California Department of Education
Assessment Development and Division

California Assessment of
Student Performance and Progress

California Assessment of Student Performance and Progress

California Science Test
Pilot Technical Report
2016–17 Administration

Final Submitted May 29, 2018

Educational Testing Service

Contract No. CN150012

Table of Contents

Tables

Figures

Acronyms and Initialisms Used in the CAST Technical Report

Abbreviation	Term
AD	Assessment Development
AERA	American Educational Research Associatin ion
AI	artificial intelligence
AIR	American Institutes for Research
AIS	average item score
APA	American Psychological Association
ASL	American Sign Language
CA NGSS	California Next Generation Science Standards
CAA	California Alternate Assessment
CAASPP	California Assessment of Student Performance and Progress
CALPADS	California Longitudinal Pupil Achievement Data System
CalTAC	California Technical Assistance Center
CAST	California Science Test
CCC	crosscutting concepts
CCR	*California Code of Regulations*
CCSS	Common Core State Standards
CDE	California Department of Education
CR	constructed response
CR Scoring Systems and Capabilities	CR Scoring S&C
DCI	disciplinary core ideas
DIF	differential item functioning
DOK	Depth of Knowledge
EC	*Education Code*
ECD	Evidence-Centered Design
EL	English learner
eSKM	Enterprise Score Key Management System
ETS	Educational Testing Service
IA	item analyses
ICC	intraclass correlation
ID	item identification
IFEP	Initially Fluent English Proficient
IEP	individualized education program
IMS	*Instructional Management Systems*
IRC	Item Review Committee
KSA	knowledge, skills, and abilities
LEA	local educational agency
MC	multiple choice
MH DIF	Mantel-Haenszel Differential Item Functioning

Abbreviation	Term
NCME	National Council on Measurement in Education
ONE	Online Network for Evaluation
OTI	Office of Testing Integrity
PE	performance expectation
PT	performance task
QA	quality assurance
QC	quality control
QTI	*Question and Test Interoperability*
RFEP	Reclassified Fluent English Proficient
SBE	State Board of Education
SD	standard deviation
SEP	science and engineering practice
SFTP	secure file transfer protocol
SMD	standardized mean difference
SSID	Statewide Student Identifier
STAIRS	Security and Test Administration Incident Reporting System
STS	Standards-based Tests in Spanish
TBD	to be determined
TDS	test delivery system
TE	technology enhanced
TEI	technology-enhanced items
TOMS	Test Operations Management System
UAT	User Acceptance Testing
USC	United States Code

Chapter 1: Introduction

1.1 Background

In October 2013, Assembly Bill 484 established the California Assessment of Student Performance and Progress (CAASPP) as the new student assessment system that replaced the Standardized Testing and Reporting program. The primary purpose of the CAASPP System of assessments is to assist teachers, administrators, and students and their parents/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 started using the California Next Generation Science Standards (CA NGSS) in its curriculum in September 2013. The California Science Test (CAST) is an online assessment aligned to the CA NGSS and was administered as a pilot for the first time during the 2016–17 CAASPP administration. This new assessment is for students in grades five, eight, and high school. For the CAST pilot, students in grades ten, eleven, or twelve were tested based on the high school test assignment plan, which is discussed in subsection 1.4.

During the 2016–17 administration, the overall CAASPP System had the following components:

- Smarter Balanced assessments and tools:
  - Summative Assessments—Online assessments for English language arts/literacy (ELA) and mathematics in grades three through eight and grade eleven
  - Interim Assessments—Optional resources developed for grades three through eight and grade eleven designed to inform and promote teaching and learning by providing information that can be used to monitor student progress toward mastery of the Common Core State Standards (CCSS) and that may be administered to students at any grade level
  - Digital Library—Tools and practices designed to help teachers utilize formative assessment processes for improved teaching and learning in all grades
- California Alternate Assessments (CAAs) for ELA and mathematics in grades three through eight and grade eleven
- CAA for Science in grades five, eight, and ten, eleven, or twelve
- CAST in grades five, eight, and high school
- A primary language assessment, the Standards-based Tests (STS) in Spanish for Reading/Language Arts, in grades two through eleven (optional for eligible Spanish-speaking English learners)

The STS assessment is paper-pencil, while all other assessments are online assessments. 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 Purposes of the CAST Pilot Test

The purposes of the CAST pilot were to

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

The CAST pilot was intended to assess item performance and not student performance. Although data were collected at the item level, student scores were not reported for the CAST pilot.

1.3 Pilot Test Content

Each grade/grade span assessed had two general fixed forms, where students were given the same questions regardless of their responses or ability. Both discrete items and performance tasks were included in the tests.

Each test form consisted of 12–15 discrete items and one performance task containing six items. The performance task 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. The discrete items included traditional multiple-choice items and constructed-response (CR) items and innovative technology enhanced (TE) items (refer to subsection *3.1 Item Writing*). One accessibility form was available for students who needed one or more designated supports and/or accommodations (refer to subsection *2.3 Test Administration*). The accessibility form was one fixed form per grade span; each form was composed of seven to eight discrete items.

Table 1.1 lists the number of discrete items of each form. Totals do not include any performance task items.

Table 1.1. Forms in the CAST Pilot

Form	Grade 5	Grade 8	High School
Form 1	13	15	12
Form 2	12	14	14
Accessibility Form	7	7	8

There were three performance tasks in total, with one administered in each grade span. Forms 1 and 2 for a given grade used the same performance task.

1.4 Intended Population

The CAST pilot was a census test administered to an estimated 1.4 million students in the general population. The intended population was all students in grades five and eight, and high school students within a selected grade that were assigned for each school based on CDE guidelines (refer to subsection *4.1 Assignment Design* for more details about the high school grade assignments).

Students eligible for alternate assessments took the CAA for Science in grades five and eight. Students in high school took the CAA for Science in grades ten, eleven, or twelve, depending on high school grade assignments.

For the CAST, no exceptions were made for English learner (EL) students—English learners (ELs) were expected to participate in the CAST unless they are designated to take the CAA for Science. ELs are defined as follows:

> "English learner students are those students for whom there is a report of a primary language other than English on the state-approved Home Language Survey **and** who, on the basis of the state approved oral language (grades kindergarten through grade twelve) assessment procedures and literacy (grades three through twelve only), have been determined to lack the clearly defined English language skills of listening comprehension, speaking, reading, and writing necessary to succeed in the school's regular instructional programs."[1]

1.5 Testing Window and Times

The CAST pilot was administered during a testing window selected by the local educational agency (LEA), with the first possible date of administration being March 20, 2017, and the last possible date being July 17, 2017. The testing dates fell within the LEA's selected testing window.

Similar to other CAASPP assessments, the CAST pilot was untimed for test takers. A student could take the CAST pilot test within the LEA's testing window over as many days as required to meet a student's needs (*California Code of Regulations,* Title 5*,* Education, Division 1, Chapter 2, Subchapter 3.75, Article 2*,* Section 855[a][3]).

1.6 Preparation for LEAs

To ensure the 2016–17 test administration was a successful experience for CAST test administrators and students, Educational Testing Service (ETS) provided onsite test administration workshops in various locations throughout California in January and February 2017 and produced Webcasts and videos with detailed information on CAASPP test administration procedures. In addition, ETS provided a number of test administration resources to schools and LEAs. These resources included detailed information on topics such as technology readiness, test administration, test security, accommodations, TDS, and other general testing rules.

1.7 Limitations of the Assessment

Due to the innovative item types being used to assess these new standards, providing full accessibility features for every item was difficult. To make the test accessible for all students, a separate accessibility form was provided. The accessibility features pilot forms were intended to test the accessibility features and had fewer items than the forms given to the general test population, because not all the items could be embedded with accessibility features as needed. However, the capacity to include accessibility features will increase

[1] "English Learner (EL) Students (Formerly Known as Limited-English-Proficient or LEP)," from the CDE Glossary of Terms Web page at http://www.cde.ca.gov/ds/sd/cb/glossary.asp.

during the preoperational tests, to allow for full accessibility features available in full-length forms in the operational administrations for students needing these resources.

Another unique challenge of the CAST pilot test was the alignment of the tests to the curriculum, because California recently adopted the CA NGSS, which are distinctly different from the previous California science standards. In addition, it was also challenging to align the pilot tests to the assessment model. Because the purpose of the pilot tests was to evaluate the items rather than students' knowledge, skills, and abilities (KSA), the pilot tests were not a full representation of the assessment model of the CA NGSS. Refer to subsection *3.2 Item Writing* for the model.

1.8 Groups and Organizations Involved with the CAST Assessment

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, establishes regulations to implement legislation, and has the authority to grant waivers of the *Education Code*.

In addition to adopting the rules and regulations for itself, its appointees and California's public schools, the SBE also is the state educational agency responsible for overseeing California's compliance of 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, 2016).

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 6,200,000 children and young adults in more than 10,500[2] schools. California aims to provide a world-class education for all students, from early childhood to adulthood. The CDE serves the state by innovating and collaborating with educators, school staff, parents/guardians, and community partners which together, as a team, prepares students to live, work, and thrive in a highly connected world.

Within the CDE, it is the Performance, Planning, and Technology Branch that oversees programs promoting innovation and improving student achievement. Programs include oversight of statewide assessments and the collection and reporting of educational data (CDE, 2017b).

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 entire CAST assessment development process. California educators participated in tasks that included defining the purpose and scope of

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

the assessment, assessment design, item development, and scoring the constructed-response items.

1.8.4. Contractors

1.8.3.1. ETS

The CDE and the SBE contract with ETS to develop, administer, and report the CAST, although for the 2016–17 administration, student results were not reported. As the prime contractor, ETS has the overall responsibility of working with the CDE to implement and maintain an effective assessment system and to coordinate the work of ETS with its subcontractors. Activities that ETS directly conducts include but are not limited to the following:

- Managing program activities
- Supporting and training counties, LEAs, and direct funded charter schools
- Providing tiered help desk support to LEAs
- Developing all CAST test items
- Constructing, producing, and controlling the quality of CAASPP test forms and related test materials
- Hosting and maintaining a Web site with resources for LEA CAASPP coordinators, CAASPP test site coordinators, and test administrators
- Developing, hosting, and providing support for the Test Operations Management System (TOMS)
- Processing student test assignments
- Producing and distributing student score reports
- Developing a summary score reporting Web site that can be viewed by the public
- Completing all psychometric procedures

1.8.3.2. American Institutes for Research (AIR)

ETS also monitors and manages the work of AIR, ETS's 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 test;
- Hosting and providing support for its TDS and the Online Reporting System, a component of the overall CAASPP Assessment Delivery System;
- Scoring machine-scorable items; and
- Providing Level 3 technology help desk support to LEAs.

1.9 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, which, for the CAST pilot, included but were not limited to the following:

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

TOMS receives student enrollment data and LEA/school hierarchy data from the California Longitudinal Pupil Achievement Data System (CALPADS) via a daily feed. CALPADS is "a longitudinal data system used to maintain individual-level data including student demographics, course data, discipline, assessments, staff assignments, and other data for state and federal reporting."[3] LEA staff involved in the administration of the CAST assessments, such as LEA coordinators, test site coordinators, test administrators, and test examiners are assigned varying levels of access to TOMS. For example, only an LEA coordinator has permission to set up the LEA's test administration window; a test administrator cannot download student reports. A description of user roles is more extensively explained in the *2016–17 Online Test Administration Manual* (CDE, 2017a).

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

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

1.9.3. Training Test

The publicly available training test is provided to prepare students for the summative assessment. These tests simulate the experience of the CAST Online Assessments. The training tests align to performance expectations, gauge student success on the operational test, or produce scores. Students may access them using a Web browser.

The purposes of the training test are to

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

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

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

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

For the CAST pilot, targeted efforts were made to hire qualified raters from existing CAASPP rater pools and California science teachers. The hired human raters were provided in-depth training and were certified before starting the scoring process. Human raters were organized under a scoring leader and were provided CAST scoring materials such as anchor sets, scoring rubrics, validity samples, qualifying sets, and condition codes for unscorable responses within the interface. The quality control processes for CR scoring are explained further in *Chapter 7 Quality Control*.

1.10 Overview of the Technical Report

This technical report addresses the characteristics of the CAST administered in spring 2017. It contains seven additional chapters as follows:

- Chapter 2 presents an overview of processes involved in the CAST pilot, including a description of item types developed, item development specifications, form assembly, pilot administration, participation, and psychometric analysis.
- Chapter 3 discusses the detailed procedures of item development, item review, and pilot test assembly for the 2016–17 administration. In particular, new item types and features that differ from traditional item types are described.
- Chapter 4 describes the details of administering the CAST pilot forms, as well as the procedures ETS followed to ensure test security.
- Chapter 5 summarizes the types of scoring approaches that are used for each type of item in the CAST pilot forms, including the process for building AI scoring models.
- Chapter 6 summarizes the results of the analyses on a sample of data from the CAST pilot administration, including classical item analyses, test completion analyses, response time analyses, interrater reliability analyses, and differential item functioning analyses.
- Chapter 7 highlights the quality control processes used at various stages of administration.
- Chapter 8 describes the development and administration of the survey questionnaires to test administrators and students and the results of analyses on their responses.

References

California Code of Regulations, Title 5, Education, Division 1, Chapter 2, Subchapter 3.75, Article 2, § 855. Retrieved from http://www.cde.ca.gov/re/lr/rr/caaspp.asp

California Department of Education. (2016, June). *State Board of Education responsibilities.* Retrieved from http://www.cde.ca.gov/be/ms/po/sberesponsibilities.asp

California Department of Education. (2017a). *CAASPP online test administration manual, 2015–16 administration.* Sacramento, CA: California Department of Education. Retrieved from http://www.caaspp.org/rsc/pdfs/CAASPP.online_tam.2016-17.pdf

California Department of Education. (2017b, August). *Organization.* Retrieved from http://www.cde.ca.gov/re/di/or/

Chapter 2: An Overview of California Science Test (CAST) Processes

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

2.1. Item Development

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

2.1.1. Design Guidelines

ETS content specialists referred to design patterns and task templates as part of the incipient Evidence-Centered Design (ECD) documentation created by ETS researchers and based on current educational research to properly frame the construct measured in each item. As such, all items developed and used in the 2016–17 CAST pilot administration are appropriate for the grade level and aligned with the California Next Generation Science Standards (CA NGSS).

The design patterns were developed to define and further unpack each of the eight science-focused science and engineering practices (SEPs) and the two engineering-focused SEPs and to identify characteristics of the practice. The SEP was used as an entry point for item development, both because it represents a fundamental difference between previous science standards and the CA NGSS, as well as because the practice is less familiar to item developers. During the development of the design patterns, it was determined that each SEP could be further unpacked into several subpractices and that each subpractice could include a set of associated focal knowledge, skills, and abilities.

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

2.1.2. Content Guidelines

Throughout the item writing process, ETS developers adhered to ETS's foundational guidelines for quality item writing. These guidelines formed the basis for training of item writers and the rigorous review process that is implemented for every item. Additionally, task models and the CA NGSS PEs were used to guide the writing of items for the pilot. Refer to subsection *3.2 Item Writing* for the guidelines of item writing, including the item development specifications.

ETS trained California science teachers to develop items for the CAST pilot during an item writing workshop in April 2016 (see subsections *3.3 Item Writer Training* and *3.4 Selection of Item Writers*). California science teachers were instructed to produce items that spanned a variety of SEPs and science domains (i.e., Life Sciences, Physical Sciences, and Earth and Space Sciences) to provide as wide an array of items as possible for the pilot forms construction.

2.1.3. Item Types Guidelines

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

The CAST pilot was designed to assess the CA NGSS using discrete items, single and multipoint items, and performance tasks. There were a variety of item types, including traditional multiple-choice, CR items, some familiar TE types, as well as some new TE types that utilized simulations and animations. Refer to subsection *3.2 Item Writing* for more details on item volumes developed; refer to subsection *3.5 Item Types and Features* for the types of items used in the CAST pilot.

2.2. Test Assembly

2.2.1. General Forms

Each grade level had two general pilot forms with 19–21 unique items per form with a shared performance task consisting of a stimulus and six items. See subsection *3.7 Test Assembly and Length* for more information on test assembly.

2.2.2. Forms with Accessibility Features

ETS developed one accessible form per grade using items in common with the general pilot forms. The accessibility features forms supported five features, which subsection *2.3.2 Accessibility Features* and subsection *3.7 Test Assembly and Length* describe in detail. Items were selected for use on the accessibility features pilot form based on item type, item content, and type of stimulus. Item types that relied heavily on visual input, such as drag-and-drop or hot spot, were not used on the accessibility features pilot form. Another key consideration was the degree to which visual stimuli could be described clearly. Note that the accessible forms did not include any performance tasks.

2.3. Test Administration

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

2.1.4. Test Security and Confidentiality

All tests within the CAASPP System are secure. For the CAST pilot, every person with access to test materials maintained the security and confidentiality of the tests. ETS's 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. In the pursuit of enforcing secure practices, ETS strives to safeguard the various processes

involved in a test development and administration cycle. The practices related to each of the following security processes are discussed in detail in subsection *4.6 Test Security and Confidentiality*.

2.1.5. Accessibility Features

ETS administered one unique accessibility features form per grade span during the CAST pilot. Each form consisted of seven to eight discrete items. The forms were administered to eligible students in grades five, eight, and a preselected cohort of high school students in grades ten, eleven, and twelve.

The purpose of the accessibility features form was to evaluate the functionality of the selected accessibility features so that all necessary refinements could be made to maximize the accessibility of the 2017–18 field test. The five accessibility features available during the CAST pilot were as follows:

1. American Sign Language (ASL)
2. Text-to-speech
3. Refreshable braille
4. Braille embosser
5. Print on demand

The accessibility features selected for these forms, with the exception of print on demand, were enabled via item-embedded features. This means that the necessary resource for the feature (e.g., alternate text descriptions for screen readers and refreshable braille and pronunciation cues for text-to-speech–based read aloud) were authored as part of each unique item. These resources work in conjunction with accessibility tools within the test delivery system (TDS) or by means of third-party assistive technology software or devices (e.g., screen-reading software such as Job Access with Speech, refreshable braille displays such as the ALVA USB 640 40-cell braille display from Vision Cue, and braille embossers such as the ViewPlus Tiger Max Embosser).

Based on 2014–15 Smarter Balanced participation data and 2015–16 administration data from the California Longitudinal Pupil Achievement Data System (CALPADS) and the Test Operations Management System (TOMS), approximately 180,000 students were expected to be eligible for the accessibility features forms on the CAST pilot. The support(s) provided to each individual student was determined 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. Selecting one of the five accessibility features in TOMS automatically assigned the student to the appropriate grade span accessibility features form at the time of testing.

2.1.6. Summary of Special Services

The CDE maintains a list of universal tools, designated supports, and accommodations that were permitted for use in the CAST pilot in its Web document "Matrix One: Universal Tools, Designated Supports, and Accommodations for the CAASPP System" (CDE, 2017). Part 1 includes the embedded universal tools, designated supports, and accommodations that are available for online testing. The five accessibility features listed in subsection *2.3.2 Accessibility Features* are the embedded designated supports and accommodations that were available for the CAST pilot. Parts 2 and 3 of Matrix One include the non-embedded universal tools, designated supports, and accommodations. School-level personnel, IEP teams, and Section 504 teams use Matrix One when deciding how best to support the student's test-taking experience. *Note that this technical report is based on the*

version of Matrix One that was available during the 2016–17 CAST pilot administration. Refer to subsection *2.3.2 Accessibility Features* for the list of accommodations that were available for the CAST pilot.

Appendix 2.A presents counts and percentages of students who were provided with designated supports and accommodations for grade five, grade eight, and high school tests. The majority of students did not use any designated supports, accommodations, or unlisted resources. Note that the tables in Appendix 2.A were created using student demographic data that are in version 2 of the production data file ("P2") that was updated on August 29, 2017.

2.4. Participation

All students enrolled in grades five and eight were required to participate in CAST pilot except for students with the most significant cognitive disabilities who meet the criteria for the California Alternate Assessment (CAA) for Science as indicated by the student's IEP team. For the CAST pilot, a single grade level (i.e., ten, eleven, or twelve) was assigned to each high school. All students in the assigned grade eligible for the general science assessments (i.e., not eligible for the CAA for Science) were required to participate. Details of participation are shown in subsection *4.4 Participation*.

2.5. Scoring

The CAST pilot contained traditional multiple-choice (MC) items, TE items, and CR items. The MC items and TE items were machine-scored through the TDS. The CR items were scored by trained raters. In addition, artificial intelligence (AI) scoring models were built using the CAST pilot data and will be implemented for future AI scoring. *Chapter 5 Scoring* provides details on scoring samples, machine scoring in the TDS, the human scoring process, and AI scoring models.

2.6. Psychometric Analyses

Some psychometric analyses were conducted on the data from the CAST pilot administration, including classical item analyses, test completion analyses, response time analyses, interrater reliability analyses, and differential item functioning (DIF) analyses. *Chapter 6 Analyses* provides details. The results of these analyses support the understanding of the item performances and internal structure and provide the validity evidence for both the response processes and scoring.

Appendix 2.A Summary of Special Services

Note: The tables in Appendix 2.A present counts and percentages of students who were provided with designated supports and accommodations for the grades five, grade eight, and high school California Science Tests.

Table 2.A.1 Special Services Summary for CAST Grade Five, Grade Eight, High School—All Tested

Accessibility Resource	Grade 5	Pct. of Total	Grade 8	Pct. of Total	High School	Pct. of Total
Embedded accommodation—American Sign Language	157	0.03	188	0.04	149	0.03
Embedded accommodation—Braille	16	0.00	20	0.00	18	0.00
Non-Embedded accommodation—Abacus	12	0.00	10	0.00	7	0.00
Non-Embedded accommodation—Calculator	405	0.09	3,742	0.83	2,534	0.55
Non-Embedded accommodation—Mathematics Tools	92	0.02	185	0.04	206	0.04
Non-Embedded accommodation—Multiplication Table	16,721	3.54	11,140	2.46	1,775	0.39
Non-Embedded accommodation—Print on Demand	372	0.08	294	0.06	171	0.04
Non-Embedded accommodation—Speech to Text	2,843	0.60	2,207	0.49	658	0.14
Non-Embedded accommodation—100s Number Table	2,451	0.52	589	0.13	135	0.03
Embedded designated support—Permissive Mode	615	0.13	495	0.11	170	0.04
Embedded designated support—Print Size	1,784	0.38	883	0.19	297	0.06
Embedded designated support—Text-to-Speech	36,611	7.75	17,815	3.93	7,623	1.66
Non-Embedded designated support—Color Overlay	54	0.01	54	0.01	12	0.00
Non-Embedded designated support—Magnification	240	0.05	201	0.04	74	0.02
Non-Embedded designated support—Noise Buffers	1,295	0.27	576	0.13	178	0.04
Non-Embedded designated support—Read Aloud	2,853	0.60	1,612	0.36	479	0.10
Non-Embedded designated support—Scribe	300	0.06	190	0.04	54	0.01

Accessibility Resource	Grade 5	Pct. of Total	Grade 8	Pct. of Total	High School	Pct. of Total
Non-Embedded designated support—Separate Setting	24,389	5.16	16,853	3.72	10,111	2.20
Non-Embedded designated support—Simplified Test Directions	5,756	1.22	2,512	0.55	5,148	1.12
Other—Designated support, or accommodation is in IEP	17,974	3.80	18,384	4.06	29,336	6.39
Other—Designated support, or accommodation is in Section 504 plan	4,767	1.01	7,011	1.55	8,382	1.83

Table 2.A.2 Special Services Summary for CAST Grade Five, Grade Eight, High School—Students with No Special Education Services

Accessibility Resource	Grade 5	Pct. of Total	Grade 8	Pct. of Total	High School	Pct. of Total
Embedded accommodation—American Sign Language	0	0.00	0	0.00	0	0.00
Embedded accommodation—Braille	0	0.00	0	0.00	0	0.00
Non-Embedded accommodation—Abacus	0	0.00	0	0.00	0	0.00
Non-Embedded accommodation—Calculator	7	0.00	56	0.01	36	0.01
Non-Embedded accommodation—Mathematics Tools	3	0.00	5	0.00	2	0.00
Non-Embedded accommodation—Multiplication Table	199	0.05	99	0.02	12	0.00
Non-Embedded accommodation—Print on Demand	3	0.00	6	0.00	2	0.00
Non-Embedded accommodation—Speech to Text	46	0.01	29	0.01	9	0.00
Non-Embedded accommodation—100s Number Table	35	0.01	5	0.00	8	0.00
Embedded designated support—Permissive Mode	182	0.04	206	0.05	43	0.01
Embedded designated support—Print Size	721	0.17	221	0.05	22	0.01
Embedded designated support—Text-to-Speech	16,785	4.02	5,602	1.38	4,664	1.12
Non-Embedded designated support—Color Overlay	24	0.01	6	0.00	4	0.00
Non-Embedded designated support—Magnification	123	0.03	58	0.01	5	0.00
Non-Embedded designated support—Noise Buffers	750	0.18	309	0.08	101	0.02
Non-Embedded designated support—Read Aloud	763	0.18	256	0.06	71	0.02
Non-Embedded designated support—Scribe	115	0.03	101	0.02	28	0.01
Non-Embedded designated support—Separate Setting	2,164	0.52	1,172	0.29	634	0.15
Non-Embedded designated support—Simplified Test Directions	3,669	0.88	977	0.24	4,194	1.01

Accessibility Resource	Grade 5	Pct. of Total	Grade 8	Pct. of Total	High School	Pct. of Total
Other—Designated support, or accommodation is in IEP	0	0.00	0	0.00	0	0.00
Other—Designated support, or accommodation is in Section 504 plan	4,523	1.08	6,693	1.65	8,041	1.93

Table 2.A.3 Special Services Summary for CAST Grade Five, Grade Eight, High School—Students with Special Education Services

Accessibility Resource	Grade 5	Pct. of Total	Grade 8	Pct. of Total	High School	Pct. of Total
Embedded accommodation—American Sign Language	157	0.28	188	0.40	149	0.35
Embedded accommodation—Braille	16	0.03	20	0.04	18	0.04
Non-Embedded accommodation—Abacus	12	0.02	10	0.02	7	0.02
Non-Embedded accommodation—Calculator	398	0.72	3,686	7.76	2,498	5.85
Non-Embedded accommodation—Mathematics Tools	89	0.16	180	0.38	204	0.48
Non-Embedded accommodation—Multiplication Table	16,522	29.93	11,041	23.25	1,763	4.13
Non-Embedded accommodation—Print on Demand	369	0.67	288	0.61	169	0.40
Non-Embedded accommodation—Speech to Text	2,797	5.07	2,178	4.59	649	1.52
Non-Embedded accommodation—100s Number Table	2,416	4.38	584	1.23	127	0.30
Embedded designated support—Permissive Mode	433	0.78	289	0.61	127	0.30
Embedded designated support—Print Size	1,063	1.93	662	1.39	275	0.64
Embedded designated support—Text-to-Speech	19,826	35.92	12,213	25.72	2,959	6.93
Non-Embedded designated support—Color Overlay	30	0.05	48	0.10	8	0.02
Non-Embedded designated support—Magnification	117	0.21	143	0.30	69	0.16
Non-Embedded designated support—Noise Buffers	545	0.99	267	0.56	77	0.18
Non-Embedded designated support—Read Aloud	2,090	3.79	1,356	2.86	408	0.96
Non-Embedded designated support—Scribe	185	0.34	89	0.19	26	0.06
Non-Embedded designated support—Separate Setting	22,225	40.26	15,681	33.03	9,477	22.21
Non-Embedded designated support—Simplified Test Directions	2,087	3.78	1,535	3.23	954	2.24

Accessibility Resource	Grade 5	Pct. of Total	Grade 8	Pct. of Total	High School	Pct. of Total
Other—Designated support, or accommodation is in IEP	17,974	32.56	18,384	38.72	29,336	68.74
Other—Designated support, or accommodation is in Section 504 plan	244	0.44	318	0.67	341	0.80

Table 2.A.4 Special Services Summary for CAST Grade Five, Grade Eight, High School—English Only Students

Accessibility Resource	Grade 5	Pct. of Total	Grade 8	Pct. of Total	High School	Pct. of Total
Embedded accommodation—American Sign Language	98	0.04	135	0.05	99	0.04
Embedded accommodation—Braille	7	0.00	9	0.00	10	0.00
Non-Embedded accommodation—Abacus	7	0.00	4	0.00	5	0.00
Non-Embedded accommodation—Calculator	235	0.09	2,335	0.95	1,558	0.64
Non-Embedded accommodation—Mathematics Tools	54	0.02	113	0.05	138	0.06
Non-Embedded accommodation—Multiplication Table	9,468	3.56	6,135	2.50	994	0.41
Non-Embedded accommodation—Print on Demand	213	0.08	153	0.06	118	0.05
Non-Embedded accommodation—Speech to Text	1,664	0.63	1,227	0.50	352	0.14
Non-Embedded accommodation—100s Number Table	1,254	0.47	291	0.12	79	0.03
Embedded designated support—Permissive Mode	368	0.14	269	0.11	122	0.05
Embedded designated support—Print Size	915	0.34	460	0.19	151	0.06
Embedded designated support—Text-to-Speech	17,102	6.43	8,484	3.45	3,248	1.33
Non-Embedded designated support—Color Overlay	26	0.01	44	0.02	7	0.00
Non-Embedded designated support—Magnification	122	0.05	105	0.04	44	0.02
Non-Embedded designated support—Noise Buffers	549	0.21	223	0.09	71	0.03
Non-Embedded designated support—Read Aloud	1,396	0.52	715	0.29	232	0.10
Non-Embedded designated support—Scribe	109	0.04	58	0.02	17	0.01
Non-Embedded designated support—Separate Setting	14,330	5.39	9,656	3.93	6,005	2.46
Non-Embedded designated support—Simplified Test Directions	2,305	0.87	942	0.38	2,030	0.83

Accessibility Resource	Grade 5	Pct. of Total	Grade 8	Pct. of Total	High School	Pct. of Total
Other—Designated support, or accommodation is in IEP	11,632	4.37	11,141	4.53	16,993	6.96
Other—Designated support, or accommodation is in Section 504 plan	3,910	1.47	5,774	2.35	6,878	2.82

Table 2.A.5 Special Services Summary for CAST Grade Five, Grade Eight, High School—Initially Fluent English Proficient Students

Accessibility Resource	Grade 5	Pct. of Total	Grade 8	Pct. of Total	High School	Pct. of Total
Embedded accommodation—American Sign Language	7	0.04	8	0.04	3	0.01
Embedded accommodation—Braille	0	0.00	2	0.01	1	0.00
Non-Embedded accommodation—Abacus	0	0.00	0	0.00	0	0.00
Non-Embedded accommodation—Calculator	0	0.00	37	0.17	42	0.17
Non-Embedded accommodation—Mathematics Tools	0	0.00	4	0.02	5	0.02
Non-Embedded accommodation—Multiplication Table	151	0.79	113	0.51	22	0.09
Non-Embedded accommodation—Print on Demand	5	0.03	4	0.02	1	0.00
Non-Embedded accommodation—Speech to Text	33	0.17	26	0.12	11	0.04
Non-Embedded accommodation—100s Number Table	16	0.08	9	0.04	4	0.02
Embedded designated support—Permissive Mode	10	0.05	9	0.04	0	0.00
Embedded designated support—Print Size	42	0.22	22	0.10	3	0.01
Embedded designated support—Text-to-Speech	481	2.53	259	1.16	277	1.12
Non-Embedded designated support—Color Overlay	0	0.00	0	0.00	2	0.01
Non-Embedded designated support—Magnification	6	0.03	6	0.03	1	0.00
Non-Embedded designated support—Noise Buffers	23	0.12	10	0.04	12	0.05
Non-Embedded designated support—Read Aloud	32	0.17	17	0.08	5	0.02
Non-Embedded designated support—Scribe	0	0.00	3	0.01	2	0.01
Non-Embedded designated support—Separate Setting	271	1.43	245	1.10	178	0.72
Non-Embedded designated support—Simplified Test Directions	83	0.44	30	0.13	251	1.01

