks. Because it would take extensive amount of time to collect sufficient volume by form to run the PIA for the CAST, the PIA was, instead, run first for the combined two Segment A blocks and then for segment B and C blocks separately.

For the combined Segment A block run, the total number of raw score points for the two Segment A blocks was used as the criterion score in calculating the item-total correlations. Since the blocks in segments B and C were shorter, the total score from the block being analyzed might not be stable enough to be used as a criterion score. Therefore, for the other runs—separately for segment B and C blocks—the total number of raw score points of the machine-scored items from the two Segment A blocks, plus the scores from the segment B or C block being analyzed, 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

$$r_{polyreg} = \frac{\hat{\beta} s_{tot}}{\sqrt{\hat{\beta}^2 s_{tot}^2 + 1}} \tag{8.3}$$

*Refer to the Alternative Text for Equation 8.3 for a description of this equation.*

where,

$\beta$ 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 (SD) of the criterion (the students' total score).

For a polytomous item, there is a regression for each boundary between item scores, with all regressions for the same item sharing a common slope, *β*. For a polytomous item with $m$ possible score values, there are $k$-1 regressions. Beta (*β*) is the common slope for all $m$-1 regressions.

Acceptable values for this correlation are positive and greater than 0.20. A relatively high item-total correlation coefficient value is preferred, 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.

### 8.2.3. Distribution of Item Scores

For polytomous items, examination of the distribution of scores assists in showing how well the items performed. If no students were given the highest possible score, the item may not be functioning as expected because 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 3 percent) of students obtaining any score point were flagged for review.

## 8.2.4. Omission Rates

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.

## 8.2.5. Distractor Analyses

### 8.2.5.1. The Proportion of Students Choosing Each Distractor

For the CAST, distractor analyses were conducted on selected-response (SR) 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).

### 8.2.5.2. Distractor-Total Correlation

For SR 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 was calculated for the distractors, like the item-total correlation previously described, except that the regressions were implemented on the distractors rather than the keys. Items with positive distractor-total correlations were flagged for review, as these items may have multiple correct answers, be miskeyed, or have other content issues.

## 8.2.6. Summary of Classical Item Analyses Flagging Criteria

In summary, items were 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 an item-total polyserial correlation less than 0.20 are flagged.
- An **omit flag** is set if the nonresponse rates are greater than 5 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 3 percent of the students at any score level.

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

## 8.2.7. Classical Item Analysis Results

This subsection provides the summary tables of operational item distributions for the item difficulty and the item discrimination statistics. The overall item difficulty distributions are presented in table 8.1. Across grade levels, most items had *p*-value between 0.2 and 0.8 and only a few were outside of this range. Item difficulty distributions by item type are shown in table 8.A.1; item difficulty distributions by content domain are presented in table 8.A.2.

### Table 8.1 Item Difficulty Distributions

| Grade or Grade Level | $0 \le p < 0.2$ | $0.2 \le p < 0.4$ | $0.4 \le p < 0.6$ | $0.6 \le p < 0.8$ | $0.8 \le p \le 1.0$ | Total Number of Items |
|---|---|---|---|---|---|---|
| Grade 5 | 4 | 20 | 24 | 15 | 1 | 64 |
| Grade 8 | 3 | 31 | 21 | 11 | 1 | 67 |
| High school—Grade 10 | 3 | 26 | 19 | 3 | 1 | 52 |
| High school—Grade 11 | 2 | 24 | 19 | 5 | 2 | 52 |
| High school—Grade 12 | 2 | 29 | 17 | 4 | 0 | 52 |
| High school—All grades | 2 | 26 | 18 | 5 | 1 | 52 |

Overall item-total correlation distributions are presented in table 8.2. Across grade levels, the item-total correlations were 0.2 or higher, except for only a few items. No item-total correlations were negative. Item-total correlation distributions by item type are shown in table 8.B.1; item-total correlation distributions by content domain are presented in table 8.B.2.

### Table 8.2 Item-Total Correlation Distributions

| Grade or Grade Level | $r < 0$ | $0 \le r < 0.2$ | $0.2 \le r < 0.3$ | $0.3 \le r < 0.4$ | $0.4 \le r < 0.5$ | $r \ge 0.5$ | Total Number of Items |
|---|---|---|---|---|---|---|---|
| Grade 5 | 0 | 0 | 4 | 5 | 10 | 45 | 64 |
| Grade 8 | 0 | 3 | 2 | 12 | 10 | 40 | 67 |
| High school—Grade 10 | 0 | 6 | 4 | 5 | 13 | 24 | 52 |
| High school—Grade 11 | 0 | 5 | 4 | 7 | 9 | 27 | 52 |
| High school—Grade 12 | 0 | 5 | 3 | 7 | 9 | 28 | 52 |
| High school—All grades | 0 | 5 | 4 | 6 | 9 | 28 | 52 |

# 8.3. 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 2018–2019 CAST items that meet the sample size requirements. 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 to 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.

## 8.3.1. DIF procedure for 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 8.4 (Dorans & Holland, 1993). The resulting estimate is interpreted as the relative probability of success on a particular item for members of two groups when matched on ability.

$$\alpha_{MH} = \frac{\left(\sum_m R_{rm} \frac{W_{fm}}{N_{tm}}\right)}{\left(\sum_m R_{fm} \frac{W_{rm}}{N_{tm}}\right)} \tag{8.4}$$

*Refer to the Alternative Text for Equation 8.4 for a description of this equation.*

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):

$$MH\ D\text{-}DIF = -2.35 \ln\left[\alpha_{MH}\right] \tag{8.5}$$

Refer to the *Alternative Text for Equation 8.5* for a description of this equation.

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).

## 8.3.2. 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:

$$SMD = \frac{\sum_{m=1}^{M} N_{fm} \times E_f(Y \mid X = m)}{\sum_{m=1}^{M} N_{fm}} - \frac{\sum_{m=1}^{M} N_{fm} \times E_r(Y \mid X = m)}{\sum_{m=1}^{M} N_{fm}} = \frac{\sum_{m=1}^{M} D_m}{\sum_{m=1}^{M} N_{fm}} \tag{8.6}$$

Refer to the *Alternative Text for Equation 8.6* for a description of this equation.

where,

$X$ is the criterion score (total raw score),

$Y$ is the item score,

$M$ is the number of score levels on $X$,

$D_m$ 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).

### 8.3.3. 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 8.3; the flagging criteria for polytomous items are provided in table 8.4.

**Table 8.3 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.<br>• 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<br>• Absolute value of MH D-DIF is significantly different from one, but is less than 1.5.<br>• 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.<br>• Positive values are classified as "C+" and negative values as "C-." |

**Table 8.4 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 met the sample size requirements. These comparison groups are specified in table 8.5.

**Table 8.5 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 |

Table 8.5 *(continuation)*

| DIF Type | Focal Group | Reference Group |
|---|---|---|
| English fluency | English learner | English only |
| Disability | Special education services | No special education services |
| Economic status | Economically disadvantaged | Not economically disadvantaged |

### 8.3.4. Differential Item Functioning Analysis Results

Summarized DIF results for the operational items are given in table 8.6, table 8.7, and table 8.8 for grade five, grade eight, and high school, respectively. If the sample size requirement for conducting DIF analyses was not met, that item was categorized in "insufficient counts."

One item was identified with C-level DIF in each of grades five and eight, and three items with C-level DIF were identified in high school. Items that show C-level DIF and are considered biased by the DIF review panel will be deactivated for future use.

Test developers are instructed to avoid selecting C-level items considered unbiased by the DIF review panel for future test forms unless their inclusion is deemed essential to meeting test-content specifications.

**Table 8.6 Number of Items 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 | 61 | 0 | 2 | 0 | 1 | 0 |
| Asian–White | 64 | 0 | 0 | 0 | 0 | 0 |
| Black–White | 64 | 0 | 0 | 0 | 0 | 0 |
| Hispanic–White | 64 | 0 | 0 | 0 | 0 | 0 |
| American Indian or Alaska Native–White | 64 | 0 | 0 | 0 | 0 | 0 |
| English learner–English only | 64 | 0 | 0 | 0 | 0 | 0 |
| Special education services–No special education services | 63 | 1 | 0 | 0 | 0 | 0 |
| Economically disadvantaged–Not economically disadvantaged | 64 | 0 | 0 | 0 | 0 | 0 |

**Table 8.7 Number of Items 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 | 63 | 2 | 1 | 0 | 1 | 0 |
| Asian–White | 67 | 0 | 0 | 0 | 0 | 0 |
| Black–White | 67 | 0 | 0 | 0 | 0 | 0 |
| Hispanic–White | 67 | 0 | 0 | 0 | 0 | 0 |
| American Indian or Alaska Native–White | 67 | 0 | 0 | 0 | 0 | 0 |
| English learner–English only | 66 | 1 | 0 | 0 | 0 | 0 |
| Special education services–No special education services | 67 | 0 | 0 | 0 | 0 | 0 |
| Economically disadvantaged–Not economically disadvantaged | 67 | 0 | 0 | 0 | 0 | 0 |

**Table 8.8 Number of Items 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 | 51 | 0 | 1 | 0 | 0 | 0 |
| Asian–White | 51 | 0 | 1 | 0 | 0 | 0 |
| Black–White | 52 | 0 | 0 | 0 | 0 | 0 |
| Hispanic–White | 52 | 0 | 0 | 0 | 0 | 0 |
| American Indian or Alaska Native–White | 52 | 0 | 0 | 0 | 0 | 0 |
| English learner–English only | 46 | 2 | 1 | 3 | 0 | 0 |
| Special education services–No special education services | 52 | 0 | 0 | 0 | 0 | 0 |
| Economically disadvantaged–Not economically disadvantaged | 52 | 0 | 0 | 0 | 0 | 0 |

## 8.4. Test Dimensionality Analyses

The California Next Generation Science Standards (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 performance tasks (PTs) measure something different than the discrete items?
- Do the technology-enhanced items (TEIs) measure anything different from the traditional item types (e.g., multiple-choice [MC] 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 focus of this dimensionality study is to examine the dimensional structure of the CAST and find a model that best fits the data to provide a practical guideline in terms of calibration and reporting.

The methodology and results of this study are reviewed in *Chapter 12: Test Dimensionality Study Addendum*. Refer also to chapter 3 of the report *Informing the California Science Test (CAST) Blueprint Improvements: Results from the Psychometric Studies* (ETS, 2019) for additional information. Note that the report was developed to provide empirical evidence to support the test blueprint improvement. Therefore, in addition to evaluating the test dimensionality with the operational items, additional conditions that involve a field test PT were included. It should be noted that this technical report only describes the study with the operational items.

### 8.4.1. Form Selection

The CAST operational test used a block design, where each segment included multiple blocks and each student was randomly assigned a portion of the blocks. For each grade-level assessment, there were two blocks in Segment A, so all students took the same two Segment A blocks. However, because there were multiple blocks from segments B and C, each student was randomly assigned two PTs from two different domains in Segment B and randomly assigned either one discrete block or one PT from Segment C. This created multiple combinations of blocks in Segment A and Segment B that a student could receive.

Instead of conducting the analyses on all possible combinations of PTs, three forms (i.e., block combinations) were carefully selected for evaluation for each grade using the following guidelines:

- All items performed reasonably well based on the item-analysis results.
- The Segment B PTs in three forms cover three different combinations of the content domains (i.e., Life Sciences and Physical Sciences PTs for form 1, Life Sciences and Earth and Space Sciences PTs for form 2, and Physical Sciences and Earth and Space Sciences PTs for form 3).

The analysis sample includes all students who have taken the selected forms.

### 8.4.2. 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 chapter 3 of the *Informing the California Science Test (CAST) Blueprint Improvements: Results from the Psychometric Studies* (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., MC items vs. TEIs)
5. A division of the discrete items from those assigned to PTs

Evaluating the dimensionality of a test is a subjective judgment, based on analytical results, 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.

### 8.4.3. 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 details on the results, refer to chapter 3 of the report *Informing the California Science Test (CAST) Blueprint Improvements: Results from the Psychometric Studies* (ETS, 2019).

## 8.5. IRT Analyses

IRT is a family of mathematical models 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, link item parameters, scale or equate test scores across different forms or test administrations, evaluate item performance, build an item bank, and assemble test forms.

This section describes how IRT models are used in the CAST for calibrating items. Only items that were not rejected by both the data review and the CDE were included in the calibration.

The purpose of the IRT calibration for the CAST is to provide item parameters that are on the same scale for score reporting and future form assembly. For details on scale scores and achievement levels reported for the CAST, refer to *Chapter 7: Scoring and Reporting*.

## 8.5.1. Item Response Theory Models

On the basis of the results from the test dimensionality study, a unidimensional model was used to calibrate the CAST items. The two-parameter logistic item response theory (2PL-IRT) model was used to calibrate the dichotomous items (i.e., items worth 1 point) and the generalized partial credit model (GPC) model (Muraki, 1992) was used to calibrate the polytomous items (i.e., items worth more than 1 point). The 2PL-IRT model is a special case of the GPC model 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 GPC model (Muraki, 1992) is

$$P_{ih}(\theta_j) = \begin{cases} \dfrac{\exp[\sum_{v=1}^{h} Da_i(\theta_j - b_i + d_{iv})]}{1 + \sum_{c=1}^{n_i} \exp[\sum_{v=1}^{c} Da_i(\theta_j - b_i + d_{iv})]}, & \text{if score } h = 1, 2, ...., n_i \\ \dfrac{1}{1 + \sum_{c=1}^{n_i} \exp[\sum_{v=1}^{c} Da_i(\theta_j - b_i + d_{iv})]}, & \text{if score } h = 0 \end{cases} \quad (8.7)$$

*Refer to the Alternative Text for Equation 8.7 for a description of this equation.*

where,

$P_{ih}(\theta_j)$ is the probability of student with proficiency $\theta_j$ 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 $i$,

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

$D$ is a scaling constant of 1.7 that makes the logistic model approximate the normal ogive model.

When $n_i = 1$, equation 8.7 becomes an expression of the two-parameter logistic model for dichotomous items.

## 8.5.2. Data Preparation

Operational item responses and field test item responses were combined into a sparse matrix for concurrent calibration. Items that were rejected by the CDE's final decision informed by the data review committee were excluded from the calibration.

The sample used in the item calibration includes all students who took the CAST during the 2018–2019 administration.

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

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

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. Such items would not be eligible for use in future forms based on the statistical specifications for form assembly.

## 8.5.3. Summary of IRT Parameters

The overall summary of the IRT *a*-parameter estimates is shown in table 8.9. The number of items in each of the *a*-parameter intervals is shown for each grade, as well as the summary statistics, such as the minimum, maximum, mean, and SD values for each grade level.

The range of *a*-parameter estimates was between 0.05 and 1.42 across grade levels. The means of *a*-parameter estimates were 0.62, 0.55, and 0.52 for grade five, grade eight, and high school, respectively, indicating that the mean item discrimination level decreased slightly as the grade level increased. In addition, the summaries of the IRT *a*-parameter estimates for each grade-level assessment are presented in table 8.C.1 through table 8.C.3 by item type; and table 8.C.4 through table 8.C.6 by content domain for grade five, grade eight, and high school, respectively.

