California Department of Education
Assessment Development and Administration Division

California Standards Tests for Science
Technical Report
2015–16 Administration

Submitted March 1, 2017

Educational Testing Service

Contract No. CN150012

Table of Contents

Tables

Figures

Acronyms and Initialisms Used in the *CSTs for Science Technical Report*

ADA	Americans with Disabilities Act	I-FEP	initially fluent English proficient
AERA	American Educational Research Association	IRT	item response theory
ARP	Assessment Review Panel	IT	Information Technology
CAASPP	California Assessment of Student Performance and Progress	LEA	local educational agency
CAHSEE	California High School Exit Examination	MH DIF	Mantel-Haenszel DIF
CalTAC	California Technical Assistance Center	NCME	National Council on Measurement in Education
CAPA	California Alternate Performance Assessment	NPS	nonpublic, nonsectarian school
CCR	*California Code of Regulations*	OIB	ordered item booklet
CDE	California Department of Education	OTI	Office of Testing Integrity
CDS	county/district/school	*p*-value	item proportion correct
CELDT	California English Language Development Test	Pt-Bis	point-biserial correlations
CI	confidence interval	QC	quality control
CMA	California Modified Assessment	R-FEP	reclassified fluent English proficient
CSEMs	conditional standard errors of measurement	SBE	State Board of Education
CSTs	California Standards Tests	SD	standard deviation
DFA	*Directions for Administration*	SEM	standard error of measurement
DIF	differential item functioning	SFTP	secure file transfer protocol
DOK	depth of knowledge	SGID	School and Grade Identification sheet
EC	*Education Code*	SKM	score key management
EL	English learner	SPAR	Statewide Pupil Assessment Review
ELA	English–language arts	STAR	Standardized Testing and Reporting
EOC	end-of-course	STS	Standards-based Tests in Spanish
ETS	Educational Testing Service	TBD	To Be Determined
FIA	final item analysis	TIF	test information function
GENASYS	Generalized Analysis System	TOMS	Test Operations Management System
HumRRO	Human Resource Research Organization	USDOE	United States Department of Education
ICC	item characteristic curve	WRMSD	weighted root-mean-square difference
IEP	individualized education program		

Chapter 1: Introduction

Background

In 1997 and 1998, the California State Board of Education (SBE) adopted content standards in four major content areas: English–language arts, mathematics, history–social science, and science. These standards were designed to provide state-level input into instruction curricula and serve as a foundation for the state's school accountability programs.

In order to measure and evaluate student achievement of the content standards, the state instituted the Standardized Testing and Reporting (STAR) Program. This Program, administered annually as paper-pencil assessments, was authorized in 1997 by state law (Senate Bill 376). In 2013, Assembly Bill 484 was introduced to establish California's new student assessment system, now known as the California Assessment of Student Performance and Progress (CAASPP). The CAASPP System of assessments replaced the STAR Program. The new assessment system includes computer-based tests for English language arts/literacy and mathematics; and paper-pencil tests in science for the California Standards Tests (CSTs), California Modified Assessment (CMA), and California Alternate Performance Assessment (CAPA), and reading/language arts for the Standards-based Tests in Spanish (STS).

During the 2015–16 administration, the CAASPP System had four components for the paper-pencil tests:

- CSTs for Science, produced for California public schools to assess the California content standards for science in grades five, eight, and ten
- CMA for Science, an assessment of students' achievement of California's content standards for science in grades five, eight, and ten, developed for students with an individualized education program (IEP) who meet the CMA eligibility criteria approved by the SBE
- CAPA for Science, produced for students with an IEP and who have significant cognitive disabilities in grades five, eight, and ten and are not able to take the CSTs for Science with accommodations and/or non-embedded accessibility supports or the CMA for Science with accommodations
- STS for Reading/Language Arts, an optional assessment of students' achievement of California's content standards for Spanish-speaking English learners (ELs) that is administered as the CAASPP System's designated primary language test

Test Purpose

The three grade-level CSTs for Science form the cornerstone of the paper-pencil tests of the CAASPP System. The CSTs for Science, given in English, are designed to show how well students in grades five, eight, and ten are performing with respect to California's content standards in science that were adopted by the SBE in 1998. These standards describe what students should know and be able to do at each grade level.

Test Content

The CSTs for Science are administered in grades five, eight, and ten. The grade five test assesses science content standards in grades four and five. The grade eight test assesses the grade-level standards. Finally, the CST for Life Science administered in grade ten assesses science content standards in grades six, seven, eight, and high school biology. For

a list of the CST for Science reporting clusters and the standards they assess, see Appendix 2.B—Reporting Clusters on page 20.

Intended Population

Each grade-level CST for Science was administered to approximately 437,000 to 462,000 test-takers during the 2015–16 administration.

All students enrolled in grades five, eight, and ten in California public schools on the day testing begins are required to take a CST for Science assessment or, for eligible students, a CMA for Science assessment; or, for students who meet the eligibility requirements, the CAPA for Science. This requirement includes ELs regardless of the length of time they have been in U.S. schools or their fluency in English, as well as students with disabilities who receive special education services. For students with cognitive disabilities, the decision to administer a CST for Science, the CMA for Science, or the CAPA for Science is made by their IEP team.

Parents/Guardians may submit a written request to have their child exempted from taking any or all parts of the tests within the CAASPP System. Only students whose parents/Guardians submit a written request may be exempted from taking the tests (*California Education Code [EC]* Section 60615).

Intended Use and Purpose of Test Scores

The results for tests within the CAASPP System are used for two primary purposes, described in *EC* sections 60602.5 (a) and (a)(4). Sections 60602.5 (c) and (d) provide additional background on the tests. (Excerpted from the *EC* Section 60602 Web page at http://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?lawCode=EDC&division=4.& title=2.&part=33.&chapter=5.&article=1).