Accessibility Resource	Grade 5	Pct. of Total	Grade 8	Pct. of Total	High School	Pct. of Total
Other—Designated support, or accommodation is in IEP	278	1.46	348	1.56	572	2.31
Other—Designated support, or accommodation is in Section 504 plan	131	0.69	198	0.89	251	1.01

Table 2.A.6 Special Services Summary for CAST Grade Five, Grade Eight, High School—English Learner Students

Accessibility Resource	Grade 5	Pct. of Total	Grade 8	Pct. of Total	High School	Pct. of Total
Embedded accommodation—American Sign Language	46	0.05	35	0.06	31	0.06
Embedded accommodation—Braille	6	0.01	6	0.01	2	0.00
Non-Embedded accommodation—Abacus	4	0.00	3	0.01	2	0.00
Non-Embedded accommodation—Calculator	158	0.16	1,030	1.89	662	1.33
Non-Embedded accommodation—Mathematics Tools	36	0.04	51	0.09	42	0.08
Non-Embedded accommodation—Multiplication Table	6,524	6.56	3,719	6.82	580	1.17
Non-Embedded accommodation—Print on Demand	136	0.14	112	0.21	34	0.07
Non-Embedded accommodation—Speech to Text	1,053	1.06	744	1.36	212	0.43
Non-Embedded accommodation—100s Number Table	1,092	1.10	205	0.38	37	0.07
Embedded designated support—Permissive Mode	165	0.17	99	0.18	32	0.06
Embedded designated support—Print Size	661	0.67	243	0.45	99	0.20
Embedded designated support—Text-to-Speech	15,574	15.67	6,273	11.50	2,294	4.62
Non-Embedded designated support—Color Overlay	25	0.03	9	0.02	2	0.00
Non-Embedded designated support—Magnification	78	0.08	54	0.10	20	0.04
Non-Embedded designated support—Noise Buffers	479	0.48	178	0.33	39	0.08
Non-Embedded designated support—Read Aloud	1,218	1.23	711	1.30	174	0.35
Non-Embedded designated support—Scribe	182	0.18	110	0.20	32	0.06
Non-Embedded designated support—Separate Setting	8,794	8.85	4,970	9.11	2,714	5.46
Non-Embedded designated support—Simplified Test Directions	2,665	2.68	1,255	2.30	1,290	2.60

Accessibility Resource	Grade 5	Pct. of Total	Grade 8	Pct. of Total	High School	Pct. of Total
Other—Designated support, or accommodation is in IEP	4,659	4.69	4,223	7.74	7,010	14.11
Other—Designated support, or accommodation is in Section 504 plan	417	0.42	308	0.56	296	0.60

Table 2.A.7 Special Services Summary for CAST Grade Five, Grade Eight, High School—Reclassified Fluent English Proficient Students

Accessibility Resource	Grade 5	Pct. of Total	Grade 8	Pct. of Total	High School	Pct. of Total
Embedded accommodation—American Sign Language	6	0.01	10	0.01	16	0.01
Embedded accommodation—Braille	3	0.00	3	0.00	5	0.00
Non-Embedded accommodation—Abacus	1	0.00	3	0.00	0	0.00
Non-Embedded accommodation—Calculator	12	0.01	340	0.26	272	0.19
Non-Embedded accommodation—Mathematics Tools	2	0.00	17	0.01	21	0.02
Non-Embedded accommodation—Multiplication Table	573	0.66	1,167	0.90	179	0.13
Non-Embedded accommodation—Print on Demand	18	0.02	25	0.02	18	0.01
Non-Embedded accommodation—Speech to Text	90	0.10	209	0.16	83	0.06
Non-Embedded accommodation—100s Number Table	86	0.10	84	0.06	15	0.01
Embedded designated support—Permissive Mode	71	0.08	116	0.09	15	0.01
Embedded designated support—Print Size	165	0.19	158	0.12	44	0.03
Embedded designated support—Text-to-Speech	3,367	3.86	2,765	2.13	1,787	1.28
Non-Embedded designated support—Color Overlay	3	0.00	1	0.00	1	0.00
Non-Embedded designated support—Magnification	34	0.04	36	0.03	9	0.01
Non-Embedded designated support—Noise Buffers	241	0.28	164	0.13	55	0.04
Non-Embedded designated support—Read Aloud	206	0.24	168	0.13	68	0.05
Non-Embedded designated support—Scribe	9	0.01	19	0.01	3	0.00
Non-Embedded designated support—Separate Setting	963	1.10	1,970	1.52	1,205	0.86
Non-Embedded designated support—Simplified Test Directions	683	0.78	268	0.21	1,565	1.12
Other—Designated support, or accommodation is in IEP	1,396	1.60	2,656	2.05	4,741	3.39
Other—Designated support, or accommodation is in Section 504 plan	309	0.35	729	0.56	952	0.68

Chapter 3: Item Development and Assembly

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

3.1. Use of Evidence-Centered Design (ECD)

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

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

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

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

During the drafting stage, ETS further specified approaches to the task templates designed to engage students meaningfully with the SEP by specifying item characteristics, work products, and observations that can be made about student proficiency from those work products. This documentation was used during both item development and revision to ensure that the student responses elicited by the items validly reflected the integrated science understanding specified in the targeted PEs. In addition to task models and task templates, a draft blueprint was developed to identify the specific claims and quantify the number of items necessary to collect the evidence to support the claims.

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 pilot 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.

Similarly, the definition of claims for the CAST is an ongoing and iterative process, one informed both by data collected from the CAST pilot and the future data collection from the field test administration in 2017–18. Comprehensive documentation of this process is being captured in an in-progress white paper titled “Use of Evidence-Centered Design in CAST Item and Test Development.”

3.2. Item Writing

The initial item development plan for the CAST pilot focused on developing items that integrated at least two of the three dimensions of the CA NGSS—disciplinary core ideas (DCIs), SEPs, and crosscutting concepts (CCCs). The plan incorporated a diverse selection of PEs to incorporate a range of SEPs, DCIs, and CCCs.

Table 3.1 shows the total number of items developed per grade to accommodate the pilot and training tests, as described in subsection *4.3 Training Test*.

Table 3.1 Total Number of Items Developed per Grade for the CAST Pilot

Item Type	Grade 5	Grade 8	High School
Standard discrete item types (non-CR)	30	30	30
Discrete CR	4	4	4
Custom discrete interactive items	4	4	4
Performance tasks (three tasks, six of each task)	18	18	18
TOTAL	**56**	**56**	**56**

Constructed response (CR) items included text entry and extended text item types are shown in Table 3.2. Discrete items included traditional multiple-choice (MC) items, CR items, and some familiar technology-enhanced (TE) item types (e.g., match, inline choice list, zone or hot spot, etc.), as well as some new TE item types that utilized simulations and animations, which are also indicated as custom discrete interactive items. The performance task, which contained six items for the CAST pilot, is designed to provide students with an opportunity to demonstrate their ability to apply knowledge and higher-order thinking skills to explore and analyze a complex, real-world scenario.

Table 3.2 Selected Item Types in the CAST Pilot

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

Feature	Description
Text Entry	Items that require a constrained response from a student using the keyboard (e.g., the numerator and denominator in a fraction, a blank within a longer sentence)
Inline Choice	Items that provide multiple choices for filling in one or more blanks within a sentence or paragraph
Associate	Items where a student creates pairings among items on a list (e.g., student identifies multiple chemical elements with common chemical properties)
Hot Text	Items where a student selects text (e.g., within a paragraph) as the response
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)

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

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

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

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

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

- Subpractice
- Subpractice Assessment Targets

- DCI Assessment Targets
- CCC Assessment Targets
- Possible Phenomena or Contexts
- Examples of Integration of Assessment Targets and Evidence
- Common Misconceptions
- Additional Assessment Boundaries

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

3.3. 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 pilot, 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:

1. to 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. to ensure that teachers who successfully completed the training were ready to develop high-quality items for the CAST pilot; and
3. to leverage the experiences, perspectives, and expertise of the teachers in writing items for the CAST pilot.

ETS held an item writer–training workshop in April 2016 in Sacramento, California, to provide prospective item writers with professional development in several areas. A review of the general assessment development process gave trainees a sense of the total lifecycle of an item. The dimensions of the CA NGSS (i.e., DCI, CCC, and SEP) were analyzed and explored to focus on the three dimensions of the CA NGSS that items for the CAST pilot 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 e-mail and phone calls to trained item writers on further item samples and ideas submitted ahead of contractual item submissions.

3.4. Selection of Item Writers

Senior ETS content staff screened applications for item writers for the CAST pilot, and ETS approved only those with strong content and teaching backgrounds for the item writing

training program. ETS selected item writers after the training, but not all recipients of the training became an item writer.

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

- Possession of a bachelor's degree in science or in the field of education with special focus on a particular scientific domain; an advanced degree in the relevant content was desirable
- Had 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
- Had previous experience or training in writing items in the grades and content areas covered by the CAST pilot
- Had familiarity and understanding of the CA NGSS

3.5. Item Types and Features

Every item assessed a CA NGSS DCI as well as at least one of the other two CA NGSS dimensions (i.e., SEP or 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, 2016b).

ETS used item types, individually and/or 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/or 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 2016–17 pilot that were supported by *Instructional Management Systems (IMS) Global Question and Test Interoperability (QTI)* standards (IMS, 2016).

Table 3.2 on page 27 outlines the major categories of QTI item types that were included in the CAST pilot. This includes item types ranging from traditional MC and CR (i.e., extended text or text entry) to new TE types (the rest of the item types).

3.6. Item Review Process

ETS placed items developed for the CAST pilot through an extensive internal item review process. This section summarizes the item review process that confirmed the quality of CAST pilot 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 measured 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 Committees (IRCs), which are described in more detail in subsection *3.6.5 Content Expert Reviews*.

The ETS review process for the CAST included the following:

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

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

3.6.1. ETS Content Review

ETS content-area assessment specialists conducted three reviews on items and stimuli. These assessment specialists verified that the items and stimuli were in compliance with ETS's written guidelines for clarity, style, accuracy, and appropriateness for California students as well as in compliance with the task models. Assessment specialists reviewed each item in terms of 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 (DOK)
- 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.

3.6.2. ETS Research Review

Internal science researchers, who also contributed to the ECD documentation, reviewed items with a focus on the alignment issues at the item level and provided potential refinement solutions to improving 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 iterative development process of items.

3.6.3. 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 IRCs. 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.

3.6.4. ETS Fairness Review

ETS assessment specialists who were specially trained to identify and eliminate questions that contained content or wording that could be construed to be offensive to or biased

against members of specific ethnic, racial, or gender groups conducted the next level of review (ETS, 2014, 2016). These trained staff members reviewed every item before the CDE and IRC reviews.

The review process promoted 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-language learners

3.6.5. Content Expert Reviews

3.6.5.1. IRCs

In addition to the ETS internal content reviews, items went through the content expert review by IRCs before being placed on a pilot assessment. The IRCs are advisory panels to the CDE and ETS who provided guidance on matters related to item development for the CAST pilot. The IRCs were responsible for reviewing all newly developed items for alignment to the CA NGSS. The IRCs also reviewed the items for accuracy of content, clarity of phrasing, and overall quality. In their examination of test items, committee members could have raised concerns related to age/grade appropriateness and gender, racial, ethnic, and/or socioeconomic bias.

3.6.5.2. Composition of IRCs

The IRCs for the pilot items were comprised of current and former teachers, resource specialists, administrators, curricular experts, and other education professionals. Members had to have the following minimum qualifications to serve on the IRCs for the CAST pilot:

- Three or more years of general teaching experience in grades kindergarten through twelve and in science
- Bachelor's or higher degree in science or education
- Knowledge of and experience with the CA NGSS

School administrators, local educational agency (LEA)/county content/program specialists, or university educators who met the following qualifications could serve on IRCs for the CAST pilot:

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

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

3.6.5.3. IRC Meetings for Review of CAST Pilot Items

ETS content-area assessment specialists facilitated CAST pilot IRC 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 pilot
- Overview of the CAST pilot test design specifications
- Overview of criteria for evaluating test items
- Review and evaluation of items for fairness issues

The criteria for reviewing items included the following:

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

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

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

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

3.7. Test Assembly and Length

ETS designed the general pilot forms to be taken in approximately one hour. ETS used historical timing data from previous assessments that had the same items types to estimate the amount of time needed to complete MC, CR, and TE item types.

Each form covered a broad range of PEs; these are shown in the sequence in which they were assessed in Table 3.3, Table 3.4, and Table 3.5 for grade five, grade eight, and high school respectively.

In these tables, the superscript "[1]" indicates items for PEs common between the accessible form and general form 1. The superscript "[2]" indicates items for PEs common between the accessible form and general form 2. Asterisks indicate items common between both general forms that were assessed with a performance task.

Table 3.3 Performance Expectations Assessed on CAST Pilot—Grade Five

Accessible	General: Form1	General: Form 2
5-ESS1-1[1]	5-ESS1-1[1]	5-LS1-1
4-LS1-1[1]	4-LS1-1[1]	4-LS1-1
5-PS2-1[1]	5-PS2-1[1]	5-PS1-1
5-PS3-1[2]	5-PS2-1	5-PS3-1[2]
5-ESS1-1[2]	5-PS1-4	5-ESS1-1[2]
5-PS1-2[2]	5-PS1-1	5-PS1-2[2]
5-ESS1-2[2]	5-PS1-1	5-ESS1-2[2]
	5-ESS2-1	5-PS1-2
	5-PS3-1	5-ESS2-1
	5-ESS2-1	5-PS3-1
	5-PS1-2	4-ESS3-1
	5-LS2-1	5-PS1-4
	4-PS4-2	5-PS1-3*
	5-PS1-3*	5-PS1-3*
	5-PS1-3*	5-PS1-3*
	5-PS1-3*	5-PS1-4*
	5-PS1-4*	5-PS1-4*
	5-PS1-4*	5-PS1-2*
	5-PS1-2*	

Table 3.4 Performance Expectations Assessed on CAST Pilot—Grade Eight

Accessible	General: Form 1	General: Form 2
MS-LS2-3[1]	MS-LS2-3[1]	MS-ESS2-3
MS-ESS3-5[1]	MS-ESS3-5[1]	MS-LS3-1
MS-PS3-2[1]	MS-PS3-2[1]	MS-PS3-1
MS-PS4-1[2]	MS-PS1-1	MS-PS4-1[2]
MS-PS3-4[2]	MS-ESS2-6	MS-PS3-4[2]
MS-ESS2-4[2]	MS-PS2-1	MS-ESS2-4[2]
MS-PS1-4[2]	MS-LS2-4	MS-PS1-4[2]

Accessible	General: Form 1	General: Form 2
	MS-PS4-2	MS-LS4-2
	MS-ESS2-4	MS-ESS2-6
	MS-PS1-2	MS-ESS2-6
	MS-PS3-1	MS-PS1-2
	MS-PS2-1	MS-PS3-2
	MS-PS4-1	MS-PS2-2
	MS-PS2-5	MS-ESS2-1
	MS-LS4-1	MS-LS4-6
	MS-ESS2-1*	MS-PS2-1
	MS-ESS2-1*	MS-ESS2-1*
	MS-ESS2-1*	MS-ESS2-1*
	MS-ESS2-3*	MS-ESS2-1*
	MS-ESS2-2*	MS-ESS2-3*
	MS-ESS2-2*	MS-ESS2-2*
		MS-ESS2-2*

Table 3.5 Performance Expectations Assessed on CAST Pilot—High School

Accessible	General: Form 1	General: Form 2
HS-LS3-3[1]	HS-LS3-3[1]	HS-PS4-2
HS-PS4-1[1]	HS-PS4-1[1]	HS-PS1-8
HS-LS2-3[1]	HS-LS2-3[1]	HS-LS3-2
HS-PS1-2[1]	HS-PS1-2[1]	HS-PS1-7
HS-ESS3-5[2]	HS-LS1-3	HS-ESS3-5[2]
HS-ESS1-3[2]	HS-ESS2-3	HS-ESS1-3[2]
HS-LS1-1[2]	HS-PS2-2	HS-LS1-1[2]
HS-ESS1-5[2]	HS-PS2-6	HS-ESS1-5[2]
	HS-LS4-1	HS-PS1-4
	HS-ESS2-6	HS-LS2-5
	HS-LS2-1	HS-ESS2-5
	HS-PS2-1	HS-PS1-8
	HS-ESS1-2	HS-LS1-3
	HS-PS2-2	HS-PS2-4
	HS-PS2-4	HS-PS1-8
	HS-ESS1-5*	HS-ESS1-5*
	HS-ESS1-5*	HS-ESS1-5*
	HS-ESS1-5*	HS-ESS1-5*
	HS-ESS2-1*	HS-ESS2-1*
	HS-ESS2-1*	HS-ESS2-1*
	HS-ESS2-3*	HS-ESS2-3*

3.8. Data Review

ETS conducted data review meetings with California teachers and the CDE after the pilot data analysis was complete. Reviewers examined items that were flagged for item difficulty, item-total correlation, item response distribution, and/or differential item functioning according to predefined criteria. The ETS facilitator led discussions about each flagged item and reviewed the content of the item to reach consensus on whether items should be accepted as is, accepted with revision, or rejected.

For items that were accepted with revision, California teachers participated in making suggested edits. As time allowed, ETS showed the statistics for items that were not flagged to determine if there were any edits that the stakeholders felt should be made prior to field testing the items. Refer to Table 3.6 for the results of data review, showing the number of items that were accepted with and without edits and the number of items that were rejected outright.

Table 3.6 Data Review Results

Grade	Accept As Is	Accept with Edits	Reject	Total Items
5	5	23	3	31
8	13	24	1	38
High School	7	28	1	36

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.

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

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

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

4.1. Assignment Design

The CAST pilot was administered to students in grades five, eight, and high school, including grades ten, eleven, and twelve. For grades five and eight, a census pilot was conducted. That is, all students in California enrolled in grades five and eight who were eligible for the general science assessments (i.e., not eligible for the California Alternate Assessment [CAA] for Science) were required to participate in the pilot. With a consideration to balance the administration efforts among high schools, schools with grades ten, eleven, and twelve had one of these grades assigned to participate in the CAST pilot. Subsection *4.1.1 Grade Assignment for High School Grades Ten to Twelve* outlines the process for grade assignment in the CAST pilot for high school students. For all tested students across grade levels, those who are eligible for certain types of designated supports or accommodations were assigned the accessibility form appropriate for their grade level (see subsection *4.2 Accessibility Features Pilot* for the details on eligibility). All other students were randomly assigned one of the two general forms that were available for their grade level.

4.1.1. Grade Assignment for High School Grades Ten to Twelve

The CAST pilot for high school was administered to students in grades ten, eleven, or twelve. However, to balance the administration efforts for schools participating in the CAST pilot and to facilitate the computation of the test participation rate for federal accountability, only a single grade level (i.e., ten, eleven, or twelve) was assigned to each high school in the state. That is, all California high schools participated but they each tested students in a single grade level (i.e., ten, eleven, or twelve). All students in the assigned grade eligible for the general science assessments (i.e., not eligible for the CAA for Science) were required to participate.

Concerning the assignment for the high school grade levels, the California Department of Education (CDE) provided the following guidelines for prioritization for selecting a high school grade for CAST administration:

1. Grade ten, and then
2. A grade level that allowed for administering both the CAST and CAA for Science, and then
3. Schools with a grade twelve population larger than the grade ten and grade eleven populations.

Eligible schools were first determined using the California Longitudinal Pupil Achievement Data System (CALPADS) data set available on October 19, 2016, which reflected locally entered enrollment information on students up to that date. Any school that had at least one student enrolled in grades ten, eleven, and twelve was considered eligible for a high school

grade-level assignment. Schools that entered information in CALPADS between October 19 and December 1, 2016, were then added to the assignment plan.

After assigning schools to meet the prioritizing criteria, CDE reviewed the assignments and, in a few instances, made changes to the assigned grade level to accommodate extenuating circumstances of the school. The high school assignments were then posted on the California Assessment of Student Performance and Progress (CAASPP) Web site at http://www.caaspp.org/ the first week of December 2016. Some schools serving grades ten, eleven, and/or twelve had not entered their enrollment information by December 1, 2016. Such schools were identified by March 16, 2017, just prior to the opening of the testing window on March 20, 2017, and assigned grade eleven, given that that grade had the lowest representation in the assignment plan. However, in cases in which the school did not serve any grade eleven students, it was assigned to the grade level in which it did have available students to test. Table 4.1 provides the final composition of the high school assignments by school and student.

Table 4.1 Grade Level Assignments for the CAST High School Pilot

Group	Grade 10	Grade 11	Grade 12	Total
Number of Schools	2,102	822	1,342	4,266
Percent of Schools	49%	19%	31%	100%

Note: Total does not equal 100 percent due to rounding.

4.2. Accessibility Features Pilot

The goal of the accessibility features pilot was to ensure that every student would have a positive and productive testing experience. A valid assessment that provides an accurate measurement of a student's academic achievement is one of multiple sources of information that can be used to improve student learning. In order to ensure that all students, including English learners and students with disabilities, have a fair opportunity to demonstrate what they know and are able to do, it is necessary that assessments be designed to be inclusive for a diverse student population.

The accessibility features pilot form provided information about the way the California Next Generation Science Standards (CA NGSS) content could be portrayed using the accessibility features selected for the pilot.

4.2.1 Accessibility Resources

Accessibility resources were tested in separate pilot forms at each grade level to evaluate the functionality of the following embedded designated supports and accommodations selected for the CAST pilot. These were as follows:

- Designated supports:
  - Permissive mode settings
  - Print size
  - Text-to-speech (items and stimuli)
- Accommodations:
  - American Sign Language
  - Braille (refreshable or via embosser)

Students were routed to the accessibility features pilot if their local educational agency (LEA) CAASPP coordinator assigned one of the embedded CAST test settings on their behalf in the Test Operations Management System (TOMS). Students participating in the accessibility features pilot could not take the items in the primary item pilot assessment (i.e., general form 1 or 2).

4.2.2 Guidelines for Accessibility and Accommodations

Educational Testing Service (ETS) and the CDE collaborated in establishing guidelines for assessment development for the CAST pilot that followed an approach consistent with the Smarter Balanced Assessment Consortium's *Accessibility and Accommodations Framework* (Smarter Balanced, 2014). In addition, the principles of universal design were applied in the development of the CAST pilot items to make them accessible for student accessibility needs.

4.2.3 Construction of Accessibility Forms

One accessibility form was delivered per grade span; the number of discrete items per form were as follows:

- Grade five—Seven items
- Grade eight—Seven items
- High school—Eight items

The accessibility forms were presented through the same test delivery system as the general forms. Refer to subsection *3.7 Test Assembly and Length* for additional information about item selection for accessibility forms, and Table 6.A.1 through Table 6.A.13 for the average form difficulty.

4.3. Training Test

The training test was designed to provide students with an opportunity to engage with CA NGSS–aligned items, including technology-enhanced (TE) items. It also allowed students to familiarize themselves with the test settings, including universal tools, available for the pilot. A *Training Items Scoring Guide* was available for test administrators to offer details about the items, student response types, correct responses, and related scoring considerations for the included sample of training items (CDE, 2017b). In addition, the training test allowed educators to familiarize themselves with the organization of the CAST pilot and help maintain the standardization of test administration.

A single training test was released that contained content from each grade level (i.e., grades five and eight and high school), and consisted of 17 items in total as well as instructional items that informed the test taker at what grade level the subsequent items would be. This training test was available through the Practice and Training Test Portal at http://www.caaspp.org/practice-and-training/index.html through the test delivery system (CDE, 2017c). The training test and the scoring guide for the training test were available to anyone with Internet access.

4.4. Participation

All students enrolled in grades five and eight were required to participate in the CAST pilot except for students with the most significant cognitive disabilities who met the criteria for the CAA for Science (approximately one percent or fewer of the student population) as indicated by the student's individualized education program team.

The CAST pilot for high school students was intended for students in grades ten, eleven, or twelve. Subsection *4.1.1 Grade Assignment for High School Grades Ten to Twelve* outlines the process for assignment in the CAST pilot for high school students. Table 4.2 provides the composition of the test-taker population for the CAST pilot for high school students.

Table 4.2 Composition of Test-taker Population for the CAST Pilot for High School Students

Group	Grade 10	Grade 11	Grade 12	Total
Number of Schools	1,562	406	678	2,646
Percent of Schools	59%	15%	26%	100%
Number of Students	371,949	43,282	43,988	459,219
Percent of Students	81%	9%	10%	100%

There were 1,620 schools that were assigned but did not participate in the pilot. The vast majority of these schools—1,511, or 93 percent—had 10 or fewer students enrolled in grade ten, eleven, or twelve at the time of assignment. The remaining schools may have only had students eligible for the CAA for Science or chose not to participate in the CAST pilot for another reason.

Table 4.3 presents the participation rates of each test and also participation rates for each grade for the high school test. Appendix 4.A shows the participation rates of selected demographic student groups in each test. The demographic student groups included gender, ethnicity, English language fluency, economic status (disadvantaged or not), special education services status, and migrant status, the list of which is shown in Table 4.4.

Table 4.3 CAST Pilot Participation Rates of the Full Population

Group	Grade 5	Grade 8	HS—Grade 10	HS—Grade 11	HS—Grade 12	HS—All Grades
Number of Enrolled Students	484,701	470,237	397,736	48,230	71,572	517,538
Number of Participants*	472,608	452,936	371,949	43,282	43,988	459,219
Percent of Participation	97.51	96.32	93.52	89.74	61.46	88.73

* “Participants” are enrolled students who log on to the test.

4.5. Demographic Summaries

The number and the percent of students for selected groups are provided starting in Table 4.B.1 through Table 4.B.3 for each grade-level test, and in Table 4.B.4 through Table 4.B.6 for each high school grade (i.e., grades ten, eleven, and twelve). In the tables, students are grouped by demographic characteristics, including gender, ethnicity, English language fluency, economic status (disadvantaged or not), special education services status, migrant status, and ethnicity by economic status. Demographic student groups are shown in Table 4.4.

Percents of student groups for the population, along with the percents of student groups for the participants, are presented for Table 4.B.3 through Table 4.B.6, because high schools with grades ten, eleven, and twelve were sampled to the CAST pilot based on an assignment plan (refer to subsection *4.1 Assignment Design*). However, percents of student groups for the population were not presented for Table 4.B.1 and Table 4.B.2 because all

students in California enrolled in grades five and eight who were eligible for the general science assessments were required to participate in the pilot.

Table 4.4 Demographic Student Groups to Be Reported

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

4.6. Test Security and Confidentiality

All tests within the CAASPP System are secure. For the CAST pilot, every person who had access to test materials maintained the security and confidentiality of the tests. ETS's internal Code of Ethics requires that all test information, including tangible materials associated with the CAST pilot test (e.g., test questions and test results), confidential files, processes, and activities be kept secure. To ensure security for all tests that ETS develops or handles, ETS maintains an Office of Testing Integrity (OTI).

In the pursuit of enforcing secure practices, ETS strives to safeguard the various processes involved in a test development and administration cycle. Those processes are as follows:

- Test delivery
- Security of electronic files using a firewall
- Transfer of scores via secure data exchange
- Data management
- Statistical analysis
- Student confidentiality

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

For the CAST pilot, every person who worked with the assessments, communicated test results, and/or received testing information was responsible for maintaining the security and confidentiality of the tests, including CDE staff, ETS staff, ETS subcontractors, LEA assessment coordinators, school assessment coordinators, students, parents, teachers, and cooperative educational service agency staff. ETS's 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 configurations of secure servers) be kept secure. ETS has systems in place that maintained tight security for test items and test results, as well as for student data.

4.6.1 Office of Testing Integrity (OTI)

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

The OTI's mission is to

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

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

4.6.2 Procedures to Maintain Standardization of Test Security

Test security requires the accounting of all secure materials—including online summative test items and student data—before, during, and after each test administration. For the CAST pilot, 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 items and/or passages generated by the print-on-demand feature of the test delivery system (CDE, 2017a).

The following measures ensured the security of CAASPP System assessments administered in 2016–17:

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

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

4.6.3 Security of Electronic Files Using a Firewall

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

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

4.6.4 Transfer of Scores via Secure Data Exchange

Due to the confidential nature of test results, ETS currently uses secure file transfer protocol (SFTP) and encryption for all data file transfers, including student data files. 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.

ETS enters information about the files posted to the SFTP server in a Web form on a SharePoint Web site; 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.

Data are always transmitted to the SFTP server in an encrypted format; test data are never sent via e-mail. The SFTP server is used as a conduit for the transfer of files; secure test data are only temporarily stored on the shared SFTP server.

4.6.5 Data Management

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 these data are collected during operational testing. Only individuals with the appropriate credentials can access these 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.

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

All stored test content and student data are 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).

4.6.6 Statistical Analysis

During CAASPP testing, the information technology staff at ETS retrieves data files from the American Institutes for Research (AIR) and loads them into a database. The ETS Data Quality Services staff extracts the data from the database and performs quality control procedures (e.g., the values of all variables are as expected) before passing files to the ETS statistical analysis group (refer to subsection *7.5 Quality Control of Psychometric Processes* for data validation processes undertaken by ETS Data Quality Services). The statistical analysis staff stores the files on secure servers. All staff involved with the data adheres to the ETS Code of Ethics and the ETS Information Protection Policies to prevent any unauthorized access to data.

4.6.7 Student Confidentiality

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

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

Test security incidents, such as improprieties, irregularities, and breaches, are prohibited behaviors that give a student an unfair advantage or compromise the secure administration of the tests, which, in turn, compromise the reliability and validity of test results (CDE, 2016c). Whether intentional or unintentional, failure by staff or students to comply with security rules constitutes a test security incident. Test security incidents have impacts on scoring and affect students' performance on the test.

For the CAST pilot, LEA CAASPP coordinators and CAASPP test site coordinators verified that all test security and summative administration incidents were documented by filling out the secure *CAASPP STAIRS* form for reporting; this form contained selectable options to guide coordinators in their submittal. 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 e-mail in response to the *CAASPP STAIRS* form submittal.

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

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

The CDE reviewed appeals requests. Appeals could not be requested without a STAIRS case number (CDE, 2017d).

Table 4.5 describes types of appeals available during the 2016–17 CAASPP administration.