**Table 8.9 Item Discrimination Parameter Distribution by Grade**

| IRT-a Range | Grade 5 | Grade 8 | High School |
|---|---|---|---|
| a < 0 | 0 | 0 | 0 |
| 0 ≤ a < 0.2 | 4 | 5 | 7 |
| 0.2 ≤ a < 0.4 | 8 | 14 | 10 |
| 0.4 ≤ a < 0.6 | 20 | 25 | 17 |
| 0.6 ≤ a < 0.8 | 18 | 11 | 11 |
| 0.8 ≤ a < 1.0 | 10 | 9 | 4 |
| 1.0 ≤ a < 1.2 | 3 | 2 | 2 |
| 1.2 ≤ a < 1.4 | 1 | 0 | 1 |
| 1.4 ≤ a < 1.6 | 0 | 1 | 0 |

Table 8.9 *(continuation)*

| IRT-a Range | Grade 5 | Grade 8 | High School |
|---|---|---|---|
| 1.6 ≤ a < 1.8 | 0 | 0 | 0 |
| 1.8 ≤ a < 2.0 | 0 | 0 | 0 |
| a ≥ 2.0 | 0 | 0 | 0 |
| Minimum | 0.15 | 0.08 | 0.05 |
| Maximum | 1.36 | 1.42 | 1.31 |
| Mean | 0.62 | 0.55 | 0.52 |
| SD | 0.25 | 0.25 | 0.28 |
| **Number of Items** | **64** | **67** | **52** |

Similar information for the IRT *b*-parameter estimates is shown in table 8.10 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 for each grade level. The means of *b*-parameter estimates were 0.38, 0.64, and 1.11 for grade five, grade eight, and high school, respectively, indicating that the mean item difficulty level increased as the grade level increased.

There are a small number of hard items with large *b*-parameters. Although these items were accepted by both the content review and data review panelists, they will not be included in future forms unless they are deemed necessary and approved by both the assessment specialists and the CDE.

The summaries of *b*-parameter estimates, broken down by item type, are shown in table 8.D.1 through table 8.D.3 and by content domain in table 8.D.4 through table 8.D.6 for the three grade levels, respectively.

### Table 8.10 Item Difficulty Parameter Distribution by Grade

| IRT-b Range | Grade 5 | Grade 8 | High School |
|---|---|---|---|
| b < −3.5 | 0 | 0 | 1 |
| −3.5 ≤ b < −3.0 | 0 | 0 | 0 |
| −3.0 ≤ b < −2.5 | 0 | 0 | 0 |
| −2.5 ≤ b < −2.0 | 0 | 1 | 0 |
| −2.0 ≤ b < −1.5 | 0 | 1 | 0 |
| −1.5 ≤ b < −1.0 | 4 | 2 | 2 |
| −1.0 ≤ b < −0.5 | 9 | 6 | 2 |
| −0.5 ≤ b < 0 | 13 | 12 | 7 |
| 0 ≤ b < 0.5 | 12 | 9 | 13 |
| 0.5 ≤ b < 1.0 | 10 | 15 | 8 |
| 1.0 ≤ b < 1.5 | 8 | 10 | 10 |
| 1.5 ≤ b < 2.0 | 3 | 7 | 2 |
| 2.0 ≤ b < 2.5 | 2 | 2 | 0 |
| 2.5 ≤ b < 3.0 | 1 | 0 | 0 |
| 3.0 ≤ b < 3.5 | 1 | 0 | 2 |

Table 8.10 *(continuation)*

| IRT-b Range | Grade 5 | Grade 8 | High School |
|---|---|---|---|
| b ≥ 3.5 | 1 | 2 | 5 |
| Min | -1.47 | -2.01 | -8.21 |
| Max | 4.03 | 8.88 | 13.01 |
| Mean | 0.38 | 0.64 | 1.11 |
| SD | 1.07 | 1.49 | 2.87 |
| **Number of Items** | **64** | **67** | **52** |

## 8.6. Testing Time Analyses

The CAST includes three segments: Segment A, Segment B, and Segment C. Each student received two blocks in Segment A, two PTs in Segment B, and either one PT or one block of discrete items in Segment C. The CAST is an untimed assessment.

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 it took students to complete a test was recorded and analyzed.[10]

Testing time analyses were based on students who logged on the test and whose total testing time at the test level did not equal zero. According to the test design, half of the students received a PT block and the other half of the students received a discrete item block in Segment C. Therefore, testing time analyses for Segment C were conducted separately for the PT block and the discrete item block.

Because the testing time for a discrete block was typically longer than that of a PT block, the testing time for the total test in table 8.11 was broken down for students who received a PT in Segment C (i.e., two Segment A blocks + two PTs + one field test PT) and those who received a discrete block in Segment C (i.e., two Segment A blocks + two PTs + one field test discrete block). The unit of testing time is minutes.

The medians (50th percentile) are used to interpret the results because medians are less impacted by the extreme values and therefore are more meaningful. The median of total testing time for students who received a discrete Segment C block versus students who received a PT Segment C block was 133.8 minutes versus 130.2 minutes in grade five, 114.3 minutes versus 106.7 minutes in grade eight, and 67.1 minutes versus 63.4 minutes in high school. The corresponding total testing time for grade ten, grade eleven, and grade twelve students was 75.1 minutes versus 71.6 minutes, 76.5 minutes versus 72.3 minutes, and 58.7 minutes versus 55.7 minutes, respectively.

The total testing time was longer for students receiving a form with a discrete Segment C block than students receiving a form with a PT Segment C block. The total testing time also decreased as the grade level increased.

Note that the criterion for students to be included in table 8.11 is that they had no “Not Seen” items.

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

Summaries of the times that students spent for the total test are given in table 8.11.

**Table 8.11 Testing Time (in Minutes) for the Total Test**

| Grade or Grade Level | Segment | N | Mean | SD | Min | Max | 1st Percentile | 10th Percentile | 25th Percentile | 50th Percentile | 75th Percentile | 90th Percentile | 99th Percentile |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Grade 5 | 2A Blocks + 2 PTs + 1 Field Test Discrete Block | 227,032 | 149.2 | 75.0 | 4.6 | 1244.3 | 37.3 | 75.5 | 99.7 | 133.8 | 180.7 | 240.4 | 409.4 |
| Grade 5 | 2A Blocks + 2 PTs + 1 Field Test PT | 227,464 | 145.6 | 74.1 | 5.2 | 1251.7 | 35.3 | 72.7 | 96.6 | 130.2 | 176.8 | 235.9 | 400.9 |
| Grade 8 | 2A Blocks + 2 PTs + 1 Field Test Discrete Block | 229,491 | 123.6 | 57.5 | 3.9 | 1042.8 | 24.8 | 64.3 | 86.4 | 114.3 | 149.6 | 192.6 | 315.8 |
| Grade 8 | 2A Blocks + 2 PTs + 1 Field Test PT | 229,907 | 115.8 | 54.2 | 3.4 | 917.9 | 22.9 | 60.0 | 80.5 | 106.7 | 140.3 | 181.3 | 297.5 |
| High school—Grade 10 | 2A Blocks + 2 PTs + 1 Field Test Discrete Block | 11,519 | 78.6 | 35.5 | 5.0 | 526.2 | 14.3 | 38.1 | 56.0 | 75.1 | 96.3 | 120.0 | 192.8 |
| High school—Grade 10 | 2A Blocks + 2 PTs + 1 Field Test PT | 11,455 | 74.9 | 34.2 | 4.3 | 430.8 | 12.2 | 36.2 | 53.2 | 71.6 | 91.9 | 114.3 | 183.0 |

Table 8.11 *(continuation)*

| Grade or Grade Level | Segment | N | Mean | SD | Min | Max | 1st Percentile | 10th Percentile | 25th Percentile | 50th Percentile | 75th Percentile | 90th Percentile | 99th Percentile |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| High school—Grade 11 | 2A Blocks + 2 PTs + 1 Field Test Discrete Block | 127,426 | 81.0 | 40.8 | 1.9 | 904.5 | 9.8 | 34.7 | 55.1 | 76.5 | 100.3 | 129.9 | 210.1 |
| High school—Grade 11 | 2A Blocks + 2 PTs + 1 Field Test PT | 127,632 | 76.8 | 38.6 | 2.8 | 685.6 | 9.1 | 33.3 | 52.2 | 72.3 | 95.2 | 123.3 | 199.5 |
| High school—Grade 12 | 2A Blocks + 2 PTs + 1 Field Test Discrete Block | 136,721 | 61.3 | 31.3 | 2.6 | 636.4 | 7.8 | 24.4 | 40.4 | 58.7 | 77.6 | 97.8 | 157.1 |
| High school—Grade 12 | 2A Blocks + 2 PTs + 1 Field Test PT | 136,803 | 58.0 | 29.6 | 2.5 | 892.7 | 7.1 | 22.9 | 38.3 | 55.7 | 73.7 | 92.9 | 148.1 |
| High school—All grades | 2A Blocks + 2 PTs + 1 Field Test Discrete Block | 275,666 | 71.1 | 37.4 | 1.9 | 904.5 | 8.6 | 28.5 | 46.6 | 67.1 | 89.2 | 115.8 | 190.1 |
| High school—All grades | 2A Blocks + 2 PTs + 1 Field Test PT | 275,890 | 67.4 | 35.5 | 2.5 | 892.7 | 7.9 | 27.0 | 44.2 | 63.4 | 84.7 | 109.7 | 180.1 |

Table 8.E.1 shows the testing time by segment. The testing time for a discrete block was longer than that for a PT block for all the percentiles considered.

Table 8.E.2 shows the testing time for four item types: MC or single selection, CR, TE, and composite. Because testing time was recorded at the page level, items that were on a page with multiple items were excluded from the analysis. The testing time for a CR item was the longest for each percentile.

# 8.7. Reliability

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 of the CR scoring is the extent to which two different raters give 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 (QWK) coefficient.

## 8.7.1. Test Reliability

### 8.7.1.1. Marginal Reliability

In a specified population of students, the reliability of test scores, $X$, is defined as the proportion of the test score variance that is attributable to true differences in student abilities and is sometimes operationalized as the correlation between scores on two replications of the same testing procedure.

Reliability coefficients range from 0 to 1. The higher the reliability coefficient for a set of scores, the more likely students would be to obtain very similar scores if they were retested. 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; however, these reliability indices are not directly applicable because the CAST has multiple forms.

Instead, an IRT-based approach called marginal reliability (Green, Bock, Humphreys, Linn, & Reckase, 1984) can be used to estimate the reliability of CAST scores. The estimates of reliability coefficients reported here are for IRT model–based ability estimates.

This reliability coefficient for theta estimates, $\rho_{\theta\theta'}$, is defined, based on a single test administration, as shown in equation 8.8:

$$\rho_{\theta\theta'} = 1 - \frac{M_{SE_\theta^2}}{s_\theta^2} \quad (8.8)$$

*Refer to the Alternative Text for Equation 8.8 for a description of this equation.*

where,

$\theta$ is the ability estimate,

$s_\theta^2$ is the measure of variance in ability estimates, and

$M_{SE_\theta^2}$ is an average of the variance of the ability estimates.

The standard error of measurement (SEM) of the test on the theta scale is defined as

$$SEM_\theta = \sqrt{M_{SE_\theta^2}} \quad (8.9)$$

*Refer to the Alternative Text for Equation 8.9 for a description of this equation.*

and the SEM of the test on the scale score metric is defined as

$$SEM_{\text{Scale Score}} = \sqrt{M_{CSEM^2_{\text{Scale score}}}} \quad (8.10)$$

*Refer to the Alternative Text for Equation 8.10 for a description of this equation.*

where,

$M_{SE_\theta^2}$ and $M_{CSEM^2_{\text{Scale score}}}$ are the mean estimation variance of theta and scale score, respectively.

Table 8.12 provides the total score reliability for theta as well as the mean, SD, and SEM of both thetas and scale scores for each grade, along with the number of students upon which those analyses were performed. Note that in the case of the total test reliability, the reliability is for the total test on the theta score scale; it is calculated using the total test theta score of individual students. The test reliability ranged from 0.88 to 0.91 across grade levels, exhibiting high levels of reliability.

**Table 8.12 Summary Statistics for Scale Scores and Theta Scores, Reliability, and SEMs**

| Grade or Grade Level | Number of Students | Reliability | Scale Score Mean | Scale Score SD | Scale Score SEM | Theta Score Mean | Theta Score SD | Theta Score SEM |
|---|---|---|---|---|---|---|---|---|
| Grade 5 | 455,033 | 0.91 | 201 | 22.1 | 6.10 | 0.00 | 1.06 | 0.31 |
| Grade 8 | 462,015 | 0.90 | 401 | 22.4 | 6.69 | 0.00 | 1.08 | 0.34 |
| Grade 10 | 23,313 | 0.88 | 597 | 21.4 | 7.01 | -0.21 | 1.04 | 0.36 |
| Grade 11 | 256,472 | 0.89 | 601 | 22.3 | 6.91 | -0.01 | 1.09 | 0.36 |
| Grade 12 | 275,679 | 0.89 | 597 | 22.7 | 6.92 | -0.17 | 1.12 | 0.37 |
| High school | 555,464 | 0.89 | 599 | 22.5 | 6.92 | -0.10 | 1.10 | 0.36 |

### 8.7.1.2. Student Group Reliabilities and Standard Errors of Measurement

The reliabilities of the total test scores were examined for various student groups within the student population. These student groups included demographic groups and subgroups of students who took both the CAST and the English Language Proficiency Assessments for California (ELPAC).

#### *8.7.1.2.1. Reliabilities by Demographic Groups*

The student groups included in these analyses are defined by gender, economic status, special education services status, accommodations for students with special education services, English language fluency, primary ethnicity, migrant status, parent military status, and homeless status. The reliability analyses are also presented by primary ethnicity within economic status.

Reliabilities, theta-based SEMs, and theta score variances for the total test scores are reported for each student group analysis. Table 8.F.1 through table 8.F.6 in appendix 8.F present the overall test reliabilities for student groups defined by student gender, economic status, special education services status, English language fluency, primary ethnicity, migrant status, parent military status, homeless status, and crosstab of primary ethnicity and economic status. Most student groups have reliability of 0.87 or greater in grades five and eight and 0.86 or greater in high school. The exceptions were English learners in grade five; English learners, migrants, and students receiving special education services in grade eight; and English learners, African Americans, migrants, homeless, and students receiving special education services in high school. Among these groups, the reliability coefficients ranged from 0.69 to 0.85, with English learners having the lowest reliability in each grade. Small sample sizes and small variance of the low scores of those groups may contribute to these low reliabilities.

Note that the reliabilities are reported only for samples that are comprised of 11 or more students. Also, in some cases, score reliabilities are not estimable and are presented in the tables as “N/A.” The reliability estimates for some of the student groups are negative due to small variation in scale scores and large conditional standard errors of measurement for extreme score values. These negative reliabilities and their associated SEMs also are presented as “N/A.”

#### *8.7.1.2.2. Reliabilities by English Language Proficiency Assessments for California (ELPAC) Performance Levels*

A subset of students who took the CAST also took the Summative ELPAC, which is the required state test for English language proficiency that must be given to students whose primary language is a language other than English. The Summative ELPAC results show the overall English performance level attained by students, and the performance levels are reported as the following:

- Level 1: Minimally Developed
- Level 2: Somewhat Developed
- Level 3: Moderately Developed
- Level 4: Well Developed

Detailed descriptions of these ELPAC performance levels can be found on the Summative ELPAC General PLDs web page at https://www.cde.ca.gov/ta/tg/ep/elpacgpld.asp.

Table 8.F.7 shows the CAST student group reliabilities, SEM, and theta score variances by the four ELPAC performance levels. The results show that the low group reliabilities were associated with both low performance on the ELPAC and small theta score variance.

### 8.7.1.3. Decision Classification Analyses

When an assessment uses achievement levels as the primary method to report test results, accuracy and consistency of decisions become key indicators of the quality of the assessment.

Decision accuracy is the extent to which students are classified in the same way as they would be if each student's score were the average over all possible forms of the test (the student's true score). Decision accuracy answers the following question: How closely does the actual classification of test takers, based on their single-form scores, agree with the classification that would be made on the basis of their true scores, if their true scores could somehow be known?

Decision consistency is the extent to which students are classified in the same way as they would be on the basis of a different form of a test. Decision consistency answers the following question: What is the agreement between the classifications based on two nonoverlapping, equally difficult forms of the test?