“60602.5 (a) It is the intent of the Legislature in enacting this chapter to provide a system of assessments of pupils that has the primary purposes of assisting teachers, administrators, and pupils and their parents; improving teaching and learning; and promoting high-quality teaching and learning using a variety of assessment approaches and item types. The assessments, where applicable and valid, will produce scores that can be aggregated and disaggregated for the purpose of holding schools and local educational agencies accountable for the achievement of all their pupils in learning the California academic content standards.”

“60602.5 (a) (4) Provide information to pupils, parents or guardians, teachers, schools, and local educational agencies on a timely basis so that the information can be used to further the development of the pupil and to improve the educational program.”

“60602.5 (c) It is the intent of the Legislature that parents, classroom teachers, other educators, pupil representatives, institutions of higher education, business community members, and the public be involved, in an active and ongoing basis, in the design and implementation of the statewide pupil assessment system and the development of assessment instruments.”

“60602.5 (d) It is the intent of the Legislature, insofar as is practically feasible and following the completion of annual testing, that the content, test structure, and test items in the assessments that are part of the statewide pupil assessment system become open and transparent to teachers, parents, and pupils, to assist stakeholders in working together to demonstrate improvement in pupil academic achievement. A planned change in annual test

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

Testing Window

The CSTs for Science are administered within a 25-day window, which begins 12 instructional days before and ends 12 instructional days after the day on which 85 percent of the instructional year is completed. Local educational agencies (LEAs) may use all or any part of the 25 days for testing but are encouraged to schedule testing over no more than a 10- to 15-day period (*California Code of Regulations [CCR], Title 5,* Education, Division 1, Chapter 2, Subchapter 3.75, Article 2, § 855 [a] [2]; please note this section of *5 CCR* has been updated since the 2015–16 CAASPP administration).

Significant CAASPP Developments in 2015–16

Online Reporting System (ORS)

Results for all paper-pencil tests, including the CSTs for Science, now can be accessed by LEA CAASPP coordinators and CAASPP test site coordinators in the ORS.

Testing Window

Pursuant to *5 CCR*, Section 855 (a)(3), the testing windows for the CSTs for Science were recalculated to start on the day of completion of 85 percent of instruction (rather than the day after completion).

Unlisted Resources

The term "individualized aid" has been replaced with "unlisted resource." An unlisted resource is an instructional support that a pupil regularly uses in daily instruction and/or assessment that has not been previously identified as a universal tool, designated support or accommodation. Because an unlisted resource has not been previously identified as a universal tool, designated support, or accommodation, it may or may not change the construct of the assessment (*5 CCR*, Section 850 [ak]). When an unlisted resource has been determined to change the construct, its use invalidates the results for the purpose of accountability. A student score is provided with a statement that the test was administered under conditions that resulted in a score that may not be an accurate representation of the student's achievement.

Web Reporting

Statewide results were released via a newly designed Public Reporting Web site at http://caaspp.cde.ca.gov/, which is available to view summary results. Two new features include the ability to see change over time (e.g., view grade four summary results and review results from grade three from the previous year), and the ability to view results from up to three entities (i.e., schools, district, county, or state).

Limitations of the Assessment

Score Interpretation

An LEA may use CST for Science results to help make decisions about student placement, promotion, retention, or other considerations related to student achievement. However, it is important to remember that a single test can provide only limited information. Other relevant information should be considered as well. It is advisable for parents/guardians to evaluate their child's strengths and weaknesses in the relevant topics by reviewing local assessments, classroom tests, student grades, classroom work, and teacher

recommendations in addition to the child's CST for Science results (California Department of Education [CDE], 2013).

Out-of-Level Testing

Each CST for Science is designed to measure the content corresponding to a specific grade or course and is appropriate for students in the specific grade or course. Testing below a student's grade is not allowed for the CSTs for Science or any test in the CAASPP System; all students in grades five, eight, and ten are required to take the science test for the grade in which they are enrolled. LEAs are advised to review all IEPs to ensure that any provision for testing below a student's grade level has been removed.

Score Comparison

When comparing scale score results for the CSTs for Science, the reviewer is limited to comparing results only within the same content area and grade. For example, it is appropriate to compare scores obtained by students and/or schools on the 2015–16 grade five science test; it would not be appropriate to compare scores obtained on the grade five science test with those obtained on the grade ten science test. The reviewer may compare results for the same content area and grade, within a school, between schools, or between a school and its district, its county, or the state within the same year or to previous years.

Groups and Organizations Involved with the CAASPP System

State Board of Education (SBE)

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

In addition to adopting the rules and regulations for itself, its appointees, and California's public schools, the SBE is also the state educational agency responsible for overseeing California's compliance with programs that meet the requirements of the federal Every Student Succeeds Act and the state's Public School Accountability Act, which measure the academic performance and growth of schools on a variety of academic metrics (CDE, 2015).

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 9,800 schools. California aims to provide a world-class education for all students, from early childhood to adulthood. The Department of Education serves California by innovating and collaborating with educators, schools, 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 District, School & Innovation Branch that oversees programs promoting innovation and improved student achievement. Programs include oversight of statewide assessments and the collection and reporting of educational data (CDE, 2016).

Contractor—Educational Testing Service (ETS)

The CDE and the SBE contract with ETS to develop, administer, and report the CAASPP assessments. ETS has the overall responsibility for working with the CDE to implement and maintain an effective assessment system as well as having responsibility for producing and

distributing materials, processing the tests, and producing reports. Activities directly conducted by ETS include the following:

- Overall management of the program activities;
- Development of all test items;
- Construction and production of test booklets and related test materials;
- Support and training provided to counties, LEAs, and independently testing charter schools;
- Implementation and maintenance of the Test Operations Management System for orders of materials and pre-identification services;
- Completion of all psychometric activities;
- Production of all scannable test materials;
- Packaging, distribution, and collection of testing materials to LEAs and independently testing charter schools;
- Scanning and scoring of all responses; and
- Production of all score reports and data files of test results.