Table 4.5 Types of Appeals

Type of Appeal	Description
Reset	Resetting a student's summative test removed that test from the system and enabled the student to start a new test from the beginning.
Invalidation	Invalidated summative tests were scored and reported with a note that an irregularity occurred. (Note that for the 2016–17 administration, results of the CAST pilot were not reported.) The student(s) were counted as participating in the calculation of the school's participation rate for federal accountability purposes.
Re-open	Reopening a summative test allowed a student to access a test that was already submitted or has expired.
Restore	Restoring a summative test returned a test from the Reset status to its prior status. This action could only be performed on tests that were reset.
Grace Period Extension	Permitting a grace period extension allowed the student to review previously answered questions upon logging back on to the assessment after expiration of the pause rule.

4.6.8.1. Impropriety

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

4.6.8.2. Irregularity

A testing irregularity is an unusual circumstance that impacts an individual or a group of students who are testing and may potentially affect student performance on the test or impact test security or test validity. For the CAST pilot, 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 and was directed to submit an appeal if these circumstances could not be corrected and contained at the local level.

4.6.8.3. Breach

A testing breach is an event that poses a threat to the validity of the test and require immediate attention and escalation to the CDE via telephone. For the CAST pilot, following the call, the CAASPP test site coordinator or LEA CAASPP coordinator completed the online *CAASPP STAIRS* form within 24 hours. Examples may have included such situations as a release of secure materials or a security/system risk because 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 bank. A breach incident should have been reported to the LEA CAASPP coordinator immediately.

4.6.8.4. Appeals

For incidents that result in a need to reset, re-open, invalidate, or restore individual online student assessments, the CDE must approve the request. In most instances, an appeal will be submitted to address a test security breach or irregularity. For the CAST pilot, the LEA CAASPP coordinator or CAASPP test site coordinator may have submitted appeals in TOMS. All submitted appeals were 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. An appeal could be requested only by the LEA CAASPP coordinator or CAASPP test site coordinator if directed in the e-mail response to the *CAASPP STAIRS* form (CDE, 2017d). Table 4.5 describes the types of appeals available during the 2016–17 CAASPP administration.

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. (2017c). *CAASPP Smarter Balanced online test administration manual, 2016–17 administration.* Sacramento, CA: California Department of Education. Retrieved from http://www.caaspp.org/rsc/pdfs/CAASPP.online_tam.2016-17.pdf

California Department of Education. (2017a). *2016–17 California Science Test training items scoring guide.* Sacramento, CA: California Department of Education. Retrieved from http://www.caaspp.org/rsc/resources/CAST.training-scoring-guide.2016-17.pdf

California Department of Education. (2017b). CAST Grades 5, 8 and HS Training Test. Retrieved from http://capt.tds.airast.org/student

California Department of Education. (2017d). *Security incidents and appeals procedure guide, 2016–17 administration.* Sacramento, CA: California Department of Education. Retrieved from http://www.caaspp.org/rsc/pdfs/CAASPP.stairs-appeals-guide.2016-17.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, Measured Progress, & National Center on Educational Outcomes. (2014, January 26). *Smarter Balanced Assessment Consortium: Accessibility and accommodations framework*. Retrieved from https://portal.smarterbalanced.org/library/en/accessibility-and-accommodations-framework.pdf

Appendix 4.A Participation Rates

Notes:

- This set of tables show the percent of participants of selected demographic student groups in each test.
- The total numbers of registered students are derived from the version 2 of the production data file ("P2") that were delivered on August 29, 2017.
- A student is considered a participant if he or she was enrolled during the active testing window and logged on to the test and the student survey.
- High school grades are ten, eleven, and twelve.

Table 4.A.1 CAST Pilot Participation Rates for Grade Five—Gender

Group	All	Male	Female
Number of students	484,701	247,725	236,976
Number of participants	472,608	241,178	231,430
Percent of participation	97.51	97.36	97.66

Table 4.A.2 CAST Pilot Participation Rates for Grade Five—English Language Fluency

Group	All	English Learner	English Only	Reclassified Fluent English Proficient	Initially Fluent English Proficient	To Be Determined	No Response
Number of students	484,701	101,765	274,290	88,052	19,275	380	939
Number of participants	472,608	99,376	266,056	87,213	19,004	280	679
Percent of participation	97.51	97.65	97.00	99.05	98.59	73.68	72.31

Table 4.A.3 CAST Pilot Participation Rates for Grade Five—Economic Status

Group	Economically Disadvantaged	Not Economically Disadvantaged
Number of students	299,216	185,485
Number of participants	292,601	180,007
Percent of participation	97.79	97.05

Table 4.A.4 CAST Pilot Participation Rates for Grade Five—Ethnicity

Group	American Indian/ Alaskan Native	Asian	Pacific Islander	Filipino	Hispanic or Latino	African American	White	Two or More Races	Unknown
Number of students	2,498	43,351	2,194	10,202	267,420	26,368	111,653	17,724	3,291
Number of participants	2,372	42,724	2,138	10,058	262,131	25,388	107,636	17,159	3,002
Percent of participation	94.96	98.55	97.45	98.59	98.02	96.28	96.40	96.81	91.22

Table 4.A.5 CAST Pilot Participation Rates for Grade Five—Primary Disability

Group	Special Education Services	No Special Education Services
Number of students	58,171	426,530
Number of participants	55,201	417,407
Percent of participation	94.89	97.86

Table 4.A.6 CAST Pilot Participation Rates for Grade Five—Migrant Status

Group	Migrant	Not Migrant
Number of students	4,285	480,416
Number of participants	4,185	468,423
Percent of participation	97.67	97.5

Table 4.A.7 CAST Pilot Participation Rates for Grade Eight—Gender

Group	All	Male	Female
Number of students	470,237	240,486	229,751
Number of participants	452,936	231,342	221,594
Percent of participation	96.32	96.20	96.45

Table 4.A.8 CAST Pilot Participation Rates for Grade Eight—English Language Fluency

Group	All	English Learner	English Only	Reclassified Fluent English Proficient	Initially Fluent English Proficient	To Be Determined	No Response
Number of students	470,237	57,368	257,158	131,766	22,779	359	807
Number of participants	452,936	54,544	245,759	129,509	22,299	278	547
Percent of participation	96.32	95.08	95.57	98.29	97.89	77.44	67.78

Table 4.A.9 CAST Pilot Participation Rates for Grade Eight—Economic Status

Group	Economically Disadvantaged	Not Economically Disadvantaged
Number of students	278,695	191,542
Number of participants	268,695	184,241
Percent of participation	96.41	96.19

Table 4.A.10 CAST Pilot Participation Rates for Grade Eight—Ethnicity

Group	American Indian/ Alaskan Native	Asian	Pacific Islander	Filipino	Hispanic or Latino	African American	White	Two or More Races	Unknown
Number of students	2,561	43,938	2,286	12,007	251,267	26,592	114,370	14,216	3,000
Number of participants	2,362	43,247	2,207	11,756	243,179	25,078	108,923	13,542	2,642
Percent of participation	92.23	98.43	96.54	97.91	96.78	94.31	95.24	95.26	88.07

Table 4.A.11 CAST Pilot Participation Rates for Grade Eight—Primary Disability

Group	Special Education Services	No Special Education Services
Number of students	51,432	418,805
Number of participants	47,481	405,455
Percent of participation	92.32	96.81

Table 4.A.12 CAST Pilot Participation Rates for Grade Eight—Migrant Status

Group	Migrant	Not Migrant
Number of students	3,684	466,553
Number of participants	3,553	449,383
Percent of participation	96.44	96.32

Table 4.A.13 CAST Pilot Participation Rates for High School—Gender

Group	All	Male	Female
Number of students	517,538	268,389	249,149
Number of participants	459,219	236,838	222,381
Percent of participation	88.73	88.24	89.26

Table 4.A.14 CAST Pilot Participation Rates for High School—English Language Fluency

Group	All	English Learner	English Only	Reclassified Fluent English Proficient	Initially Fluent English Proficient	To Be Determined	No Response
Number of students	517,538	60,000	276,731	151,260	27,251	439	1,857
Number of participants	459,219	49,671	244,054	139,652	24,751	283	808
Percent of participation	88.73	82.79	88.19	92.33	90.83	64.46	43.51

Table 4.A.15 CAST Pilot Participation Rates for High School—Economic Status

Group	Economically Disadvantage	Not Economically Disadv
Number of students	302,423	215,115
Number of participants	265,635	193,584
Percent of participation	87.84	89.99

Table 4.A.16 CAST Pilot Participation Rates for High School—Ethnicity

Group	American Indian/ Alaskan Native	Asian	Pacific Islander	Filipino	Hispanic or Latino	African American	White	Two or More Races	Unknown
Number of students	3,209	44,915	2,577	13,596	279,396	31,972	122,466	14,806	4,601
Number of participants	2,619	42,417	2,221	12,814	246,696	26,406	109,535	13,034	3,477
Percent of participation	81.61	94.44	86.19	94.25	88.30	82.59	89.44	88.03	75.57

Table 4.A.17 CAST Pilot Participation Rates for High School—Primary Disability

Group	Special Education Services	No Special Education Services
Number of students	53,020	464,518
Number of participants	42,679	416,540
Percent of participation	80.50	89.67

Table 4.A.18 CAST Pilot Participation Rates for High School—Migrant Status

Group	Migrant	Not Migrant
Number of students	3,499	514,039
Number of participants	3,155	456,064
Percent of participation	90.17	88.72

Appendix 4.B Demographic Summary

Notes:

1. This set of tables are presented separately for grades five, eight, and high school, which shows categories for grades ten, eleven, and twelve.
2. All students in California enrolled in grades five and eight who were eligible for the general science assessments were required to participate in the pilot.
3. Percents of student groups for the population are presented for Table 4.B.3 through Table 4.B.6, because high schools with grades ten, eleven, and twelve were assigned to the CAST pilot based on an assignment plan.
4. The percents of student groups may not sum to 100 due to rounding.

Table 4.B.1 Grade Five Demographic Summary

Demographic Student Group	Number Tested	Percent
All valid scores	472,608	100.00
Male	241,178	51.03
Female	231,430	48.97
American Indian or Alaska Native (All)	2,372	0.50
Asian (All)	42,724	9.04
Native Hawaiian or Other Pacific Islander (All)	2,138	0.45
Filipino (All)	10,058	2.13
Hispanic or Latino (All)	262,131	55.46
Black or African American (All)	25,388	5.37
White (All)	107,636	22.77
Two or more races (All)	17,159	3.63
Ethnicity unknown (All)	3,002	0.64
English only	266,056	56.30
Initially fluent English proficient	19,004	4.02
English learner	99,376	21.03
Reclassified fluent English proficient	87,213	18.45
To be determined	280	0.06
English proficiency unknown	679	0.14
No special education services	417,407	88.32
Special education services	55,201	11.68
Not economically disadvantaged	180,007	38.09
Economically disadvantaged	292,601	61.91
Migrant	4,185	0.89
Not migrant	468,423	99.11
American Indian or Alaska Native (Primary ethnicity—not economically disadvantaged)	784	0.17
Asian (Primary ethnicity—not economically disadvantaged)	27,633	5.85

Demographic Student Group	Number Tested	Percent
Native Hawaiian or Other Pacific Islander (Primary ethnicity—not economically disadvantaged)	732	0.15
Filipino (Primary ethnicity—not economically disadvantaged)	6,494	1.37
Hispanic or Latino (Primary ethnicity—not economically disadvantaged)	48,935	10.35
Black or African American (Primary ethnicity—not economically disadvantaged)	6,214	1.31
White (Primary ethnicity—not economically disadvantaged)	76,614	16.21
Two or more races (Primary ethnicity—not economically disadvantaged)	11,251	2.38
Ethnicity unknown (Primary ethnicity—not economically disadvantaged)	1,350	0.29
American Indian or Alaska Native (Primary ethnicity—economically disadvantaged)	1,588	0.34
Asian (Primary ethnicity—economically disadvantaged)	15,091	3.19
Native Hawaiian or Other Pacific Islander (Primary ethnicity—economically disadvantaged)	1,406	0.30
Filipino (Primary ethnicity—economically disadvantaged)	3,564	0.75
Hispanic or Latino (Primary ethnicity—economically disadvantaged)	213,196	45.11
Black or African American (Primary ethnicity—economically disadvantaged)	19,174	4.06
White (Primary ethnicity—economically disadvantaged)	31,022	6.56
Two or more races (Primary ethnicity—economically disadvantaged)	5,908	1.25
Ethnicity unknown (Primary ethnicity—economically disadvantaged)	1,652	0.35

Table 4.B.2 Grade Eight Demographic Summary

Demographic Student Group	Number Tested	Percent
All valid scores	452,936	100.00
Male	231,342	51.08
Female	221,594	48.92
American Indian or Alaska Native (All)	2,362	0.52
Asian (All)	43,247	9.55
Native Hawaiian or Other Pacific Islander (All)	2,207	0.49
Filipino (All)	11,756	2.60
Hispanic or Latino (All)	243,179	53.69
Black or African American (All)	25,078	5.54
White (All)	108,923	24.05
Two or more races (All)	13,542	2.99
Ethnicity unknown (All)	2,642	0.58
English only	245,759	54.26
Initially fluent English proficient	22,299	4.92
English learner	54,544	12.04
Reclassified fluent English proficient	129,509	28.59
To be determined	278	0.06
English proficiency unknown	547	0.12
No special education services	405,455	89.52
Special education services	47,481	10.48
Not economically disadvantaged	184,241	40.68
Economically disadvantaged	268,695	59.32
Migrant	3,553	0.78
Not migrant	449,383	99.22
American Indian or Alaska Native (Primary ethnicity—not economically disadvantaged)	867	0.19
Asian (Primary ethnicity—not economically disadvantaged)	27,423	6.05
Native Hawaiian or Other Pacific Islander (Primary ethnicity—not economically disadvantaged)	825	0.18
Filipino (Primary ethnicity—not economically disadvantaged)	7,750	1.71
Hispanic or Latino (Primary ethnicity—not economically disadvantaged)	49,991	11.04
Black or African American (Primary ethnicity—not economically disadvantaged)	7,175	1.58
White (Primary ethnicity—not economically disadvantaged)	79,809	17.62
Two or more races (Primary ethnicity—not economically disadvantaged)	9,094	2.01
Ethnicity unknown (Primary ethnicity—not economically disadvantaged)	1,307	0.29

Demographic Student Group	Number Tested	Percent
American Indian or Alaska Native (Primary ethnicity—economically disadvantaged)	1,495	0.33
Asian (Primary ethnicity—economically disadvantaged)	15,824	3.49
Native Hawaiian or Other Pacific Islander (Primary ethnicity—economically disadvantaged)	1,382	0.31
Filipino (Primary ethnicity—economically disadvantaged)	4,006	0.88
Hispanic or Latino (Primary ethnicity—economically disadvantaged)	193,188	42.65
Black or African American (Primary ethnicity—economically disadvantaged)	17,903	3.95
White (Primary ethnicity—economically disadvantaged)	29,114	6.43
Two or more races (Primary ethnicity—economically disadvantaged)	4,448	0.98
Ethnicity unknown (Primary ethnicity—economically disadvantaged)	1,335	0.29

Table 4.B.3 High School Demographic Summary

Demographic Student Group	Number Tested	Percent	Population Percent
All valid scores	459,219	100.00	100.00
Male	236,838	51.57	51.30
Female	222,381	48.43	48.70
American Indian or Alaska Native (All)	2,619	0.57	0.59
Asian (All)	42,417	9.24	9.32
Native Hawaiian or Other Pacific Islander (All)	2,221	0.48	0.51
Filipino (All)	12,814	2.79	2.85
Hispanic or Latino (All)	246,696	53.72	52.40
Black or African American (All)	26,406	5.75	6.00
White (All)	109,535	23.85	24.70
Two or more races (All)	13,034	2.84	2.85
Ethnicity unknown (All)	3,477	0.76	0.77
English only	244,054	53.15	54.13
Initially fluent English proficient	24,751	5.39	6.55
English learner	49,671	10.82	10.48
Reclassified fluent English proficient	139,652	30.41	28.46
To be determined	283	0.06	0.07
English proficiency unknown	808	0.18	0.31
No special education services	416,540	90.71	88.70
Special education services	42,679	9.29	11.30
Not economically disadvantaged	193,584	42.16	43.81
Economically disadvantaged	265,635	57.84	56.19
Migrant	3,155	0.69	0.68
Not migrant	456,064	99.31	99.32
American Indian or Alaska Native (Primary ethnicity—not economically disadvantaged)	1,061	0.23	0.25
Asian (Primary ethnicity—not economically disadvantaged)	26,154	5.70	5.72
Native Hawaiian or Other Pacific Islander (Primary ethnicity—not economically disadvantaged)	907	0.20	0.22
Filipino (Primary ethnicity—not economically disadvantaged)	8,673	1.89	1.93
Hispanic or Latino (Primary ethnicity—not economically disadvantaged)	56,400	12.28	12.74
Black or African American (Primary ethnicity—not economically disadvantaged)	8,414	1.83	2.01
White (Primary ethnicity—not economically disadvantaged)	81,735	17.80	18.62

Demographic Student Group	Number Tested	Percent	Population Percent
Two or more races (Primary ethnicity—not economically disadvantaged)	8,696	1.89	1.93
Ethnicity unknown (Primary ethnicity—not economically disadvantaged)	1,544	0.34	0.39
American Indian or Alaska Native (Primary ethnicity—economically disadvantaged)	1,558	0.34	0.34
Asian (Primary ethnicity—economically disadvantaged)	16,263	3.54	3.60
Native Hawaiian or Other Pacific Islander (Primary ethnicity—economically disadvantaged)	1,314	0.29	0.29
Filipino (Primary ethnicity—economically disadvantaged)	4,141	0.90	0.92
Hispanic or Latino (Primary ethnicity—economically disadvantaged)	190,296	41.44	39.66
Black or African American (Primary ethnicity—economically disadvantaged)	17,992	3.92	3.99
White (Primary ethnicity—economically disadvantaged)	27,800	6.05	6.08
Two or more races (Primary ethnicity—economically disadvantaged)	4,338	0.94	0.92
Ethnicity unknown (Primary ethnicity—economically disadvantaged)	1,933	0.42	0.37

Table 4.B.4 High School Grade Ten Demographic Summary

Demographic Student Group	Number Tested	Percent	Population Percent
All valid scores	371,949	100.00	100.00
Male	189,765	51.02	51.25
Female	182,184	48.98	48.75
American Indian or Alaska Native (All)	1,967	0.53	0.58
Asian (All)	36,644	9.85	9.09
Native Hawaiian or Other Pacific Islander (All)	1,782	0.48	0.48
Filipino (All)	10,924	2.94	2.76
Hispanic or Latino (All)	198,426	53.35	53.31
Black or African American (All)	20,857	5.61	5.87
White (All)	87,713	23.58	24.14
Two or more races (All)	10,631	2.86	2.90
Ethnicity unknown (All)	3,005	0.81	0.86
English only	195,390	52.53	53.56
Initially fluent English proficient	18,906	5.08	4.90
English learner	40,074	10.77	11.71
Reclassified fluent English proficient	116,690	31.37	29.47
To be determined	254	0.07	0.09
English proficiency unknown	635	0.17	0.27
No special education services	337,656	90.78	89.20
Special education services	34,293	9.22	10.80
Not economically disadvantaged	158,628	42.65	42.45
Economically disadvantaged	213,321	57.35	57.55
Migrant	2,401	0.65	0.70
Not migrant	369,548	99.35	99.30
American Indian or Alaska Native (Primary ethnicity—not economically disadvantaged)	808	0.22	0.22
Asian (Primary ethnicity—not economically disadvantaged)	22,340	6.01	5.67
Native Hawaiian or Other Pacific Islander (Primary ethnicity—not economically disadvantaged)	740	0.20	0.20
Filipino (Primary ethnicity—not economically disadvantaged)	7,375	1.98	1.87
Hispanic or Latino (Primary ethnicity—not economically disadvantaged)	45,698	12.29	12.27
Black or African American (Primary ethnicity—not economically disadvantaged)	6,794	1.83	1.87
White (Primary ethnicity—not economically disadvantaged)	66,327	17.83	18.00

Demographic Student Group	Number Tested	Percent	Population Percent
Two or more races (Primary ethnicity—not economically disadvantaged)	7,237	1.95	1.95
Ethnicity unknown (Primary ethnicity—not economically disadvantaged)	1,309	0.35	0.40
American Indian or Alaska Native (Primary ethnicity—economically disadvantaged)	1,159	0.31	0.36
Asian (Primary ethnicity—economically disadvantaged)	14,304	3.85	3.42
Native Hawaiian or Other Pacific Islander (Primary ethnicity—economically disadvantaged)	1,042	0.28	0.28
Filipino (Primary ethnicity—economically disadvantaged)	3,549	0.95	0.89
Hispanic or Latino (Primary ethnicity—economically disadvantaged)	152,728	41.06	41.04
Black or African American (Primary ethnicity—economically disadvantaged)	14,063	3.78	4.00
White (Primary ethnicity—economically disadvantaged)	21,386	5.75	6.15
Two or more races (Primary ethnicity—economically disadvantaged)	3,394	0.91	0.95
Ethnicity unknown (Primary ethnicity—economically disadvantaged)	1,696	0.46	0.46

Table 4.B.5 High School Grade Eleven Demographic Summary

Demographic Student Group	Number Tested	Percent	Population Percent
All valid scores	43,282	100.00	100.00
Male	21,980	50.78	51.19
Female	21,302	49.22	48.81
American Indian or Alaska Native (All)	288	0.67	0.59
Asian (All)	4,279	9.89	9.66
Native Hawaiian or Other Pacific Islander (All)	226	0.52	0.52
Filipino (All)	1,236	2.86	2.86
Hispanic or Latino (All)	21,942	50.69	52.08
Black or African American (All)	2,143	4.95	5.87
White (All)	11,680	26.99	24.77
Two or more races (All)	1,217	2.81	2.83
Ethnicity unknown (All)	271	0.63	0.82
English only	24,381	56.33	53.92
Initially fluent English proficient	3,378	7.80	7.17
English learner	3,913	9.04	10.56
Reclassified fluent English proficient	11,532	26.64	27.87
To be determined	13	0.03	0.07
English proficiency unknown	65	0.15	0.41
No special education services	39,531	91.33	89.63
Special education services	3,751	8.67	10.37
Not economically disadvantaged	20,595	47.58	44.23
Economically disadvantaged	22,687	52.42	55.77
Migrant	447	1.03	0.69
Not migrant	42,835	98.97	99.31
American Indian or Alaska Native (Primary ethnicity—not economically disadvantaged)	126	0.29	0.25
Asian (Primary ethnicity—not economically disadvantaged)	2,915	6.73	5.94
Native Hawaiian or Other Pacific Islander (Primary ethnicity—not economically disadvantaged)	99	0.23	0.23
Filipino (Primary ethnicity—not economically disadvantaged)	896	2.07	1.94
Hispanic or Latino (Primary ethnicity—not economically disadvantaged)	5,378	12.43	12.76
Black or African American (Primary ethnicity—not economically disadvantaged)	826	1.91	2.01
White (Primary ethnicity—not economically disadvantaged)	9,297	21.48	18.72

Demographic Student Group	Number Tested	Percent	Population Percent
Two or more races (Primary ethnicity—not economically disadvantaged)	898	2.07	1.93
Ethnicity unknown (Primary ethnicity—not economically disadvantaged)	160	0.37	0.46
American Indian or Alaska Native (Primary ethnicity—economically disadvantaged)	162	0.37	0.34
Asian (Primary ethnicity—economically disadvantaged)	1,364	3.15	3.73
Native Hawaiian or Other Pacific Islander (Primary ethnicity—economically disadvantaged)	127	0.29	0.29
Filipino (Primary ethnicity—economically disadvantaged)	340	0.79	0.92
Hispanic or Latino (Primary ethnicity—economically disadvantaged)	16,564	38.27	39.32
Black or African American (Primary ethnicity—economically disadvantaged)	1,317	3.04	3.85
White (Primary ethnicity—economically disadvantaged)	2,383	5.51	6.05
Two or more races (Primary ethnicity—economically disadvantaged)	319	0.74	0.90
Ethnicity unknown (Primary ethnicity—economically disadvantaged)	111	0.26	0.36

Table 4.B.6 High School Grade Twelve Demographic Summary

Demographic Student Group	Number Tested	Percent	Population Percent
All valid scores	43,988	100.00	100.00
Male	25,093	57.05	51.47
Female	18,895	42.95	48.53
American Indian or Alaska Native (All)	364	0.83	0.60
Asian (All)	1,494	3.40	9.22
Native Hawaiian or Other Pacific Islander (All)	213	0.48	0.53
Filipino (All)	654	1.49	2.94
Hispanic or Latino (All)	26,328	59.85	51.78
Black or African American (All)	3,406	7.74	6.29
White (All)	10,142	23.06	25.19
Two or more races (All)	1,186	2.70	2.83
Ethnicity unknown (All)	201	0.46	0.62
English only	24,283	55.20	54.93
Initially fluent English proficient	2,467	5.61	7.64
English learner	5,684	12.92	9.13
Reclassified fluent English proficient	11,430	25.98	28.00
To be determined	16	0.04	0.05
English proficiency unknown	108	0.25	0.26
No special education services	39,353	89.46	87.23
Special education services	4,635	10.54	12.77
Not economically disadvantaged	14,361	32.65	44.80
Economically disadvantaged	29,627	67.35	55.20
Migrant	307	0.70	0.64
Not migrant	43,681	99.30	99.36
American Indian or Alaska Native (Primary ethnicity—not economically disadvantaged)	127	0.29	0.27
Asian (Primary ethnicity—not economically disadvantaged)	899	2.04	5.55
Native Hawaiian or Other Pacific Islander (Primary ethnicity—not economically disadvantaged)	68	0.15	0.24
Filipino (Primary ethnicity—not economically disadvantaged)	402	0.91	1.99
Hispanic or Latino (Primary ethnicity—not economically disadvantaged)	5,324	12.10	13.21
Black or African American (Primary ethnicity—not economically disadvantaged)	794	1.81	2.16
White (Primary ethnicity—not economically disadvantaged)	6,111	13.89	19.15

Demographic Student Group	Number Tested	Percent	Population Percent
Two or more races (Primary ethnicity—not economically disadvantaged)	561	1.28	1.91
Ethnicity unknown (Primary ethnicity—not economically disadvantaged)	75	0.17	0.33
American Indian or Alaska Native (Primary ethnicity—economically disadvantaged)	237	0.54	0.33
Asian (Primary ethnicity—economically disadvantaged)	595	1.35	3.67
Native Hawaiian or Other Pacific Islander (Primary ethnicity—economically disadvantaged)	145	0.33	0.30
Filipino (Primary ethnicity—economically disadvantaged)	252	0.57	0.95
Hispanic or Latino (Primary ethnicity—economically disadvantaged)	21004	47.75	38.57
Black or African American (Primary ethnicity—economically disadvantaged)	2612	5.94	4.13
White (Primary ethnicity—economically disadvantaged)	4031	9.16	6.04
Two or More Races (Primary ethnicity—economically disadvantaged)	625	1.42	0.92
Ethnicity Unknown (Primary ethnicity—economically disadvantaged)	126	0.29	0.29

Chapter 5: Scoring

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

5.1. Constructed Response (CR) Scoring Sampling Process

5.1.1. Sample Size and Number of CR items

A sample of responses to each piloted CR item was selected for human scoring. Because the purposes of the CAST pilot were to provide information on the performance of the newly developed items and test the functionality of the systems, only a small percentage of response data were needed. In addition, the timing and budget constraints also rendered it unrealistic to score all responses.

To allow for building AI scoring models, at least 1,800 human-scored responses were needed with approximately 600 of those, at minimum, double-scored by two independent raters. The same sample drawn for CR scoring was used for item analysis. This procedure yielded completely scored responses of both CR and machine-scored MC items, which enables Educational Testing Service (ETS) to conduct classical item analyses (IA) and differential item functioning (DIF) analyses for various student groups of interest. Refer to subsection *6.1 Samples Used for Item Analyses* for more details about the sample used for IA and DIF analyses.

Human scoring was conducted on four piloted CR items. ETS developed at least two CR items for each grade span, but only four CR items were approved by the California Department of Education (CDE) for scoring, including three CR items on the grade five pilot and one CR item on the high school pilot. There were no approved CR items for scoring for the grade eight pilot. For each CR, 1,800 responses were scored per form, as shown in Table 5.1. In all but one case, the CR items appeared on both the general forms because it was part of the performance task, which was common between the two general forms at each tested grade level. As a result, each of these three CR items was scored with 3,600 responses; for the CR item that was not shared between the two general forms, 1,800 responses were scored, as shown in Table 5.1. “N/A” indicates there are no CR item responses scored on a particular form.

Table 5.1 Sample of Responses for Human Scoring the Piloted CRs

Grade	Client ID	Accession Number	General Form 1	General Form 2	Accessible Form	Sample for Human Scoring
5	CNGSE0360S	VH689376	X	N/A	N/A	1,800
5	CNGSE0371S	VH667949	X		N/A	1,800
5	CNGSE0371S	VH667949		X	N/A	1,800
5	CNGSE0372S	VH668026	X		N/A	1,800
5	CNGSE0372S	VH668026		X	N/A	1,800

Grade	Client ID	Accession Number	General Form 1	General Form 2	Accessible Form	Sample for Human Scoring
HS	CNGSH0291S	VH651810	X		N/A	1,800
HS	CNGSH0291S	VH651810		X	N/A	1,800

5.1.2. Selection Criteria

Given the sample size constraint for analyzing the pilot data, it was imperative to identify the most appropriate student sample. Specifically, it was optimal to include students who were most likely to be exposed to the California Next Generation Science Standards (CA NGSS). Therefore, the sampling procedure prioritized selecting students from local educational agencies (LEAs) that were known early adopters of the CA NGSS. Students from such LEAs were more likely to have been exposed to curriculum aligned to the CA NGSS and were thus more likely to be measured more validly by materials related to the CA NGSS. In contrast, students in LEAs that were still in the process of adopting the CA NGSS may not have been exposed to the materials related to the CA NGSS; those students may not have been ready for the CAST pilot. Table 5.2 and Table 5.4 in the next subsection present the numbers of students who were assigned to take the CAST in schools that were early adopters and in schools that were not early adopters, and the numbers in both categories who took the assessments.

Another concern was that poor item performance could be due to the lack of motivation for a pilot assessment. Therefore, “motivated students” was the second criterion for sampling. Motivated students were defined as those who completed all of the machine-scorable, multiple-choice items on the form and who completed the test in a realistic length of time. The minimum testing time threshold was calculated for each form by determining that students should spend at least 20 seconds on single-select MC items and at least 30 seconds on all other items. The minimum testing time for a student to qualify as a motivated student is obtained by cumulating the minimum item times.

5.1.3. Sampling Procedures

The sampling proceeded as follows for each of the grade five general forms based on the pilot data extracted on May 9, 2017, after screening out unmotivated students.

1. Draw a simple random sample of as many early-adopter LEA students as available, up to 1,800 students. If the entire sample was selected from early-adopter LEAs, then the sampling process is complete; otherwise, continue to step 2.
2. The remaining student sample needed for analysis was the difference between 1,800 and the number of students selected in step 1. This sample was selected by taking a simple random sample of non-early-adopter LEA students. For example, if there were only 800 early-adopter LEA students selected in step 1, an additional 1,000 non-early-adopter LEA students were selected.

The responses of the sampled students to the approved CR items were then human-scored.