The methodology used for estimating the reliability of classification decisions is described in Livingston and Lewis (1995). The necessary input information includes only the maximum and minimum possible scores on the test and the observed score distribution and the reliability coefficient for the group of students that the estimates will refer to. The method was implemented by the ETS proprietary computer program RELCLASS-COMP (Version 4.14).

Decision accuracy at a particular threshold is estimated by partitioning the estimated bivariate distribution of true scores and observed scores (refer to table 8.13) into a two-by-two table, using the same threshold score on both variables. The decision accuracy statistic is the sum of the proportions in the cells representing consistent classifications—above the cut on both variables or below the cut on both variables. Decision consistency is estimated in the same way, by partitioning the estimated bivariate distribution of observed scores on two forms of the test (refer to table 8.14). Decision consistency values are always lower than the corresponding decision accuracy values because in decision consistency, both of the classifications of the student are based on scores that depend on which form of the test the student took. In decision accuracy, only one of the classifications is based on a score that can vary in this way.

**Table 8.13 Decision Accuracy for Reaching an Achievement Level**

| Achievement level status | Does not reach an achievement level based on true score | Reaches an achievement level based on true score |
|---|---|---|
| Does not reach an achievement level | Correct classification | Misclassification |
| Reaches an achievement level | Misclassification | Correct classification |

**Table 8.14 Decision Consistency for Reaching an Achievement Level**

| Achievement level status | Does not reach an achievement level based on an alternate form | Reaches an achievement level based on an alternate form |
|---|---|---|
| Does not reach an achievement level | Correct classification | Misclassification |
| Reaches an achievement level | Misclassification | Correct classification |

For a test with three threshold scores, the classification is a partition of the distributions of true scores and observed scores into a four-by-four table with the diagonal elements representing consistent classifications based on the two score distributions. The analysis results of decision accuracy and consistency for the CAST by grade levels are presented in table 8.G.1 through table 8.G.12 in appendix 8.G. The proportion of students accurately classified into the four achievement levels is the sum of the main diagonal elements of decision accuracy tables. The proportion of students consistently classified is the sum of the main diagonal elements of the decision consistency tables. The percentages of students who were classified accurately and consistently ranged from 0.79 to 0.81 and from 0.70 to 0.73, respectively.

Using the threshold of Standard Met, the classifications are collapsed to *Standard Not Met* and *Standard Nearly Met* versus *Standard Met* and *Standard Exceeded*, which are the critical categories for accountability. The resulting table is a two-by-two table with diagonal elements representing consistent classifications. The overall decision accuracy ranged from 0.92 to 0.93 and the overall decision consistency ranged from 0.88 to 0.90 for all grade levels. These are considered high levels of accuracy and consistency.

## 8.7.2. Scoring Reliability

### 8.7.2.1. Interrater Agreement

#### *8.7.2.1.1. 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.

#### *8.7.2.1.2. Kappa*

Interrater reliability or consistency is an indicator of homogeneity and is most frequently measured using Cohen's Kappa statistic (1960) which takes chance agreement. For a human-scored item with $m+1$ categories, one can construct an $(m+1) \times (m+1)$ rating table with scores provided by two raters, *X* and *Y*, as shown in table 8.15. Let $n_{st}$ denote the number of responses for which rater *X's* score = $s$ and rater *Y's* score = $t$, $n_{s+}$ is the number of responses for which rater *X's* score = $s$, $n_{+t}$ is the number of responses for which rater *Y's* score = $t$, and $n_{++}$ is the number of all responses.

**Table 8.15 Frequencies of Ratings**

| Rating | Y = 0 | Y = 1 | Y = 2 | …** | Y = m* |
|---|---|---|---|---|---|
| X = 0 | $n_{00}$ | $n_{01}$ | $n_{02}$ | … | $n_{1m}$ |
| X = 1 | $n_{20}$ | $n_{21}$ | $n_{22}$ | … | $n_{2m}$ |
| X = 2 | $n_{30}$ | $n_{31}$ | $n_{32}$ | … | $n_{3m}$ |
| … | … | … | … | … | … |
| X = m | $n_{m0}$ | $n_{m1}$ | $n_{m2}$ | … | $n_{mm}$ |

* m is the number of score categories of an item.

** An ellipsis (…) signifies that there might be more rows (or columns) in the table.

The kappa statistic is defined as

$$kappa = \frac{p_{obs} - p_{\exp}}{1 - p_{\exp}} \tag{8.11}$$

*Refer to the Alternative Text for Equation 8.11 for a description of this equation.*

$$p_{obs} = \frac{1}{n_{++}} \sum_{s=0}^{m} n_{ss} \tag{8.12}$$

*Refer to the Alternative Text for Equation 8.12 for a description of this equation.*

$$p_{\exp} = \frac{1}{n_{++}^2} \sum_{s=0}^{m} n_{s+} n_{+s} \tag{8.13}$$

*Refer to the Alternative Text for Equation 8.13 for a description of this equation.*

where,

$p_{obs}$ is the observed agreement, and

$p_{\exp}$ is the expected agreement between *X* and *Y*.

When $p_{obs}$ and $p_{\exp}$ agree only at the chance level, the value of kappa is 0. When the two measurements agree perfectly, the value of kappa is 1.0.

#### *8.7.2.1.3. Quadratic-Weighted Kappa*

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

$$\kappa_{st} = \frac{\left( \sum_{s=0}^{m} \sum_{t=0}^{m} w_{st} \frac{n_{st}}{n_{++}} \right) - \left( \sum_{s=0}^{m} \sum_{t=0}^{m} w_{st} \frac{n_{s+} n_{+t}}{n_{++}^2} \right)}{1 - \left( \sum_{s=0}^{m} \sum_{t=0}^{m} w_{st} \frac{n_{s+} n_{+t}}{n_{++}^2} \right)} \tag{8.14}$$

*Refer to the Alternative Text for Equation 8.14 for a description of this equation.*

For QWK, the weights are as follows:

$$w_{st} = 1 - \frac{(s-t)^2}{m^2} \quad (8.15)$$

*Refer to the Alternative Text for Equation 8.15 for a description of this equation.*

#### 8.7.2.2. Summary of Scoring Reliabilities

The interrater reliabilities for operational CR items are shown in table 7.3 through table 7.8; table 7.3 through table 7.5 show human-scored CR items; and table 7.6 through table 7.8 show artificial intelligence (AI)–scored CR items. The QWK ranged from 0.51 to 0.94 for human-scored items and 0.56 to 0.90 for AI-scored items, which indicated a moderate to high level of agreement between two raters. Detailed information on interrater reliability results can be found in subsection *7.1.2.9 Interrater Reliability for Operational Items*.

## 8.8. Validity Evidence

Validity refers to the degree to which each interpretation or use of a test score is supported by the accumulated evidence (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014; ETS, 2014). It constitutes the central notion underlying the development, administration, and scoring of a test and the uses and interpretations of test scores.

Validation is the process of accumulating evidence to support each proposed score interpretation or use. This validation process does not rely on a single study or gathering only one type of evidence. Rather, validation involves multiple investigations and different kinds of supporting evidence (AERA, APA, & NCME, 2014; Cronbach, 1971; ETS, 2014; Kane, 2006). It begins with the test design and is implicit throughout the entire assessment process, which includes item development and field testing, analyses of items, test scaling and linking, scoring, reporting, and score usage.

In this section, the evidence gathered is presented to support the intended uses and interpretations of scores for the CAST. This section is organized primarily around the principles prescribed by AERA, APA, and NCME's *Standards for Educational and Psychological Testing* (2014). These *Standards* require a clear definition of the purpose of the test, a description of the constructs to be assessed, and the population to be assessed, as well as how the scores are to be interpreted and used.

The *Standards* identify five kinds of evidence that can provide support for score interpretations and uses:

1. Evidence based on test content
2. Evidence based on response processes
3. Evidence based on internal structure
4. Evidence based on relations to other variables
5. Evidence based on consequences of testing

The next subsection defines the purpose of the CAST, followed by a description and discussion of the kinds of validity evidence that have been gathered.

## 8.8.1. Evidence in the Design of the CAST

### 8.8.1.1. Purpose

The CAST is designed to measure performance on the CA NGSS. The goal of the CAST is to measure what students can do in science. The CAST covers information across the three science domains of Life Sciences, Physical Sciences, and Earth and Space Sciences.

### 8.8.1.2. Constructs to Be Measured

The CAST is designed to show how well students perform relative to the CA NGSS. These standards describe what students should know and be able to do at each grade level.

Test blueprints define the procedures used to measure the domains and standards. These blueprints are provided in table 4.1. They also provide an operational definition of the construct to which each set of standards refers. That is, they define, for each content area, the subject to be assessed, the tasks to be presented, the administration instructions to be given, and the rules used to score student responses. The test blueprints control as many aspects of the measurement procedure as possible so that the testing conditions will remain the same over test administrations (Cronbach, 1971) in order to minimize construct-irrelevant score variance (Messick, 1989).

### 8.8.1.3. Interpretations and Uses of the Scores

Overall student performance is expressed as scale scores and achievement levels. An inference is drawn about how much knowledge and skill in the CAST the student has, on the basis of a student's total score. The total score is also used to classify students in terms of their level of knowledge and skill in the CAST. These levels are called achievement levels and are labeled *Standard Exceeded*, *Standard Met*, *Standard Nearly Met*, and *Standard Not Met*. The descriptions reflecting the level of expectation on students' performance of these achievement levels can be found in subsection *7.3.3 Achievement Levels*. A detailed description of the uses and applications of the CAST scores is presented in *Chapter 7: Scoring and Reporting*.

The CAST results have four primary purposes:

1. Help facilitate conversations between parents/guardians and teachers about student performance
2. Serve as a tool to help parents/guardians and teachers work together to improve student learning
3. Help staff from schools and local educational agencies (LEAs) identify strengths and areas that need improvement in their educational programs
4. Provide the public and policymakers with information about student achievement

More detailed descriptions regarding score use can be found in the *Education Code* Section 60602 web page at http://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?lawCode=EDC&division=4.&title=2.&part=33.&chapter=5.&article=1 (outside source).

### 8.8.1.4. Intended Test Population

Students enrolled in grades five and eight are required to take part in the CAST, unless they are eligible to participate in the alternate assessments. Students enrolled in high school are required to take part in the CAST once while in high school, unless they are eligible to participate in the alternate assessments.

## 8.8.2. Evidence Based on Test Content

Evidence based on test content refers to traditional forms of content validity evidence, such as the rating of test specifications and test items (Crocker, Miller, & Franks, 1989; Sireci, 1998), as well as alignment methods for educational tests that evaluate the interactions between curriculum frameworks, testing, and instruction (Rothman, Slattery, Vranek, & Resnick, 2002; Bhola, Impara & Buckendahl, 2003; Martone & Sireci, 2009).

### 8.8.2.1. Description of California Next Generation Science Standards

As noted in section *1.1 Background*, the CAST is aligned with the CA NGSS. There are three main domains at each grade level: Life Sciences, Physical Sciences, and Earth and Space Sciences. Performance expectations (PEs) within the CA NGSS are assessable statements of what students should know and be able to do in each domain. Overall, the alignment study results provide strong support that the CAST system produces aligned test forms (CDE, 2019a).

### 8.8.2.2. Item Specifications

Item specifications describe the characteristics of items that are written to measure each content standard. Specifications were developed for each PE at each grade level. Details on item specifications can be found in subsection *3.2.3 Specifications*.

### 8.8.2.3. Assessment Blueprints

The CAST blueprints provided in table 4.1 describe the content of the science assessments for all grades tested and how that content is assessed. The CAST blueprints reflect the depth and breadth of the PEs of the CA NGSS. The test blueprints have information about the number of items and depth of knowledge for items associated with each assessment target. Each test is described by a single blueprint for each segment of the test. For details about CAST blueprints, please refer to subsection *4.2.1 Test Blueprints*.

### 8.8.2.4. Alignment Study

A strong alignment between CAST and the CA NGSS is fundamental to the meaningful measurement of student achievement and instructional effectiveness. Alignment results demonstrate that the CAST represents the full range of the content standards and measures student knowledge in the same manner and at the same level of complexity as expected in the content standards. For detailed information on the alignment study conducted, refer to the *California Science Test (CAST) Alignment Study Report* (CDE, 2019a).

### 8.8.2.5. Form Assembly Process

The content standards and blueprints are the basis for choosing items for each assessment. Additionally, item difficulty and item-total score correlations are provided to evaluate the statistical characteristics of test forms. All CAST form assembly meets all the content and statistical specifications. Refer to *Chapter 4: Test Assembly* for additional information.

## 8.8.3. Evidence Based on Response Processes

Validity evidence based on response processes refers to “evidence concerning the fit between the construct and the detailed nature of performance or response actually engaged in by test takers” (AERA et al., 2014, p. 15). This type of evidence generally includes documentation of activities such as:

- interviews with students concerning their responses to test items (i.e., think alouds);
- systematic observations of test response behavior;
- evaluation of the criteria used by judges when scoring CRs, analysis of student item-response-time data, and features scored by automated algorithms; and
- evaluation of the reasoning processes students employ when solving test items (Embretson, 1983; Messick, 1989; Mislevy, 2009).

#### 8.8.3.1. Analysis of Testing Time

Testing times for each administration can be evaluated for consistency, with the expected response processes for the tasks presented to students. The length of time it takes students to take a test is recorded and analyzed to build a profile describing what a typical testing event looks like for each grade. In addition, variability in testing time is investigated to determine whether a student's testing time should be viewed as unusual or irregular. It should be noted that the CAST is untimed.

The descriptive statistics—e.g., the number of students, mean, SD, minimum and maximum, and percentiles—of the following time variables are computed for each grade:

- Time required to complete each segment
- Time required to complete each item type
- Time required to complete the total test

Some cases of extremely long testing times may be attributed to students with special needs taking longer to complete the tests, or the test not being closed down properly. Therefore, mean testing times may be misleading. The medians (50th percentile) are more meaningful in the interpretation of the time comparisons because medians are less impacted by the extreme values than means.

Table 8.11 provides the summary of times that students spent for the total test. Table 8.E.1 and table 8.E.2 in appendix 8.E provide the summary of testing time by segment and by item type, respectively. These tables include various percentiles of test times (in minutes). The maximum median testing times were 134, 114, and 67 minutes for grade five, grade eight, and high school, respectively. Given that the estimated testing time was two hours, approximately 50 percent of the students in grade five and the majority of students in grade eight and high school were able to complete the test within the estimated testing time. Detailed information on students' testing time can be found in section *8.6 Testing Time Analyses*.

#### 8.8.3.2. Student Survey

The student survey questions were administered at the end of the test. The survey questions were focused on gathering information about how the science content on the CAST compared to the science content presented in the classroom. The majority of students responded that compared with what they had taught in their science classes, 1) all or most of the topics on the test were taught; 2) some items on the test were different from what they were taught; and 3) questions on the test were comparable with what they were taught. The student survey results show that the CAST reflects what students were taught in the classroom. Detailed information on the student survey can be found in *Chapter 10: Student Survey*.

## 8.8.4. Evidence Based on Internal Structure

Evidence based on *internal structure* refers to the statistical analysis of item and score subdomains to investigate the primary and secondary (if any) dimensions measured by an assessment. A dimensionality study was conducted for the CAST based on 2018–2019 test data.

Analysis of the internal structure evidence also includes indices of measurement precision such as DIF analysis, test reliability, student group reliability, decision accuracy and consistency, interrater agreement, conditional and unconditional SEMs, and test information functions (TIFs).

### 8.8.4.1. Dimensionality

The CAST assesses PEs as they appear in the CA NGSS, and the PEs represent a complete integration of the three dimensions, not three dimensions that coincide together. A dimensionality study was conducted during the 2018–2019 administration to determine the factor structure of the assessments. Results suggested the test is essentially unidimensional, which is consistent with the notion of CAST design in that it measures the integration of the dimensions. Details on the dimensionality study can be found in section *8.4 Test Dimensionality Analyses*. The study itself can be found in *Chapter 12: Test Dimensionality Study Addendum*.