Overview of the Technical Report

This technical report addresses the characteristics of the 2015–16 CSTs for Science. The technical report contains nine additional chapters as follows:

- Chapter 2 presents a conceptual overview of the processes involved in a testing cycle for a CST for Science form. This includes test construction, test administration, generation of test scores, and dissemination of score reports. Information about the distributions of scores aggregated by subgroups based on demographics and the use of special services is included, as are the references to various chapters that detail the processes briefly discussed in this chapter.
- Chapter 3 describes the procedures followed during the development of valid CST for Science items, when newly developed items were used in forms—in 2015–16, the test forms from previous STAR administrations from different years were reused and there was no new item development. The chapter also explains the process of field-testing new items and the review of items by contractors and content experts.
- Chapter 4 details the content and psychometric criteria that guided the construction of the CST for Science forms reused in 2015–16.
- Chapter 5 presents the processes involved in the actual administration of the 2015–16 CSTs for Science with an emphasis on efforts made to ensure standardization of the tests. It also includes a detailed section that describes the procedures that were followed by ETS to ensure test security.
- Chapter 6 describes the standard-setting process previously conducted to establish cut scores for newly introduced CSTs for Science.
- Chapter 7 details the types of scores and score reports that are produced at the end of each administration of the CSTs for Science.
- Chapter 8 summarizes the results of the test- and item-level analyses performed during the 2015–16 administration of the tests. These include the classical item analyses, the reliability analyses that include assessments of test reliability and the consistency and accuracy of the CST for Science performance-level classifications, and the procedures

designed to ensure the validity of CST for Science score uses and interpretations. Also discussed in this chapter are item response theory, CST for Science conversion tables, and the considerations and processes involved in pre-equating.

- Chapter 9 highlights the importance of controlling and maintaining the quality of the CSTs for Science.
- Chapter 10 presents historical comparisons of various item- and test-level results for the past three years and for the base year of each test, which vary according to test.

Each chapter contains summary tables in the body of the text. However, extended appendixes that give more detailed information are provided at the end of the relevant chapters.

References

California Code of Regulations, Title 5, Education, Division 1, Chapter 2, Subchapter 3.75, Article 2, §§ 853.5 and 855. Retrieved from http://www.cde.ca.gov/re/lr/rr/caaspp.asp

California Department of Education. (2013). *STAR Program information packet for school district and school staff* (p. 15). Sacramento, CA.

California Department of Education, EdSource, & the Fiscal Crisis Management Assistance Team. (2014). *Fiscal, demographic, and performance data on California's K–12 schools*. Sacramento, CA: Ed-Data. Retrieved from http://www.ed-data.k12.ca.us/App_Resx/EdDataClassic/fsTwoPanel.aspx?#!bottom=/_layouts/EdDataClassic/profile.asp?Tab=1&Level=04&reportNumber=16

California Department of Education. (2015, May). *State Board of Education responsibilities.* Retrieved from http://www.cde.ca.gov/be/ms/po/sberesponsibilities.asp

California Department of Education. (2016, January). *Organization.* Retrieved from http://www.cde.ca.gov/re/di/or/

Chapter 2: An Overview of CST for Science Processes

This chapter provides an overview of the processes involved in a typical test development and administration cycle for a California Standards Test (CST) for Science. Also described are the specifications maintained by Educational Testing Service (ETS) to implement each of those processes. In 2015–16, three CSTs for Science were administered. Intact forms from 2011–12 were used. The CSTs for Science in grades five, eight, and ten are considered pre-equated.

The chapter is organized to provide a brief description of each process followed by a summary of the associated specifications. More details about the specifications and the analyses associated with each process are described in other chapters that are referenced in the sections that follow.

Item Development

Item Formats

All CSTs for Science administered in 2015–16 contain four-option multiple-choice items.

Item Specifications

There was no new item development for the 2015–16 administration. Prior to the 2013–14 administration, the CST for Science items were developed to measure California content standards adopted by the state in 1997 and 1998 and designed to conform to principles of item writing defined by ETS (ETS, 2002). ETS maintained and updated an item specifications document, otherwise known as “item writer guidelines,” for each CST for Science and used an item utilization plan to guide the development of the items for each content area. Item writing emphasis was determined in consultation with the California Department of Education (CDE).

The item specifications describe the characteristics of the items that should be written to measure each content standard; items of the same type should consistently measure the content standards in the same way. The item specifications helped ensure that the items on the CSTs for Science measure the content standards in the same way. To achieve this, the item specifications provided detailed information to item writers who developed items for the CSTs for Science.

The items selected for each CST for Science underwent an extensive item review process that is designed to provide the best standards-based tests possible. Details about the item specifications, the item review process, and the item utilization plan are presented in Chapter 3, starting on page 29.

Item Banking

Newly developed items were placed into the item bank in years when items were developed. Before this was done, ETS prepared them for review by content experts and various external review committees such as the Assessment Review Panels (ARPs), which are described in Chapter 3, starting on page 29; and the Statewide Pupil Assessment Review panel, described in Chapter 3, starting on page 29.

Once the ARP review was complete, the items were placed in the item bank along with the associated information obtained at the review sessions. Items that were accepted by the content experts were updated to a “field-test ready” status. ETS then delivered the items to

the CDE by means of a delivery of the California electronic item bank. Items were subsequently field-tested to obtain information about item performance and item statistics that could be used to assemble operational forms.

The CDE then reviewed those items with their statistical data flagged to determine whether they should be used operationally (see page 36 for more information about the CDE's data review). Any additional updates to item content and statistics were based on data collected from the operational use of the items. However, only the latest content of the item is retained in the bank at any time, along with the administration data from every administration that has included the item.

Further details on item banking are presented on page 36 in Chapter 3.

Item Refresh Rate

Prior to the first time intact forms were reused, during the 2012–13 administration, the item utilization plan required that each year, 35 percent of items on an operational form were refreshed (replaced); these items remained in the item bank for future use. Because the forms were reused, there were no items refreshed in the 2015–16 administration.

Test Assembly

Test Length

The CST for Science grade-level tests are composed of 60 operational items each. The considerations used in deciding the test length are described on page 39 in Chapter 4.