A similar procedure was used for each of the high school general forms. However, the sampling was stratified by grade level to ensure equal representation of each of the three tested grade levels: ten, eleven, and twelve. Using data extracted on May 9, 2017, and after screening out unmotivated students, the following sampling procedure was followed at each grade level, for each general form.

1. Draw a simple random sample of as many early-adopter LEA students as available, up to 600 for each grade. If the entire sample was selected from early-adopter LEAs, then the sampling process was complete; otherwise, continue to step 2.
2. The remaining student sample needed for analysis was the difference between 600 and the number of students selected in step 1. This sample was selected by taking a simple random sample of non-early-adopter LEA students to reach 600 students. For instance, if there were only 200 early-adopter LEA students selected in step 1, an additional 400 non-early-adopter LEA students were selected.

The responses of the sampled students to the approved CR items were then human-scored.

Table 5.2 and Table 5.3 summarize the resulting samples from these sampling procedures. They show that for grades five and ten, all sampled students were from early-adopter LEAs. However, for grades eleven and twelve, the sample was split between early-adopter and non-early-adopter LEA students. (Because there were no approved CR items for scoring for the grade eight pilot, there are no data to present for this grade.)

Table 5.2 Composition of Student Samples for CR Scoring—Students Available as of May 9, 2017

Tested Grade	Form	Early Adopter LEAs	Non-Early Adopter LEAs
5	Gr. 5 Form 1	3,180	108,888
5	Gr. 5 Form 2	3,218	109,358
10	HS Form 1	3,952	121,422
11	HS Form 1	72	13,561
12	HS Form 1	216	14,044
10	HS Form 2	3,847	118,283
11	HS Form 2	66	13,134
12	HS Form 2	213	13,331

Table 5.3 Composition of Student Samples for CR Scoring—Students Sampled

Tested Grade	Form	Early Adopter LEAs	Non-Early Adopter LEAs
5	Gr. 5 Form 1	1,800	0
5	Gr. 5 Form 2	1,800	0
10	HS Form 1	600	0
11	HS Form 1	72	528
12	HS Form 1	216	384
10	HS Form 2	600	0
11	HS Form 2	66	534
12	HS Form 2	213	387

5.2. Human Scoring

ETS conducted human scoring for piloted CR items using the Online Network for Evaluation (ONE) system and qualified raters from existing California Assessment of Student Performance and Progress (CAASPP) rater pools, as well as using California science teachers. The next subsections detail aspects of the human scoring process, from range finding and recruiting through the quality monitoring processes.

5.2.1 Rater Recruitment and Certification Process

Several weeks prior to the start of CR scoring, ETS recruited a pool of eligible CAST raters from the current CAASPP Smarter Balanced pool of eligible raters and invited science teachers from California. All CAST raters were required to have a bachelor's degree to be eligible to attempt certification. Out of the original eligible CAST pool of 116 raters, 54 raters completed scoring responses for three prompts for grade five and one prompt for high school over a two-day scoring window. Twenty-seven percent of the scoring pool consisted of California educators; the remaining pool of raters represented a variety of backgrounds in business, education, and other fields.

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

Raters were required to achieve an 80 percent exact match to the CDE-approved rating for the responses on at least one of the certification sets to be eligible for calibration on a specific grade-level test prompt. If raters did not pass either certification set, they were excused from scoring for the 2016–17 CAST pilot items.

5.2.2 Scoring Leader and Rater Training

ETS selected scoring leaders to oversee a group of raters during the scoring process. Scoring leaders are experienced raters who have demonstrated high scoring accuracy from previous scoring projects at ETS and are invited to act as a scoring leader on a project. For the CAST pilot, 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 CR Scoring S&C with scoring quality.

5.2.2.1. Training for Scoring Leaders

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

5.2.2.2. Training for Raters

Training for raters occurred within the ONE system. Raters were provided ONE system training documents as well as program-specific information that they could refer to at any time. Prior to attempting calibration, raters were given a window of time to review all training

materials in the system and practice scoring using the prescored training sets. After raters completed a training set, they were provided with annotations for each response as a rationale for the rating assigned.

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

1. **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**], and the benchmarks, also called anchors, were accessed in ONE through a tab called [**Benchmarks**]. The benchmarks had annotations associated with them to call the rater's attention to specific content in the sample responses.

2. **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 CAST pilot 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.

3. **Set calibration.**

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

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

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

   For the CAST pilot, 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.

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

- ETS psychometric staff provided a sampling plan of 1,800 specific responses per prompt per form to be scored. See subsection *5.1 Constructed Response (CR) Scoring Sampling Process* for the sampling plan.
- The sampling plan was uploaded to ONE to activate the responses for scoring.
- Approximately 33 percent of responses were double-scored to facilitate the building of AI scoring models.
  - ONE randomly selected the responses to receive double scoring with the second rating conducted by a different rater.
  - Raters were not aware when a second scoring was occurring and so did not have access to the first score.
- Raters did not have access to condition codes and were instructed during training to defer any nonscorable responses to their scoring leader for scoring. The condition codes were:
  - Blank (B): The response area was completely blank.
  - Insufficient (I): The response had no meaningful response or even a guess at a possible answer (e.g., random keystrokes, opinions of the test).
  - Nonscorable Language (L): The language of the response was not English.
- Scoring leaders were trained to apply condition codes to nonscorable responses.
- Raters were instructed to apply zero (0) scores when there was an attempt to answer the question but the information was incorrect so could not earn the minimum score. If the rater was unsure, he or she deferred responses to the scoring leader.

5.2.3 Range Finding

Soon after receiving a large volume of CR responses from California schools, ETS began the range finding process of randomly selecting a wide variety of student response samples for the purpose of having 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. Part of the range finding process also 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 Development (AD) staff used the rubric (scoring guide) for each item to randomly select and score responses to represent each score point on an item's rubric. The number of responses selected varied by prompt and was based on the number of points and the prompts that were preselected for certifying and training raters. Scored samples were needed for the various purposes summarized in Table 5.4.

Table 5.4 CAST Pilot Sample Selection for Human Scoring Procedures

Sample Type	Purpose	Number of Sets and Samples in Sets	Configuration of Sets
Certification	Certification samples for verifying scoring accuracy of potential raters and Scoring Leaders	• Two sets of 10 samples per set for **one high school 2-point prompt** • Mixed score points	Three to five samples for each score point per set
Training	Training samples with annotations for rater training and scoring practice	• Two sets of seven samples per set per grade **for one prompt at each grade** • Mixed score points • High school prompt with an additional set of six samples to be used as postcertification training prior to calibration	Two to three samples for each score point per set
Benchmarks	Benchmark samples with annotations that represent exemplar responses at each score point on the rubric	• One set of four to nine samples per unique prompt per grade (eight unique prompts total)	Two to three samples for each score point
Calibration	Calibration samples for evaluating rater scoring performance on specific prompts	• Two sets of five samples per set for one prompt per grade • Mixed score points	One to three samples for each score point per set
Validity	Validity samples inserted into rater's scoring queue to monitor the quality of scoring	• One set of 20 papers per prompt • Mixed score points	Six through 12 samples for each score point

2. Responses were scored by two independent, experienced raters using the ONE system. ETS AD staff also wrote annotations, or short notes, with each score point to explain why a response earned a particular rating. Annotations helped raters make explicit connections between the scoring guide and responses, and thus informed their careful and accurate scoring of responses. ETS provided the CDE with the independent ratings, scored samples, annotations, and recommendations for which responses would go in the different scoring materials (i.e., certification, benchmark, training, calibration, and validity, as summarized in Table 5.4).
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, training papers, and annotations from seven California science teachers from grades five, eight, and high school. The teachers were recruited from the existing 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 then created all final sample sets in the ONE system and used these samples as part of a system of training and controls for verifying the quality and consistency of pilot scoring.

5.2.4 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 backread 10 percent or more 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 and/or benchmarks to explain why a response should have received a different score. When a rater was scoring inconsistently, the backreading proportion may be more than 10 percent.

Prescored responses from validity sets were also inserted into the rater's queue for every 10 responses scored. These were inserted in random positions and not fixed so a rater was unaware which response was a validity response. The ETS CR Scoring S&C team 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 utilized to monitor the progress and accuracy of individual raters and raters in aggregate. Reports produced to show rater productivity and performance presented how many responses a rater scored during a shift and how two raters scored the same response (i.e., interrater reliability).

5.2.5 Interrater Reliability

The ONE system captures interrater reliability by monitoring data for responses that are double-scored. For the CAST pilot, 33 percent of responses were double-scored for studies to be made for possible AI scoring. The interrater reliability reports included the number and percent of exact matches for each rater and the number and percent of adjacent and discrepant scores. Scoring management reviewed the interrater reliability statistics for each prompt to determine if there were any issues that needed to be addressed during scoring. The interrater reliability statistics are included in subsection *6.5 Interrater Reliability Analyses*.

5.2.6 Validity Responses and Sets

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

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

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

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

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

Features such as backreading, interrater reliability reporting functions, and validity response insertion and reporting functions allowed scoring leaders to quickly identify inaccurate scoring patterns and take appropriate corrective actions. Such actions may have included retraining individuals and groups of raters, dismissal of raters if quality standards could not be met, and rescoring of responses when anomalous scoring patterns were identified.

5.2.7 Scoring Metrics Development

ETS's AD group developed six performance tasks—two per grade span—to be included in the 2016–17 CAST pilot, as appropriate, for measuring more complex skills. During item development, draft scoring metrics (rubrics) were created with the point scale and descriptions. ETS included these rubrics with the associated items in the internal and external review processes described in subsection *3.6 Item Review Process*. Rubrics were edited as needed on the basis of feedback from the CDE and California teachers during the item review and range finding processes. Exemplar responses of each score point were provided for scoring guidance as benchmarks.

5.3. Machine Scoring

CAST pilot assessments included machine-scorable, traditional MC items and TE items that were scored by the test delivery system (TDS). In the TDS, responses to the test forms were compared with the answer keys or scoring rubrics embedded in the TDS to determine the score points. A real-time, quality-monitoring component was built into the TDS. After a test was administered to a student, the TDS passed the resulting data to the Quality Assurance system to ensure a score from the machine-scoring system was scored accurately. The details of the quality control are provided in subsection *7.4.1 Quality Control of Machine-Scoring Procedures*.

5.4. AI Scoring Model Building

For the pilot administration, the purpose of AI scoring model building was to assess the feasibility of AI scoring of future CAST CR items. The main stages involved were

1. data collection,
2. model training, and
3. model evaluation.

5.4.1. Data Collection

After the pilot, ETS collected a sample of students' responses to four CR items with human score(s) assigned, as described in subsection *5.1 Constructed Response (CR) Scoring Sampling Process*. ETS also collected student demographic information such as gender, ethnicity, and economic status to use for student group analysis during the model-evaluation stage.

The number of responses and the percentage of double-scoring for the four CR items are shown in Table 5.5. The item VH689376 has half the number of responses as the other three items because it appeared on one form instead of two.

Table 5.5 The Number of CR Responses Scored and Double Scored

Item ID	N Responses	% of Double-Scored Responses
VH668026	3,600	32.6
VH667949	3,600	34.0
VH689376	1,800	32.3
VH651810	3600	33.7

5.4.2. 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 these 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.[4] Each model then went through an evaluation stage with multiple statistical criteria, such as Pearson's *r* and Quadratic-weighted Kappa, using an independent data set. The evaluation will be performed at the overall data set level as well as student group–level and reported in a separate report. Figure 5.1 illustrates the primary steps.

[4] A Support Vector Machine (SVM) performs classification by finding the hyperplane that maximizes the margin between two classes. The vectors (cases) that define the hyperplane are the support vectors. The details can be found in Vapnik, Vladimir N.; *The Nature of Statistical Learning Theory*, Springer-Verlag, 1995. The Support Vector Regression (SVR) is an extension of SVMs and uses the same principles as the SVM for classification, with only a few minor differences. For details of SVR, see Drucker, Harris; Burges, Christopher J. C.; Kaufman, Linda; Smola, Alexander J.; and Vapnik, Vladimir N. (1997); "Support Vector Regression Machines," in *Advances in Neural Information Processing Systems 9*, NIPS 1996, 155–161, MIT Press.

See the Alternative Text for Figure 5.1 for a description of this figure.

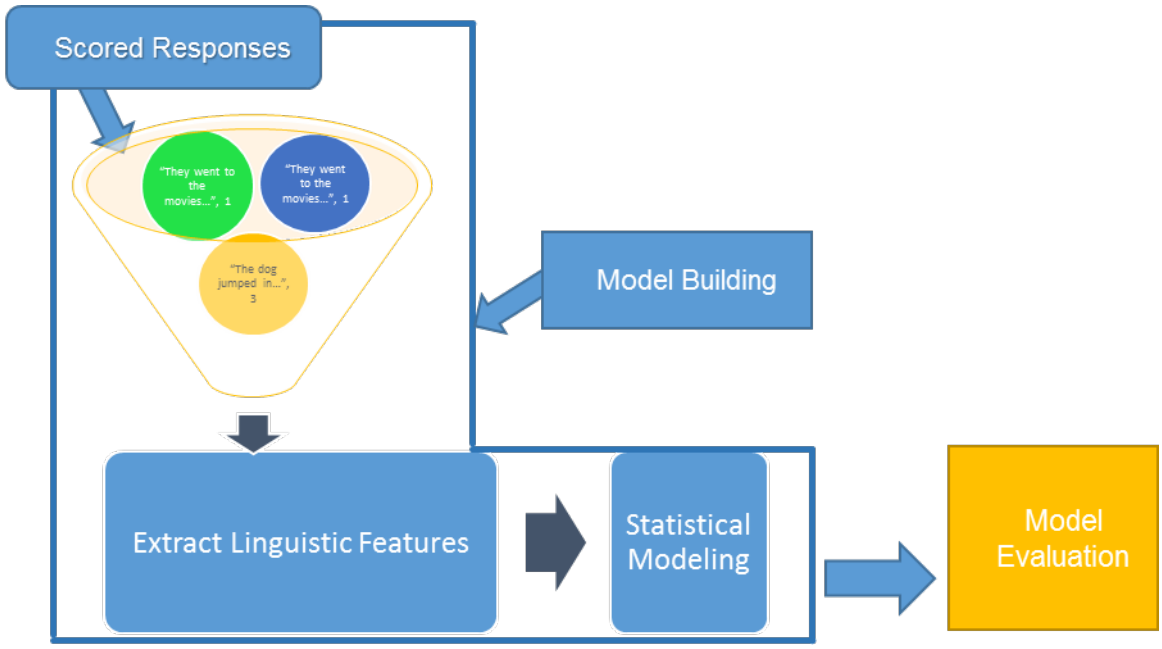

Figure 5.1 Model Building Steps

5.4.3. Model Evaluation

One of the important factors in building AI scoring models with good performance was the use of data with reliable human scores. A commonly used indicator for evaluating human scoring reliability is to use more than one rater on a large enough sample of responses and evaluate the extent to which they agree with each other. The agreement rates between two human raters for the samples for the four pilot CR items are shown in Table 6.5 in subsection *6.5 Interrater Reliability Analyses*.

Quadratic-weighted Kappa is consistently higher when condition codes were excluded from the calculation, as shown in Table 5.6. (Refer to subsection *6.5 Interrater Reliability Analyses* for the definition of Quadratic-weighted Kappa.) This is largely due to the fact that human raters showed significant confusion between the condition code "I" and the score of "0." Refer to subsection *5.2.2 Scoring Leader and Rater Training* for details on the condition codes. The case for condition codes is shown for each item in the confusion matrices. The shaded cells of Table 5.7 through Table 5.10 show where the human raters disagreed on condition codes.

Table 5.6 Human Rater Agreement

Item ID	Quadratic Weighted Kappa (includes condition codes)	Quadratic Weighted Kappa (0, 1 only)	Correlation (0, 1 only)
VH668026	0.757	0.772	0.772
VH667949	0.773	0.785	0.785
VH689376	0.739	0.771	0.771
VH651810	0.814	0.827	0.828

Table 5.7 Confusion Matrix for Two Human Raters, VH667949 (Grade Five)

	00	01	B	I	L
00	616	67	1	6	1
01	59	451	0	1	0
B	0	0	0	0	0
I	6	0	0	13	0
L	0	0	0	0	2

Table 5.8 Confusion Matrix for Two Human Raters, VH668026 (Grade Five)

	00	01	I	L
00	497	64	10	1
01	65	506	0	0
I	10	0	16	0
L	0	0	0	4

Table 5.9 Confusion Matrix for Two Human Raters, VH689376 (Grade Five)

	00	01	I
00	210	30	3
01	33	289	0
I	12	0	5

Table 5.10 Confusion Matrix for Two Human Raters, VH651810 (High School)

	00	01	I	L
00	592	37	9	0
01	61	460	0	0
I	11	0	42	3
L	0	0	0	0

In human scoring, all condition codes were assigned the score of “00” in section-level or total-score calculation. Following this practice, they were converted to “00” during model building and evaluation. Rather than training AI scoring models to detect nonscorable responses, a separate filtering tool will be created to detect them. The AI scoring study was concluded at the end of November 2017; the final results will be described in a separate report.

5.5. Raw Scores Distributions

For all of the CAST pilot assessments, the total test raw score equaled the sum of the points of test items answered correctly. Table 5.11 through Table 5.13 show the raw score distributions of each form for each test using the analysis sample for grades five, eight, and high school respectively. They are based on an item analysis sample as described in section *6.1 Samples Used for Item Analyses*. Hyphens (" - ") indicate where there are no students at a particular score.

Table 5.11 Grade Five Raw Score Frequency Distribution

Raw Score	Form 1 Frequency	Form 1 Percent	Form 1 Cumulative Percent	Form 2 Frequency	Form 2 Percent	Form 2 Cumulative Percent	Accessibility Form Frequency	Accessibility Form Percent	Accessibility Form Cumulative Percent
0	5	0.28	0.28	4	0.22	0.22	83	4.61	4.61
1	32	1.78	2.06	14	0.78	1.00	275	15.28	19.89
2	65	3.61	5.67	47	2.61	3.61	472	26.22	46.11
3	109	6.06	11.72	79	4.39	8.00	422	23.44	69.56
4	128	7.11	18.83	112	6.22	14.22	287	15.94	85.50
5	170	9.44	28.28	139	7.72	21.94	182	10.11	95.61
6	150	8.33	36.61	181	10.06	32.00	66	3.67	99.28
7	165	9.17	45.78	181	10.06	42.06	13	0.72	100.00
8	164	9.11	54.89	186	10.33	52.39	-	-	-
9	178	9.89	64.78	166	9.22	61.61	-	-	-
10	153	8.50	73.28	148	8.22	69.83	-	-	-
11	143	7.94	81.22	150	8.33	78.17	-	-	-
12	108	6.00	87.22	111	6.17	84.33	-	-	-
13	107	5.94	93.17	101	5.61	89.94	-	-	-
14	54	3.00	96.17	79	4.39	94.33	-	-	-
15	41	2.28	98.44	49	2.72	97.06	-	-	-
16	18	1.00	99.44	30	1.67	98.72	-	-	-
17	7	0.39	99.83	17	0.94	99.67	-	-	-
18	2	0.11	99.94	6	0.33	100.00	-	-	-
19	1	0.06	100.00	-	-	-	-	-	-

Table 5.12 Grade Eight Raw Score Frequency Distribution

Raw Score	Form 1 Frequency	Form 1 Percent	Form 1 Cumulative Percent	Form 2 Frequency	Form 2 Percent	Form 2 Cumulative Percent	Accessibility Form Frequency	Accessibility Form Percent	Accessibility Form Cumulative Percent
0	1	0.06	0.06	0	0.00	0.00	281	15.61	15.61
1	2	0.11	0.17	4	0.22	0.22	578	32.11	47.72
2	19	1.06	1.22	21	1.17	1.39	498	27.67	75.39
3	54	3.00	4.22	44	2.44	3.83	299	16.61	92.00
4	108	6.00	10.22	112	6.22	10.06	98	5.44	97.44
5	161	8.94	19.17	157	8.72	18.78	39	2.17	99.61
6	205	11.39	30.56	216	12.00	30.78	7	0.39	100.00
7	203	11.28	41.83	222	12.33	43.11	-	-	-
8	217	12.06	53.89	227	12.61	55.72	-	-	-
9	196	10.89	64.78	225	12.50	68.22	-	-	-
10	170	9.44	74.22	168	9.33	77.56	-	-	-
11	128	7.11	81.33	143	7.94	85.50	-	-	-
12	108	6.00	87.33	84	4.67	90.17	-	-	-
13	77	4.28	91.61	74	4.11	94.28	-	-	-
14	59	3.28	94.89	47	2.61	96.89	-	-	-
15	35	1.94	96.83	27	1.50	98.39	-	-	-
16	28	1.56	98.39	18	1.00	99.39	-	-	-
17	17	0.94	99.33	7	0.39	99.78	-	-	-
18	9	0.50	99.83	4	0.22	100.00	-	-	-
19	2	0.11	99.94	-	-	-	-	-	-
20	0	0.00	99.94	-	-	-	-	-	-
21	1	0.06	100.00	-	-	-	-	-	-

Table 5.13 High School Raw Score Frequency Distribution

Raw Score	Form 1 Frequency	Form 1 Percent	Form 1 Cumulative Percent	Form 2 Frequency	Form 2 Percent	Form 2 Cumulative Percent	Accessibility Form Frequency	Accessibility Form Percent	Accessibility Form Cumulative Percent
0	5	0.28	0.28	4	0.22	0.22	148	8.22	8.22
1	20	1.11	1.39	32	1.78	2.00	355	19.72	27.94
2	87	4.84	6.23	104	5.78	7.79	542	30.11	58.06
3	150	8.34	14.56	176	9.79	17.58	434	24.11	82.17
4	202	11.23	25.79	269	14.96	32.54	236	13.11	95.28
5	238	13.23	39.02	286	15.91	48.44	65	3.61	98.89
6	261	14.51	53.53	283	15.74	64.18	19	1.06	99.94
7	209	11.62	65.15	222	12.35	76.53	1	0.06	100.00
8	192	10.67	75.82	146	8.12	84.65	-	-	-
9	165	9.17	84.99	103	5.73	90.38	-	-	-
10	125	6.95	91.94	64	3.56	93.94	-	-	-
11	66	3.67	95.61	39	2.17	96.11	-	-	-
12	43	2.39	98.00	26	1.45	97.55	-	-	-
13	21	1.17	99.17	23	1.28	98.83	-	-	-
14	8	0.44	99.61	12	0.67	99.50	-	-	-
15	6	0.33	99.94	7	0.39	99.89	-	-	-
16	1	0.06	100.00	1	0.06	99.94	-	-	-
17	-	-	-	0	0.00	99.94	-	-	-
18	-	-	-	1	0.06	100.00	-	-	-

Table 5.14 through Table 5.19 present, for the grade five, grade eight, and high school assessments, the number of items, the number of tested students and the summary statistics of each form—two general forms and one accessible form—as well as the common set between the two main forms. The summary statistics presented include the mean and standard deviation (SD). Table 5.14 through Table 5.19 are based on an item analysis sample as described in subsection *6.1 Samples Used for Item Analyses*.

Table 5.14 Summary Statistics of the Raw Scores—Grade Five

Form	N Items	N Points	N Students	Mean	Mean as % of Total	SD
1	19	19	1,800	8.02	42	3.63
2	18	19	1,800	8.51	45	3.63
Set of common items between 1 & 2	6	6	1,800	2.42	40	1.49
Set of common items between 1 & 2	6	6	1,800	2.45	41	1.59
Accessibility	7	7	1,800	2.79	40	1.50

Table 5.15 Summary Statistics of the Raw Scores—Grade Eight

Form	N Items	N Points	N Students	Mean	Mean as % of Total	SD
1	21	22	1,800	8.50	39	3.33
2	20	21	1,800	8.26	39	3.05
Set of common items between 1 & 2	6	7	1,800	3.00	43	1.40
Set of common items between 1 & 2	6	7	1,800	2.92	42	1.36
Accessibility	6	6	1,800	1.72	29	1.23

Table 5.16 Summary Statistics of the Raw Scores—All High School

Form	N Items	N Points	N Students	Mean	Mean as % of Total	SD
1	18	18	1,799	6.49	36	2.77
2	20	21	1,798	5.90	28	2.68
Set of common items between 1 & 2	5	5	1,799	1.53	31	1.01
Set of common items between 1 & 2	5	5	1,798	1.53	31	0.98
Accessibility	7	7	1,800	2.30	33	1.31

Table 5.17 Summary Statistics of the Raw Scores—Grade Ten

Form	N Items	N Points	N Students	Mean	Mean as % of Total	SD
1	18	18	600	6.47	36	2.65
2	20	21	600	5.91	28	2.59
Set of common items between 1 & 2	5	5	600	1.51	30	0.98
Set of common items between 1 & 2	5	5	600	1.53	31	0.94
Accessibility	7	7	794	2.20	31	1.24

Table 5.18 Summary Statistics of the Raw Scores—Grade Eleven

Form	N Items	N Points	N Students	Mean	Mean as % of Total	SD
1	18	18	599	7.09	39	2.88
2	20	21	599	6.56	31	2.85
Set of common items between 1 & 2	5	5	599	1.68	34	1.00
Set of common items between 1 & 2	5	5	599	1.69	34	1.00
Accessibility	7	7	212	2.15	31	1.23

Table 5.19 Summary Statistics of the Raw Scores—Grade Twelve

Form	N Items	N Points	N Students	Mean	Mean as % of Total	SD
1	18	18	600	5.92	33	2.67
2	20	21	599	5.23	25	2.41
Set of common items between 1 & 2	5	5	600	1.40	28	1.03
Set of common items between 1 & 2	5	5	599	1.37	27	0.98
Accessibility	7	7	794	2.43	35	1.38

Accessibility Information

Alternative Text for Figure 5.1

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

Chapter 6: Analyses

This chapter summarizes the results of the item- and test-level analyses on samples from the 2016–17 California Science Test (CAST) pilot test administration. Analyses include the following:

- Classical item analyses
- Form completion analyses
- Response time analyses
- Interrater reliability analyses
- Differential item functioning (DIF) analyses

Note that the analyses results of the scores are not intended for reporting but are used only for research and future test development.

6.1. Samples Used for Item Analyses

The same samples that were drawn for constructed-response (CR) scoring for the grade five and high school general forms were used for item analysis. These analyses allow item-total correlations to be calculated across the whole forms.

The same sampling procedure as that used for drawing the CR scoring samples for the grade five general forms (refer to subsection *5.1 Constructed Response (CR) Scoring Sampling Process*) was used for the grade five accessibility features form and all of the grade eight general forms, which did not have any approved CR items for scoring.

The same sampling procedure as that used for the high school general forms, detailed in subsection *5.1 Constructed Response (CR) Scoring Sampling Process*, was intended for use for the high school accessibility form as well. However, as seen in Table 6.1 and Table 6.2, the accessibility form of grade eleven—"HS Form A"— did not have sufficient numbers of students available, even from the combination of early-adopter and non-early-adopter LEA students, to reach the desired count of 600 students. Accordingly, all 212 available grade eleven students who were administered the accessibility form were selected, and the remaining 1,588 of the needed 1,800 were split between the other two grades, with first selecting any early-adopter local educational agency (LEA) students, and then randomly selecting from the available non-early-adopter LEAs.

Table 6.1 Composition of Student Samples for Item Analysis—Students Available as of May 9, 2017

Form	Early-Adopter LEAs	Non-Early-Adopter LEAs
Grade 5 Form A	661	16,746
Grade 8 Form 1	3,624	114,297
Grade 8 Form 2	3,591	113,617
Grade 8 Form A	349	9,206
HS Form A		
Grade 10	11	2,399
Grade 11	0	212
Grade 12	1	1,015

Table 6.2 Composition of Student Samples for Item Analysis—Students Sampled

Form	Early-Adopter LEAs	Non-Early-Adopter LEAs
Grade 5 Form A	661	1,139
Grade 8 Form 1	1,800	0
Grade 8 Form 2	1,800	0
Grade 8 Form A	349	1,451
HS Form A		
Grade 10	11	783
Grade 11	0	212
Grade 12	1	793

6.2. Classical Item Analyses

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

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

For dichotomous items, item difficulty is indicated by its *p*-value, which is the proportion of students who answer the item correctly. The range of *p*-values is from 0.00 to 1.00. Items with high *p*-values are easier items; those with low *p*-values are more difficult items. Dichotomous items are flagged for review if they have *p*-values above 0.95 (i.e., too easy) or below 0.10 (i.e., too difficult).

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

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

See the Alternative Text for Equation 6.1 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). The AIS can range from 0.00 to the maximum total possible points for an item. Desired AIS values for polytomous items generally fall within the range of 10 percent to 95 percent of the maximum obtainable item score; items with values outside this range are flagged for review. To facilitate the interpretation, the AIS values for polytomous items are often expressed as the proportion of the maximum possible score, which are equivalent to the *p*-values of dichotomous items.

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

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

See the Alternative Text for Equation 6.2 for a description of this equation.

where,

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

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

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

6.2.2 Item-Total Correlations

The item-total correlation statistic describes the relationship between students' performance on a specific item and its performance on the total test and 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).

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:

$$r_{polyreg} = \frac{\hat{\beta} s_{tot}}{\sqrt{\hat{\beta}^2 s_{tot}^2 + 1}} \quad (6.3)$$

See the Alternative Text for Equation 6.3 for a description of this equation.

where,

S_{tot} is the standard deviation of the students' total test scores as a criterion score, and

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

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

Desired polyserial correlation values of items are positive and larger than 0.20. A relatively high item-total correlation coefficient value is desired, as it indicates that students with higher total raw scores on the overall test tend to perform better on the item than students with lower total raw scores. However, an item with a negative item-total correlation typically signifies a problem with the item, as that indicates that (1) the higher-ability students on the

overall test tend to respond incorrectly to the item (if dichotomous) or are assigned a low score for the item (if polytomous), or (2) the lower-ability students on the overall test are responding correctly to the item (if dichotomous) or are assigned a high score for that item (if polytomous).

6.2.3 Distribution of Item Scores

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

6.2.4 Omission

An item is considered "omitted" if it was seen but was not answered (i.e., it was left blank) in the middle of an administered assessment. Another way of saying this is that the student viewed an item without responding but moved on to respond to successive items.

6.2.5 Distractor Analyses

6.2.5.1. The Proportion of Students Choosing Each Distractor

For the CAST, distractor analyses were conducted on multiple-choice (MC) items (i.e., items that were not constructed-response items). 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 mis-keyed).

6.2.5.2. Distractor-Total Correlation

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

6.2.6 Summary of Classical Item Analyses Flagging Criteria

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

- **Difficulty flags** indicate extreme values of the proportion-correct (for dichotomous items) or the proportion of the possible maximum points earned (for polytomous items).
  - A value less than 0.1 suggests that the item might be too difficult.
  - A value greater than 0.95 suggests that the item might be too easy.
- A **discrimination flag** indicates that the item does not discriminate effectively between high- and low-ability students. Items with a polyserial correlation less than 0.20 are flagged.

- An **omit flag** is set for dichotomous items with nonresponse rates greater than five percent and polytomous items with nonresponse rates greater than 20 percent.
- A **distractor flag** is used for any distractors having positive correlation with the criterion score.
- A **mis-key flag** is used for MC items when more of the high-ability examinee group—the top 20 percent of examinees on the total test—choose any distractor rather than choosing the response keyed as correct.
- The **underrepresented score point flag** is used for any item that has less than three percent of the students at any score level.