### 8.8.4.2. Differential Item Functioning (DIF)

Analysis of item functioning using IRT and DIF falls under the internal structure category. For the CAST, DIF analyses were conducted to assess differences in the item performance of groups of students who differ in their demographic characteristics. For the CAST 2018–2019 administration, few items were identified as having significant levels of DIF. The details on DIF analyses performed can be found in section 8.3 Differential Item Functioning (DIF) Analyses.

### 8.8.4.3. Overall Reliability Estimates

The results of marginal reliability analyses on the total theta scores for the CAST are presented in table 8.12. The results indicate that the reliability estimates for the CAST total scores are high, ranging from 0.88 to 0.91 across all grade levels.

### 8.8.4.4. Student Group Reliability Estimates

The reliabilities also are examined for various student groups within the student population that differ in their demographic characteristics. The characteristics considered are gender, ethnicity, economic status, special education services status, migrant status, English language fluency, parent military status, homeless status, and ethnicity by economic status (refer to table 5.3 for the demographic groups reported). Reliability estimates and SEM information for the total test theta scores are reported for each student group in table 8.F.1 through table 8.F.6 in appendix 8.F.

### 8.8.4.5. Reliability of Performance Classifications

The methodology used for estimating the reliability of classification decisions is described with the decision classification analyses in subsection *8.7.1.3 Decision Classification Analyses*. The results of these analyses are presented in table 8.G.1 through table 8.G.12 in appendix 8.G.

The proportions of students who were classified accurately and consistently ranged from 0.79 to 0.81 and 0.70 to 0.73, respectively. When the classifications were collapsed to

below *Standard Met* versus *Standard Met* and above, which are the critical categories for accountability analyses, the estimated proportion of students who were classified accurately ranged from 0.92 to 0.93 across all grade levels, and the estimated proportion of students who were classified consistently ranged from 0.88 to 0.90. These are considered high levels of accuracy and consistency.

#### 8.8.4.6. Interrater Reliability

Cohen's Kappa statistics provide evidence of the degree to which a student's score is consistent from one rater to another. Research has shown values of kappa between 0.41 and 0.60 exhibit moderate levels of agreement between the two ratings (Landis & Koch, 1977; Flack, Afifi, Lachenbruch, & Schouten, 1988) and that values of QWK greater than 0.70 indicate excellent agreement (Williamson, Xi, & Breyer, 2012).

The results in table 7.3 through table 7.8 in *Chapter 7: Scoring and Reporting* show the QWK ranges from 0.51 to 0.94 for human-scored items and 0.56 to 0.90 for AI-scored items, which indicate moderate to high levels of agreement between two raters.

## 8.8.5. Evidence Based on Relations to Smarter Balanced Test Scores

Evidence based on *relations to other variables* refers to traditional forms of criterion-related validity evidence such as concurrent and predictive validity, as well as more comprehensive investigations of the relationships among test scores and other variables such as multitrait-multimethod studies (Campbell & Fiske, 1959). External variables can be used to evaluate hypothesized relationships between test scores and other measures of student achievement (e.g., test scores on other tests) to evaluate the degree to which different tests actually measure different skills and the utility of test scores for predicting specific criteria (e.g., college grades). This type of evidence is essential for supporting the validity of certain inferences based on CAST scores.

Most students from grades five, eight, and eleven who took the CAST also took the Smarter Balanced English language arts/literacy (ELA) and mathematics assessments. Table 8.H.1 through table 8.H.6 in appendix 8.H show these correlations for both ELA and mathematics test scores by demographic groups. Table 8.H.1 and table 8.H.2 show data for grade five, table 8.H.3 and table 8.H.4 show data for grade eight, and table 8.H.5 and table 8.H.6 show data for grade eleven. For the total student group, higher correlations were observed in grades five and eight (0.82 or greater for both ELA and mathematics) and slightly lower correlations were observed in grade eleven (0.78 for both ELA and mathematics).

## 8.8.6. Evidence Based on Consequences of Testing

Evidence based on *consequences of testing* refers to the evaluation of the intended and unintended consequences associated with a testing program. Examples of evidence based on testing consequences include investigations of adverse impact, evaluation of the effects of testing on instruction, and evaluation of the effects of testing studies on issues such as high school dropout rates. With respect to educational tests, the *Standards* stress the importance of evaluating test consequences. For example, they state the following:

> "When educational testing programs are mandated by school, district, state, or other authorities, the ways in which test results are intended to be used should be clearly described by those who mandate the tests. It is also the responsibility of those who mandate the use of tests to monitor their impact and to identify and minimize potential negative consequences as feasible. Consequences resulting from the use of the test, both

> intended and unintended, should also be examined by the test developer and/or user." (AERA et al., 2014, p. 195)

Investigations of testing consequences relevant to the CAST goals include analyses of students' opportunity to learn the CA NGSS and analyses of changes in textbooks and instructional approaches. Unintended consequences, such as diminished morale among teachers and students, increased pressure on students leading to increased dropout rates, or the pursuit of college majors and careers that are less challenging can be evaluated. These sorts of investigations require information beyond what has been available to the CAST program to date.

# 8.9. Research Studies

## 8.9.1. Multistage Adaptive Test (MST) Practicality Study

### 8.9.1.1. 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, ETS conducted a study using the item pool from the 2018–2019 administration to evaluate, 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. It must be noted that there is no absolute threshold separating improvements judged as substantial from those that are not. Rather, the judgment is whether the improvement is great enough to offset the complexity and risk inherent in all adaptive testing.

The methodology and results are briefly reviewed in this subsection. Refer to chapter 4 in *Informing the California Science Test (CAST) Blueprint Improvements: Results from the Psychometric Studies* (ETS, 2019) for the full details.

### 8.9.1.2. Methodology

The MST practicality study was conducted using the item parameters and student distributions estimated from the 2018–2019 operational test data. The items used had good classical item analysis results—well fit by their IRT 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.

Statistical requirements were also imposed on each item block. This was done by setting targets for the IRT information that each block was to contribute to measurement. The information functions for the item blocks that comprise an MST can provide a concise visual impression of how the test will perform. The heights of the information functions are proportional to test precision, while the separation between higher-stage blocks indicates the breadth of the proficiency range across which measurement is good. The target

information functions were roughly bell-shaped and centered at midrange proficiency values for the router and the medium-difficulty second-stage blocks.

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:

- **TIFs:** The information functions for the second stage levels; expected to be reasonably distinctive for the panel to be meaningful
- **Measurement precision:** Measured by the conditional standard error of measurement (i.e., the SD 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
- **Relative efficiency:** Ratio of the information function for the two designs being compared at every true ability level
- **Reliability**

#### 8.9.1.3. Results

For grade five, grade eight, and high school, MST performance exceeds that of linear forms for both the lower and higher end of the ability continuum. The improvement is relatively modest except at the far extremes of the proficiency range.

The relative strengths and weaknesses of the assembly pools are largely reflected in the degree and location of the measurement improvement. The grade five pool is strongest at lower levels of student performance; that is where the MST is strongest as well. The grade eight pool is well balanced in terms of item difficulty and MST performance, particularly at both lower and upper proficiency ranges. The high school pool also shows improvement at both the lower and upper ranges.

#### 8.9.1.4. Limitations of the Study

The MST practicality study was conducted based on item parameters and empirical student ability distributions estimated from the first-operational-year data. There are a few limitations that would impact the generalizability of the results.

Not all LEAs 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 the first operational assessment and future ones.

In addition, several characteristics of the item pool also limited the generalizability of the results. For example, items tended to be difficult for high school, and the bulk of the information in the current assembly pool for grade five was on the easier end of the performance spectrum; it was not possible to create a block that was difficult enough and still met content requirements.

Given these limitations, and since the full benefit of the MST panel can be realized once the pool is expanded to support more differentiable easy and hard blocks, it is recommended to hold off on the implementation of the MST design and reevaluate once the pool expands beyond the current level.

## 8.9.2. Content Screen-Out Study

### 8.9.2.1. Description

In the 2018–2019 operational administration, students received two PTs in two different domains 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.

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 science content domains with which they were interested and experienced and they happened to be assigned PTs in the domains in which they were most or least familiar. Alternatively, performance in Segment A, which is comprised of 32–34 items, roughly spread evenly across the three domains, could have been used to screen out PTs in the domain in which the student demonstrated the weakest performance, so as not to disadvantage any student in the assignment of PTs in Segment B. For instance, students who performed conspicuously poorly on Life Sciences items in Segment A would have been assigned PTs in the other two domains, whereas students who performed similarly across all three domains in Segment A would have been 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 2018–2019 operational test data.

The methodology and results of this study are briefly reviewed. Refer to *Informing the California Science Test (CAST) Blueprint Improvements: Results from the Psychometric Studies* (ETS, 2019, Chapter 5) for the full details.

### 8.9.2.2. Methodology

This study used all items from the operational items in segments A and B. The sample included all students who completed two Segment A blocks and two Segment B blocks. Unmotivated students were identified and removed from the analyses.

For grade five, there were five PTs total: one for Life Sciences, one for Earth and Space Sciences, and three for Physical Sciences, resulting in seven possible pairings of PTs from two different contents for Segment B. For grade eight, there were six PTs total—two for each content domain, resulting in 12 possible pairings of PTs for Segment B. For high school, there were only three PTs—one for each content domain, resulting in three possible pairings of PTs for Segment B. Students were roughly evenly distributed across the possible Segment B PT pairings. There were a total of 454,657 students in the grade five dataset, 495,940 students in the grade eight dataset, and 551,757 students in the high school dataset.

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, 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 *Informing the California Science Test (CAST) Blueprint Improvements: Results from the Psychometric Studies* (ETS, 2019, Chapter 5).

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.

#### 8.9.2.3. 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 (.77 to .92) across the three grade levels, suggesting that, generally, students performed comparably across the domains with only some differential performance.

The dimensionality study (ETS, 2019, Chapter 3), which uses MIRT models, also provided estimated true-score correlations by domain. Although it pooled across both segments A and B, using a correlated three-factor MIRT model among the three domains, the dimensionality study estimated true-score correlations of .88 to .97 among the domains.

Such high correlations provide evidence against much variation in performance by science content domain on the CAST, making the utility of a screener unlikely. However, further analysis is needed, as there may still be outlying students who would benefit from a screener. In general, it is expected that student performance across domains is comparable and no screener would be needed (i.e., students would be randomly assigned two PTs from two different domains), but the interest is in the outlying students and the extent that a screener would benefit them.

The results of the linear models further suggested a screener may not be useful, as all three models had $R^2$ differences within 0.001, ranging from 0.4097 to 0.4100 in grade five, 0.4199 to 0.4204 in grade eight, and 0.3835 to 0.3845 for high school. 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.

The results were mixed. There were several cases for which students performed as well as or better, on average, when assigned a PT in the students' conspicuously weak Segment A domain than when not assigned a PT in the weakest domain. In most cases, the rank ordering of performance on the Segment B PT pairs generally corresponded to that of the overall population and was more indicative of the difficulty of the particular PTs than of whether the students were assigned a PT in the students' weak Segment A domain.

#### 8.9.2.4. Limitations of the Study

The CAST 2018–2019 operational administration offered rich data and large sample sizes to use for this analysis. However, for grade five and high school, two to three of the content domains only had one PT, limiting the extent that results are generalizable beyond these particular PTs.

# 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.

Bhola, D. S., Impara, J. C., & Buckendahl, C. W. (2003). Aligning tests with states' content standards: Methods and issues. *Educational Measurement: Issues and Practice, 22*, 21–29.

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.

California Department of Education. (2019a). California Science Test (CAST) Alignment Study Report. [Unpublished report]. Sacramento, CA: California Department of Education.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105.

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

Crocker, L. M., Miller, D., & Franks, E. A. (1989). Quantitative methods for assessing the fit between test and curriculum. *Applied Measurement in Education, 2*, 179–94.

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

Cronbach, L. J. (1971). Test Validation. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.

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. (2014). *ETS Standards for Quality and Fairness.* Princeton, NJ: Educational Testing Service.

Educational Testing Service. (2019). *Informing the California Science Test (CAST) blueprint improvements: Results from the psychometric studies*. [Draft manuscript]. Princeton, NJ: Educational Testing Service.

Embretson (Whitley), S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179–97.

Flack, V. F., Afifi, A. A., Lachenbruch, P. A., & Schouten, H. J. A. (1988). Sample size determinations for the two rater Kappa statistics. *Psychometrika, 53*(3), 321–25.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21*, 347–60.

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.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Washington, DC: American Council on Education/Praeger.

Landis, J. R., & Koch, G. G. (1977). The measurement of interrater agreement for categorical data. *Biometrics*, *33*, 159–74.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classification based on test scores. *Journal of Educational Measurement*, *32,* 179–97.

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.

Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessments, and instruction. *Review of Educational Research, 4*, 1332–61.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed.). Washington, DC: American Council on Education.

Mislevy, R. J. (2009). Validity from the perspective of model-based reasoning. *CRESST Report 752*. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159–76.

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

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, *21*(2), 137–50.

Rothman, R., Slattery, J. B., Vranek, J. L., & Resnick, L. B. (2002). *Benchmarking and alignment of standards and testing.* [Technical Report 566]. Washington, DC: Center for the Study of Evaluation.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*(1), 145–54.

Sireci, S. G. (1998). Gathering and analyzing content validity data. *Educational Assessment*, *5*, 299–321.

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), 613–25.

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), 321–44.

# Accessibility Information

## Alternative Text for Equation 8.1

P value sub dich equals the fraction with the numerator the sum of X sub ic and the denominator N sub i end fraction.

## Alternative Text for Equation 8.2

P value sub poly equals the fraction with the numerator the sum of X sub ij over j and the denominator N sub i times Max of X sub i end fraction.

## Alternative Text for Equation 8.3

R sub polyreg equals the fraction with the numerator Beta hat times S sub tot and the denominator the square root of beta hat squared times s sub tot squared plus 1 I end fraction.

## Alternative Text for Equation 8.4

Alpha sub MH equals the fraction with the numerator open bracket the sum from m of R sub rm times W sub fm divided by N sub tm close bracket and the denominator open bracket the sum from m of R sub fm times W sub rm divided by N sub tm close bracket end fraction.

## Alternative Text for Equation 8.5

MH D - DIF equals negative 2.35 times the natural log of open bracket alpha sub MH close bracket.

## Alternative Text for Equation 8.6

SMD equals the fraction with numerator the sum from m equals 1 to M of N sub fm times E sub f of Y from X equals m and denominator the sum from m equals 1 to M of N sub fm end fraction minus the fraction with numerator the sum from m equals 1 to M of N sub fm times E sub r of Y from X equals m and denominator the sum from m equals 1 to M of N sub fm end fraction equals the fraction with the numerator the sum from m equals 1 to M of D sub fm and the denominator m equals1 to M of N sub fm end fraction.

## Alternative Text for Equation 8.7

P sub ih of theta sub j equals the numerator exp open parenthesis the sum from v equals 1 to h of D times a sub i of the quantity open parenthesis theta sub j minus b sub i plus d sub iv close parenthesis close parenthesis and denominator 1 plus the sum from c equals 1 to n sub i exp open parenthesis the sum from v equals 1 to c D times a sub i of the quantity open parenthesis theta sub j minus b sub i plus d sub iv close parenthesis close parenthesis, if score h equals 1, 2, …., n sub i.

P sub ih of theta sub j equals 1 divided by denominator 1 plus the sum from c equals 1 to n sub i exp open parenthesis the sum from v equals 1 to c D times a sub i of the quantity open parenthesis theta sub j minus b sub i plus d sub iv close parenthesis close parenthesis, if score h equals 0.