Each CST for Science also includes six field-test items in addition to the operational items. Although there was no new item development for the 2015–16 administration, the field-test items were included as part of the intact forms but did not contribute to students' scores. The total number of items, including operational and field-test items, in each CST for Science form and the estimated time to complete a test form are presented in Appendix 2.A on page 19.

Test Blueprints

ETS selected all CST for Science items to conform to the State Board of Education (SBE)-approved California content standards and test blueprints. The test blueprints for the CSTs for Science, adopted in 2002 by the SBE, can be found on the California Department of Education (CDE) Standardized Testing and Reporting CST Blueprints Web page at http://www.cde.ca.gov/ta/tg/sr/blueprints.asp.

Because the test blueprints specify the number of items at the individual standard level, scores for the CST for Science items are grouped into subcontent areas referred to as "reporting clusters." For each CST for Science reporting cluster, the percentage of questions correctly answered was reported on a student's score report prior to the 2015–16 administration. Although only the total test scale score are reported and cluster scores are no longer included in the score report, a description of the CST for Science reporting clusters and the standards that comprise each cluster are provided in Appendix 2.B, which starts on page 20.

Content Rules and Item Selection

Intact forms from 2011–12 were used during the 2015–16 administration. (See Table 8.4 on page 90 for administration years.) In a typical development cycle prior to the 2012–13 administration, test developers followed a number of rules when developing a new test form for a given grade and content area. First and foremost, they selected items that met the

blueprint for that grade and content area. Using an electronic item bank, assessment specialists began by identifying a number of linking items. These were items that had appeared in previous operational test administrations and were then used to equate subsequent (new) test forms. After the linking items were approved, assessment specialists populated the rest of the test form.

Linking items were selected to proportionally represent the full blueprint. Each CST for Science form was a collection of test items designed for a reliable, fair, and valid measure of student achievement within well-defined course content.

Another consideration was the difficulty of each item. Test developers strived to ensure that there were some easy and some hard items and that there were a number of items in the middle range of difficulty. The detailed rules are presented in Chapter 4, which begins on page 39.

Psychometric Criteria

The staff assessed the projected test characteristics during the preliminary review of the assembled forms. The statistical targets used to develop the 2015–16 forms and the projected characteristics of the assembled forms are presented starting from page 40 in Chapter 4.

The items in test forms were organized and sequenced differently according to the requirements of the content area. Further details on the arrangement of items during test assembly are also described on page 39 in Chapter 4.

All the forms in the 2015–16 CST for Science administration were used in prior operational test administrations. See Table 8.4 on page 90 for the list containing the administration in which each CST for Science was originally administered.

Test Administration

It is of utmost priority to administer the CSTs for Science in an appropriate, consistent, secure, confidential, and standardized manner.

Test Security and Confidentiality

All tests within the California Assessment of Student Performance and Progress (CAASPP) System are secure documents. For the CST for Science administration, every person having access to test materials maintains the security and confidentiality of the tests. ETS's Code of Ethics requires that all test information, including tangible materials (such as test booklets, test questions, test results), confidential files, processes, and activities are kept secure. To ensure security for all tests that ETS develops or handles, ETS maintains an Office of Testing Integrity (OTI). A detailed description of the OTI and its mission is presented in Chapter 5 on page 48.

In the pursuit of enforcing secure practices, ETS and its OTI strive to safeguard the various processes involved in a test development and administration cycle. Those processes are listed below. The practices related to each of the following processes are discussed in detail in Chapter 5, starting on page 48.

- Test development
- Item and data review
- Item banking
- Transfer of forms and items to the CDE

- Security of electronic files using a firewall
- Printing and publishing
- Test administration
- Test delivery
- Processing and scoring
- Data management
- Statistical analysis
- Reporting and posting results
- Student confidentiality
- Student test results

Procedures to Maintain Standardization

The CST for Science processes are designed so that the tests are administered and scored in a standardized manner. ETS takes all necessary measures to ensure the standardization of the CSTs for Science, as described in this section.

Test Administrators

The CSTs for Science are administered in conjunction with the other tests that comprise the CAASPP System. ETS employs personnel who facilitate various processes involved in the standardization of an administration cycle.

Staff at LEAs who are central to the processes include LEA CAASPP coordinators, CAASPP test site coordinators, test administrators, proctors, and scribes. The responsibilities of each of the staff members are included in the *CAASPP Paper-Pencil Testing Test Administration Manual* (CDE, 2016a); see page 53 in Chapter 5 for more information.

Test Directions

A series of instructions compiled in detailed manuals is provided to the test administrators. Such documents include, but are not limited to, the following:

***Directions for Administration* (*DFA*s)—**Manuals used by test administrators to administer the CSTs for Science to students to be followed exactly so that all students have an equal opportunity to demonstrate their academic achievement (See page 54 in Chapter 5 for more information.)

***CAASPP Paper-Pencil Testing Test Administration Manual*—**Test administration procedures for LEA CAASPP coordinators and CAASPP test site coordinators (See page 54 in Chapter 5 for more information.)

Test Operations Management System (TOMS) manuals—Instructions for the Web-based modules that allow LEA CAASPP coordinators to set up test administrations, assign tests, and assign student test settings; every module has its own user manual with detailed instructions on how to use TOMS (See page 55 in Chapter 5 for more information.)

Universal Tools, Designated Supports, and Accommodations

All public school students participate in the CAASPP System, including students with disabilities and ELs. Most students with individualized education programs (IEPs) and most English learners (ELs) take the CSTs for Science under standard conditions. However,

some students with IEPs and some ELs may need assistance when taking the CSTs for Science. This assistance takes the form of universal tools, designated supports, and accommodations. All students in these categories may have test administration directions simplified or clarified.