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

6.2.7 Classical Item Analysis Results Summary

The summary statistics of the item analyses, which include the means and medians of *p*-values and item-total correlations, are presented in Table 6.3 for all items in each form.

Table 6.3 Mean and Median Proportion Correct (*p*-value) and Item Correlation (Polyserial Correlation) in Each Form

Grade	Form	N Items	N Students	*p*-value	Mean Polyserial	Median *p*-value	Median Polyserial
5	1	19	1,800	0.42	0.53	0.42	0.54
5	2	18	1,800	0.46	0.56	0.45	0.60
5	A	7	1,800	0.40	0.57	0.34	0.59
8	1	21	1,800	0.38	0.44	0.38	0.43
8	2	20	1,800	0.39	0.42	0.38	0.43
8	A	6	1,800	0.29	0.58	0.31	0.58
HS	1	18	1,799	0.38	0.45	0.37	0.46
HS	2	20	1,799	0.30	0.43	0.31	0.40
HS	A	7	1,800	0.33	0.53	0.35	0.52

The results of item analyses for each item in each form are presented in Table 6.A.1 through Table 6.A.9 for dichotomous items and Table 6.A.10 for polytomous items. The item statistics, including the AIS for polytomous items or *p*-value for dichotomous items, polyserial correlation, omit rates, and the distribution of score points on each polytomous item are listed in those tables.

Table 6.A.11 through Table 6.A.19 present the summary statistics of *p*-values by item types—such as MC, technology-enhanced (TE), and CR items as shown in Table 3.2 of subsection *3.5 Item Types and Features*—including the minimum, mean, and maximum, as well as the number of items in low, medium, and high range of the *p*-values in each form. It seems that the TE items are more difficult on average than MC items within each form. Table 6.A.20 through Table 6.A.28 present the summary statistics of polyserial correlations by item types.

Table 6.A.29 through Table 6.A.34 provide the summary of the *p*-values and item-total correlations respectively by grades (ten, eleven, and twelve) for the high school assessment. Table 6.A.29 through Table 6.A.31 show that, on the two general forms, the average *p*-value of grade eleven is slightly higher than the average *p*-value for grade ten, which, in turn, is slightly higher than the average *p*-value for grade twelve. It suggests that, given the same test form, performance by students in grade eleven is slightly better than that of students in either grade ten or grade twelve for the analysis sample.

6.3. Form Completion

Completion rates indicate the proportion of students who completed the whole test. For the purposes of this short pilot, a student's record for the test was not considered complete unless the student answered all items.

Table 6.4 presents the number and percent of students who answered all items in each form in the full test-taker population for each tested grade. The completion rate was over 99 percent for each form. Table 6.B.1 through Table 6.B.18 present the number and percent of students in selected demographic student groups (i.e., the groups listed in Table 4.4) who answered all items.

Table 6.4 Form Completion Rates

Grade	Form	Percent Completion	Count of Completion	Total Count (N)
5	1	99.54%	216,688	217,688
5	2	99.60%	217,320	218,185
5	A	99.90%	36,700	36,735
8	1	99.15%	216,004	217,866
8	2	99.05%	214,892	216,951
8	A	99.83%	18,089	18,119
10	1	99.07%	181,327	183,033
10	2	99.22%	182,494	183,930
10	A	99.82%	4,977	4,986
11	1	99.10%	21,152	21,344
11	2	99.01%	21,256	21,468
11	A	99.36%	467	470
12	1	99.20%	20,668	20,834
12	2	99.37%	20,577	20,708
12	A	99.92%	2,444	2,446
All HS	1	99.08%	223,147	225,211
All HS	2	99.21%	224,327	226,106
All HS	A	99.82%	7,888	7,902

6.4. Response Time Analyses

The length of time it took students to complete a test was recorded and analyzed to build a profile describing what a typical testing event looked like for each test. In addition, variability in testing time was investigated to determine whether a student's testing time should be

viewed as unusual or irregular for further investigation. It should be noted that the CAST assessments, including the pilot forms, are untimed.

In these analyses, all students are considered. Only 104 students across all grade levels who have zero testing time due to system reset are removed from the analysis. The remaining testing population is partitioned into quartiles based on raw scores on the MC items, because only a sample of students had their CR item responses scored.

The descriptive statistics—for example, the number of students, mean, standard deviation, minimum and maximum, percentiles—of the time required to complete the total test are computed for each of the four quartile groups for each tested grade (e.g., grades five, eight, ten, eleven, and twelve).

Some cases of extremely long testing time may be attributed to students with special needs who took longer to complete the tests, or the test not being closed down properly. Given the presence of a few outlying, extreme values, the results should be interpreted with caution. Mean testing times should not be interpreted as typical. Rather, the medians (i.e., 50th percentile) are more meaningful in the interpretation of the time comparisons because medians are less impacted by the extreme values than means.

Appendix 6.C summarizes results of testing time analysis. Table 6.C.1 provides descriptive statistics of total testing time for the full student population at each ability level. The unit of testing time is the minute; for example, in Table 6.C.1, the median (i.e., 50th percentile) of the testing time for form 1 of the grade five Q1 group is 32.59 minutes. Overall, based on the medians of total testing time from grade five and grade eight tests, students in the highest ability level (4th quartile, Q4) have shorter testing times than students in the other groups.

However, for students who took the high school test, students in the lowest ability level (1st quartile, Q1) have shorter testing times than students in the other groups, and the median of total testing time generally increases with ability level from Q1 to Q4. This trend can be observed within each high school grade (i.e., grades ten, eleven, and twelve). On the two general forms of each test, students at the 50th percentile within each ability quartile spent 29 to 35 minutes on the grade five test, 31 to 35 minutes on grade eight test, and 18 to 30 minutes on the high school test across grades ten, eleven, and twelve. Accessible forms are much shorter and thus the time spent on them is not comparable to time spent on the general forms.

Table 6.C.2 provides the testing time statistics of the performance tasks (PTs) for the overall population at each ability level. Each grade level had a six-item PT shared between two general forms. This table includes the mean and standard deviation, the minimum and maximum of the PT time, as well as percentile information at each ability level. The unit of testing time is a minute. Based on the median (i.e., 50th percentile) of the testing time, students who took the grade five and grade eight tests spent about nine minutes on the PT while students who took the high school test spent about four minutes on the PT. In general, high-performing students (i.e., Q3 and Q4) tended to take a little more time on PTs than low-performing students (i.e., Q1 and Q2).

Table 6.C.3 through Table 6.C.5 present the average item time by item types for each form. Only discrete items that are not affiliated with any PTs are included in this analysis. For MC items, students who took the general forms of the grade five and grade eight tests spent about one minute on average for each MC item, while high school students took about 0.7 minute to finish an MC item, on average, in general forms. Grades ten and eleven students

spent a little longer on MC items than grade twelve students. Students who took the accessibility features pilot forms spent only half of the time, on average, on MC items than students who took the general forms. These data are presented in Table 6.C.3.

For TE items, the average time spent on each TE item of the general forms was about two minutes; the amount of time decreased from grades ten to twelve for TE items on the general forms. Students who took accessibility feature pilot forms spent a little longer on each TE item than students who took the general forms. These data are presented in Table 6.C.4.

A composite group is a group of items of different types (e.g., one MC item and one TE item) appearing on the same page of the assessment. Because testing time was recorded for every page, the time for the composite groups cannot be partitioned into any item types. These data are presented in Table 6.C.5.

6.5. Interrater Reliability Analyses

To monitor the consistency of ratings assigned to students' responses by human raters, approximately 33 percent of the human-scored CR responses received a second rating ("backreading"); the responses in this subsample were randomly selected and scored by two raters. The two sets of ratings are used to compute statistics describing the consistency (reliability) of the human ratings. This interrater consistency is described in three ways:

1. percentage agreement between two human raters,
2. Cohen's Kappa, and
3. Quadratic-weighted Kappa coefficient.

6.5.1. 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.

6.5.2. Kappa

Interrater reliability or consistency is an indicator of homogeneity in scoring and is measured using an intraclass correlation (ICC) that incorporates the exact agreement between raters over and above that expected by chance. The index is defined as:

$$ICC = r_I = \left(ms_{between} - ms_{within}\right) / \left(ms_{between} + \left[k - 1\right] ms_{within}\right) \quad (6.4)$$

See the Alternative Text for Equation 6.4 for a description of this equation.

where,

$ms_{between}$ is the mean-square estimate of between-subjects variance,

ms_{within} is the mean-square estimate of within-subjects variance, and

k is the number of classes, in this case, the number of raters.

For categorical ratings, Cohen's Kappa statistic (1960) has the properties of an ICC and can be used for interrater reliability. The definition of Cohen's Kappa is:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{6.5}$$

See the Alternative Text for Equation 6.5 for a description of this equation.

where,

p_o is the relative observed agreement among raters, and

p_e is the hypothetical probability of chance agreement, in which the observed data are used to calculate the probabilities of each observer randomly seeing each category m:

$$p_e = \frac{1}{N^2}\sum_m n_{m1} n_{m2} \tag{6.6}$$

See the Alternative Text for Equation 6.6 for a description of this equation.

where,

N is the number of items,

n_{m1} and n_{m2} are the number of times rater 1 and rater 2 provide a score of category m respectively. Cohen's Kappa ranges from 0.0 to 1.0, with complete agreement being equal to 1.0.

6.5.3. Quadratic-weighted Kappa

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

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

$$\kappa_{st} = \frac{\left(\sum_{s=0}^{m}\sum_{t=0}^{m} w_{st}\frac{n_{st}}{n_{++}}\right) - \left(\sum_{s=0}^{m}\sum_{t=0}^{m} w_{st}\frac{n_{s+}n_{+t}}{n_{++}^2}\right)}{1 - \left(\sum_{s=0}^{m}\sum_{t=0}^{m} w_{st}\frac{n_{s+}n_{+t}}{n_{++}^2}\right)} \tag{6.7}$$

See the Alternative Text for Equation 6.7 for a description of this equation.

For Quadratic-weighted Kappa, the weights are:

$$w_{st} = 1 - \frac{(s-t)^2}{m^2} \tag{6.8}$$

See the Alternative Text for Equation 6.8 for a description of this equation.

6.5.4. Results

Table 6.5 presents the results of the interrater analyses and descriptive statistics of the ratings by the two raters on CR items, including the following:

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

All four items—three items in the grade five test and one item in the high school test—that were human-scored are 1-point items. As such, Cohen's Kappa equals Quadratic-weighted Kappa for 1-point items; percent of adjacent agreement is not presented, because with only 0 or 1 point, the adjacent agreement is 100 percent.

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

Given the criteria mentioned, the results of these four items in Table 6.5 show excellent interrater reliability, with Quadratic-weighted Kappa ranging from 0.77 to 0.83. The interrater agreement is also high, with the percent of exact agreement ranging from 88.6 percent to 91.5 percent.

Note the following about the data presented in Table 6.5:

1. The item analysis sample, rather than the full population, was used.
2. Only responses that were scored 0 or 1 are included, and condition codes are excluded.
3. Cohen's Kappa and Quadratic-weighted Kappa are equal for items with 1 point.
4. Rater 2 is the backup rater.

Table 6.5 Interrater Reliability for Human-Scored Constructed-Response (CR) Items

Grade	Form	Item Identification (ID)	Score Points	Quadratic-weighted Kappa (0,1 only)	Percent Exact	Rater 1 Total Count	Rater 1 Mean	Rater 1 SD	Rater 2 Total Count	Rater 2 Mean	Rater 2 SD
5	1&2	VH668026	1	0.772	88.6	3484	0.501	0.500	1141	0.499	0.500
5	1&2	VH667949	1	0.784	89.4	3533	0.434	0.496	1198	0.432	0.495
5	1	VH689376	1	0.771	88.8	1757	0.586	0.493	574	0.556	0.497
HS	1&2	VH651810	1	0.827	91.5	3436	0.450	0.497	1159	0.427	0.495

6.6. Differential Item Functioning (DIF) Analyses

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 other than the intended construct. An item may be biased if it contains content or language that is differentially familiar to groups of students (item 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 (item impact) or statistical Type I error, which might falsely assert DIF exists for an item. As a result, DIF statistics are used to identify *potential* item bias. Subsequent reviews by content experts and bias/sensitivity experts are required to determine the source and meaning of performance differences.

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

6.6.1. DIF Procedure for Dichotomous Items

The Mantel-Haenszel (MH) DIF statistic was calculated for dichotomous items. For this method, students are classified to relevant student groups of interest (e.g., gender or ethnicity). Students at each total score level in the focal group (e.g., females) are compared with examinees at each total score level in the reference group (e.g., males). The common odds ratio—that is, the proportion of correct response over the proportion of incorrect response—is estimated across all levels of matched student ability using the formula in Equation 6.9 (Dorans & Holland, 1993). The resulting estimate is interpreted as the relative probability of success on a particular item for members of two groups when matched on ability.

$$\alpha_{MH} = \frac{\left(\sum_m R_{rm} \frac{W_{fm}}{N_{tm}}\right)}{\left(\sum_m R_{fm} \frac{W_{rm}}{N_{tm}}\right)} \tag{6.9}$$

See the Alternative Text for Equation 6.9 for a description of this equation.

where,

m = the number of score categories of the total test,

R_{rm} = the number of students in the reference group who answer the item correctly,

W_{fm} = the number of students in the focal group who answer the item incorrectly,

R_{fm} = the number of students in the focal group who answer the item correctly,

W_{rm} = the number of students in the reference group who answer the item incorrectly, and

N_{tm} = the total number of students.

To facilitate the interpretation of MH results, the common odds ratio is frequently transformed to the delta scale using the following formula (Holland & Thayer, 1988):

$$MH\ D\text{-}DIF = -2.35 \ln\left[\alpha_{MH}\right] \quad (6.10)$$

See the Alternative Text for Equation 6.10 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).

6.6.2. 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:

$$SMD = \frac{\sum_{m=1}^{M} N_{fm} \times E_f(Y \mid X = m)}{\sum_{m=1}^{M} N_{fm}} - \frac{\sum_{m=1}^{M} N_{fm} \times E_r(Y \mid X = m)}{\sum_{m=1}^{M} N_{fm}} = \frac{\sum_{m=1}^{M} D_m}{\sum_{m=1}^{M} N_{fm}} \quad (6.11)$$

See the Alternative Text for Equation 6.11 for a description of this equation.

where,

X = the criterion score (total raw score),

Y = the item score,

M = the number of score levels on X,

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

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

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

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

E_f = the expected item score for the focal group.

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

6.6.3. Classification

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

The flagging criteria for dichotomous items are presented in Table 6.6; the flagging criteria for polytomous items are provided in Table 6.7.

Table 6.6 DIF Categories for Dichotomous Items

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

Table 6.7 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 or 0.17< |SMD/SD| ≤ 0.25
C (large)	• Mantel Chi-square *p value* > 0.05 or |SMD/*SD*| > 0.25

Note: SMD = standardized DIF; SD = total group standard deviation of item score.

DIF analyses were conducted on each test for designated comparison groups defined on the basis of demographic variables, including gender, race/ethnicity, and primary disabilities. These comparison groups are specified in Table 6.8. Note that the comparison between the high school grades was intended to examine whether items performed differently across high school grade level and that grade ten is designated as the reference group. However, none of the grade levels were regarded as advantaged or disadvantaged groups.

Table 6.8 Student Groups for DIF Comparison

Group Type	Focal Group	Reference Group
Gender	Female	Male
Ethnicity	American Indian/Alaska Native	White
Ethnicity	Asian	White
Ethnicity	Black or African American	White
Ethnicity	Filipino	White
Ethnicity	Hispanic/Latino	White
Ethnicity	Native Hawaiian or Other Pacific Islander	White
English fluency	English learner	English only
Disability	Special education services	No special education services
Economic status	Economically disadvantaged	Not economically disadvantaged
High School Grade Level	Grade 11	Grade 10
High School Grade Level	Grade 12	Grade 10

6.6.4. Items Exhibiting Significant DIF

Appendix 6.D provides detailed DIF results. Table 6.D.1 through Table 6.D.3 list the DIF classification for every item in each test. The items that exhibit significant DIF (C-DIF) are highlighted. Test developers are instructed to avoid selecting items flagged as showing DIF that disadvantages a focal group (negative C-DIF) for future test forms unless their inclusion is deemed essential to meeting test-content specifications.

Table 6.D.4 through Table 6.D.88 show the distributions of items across the DIF category classifications for each form of the CAST pilot. Data in the *N* column show the number of item occurrences with sufficient sample sizes to be included in DIF analyses. In addition, "Small N" indicates that the DIF analysis was not performed due to insufficient sample size.

References

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

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.

Holland, P. W., & Thayer, D. T. (1988). *An alternative definition of the ETS delta scale of item difficulty* (Research Report 85–43). Princeton, NJ: Educational Testing Service.

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.

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

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

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

Accessibility Information

Alternative Text for Equation 6.1

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 6.2

P value sub poly equals the fraction with the numerator X sub ij and the denominator N sub i times Max of X sub I end fraction.

Alternative Text for Equation 6.3

R sub polyreg equals the fraction Beta sub hat times S tot divided by the square root of beta sub hat squared times s sub tot squared plus 1n

Alternative Text for Equation 6.4

ICC equals r sub I equals open parenthesis ms sub between minus ms sub within close parenthesis divided by open parenthesis ms sub between plus open bracket k minus 1 close bracket tiems ms sub within close parenthesis.

Alternative Text for Equation 6.5

Kappa equals the fraction p sub o minus p sub e divided by 1 minus p sub en.

Alternative Text for Equation 6.6

P sub e equals the reciprocal of N squared times the summation of n sub m1 times n sub m2 from 1 to mn.

Alternative Text for Equation 6.7

K sub ij equals open parenthesis the sum from i equals zero to m the sum from j equals zero to m of w sub ij times n sub ij divided by n sub plus plus close parenthesis minus open parenthesis the sum from i equals zero to m the sum from j equals zero to m of w sub ij times n sub iplus times n sub plusj divided by n squared sub plusplus close parenthesis divided open parenthesis 1 minus open parenthesis the sum from i equals zero to m the sum from j equals zero to m of w sub ij times n sub iplus times n sub plusj divided by n squared sub plusplus close parenthesis close parenthesis.

Alternative Text for Equation 6.8

W sub st equals 1 minus open parenthesis s minus t close parenthesis squared divided by m squared.

Alternative Text for Equation 6.9

Alpha sub MH equals open parenthesis the sum from m of R sub rm times W sub fm divided by N sub tm close parenthesis divided by open parenthesis the sum from m of R sub fm times W sub rm divided by N sub tm close parenthesis.

Alternative Text for Equation 6.10

MH D - DIF equals negative 2.35 times the natural log of open bracket alpha sub MH close bracket.

Alternative Text for Equation 6.11

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 suf fm end fraction.

Appendix 6.A Classical Item Analyses

Notes:

1. An “r” indicates polyserial correlation, which is a statistical index of the item-total correlation.
2. “A” indicates the accessibility feature form.
3. A hyphen (“-”) in Table 6.A.1 through Table 6.A.9 indicates polytomous items, the statistics that are presented in Table 6.A.10.

Table 6.A.1 Form 1 Item-by-Item *p*-value, Point Biserial (Pt-Bis), and Omit Rate of Dichotomous Items—Grade Five

Item Sequence	*p*-value	Pt-Bis	Omit Rate
1	0.38	0.39	0.00
2	0.25	0.26	0.00
3	0.66	0.48	0.00
4	0.42	0.51	0.00
5	0.56	0.41	0.00
6	0.68	0.42	0.00
7	0.57	0.47	0.00
8	0.58	0.49	0.00
9	0.28	0.51	0.00
10	0.46	0.41	0.00
11	0.38	0.35	0.00
12	0.22	0.30	0.00
13	0.16	0.26	0.00
14	0.74	0.42	0.00
15	0.41	0.52	0.00
16	0.49	0.55	0.00
17	0.58	0.54	0.00
18	0.15	0.30	0.00
19	0.06	0.25	0.00
20			
21			

Table 6.A.2 Form 1 Item-by-Item *p*-value, Point Biserial (Pt-Bis), and Omit Rate of Dichotomous Items—Grade Eight

Item Sequence	*p*-value	Pt-Bis	Omit Rate
1	0.52	0.44	0.00
2	0.33	0.48	0.00
3	0.40	0.29	0.00
4	0.22	0.16	0.00
5	0.15	0.30	0.00
6	0.58	0.38	0.00
7	0.43	0.50	0.00
8	0.38	0.35	0.00
9	0.31	0.24	0.00
10	0.33	0.23	0.00
11	0.39	0.33	0.00
12	0.61	0.30	0.00
13	0.20	0.29	0.00
14	0.40	0.40	0.00
15	0.25	0.19	0.00
16	0.15	0.48	0.00
17	0.14	0.33	0.00
18	-	-	-
19	0.59	0.45	0.00
20	0.76	0.37	0.00
21	0.38	0.37	0.00

Table 6.A.3 Form 1 Item-by-Item *p*-value, Point Biserial (Pt-Bis), and Omit Rate of Dichotomous Items—High School

Item Sequence	*p*-value	Pt-Bis	Omit Rate
1	0.58	0.38	0.00
2	0.24	0.25	0.00
3	0.17	0.28	0.00
4	0.48	0.37	0.00
5	0.33	0.42	0.00
6	0.24	0.27	0.00
7	0.45	0.48	0.00
8	0.30	0.26	0.00
9	0.45	0.34	0.00
10	0.42	0.13	0.00
11	0.68	0.46	0.00
12	0.36	0.41	0.00

Item Sequence	*p*-value	Pt-Bis	Omit Rate
13	0.28	0.39	0.00
14	0.43	0.46	0.00
15	0.65	0.50	0.00
16	0.14	0.21	0.00
17	0.37	0.35	0.00
18	0.36	0.39	0.00
19			
20			
21			

Table 6.A.4 Form 2 Item-by-Item *p*-value, Point Biserial (Pt-Bis), and Omit Rate of Dichotomous Items—Grade Five

Item Sequence	*p*-value	Pt-Bis	Omit Rate
1	0.83	0.40	0.00
2	0.79	0.48	0.00
3	0.09	0.20	0.00
4	0.79	0.43	0.00
5	0.65	0.48	0.00
6	0.37	0.26	0.00
7	0.46	0.48	0.00
8	-	-	-
9	0.40	0.39	0.00
10	0.46	0.42	0.00
11	0.29	0.21	0.00
12	0.41	0.35	0.00
13	0.71	0.43	0.00
14	0.44	0.58	0.00
15	0.48	0.54	0.00
16	0.57	0.58	0.00
17	0.15	0.33	0.00
18	0.09	0.39	0.00
19			
20			

Table 6.A.5 Form 2 Item-by-Item *p*-value, Point Biserial (Pt-Bis), and Omit Rate of Dichotomous Items—Grade Eight

Item Sequence	*p*-value	Pt-Bis	Omit Rate
1	0.23	0.32	0.00
2	0.51	0.38	0.00
3	0.38	0.22	0.00
4	0.17	0.26	0.00
5	0.57	0.47	0.00
6	0.32	0.19	0.00
7	0.55	0.44	0.00
8	0.59	0.35	0.00
9	0.27	0.28	0.00
10	0.27	0.09	0.00
11	0.42	0.31	0.00
12	0.23	0.24	0.00
13	0.38	0.36	0.00
14	0.45	0.44	0.00
15	0.15	0.47	0.00
16	0.11	0.33	0.00
17	-	-	-
18	0.59	0.43	0.00
19	0.71	0.36	0.00
20	0.37	0.33	0.00

Table 6.A.6 Form 2 Item-by-Item *p*-value, Point Biserial (Pt-Bis), and Omit Rate of Dichotomous Items—High School

Item Sequence	*p*-value	Pt-Bis	Omit Rate
1	0.22	0.25	0.00
2	0.19	0.36	0.00
3	0.47	0.37	0.00
4	0.16	0.41	0.00
5	0.46	0.27	0.00
6	0.41	0.28	0.00
7	0.51	0.39	0.00
8	0.29	0.31	0.00
9	0.08	0.19	0.00
10	0.05	0.34	0.00
11	-	-	-
12	0.31	0.30	0.00
13	0.33	0.47	0.00

Item Sequence	*p*-value	Pt-Bis	Omit Rate
14	0.13	0.24	0.00
15	0.15	0.26	0.00
16	0.44	0.45	0.00
17	0.68	0.43	0.00
18	0.14	0.18	0.00
19	0.35	0.34	0.00
20	0.36	0.37	0.00

Table 6.A.7 Accessibility Features Form Item-by-Item *p*-value, Point Biserial (Pt-Bis), and Omit Rate of Dichotomous Items—Grade Five

Item Sequence	*p*-value	Pt-Bis	Omit Rate
1	0.30	0.40	0.00
2	0.24	0.39	0.00
3	0.51	0.49	0.00
4	0.62	0.53	0.00
5	0.48	0.48	0.00
6	0.34	0.38	0.00
7	0.32	0.49	0.00

Table 6.A.8 Accessibility Features Form Item-by-Item *p*-value, Point Biserial (Pt-Bis), and Omit Rate of Dichotomous Items—Grade Eight

Item Sequence	*p*-value	Pt-Bis	Omit Rate
1	0.35	0.51	0.00
2	0.20	0.46	0.00
3	0.33	0.46	0.00
4	0.16	0.39	0.00
5	0.39	0.56	0.00
6	0.29	0.36	0.00
7			

Table 6.A.9 Accessibility Features Form Item-by-Item *p*-value, Point Biserial (Pt-Bis), and Omit Rate of Dichotomous Items—High School

Item Sequence	*p*-value	Pt-Bis	Omit Rate
1	0.52	0.47	0.00
2	0.21	0.35	0.00
3	0.13	0.31	0.00
4	0.39	0.45	0.00
5	0.35	0.41	0.00
6	0.41	0.47	0.00
7	0.28	0.40	0.00

Table 6.A.10 Item-by-Item Proportion of Students at Each Score Point, Polyserial Correlation, and Omit Rate of Polytomous Items

Item Sequence	Grade	Form	Score Points	r	Omit Rate	Proportion of 0 Point	Proportion of 1 Point	Proportion of 2 Points
8	5	2	3	0.73	0.00	0.72	0.04	0.24
18	8	1	3	0.33	0.00	0.22	0.57	0.21
19	8	2	3	0.33	0.00	0.21	0.59	0.20
11	HS	2	3	0.32	0.00	0.49	0.41	0.10

Table 6.A.11 Summary of *p*-values for Multiple Choice Items for Each Test Form—Grade Five

Form	Low (0, 0.3)	Medium (0.3, 0.7)	High (0.7, 1)	Min	Max	Mean
1	3	6	1	0.15	0.74	0.43
2	2	5	4	0.15	0.83	0.52
A	1	5	0	0.24	0.62	0.39

Table 6.A.12 Summary of *p*-values for Multiple Choice Items for Each Test Form—Grade Eight

Form	Low (0, 0.3)	Medium (0.3, 0.7)	High (0.7, 1)	Min	Max	Mean
1	2	8	0	0.22	0.61	0.37
2	5	8	0	0.17	0.57	0.37
A	3	2	0	0.16	0.39	0.28

Table 6.A.13 Summary of *p*-values for Multiple Choice Items for Each Test Form—High School

Form	Low (0, 0.3)	Medium (0.3, 0.7)	High (0.7, 1)	Min	Max	Mean
1	2	7	0	0.17	0.68	0.42
2	3	7	0	0.13	0.51	0.35
A	2	4	0	0.13	0.52	0.35

Table 6.A.14 Summary of *p*-values for Constructed-Response Items for Each Test Form—Grade Five

Form	Low (0, 0.3)	Medium (0.3, 0.7)	High (0.7, 1)	Min	Max	Mean
1	0	3	0	0.41	0.57	0.49
2	0	2	0	0.44	0.48	0.46
A	0	0	0	-	-	-

Table 6.A.15 Summary of *p*-values for Constructed-Response Items for Each Test Form—Grade Eight

Form	Low (0, 0.3)	Medium (0.3, 0.7)	High (0.7, 1)	Min	Max	Mean
1	0	0	0	-	-	-
2	0	0	0	-	-	-
A	0	0	0	-	-	-

Table 6.A.16 Summary of *p*-values for Constructed-Response Items for Each Test Form—High School

Form	Low (0, 0.3)	Medium (0.3, 0.7)	High (0.7, 1)	Min	Max	Mean
1	0	1	0	0.43	0.43	0.43
2	0	1	0	0.44	0.44	0.44
A	0	0	0	-	-	-

Table 6.A.17 Summary of *p*-values for Technology-Enhanced Items for Each Test Form—Grade Five

Form	Low (0, 0.3)	Medium (0.3, 0.7)	High (0.7, 1)	Min	Max	Mean
1	3	3	0	0.06	0.68	0.37
2	3	2	0	0.09	0.65	0.31
A	0	1	0	0.48	0.48	0.48

Table 6.A.18 Summary of *p*-values for Technology-Enhanced Items for Each Test Form—Grade Eight

Form	Low (0, 0.3)	Medium (0.3, 0.7)	High (0.7, 1)	Min	Max	Mean
1	4	6	1	0.14	0.76	0.39
2	2	4	1	0.11	0.71	0.43
A	0	1	0	0.33	0.33	0.33

Table 6.A.19 Summary of *p*-values for Technology-Enhanced Items for Each Test Form—High School

Form	Low (0, 0.3)	Medium (0.3, 0.7)	High (0.7, 1)	Min	Max	Mean
1	4	4	0	0.14	0.65	0.34
2	6	3	0	0.05	0.68	0.23
A	1	0	0	0.21	0.21	0.21

Table 6.A.20 Summary of Polyserial Correlations for Multiple-Choice Items for Each Test Form—Grade Five

Form	Low (0, 0.3)	Medium (0.3, 0.7)	High (0.7, 1)	Min	Max	Mean
1	0	10	0	0.34	0.66	0.50
2	1	8	2	0.26	0.71	0.53
A	0	6	0	0.48	0.69	0.57

Table 6.A.21 Summary of Polyserial Correlations for Multiple-Choice Items for Each Test Form—Grade Eight

Form	Low (0, 0.3)	Medium (0.3, 0.7)	High (0.7, 1)	Min	Max	Mean
1	4	6	0	0.22	0.59	0.39
2	3	10	0	0.11	0.60	0.39
A	0	5	0	0.46	0.69	0.58

Table 6.A.22 Summary of Polyserial Correlations for Multiple-Choice Items for Each Test Form—High School

Form	Low (0, 0.3)	Medium (0.3, 0.7)	High (0.7, 1)	Min	Max	Mean
1	0	9	0	0.38	0.62	0.49
2	0	10	0	0.32	0.58	0.41
A	0	6	0	0.47	0.59	0.54

Table 6.A.23 Summary of Polyserial Correlations for Constructed-Response Items for Each Test Form—Grade Five

Form	Low (0, 0.3)	Medium (0.3, 0.7)	High (0.7, 1)	Min	Max	Mean
1	0	3	0	0.58	0.67	0.62
2	0	2	0	0.66	0.70	0.68
A	0	0	0	-	-	-

Table 6.A.24 Summary of Polyserial Correlations for Constructed-Response Items for Each Test Form—Grade Eight

Form	Low (0, 0.3)	Medium (0.3, 0.7)	High (0.7, 1)	Min	Max	Mean
1	0	0	0	-	-	-
2	0	0	0	-	-	-
A	0	0	0	-	-	-

Table 6.A.25 Summary of Polyserial Correlations for Constructed-Response Items for Each Test Form—High School

Form	Low (0, 0.3)	Medium (0.3, 0.7)	High (0.7, 1)	Min	Max	Mean
1	0	1	0	0.57	0.57	0.57
2	0	1	0	0.56	0.56	0.56
A	0	0	0	-	-	-

Table 6.A.26 Summary of Polyserial Correlations for Technology-Enhanced Response Items for Each Test Form—Grade Five

Form	Low (0, 0.3)	Medium (0.3, 0.7)	High (0.7, 1)	Min	Max	Mean
1	0	6	0	0.40	0.64	0.53
2	0	4	1	0.33	0.73	0.57
A	0	1	0	0.59	0.59	0.59

Table 6.A.27 Summary of Polyserial Correlations for Technology-Enhanced Response Items for Each Test Form—Grade Eight

Form	Low (0, 0.3)	Medium (0.3, 0.7)	High (0.7, 1)	Min	Max	Mean
1	0	11	0	0.33	0.65	0.48
2	0	7	0	0.33	0.65	0.48
A	0	1	0	0.58	0.58	0.58

Table 6.A.28 Summary of Polyserial Correlations for Technology-Enhanced Response Items for Each Test Form—High School

Form	Low (0, 0.3)	Medium (0.3, 0.7)	High (0.7, 1)	Min	Max	Mean
1	1	7	0	0.17	0.65	0.39
2	2	7	0	0.26	0.62	0.43
A	0	1	0	0.48	0.48	0.48

Table 6.A.29 Summary of *p*-values for Items for Grade Ten on the High School Test

Form	Low (0, 0.3)	Medium (0.3, 0.7)	High (0.7, 1)	Min	Max	Mean	SD	N
1	6	12	0	0.15	0.68	0.38	0.15	600
2	9	11	0	0.05	0.69	0.30	0.17	600
A	3	4	0	0.11	0.49	0.31	0.13	794

Table 6.A.30 Summary of *p*-values for Items for Grade Eleven on the High School Test

Form	Low (0, 0.3)	Medium (0.3, 0.7)	High (0.7, 1)	Min	Max	Mean	SD	N
1	5	11	2	0.15	0.75	0.42	0.17	599
2	9	10	1	0.08	0.76	0.34	0.18	599
A	3	4	0	0.10	0.48	0.31	0.12	212

Table 6.A.31 Summary of *p*-values for Items for Grade Twelve on the High School Test

Form	Low (0, 0.3)	Medium (0.3, 0.7)	High (0.7, 1)	Min	Max	Mean	SD	N
1	7	11	0	0.13	0.62	0.35	0.14	600
2	10	10	0	0.02	0.59	0.27	0.15	600
A	2	5	0	0.15	0.56	0.35	0.14	794

Table 6.A.32 Summary of Polyserial Correlations for Items for Grade Ten on the High School Test

Form	Low (0, 0.3)	Medium (0.3, 0.7)	High (0.7, 1)	Min	Max	Mean	SD	N
1	2	16	0	0.29	0.58	0.43	0.10	600
2	3	17	0	0.22	0.61	0.41	0.12	600
A	0	7	0	0.43	0.60	0.51	0.06	794

Table 6.A.33 Summary of Polyserial Correlations for Items for Grade Eleven on the High School Test

Form	Low (0, 0.3)	Medium (0.3, 0.7)	High (0.7, 1)	Min	Max	Mean	SD	N
1	1	17	0	0.13	0.64	0.47	0.13	599
2	2	18	0	0.27	0.68	0.44	0.11	599
A	0	7	0	0.43	0.57	0.51	0.04	212

Table 6.A.34 Summary of Polyserial Correlations for Items for Grade Twelve on the High School Test

Form	Low (0, 0.3)	Medium (0.3, 0.7)	High (0.7, 1)	Min	Max	Mean	SD	N
1	2	15	1	0.14	0.70	0.44	0.14	600
2	3	17	0	0.21	0.57	0.40	0.11	600
A	0	7	0	0.49	0.63	0.55	0.05	794

Appendix 6.B Form Completion

Notes:

- The total numbers of registered students are derived from version 2 of the production data file ("P2") that was delivered on August 29, 2017.
- A test is considered to be complete if a student answered all items.
- High school includes grades ten, eleven, and twelve.