## Alternative Text for Equation 8.8

Rho sub theta prime equals 1 minus M sub SEM squared sub theta divided by s squared sub theta.

## Alternative Text for Equation 8.9

SEM sub theta equals square root of M sub open parenthesis of SE sub theta close parenthesis square.

## Alternative Text for Equation 8.10

SEM sub scale score equals square root of M sub open parenthesis of CSEM sub scale score close parenthesis square.

## Alternative Text for Equation 8.11

kappa equals the fraction with the numerator p sub obs minus p sub exp the denominator 1 minus p sub exp.

## Alternative Text for Equation 8.12

p sub obs equals 1 divided by n times the sum from s equals 0 to m n sub ss.

## Alternative Text for Equation 8.13

p sub exp equals 1 divided by n square times the sum from s equals 0 to m n sub s plus times n sub plus s.

## Alternative Text for Equation 8.14

K sub st equals the fraction with numerator open parenthesis the sum from s equals zero to m the sum from t equals zero to m of w sub st times n sub st divided by n sub plus plus close parenthesis minus open parenthesis the sum from s equals zero to m the sum from t equals zero to m of w sub st times n sub splus times n sub plust divided by n squared sub plusplus close parenthesis and the denominator 1 minus open parenthesis the sum from s equals zero to m the sum from t equals zero to m of w sub st times n sub splus times n sub plus t divided by n squared sub plus plus close parenthesis end fraction.

## Alternative Text for Equation 8.15

W sub st equals 1 minus fraction with numerator open parenthesis s minus t close parenthesis squared and denominator m squared end fraction.

# Chapter 9: 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 reporting processes for the California Science Test (CAST). 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 goals of delivering technically sound, fair, and useful products and services; and assisting the public and auditors evaluating those products and services. Quality control procedures are outlined in this chapter.

## 9.1. Quality Control of Test Materials

### 9.1.1. Developing Test Administration Instructions

CAST content was incorporated to fit the California Assessment of Student Performance and Progress (CAASPP) System specifications used to administer Smarter Balanced Summative Assessments. ETS staff consulted with internal subject matter experts and conducted validation checks to verify that test instructions accurately matched the testing processes. Copy editors and content editors reviewed each document for spelling, grammar, accuracy, and adherence to CDE style and usage requirements as well as the CDE accessibility standards. All CAASPP documents were approved by the CDE before they could be published to the CAASPP website at http://www.caaspp.org/. Only nonsecure documents were posted to this website.

### 9.1.2. Processing Test Materials

Online tests that were submitted by students were transmitted from the American Institutes for Research (AIR, now Cambium Assessment) 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.

## 9.2. 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 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 was accepted for authoring—that is, once it was entered into ETS' item bank and formatted for use in an assessment—ETS employed a series of internal and external reviews. These reviews used established criteria and specifications to judge the quality of item content and ensured that each item measured what it was intended to measure. These reviews also examined the overall quality of the test items before presentation to the CDE and item reviewers. To finish the process, a group of California educators reviewed the items and PTs for accessibility, bias and sensitivity, and content,

and made recommendations for item enhancement. The details on quality control of item development are described in section *3.3 Item Review Process*.

When student response data on each item became available, ETS Psychometric Analysis & Research (PAR) staff conducted item analysis and a key check to examine whether the items performed as expected. ETS psychometric staff conducted a thorough item analysis and evaluated all items carefully using the statistical criteria described in subsection *8.2.6 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.

## 9.3. Quality Control of Test Form Development

ETS conducted multiple levels of quality assurance checks on each assembled CAST form to ensure it met the form-building specifications. Both ETS Assessment & Learning Technology Development (ALTD) and PAR staff reviewed and signed off on the accuracy of test forms before the forms were put into production for administration in the CAST. Detailed information related to test assembly can be found in *Chapter 4: Test Assembly*.

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

- all answers were correct;
- answers were scored correctly in the item bank;
- all items matched the standard;
- all content in the item was correct;
- all items met the statistical criteria;
- distractors were plausible;
- multiple-choice item options were parallel in structure;
- language was grade-level appropriate;
- no more than three multiple-choice items in a row had the same key;
- all art was correct;
- there were no mechanical errors in grammar, spelling, punctuation, and the like; and
- items adhered to the approved style guide.

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

## 9.4. Quality Control of Test Administration

The quality of test administration for the CAST, 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 and 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. 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 *5.7.1 ETS' Office of Testing Integrity (OTI)*.

# 9.5. Quality Control of Scoring

## 9.5.1. Development of Scoring Specifications

A number of measures were taken to ascertain that the scoring keys were applied to the student responses as intended and the student scores were computed accurately. ETS built and reviewed 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 ALTD 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 were sent electronically to the ETS Online Network for Evaluation for scoring by trained, qualified raters. In addition, the student's original response string was stored for data verification and auditing purposes. Standard quality inspections were 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.

## 9.5.2. Quality Control of Machine-Scoring Procedures

AIR (now Cambium Assessment), 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 conducted 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 checked to ensure that the test record contained no data from items that might have been invalidated.

Data passed 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 was pulled and transmitted to ETS in a predetermined results format.

## 9.5.3. Quality Control of Human Scoring

For human scoring, ETS employed multiple quality controls including

- raters being required to successfully pass calibration, described in subsection *7.1.2.6 Training for Raters*, prior to beginning scoring at each grade level;
- scoring leaders conducting backreads during each scoring shift;

- ETS reviewing statistics on validity papers; and
- ETS reviewing interrater reliability statistics.

Refer to subsection *7.1.2 Human Scoring* for the topics; refer to subsections *7.1.2.8 Scoring Monitoring and Quality Management*, *7.1.2.9 Interrater Reliability for Operational Items*, and *7.1.2.10 Validity Responses and Sets* for more specific details on these tools used for quality control of human scoring.

### 9.5.4. Artificial Intelligence Scoring Verification

To ensure the quality of artificial intelligence scoring, ETS maintained a QA system where a random sample of human ratings was also obtained and used for rater agreement analyses. The results of the agreement analyses are presented in subsection *7.1.2.9 Interrater Reliability for Operational Items*. Also refer to subsection *7.1.2.8 Scoring Monitoring and Quality Management* for more information.

### 9.5.5. Enterprise Score Key Management (eSKM) System Processing

Prior to the test administration, ETS ALTD staff reviewed and verified the keys and scoring rubrics for all items. Then, these keys and rubrics were provided to AIR for its machine-scoring implementation. After AIR finished machine-scoring, those scores and responses were delivered to ETS.

ETS' Centralized Repository Distribution System and Enterprise Service Bus departments collected and parsed .xml files that contained student response data from AIR. ETS' eSKM system collected and calculated individual students' overall scores (total raw scores) and generated student scores in the approved statistical extract format. These data extracts were sent to ETS' Data Quality Services for data validation. Following successful validation, the student response statistical extracts were made available to the psychometric team.

## 9.6. Quality Control of Psychometric Processes

### 9.6.1. Development of Psychometric Specifications

The psychometric procedures for the CAST 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 analysis steps such as sample inclusion, analyses methods, and special handling of the data.

### 9.6.2. Quality Control of Psychometric Analyses

ETS developed two parallel scoring systems to produce and verify student scores: the eSKM scoring system received an individual student's item scores and item responses from AIR and calculated individual student scores for ETS' reporting systems. The PAR team also computed individual student scores based on item scores delivered by AIR. The scores from the two sources were then compared for internal quality control. Any differences in the scores were discussed and resolved. All scores complied with the ETS scoring specifications and passed the parallel scoring process to ensure the quality and accuracy of scoring and to support the transfer of scores into TOMS, the database of the student records scoring.

### 9.6.3. Psychometric Processing

Psychometricians conducted extensive analyses including item analyses, differential item functioning, item response theory (IRT) calibration, linking, and scaling.

The psychometric analyses conducted at ETS underwent comprehensive quality checks by a team of psychometricians and data analysts. Detailed checklists were developed by members of the team for each of the statistical procedures performed on the CAST. Classical item analyses were performed to evaluate the performance of items, such as item difficulty and correlation between item scores and total scores. Items that were flagged for questionable statistical attributes were sent to ALTD staff for review; ALTD comments were then reviewed by the psychometricians before items were approved for inclusion in calibration.

During the calibration process, checks were made to ascertain that the input files were established accurately. Checks were also made on the number of items, number of examinees with valid scores, IRT item difficulty estimates, standard errors for the item difficulty estimates, and the linking and scaling process. Two psychometricians conducted parallel calibration processing and compared the results to check for any inconsistency. Psychometricians also performed detailed reviews of relevant statistics to determine whether the chosen IRT model fit the data. ETS then presented and reviewed the calibration results with the CDE for approval.

## 9.7. Quality Control of Reporting

To ensure the quality of the CAST results for both individual student and summary reports, four general areas were evaluated:

1. Comparison of report formats with input sources from the CDE-approved samples
2. Validation of the report data through quality control checks performed by ETS' Data Quality Services and Resolutions teams, as well as running of all the student score reports through ETS' patented Quality Control Integrator software
3. Evaluation of the production of all Student Score Reports—available in paper and electronic versions—by verifying the print quality, comparing the number of report copies, sequence of the report order, and offset characteristics to the CDE requirements
4. Proofreading of the pilot and production reports by the CDE and ETS prior to any LEA mailings

All reports were required to include a single, accurate LEA code, a charter school number (if applicable), an LEA name, and a school name. All elements conformed to the CDE's official county/district/school (CDS) code and naming records. From the start of processing through scoring and reporting, the CDS Master File was used to verify and confirm the accuracy of codes and names. The CDE provided a revised LEA Master File to ETS throughout the year as updates became available.

After the reports were validated in accordance with CDE requirements, a set of reports representing all possible grades, content areas, and reporting outcomes was provided to the CDE and ETS for review and approval. Electronic reports were sent on the actual report template, organized as they were expected to look in production.

Upon the CDE's approval of the reports generated for the initial review, ETS proceeded with the first batch of report production. The first production batch was inspected to validate a subset of LEAs that contained key reporting characteristics and demographics of the state. The first production batch incorporated selected LEAs and provided the final check prior to generating all reports and making them electronically available for download in TOMS and for student information systems through an application programming interface, as well as mailing them to the LEAs that requested printed Student Score Reports.

### 9.7.1. Exclusion of Student Scores from Summary Reports

ETS provided the CDE with reporting specifications that documented when to exclude student scores from summary reports. These specifications included the logic for handling submitted tests and answer documents that, for example, identified students who tested but responded to no items, who were not tested due to parent/guardian request, or who did not complete the test due to illness. The methods for handling other anomalies were also covered in the specifications. These anomalies are described in more detail in the subsection *7.4.2 Special Cases*.

### 9.7.2. End-to-End Testing for Operational Administration

ETS conducted end-to-end testing prior to the start of the test administration. The purpose of this testing was to verify that all systems, processes, and resources were ready for the operational administration. ETS employed a number of approaches to verify ongoing systems performance, including monitoring of system availability and online system usage. Time was allotted for user acceptance testing to confirm that the systems met requirements and to make identified corrections before final deployment. To accomplish system acceptance and sign off, ETS deployed systems to a staging area, which mirrored the final production environment, for operational and user acceptance testing. Final approval by the CDE triggered the final deployment of the system.

# Reference

Educational Testing Service. (2014). *ETS Standards for Quality and Fairness*. Princeton, NJ: Educational Testing Service. Retrieved from https://www.ets.org/s/about/pdf/standards.pdf

# Chapter 10: Student Survey

This chapter describes the development and administration of the survey questionnaires presented to students during the 2018–2019 California Science Test (CAST) operational test administration. The summary of findings and results of analyses from the survey are included.

Student survey questions were developed by Human Resources Research Organization (HumRRO) in consultation with the California Department of Education (CDE). HumRRO, as part of federal requirements, is the state's independent evaluator and provided survey questions to gather data from students for use in their evaluation of the CAST. In 2018–2019, the first operational year for the CAST, HumRRO gathered information about how the science content on the CAST compared to the science content presented in the classroom.

## 10.1. Student Survey Questions

The survey questionnaire was administered after students completed the CAST. There were four survey questions:

1. Did you learn about the topics on the test in your science classes?
2. Were any questions on the test different from the types of questions you see in science classes?
3. How hard were questions on this test compared to questions you see in science classes?
4. Do you think you will be enrolling in any more science classes in high school?

Questions one through three were for all students and question four was applicable only to high school students in grades ten and eleven. Refer to appendix 10.A for details on the options for each question and student response frequencies.

The questions were in braille for students who used the braille accommodation.

## 10.2. Student Survey Results

Table 10.A.1 through table 10.A.3 show the survey results for the three survey questions for grade five students. Table 10.A.4 through table 10.A.6 show the survey results for grade eight students. Table 10.A.7 through table 10.A.21 show the survey results for the survey questions by grade level and for high school students overall. Table 10.A.21 shows results only for grades ten and eleven students, since the fourth survey question does not apply to grade twelve students.

With regard to the first survey question, "Did you learn about the topics on the test in your science classes?," the majority of students agreed that most or all of the topics were taught in their science classes. Specifically, this was endorsed by 87 percent of grade five students, 84 percent of grade eight students, and 70 percent of high school students. Among high school students, this was endorsed by 66 percent of grade ten students, 70 percent of grade eleven students, and 70 percent grade twelve students.

With regard to the second survey question, "Were any questions on the test different from the types of questions you see in science classes?," 47 percent to 61 percent of students across grades agreed that some of the questions on the test were different. Across grades,

33 percent to 47 percent of students responded that most of the questions on the test were different. These percentages were similar in grades five and eight but they were lower than those in high school grades.

With regard to the third survey question, “How hard were questions on this test compared to questions you see in science classes?,” 42 percent to 57 percent of students across grades felt that the test questions were about as hard as the questions in their science classes. Across grades, 32 percent to 51 percent of students felt that the test questions were harder than most questions in their science classes. The percentages in high school grades were higher than those in grades five and eight.

With regard to the fourth survey question, “Do you think you will be enrolling in any more science classes in high school?,” 69 percent of grade ten students and 54 percent of grade eleven students responded that they would enroll in more science classes during high school.

The correlations between student survey responses and their scale scores across grade levels ranged from -0.11 to 0.27, indicating there was no clear relationship between student responses on the survey and their performance on the CAST.

# Chapter 11: Continuous Improvement

This chapter documents the process whereby continuous improvements are ensured and includes the results of this process in the current year.

## 11.1. Test Design

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

The operational test forms for the 2019–2020 CAST administration will be delivered in accordance with the approved blueprint, as were the test forms for the 2018–2019 administration. However, in planning for the 2019–2020 administration and as a result of related research studies, Educational Testing Service (ETS), in collaboration with the CDE, developed an updated blueprint that was presented to the State Board of Education in January 2020. The updated blueprint was approved, and the administration of the 2020–2021 CAST will differ in a few key areas, such as that Segment B will contain three performance tasks.

## 11.2. Item Development

For the 2018–2019 item development cycle, the ETS content team used item specifications that make the alignment of all three dimensions of the California Next Generation Science Standards—disciplinary core ideas, science and engineering practices, and crosscutting concepts—clearer on CAST items. These item specifications were shared with the public during the 2018–2019 administration and met with positive feedback.

Item performance data from the 2017–2018 CAST field test administration provided additional information with which to review items during and after a data review meeting with teachers in the field in 2018. This additional information was key to making constructive changes to item development processes internal to ETS.

Field test item data showed what worked and what did not work, especially in terms of the language used in items. Both positive and negative exemplar items from the 2017–2018 field test administration were used in item developer training, including adjustments to how developers approach constructed-response items. Responses from students validated some aspects of development processes and pointed out deficits in other aspects of the assessments. One specific example of a process improvement was to provide concise language in items to explain the item type functionality.

As a result of the feedback that ETS received, the rejection rates for the newly developed items decreased significantly from the 2017–2018 item development cycle to the 2018–2019 cycle.