Appendix 2.C on page 21 presents an adaptation of Matrix One of the "Universal Tools, Designated Supports, and Accommodations for the California Assessment of Student Performance and Progress" (CDE, 2016b). Part 2 of Matrix One, found in Table 2.C.1, includes the non-embedded supports; Part 3, also in Table 2.C.1, includes the non-embedded accessibility supports that can be used for the paper-pencil tests. Appendix 2.C shows only the supports that were allowed for the CSTs for Science during the 2015–16 administration and, because they were mapped to CST for Science answer documents, had results data. Table 2.C.1 describes the Section A3 answer document options that frequencies and percentages are reported for in Appendix 2.D. Frequencies and percentages are also included in Appendix 2.D for students who did not map to a specific universal tool, designated support, or accommodation, as well as the reported answer document options in section A4 that are unmapped.

The purpose of universal tools, designated supports, and accommodations in testing is to allow *all* students the opportunity to demonstrate what they know and what they are able to do, rather than give students using them an advantage over other students or artificially inflate their scores. Universal tools, designated supports, and accommodations minimize or remove the barriers that could otherwise prevent students from generating results that reflect their achievement in the content area.

Non-embedded Supports

Non-embedded supports—universal tools, designated supports, and accommodations—do not change the construct being measured. For example, if students used a non-embedded support, such as a large-print version of any CAASPP test, the accommodation does not change what was tested. Accommodations are available to students with documented need; these must be identified, approved, and listed in the student's IEP or Section 504 plan. The use of non-embedded supports does not change the way scores are reported.

Unlisted Resources

Unlisted resources are those that fundamentally change what is being tested and may interfere with the construct being measured. All unlisted resources must be identified, approved, and listed in the student's IEP or Section 504 plan. Unlisted resources, when approved, are marked as option Y in Appendix 2.D.

Special Services Summaries

The percentage of students using various universal tools, designated supports, and accommodations during the 2015–16 administration of the CSTs for Science is presented in Appendix 2.D, which starts on page 22. The data are organized into three sections within each table. The first section presents the percentages of students using each accommodation or modification in the total testing population. The next section presents the results for students in special education and for those not in special education. The final section presents the results for various categories based on the following levels of English-language fluency:

- **English only (EO)—**A student for whom there is a report of English as the primary language (i.e., language first learned, most frequently used at home, or most frequently spoken by the parents or adults in the home) on the "Home Language Survey"

- **Initially fluent English proficient (I-FEP)—**A student whose primary language is a language other than English who initially met the LEA criteria for determining proficiency in English
- **English learner (EL)—**A student who first learned or has a home language other than English who was determined to lack sufficient fluency in English on the basis of state oral language (K–12) and literacy (3–12) assessments to succeed in the school's regular instructional program (For students tested for initial classification prior to May 2001, this determination is made on the basis of the state-approved instrument the LEA was using. For students tested after May 2001, use the California English Language Development [CELDT] results.)
- **Reclassified fluent English proficient (R-FEP)—**A student whose primary language is a language other than English who was reclassified from EL to fluent-English proficient

The information within each section is presented for the relevant grades. Most variations, accommodations, and modifications are common across CSTs for Science.

Scores

Total test raw scores for the CSTs for Science equal the sum of students' scores on the operational multiple-choice test items.

Total test raw scores on each CST for Science are converted to three-digit scale scores using the pre-equating process described starting on page 14. CST for Science results are reported through the use of these scale scores; the scores range from 150 to 600 for each test. Also reported are performance levels obtained by categorizing the scale score into one of the following levels: far below basic, below basic, basic, proficient, or advanced. Scale scores of 300 and 350 correspond to the cut scores for the basic and proficient performance levels, respectively. The state's target is for all students to score at the proficient or advanced level.

While no longer reported to students, performance on reporting clusters is provided in this technical report. The subscore or reporting cluster score is obtained by summing an examinee's scores on the items in each reporting cluster. That information is reported in terms of a percent-correct score.

Detailed descriptions of CST for Science scores are found in Chapter 7, which starts on page 64.

Aggregation Procedures

In order to provide meaningful results to the stakeholders, CST for Science scores for a given grade are aggregated at the school, independently testing charter school, district, county, and state levels. The aggregated scores are generated for both individual students and demographic subgroups. The following sections describe the summary results of individual and demographic subgroup CST for Science scores aggregated at the state level.

Please note that aggregation is performed on valid scores only, which are cases where students met all of the following criteria:

1. Met attemptedness criteria
2. Did not have a parental exemption
3. Did not miss any part of the test due to illness or medical emergency

4. Did not take a modified test
5. Did not test out of level (grade inappropriate)

Individual Scores

Table 7.1 and Table 7.2 starting on page 66 in Chapter 7 offer summary statistics for individual scores aggregated at the state level, describing overall student performance on each CST for Science. Included in the tables are the means and standard deviations of student scores expressed in terms of both raw scores and scale scores; the raw score means and standard deviations expressed as percentages of the total raw score points in each test; and the percentages of students in each performance level.

Statistics summarizing CST for Science student performance by grade are provided in Table 7.A.1 on page 72 in Appendix 7.A.

Demographic Subgroup Scores

In Table 7.B.1 through Table 7.B.2 starting on page 73 in Appendix 7.B, students are grouped by demographic characteristics, including gender, ethnicity, English-language fluency, use of special education services, and economic status. The tables show the numbers of students with valid scores in each group, scale score means and standard deviations, and percent in a performance level, as well as percent correct for each reporting cluster for each demographic group. Table 7.3 on page 67 provides definitions for the demographic groups included in the tables.

Equating

Post-Equating

In the years when the new forms were developed prior to the 2012–13 administration, the CST for Science scores were equated to reference form scores using a linking items nonequivalent groups data collection design and post-equating methods based on item response theory (IRT) (Hambleton & Swaminathan, 1985). The “base” or “reference” calibrations for the CSTs for Science were established by calibrating samples of item response data from a specific administration, through which item parameter estimates for the items in the forms were placed on the reference scale using a set of linking items selected from the previous year. Doing so established a scale to which subsequent item calibrations could be linked.