Table 6.B.1 CAST Form Completion by Student Groups for Grade Five—Gender

Form	All	Male	Female
Total N	470708	240,252	230,456
Total %	99.60	99.62	99.58
Form 1 N	216,688	109,487	107,201
Form 1 %	99.54	99.54	99.54
Form 2 N	217,320	109,337	107,983
Form 2 %	99.60	99.63	99.58
Accessibility Form N	36,700	21,428	15,272
Accessibility Form %	99.90	99.92	99.88

Table 6.B.2 CAST Form Completion by Student Groups for Grade Five—English Language Fluency

Form	All	English Learner (EL)	English Only (EO)	Reclassified Fluent English Proficient (RFEP)	Initially Fluent English Proficient (IFEP)	To Be Determined (TBD)	No Response
Total N	470,708	98,889	265,018	86,917	18,949	273	662
Total %	99.60	99.51	99.61	99.66	99.71	97.50	97.50
Form 1 N	216,688	41,846	123,511	41,789	9,122	109	311
Form 1 %	99.54	99.35	99.57	99.63	99.63	98.20	97.49
Form 2 N	217,320	41,473	124,316	41,757	9,344	133	297
Form 2 %	99.60	99.52	99.61	99.67	99.78	96.38	97.06
Accessibility Form N	36,700	15,570	17,191	3,371	483	31	54
Accessibility Form %	99.90	99.92	99.88	99.94	100.00	100.00	100.00

Table 6.B.3 CAST Form Completion by Student Groups for Grade Five—Economic Status

Form	All	Economically Disadvantaged	Not Economically Disadvantaged
Total N	470,708	291,358	179,350
Total %	99.60	99.58	99.64
Form 1 N	216,688	131,397	85,291
Form 1 %	99.54	99.51	99.59
Form 2 N	217,320	131,319	86,001
Form 2 %	99.60	99.57	99.66
Accessibility Form N	36,700	28,642	8,058
Accessibility Form %	99.90	99.91	99.89

Table 6.B.4 CAST Form Completion by Student Groups for Grade Five—Ethnicity

Form	American Indian/ Alaskan Native	Asian	Pacific Islander	Filipino	Hispanic or Latino	African American	White	Two or More Races	Unknown
Total %	2,352	42,639	2,132	10,032	261,034	25,242	107,200	17,097	2,980
Form 1 N	99.16	99.80	99.72	99.74	99.58	99.42	99.59	99.64	99.27
Form 1 %	1,080	20,646	1,009	4,869	117,870	11,331	50,544	7,972	1,367
Form 2 N	98.90	99.80	99.61	99.73	99.51	99.37	99.54	99.65	99.06
Form 2 %	1,063	20,581	986	4,785	117,879	11,509	50,791	8,288	1,438
Accessibility Form N	99.35	99.80	99.80	99.73	99.58	99.40	99.61	99.59	99.38
Accessibility Form %	209	1,412	137	378	25,285	2,402	5,865	837	175
Total %	99.52	99.79	100.00	100.00	99.92	99.79	99.88	100.00	100.00

Table 6.B.5 CAST Form Completion by Student Groups for Grade Five—Disability

Form	Special Education Services	No Special Education Services
Total %	54,916	415,792
Form 1 N	99.48	99.61
Form 1 %	17,359	199,329
Form 2 N	99.14	99.58
Form 2 %	17,600	199,720
Accessibility Form N	99.37	99.62
Accessibility Form %	19,957	16,743
Total %	99.88	99.93

Table 6.B.6 CAST Form Completion by Student Groups for Grade Five—Migrant Status

Form	Migrant	Not Migrant
Total %	4,161	466,547
Form 1 N	99.43	99.60
Form 1 %	1,855	214,833
Form 2 N	99.46	99.54
Form 2 %	1,803	215,517
Accessibility Form N	99.23	99.61
Accessibility Form %	503	36,197
Total %	100.00	99.90

Table 6.B.7 CAST Form Completion by Student Groups for Grade Eight—Gender

Form	All	Male	Female
Total %	448,985	229,303	219,682
Form 1 N	99.13	99.12	99.14
Form 1 %	216,004	109,315	106,689
Form 2 N	99.15	99.12	99.17
Form 2 %	214,892	108,908	105,984
Accessibility Form N	99.05	99.04	99.06
Accessibility Form %	18,089	11,080	7,009
Total %	99.83	99.86	99.80

Table 6.B.8 CAST Form Completion by Student Groups for Grade Eight—English Language Fluency

Form	All	EL	EO	RFEP	IFEP	TBD	No Response
Total %	448,985	53,904	243,860	128,277	22,142	267	535
Form 1 N	99.13	98.83	99.23	99.05	99.30	96.04	97.81
Form 1 %	216,004	23,861	117,826	62,967	10,946	126	278
Form 2 N	99.15	98.78	99.22	99.12	99.31	99.21	97.89
Form 2 %	214,892	23,706	117,366	62,527	10,927	125	241
Accessibility Form N	99.05	98.62	99.18	98.95	99.27	92.59	97.57
Accessibility Form %	18,089	6,337	8,668	2,783	269	16	16
Total %	99.83	99.78	99.90	99.78	99.63	100.00	100.00

Table 6.B.9 CAST Form Completion by Student Groups for Grade Eight—Economic Status

Form	All	Economically Disadvantaged	Not Economically Disadvantaged
Total %	448,985	265,908	183,077
Form 1 N	99.13	98.96	99.37
Form 1 %	216,004	126,252	89,752
Form 2 N	99.15	98.98	99.38
Form 2 %	214,892	125,795	89,097
Accessibility Form N	99.05	98.86	99.33
Accessibility Form %	18,089	13,861	4,228
Total %	99.83	99.82	99.88

Table 6.B.10 CAST Form Completion by Student Groups for Grade Eight—Ethnicity

Form	American Indian/ Alaskan Native	Asian	Pacific Islander	Filipino	Hispanic or Latino	African American	White	Two or More Races	Unknown
Total %	2,340	43,036	2,185	11,713	240,646	24,724	108,283	13,441	2,617
Form 1 N	99.07	99.51	99.00	99.63	98.96	98.59	99.41	99.25	99.05
Form 1 %	1,152	21,076	1,092	5,804	114,340	11,722	52,926	6,623	1,269
Form 2 N	98.97	99.57	98.91	99.55	98.98	98.52	99.43	99.27	98.91
Form 2 %	1,062	21,262	1,021	5,765	114,335	11,569	52,143	6,478	1,257
Accessibility Form N	99.07	99.44	99.03	99.71	98.85	98.49	99.37	99.20	99.13
Accessibility Form %	126	698	72	144	11,971	1,433	3,214	340	91
Total %	100.00	100.00	100.00	100.00	99.79	99.93	99.88	100.00	100.00

Table 6.B.11 CAST Form Completion by Student Groups for Grade Eight—Primary Disability

Form	No Special Education Services	Special Education Services
Total %	47,000	401,985
Form 1 N	98.99	99.14
Form 1 %	17,293	198,711
Form 2 N	98.76	99.18
Form 2 %	17,207	197,685
Accessibility Form N	98.61	99.09
Accessibility Form %	12,500	5,589
Total %	99.83	99.84

Table 6.B.12 CAST Form Completion by Student Groups for Grade Eight—Migrant Status

Form	Migrant	Not Migrant
Total %	3,525	445,460
Form 1 N	99.21	99.13
Form 1 %	1,646	214,358
Form 2 N	99.10	99.15
Form 2 %	1,636	213,256
Accessibility Form N	99.21	99.05
Accessibility Form %	243	17,846
Total %	100.00	99.83

Table 6.B.13 CAST Form Completion by Student Groups for High School—Gender

Form	All	Male	Female
Total %	455,362	234,867	220,495
Form 1 N	99.16	99.17	99.15
Form 1 %	223,147	114,864	108,283
Form 2 N	99.08	99.10	99.07
Form 2 %	224,327	115,536	108,791
Accessibility Form N	99.21	99.21	99.22
Accessibility Form %	7,888	4,467	3,421
Total %	99.82	99.80	99.85

Table 6.B.14 CAST Form Completion by Student Groups for High School—English Language Fluency

Form	All	EL	EO	RFEP	IFEP	TBD	No Response
Total %	455,362	48,998	242,278	138,455	24,580	274	777
Form 1 N	99.16	98.65	99.27	99.14	99.31	96.82	96.16
Form 1 %	223,147	23,280	118,914	68,208	12,213	132	400
Form 2 N	99.08	98.55	99.19	99.07	99.29	97.06	95.92
Form 2 %	224,327	23,380	119,922	68,434	12,088	140	363
Accessibility Form N	99.21	98.63	99.34	99.19	99.32	96.55	96.54
Accessibility Form %	7,888	2,338	3,442	1,813	279	2	14
Total %	99.82	99.74	99.85	99.94	99.64	100.00	93.33

Table 6.B.15 CAST Form Completion by Student Groups for High School—Economic Status

Form	All	Economically Disadvantaged	Not Economically Disadvantaged
Total %	455,362	263,057	192,305
Form 1 N	99.16	99.03	99.34
Form 1 %	223,147	128,557	94,590
Form 2 N	99.08	98.92	99.30
Form 2 %	224,327	129,137	95,190
Accessibility Form N	99.21	99.10	99.36
Accessibility Form %	7,888	5,363	2,525
Total %	99.82	99.80	99.88

Table 6.B.16 CAST Form Completion by Student Groups for High School—Ethnicity

Form	American Indian/ Alaskan Native	Asian	Pacific Islander	Filipino	Hispanic or Latino	African American	White	Two or More Races	Unknown
Form 1 N	2,604	42,147	2,188	12,721	244,331	26,025	108,843	12,933	3,416
Form 1 %	99.47	99.38	98.56	99.31	99.07	98.64	99.41	99.29	98.22
Form 2 N	1,259	20,891	1,092	6,333	119,039	12,683	53,667	6,448	1,653
Form 2 %	99.45	99.38	98.03	99.19	98.99	98.39	99.38	99.22	97.87
Accessibility Form N	1,313	20,927	1,070	6,195	119,818	12,949	53,944	6,323	1,716
Accessibility Form %	99.47	99.37	99.07	99.41	99.11	98.85	99.44	99.36	98.51
Total %	32	329	26	193	5,474	393	1,232	162	47
Form 1 N	100.00	99.70	100.00	100.00	99.78	100.00	99.84	100.00	100.00

Table 6.B.17 CAST Form Completion by Student Groups for High School—Disability

Form	No Special Education Services	Special Education Services
Form 1 N	42,181	413,181
Form 1 %	98.83	99.19
Form 2 N	19,278	203,869
Form 2 %	98.76	99.11
Accessibility Form N	19,672	204,655
Accessibility Form %	98.76	99.26
Total %	3,231	4,657
Form 1 N	99.75	99.87

Table 6.B.18 CAST Form Completion by Student Groups for High School—Migrant Status

Form	Migrant	Not Migrant
Form 1 N	3,132	452,230
Form 1 %	99.27	99.16
Form 2 N	1,487	221,660
Form 2 %	98.87	99.08
Accessibility Form N	1,477	222,850
Accessibility Form %	99.60	99.21
Total %	168	7,720
Form 1 N	100.00	99.82

Appendix 6.C Response Time Analysis

Note the following about Table 6.C.1:

- Raw scores for machine scorable items were used to partition students into quartiles. All students who tested and have unrounded test time greater than 0 are included.
- Form 1 and Form 2 are general forms while A indicates accessibility features pilot forms.

Table 6.C.1 Total Testing Time (in Minutes) at Each Raw Score Interval

Grade	Form	Raw Score Interval	Descriptive N	Descriptive Mean	Descriptive Standard Deviation (SD)	Descriptive Minium	Descriptive Maximum	Percentile Point 1	Percentile Point 10	Percentile Point 25	Percentile Point 50	Percentile Point 75	Percentile Point 90	Percentile Point 99
5	1	Q 1	33,226	36.07	19.61	0.06	337.13	6.17	16.00	23.25	32.59	44.63	59.31	102.25
5	1	Q 2	74,126	38.37	18.70	2.34	282.87	10.68	20.08	26.09	34.62	46.06	60.66	104.04
5	1	Q 3	47,979	37.80	17.85	2.91	379.63	13.94	20.95	26.22	33.95	44.69	58.53	101.18
5	1	Q 4	62,344	34.27	16.27	1.99	309.97	13.67	19.36	23.86	30.57	40.19	52.92	93.36
5	2	Q 1	36,086	35.67	20.06	0.06	348.97	5.29	15.20	22.77	32.16	44.38	59.49	103.13
5	2	Q 2	47,146	37.85	19.14	1.67	333.13	9.25	18.95	25.37	34.14	45.81	60.69	103.76
5	2	Q 3	72,557	37.18	18.11	1.71	335.19	12.85	20.06	25.42	33.24	44.25	58.28	102.03
5	2	Q 4	62,384	32.87	16.02	2.09	299.89	12.86	18.34	22.71	29.23	38.47	50.95	91.10
5	A	Q 1	8,449	14.75	10.89	0.08	148.71	1.00	4.78	7.99	12.21	18.42	27.07	55.26
5	A	Q 2	9,302	15.43	10.77	0.35	143.22	1.60	5.91	8.91	12.80	18.69	27.60	57.96
5	A	Q 3	8,800	15.68	10.57	0.36	151.93	2.03	6.65	9.24	13.09	19.03	27.42	55.24
5	A	Q 4	10,177	15.46	10.42	0.45	187.00	3.54	6.94	9.12	12.58	18.38	26.97	55.94

Grade	Form	Raw Score Interval	Descriptive N	Descriptive Mean	Descriptive Standard Deviation (SD)	Descriptive Minium	Descriptive Maximum	Percentile Point 1	Percentile Point 10	Percentile Point 25	Percentile Point 50	Percentile Point 75	Percentile Point 90	Percentile Point 99
8	**1**	Q 1	38,524	33.90	17.07	0.05	277.70	5.32	14.76	22.60	31.71	42.36	54.79	86.88
8	**1**	Q 2	48,403	35.81	16.22	2.28	387.59	8.06	18.47	25.14	33.32	43.66	55.66	88.11
8	**1**	Q 3	69,792	36.35	14.94	1.90	263.95	12.05	20.90	26.39	33.75	43.31	54.73	85.31
8	**1**	Q 4	61,141	34.15	12.91	1.86	199.46	14.94	21.05	25.50	31.71	39.79	49.97	77.33
8	**2**	Q 1	35,660	36.60	17.67	0.04	248.47	6.24	17.03	25.04	34.14	45.46	58.38	93.14
8	**2**	Q 2	49,112	37.21	16.62	1.76	253.18	7.88	19.23	26.20	34.73	45.50	57.82	90.46
8	**2**	Q 3	72,256	37.54	15.74	0.93	303.38	11.49	21.43	27.16	34.82	44.85	56.73	88.66
8	**2**	Q 4	59,913	35.76	13.74	2.45	290.73	15.44	21.90	26.66	33.16	41.68	52.47	81.55
8	**A**	Q 1	2,871	13.34	9.55	0.14	182.87	1.08	4.89	7.53	11.34	16.51	23.63	49.89
8	**A**	Q 2	5,793	13.30	8.76	0.18	145.67	1.15	4.97	7.83	11.33	16.54	23.67	45.54
8	**A**	Q 3	0	-	-	-	-	-	-	-	-	-	-	-
8	**A**	Q 4	9,447	13.20	7.97	0.25	151.02	1.99	5.86	8.17	11.32	16.09	22.80	41.16
HS 10	**1**	Q 1	41,986	27.93	13.36	0.05	149.58	4.84	12.52	18.68	26.35	35.10	44.62	68.38
HS 10	**1**	Q 2	24,554	29.61	12.58	2.25	154.20	7.49	15.37	20.99	28.04	36.30	45.22	67.81
HS 10	**1**	Q 3	51,267	30.69	12.04	2.33	364.89	9.56	17.40	22.53	29.07	36.98	45.72	66.83
HS 10	**1**	Q 4	65,216	31.52	11.11	2.48	205.60	13.04	19.56	23.98	29.86	37.11	45.26	65.76
HS 10	**2**	Q 1	31,792	25.61	12.58	0.07	167.33	4.29	11.46	16.87	23.89	32.32	41.47	64.24
HS 10	**2**	Q 2	54,758	26.39	12.03	1.32	186.53	5.62	12.93	18.17	24.72	32.70	41.54	64.06
HS 10	**2**	Q 3	48,990	26.92	11.31	2.34	150.97	7.22	14.53	19.18	25.27	32.85	41.21	61.74
HS 10	**2**	Q 4	48,374	27.11	10.37	2.39	194.24	10.35	16.15	20.04	25.39	32.17	40.11	59.76

Grade	Form	Raw Score Interval	Descriptive N	Descriptive Mean	Descriptive Standard Deviation (SD)	Descriptive Minium	Descriptive Maximum	Percentile Point 1	Percentile Point 10	Percentile Point 25	Percentile Point 50	Percentile Point 75	Percentile Point 90	Percentile Point 99
HS 10	**A**	Q 1	366	14.38	9.39	0.92	61.62	1.32	4.63	8.00	12.14	18.79	25.05	47.52
HS 10	**A**	Q 2	1,032	14.38	8.84	0.72	68.80	1.56	5.27	8.63	12.95	17.93	24.45	47.93
HS 10	**A**	Q 3	1,485	14.66	9.20	0.60	127.69	1.58	5.80	8.92	12.93	18.36	24.81	44.67
HS 10	**A**	Q 4	2,102	14.00	8.44	0.63	108.16	2.21	6.60	9.13	12.40	16.66	23.01	44.41
HS 11	**1**	Q 1	4,428	25.37	15.69	0.10	157.37	2.99	8.22	14.42	22.79	32.76	44.64	75.75
HS 11	**1**	Q 2	5,484	28.76	13.91	2.42	167.61	5.76	13.18	19.37	26.83	35.52	45.47	73.24
HS 11	**1**	Q 3	5,326	30.65	13.23	3.20	134.73	8.76	16.60	21.84	28.60	36.76	46.82	75.72
HS 11	**1**	Q 4	6,102	31.64	12.38	5.06	154.83	12.55	18.75	23.33	29.59	37.33	46.84	72.73
HS 11	**2**	Q 1	3,369	23.23	14.21	0.13	150.93	2.95	7.79	13.28	20.80	30.01	40.81	69.80
HS 11	**2**	Q 2	5,826	24.97	13.31	1.33	140.59	3.73	10.34	16.00	22.83	31.65	41.33	69.47
HS 11	**2**	Q 3	5,467	26.82	12.78	1.76	118.33	4.90	13.23	18.14	24.83	32.74	42.74	68.43
HS 11	**2**	Q 4	6,806	27.38	11.33	0.00	108.48	9.02	15.76	19.81	25.32	32.52	41.47	65.97
HS 11	**A**	Q 1	38	16.41	12.13	1.00	51.13	1.00	3.35	7.27	14.84	18.83	37.25	51.13
HS 11	**A**	Q 2	108	15.12	11.80	0.42	56.34	0.81	2.90	4.71	13.32	21.88	30.40	52.05
HS 11	**A**	Q 3	138	15.62	11.74	2.07	61.49	2.48	4.09	7.35	11.55	20.91	30.60	51.90
HS 11	**A**	Q 4	186	15.40	10.99	1.37	90.39	1.64	5.21	8.36	12.96	20.63	27.70	60.01
HS 12	**1**	Q 1	4,173	20.70	13.24	0.10	157.29	2.82	6.61	11.25	18.39	26.90	37.04	63.32
HS 12	**1**	Q 2	5,971	24.32	13.43	1.64	128.13	4.34	9.59	15.05	22.10	30.98	40.99	69.64
HS 12	**1**	Q 3	2,998	26.88	13.77	2.12	162.73	5.75	12.72	17.61	24.64	33.03	43.81	71.76
HS 12	**1**	Q 4	7,685	29.72	13.61	3.80	187.72	8.52	15.98	20.86	27.35	35.48	45.62	77.48

Grade	Form	Raw Score Interval	Descriptive N	Descriptive Mean	Descriptive Standard Deviation (SD)	Descriptive Minium	Descriptive Maximum	Percentile Point 1	Percentile Point 10	Percentile Point 25	Percentile Point 50	Percentile Point 75	Percentile Point 90	Percentile Point 99
HS 12	**2**	Q 1	4,957	20.36	12.46	0.11	148.03	2.97	7.07	11.53	18.12	26.28	36.16	61.52
HS 12	**2**	Q 2	3,597	21.69	12.43	1.42	126.62	3.32	7.89	12.88	19.72	27.97	37.41	59.78
HS 12	**2**	Q 3	6,566	22.93	12.74	1.99	146.00	3.65	8.99	14.12	20.93	28.97	38.48	65.45
HS 12	**2**	Q 4	5,578	25.31	12.54	1.83	177.27	5.28	12.54	16.99	23.41	30.73	40.15	67.61
HS 12	**A**	Q 1	607	10.32	6.65	0.57	60.17	0.91	3.10	5.85	9.61	12.79	18.15	32.06
HS 12	**A**	Q 2	0	-	-	-	-	-	-	-	-	-	-	-
HS 12	**A**	Q 3	648	11.53	6.46	1.01	58.29	1.37	4.68	7.48	10.58	14.45	19.40	32.10
HS 12	**A**	Q 4	1,191	11.15	5.79	0.79	60.11	1.55	5.38	7.55	10.09	13.49	17.70	32.38

Note the following about Table 6.C.2:

- Raw scores for machine scorable items were used to partition students into quartiles. All students who tested and have unrounded test time greater than 0 are included.
- Performance tasks (PTs) only exist in general forms but not in accessibility features pilot forms.

Table 6.C.2 Performance Task Testing Time (in Minutes) at Each Raw Score Interval Using Total Raw Score for General Forms

Grade	Form	Raw Score Interval	Descriptive N	Descriptive Mean	Descriptive Standard Deviation	Descriptive Min	Descriptive Max	Percentile Point 1	Percentile Point 10	Percentile Point 25	Percentile Point 50	Percentile Point 75	Percentile Point 90	Percentile Point 99
5	1	Q 1	33,226	9.15	6.72	0.00	161.57	0.00	2.74	4.78	7.76	11.78	16.78	32.80
5	1	Q 2	74,126	10.40	6.68	0.00	115.07	1.46	3.98	6.08	9.06	13.00	18.02	34.22
5	1	Q 3	47,979	11.02	6.57	0.00	235.95	2.38	4.86	6.77	9.64	13.56	18.43	34.66
5	1	Q 4	62,344	10.82	6.22	0.00	129.92	2.94	4.95	6.71	9.45	13.26	18.03	33.49
5	2	Q 1	36,086	9.89	7.39	0.00	189.55	0.80	2.99	5.11	8.37	12.73	18.18	35.73
5	2	Q 2	47,146	10.89	7.18	0.00	137.97	1.36	3.95	6.18	9.43	13.78	19.19	36.27
5	2	Q 3	72,557	11.45	6.91	0.00	156.42	2.18	4.89	6.98	10.03	14.14	19.32	35.99
5	2	Q 4	62,384	11.15	6.38	0.00	155.63	2.96	5.15	6.99	9.78	13.60	18.43	33.92
8	1	Q 1	38,524	8.25	5.12	0.00	113.16	0.00	3.03	4.82	7.46	10.56	14.19	25.00
8	1	Q 2	48,403	9.51	4.95	0.00	108.11	1.98	4.22	6.21	8.80	11.81	15.28	25.48
8	1	Q 3	69,792	10.50	4.70	0.00	131.17	2.71	5.61	7.50	9.79	12.63	16.00	25.57
8	1	Q 4	61,141	10.82	4.10	0.00	86.13	4.36	6.69	8.14	10.09	12.65	15.71	24.28
8	2	Q 1	35,660	8.10	4.97	0.00	87.46	0.00	3.00	4.78	7.33	10.43	14.01	24.69
8	2	Q 2	49,112	9.16	4.93	0.00	92.04	1.85	3.91	5.88	8.44	11.47	14.90	24.90
8	2	Q 3	72,256	10.29	4.77	0.00	113.77	2.47	5.29	7.28	9.63	12.47	15.76	25.55
8	2	Q 4	59,913	10.97	4.24	0.00	130.20	4.36	6.73	8.25	10.25	12.80	15.89	24.80

Grade	Form	Raw Score Interval	Descriptive N	Descriptive Mean	Descriptive Standard Deviation	Descriptive Min	Descriptive Max	Percentile Point 1	Percentile Point 10	Percentile Point 25	Percentile Point 50	Percentile Point 75	Percentile Point 90	Percentile Point 99
HS 10	**1**	Q 1	41,986	4.99	3.90	0.00	90.06	0.00	1.56	2.41	3.91	6.56	9.66	18.68
HS 10	**1**	Q 2	24,554	5.21	3.84	0.00	86.71	0.88	1.84	2.62	4.14	6.80	9.86	18.19
HS 10	**1**	Q 3	51,267	5.28	3.77	0.00	228.45	1.07	1.97	2.72	4.19	6.95	9.95	17.30
HS 10	**1**	Q 4	65,216	5.23	3.52	0.00	64.29	1.27	2.05	2.73	4.11	6.93	9.90	16.59
HS 10	**2**	Q 1	31,792	5.36	4.12	0.00	65.86	0.00	1.69	2.60	4.21	7.02	10.39	19.90
HS 10	**2**	Q 2	54,758	5.43	3.98	0.00	68.14	0.90	1.86	2.69	4.28	7.16	10.37	19.14
HS 10	**2**	Q 3	48,990	5.42	3.84	0.00	88.33	1.09	1.95	2.72	4.27	7.15	10.36	18.28
HS 10	**2**	Q 4	48,374	5.25	3.65	0.00	65.86	1.23	1.99	2.69	4.08	6.94	10.05	17.27
HS 11	**1**	Q 1	4,428	4.82	4.32	0.00	61.47	0.00	1.26	2.11	3.60	6.24	9.66	20.84
HS 11	**1**	Q 2	5,484	5.34	4.08	0.00	87.87	0.81	1.78	2.63	4.21	7.04	10.15	19.02
HS 11	**1**	Q 3	5,326	5.58	4.18	0.00	55.68	1.00	1.99	2.77	4.39	7.30	10.47	20.17
HS 11	**1**	Q 4	6,102	5.54	3.85	0.00	50.35	1.17	2.04	2.80	4.38	7.39	10.42	18.13
HS 11	**2**	Q 1	3,369	5.02	4.24	0.00	80.80	0.00	1.43	2.33	3.83	6.53	9.96	20.26
HS 11	**2**	Q 2	5,826	5.34	4.20	0.00	54.59	0.62	1.68	2.56	4.20	6.98	10.40	19.70
HS 11	**2**	Q 3	5,467	5.60	4.15	0.00	45.36	0.87	1.85	2.65	4.42	7.36	10.71	19.74
HS 11	**2**	Q 4	6,806	5.65	4.13	0.00	58.28	1.01	1.93	2.73	4.44	7.60	10.70	19.46
HS 12	**1**	Q 1	4,173	4.07	3.64	0.00	63.08	0.00	1.10	1.77	3.07	5.22	8.20	17.39
HS 12	**1**	Q 2	5,971	4.68	3.69	0.00	64.40	0.71	1.49	2.21	3.65	6.09	9.11	18.50
HS 12	**1**	Q 3	2,998	5.04	3.96	0.00	57.02	0.85	1.62	2.48	3.98	6.51	9.66	18.90
HS 12	**1**	Q 4	7,685	5.37	3.94	0.00	67.27	1.06	1.86	2.68	4.32	7.01	9.93	19.53

Grade	Form	Raw Score Interval	Descriptive N	Descriptive Mean	Descriptive Standard Deviation	Descriptive Min	Descriptive Max	Percentile Point 1	Percentile Point 10	Percentile Point 25	Percentile Point 50	Percentile Point 75	Percentile Point 90	Percentile Point 99
HS 12	**2**	Q 1	4,957	4.60	3.95	0.00	83.94	0.00	1.40	2.20	3.55	5.81	8.95	18.52
HS 12	**2**	Q 2	3,597	4.87	3.86	0.00	51.91	0.73	1.48	2.32	3.85	6.33	9.43	17.96
HS 12	**2**	Q 3	6,566	4.99	3.61	0.00	47.04	0.83	1.62	2.40	4.02	6.66	9.42	17.22
HS 12	**2**	Q 4	5,578	5.26	3.91	0.00	49.99	0.91	1.72	2.52	4.17	6.90	9.96	18.75

Note the following about Table 6.C.3, Table 6.C.4, and Table 6.C.5:

1. Discrete items indicate items that are not affiliated with a PT.
2. Items types that are not multiple choice, or extended response, or fill-in-blanks are categorized as technology-enhanced (TE) items. Refer to Table 3.2 in subsection *3.5 Item Types and Features*.
3. Testing time is recorded for every page in the test delivery system. While most pages contain only one item, there are pages that contain more than one item where each item belongs to a different item type—for example, one multiple-choice item and one TE item might be on the same page. This circumstance is considered a composite group.