Table 11.1 shows, for each item development cycle, the total number of unique items put on the forms, the number of rejected items, and the rejection rates. For example, the rejection rate for grade five decreased from 24 percent to 3 percent from the 2017–2018 cycle to the 2018–2019 cycle.

**Table 11.1 Item Development Results**

| Grade Level | Total Number of Unique Items Field-Tested in 2017–2018 Administration | Number of Rejected Items From 2017–2018 Cycle | Rejection Rate From 2018 Data Review | Total Number of Unique Items Field-Tested in 2018–2019 Administration | Number of Rejected Items From 2018–2019 | Rejection Rate From 2019 Data Review |
|---|---|---|---|---|---|---|
| Grade 5 | 209 | 51 | 24% | 234 | 7 | 3% |
| Grade 8 | 211 | 10 | 5% | 228 | 11 | 5% |
| High school | 224 | 50 | 22% | 224 | 7 | 3% |

Improvements to item development that resulted in higher data review acceptance rates also resulted from the creation of the first iteration of item specifications. Refer to subsection *3.1.3 Incorporation into Item Development Processes* for more information on item specifications.

Work to refresh the CAST item bank will continue through subsequent development cycles with the goal of developing items of low, medium, and high complexity for the 2021–2022 administration.

## 11.3. Administration and Test Delivery

Test administration became more streamlined during the 2018–2019 testing year, with some improvements to the test delivery system. Test administrators were given visibility into what type of segment a student was on, whether it be discrete items or a performance task, allowing them to make a more informed decision about when to have a student pause a test.

ETS continues to improve in this area, and the 2019–2020 administration will introduce a progress bar for the student as the student progresses through their test. In contrast, to contend with this in 2018–2019, the number of items being taken by the student was removed from the student view to prevent that denominator from increasing as it did in previous administrations because of the delivery of the random blocks comprising the test.

# 11.4. Constructed-Response Item Scoring

## 11.4.1. Human Scoring Activities

Continuous improvements for the 2019–2020 administration included that rater agreement and validity statistics for rater agreement were monitored during each scoring shift that was worked. Scoring leaders provided feedback to ETS Assessment & Learning Technology Development to determine what adjustments to training or samples were to be made.

ETS will use standardized training to assist the scoring leader in using the performance indicator panels, which allow easier access to quantitative feedback regarding individual raters.

Improved training for scoring leaders will be conducted via online learning courses, as opposed to the WebEx sessions used previously. Online learning courses provide the following expected benefits:

- Standardized explanation on what information is available within the performance indicator panel as well as how to use the information
- Standardized format for providing feedback to individual raters to ensure the area of improvement needed is clear and consistent regardless of which score leader may be monitoring an individual rater on any given day
- Automatic restriction of scoring leaders from monitoring raters until training requirements from 2019–2020 have been satisfied (previously done manually)

# 11.5. Psychometric Analyses

Preequating will be implemented in the 2019–2020 test administration to allow an early reporting timeline. For details about the benefits, feasibility, and concerns of the preequating plan, refer to the *Feasibility of a Preequating Plan for the California Science Test (CAST) for the 2019–2020 Administration* (ETS, 2020).

A comprehensive scoring quality control process, necessary for the preequating process, is being developed and implemented for the 2019–2020 test administration to ensure the accuracy of all scores in early reporting.

ETS will also explore the possibility of providing more graphs and plots, rather than primarily using tables, in future technical reports, to make the technical reports more visualized and user friendly.

# 11.6. Accessibility

ETS increased the number of accessibility resources available, as evidenced by adding Green Hmong as a translation glossary in the items. As for the items themselves, ETS remains focused on making items accessible from the outset, reducing the need to provide extensive adaptations to make the items accessible to students with visual impairment.

# Reference

Educational Testing Service. (2020). *Feasibility of a preequating plan for the California Science Test (CAST) for the 2019–2020 administration*. [Unpublished memorandum]. Princeton, NJ: Educational Testing Service.

# Chapter 12: Test Dimensionality Study Addendum

## 12.1. Study Purpose

The California Next Generation Science Standards (CA NGSS) are considered three dimensional (3D), given the interrelationships between the disciplinary core ideas (DCIs), science and engineering practices (SEPs), and crosscutting concepts (CCCs). The California Science Test (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, which is directed by the CAST blueprint.

Several questions must be addressed to report reliable student scores that allow valid inferences about students' mastery of the CA NGSS. Examples of these questions include the following:

- Is the test essentially unidimensional, or are the integrated DCI, SEP, and CCC clearly distinguished?
- Similarly, do the performance tasks (PTs) measure something different than the discrete items?
- Finally, do the technology-enhanced items (TEIs) measure anything different from the traditional item types such as multiple-choice (MC) or constructed-response (CR) items?

These questions were addressed in a test dimensionality analysis. The dimensionality study reported here evaluated whether the CAST measures a single integrated science construct or several distinctive constructs. A determination informed how the test items were to be calibrated and how the scores were to be reported to best measure students' performance on the CAST. If the student response data from the assessment provided evidence of unidimensionality, or essential unidimensionality, all the items from different domains could be calibrated jointly, providing a total test score that could be reported based on all items the student took. If the student response data from the assessment provided evidence of multidimensionality, the items from different dimensions could be calibrated separately, and then the scores could be combined to provide a weighted composite score from these dimensions.

## 12.2. Study Design

Traditional factor analyses are typically used to evaluate the underlying structure of multiple variables. Research has shown the equivalence of the classical factor analyses and the item response theory (IRT) models (Kamata & Bauer, 2008; Takane & Leeuw, 1987).

In this study, two different models within the multidimensional IRT (MIRT) framework were used to evaluate test dimensionality: a bifactor model and a correlated factor MIRT model. Model specifications are included in subsection *12.3.1 Model Specification*.

### 12.2.1. Dimensional Structures

Five hypothesized dimensional structures were of interest in this study:

1. Content domain
2. Item type
3. SEP

4. CCC
5. Task type (i.e., discrete items versus the PTs).

Table 12.1 provides the details on the levels in each of these five hypothesized dimensional structures.

**Table 12.1 Hypothesized Dimensional Structures in the Study**

| Hypothesized Dimensional Structures | Number of Levels | Levels |
|---|---|---|
| Content domain | 3 | • Physical Sciences (PS)<br>• Life Sciences (LS)<br>• Earth and Space Sciences (ESS) |
| Item type | 3 | • CR items<br>• MC items<br>• TEIs |
| Science and engineering practice | 8 | • SEP 1: Asking questions (science), defining problems (engineering)<br>• SEP 2: Developing and using models<br>• SEP 3: Planning and carrying out investigations<br>• SEP 4: Analyzing and interpreting data<br>• SEP 5: Using mathematics and computational thinking<br>• SEP 6: Constructing explanations (science), designing solutions (engineering)<br>• SEP 7: Engaging in argument from evidence<br>• SEP 8: Obtaining, evaluating, and communicating information |
| Crosscutting concept | 7 | • CCC 1: Patterns<br>• CCC 2: Cause and effect<br>• CCC 3: Scale, proportion, and quantity<br>• CCC 4: Systems and system models<br>• CCC 5: Energy and matter<br>• CCC 6: Structure and function<br>• CCC 7: Stability and change |
| Task type | 2 | • Performance tasks<br>• Non-PTs |

## 12.2.2. Scores Reported

At the individual student level, only the total score and the content domain subscores were reported. Content domains were the most substantively intuitive classifications of items, were of practical interest, and were best supported by the test design because they had sufficient numbers of items to support the reporting of these subscores.

Although only content domain scores were included in the reporting of individual student scores, additional MIRT models by other structures (i.e., SEP, CCC, task type, and item

type) were run in this study to provide a complete picture on whether the test had any unexpected multidimensional structure that must be considered when calibrating and equating the test.

Students' scores were reported based on the items the students received. Therefore, it was most informative to run the test dimensionality study at the form level, with all the blocks that a student took contributing to reporting the student's scores, to inform individual student score reporting.

### 12.2.3. Assessment Structure

The CAST operational test is divided into segments, each of which can administer any of several alternative, randomly selected item blocks. For each grade-level assessment, there were two blocks in Segment A, so all students took the same two A blocks. However, there were multiple blocks from segments B and C. Each student was randomly assigned two PTs from two different domains in Segment B and randomly assigned either one discrete block or one PT from Segment C. This created multiple combinations of blocks in Segment A and Segment B that a student could receive. Instead of conducting the analyses on all possible combinations of PTs, three forms (i.e., block combinations) were carefully selected for evaluation using the following guidelines:

1. All items performed well based on the item-analysis results.
2. The number of items from the two PTs were balanced.
3. For PTs that had items from two domains where one was designated as the primary domain, the ones that had more items from the primary domain were selected.

Table 12.2 shows the item block combination in each of the three forms for each grade.

**Table 12.2 Forms Used in the Test Dimensionality Study**

| Grade or Grade Level | Form One | Form Two | Form Three |
|---|---|---|---|
| Grade 5 | A1–A2–B1–B2 | A1–A2–B1–B5 | A1–A2–B4–B5 |
| Grade 8 | A1–A2–B2–B4 | A1–A2–B1–B6 | A1–A2–B3–B6 |
| High school | A1–A2–B1–B2 | A1–A2–B1–B3 | A1–A2–B2–B3 |

The dimensionality study was run by form for the grade five, grade eight, and high school tests separately. For high school, the dimensionality study was run using a multigroup analysis with each grade as a group.

## 12.3. Methods

### 12.3.1. Model Specification

A bifactor model and a correlated factor MIRT model were used to evaluate the test dimensionality for each hypothesized dimensional structure. Figure 12.1 and figure 12.2 use path diagrams to show how the models were specified.

These two diagrams illustrate the relationships between the items and the general and specific factors for a test with *K* items and three specific factors. In these two figures, item 1 to item 4 are associated with specific factor 1, item 5 to item 9 are associated with specific factor 2, and the last three items are associated with specific factor 3. The single-sided arrow indicates the source that contributes to the item variance.

In figure 12.1, each item has three sources that contribute to the item variance: the general factor, the specific factor, and the error term. “Gen” refers to the general factor, “SF” refers to a specific factor, and “E” refers to the error term for each item.

Using the content domain structure as an example, the CAST has three major content domains, and each item is associated with one of the three domains. In the bifactor model presented in figure 12.1, each item has loadings on both the general factor and the content domain associated with an item. The content domains are uncorrelated in this model because their correlations have been reflected by the general factor.

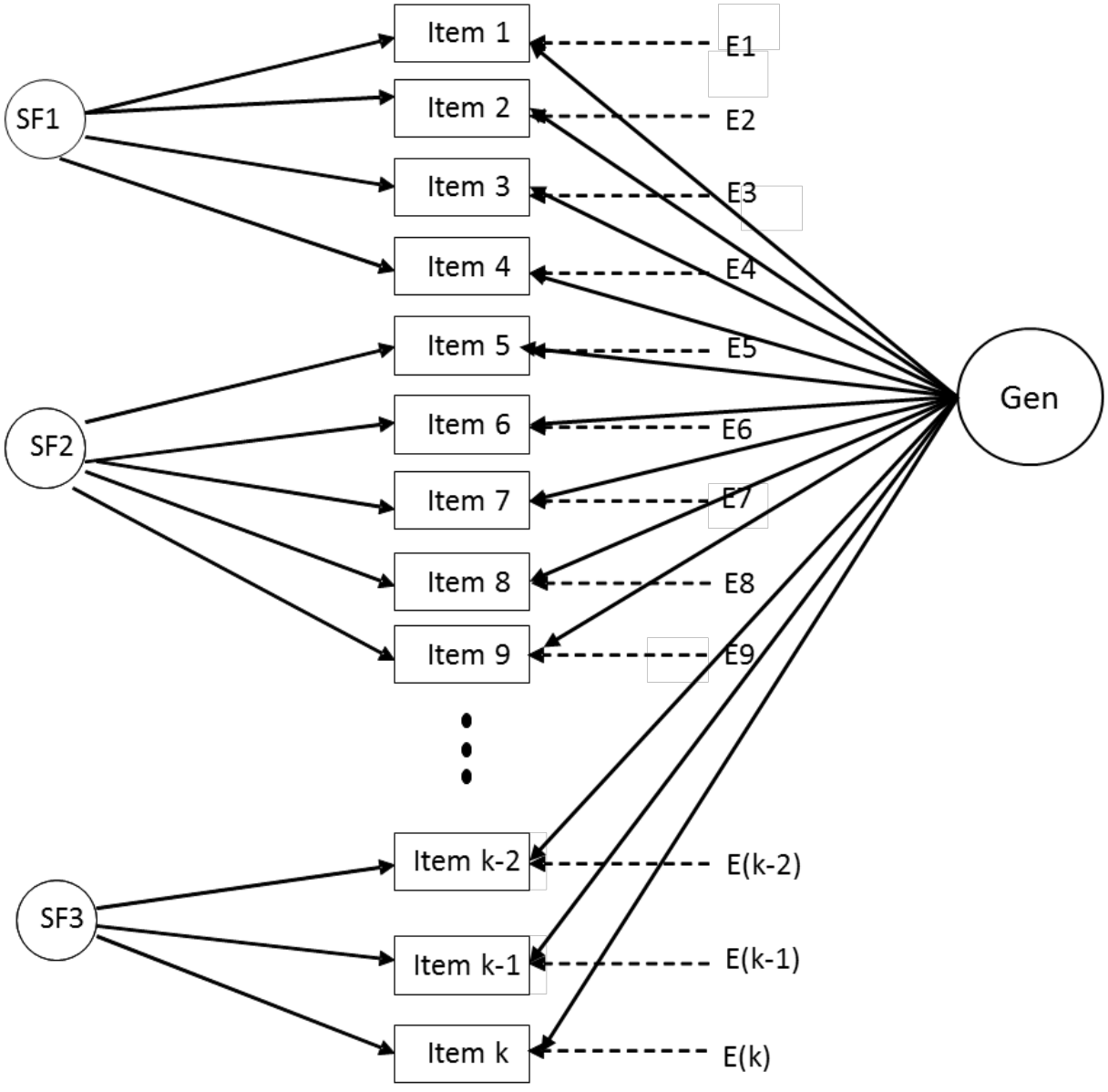


**Figure 12.1 The Path Diagram for a Bifactor Model**

In figure 12.2, each item has two sources that contribute to the item variance: the specific factor and the error term. There are double-sided arrows in figure 12.2 connecting each pair of the specific factors, indicating that the correlations between the specific factors are freely

estimated in this model. “F” refers to a specific factor and “E” refers to the error term for each item.

In the correlated factor MIRT model presented in figure 12.2, however, each item has a loading on only the associated content domain. The correlations of the three content domain scores are freely estimated in this model.

Note that even though the same items are associated with SF1 in figure 12.1 and F1 in figure 12.2 in these two models, the interpretation of SF1 and F1 is different. Using PS as an example, F1 in figure 12.2 indicates what students know about PS. However, the SF1 in figure 12.1 indicates what students know about PS that is not part of what they know about science.