The procedure used for post-equating the CSTs for Science prior to 2013 involved three steps: item calibration, item parameter scaling, and true score equating. Each of those steps, as described below, was applied to all CSTs for Science during the tests’ original years of administration. Results were not post-equated for the 2015–16 administration.

Pre-Equating

During the 2015–16 administration, because all the test items were used in previous operational administrations, pre-equating was conducted prior to administration of the tests. Based on the sample invariant property of IRT, all the item parameter estimates were placed on the reference scale in their previous administrations through the post-equating procedure described previously. Item parameters derived in such a manner can be used to create raw-score-to-scale-score conversion tables prior to test administration. Neither calibration nor scaling was implemented in the pre-equating process.

Since all CSTs for Science were intact forms without any edits or replacement to items, the original conversion tables from the previous administrations when the forms were originally used are directly applied to the current administration.

Table 8.4 on page 90 shows the years the forms were introduced for each test.

Calibration

To conduct item calibrations during the initial administration of each form, a proprietary version of the PARSCALE program was used. The estimation process was constrained by setting a common discrimination value for all items equal to 1.0 / 1.7 (or 0.588) and by setting the lower asymptote for all multiple-choice items to zero. The resulting estimation was equivalent to the Rasch model for multiple-choice items. This approach was in line with previous CST for Science equating and scaling procedures achieved using the WINSTEPS program (Linacre, 2000). For the purpose of equating, only the operational items were calibrated for each test.

The PARSCALE calibrations were run in two stages following procedures used with other ETS testing programs. In the first stage, estimation imposed normal constraints on the updated prior-ability distribution. The estimates resulting from this first stage were used as starting values for a second PARSCALE run, in which the subject prior distribution was updated after each expectation maximization cycle with no constraints. For both stages, the metric of the scale was controlled by the constant discrimination parameters.

Scaling

In the years when the new forms were developed prior to the 2012–13 administration, calibrations of the items were linked to the previously obtained reference scale estimates using linking items and the Stocking and Lord (1983) procedure. In the case of the one-parameter model calibrations, this procedure was equivalent to setting the mean of the new item parameter estimates for the linking set equal to the mean of the previously scaled estimates. As noted earlier, the linking set was a collection of items in a current test form that also appeared in the previous year's form and was scaled at that time.

The linking process was carried out iteratively by inspecting differences between the transformed new and old (reference) estimates for the linking items and removing items for which the item difficulty estimates changed significantly. Items with large weighted root-mean-square differences (WRMSDs) between item characteristic curves based on the old and new difficulty estimates were removed from the linking set. The differences were calculated using the following formula:

$$WRMSD = \sqrt{\sum_{j=1}^{n_g} w_j \left[P_n\left(\theta_j\right) - P_r\left(\theta_j\right) \right]^2} \quad (2.1)$$

where,

abilities are grouped into intervals of 0.005 ranging from –3.0 to 3.0,

n_g is the number of intervals/groups,

θ_j is the mean of the ability estimates that fall in interval j,

w_j is a weight equal to the proportion of estimated abilities from the transformed new form in interval j,

$P_n(\theta_j)$ is the probability of correct response for the transformed new form item at ability θ_j, and

$P_r(\theta_j)$ is the probability of correct response for the old (reference) form item at ability θ_j.

Based on established procedures, any linking items for which the WRMSD was greater than 0.125 were eliminated from the linking set. This criterion has produced reasonable results over time in similar equating work done with other testing programs at ETS.

True-score Equating

Once the new item calibrations for each test were placed on the base scale after scaling, IRT true-score equating procedures were used to transform the new form number-correct scores (raw scores) to their respective reference form number-correct scale. These converted raw scores could then be transformed to scale scores through table lookup and linear interpolation.

The true-score equating procedure is based on the relationship between raw scores and ability (theta). For the CSTs for Science, which consist entirely of *n* multiple-choice items, this is the well-known relationship defined in Lord (1980; equations 4–5):

$$\xi(\theta) = \sum_{i=1}^{n} P_i(\theta) \tag{2.2}$$

where,

$P_i(\theta)$ is the probability of a correct response to item i at ability θ, and

$\xi(\theta)$ is the corresponding true score.

For each integer score ξ_n on the form after its original use, the true-score equating procedure was used to first solve for the corresponding ability estimate using equation 2.2. The procedure used this ability estimate to find the corresponding number-correct true score ξ_b on the reference form. Finally, each score ξ_b was transformed to the appropriate CST for Science scale score scale using the reference form CST for Science raw-score-to-scale-score conversion tables and linear interpolation. Complete raw-to-scale-score conversion tables for the 2015–16 CSTs for Science are presented in Table 8.C.1 through Table 8.C.5 in Appendix 8.C, starting on page 111. The raw scores and corresponding transformed scale scores are listed in those tables.

For all of the CSTs for Science, regardless of when the form was administered, scale scores were adjusted at both ends of the scale so that the minimum reported scale score was 150 and the maximum reported scale score was 600. Raw scores of zero and perfect raw scores were assigned scale scores of 150 and 600, respectively.

The scale-score ranges defining the various performance levels are presented in Table 2.1.

Table 2.1 Scale-Score Ranges for Performance Levels

CST	Far Below Basic	Below Basic	Basic	Proficient	Advanced
Grade 5 Science	150 – 267	268 – 299	300 – 349	350 – 409	410 – 600
Grade 8 Science	150 – 252	253 – 299	300 – 349	350 – 402	403 – 600
Grade 10 Life Science	150 – 268	269 – 299	300 – 349	350 – 398	399 – 600

The next section describes characteristics of the samples used to establish the 2002 reference scales as well as the equating samples used to equate the CSTs for Science in subsequent years.

Equating Samples

To establish the 2002 reference scales, ETS staff used data based on samples of students selected from the 2001–02 administration for each CST for Science. In drawing these

samples, it was necessary to account for the small portion of the complete testing data available at the time of equating. To simulate the situation in which only schools that test early are used in an equating, the complete CST for Science data were sorted according to the test administration date shown in the student records. Only students tested before a selected cutoff date were chosen. Ten thousand test-takers were randomly sampled from all available records.