Table 6.C.3 Average Item Time (in Minutes) of Discrete Items by Multiple Choice Items (excluding PT items) Using Machine-Scorable Raw Score and All Students

Grade	Form	N	Mean	SD	Min	Med	Max
5	1	217,675	0.90	0.61	0.00	0.75	13.81
5	2	218,173	1.11	0.69	0.00	0.94	12.49
5	A	36,728	0.55	0.39	0.00	0.45	7.42
8	1	217,860	1.12	0.64	0.00	0.98	17.06
8	2	216,941	1.18	0.63	0.00	1.05	16.50
8	A	18,111	0.42	0.30	0.01	0.34	6.72
10	1	183,023	0.79	0.40	0.00	0.74	8.21
10	2	183,914	0.74	0.40	0.00	0.67	9.22
10	A	4,985	0.36	0.24	0.01	0.31	3.43
11	1	21,340	0.77	0.46	0.00	0.70	6.68
11	2	21,468	0.72	0.45	0.00	0.63	7.14
11	A	470	0.39	0.33	0.01	0.31	2.55
12	1	20,827	0.67	0.44	0.00	0.59	8.19
12	2	20,698	0.64	0.43	0.00	0.56	7.43
12	A	2,446	0.27	0.17	0.01	0.24	1.88

Table 6.C.4 Average Item Time (in Minutes) of Discrete Items by Technology-Enhanced Items (excluding PT items) Using Machine-Scorable Raw Score and All Students

Grade	Form	N	Mean	SD	Min	Med	Max
5	1	217,675	1.87	1.22	0.00	1.58	45.43
5	2	218,173	2.05	1.25	0.00	1.74	45.31
5	A	36,728	2.14	1.99	0.00	1.64	62.63
8	1	217,860	1.90	0.95	0.00	1.71	21.88
8	2	216,941	1.51	1.51	0.00	1.10	62.69
8	A	18,111	1.85	1.37	0.00	1.54	45.19
10	1	183,023	2.45	1.06	0.00	2.30	29.18
10	2	183,914	1.92	0.87	0.00	1.81	22.76
10	A	4,985	3.03	2.31	0.00	2.49	39.77
11	1	21,340	2.33	1.14	0.00	2.16	15.90
11	2	21,468	1.85	0.94	0.00	1.73	13.03
11	A	470	3.27	2.87	0.08	2.45	25.35
12	1	20,827	2.07	1.17	0.00	1.90	20.57
12	2	20,698	1.59	0.96	0.00	1.45	18.45
12	A	2,446	2.42	1.67	0.10	2.05	24.21

Table 6.C.5 Average Item Time (in Minutes) of Discrete Items by Composite Group Items (excluding PT items) Using Machine-Scorable Raw Score and All Students

Grade	Form	N	Mean	SD	Min	Med	Max
5	1	217,675	3.52	2.41	0.00	2.99	56.98
5	2	-	-	-	-	-	-
5	A	-	-	-	-	-	-
8	1	-	-	-	-	-	-
8	2	216,941	3.26	2.24	0.00	2.73	71.73
8	A	-	-	-	-	-	-
10	1	-	-	-	-	-	-
10	2	-	-	-	-	-	-
10	A	-	-	-	-	-	-
11	1	-	-	-	-	-	-
11	2	-	-	-	-	-	-
11	A	-	-	-	-	-	-
12	1	-	-	-	-	-	-
12	2	-	-	-	-	-	-
12	A	-	-	-	-	-	-

Appendix 6.D Differential Item Functioning (DIF)

Note:

1. DIF is estimated based on an item analysis sample as is summarized in Table 5.2, rather than the full population.
2. A hyphen ("-") in Table 6.D.1 through Table 6.D.3 and "Small N" in Table 6.D.4 through Table 6.D.88 indicate that minimum sample size for DIF analysis was not met
3. An "A" indicates the accessibility features pilot forms. Refer to subsection *2.2 Test Assembly* for the details of the accessibility features pilot forms.
4. The items that exhibit significant DIF (C-DIF) are highlighted in gray.

Table 6.D.1 Grade Five DIF

Item ID	Item Sequence	Form	Male–Female	White–American Indian	White–Asian	White–Pacific Islander	White–Filipino	White–Hispanic	White–African American	English Only–English Learner	No Special Services–Special Services	Not Economically Disadvantaged–Economically Disadvantaged
VH640495	1	1	A-	-	A-	-	-	A-	A-	A+	A-	A-
VH640511	2	1	A+	-	A+	-	-	A+	A-	A+	A+	A+
VH640385	3	1	C-	-	A+	-	-	A-	A-	A-	A+	A-
VH644341	4	1	A-	-	A-	-	-	A-	A-	A-	A+	A-
VH644654	5	1	A+	-	A-	-	-	A-	A+	A-	A-	A-
VH689367	6	1	A+	-	A-	-	-	A+	A+	A+	A+	A+
VH689376	7	1	A+	-	A-	-	-	A+	A+	A-	A-	A+
VH640498	8	1	A-	-	B-	-	-	A-	A-	A-	A+	A-
VH642060	9	1	B-	-	A+	-	-	A-	B-	A-	A+	A-
VH647663	10	1	A-	-	A-	-	-	A-	A+	A-	A+	A-
VH644199	11	1	A-	-	A+	-	-	A+	A-	A+	A+	A+
VH640593	12	1	A-	-	B+	-	-	A+	A+	A+	A+	A+

Item ID	Item Sequence	Form	Male–Female	White–American Indian	White–Asian	White–Pacific Islander	White–Filipino	White–Hispanic	White–African American	English Only–English Learner	No Special Services–Special Services	Not Economically Disadvantaged–Economically Disadvantaged
VH640634	13	1	A+	-	A+	-	-	A-	B+	A-	B+	B-
VH659818	14	1	A+	-	A+	-	-	A+	A-	A+	A-	A+
VH667949	15	1	C+	-	A+	-	-	A+	A+	A+	A-	A+
VH668026	16	1	B+	-	A+	-	-	A+	A+	A+	A+	A+
VH667740	17	1	A+	-	A+	-	-	A-	A-	A-	A-	A-
VH684170	18	1	A-	-	A-	-	-	A-	A-	A-	A-	A+
VH690980	19	1	A-	-	A+	-	-	A+	A-	A+	A-	A+
VH640527	1	2	A+	-	A+	-	-	A-	A-	A-	B-	A-
VH640434	2	2	A-	-	A-	-	-	A-	A-	B-	B-	A-
VH642049	3	2	A+	-	A-	-	-	A-	A-	A-	A+	A-
VH644366	4	2	A-	-	A+	-	-	A+	A-	A+	A-	A+
VH642029	5	2	A-	-	A+	-	-	A-	A-	B-	A+	A-
VH647639	6	2	A+	-	A+	-	-	A+	A+	A+	A+	A+
VH640444	7	2	A-	-	A-	-	-	B-	B-	A-	A+	A-
VH659334	8	2	A-	-	A-	-	-	A-	A-	A+	A+	A+
VH630738	9	2	A-	-	A-	-	-	A-	A-	A-	A+	A-
VH644368	10	2	A-	-	A-	-	-	A+	A+	A+	A-	A+
VH640423	11	2	A-	-	A+	-	-	A+	A-	A+	A+	A+
VH644362	12	2	A+	-	A+	-	-	A+	A+	A+	A+	A+
VH659818	13	2	A-	-	A+	-	-	A+	A+	A+	A+	A-
VH667949	14	2	B+	-	A+	-	-	B+	A+	A+	A-	A+
VH668026	15	2	B+	-	C+	-	-	A+	A+	A+	A-	A+

Item ID	Item Sequence	Form	Male–Female	White–American Indian	White–Asian	White–Pacific Islander	White–Filipino	White–Hispanic	White–African American	English Only–English Learner	No Special Services–Special Services	Not Economically Disadvantaged–Economically Disadvantaged
VH667740	16	2	A+	-	A-	-	-	A+	B+	A-	A-	A-
VH684170	17	2	A+	-	A-	-	-	A+	A-	A-	A-	A+
VH690980	18	2	A-	-	B+	-	-	A-	A-	A+	A+	A-
VH647992	1	A	A-	-	-	-	-	A+	A+	A+	A+	A+
VH647999	2	A	A+	-	-	-	-	A+	A+	B+	A+	A+
VH648004	3	A	B-	-	-	-	-	B-	A-	A-	A+	B-
VH648006	4	A	A+	-	-	-	-	A-	A+	A-	B-	A-
VH648007	5	A	A-	-	-	-	-	A-	A-	A-	A-	A-
VH648009	6	A	A+	-	-	-	-	A+	A+	B+	A+	A+
VH648015	7	A	A+	-	-	-	-	A-	A+	A-	A+	A-

Table 6.D.2 Grade Eight DIF

Item ID	Item Sequence	Form	Male–Female	White–American Indian	White–Asian	White–Pacific Islander	White–Filipino	White–Hispanic	White–African American	English Only–English Learner	No Special Services–Special Services	Not Economically Disadvantaged–Economically Disadvantaged
VH641132	1	1	A-	-	A+	-	-	A-	A+	A-	A+	A-
VH642249	2	1	A+	-	A+	-	-	A-	A-	A+	A-	A-
VH648025	3	1	A+	-	A+	-	-	A+	A+	A+	A-	A+
VH646409	4	1	A-	-	A+	-	-	A+	A+	A+	C+	A+
VH650871	5	1	A-	-	A-	-	-	A+	A-	A+	A+	A-
VH642297	6	1	A-	-	A-	-	-	A-	A-	A-	A-	A+
VH649343	7	1	A-	-	A-	-	-	B-	C-	A-	A-	A-
VH646463	8	1	A-	-	A+	-	-	A-	A+	A-	A+	A-
VH646435	9	1	A-	-	A-	-	-	A+	B+	A+	A+	A+
VH646469	10	1	A-	-	A+	-	-	A-	A+	B+	A+	A+
VH650877	11	1	A-	-	C+	-	-	A+	A+	A+	A-	A+
VH642235	12	1	A+	-	B-	-	-	A-	C-	A-	A-	A+
VH649870	13	1	A+	-	A-	-	-	A+	A+	A+	A-	A-
VH642230	14	1	A+	-	A-	-	-	A+	A-	A+	A-	A+
VH646394	15	1	A-	-	A-	-	-	A+	B+	B+	A+	A+
VH655016	16	1	A+	-	A+	-	-	A-	A-	A-	A-	C-
VH649463	17	1	A+	-	A-	-	-	A-	A-	A-	A-	B-
VH655019	18	1	A+	-	A+	-	-	A+	A+	A+	A+	A+
VH649470	19	1	A+	-	A+	-	-	A+	A+	A-	B-	A+
VH649549	20	1	A-	-	A-	-	-	A+	A-	C-	A-	A+
VH655027	21	1	A+	-	A-	-	-	A+	A+	A-	A+	A+

Item ID	Item Sequence	Form	Male–Female	White–American Indian	White–Asian	White–Pacific Islander	White–Filipino	White–Hispanic	White–African American	English Only–English Learner	No Special Services–Special Services	Not Economically Disadvantaged–Economically Disadvantaged
VH642181	1	2	A-	-	A-	-	-	A+	A-	A+	A+	A+
VH642265	2	2	A+	-	A+	-	-	A+	A-	A+	A-	A+
VH642239	3	2	A-	-	A-	-	-	A-	A+	A-	A-	A+
VH646418	4	2	A-	-	A+	-	-	A+	A+	B+	A-	A+
VH642290	5	2	A+	-	B+	-	-	A+	A+	B-	A-	A-
VH642258	7	2	A-	-	B-	-	-	A+	A+	A+	B+	A+
VH642212	8	2	A-	-	B-	-	-	A-	A-	A-	A-	A-
VH655006	9	2	A+	-	A+	-	-	A-	A-	A+	A-	A-
VH642280	11	2	A+	-	A+	-	-	A+	A+	A+	A+	A+
VH642244	12	2	A-	-	A+	-	-	C+	B+	B+	A+	A+
VH646370	13	2	A+	-	A+	-	-	A-	A+	A-	A+	A-
VH642225	14	2	A+	-	A+	-	-	A+	A+	A+	A+	A+
VH646451	15	2	A-	-	A-	-	-	A+	A+	A-	A+	A+
VH642233	16	2	A-	-	A+	-	-	A-	A-	A-	A-	A-
VH655016	17	2	A-	-	A+	-	-	A-	A-	B-	A-	B-
VH649463	18	2	A-	-	A-	-	-	A+	A-	A-	A-	A-
VH655019	19	2	A-	-	A+	-	-	A+	A-	B+	A+	A-
VH649470	20	2	A+	-	A+	-	-	A-	A-	A-	A-	A-
VH649549	21	2	A-	-	A-	-	-	A-	A-	C-	A-	A-
VH655027	22	2	A+	-	A-	-	-	A-	A-	A-	A-	A+
VH648019	1	A	A-	-	-	-	-	A+	-	A-	A-	A-
VH648020	2	A	A+	-	-	-	-	A-	-	A+	A-	A+

Item ID	Item Sequence	Form	Male–Female	White–American Indian	White–Asian	White–Pacific Islander	White–Filipino	White–Hispanic	White–African American	English Only–English Learner	No Special Services–Special Services	Not Economically Disadvantaged–Economically Disadvantaged
VH648031	3	A	A+	-	-	-	-	A+	-	A+	A+	A-
VH648036	4	A	A-	-	-	-	-	A+	-	A+	A-	A+
VH648042	5	A	A-	-	-	-	-	A-	-	A-	A-	A-
VH648049	7	A	A+	-	-	-	-	A-	-	A+	A+	A+

Table 6.D.3 High School DIF

Item ID	Item Sequence	Form	Male–Female	White–American Indian	White–Asian	White–Pacific Islander	White–Filipino	White–Hispanic	White–African American	English Only–English Learner	No Special Services–Special Services	Not Economically Disadvantaged–Economically Disadvantaqed	Grade 10–Grade 11	Grade 10–Grade 12
VH631212	1	1	A+	-	A+	-	-	A+	-	A+	A+	A-	A-	A-
VH642091	2	1	A-	-	A+	-	-	A+	-	A+	A+	A-	A-	A+
VH640717	4	1	A+	-	A+	-	-	A-	-	A+	A+	A-	A-	A-
VH640714	5	1	A+	-	A-	-	-	A-	-	A+	A+	A-	A+	A+
VH631171	6	1	A-	-	A-	-	-	A-	-	A-	A-	A+	A-	A-
VH651858	7	1	A-	-	A+	-	-	A+	-	A+	A+	A+	A+	A+
VH640665	8	1	A+	-	A+	-	-	A-	-	A-	A-	A-	A+	A-
VH631223	9	1	A+	-	A-	-	-	A-	-	A+	A+	A+	A+	A+
VH631256	10	1	A+	-	A-	-	-	A-	-	A-	A-	A+	A-	A-
VH690993	11	1	A-	-	A-	-	-	A+	-	B+	A+	A+	B-	A+
VH631281	12	1	A-	-	A+	-	-	A+	-	A+	A+	A-	A-	A-
VH666105	13	1	A-	-	A+	-	-	A+	-	A+	A+	A+	A+	A+
VH631072	15	1	A+	-	B+	-	-	A+	-	A+	A+	A+	A-	A-
VH651810	16	1	A-	-	A+	-	-	A-	-	B-	B-	A-	A+	A-
VH651813	17	1	A+	-	A-	-	-	A-	-	C-	A-	A-	A+	A-
VH684305	19	1	A+	-	A-	-	-	A+	-	A-	A-	A-	A-	A+
VH684250	20	1	A-	-	A+	-	-	A+	-	A-	A+	A-	A+	B+
VH690982	21	1	A+	-	A-	-	-	A-	-	A+	A+	A-	A-	A-
VH631294	1	2	B-	-	A+	-	-	A+	A+	C+	A+	A+	A+	A+
VH641032	2	2	A+	-	A+	-	-	A+	A+	A-	A-	A+	A+	A-
VH640671	3	2	A+	-	A+	-	-	A+	A+	A-	A-	A+	A+	A+

Item ID	Item Sequence	Form	Male–Female	White–American Indian	White–Asian	White–Pacific Islander	White–Filipino	White–Hispanic	White–African American	English Only–English Learner	No Special Services–Special Services	Not Economically Disadvantaged–Economically Disadvantaged	Grade 10–Grade 11	Grade 10–Grade 12
VH651852	4	2	A+	-	B+	-	-	A-	B-	A+	A-	A+	A+	A-
VH642167	5	2	A+	-	A+	-	-	A+	B+	A+	A+	A+	A-	A+
VH631125	6	2	A+	-	A-	-	-	A+	A+	A-	A+	A-	A-	A-
VH640684	7	2	A-	-	A-	-	-	A-	A-	A-	A+	A-	A-	A-
VH642158	8	2	A-	-	A-	-	-	A-	A-	B+	A-	A+	A+	B+
VH640759	9	2	A-	-	A+	-	-	A+	A+	A+	A-	B+	A-	A-
VH649924	10	2	A+	-	A-	-	-	B-	A-	A-	A-	A-	A+	C-
VH654968	11	2	A+	-	A-	-	-	A+	A+	A+	A+	A+	A-	A+
VH631267	12	2	A-	-	A-	-	-	A-	A-	A+	A-	A+	A+	A-
VH647579	13	2	A-	-	A+	-	-	A-	A+	B-	A-	A-	A+	A-
VH631113	14	2	A-	-	A+	-	-	A+	B+	A+	A+	A+	A+	A+
VH642101	15	2	A+	-	A+	-	-	A+	A-	A+	A-	A+	A-	A-
VH651810	16	2	A-	-	A-	-	-	A+	A-	B-	A-	A-	A+	A+
VH651813	17	2	A-	-	A-	-	-	A-	B-	C-	A-	A-	A+	A-
VH684305	19	2	A-	-	A-	-	-	A-	A-	A+	B+	A-	A-	A+
VH684250	20	2	A-	-	A+	-	-	A+	A-	A+	A+	A+	A+	A+
VH690982	21	2	A-	-	C-	-	-	A-	A-	A+	A+	A-	A-	A+
VH648029	1	A	A-	-	-	-	-	A+	-	A-	A-	A+	A+	A+
VH648030	2	A	A-	-	-	-	-	A-	-	A-	A+	A-	A+	A-
VH648037	4	A	A+	-	-	-	-	A+	-	A+	A-	A-	A-	A+
VH648041	5	A	A+	-	-	-	-	A+	-	A+	A+	A+	A+	A+

Item ID	Item Sequence	Form	Male–Female	White–American Indian	White–Asian	White–Pacific Islander	White–Filipino	White–Hispanic	White–African American	English Only–English Learner	No Special Services–Special Services	Not Economically Disadvantaged–Economically Disadvantaqed	Grade 10–Grade 11	Grade 10–Grade 12
VH648048	6	A	A+	-	-	-	-	A-	-	A-	A+	A-	A+	A-
VH648051	7	A	A-	-	-	-	-	A-	-	A-	A-	A-	A-	A-
VH648052	8	A	A-	-	-	-	-	A+	-	A+	A-	A+	A+	A+

Table 6.D.4 DIF Classifications for Grade Five Form 1—Male–Female

DIF Category	N	Pct
C-	1	5
B-	1	5
A-	8	42
A+	7	37
B+	1	5
C+	1	5
Small N	0	0
Total	**19**	**100**

Table 6.D.5 DIF Classifications for Grade Five Form 1—White–American Indian

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	0	0
A+	0	0
B+	0	0
C+	0	0
Small N	19	100
Total	**19**	**100**

Table 6.D.6 DIF Classifications for Grade Five Form 1—White–Asian

DIF Category	N	Pct
C-	0	0
B-	1	5
A-	8	42
A+	9	47
B+	1	5
C+	0	0
Small N	0	0
Total	**19**	**100**

Table 6.D.7 DIF Classifications for Grade Five Form 1—White–Pacific Islander

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	0	0
A+	0	0
B+	0	0
C+	0	0
Small N	19	100
Total	**19**	**100**

Table 6.D.8 DIF Classifications for Grade Five Form 1—White–Filipino

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	0	0
A+	0	0
B+	0	0
C+	0	0
Small N	19	100
Total	**19**	**100**

Table 6.D.9 DIF Classifications for Grade Five Form 1—White–Hispanic

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	10	53
A+	9	47
B+	0	0
C+	0	0
Small N	0	0
Total	**19**	**100**

Table 6.D.10 DIF Classifications for Grade Five Form 1—White–African American

DIF Category	N	Pct
C-	0	0
B-	1	5
A-	10	53
A+	7	37
B+	1	5
C+	0	0
Small N	0	0
Total	**19**	**100**

Table 6.D.11 DIF Classifications for Grade Five Form 1—English Only–English Learner

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	10	53
A+	9	47
B+	0	0
C+	0	0
Small N	0	0
Total	**19**	**100**

Table 6.D.12 DIF Classifications for Grade Five Form 1—No Special Services–Special Services

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	8	42
A+	10	53
B+	1	5
C+	0	0
Small N	0	0
Total	**19**	**100**

Table 6.D.13 DIF Classifications for Grade Five Form 1—Not Economically Disadvantaged–Economically Disadvantaged

DIF Category	N	Pct
C-	0	0
B-	1	5
A-	8	42
A+	10	53
B+	0	0
C+	0	0
Small N	0	0
Total	**19**	**100**

Table 6.D.14 DIF Classifications for Grade Five Form 2—Male–Female

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	10	56
A+	6	33
B+	2	11
C+	0	0
Small N	0	0
Total	**18**	**100**

Table 6.D.15 DIF Classifications for Grade Five Form 2—White–American Indian

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	0	0
A+	0	0
B+	0	0
C+	0	0
Small N	18	100
Total	**18**	**100**

Table 6.D.16 DIF Classifications for Grade Five Form 2—White–Asian

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	8	44
A+	8	44
B+	1	6
C+	1	6
Small N	0	0
Total	**18**	**100**

Table 6.D.17 DIF Classifications for Grade Five Form 2—White–Pacific Islander

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	0	0
A+	0	0
B+	0	0
C+	0	0
Small N	18	100
Total	**18**	**100**

Table 6.D.18 DIF Classifications for Grade Five Form 2—White–Filipino

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	0	0
A+	0	0
B+	0	0
C+	0	0
Small N	18	100
Total	**18**	**100**

Table 6.D.19 DIF Classifications for Grade Five Form 2—White–Hispanic

DIF Category	N	Pct
C-	0	0
B-	1	6
A-	7	39
A+	9	50
B+	1	6
C+	0	0
Small N	0	0
Total	**18**	**100**

Table 6.D.20 DIF Classifications for Grade Five Form 2—White–African American

DIF Category	N	Pct
C-	0	0
B-	1	6
A-	10	56
A+	6	33
B+	1	6
C+	0	0
Small N	0	0
Total	**18**	**100**

Table 6.D.21 DIF Classifications for Grade Five Form 2—English Only—English Learner

DIF Category	N	Pct
C-	0	0
B-	2	11
A-	6	33
A+	10	56
B+	0	0
C+	0	0
Small N	0	0
Total	**18**	**100**

Table 6.D.22 DIF Classifications for Grade Five Form 2—No Special Services–Special Services

DIF Category	N	Pct
C-	0	0
B-	2	11
A-	6	33
A+	10	56
B+	0	0
C+	0	0
Small N	0	0
Total	**18**	**100**

Table 6.D.23 DIF Classifications for Grade Five Form 2—Not Economically Disadvantaged–Economically Disadvantaged

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	9	50
A+	9	50
B+	0	0
C+	0	0
Small N	0	0
Total	**18**	**100**

Table 6.D.24 DIF Classifications for Grade Five Accessibility Form—Male–Female

DIF Category	N	Pct
C-	0	0
B-	1	14
A-	2	29
A+	4	57
B+	0	0
C+	0	0
Small N	0	0
Total	**7**	**100**

Table 6.D.25 DIF Classifications for Grade Five Accessibility Form—White–American Indian

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	0	0
A+	0	0
B+	0	0
C+	0	0
Small N	7	100
Total	**7**	**100**

Table 6.D.26 DIF Classifications for Grade Five Accessibility Form—White–Asian

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	0	0
A+	0	0
B+	0	0
C+	0	0
Small N	7	100
Total	**7**	**100**

Table 6.D.27 DIF Classifications for Grade Five Accessibility Form—White–Pacific Islander

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	0	0
A+	0	0
B+	0	0
C+	0	0
Small N	7	100
Total	**7**	**100**

Table 6.D.28 DIF Classifications for Grade Five Accessibility Form—White–Filipino

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	0	0
A+	0	0
B+	0	0
C+	0	0
Small N	7	100
Total	**7**	**100**

Table 6.D.29 DIF Classifications for Grade Five Accessibility Form—White–Hispanic

DIF Category	N	Pct
C-	0	0
B-	1	14
A-	3	43
A+	3	43
B+	0	0
C+	0	0
Small N	0	0
Total	**7**	**100**

Table 6.D.30 DIF Classifications for Grade Five Accessibility Form—White–African American

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	2	29
A+	5	71
B+	0	0
C+	0	0
Small N	0	0
Total	**7**	**100**

Table 6.D.31 DIF Classifications for Grade Five Accessibility Form—English Only–English Learner

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	4	57
A+	1	14
B+	2	29
C+	0	0
Small N	0	0
Total	**7**	**100**

Table 6.D.32 DIF Classifications for Grade Five Accessibility Form—No Special Services–Special Services

DIF Category	N	Pct
C-	0	0
B-	1	14
A-	1	14
A+	5	71
B+	0	0
C+	0	0
Small N	0	0
Total	**7**	**100**

Table 6.D.33 DIF Classifications for Grade Five Accessibility Form—Not Economically Disadvantaged–Economically Disadvantaged

DIF Category	N	Pct
C-	0	0
B-	1	14
A-	3	43
A+	3	43
B+	0	0
C+	0	0
Small N	0	0
Total	**7**	**100**

Table 6.D.34 DIF Classifications for Grade Eight Form 1—Male–Female

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	11	52
A+	10	48
B+	0	0
C+	0	0
Small N	0	0
Total	**21**	**100**

Table 6.D.35 DIF Classifications for Grade Eight Form 1—White–American Indian

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	0	0
A+	0	0
B+	0	0
C+	0	0
Small N	21	100
Total	**21**	**100**

Table 6.D.36 DIF Classifications for Grade Eight Form 1—White–Asian

DIF Category	N	Pct
C-	0	0
B-	1	5
A-	10	48
A+	9	43
B+	0	0
C+	1	5
Small N	0	0
Total	**21**	**100**

Table 6.D.37 DIF Classifications for Grade Eight Form 1—White–Pacific Islander

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	0	0
A+	0	0
B+	0	0
C+	0	0
Small N	21	100
Total	**21**	**100**

Table 6.D.38 DIF Classifications for Grade Eight Form 1—White–Filipino

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	0	0
A+	0	0
B+	0	0
C+	0	0
Small N	21	100
Total	**21**	**100**

Table 6.D.39 DIF Classifications for Grade Eight Form 1—White–Hispanic

DIF Category	N	Pct
C-	0	0
B-	1	5
A-	8	38
A+	12	57
B+	0	0
C+	0	0
Small N	0	0
Total	**21**	**100**

Table 6.D.40 DIF Classifications for Grade Eight Form 1—White–African American

DIF Category	N	Pct
C-	2	10
B-	0	0
A-	7	33
A+	10	48
B+	2	10
C+	0	0
Small N	0	0
Total	**21**	**100**

Table 6.D.41 DIF Classifications for Grade Eight Form 1—English Only–English Learner

DIF Category	N	Pct
C-	1	5
B-	0	0
A-	9	43
A+	9	43
B+	2	10
C+	0	0
Small N	0	0
Total	**21**	**100**

Table 6.D.42 DIF Classifications for Grade Eight Form 1—No Special Services–Special Services

DIF Category	N	Pct
C-	0	0
B-	1	5
A-	11	52
A+	8	38
B+	0	0
C+	1	5
Small N	0	0
Total	**21**	**100**

Table 6.D.43 DIF Classifications for Grade Eight Form 1—Not Economically Disadvantaged–Economically Disadvantaged

DIF Category	N	Pct
C-	1	5
B-	1	5
A-	6	29
A+	13	62
B+	0	0
C+	0	0
Small N	0	0
Total	**21**	**100**

Table 6.D.44 DIF Classifications for Grade Eight Form 2—Male–Female

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	12	60
A+	8	40
B+	0	0
C+	0	0
Small N	0	0
Total	**20**	**100**

Table 6.D.45 DIF Classifications for Grade Eight Form 2—White–American Indian

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	0	0
A+	0	0
B+	0	0
C+	0	0
Small N	20	100
Total	**20**	**100**

Table 6.D.46 DIF Classifications for Grade Eight Form 2—White–Asian

DIF Category	N	Pct
C-	0	0
B-	2	10
A-	6	30
A+	11	55
B+	1	5
C+	0	0
Small N	0	0
Total	**20**	**100**

Table 6.D.47 DIF Classifications for Grade Eight Form 2—White–Pacific Islander

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	0	0
A+	0	0
B+	0	0
C+	0	0
Small N	20	100
Total	**20**	**100**

Table 6.D.48 DIF Classifications for Grade Eight Form 2—White–Filipino

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	0	0
A+	0	0
B+	0	0
C+	0	0
Small N	20	100
Total	**20**	**100**

Table 6.D.49 DIF Classifications for Grade Eight Form 2—White–Hispanic

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	9	45
A+	10	50
B+	0	0
C+	1	5
Small N	0	0
Total	**20**	**100**