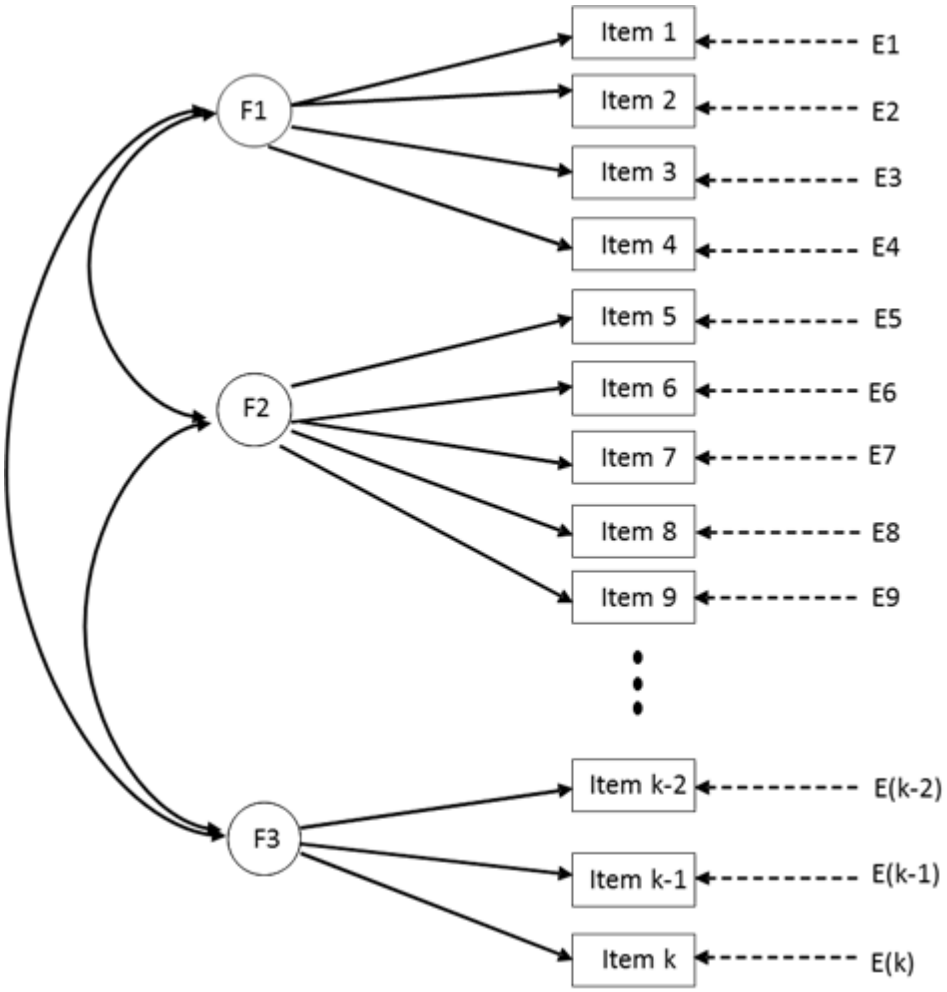


**Figure 12.2 The Path Diagram for a Correlated Factor MIRT Model**

In the IRT framework, the probability of obtaining a response pattern on $K$ items can be denoted as

$$P(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{\boldsymbol{k=1}}^{\boldsymbol{K}} P(y_k|\boldsymbol{\theta}), \tag{12.1}$$

*Refer to the Alternative Text for Equation 12.1 for a description of this equation.*

where,

$P(y_k|\boldsymbol{\theta})$ is the probability for a student with ability vector $\theta$ to get a score point $y$ on item $k$.

This probability $\pi = P(y_k|\boldsymbol{\theta})$ is related to a linear function of the latent variables $\theta$ through a link function $g(\pi)$, typically a probit or logit function. The difference in the bifactor model and the correlated factor MIRT model is in the link function. The link function (12.2) for a bifactor model is defined as

$$g(\pi) = a_{kg}\theta_g + a_{km}\theta_m + c_k, \quad (12.2)$$

*Refer to the Alternative Text for Equation 12.2 for a description of this equation.*

where,

$c_k$ is the intercept parameter for item $k$,

$a_{kg}$ is the loading on the general factor, and

$a_{km}$ is the loading on the dimension $m$.

The latent variables $\theta_g$ and $\theta_m$ are specified as uncorrelated; the link function for the correlated factor MIRT model is defined as

$$g(\pi) = a_{km}\theta_m + c_k, \quad (12.3)$$

*Refer to the Alternative Text for Equation 12.3 for a description of this equation.*

where,

the instances of $\theta_m$ are correlated.

The bifactor model and the correlated factor MIRT model provide different parameter estimates but should provide similar information on the test dimensionality. Strong unidimensionality will be reflected in the MIRT model with high correlations among all dimensions. While in the bifactor model, high correlations among the dimensions will be absorbed by the general factor, and the items will have insignificant loadings on the specific factors. The bifactor model has an added benefit of showing the proportion of variance that can be explained by the specific factor as a direct way to evaluate the test dimensionality.

A multigroup analysis was run for the high school assessment. Each grade was included in the model as a separate group. The mean and variance of the ability estimates were fixed at 0 and 1 for the grade eleven population but were freely estimated for the grades ten and twelve populations. In both models, the intercept and the loadings were set to be equal for all three grades. In the correlated factor MIRT model, the covariances among the dimensions were freely estimated for all three grades.

Both the bifactor and the correlated factor MIRT model in this study were fitted with the commercial FlexMIRT® software (Cai, 2016).

## 12.3.2. Local Dependency Evaluation

Local item dependency (LID) can cause an assessment to show secondary factors in a test dimensionality evaluation. Before fitting the data with the bifactor and the correlated factor MIRT models, an evaluation of LID was conducted to remove any construct-irrelevant dependencies among items that might have caused secondary factors.

LID is a well-known assumption for a unidimensional IRT model where the success of one item depends on the success of another item. Yen (1993) listed the factors that could cause the dependency among items:

- External assistance or interference
- Speededness
- Fatigue
- Practice effects
- Item or response format
- Passage dependence
- Item chaining
- Explanation of a previous answer
- Scoring rubrics or raters
- Content, knowledge, and abilities

Some factors, such as item or response format, passage dependence, or content and knowledge, could cause some item pairs to show dependency. The test dimensionality caused by such factors is of interest in the study, so such item(s) should not be removed. However, other factors that are unrelated to the test design and construct but might cause the item pairs to show dependency should be avoided because the test multidimensionality caused by such factors is not of interest in this study. For example, if two items measure the exact same skill, students who know the answer to one item will likely know the answer to the other item. Such dependencies should be removed prior to fitting the MIRT models to evaluate the test dimensionality.

The primary focus of this evaluation was to identify the pairs of items that measured the same or highly similar content, which led to a high correlation among the item pairs after partialling out the general ability effect. The commonly used Q3 statistic (Yen, 1984) was used in this study to evaluate LID. Q3 calculates correlation among the residual of a pair of items, where the residual was defined as the difference between students' scores and the expected item-score conditioning on a student's ability level. It is known that Q3 may not give comprehensive results for local dependencies that span more than two items, which is a common occurrence for item testlets. ETS will considering further exploration of investigating the LID using testlet models.

To control for the false discovery rate, the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995) was used to flag the pairs that displayed high correlations among the residuals. These flagged pairs were then evaluated by examining the content. If two items in a pair were highly similar in terms of the knowledge being measured, one item—typically, the less discriminating item—could be removed from the subsequent bifactor and MIRT model fitting.

Note that a common cause of LID in a passage-based test is when common stimuli are shared between items. In the CAST, all items in a PT were based on the same stimulus. However, these items were not removed simply because they shared the same stimuli.

Instead, the commonalities shared with items in a CAST PT might indicate the PTs are measuring something different than discrete items, which would be addressed through the bifactor model by task type.

After evaluating the LID, none of the items were removed from the subsequent dimensionality studies. One of the reasons is that the majority of items included in the dimensionality study are operational items that were used in the 2017–2018 field test administration. Items found to have issues in the field test were not included in the operational forms.

## 12.4. Model Evaluation Criteria

A determination of test dimensionality is a subjective judgment that weighs the different sources of empirical evidence. The following evidence was considered when determining whether the CAST follows a multidimensional, essentially unidimensional, or unidimensional structure:

- Item loadings on the general factor and the group-specific factors were examined. High loadings on the general factor and low loadings on the group-specific factors for most of the items suggest that the unidimensional model is sufficient for the data. It should be noted that the difficulties in interpreting the results of specific factors, which are not loadings on the named factor, but rather low loadings on an uninterpretable factor that is orthogonal to the general factor.
- The indices proposed by Rodriguez, Reise, and Haviland (2016) evaluate the strength of the general factor and, therefore, indirectly assess practical unidimensionality. These indices, used in this study to evaluate test dimensionality, are:
  - **Omega hierarchical (OmegaH) and Omega hierarchical subscale (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, a large amount of the variance can still be explained by the group-specific variance, which could indicate multidimensionality.
  - **Explained common variance (ECV):** ECV is the ratio of the variance explained by the general factor divided by the variance explained by the general and the group-specific factor. A high ECV value is evidence of an essentially unidimensional model (Sijtsma, 2009; Ten Berge and Socan, 2004).
  - **Relative parameter bias (RPB):** An item parameter estimate could be biased when a test with multidimensional structure is forced to conform to a unidimensional model. That is, a unidimensional model is fit to data with a multidimensional structure. RPB reflects the amount of bias in the item loadings when the test with a bifactor structure is fitted using a unidimensional model and is defined as

$$RPB = \sum(\hat{\gamma}_g - \hat{\gamma}_u)/n \quad (12.4)$$

*Refer to the Alternative Text for Equation 12.4 for a description of this equation.*

where,

$\hat{\gamma}_g$ is the item loading on the general factor in a bifactor model, and

$\hat{\gamma}_u$ is the item loading in a unidimensional model.

The mean and variance of the theta in the unidimensional model were constrained to be (0,1), and the same constraints were applied to the factors in the bifactor. This ensured the metric for the bifactor general dimension and the single dimension from the unidimensional analysis were on the same metric.

## 12.5. Results

Results from the five hypothesized dimensional structures are presented in this section.

For each hypothesized dimensional structure, the factor loading matrices created from a bifactor model for Form One of the grade five assessment are presented in appendix 12.A. The patterns of the factor loadings for the other forms in grade five and for all forms from grade eight and high school follow similar patterns and are not presented individually.

### 12.5.1. Test Dimensionality by Content Domain

The factor loading matrix for the bifactor model by content domain for grade five, Form One, is shown in table 12.A.1. Items without loadings on the specific factors are noted with a hyphen (-). To facilitate reading the factor loading matrix, the matrix is displayed by grouping the items that belong to the same dimension together.

As the data indicates, most items have higher loadings on the general factor and smaller loadings on the specific factors, which suggests that the assessment is essentially unidimensional. While the general factor is the main contributor to the total score variance, a few items have low loadings on both the general and the specific factor (e.g., items 2, 22, 36, and 47 in table 12.A.1). Possible causes of the low loadings are that these items have low discriminating power or that they measure something different than all other items in general.

#### 12.5.1.1. Test Dimensionality Indices by Content Domain

Table 12.3 provides the other indices used to evaluate the test dimensionality for content domains: OmegaH, OmegaHS, ECV, and RPB.

**Table 12.3 Evaluation Indices for the Bifactor Model by Content Domain**

| Grade and Form | OmegaH | OmegaHS: ESS | OmegaHS: LS | OmegaHS: PS | ECV | RPB |
|---|---|---|---|---|---|---|
| Grade 5, Form One | 0.94 | 0.00 | 0.00 | 0.01 | 0.94 | 0.011 |
| Grade 5, Form Two | 0.94 | 0.01 | 0.04 | 0.01 | 0.92 | 0.000 |
| Grade 5, Form Three | 0.93 | 0.00 | 0.04 | 0.00 | 0.93 | 0.000 |
| Grade 8, Form One | 0.92 | 0.00 | 0.01 | 0.02 | 0.88 | -0.001 |
| Grade 8, Form Two | 0.92 | 0.00 | 0.01 | 0.03 | 0.89 | 0.002 |
| Grade 8, Form Three | 0.92 | 0.00 | 0.03 | 0.02 | 0.87 | 0.005 |
| High school, Form One | 0.91 | 0.02 | 0.03 | 0.02 | 0.89 | 0.016 |
| High school, Form Two | 0.91 | 0.03 | 0.02 | 0.01 | 0.90 | 0.008 |
| High school, Form Three | 0.92 | 0.03 | 0.02 | 0.02 | 0.90 | 0.000 |

Using Form One from grade five as an example, OmegaH is 0.94, suggesting that 94 percent of the variance in the total score could be attributed to the single general factor. The values of OmegaHS for three content domains are close to zero, suggesting that, after accounting for the general factor, the group-specific factors only accounted for a very small proportion of the total score variance.

The ECV index shows the percent of common variance explained by the general factor and is a more direct way to measure the strength of the general factor in comparison to the content-domain specific factors. The ECV for Form One from grade five is 0.94, which means that 94 percent of the common variance was explained by the general factor and only 6 percent by the three content-domain factors.

The RPB is 0.011 for this form, indicating that if the data has a multidimensional structure but is fitted with a unidimensional model, the average bias in the factor loadings is small and negligible.

Overall, the results are consistent across forms and grades, indicating essentially unidimensional data structure based on the content-domain specification.

#### 12.5.1.2. Factor Correlations by Content Domain

Table 12.4 shows the ranges of the correlations of the group-specific factors (i.e., content domain scores) across forms and grades when the data was fitted with the correlated factor MIRT model. The high correlations among the three content domain scores indicate the unidimensionality of the test.

**Table 12.4 Correlations Among the Latent Content Domain Scores**

| Domain | ESS | LS | PS |
|---|---|---|---|
| ESS | 1.00 | 0.88–0.95 | 0.90–0.97 |
| LS | 0.88–0.95 | 1.00 | 0.91–0.95 |
| PS | 0.90–0.97 | 0.91–0.95 | 1.00 |

### 12.5.2. Test Dimensionality by SEP

The factor loading matrices for the bifactor model by SEP for grade five, Form One, is shown in table 12.A.2. Items without loadings on the specific factors are noted with a hyphen (-). The names of the SEPs are listed in table 12.1.

In the test dimensionality analysis, only SEPs with five or more items were included as specific factors, and items associated with SEPs with fewer than five items had loadings on the general factor only.

While the CA NGSS includes eight SEPs, only four SEPs had five or more items in this form. As a result, only four SEPs are included as group-specific factors in table 12.A.2. There were 14 items in this table with loadings on the general factor only and not on the specific factors. (There were fewer than five items per SEP for the other four SEPs not listed in the table.)

Similar to the results from the content domain structure, the moderate-to-large loadings on the general factor and the small loadings on the SEP-specific factors suggest that the general factor is the main contributor to the total test score variance.

#### 12.5.2.1. Test Dimensionality Indices by SEP

Table 12.5 provides the other indices used to evaluate the test dimensionality by SEP across grade-level forms: OmegaH, OmegaHS, ECV, and RPB. A SEP that was not measured on a form or had fewer than five items is noted with a hyphen (-).

**Table 12.5 Evaluation Indices for the Bifactor Model by SEP**

| Grade and Form | OmegaH | OmegaHS: SEP 2 | OmegaHS: SEP 3 | OmegaHS: SEP 4 | OmegaHS: SEP 5 | OmegaHS: SEP 6 | OmegaHS: SEP 7 | ECV | RPB |
|---|---|---|---|---|---|---|---|---|---|
| Grade 5, Form One | 0.94 | 0.03 | 0.04 | - | - | 0.03 | 0.03 | 0.91 | 0.010 |
| Grade 5, Form Two | 0.94 | 0.02 | 0.02 | 0.02 | - | 0.01 | 0.03 | 0.92 | -0.003 |
| Grade 5, Form Three | 0.93 | 0.00 | 0.07 | 0.02 | - | 0.03 | 0.04 | 0.87 | -0.001 |
| Grade 8, Form One | 0.91 | 0.04 | - | 0.03 | - | 0.04 | - | 0.88 | 0.007 |
| Grade 8, Form Two | 0.91 | 0.01 | - | 0.04 | - | 0.04 | - | 0.90 | 0.000 |
| Grade 8, Form Three | 0.92 | 0.05 | - | 0.03 | - | 0.01 | - | 0.89 | 0.004 |
| High school, Form One | 0.91 | 0.03 | - | 0.04 | 0.00 | 0.00 | - | 0.92 | 0.000 |
| High school, Form Two | 0.91 | 0.01 | - | 0.06 | - | 0.04 | 0.07 | 0.90 | 0.000 |
| High school, Form Three | 0.92 | 0.05 | - | - | 0.00 | 0.04 | 0.08 | 0.89 | 0.001 |

The OmegaH index for the total test is high, ranging from 0.91 to 0.94. The OmegaHS values are close to zero, which suggests that, after accounting for the general factor, the SEP-specific factors only accounted for a very small proportion of the total score variance.