As of 2003, equating samples were selected from available student records in a data file obtained near the end of May. As anticipated, these data comprised only 5 to 10 percent of the total CAASPP testing data that were available once testing was completed. It was necessary to use these partial student samples for equating to meet score reporting deadlines. Only test-takers with valid results on the CSTs for Science are included in the equating samples.

Due to the implementation of the pre-equating, no equating sample is necessary for the 2015–16 administration.

Equating the Braille Versions of the CSTs for Science

In some cases, it is not possible to translate all of the operational items contained in a CST for Science into braille. This situation requires that a new conversion table be developed for the resulting shortened test. To obtain this table, the shortened test is equated to the full-length operational test being used using the IRT equating methods described previously. This process ensures that the scaled cut scores established for the full-length test are used to classify students who take the shorter test.

For the 2015–16 administration, the CSTs for Science in grades five and eight had a braille version that was equated using a shortened test because of two items in the CST for Science (Grade 8) and one item in the CST for Science (Grade 5) that could not be translated into braille.

References

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

California Department of Education. (2016b). Universal Tools, Designated Supports, and Accommodations for the California Assessment of Student Performance and Progress. Sacramento, CA. http://www.cde.ca.gov/ta/tg/ai/caasppmatrix1.asp

Educational Testing Service. (2002). *ETS standards for quality and fairness*. Princeton, NJ: Author.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: principles and applications*. Boston, MA: Kluwer-Nijhoff.

Linacre, J. M. (2000). *WINSTEPS: Rasch measurement* (Version 3.23). Chicago, IL: MESA Press.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201–10.

Appendix 2.A—CST for Science Items and Estimated Time Chart

California Standards Tests	Grade 5		Grade 8		Grade 10	
	Total No. of Items	Estimated Maximum Time	Total No. of Items	Estimated Maximum Time	Total No.	Estimated Maximum Time
Science—Grade level	66	**140**	66	**120**	66	**120**
Part 1		70		60		60
Part 2		70		60		60

Appendix 2.B—Reporting Clusters for Science

Science Standards Test (Grade Five)

Physical Science	
Grade Five, Standards: 5PS1.a–i	11 items
Grade Four, Standards: 4PS1.a–g and 4IE6.a–f	8 items
Life Science	
Grade Five, Standards: 5LS2.a–g and 5IE6.a–i	13 items
Grade Four, Standards: 4LS2.a–c and 4LS3.a–d	9 items
Earth Science	
Grade Five, Standards: 5ES3.a–e, 5ES4.a–e, and 5ES5.a–c	11 items
Grade Four, Standards: 4ES4.a–b, 4ES5.a–c, and 4IE6.a–f	8 items

Science Standards Test (Grade Eight)

Motion	
Standards: 8PC1.a–f	8 items
Forces, Density, and Buoyancy	
Standards: 8PC2.a–g, 8PC8.a–d	13 items
Structure of Matter and Periodic Table	
Standards: 8PC3.a–f, 8PC7.a–c	16 items
Earth in the Solar System	
Standards: 8PC4.a–e	7 items
Reactions and the Chemistry of Living Systems	
Standards: 8PC5.a–e, 8PC6.a–c	10 items
Investigation and Experimentation	
Standards: 8PCIE9.a–g	6 items

Life Science Standards Test (Grade Ten)

Cell Biology	
Standards: 7SL1.c–e, 8PC6.b–c, and BI1.a.c.f	10 items
Genetics	
Standards: 7LS2.a, 7LS2.c–e, BI2.b, BI2.d–f, BI3.a, and BI5.a	12 items
Physiology	
Standards: 7LS5.a, 7LS5.c, 7LS6.j, BI9.a–b, and BI10.b–d	10 items
Ecology	
Standards: 6LS5.b–c, 6LS5.e, and BI6.a–f	11 items
Evolution	
Standards: 7LS3.a–c, BI7.a–d, BI8.a–b, and BI8.e	11 items
Investigation and Experimentation	
Standards: 6LSIE7.c, 6LSIE7.e, 7LSIE7.c, 8PCIE9.b–c, BIIE1.c, BIIE1.f, BIIE1.i–j	6 items

Appendix 2.C—Universal Tools, Designated Supports, and Accommodations for the California Assessment of Student Performance and Progress

Table 2.C.1 Matrix One Part 2: Non-Embedded Supports for the CSTs for Science

Option (U) Universal Tool | (D) Designated Support | (A) Accommodation

Answer Document Section A3—Accommodations and Modifications		
B	Pupil marks in paper-pencil test booklet (other than responses including highlighting)	U
C	Scribe	A
G	Braille (paper-pencil tests)	A
H	Large-print versions of a paper-pencil test (as available)	A
J	Breaks (Tested over more than one day)	U
K	Breaks (Supervised breaks within a section of the test)	U
L	Administration of the test to the pupil at the most beneficial time of day	A
M	Administered at home or in a hospital	A
O	American Sign Language	A
X	Abacus	A
Y	Unlisted resource	–
Z	Read aloud	A
Option	**(U) Universal Tool | (D) Designated Support | (A) Accommodation**	
Answer Document Section A4—English Learner (EL) Test Variations		
A	Translated test directions	D
B	Additional supervised breaks within a testing day or following each section within a test part provided that the test section is completed within a testing day. A test section is identified by a "STOP" at the end of it.	Unmapped
C	Tested separately with other English learners and was supervised directly by an employee of the school who had signed a CAASPP Test Security Affidavit. The student has been provided such a flexible setting as a part of his or her regular instruction or assessment during the school year.	Unmapped
D	Used a translation glossary/word list (English-to-primary language). Glossaries/Word lists shall not include definitions, parts of speech, or formulas.	D

Universal Tools (U)	Are available for all pupils. Pupils may turn the support(s) on/off when embedded as part of the technology platform for the computer-administered CAASPP tests or may choose to use it/them when provided as part of a paper-pencil test.
Designated Supports (D)	Are features that are available for use by any pupil for whom the need has been indicated prior to the assessment, by an educator or group of educators.
Accommodations (A)	For the CAASPP System, eligible pupils shall be permitted to take the tests with accommodations if specified in the pupil's individualized educational program (IEP) or Section 504 plan.