Table 6.D.50 DIF Classifications for Grade Eight Form 2—White–African American

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	11	55
A+	8	40
B+	1	5
C+	0	0
Small N	0	0
Total	**20**	**100**

Table 6.D.51 DIF Classifications for Grade Eight Form 2—English Only–English Learner

DIF Category	N	Pct
C-	1	5
B-	2	10
A-	8	40
A+	6	30
B+	3	15
C+	0	0
Small N	0	0
Total	**20**	**100**

Table 6.D.52 DIF Classifications for Grade Eight Form 2—No Special Services–Special Services

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	12	60
A+	7	35
B+	1	5
C+	0	0
Small N	0	0
Total	**20**	**100**

Table 6.D.53 DIF Classifications for Grade Eight Form 2—Not Economically Disadvantaged–Economically Disadvantaged

DIF Category	N	Pct
C-	0	0
B-	1	5
A-	9	45
A+	10	50
B+	0	0
C+	0	0
Small N	0	0
Total	**20**	**100**

Table 6.D.54 DIF Classifications for Grade Eight Accessibility Form—Male–Female

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	3	50
A+	3	50
B+	0	0
C+	0	0
Small N	0	0
Total	**6**	**100**

Table 6.D.55 DIF Classifications for Grade Eight Accessibility Form—White–American Indian

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	0	0
A+	0	0
B+	0	0
C+	0	0
Small N	6	100
Total	**6**	**100**

Table 6.D.56 DIF Classifications for Grade Eight Accessibility Form—White–Asian

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	0	0
A+	0	0
B+	0	0
C+	0	0
Small N	6	100
Total	**6**	**100**

Table 6.D.57 DIF Classifications for Grade Eight Accessibility Form—White–Pacific Islander

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	0	0
A+	0	0
B+	0	0
C+	0	0
Small N	6	100
Total	**6**	**100**

Table 6.D.58 DIF Classifications for Grade Eight Accessibility Form—White–Filipino

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	0	0
A+	0	0
B+	0	0
C+	0	0
Small N	6	100
Total	**6**	**100**

Table 6.D.59 DIF Classifications for Grade Eight Accessibility Form—White–Hispanic

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	3	50
A+	3	50
B+	0	0
C+	0	0
Small N	0	0
Total	**6**	**100**

Table 6.D.60 DIF Classifications for Grade Eight Accessibility Form—White–African American

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	0	0
A+	0	0
B+	0	0
C+	0	0
Small N	6	100
Total	**6**	**100**

Table 6.D.61 DIF Classifications for Grade Eight Accessibility Form—English Only–English Learner

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	2	33
A+	4	67
B+	0	0
C+	0	0
Small N	0	0
Total	**6**	**100**

Table 6.D.62 DIF Classifications for Grade Eight Accessibility Form—No Special Services–Special Services

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	4	67
A+	2	33
B+	0	0
C+	0	0
Small N	0	0
Total	**6**	**100**

Table 6.D.63 DIF Classifications for Grade Eight Accessibility Form—Not Economically Disadvantaged–Economically Disadvantaged

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	3	50
A+	3	50
B+	0	0
C+	0	0
Small N	0	0
Total	**6**	**100**

Table 6.D.64 DIF Classifications for High School Form 1—Male–Female

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	8	44
A+	10	56
B+	0	0
C+	0	0
Small N	0	0
Total	**18**	**100**

Table 6.D.65 DIF Classifications for High School Form 1—White–American Indian

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	0	0
A+	0	0
B+	0	0
C+	0	0
Small N	18	100
Total	**18**	**100**

Table 6.D.66 DIF Classifications for High School Form 1—White–Asian

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	8	44
A+	9	50
B+	1	6
C+	0	0
Small N	0	0
Total	**18**	**100**

Table 6.D.67 DIF Classifications for High School Form 1—White–Pacific Islander

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	0	0
A+	0	0
B+	0	0
C+	0	0
Small N	18	100
Total	**18**	**100**

Table 6.D.68 DIF Classifications for High School Form 1—White–Filipino

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	0	0
A+	0	0
B+	0	0
C+	0	0
Small N	18	100
Total	**18**	**100**

Table 6.D.69 DIF Classifications for High School Form 1—White–Hispanic

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	9	50
A+	9	50
B+	0	0
C+	0	0
Small N	0	0
Total	**18**	**100**

Table 6.D.70 DIF Classifications for High School Form 1—White–African American

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	0	0
A+	0	0
B+	0	0
C+	0	0
Small N	18	100
Total	**18**	**100**

Table 6.D.71 DIF Classifications for High School Form 1—English Only–English Learner

DIF Category	N	Pct
C-	1	6
B-	1	6
A-	5	22
A+	10	61
B+	1	6
C+	0	0
Small N	0	0
Total	**18**	**100**

Table 6.D.72 DIF Classifications for High School Form 1—No Special Services–Special Services

DIF Category	N	Pct
C-	0	0
B-	1	6
A-	5	28
A+	12	67
B+	0	0
C+	0	0
Small N	0	0
Total	**18**	**100**

Table 6.D.73 DIF Classifications for High School Form 1—Not Economically Disadvantaged–Economically Disadvantaged

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	11	61
A+	7	39
B+	0	0
C+	0	0
Small N	0	0
Total	**18**	**100**

Table 6.D.74 DIF Classifications for High School Form 1—Grades Ten to Eleven

DIF Category	N	Pct
C-	0	0
B-	1	6
A-	9	33
A+	8	61
B+	0	0
C+	0	0
Small N	0	0
Total	**18**	**100**

Table 6.D.75 DIF Classifications for High School Form 1—Grades Ten to Twelve

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	10	56
A+	7	39
B+	1	6
C+	0	0
Small N	0	0
Total	**18**	**100**

Table 6.D.76 DIF Classifications for High School Form 2—Male–Female

DIF Category	N	Pct
C-	0	0
B-	1	5
A-	11	55
A+	8	40
B+	0	0
C+	0	0
Small N	0	0
Total	**20**	**100**

Table 6.D.77 DIF Classifications for High School Form 2—White–American Indian

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	0	0
A+	0	0
B+	0	0
C+	0	0
Small N	20	100
Total	**20**	**100**

Table 6.D.78 DIF Classifications for High School Form 2—White–Asian

DIF Category	N	Pct
C-	1	5
B-	0	0
A-	9	45
A+	9	45
B+	1	5
C+	0	0
Small N	0	0
Total	**20**	**100**

Table 6.D.79 DIF Classifications for High School Form 2—White–Pacific Islander

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	0	0
A+	0	0
B+	0	0
C+	0	0
Small N	20	100
Total	**20**	**100**

Table 6.D.80 DIF Classifications for High School Form 2—White–Filipino

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	0	0
A+	0	0
B+	0	0
C+	0	0
Small N	20	100
Total	**20**	**100**

Table 6.D.81 DIF Classifications for High School Form 2—White–Hispanic

DIF Category	N	Pct
C-	0	0
B-	1	5
A-	8	40
A+	11	55
B+	0	0
C+	0	0
Small N	0	0
Total	**20**	**100**

Table 6.D.82 DIF Classifications for High School Form 2—White–African American

DIF Category	N	Pct
C-	0	0
B-	2	10
A-	9	45
A+	7	35
B+	2	10
C+	0	0
Small N	0	0
Total	**20**	**100**

Table 6.D.83 DIF Classifications for High School Form 2—English Only–English Learner

DIF Category	N	Pct
C-	1	5
B-	2	10
A-	5	25
A+	10	50
B+	1	5
C+	1	5
Small N	0	0
Total	**20**	**100**

Table 6.D.84 DIF Classifications for High School Form 2—No Special Services–Special Services

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	11	55
A+	8	40
B+	1	5
C+	0	0
Small N	0	0
Total	**20**	**100**

Table 6.D.85 DIF Classifications for High School Form 2—Not Economically Disadvantaged–Economically Disadvantaged

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	8	61
A+	11	39
B+	1	0
C+	0	0
Small N	0	0
Total	**20**	**100**

Table 6.D.86 DIF Classifications for High School Form 2—Grades Ten to Eleven

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	8	40
A+	12	60
B+	0	0
C+	0	0
Small N	0	0
Total	**20**	**100**

Table 6.D.87 DIF Classifications for High School Form 2—Grades Ten to Twelve

DIF Category	N	Pct
C-	1	0
B-	0	0
A-	9	56
A+	9	39
B+	1	6
C+	0	0
Small N	0	0
Total	**20**	**100**

Table 6.D.88 DIF Classifications for High School Accessibility Form—Male–Female

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	4	57
A+	3	43
B+	0	0
C+	0	0
Small N	0	0
Total	**7**	**100**

Table 6.D.89 DIF Classifications for High School Accessibility Form—White–American Indian

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	0	0
A+	0	0
B+	0	0
C+	0	0
Small N	7	100
Total	**7**	**100**

Table 6.D.90 DIF Classifications for High School Accessibility Form—White–Asian

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	0	0
A+	0	0
B+	0	0
C+	0	0
Small N	7	100
Total	**7**	**100**

Table 6.D.91 DIF Classifications for High School Accessibility Form—White–Pacific Islander

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	0	0
A+	0	0
B+	0	0
C+	0	0
Small N	7	100
Total	**7**	**100**

Table 6.D.92 DIF Classifications for High School Accessibility Form—White–Filipino

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	0	0
A+	0	0
B+	0	0
C+	0	0
Small N	7	100
Total	**7**	**100**

Table 6.D.93 DIF Classifications for High School Accessibility Form—White–Hispanic

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	3	43
A+	4	57
B+	0	0
C+	0	0
Small N	0	0
Total	**7**	**100**

Table 6.D.94 DIF Classifications for High School Accessibility Form—White–African American

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	0	0
A+	0	0
B+	0	0
C+	0	0
Small N	7	100
Total	**7**	**100**

Table 6.D.95 DIF Classifications for High School Accessibility Form—English Only–English Learner

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	4	57
A+	3	43
B+	0	0
C+	0	0
Small N	0	0
Total	**7**	**100**

Table 6.D.96 DIF Classifications for High School Accessibility Form—No Special Services–Special Services

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	4	57
A+	3	43
B+	0	0
C+	0	0
Small N	0	0
Total	**7**	**100**

Table 6.D.97 DIF Classifications for High School Accessibility Form—Not Economically Disadvantaged–Economically Disadvantaged

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	4	57
A+	3	43
B+	0	0
C+	0	0
Small N	0	0
Total	**7**	**100**

Table 6.D.98 DIF Classifications for High School Accessibility Form—Grades Ten to Eleven

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	2	29
A+	5	71
B+	0	0
C+	0	0
Small N	0	0
Total	**7**	**100**

Table 6.D.99 DIF Classifications for High School Accessibility Form—Grades Ten to Twelve

DIF Category	N	Pct
C-	0	0
B-	0	0
A-	3	43
A+	4	57
B+	0	0
C+	0	0
Small N	0	0
Total	**7**	**100**

Chapter 7: Quality Control

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

7.1. Quality Control of Item Development

ETS's goal is to provide the best standards-based and innovative items for the CAST. Items developed for the CAST pilot were subject to an extensive item review process. The item writers hired to develop CAST assessment items and tasks were trained in California Assessment of Student Performance and Progress (CAASPP) and ETS policies on quality control of item content, sensitivity, and bias guidelines, and 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's 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 to ensure that each item measured what it is 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 for the pilot items, a group of California educators reviewed the items and performance tasks for accessibility, bias/sensitivity, and content and made recommendations for item enhancement. The details on item development processes for quality control purposes are described in subsection *3.6 Item Review Process* of *Chapter 3 Item Development and Assembly*.

When student response data on each item became available, ETS Psychometric Analysis and Research staff conducted item analysis and a key check to examine whether the items performed as expected. When the CAST pilot was completed and the population data were available, psychometric staff conducted a thorough item analysis and evaluated all items carefully using the statistical criteria described in subsection *6.2.6 Summary of Classical Item Analyses Flagging Criteria* to flag items that were potentially problematic due to poor item performance, content issues, item bias, and/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 performance tasks together with their associated statistical results and made recommendations about item disposition.

7.2. Quality Control of Test Form Development

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

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

- all answers are correct,
- answers score correctly in the item bank and incorrect answers score as incorrect,
- all items align with the standard,
- all content in the item is correct,
- distractors are plausible,
- multiple-choice item options are parallel in structure,
- language is grade-level appropriate,
- no more than three multiple choice items in a row have the same key,
- all art is correct,
- there are no errors in spelling or grammar, and
- items adhere to the approved style guide.

Reviews were also conducted for functionality and sequencing during the User Acceptance Testing (UAT) process to ensure all items were functioning as expected.

7.3. Quality Control of Test Administration

The quality of test administration for the CAST, and all assessments administered as part of the CAASPP System, was monitored and controlled through several strategies. A fully staffed support center, the California Technical Assistance Center (CalTAC), supports all local educational agencies (LEAs) in the administration of CAASPP assessments. In addition to providing guidance and answering questions, CalTAC regularly conducts outreach campaigns on particular administration topics to ensure all LEAs understand correct test administration procedures. CalTAC is guided by a core group of LEA Outreach Advocacy staff that manage communications to LEAs; provide regional and Web-based trainings; and host a Web site, 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 pilot. LEAs received training on these topics and were provided tools for reporting security incidents and resolving testing discrepancies for specific testing sessions.

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

7.4. Quality Control of Scoring

7.4.1 Quality Control of Machine-Scoring Procedures

The American Institutes for Research (AIR), the CAASPP subcontractor, provided the test delivery system (TDS) and scored machine-scorable items. AIR psychometric staff members independently reviewed all CAST test forms by taking sample tests. Responses to the test forms were compared with the answer keys for each form to confirm the accuracy of scoring keys. The scores for all applicable items were recorded. A final comparison of the test map to each online form as configured in the UAT environment ensured that no changes to the form were introduced prior to operational deployment.

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, and 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.

7.4.2 Quality Control of Human Scoring

For human scoring, ETS employed multiple quality controls including

- scoring leaders conducting backreads during each scoring shift,
- review of statistics on validity papers, and
- review of interrater reliability statistics.

Refer to subsection *5.2 Human Scoring* for the topics "Scoring Monitoring and Quality Management," "Interrater Reliability," and "Validity Responses and Sets**"** for more specific details on these tools used for quality control of human scoring.

7.5. Quality Control of Psychometric Processes

7.5.1 Development of Psychometric Specifications

ETS scoring specifications for the CAST pilot were completed, reviewed, approved, and checked in advance of the receipt of student response data. Psychometric specifications contained detailed scoring procedures as well as the procedures for determining whether a student attempted a test and whether that student's response data should be included in the statistical analyses and calculations for computing summary data.

7.5.2 Development of Psychometric Procedures

ETS's Enterprise Score Key Management System (eSKM) implemented scoring procedures specified by the psychometric team. Following scoring, a series of quality control checks were carried out by ETS psychometricians to ensure the accuracy of each score.

7.5.3 eSKM Processing of Multiple-Choice (MC) Items

Prior to the test administration, ETS Assessment Development (AD) staff reviewed and verified the keys and scoring rubrics for each item. Then, these keys and rubrics were provided to AIR for implementation machine scoring of the MC items. After AIR finished machine scoring, item scores and responses were delivered to ETS.

ETS's Centralized Repository Distribution System and Enterprise Service Bus departments collected and parsed .xml files that contained student response data from AIR. ETS's 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's Data Quality Services for data validation. Following successful validation, the student response statistical extracts were made available to the psychometric team.

ETS developed two parallel scoring systems to produce and verify overall students' scores. These were used for analyses, but not for reporting, this year. The eSKM scoring system received the individual students' item scores and item responses from AIR and calculated

individual student scores for ETS's reporting systems; the psychometric 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 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.

7.5.4 Psychometric Parallel Scoring Processing

The ETS psychometric team verified the eSKM scoring by comparing the parallel scoring programs using the Statistical Analysis System. Classical item analyses and differential item functioning analyses were then conducted using verified data.

All psychometric analyses conducted at ETS underwent comprehensive quality checks by a team of psychometricians and data analysts. Detailed checklists and psychometric specifications were developed by members of the team for each of the statistical procedures performed on CAST results data. Classical item analyses were performed. These include a check of scoring keys for multiple-choice items and scoring logic. Items that were flagged for questionable statistical attributes were sent to ETS AD staff for review; their comments were reviewed by the psychometricians before the data review meetings with the CDE.

References

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 8: Surveys

This chapter describes the development and administration of the survey questionnaires presented to test administrators and students during the 2016–17 California Science Test (CAST) pilot administration. The summary of findings and results of analyses from the survey are included in this chapter.

8.1. Survey Design and Questionnaire Development

The purpose of the surveys was to provide additional insight into the student test-taking experience of the CAST pilot from both the student and test administrator perspective. The feedback from those surveys will help in the development and administration of the CAST field tests and operational tests.

Two surveys were developed by Educational Testing Service (ETS) in consultation with the California Department of Education (CDE): (1) a student survey and (2) a test administrator survey. The CDE provided guidelines in terms of the length of the surveys and the number and focus of the questions. Multiple drafts were reviewed with feedback provided, and a near-final draft shared with the several members of the CDE's Technical Advisory Group in November 2016 for additional feedback. Both survey questionnaires used in the administration are included in Appendix 8.A.

8.2. Survey Administration

Student surveys were administered as a final section at the end of each pilot form. All students received the same set of questions regardless of pilot form and grade level. Survey items were available in braille for students who needed this accommodation. The test administrators completed their survey via SurveyGizmo, an online survey software tool. The test administrator survey was the same regardless of the grade level monitored.

Results from the full population of test takers—470,223 fifth graders, 448,709 eighth graders, and 455,267 high school students—are summarized in this chapter. Of the test administrators, 10,714 surveys were returned, 8,701 of which were completed. The other 2,013 respondents indicated they had not been a CAST administrator and thus the survey ended at that point.

8.3. Survey Analysis

ETS research scientists and data analysts cleaned the data and analyzed the responses of each survey question for students or test administrators. The distributions of responses of these survey questions are presented in Appendix 8.B.

8.4. Survey Results

8.4.1 Summary of Student Survey Results

Table 8.B.1 through Table 8.B.13 provide the results for the student surveys, with the specific number of students who responded to each item presented in the tables. Fewer than 0.06 percent of students were nonresponders. Each table reports on one question and shows the frequency of responses for each grade level in adjacent columns. The key findings from the survey results are summarized in the following statements.

- In terms of difficulty of the test, regardless of grade level, the modal student response was that the test was a little more difficult than other science tests (see Table 8.B.1).

- At each grade level, the majority of students reported trying at least as hard on this test as they did on other tests, although the percentage of students who reported trying less hard increased by grade level (see Table 8.B.4).
- The majority of students also reported that it was at least somewhat important for them to do well on the test. Although a very low percentage (three and five percent) of fifth- and eighth-grade students reported that it was not at all important to do well, this increased to 13 percent for high school students (see Table 8.B.7).
- More than 70 percent of students reported that at least quite a few of the items seemed different from what they had seen previously, with only four to five percent across all grades reporting that they were not different from what they experienced in science class (see Table 8.B.10).
- In terms of enjoying science, only between six percent (fifth graders) and 16 percent (high school students) reported almost never enjoying science (see Table 8.B.13).

Overall, the student survey responses suggest that the students took the pilot seriously, found it a little harder than other science tests, and expended a reasonable amount of effort on an assessment that they recognized as being at least somewhat different from other science assessments.

8.4.2 Summary of Test Administrator Survey Results

Table 8.B.16 through Table 8.B.22 provide the results for the test administrator surveys over all grade levels. Presented next are key observations and findings for each question.

- Just over half of the test administrators (53 percent) administered the pilot to their own science students (see Table 8.B.16). They reported that more than half of the students (54 percent) did not take the training test prior to the pilot (see Table 8.B.17).
- Only 10 percent of the test administrators reported that the instructions provided to them were somewhat or very confusing (see Table 8.B.18), but that 22 percent of students found the test directions somewhat or very confusing (see Table 8.B.19).
- Regarding engagement, 83 percent of the test administrators reported that either all or more than half of the students were engaged with the pilot (see Table 8.B.20).
- In terms of timing, 95 percent of the test administrators reported that either all or more than half of the students had sufficient time to complete the pilot (see Table 8.B.21).
- In terms of the computer interface, 85 percent of the test administrators reported that either all or more than half of the students were able to navigate through the online assessment (see Table 8.B.22).

Overall, the test administrator responses suggest that while greater use could be made of the training test in the future, overall, students were engaged with the assessment, understood the test directions, could navigate the assessment interface, and had sufficient time to complete the assessment.

Taken together, the student and test administrator results provide confidence in the meaningfulness of the student data collected during the pilot administration.

Appendix 8.A: CAST Surveys

Grade Five, Grade Eight and High School Student Survey Questions

1. Compared to most other science tests you have taken this year in school, this test was:
   a. Much easier than other science tests
   b. A little easier than other science tests
   c. A little harder than other science tests
   d. Much harder than other science tests
2. How hard did you try on this test compared most other science tests you have taken this year in school?
   a. Not as hard as on other tests
   b. About as hard as on other tests
   c. Harder than on other tests
   d. Much harder than on other tests
3. How important was it to you to do well on this test?
   a. Not at all important
   b. Somewhat important
   c. Important
   d. Very important
4. Were the test questions different from the type of questions you have seen in your homework assignments or classroom tests?
   a. Yes, many were different from anything I had seen before
   b. Yes, quite a few were different from anything I had seen before
   c. No, only a few were different from anything I had seen before
   d. No, they were not different from what I have seen in my science class
5. Do you like studying science?
   a. Almost always
   b. Usually
   c. Sometimes
   d. Almost never
   e. About the same as other subjects
   f. More than other subjects

Grade Five, Grade Eight, and High School Test Administrator Survey

Directions: Please complete the survey after your students have completed the CAST.

1. Are you a test examiner for CAST?
   YES/NO (if no, skip next seven questions)
2. For the students for whom you administered the CAST, are you the student's science teacher?
   a. Yes for all of the students I tested.
   b. For more than half of the students I tested.
   c. For less than half of the students I tested.
   d. For few or none of the students I tested.

3. To your knowledge, did the students you tested have an opportunity to take the training test prior to taking the pilot?
   a. All or most of the students took the training test.
   b. More than half of the students took the training test.
   c. Less than half of the students took the training test.
   d. Few or none of the students took the training test.
4. Which of the following statements best describes the instructions provided to you, the text administrator, for the pilot?
   a. The instructions were very clear.
   b. The instructions were somewhat clear.
   c. The instructions were somewhat confusing.
   d. The instructions were very confusing.
5. Which of the following statements best describes the instructions provided to students for the pilot?
   a. The students appeared to find the instructions very clear.
   b. The students appeared to find the instructions somewhat clear.
   c. The students appeared to find the instructions somewhat confusing.
   d. The students appeared to find the instructions very confusing.
6. Which of the following statements best describes your students' engagement with the pilot?
   a. All or most of my students appeared to be fully engaged with the pilot.
   b. More than half of my students appeared to be fully engaged with the pilot.
   c. Less than half of my students appeared to be fully engaged with the pilot.
   d. Few or none of my students were engaged with the pilot.
7. Which of the following statements best describes the time allotted for the pilot?
   a. All or most of my students appeared to have sufficient time for the pilot.
   b. More than half of my students appeared to have sufficient time for the pilot.
   c. Less than half of my students appeared to have sufficient time for the pilot.
   d. Few or none of my students appeared to have sufficient time for the pilot.
8. Which of the following statements best describes the computer interface for the assessment?
   a. All or most of my students appeared to be able to easily navigate through the online assessment.
   b. More than half of my students appeared to be able to easily navigate through the online assessment.
   c. Less than half of my students appeared to be able to easily navigate through the online assessment.
   d. Few or none of my students appeared to be able to easily navigate through the online assessment.

Appendix 8.B: Survey Results

Table 8.B.1 Distribution of Grade Five Student Survey Responses to Question 1

Compared to most other science tests you have taken this year in school, this test was:	(N=470,111)
a. Much easier than other science tests	12%
b. A little easier than other science tests	30%
c. A little harder than other science tests	47%
d. Much harder than other science tests	11%

Table 8.B.2 Distribution of Grade Eight Student Survey Responses to Question 1

Compared to most other science tests you have taken this year in school, this test was:	(N=448,553)
a. Much easier than other science tests	8%
b. A little easier than other science tests	27%
c. A little harder than other science tests	47%
d. Much harder than other science tests	18%

Table 8.B.3 Distribution of High School Student Survey Responses to Question 1

Compared to most other science tests you have taken this year in school, this test was:	(N=455,119)
a. Much easier than other science tests	5%
b. A little easier than other science tests	16%
c. A little harder than other science tests	43%
d. Much harder than other science tests	36%

Table 8.B.4 Distribution of Grade Five Student Survey Responses to Question 2

How hard did you try on this test compared most other science tests you have taken this year in school?	(N=470,082)
a. Not as hard as on other tests	12%
b. About as hard as on other tests	42%
c. Harder than on other tests	29%
d. Much harder than on other tests	16%

Table 8.B.5 Distribution of Grade Eight Student Survey Responses to Question 2

How hard did you try on this test compared most other science tests you have taken this year in school?	(N=448,509)
a. Not as hard as on other tests	19%
b. About as hard as on other tests	50%
c. Harder than on other tests	23%
d. Much harder than on other tests	8%

Table 8.B.6 Distribution of High School Student Survey Responses to Question 2

How hard did you try on this test compared most other science tests you have taken this year in school?	(N=455,075)
a. Not as hard as on other tests	31%
b. About as hard as on other tests	46%
c. Harder than on other tests	17%
d. Much harder than on other tests	6%

Table 8.B.7 Distribution of Grade Five Student Survey Responses to Question 3

How important was it to you to do well on this test?	(N=470,069)
a. Not at all important	3%
b. Somewhat important	13%
c. Important	31%
d. Very important	54%

Note: Total does not equal 100 percent due to rounding.

Table 8.B.8 Distribution of Grade Eight Student Survey Responses to Question 3

How important was it to you to do well on this test?	(N=448,489)
a. Not at all important	5%
b. Somewhat important	28%
c. Important	38%
d. Very important	29%

Table 8.B.9 Distribution of High School Student Survey Responses to Question 3

How important was it to you to do well on this test?	(N=455,060)
a. Not at all important	13%
b. Somewhat important	44%
c. Important	32%
d. Very important	12%

Note: Total does not equal 100 percent due to rounding.

Table 8.B.10 Distribution of Grade Five Student Survey Responses to Question 4

Were the test questions different from the type of questions you have seen in your homework assignments or classroom tests?	(N=470,050)
a. Yes, many were different from anything I had seen before	32%
b. Yes, quite a few were different from anything I had seen before	39%
c. No, only a few were different from anything I had seen before	24%
d. No, they were not different from what I have seen in my science class	5%

Table 8.B.11 Distribution of Grade Eight Student Survey Responses to Question 4

Were the test questions different from the type of questions you have seen in your homework assignments or classroom tests?	(N=448,466)
a. Yes, many were different from anything I had seen before	32%
b. Yes, quite a few were different from anything I had seen before	39%
c. No, only a few were different from anything I had seen before	26%
d. No, they were not different from what I have seen in my science class	4%

Note: Total does not equal 100 percent due to rounding.

Table 8.B.12 Distribution of High School Student Survey Responses to Question 4

Were the test questions different from the type of questions you have seen in your homework assignments or classroom tests?	(N=455,031)
a. Yes, many were different from anything I had seen before	43%
b. Yes, quite a few were different from anything I had seen before	35%
c. No, only a few were different from anything I had seen before	18%
d. No, they were not different from what I have seen in my science class	4%

Table 8.B.13 Distribution of Grade Five Student Survey Responses to Question 5

Do you like studying science?	(N=470,031)
a. Almost always	11%
b. Usually	15%
c. Sometimes	32%
d. Almost never	6%
e. About the same as other subjects	18%
f. More than other subjects	18%

Table 8.B.14 Distribution of Grade Eight Student Survey Responses to Question 5

Do you like studying science?	(N=448,448)
a. Almost always	10%
b. Usually	20%
c. Sometimes	36%
d. Almost never	10%
e. About the same as other subjects	13%
f. More than other subjects	11%

Table 8.B.15 Distribution of High School Student Survey Responses to Question 5

Do you like studying science?	(N=455,002)
a. Almost always	10%
b. Usually	21%
c. Sometimes	37%
d. Almost never	15%
e. About the same as other subjects	9%
f. More than other subjects	8%

Table 8.B.16 Distribution of Test Administrator Responses to Question 2

For the students for whom you administered the CAST, are you the student's science teacher?	N=8,701 test administrators
a. Yes for all of the students I tested.	53%
b. For more than half of the students I tested.	5%
c. For less than half of the students I tested.	3%
d. For few or none of the students I tested.	38%
Did not respond	1%

Table 8.B.17 Distribution of Test Administrator Responses to Question 3

To your knowledge, did the students you tested have an opportunity to take the training test prior to taking the pilot?	N=8,701 test administrators
a. All or most of the students took the training test.	30%
b. More than half of the students took the training test.	7%
c. Less than half of the students took the training test.	6%
d. Few or none of the students took the training test.	54%
Did not respond	2%

Table 8.B.18 Distribution of Test Administrator Responses to Question 4

Which of the following statements best describes the instructions provided to you, the test administrator, for the pilot?	N=8,701 test administrators
a. The instructions were very clear.	53%
b. The instructions were somewhat clear.	36%
c. The instructions were somewhat confusing.	8%
d. The instructions were very confusing.	2%
Did not respond	1%

Table 8.B.19 Distribution of Test Administrator Responses to Question 5

Which of the following statements best describes the instructions provided to students for the pilot?	N=8,701 test administrators
a. The students appeared to find the instructions very clear.	37%
b. The students appeared to find the instructions somewhat clear.	41%
c. The students appeared to find the instructions somewhat confusing.	17%
d. The students appeared to find the instructions very confusing.	5%
Did not respond	1%

Table 8.B.20 Distribution of Test Administrator Responses to Question 6

Which of the following statements best describes your students' engagement with the pilot?	N=8,701 test administrators
a. All or most of my students appeared to be fully engaged with the pilot.	42%
b. More than half of my students appeared to be fully engaged with the pilot.	41%
c. Less than half of my students appeared to be fully engaged with the pilot.	12%
d. Few or none of my students were engaged with the pilot.	4%
Did not respond	1%

Table 8.B.21 Distribution of Test Administrator Responses to Question 7

Which of the following statements best describes the time allotted for the pilot?	N=8,701 test administrators
a. All or most of my students appeared to have sufficient time for the pilot.	79%
b. More than half of my students appeared to have sufficient time for the pilot.	16%
c. Less than half of my students appeared to have sufficient time for the pilot.	3%
d. Few or none of my students appeared to have sufficient time for the pilot.	1%
Did not respond	1%

Table 8.B.22 Distribution of Test Administrator Responses to Question 8

Which of the following statements best describes the computer interface for the pilot?	N=8,701 test administrators
a. All or most of my students appeared to be able to easily navigate through the online assessment.	49%
b. More than half of my students appeared to be able to easily navigate through the online assessment.	36%
c. Less than half of my students appeared to be able to easily navigate through the online assessment.	10%
d. Few or none of my students appeared to be able to easily navigate through the online assessment.	4%
Did not respond	1%