The ECV values are high, ranging from 0.87 to 0.92, suggesting that more than 85 percent of the common variance is explained by the general factor and only less than 15 percent of the common variance is spread across the SEP-specific factors. The RPB is negligible for all nine forms, indicating that even if there is a bifactor structure in the data, items can be calibrated properly using a unidimensional model.

#### 12.5.2.2. Factor Correlations by SEP

Table 12.6 shows the range of the correlations between the group-specific factors (i.e., SEPs) across forms and grades, when the data was fitted with the correlated-factor MIRT model. A hyphen (-) notes where no correlation can be estimated for this pair of SEPs either because both SEPs in the pair were not measured or there were not at least five items in all nine forms.

The correlations for all forms are above 0.83. The high correlations among the SEPs indicate that there are no clearly distinctive dimensions by SEPs for the CAST.

**Table 12.6 Correlations Among the Latent SEP Scores**

| SEP | SEP 2 | SEP 3 | SEP 4 | SEP 5 | SEP 6 | SEP 7 |
|---|---|---|---|---|---|---|
| SEP 2 | 1.00 | 0.92–0.93 | 0.87–0.94 | 0.85–0.93 | 0.86–0.95 | 0.85–0.93 |
| SEP 3 | 0.92–0.93 | 1.00 | 0.92–0.93 | - | 0.92–0.93 | 0.90–0.91 |
| SEP 4 | 0.87–0.94 | 0.92–0.93 | 1.00 | 0.87–0.90 | 0.84–0.94 | 0.89–0.92 |
| SEP 5 | 0.85–0.93 | - | 0.87–0.9 | 1.00 | 0.88–0.92 | 0.90–0.92 |
| SEP 6 | 0.86–0.95 | 0.92–0.93 | 0.84–0.94 | 0.88–0.92 | 1.00 | 0.86–0.93 |
| SEP 7 | 0.85–0.93 | 0.90–0.91 | 0.89–0.92 | 0.90–0.92 | 0.86–0.93 | 1.00 |

### 12.5.3. Test Dimensionality by CCC

The factor loading matrices from the bifactor model by CCC for grade five, Form One, are shown in table 12 A.3. Items without loadings on the specific factors are noted with a hyphen (-).

The names of the CCCs are listed in table 12.1. Similar to the dimensionality analysis performed on the SEPs, the analysis includes only the CCCs with at least five items. While the CA NGSS includes seven CCCs, only five of them have five or more items. The four items with loadings on the general factor and none on the CCC-specific factors are the ones associated with those remaining two CCCs because of the small number of items.

Most items have higher loadings on the general factor than on the CCC-specific factors, although this relationship is reversed for a few items (e.g., item 31 in table 12.A.3). The loadings on the CCC-specific factors for the items belonging to the same CCC appear to show different patterns, with some large loadings, some close to zero, and some negative. This suggests that some items might be measuring something different from other items that cannot be attributed to the same CCC.

### 12.5.3.1. Test Dimensionality Indices by CCC

Table 12.7 provides the other indices used to evaluate the test dimensionality for CCCs across grade-level forms (OmegaH, OmegaHS, ECV, and RPB).

**Table 12.7 Evaluation Indices for the Bifactor Model by CCC**

| Grade and Form | OmegaH | OmegaHS: CCC-1 | OmegaHS: CCC-2 | OmegaHS: CCC-3 | OmegaHS: CCC-4 | OmegaHS: CCC-5 | OmegaHS: CCC-6 | OmegaHS: CCC-7 | ECV | RPB |
|---|---|---|---|---|---|---|---|---|---|---|
| Grade 5, Form One | 0.94 | 0.02 | 0.00 | 0.07 | 0.01 | 0.02 | - | - | 0.92 | 0.007 |
| Grade 5, Form Two | 0.94 | 0.01 | 0.01 | 0.06 | 0.03 | - | - | - | 0.90 | 0.001 |
| Grade 5, Form Three | 0.94 | 0.01 | 0.01 | 0.01 | 0.02 | - | - | - | 0.93 | 0.000 |
| Grade 8, Form One | 0.91 | 0.04 | 0.00 | - | - | 0.06 | - | - | 0.88 | 0.002 |
| Grade 8, Form Two | 0.91 | 0.03 | 0.03 | 0.00 | - | - | - | - | 0.91 | -0.001 |
| Grade 8, Form Three | 0.92 | 0.00 | 0.03 | 0.00 | - | 0.05 | 0.08 | - | 0.83 | -0.003 |
| High school, Form One | 0.91 | 0.03 | 0.05 | 0.04 | - | - | - | 0.01 | 0.88 | 0.003 |
| High school, Form Two | 0.92 | 0.05 | 0.00 | 0.01 | - | 0.00 | - | 0.01 | 0.90 | -0.012 |
| High school, Form Three | 0.92 | 0.03 | 0.08 | - | - | 0.00 | - | 0.01 | 0.88 | -0.009 |

The OmegaH values range from 0.91 to 0.94, indicating a large proportion of the total score variance can be attributed to the general factor. The values for OmegaHS are small, indicating that, after accounting for the general factor, these CCC-specific factors contribute little to the total score variance.

The ECV values range from 0.83 to 0.93, indicating that the general factor accounts for the majority of the common variance. The small RPB indicates that items can be calibrated properly using a unidimensional model.

### 12.5.3.2. Factor Correlations by CCC

Most correlation ranges have the lower bound higher than 0.8. There are two ranges (i.e., CCC2 and CCC3, and CCC2 and CCC5) with the lower bound lower than 0.8. The evaluation of all correlations reveals that out of 103 correlations included in table 12.8, there are only three correlations that are below 0.8. Most of the correlations—62 out of 104—are above 0.9. The moderate-to-high correlations suggest the test is essentially unidimensional with the CCC structure.

Table 12.8 shows the ranges of each pair of the latent CCC scores under the correlated factor MIRT model. A hyphen (-) notes where no correlation can be estimated for this pair of CCCs because either both CCCs in the pair were not measured or there were not at least five items in all nine forms.

**Table 12.8 Correlations Among Latent CCC Scores**

| CCC | CCC 1 | CCC 2 | CCC 3 | CCC 4 | CCC 5 | CCC 6 | CCC 7 |
|---|---|---|---|---|---|---|---|
| CCC 1 | 1.00 | 0.85–0.95 | 0.85–0.94 | 0.93–0.93 | 0.82–0.92 | 0.91–0.91 | 0.9–0.94 |
| CCC 2 | 0.85–0.95 | 1.00 | 0.74–0.94 | 0.95–0.95 | 0.78–0.91 | 0.91–0.91 | 0.84–0.90 |
| CCC 3 | 0.85–0.94 | 0.74–0.94 | 1.00 | 0.91–0.91 | 0.82–0.91 | 0.89–0.89 | 0.85–0.91 |
| CCC 4 | 0.93–0.93 | 0.95–0.95 | 0.91–0.91 | 1.00 | 0.92–0.92 | - | - |
| CCC 5 | 0.82–0.92 | 0.78–0.91 | 0.82–0.91 | 0.92–0.92 | 1.00 | 0.93–0.93 | 0.83–0.90 |
| CCC 6 | 0.91–0.91 | 0.91–0.91 | 0.89–0.89 | - | 0.93–0.93 | 1.00 | - |
| CCC 7 | 0.90–0.94 | 0.84–0.90 | 0.85–0.91 | - | 0.83–0.90 | - | 1.00 |

## 12.5.4. Test Dimensionality by Item Type

Three items types (MC, CR, and TEIs) were included. The MC items included both the single-selection and multiple-selection items. The items that required extended text input were considered CR items. The rest of the items were considered the TEIs (e.g., grid, inline choice, numeric entry items).

### 12.5.4.1. Test Dimensionality Indices by Item Type

Table 12.A.4 shows the factor loading matrix from the bifactor model by item type and for grade five, Form One. Most items have high loadings on the general factor and low loadings on the item-type-specific factors. A few items—e.g., items 9, 18, and 37—have low loadings on both the general factor and the item-type-specific factor, indicating they might be measuring something different from both the general and the item-type-specific factors.

Table 12.9 provides the other indices used to evaluate the test dimensionality across forms and grades. The OmegaH values range from 0.91 to 0.94, indicating a large proportion of the total score variance can be attributed to the general factor. The ECV values range from 0.87 to 0.93, indicating that the general factor accounts for most of the common variance. Compared to the OmegaHS values for the MC and TEI dimensions, which are mostly close to zero, these values for the CR dimension are slightly larger. The slightly larger OmegaHS values for the CR dimension suggest that there might be a slight multidimensionality structure for the CR dimension; however, the small RPB indicates it is still appropriate to calibrate the data with a unidimensional model.

**Table 12.9 Evaluation Indices for the Bifactor Model by Item Type**

| Grade and Form | Omega H | Omega HS: CR | Omega HS: MC | Omega HS: TEI | ECV | RPB |
|---|---|---|---|---|---|---|
| Grade 5, Form One | 0.94 | 0.06 | 0.00 | 0.01 | 0.93 | -0.001 |
| Grade 5, Form Two | 0.94 | 0.09 | 0.01 | 0.02 | 0.91 | 0.000 |
| Grade 5, Form Three | 0.94 | 0.09 | 0.00 | 0.00 | 0.92 | -0.001 |
| Grade 8, Form One | 0.91 | 0.07 | 0.06 | 0.00 | 0.91 | 0.025 |
| Grade 8, Form Two | 0.91 | 0.08 | 0.05 | 0.00 | 0.90 | 0.005 |
| Grade 8, Form Three | 0.92 | 0.08 | 0.03 | 0.01 | 0.87 | 0.008 |

Table 12.9 *(continuation)*

| Grade and Form | Omega H | Omega HS: CR | Omega HS: MC | Omega HS: TEI | ECV | RPB |
|---|---|---|---|---|---|---|
| High school, Form One | 0.92 | 0.08 | 0.00 | 0.00 | 0.91 | -0.008 |
| High school, Form Two | 0.92 | 0.04 | 0.00 | 0.01 | 0.87 | 0.000 |
| High school, Form Three | 0.92 | 0.08 | 0.00 | 0.01 | 0.91 | -0.007 |

Table 12.10 shows the correlations among the latent item-type scores. The correlations are all higher than 0.88, suggesting an essential unidimensionality by the item-type structure.

**Table 12.10 Correlations Among the Latent Item-Type Scores**

| Item Type | MC | CR | TEI |
|---|---|---|---|
| MC | 1.00 | 0.88–0.92 | 0.89–0.94 |
| CR | 0.88–0.92 | 1.00 | 0.92–0.97 |
| TEI | 0.89–0.94 | 0.92–0.97 | 1.00 |

The correlations between MC and CR are slightly smaller than the correlations between MC and TEIs, mostly because both MC and TEIs are both selected-response items while CR items require students to produce content as part of their response.

### 12.5.5. Test Dimensionality by Task Type

To evaluate whether the PTs and the non-PTs show distinctive dimensionality, the bifactor model and the correlated factor MIRT model were fitted to the data by the task type. Table 12.A.5 shows the factor loading matrix for the bifactor model by non-PT items and PT items. Similar to the results from the other hypothesized dimensional structures, the results for most items reveal higher loadings on the general factor than the task-type-specific factor. The loadings on the task-type-specific factors are low in general, with exceptions for a few items. Overall, there is no clear evidence of multidimensionality by the task type.

Table 12.11 provides the other indices used to evaluate the test dimensionality across forms and grades.

**Table 12.11 Evaluation Indices for the Bifactor Model by Task Type**

| Grade and Form | OmegaH | OmegaHS: Non-PTs | OmegaHS: PTs | ECV | RPB |
|---|---|---|---|---|---|
| Grade 5, Form One | 0.93 | 0.01 | 0.05 | 0.92 | 0.016 |
| Grade 5, Form Two | 0.94 | 0.00 | 0.04 | 0.93 | 0.002 |
| Grade 5, Form Three | 0.93 | 0.02 | 0.01 | 0.91 | 0.005 |
| Grade 8, Form One | 0.92 | 0.00 | 0.06 | 0.87 | -0.007 |
| Grade 8, Form Two | 0.92 | 0.00 | 0.04 | 0.89 | -0.003 |
| Grade 8, Form Three | 0.92 | 0.01 | 0.05 | 0.85 | 0.006 |
| High school, Form One | 0.91 | 0.01 | 0.04 | 0.91 | 0.009 |
| High school, Form Two | 0.91 | 0.02 | 0.04 | 0.91 | 0.004 |
| High school, Form Three | 0.88 | 0.08 | 0.01 | 0.89 | 0.022 |

The OmegaH values range from 0.88 to 0.94, indicating that a large proportion of the total score variance can be attributed to the general factor. The values of the OmegaHS for the PTs and non-PTs are small. The ECV values range from 0.85 to 0.93, indicating that the

general factor accounts for most of the common variance. The small RPB indicates it is appropriate to calibrate the data with a unidimensional model. The correlations between the latent PT and non-PT scores under the correlated factor MIRT model range from 0.90 to 0.95. These high correlations also indicate an essential unidimensionality by task type.

## 12.6. Implications on Calibration and Score Reporting

Test dimensionality has implications regarding how items should be calibrated and how scores should be reported. For example, if a test shows strong multidimensionality by content domain, it would be appropriate to calibrate the items from different content domains separately and establish a separate scale for each of them. The total score reported could then be a composite score of the different content domains.

For the CAST, the results from all five hypothesized dimensional structures consistently indicate unidimensionality for this test. This reflects the fact that the CAST is designed to measure the integration of DCIs, SEPs, and CCCs, rather than the dimensions coinciding. When a test is unidimensional by content domain, all items in different content domains can be calibrated jointly using a unidimensional model. As a result, all items from different domains are put on the same IRT scale. Because student performance is largely undifferentiated across the content domains, it is appropriate to use a single integrated score based on all items to reflect students' performance on the test. The domains with less stringencies are those in which students are more proficient. Reporting at the domain level is still possible, although it might not provide much unique information from the total score for most students.

It should be noted that the focus of the study was to find a model that best fits the data to provide a practical guideline in terms of calibration and reporting. Dimensionality was assessed with two modeling approaches—the bifactor model and correlated factors model—to provide a more comprehensive exploration of any dimensionality in the test. It is possible that, in some forms, a few items do not measure the general factor as strong as other items; however, there is not strong enough evidence to support treating and scaling the test as multidimensional. Efforts were made to find some common features of those items that showed low general factor loadings, but no simple explanation could be found. It might be worthwhile to have content experts take a further look into the factor loading matrix and explore the meaning of those items with low loadings, which might help improve the item and test development in the future.

# References

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc*. *B, 57*, 289–300.

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.

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling, 15*, 136–53.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Method*s, *21*(2), 137–50.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*(1), 145–54.

Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*, 393-408.

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), 613–25.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125–45.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187–213.

# Accessibility Information

## Alternative Text for Equation 12.1

Probability of y given theta equals product from k equals 1 to K of probability of y sub k given theta.

## Alternative Text for Equation 12.2

G of pi equals a sub kg multiples theta sub g plus a sub km multiples theta sub m plus c sub k.

## Alternative Text for Equation 12.3

G of pi equals a sub km multiples theta sub m plus c sub k.

## Alternative Text for Equation 12.4

RPB equals sum of fraction with numerator open bracket gamma sub g hat minus gamma sub u hat end bracket and denominator n.