Note: The use of additional accessibility supports can be requested.

Appendix 2.D—Special Services Summary Tables

Notes:

1. To improve clarity of tables presented in this section, the columns with total number of students using each service are labeled with the particular grade or test name for which the services were utilized. For example, the column with a heading of "Grade 5 Number" in these tables present the number of students using various special services on the CST for Science in grade five. The column with the heading of "Grade 5 Pct. of Total" in the same table represents the percent of students using a service out of the total number of test-takers.
2. The total number of test-takers is the total of students listed under "*Any* universal tool, desig. support, or accommodation or EL variation" and those listed under "*No* universal tool, desig. support, or accommodation or EL variation."
3. The sum of the numbers of students across subgroups may not match exactly to the total testing population, due to the fact that only valid primary disability codes were chosen to identify those subgroups.
4. The notation "N/A" is inserted where frequencies for certain accommodations or supports that are not presented in the data. These accommodations or supports include "B: Marked responses in test booklet," "J: Breaks (Tested over more than one day)," "K: Breaks (Had supervised breaks)," "O: American Sign Language," "EL Test Variation B," and "EL Test Variation C."

Table 2.D.1 Special Services Summary for Science, Grades Five, Eight, and Ten (Life Science)—All Tested

Answer Document Option	Grade 5 Number	Grade 5 Pct. of Total	Grade 8 Number	Grade 8 Pct. of Total	Grade 10 Life Sci. Number	Grade 10 Pct. of Total
B: Marked responses in test booklet	N/A	N/A	N/A	N/A	N/A	N/A
C: Scribe	32	0.01%	24	0.01%	38	0.01%
G: Braille	15	0.00%	16	0.00%	19	0.00%
H: Large-print versions of a paper-pencil test	61	0.01%	56	0.01%	58	0.01%
J: Breaks (Tested over more than one day)	N/A	N/A	N/A	N/A	N/A	N/A
K: Breaks (Had supervised breaks)	N/A	N/A	N/A	N/A	N/A	N/A
L: Most beneficial time of day	1,247	0.28%	753	0.17%	608	0.13%
M: Administered at home or in a hospital	32	0.01%	48	0.01%	75	0.02%
O: American Sign Language	N/A	N/A	N/A	N/A	N/A	N/A
X: Abacus	32	0.01%	28	0.01%	66	0.01%
Y: Unlisted resource	28	0.01%	11	0.00%	12	0.00%
Z: Read aloud	2,991	0.67%	924	0.21%	632	0.14%
Univ. tool, desig. support, or acc. is in Section 504 plan	127	0.03%	63	0.01%	67	0.01%
Univ. tool, desig. support, or acc. is in IEP	3,693	0.83%	1,574	0.36%	1,252	0.27%
EL Test Variation A	118	0.03%	180	0.04%	736	0.16%
EL Test Variation B	N/A	N/A	N/A	N/A	N/A	N/A
EL Test Variation C	N/A	N/A	N/A	N/A	N/A	N/A
EL Test Variation D	19	0.00%	51	0.01%	819	0.18%
Any universal tool, desig. support, or accommodation or EL variation	4,075	0.92%	1,949	0.45%	2,805	0.61%
No universal tool, desig. support, or accommodation or EL variation	440,964	99.08%	435,574	99.55%	458,465	99.39%

Table 2.D.2 Special Services Summary for Science, Grades Five, Eight, and Ten (Life Science)—Students Not in Special Education

Answer Document Option	Grade 5 Number	Grade 5 Pct. of Total	Grade 8 Number	Grade 8 Pct. of Total	Grade 10 Life Sci. Number	Grade 10 Pct. of Total
B: Marked responses in test booklet	N/A	N/A	N/A	N/A	N/A	N/A
C: Scribe	9	0.00%	4	0.00%	12	0.00%
G: Braille	2	0.00%	1	0.00%	2	0.00%
H: Large-print versions of a paper-pencil test	7	0.00%	8	0.00%	5	0.00%
J: Breaks (Tested over more than one day)	N/A	N/A	N/A	N/A	N/A	N/A
K: Breaks (Had supervised breaks)	N/A	N/A	N/A	N/A	N/A	N/A
L: Most beneficial time of day	122	0.03%	73	0.02%	95	0.02%
M: Administered at home or in a hospital	7	0.00%	23	0.01%	38	0.01%
O: American Sign Language	N/A	N/A	N/A	N/A	N/A	N/A
X: Abacus	6	0.00%	6	0.00%	8	0.00%
Y: Unlisted resource	1	0.00%	0	0.00%	0	0.00%
Z: Read aloud	129	0.03%	52	0.01%	33	0.01%
Univ. tool, desig. support, or acc. is in Section 504 plan	100	0.02%	44	0.01%	56	0.01%
Univ. tool, desig. support, or acc. is in IEP	0	0.00%	0	0.00%	0	0.00%
EL Test Variation A	111	0.03%	176	0.04%	702	0.16%
EL Test Variation B	N/A	N/A	N/A	N/A	N/A	N/A
EL Test Variation C	N/A	N/A	N/A	N/A	N/A	N/A
EL Test Variation D	18	0.00%	49	0.01%	720	0.17%
Any universal tool, desig. support, or accommodation or EL variation	378	0.09%	370	0.09%	1,439	0.34%
No universal tool, desig. support, or accommodation or EL variation	412,507	99.91%	407,505	99.91%	424,747	99.66%

Table 2.D.3 Special Services Summary for Science, Grades Five, Eight, and Ten (Life Science)—Students in Special Education

Answer Document Option	Grade 5 Number	Grade 5 Pct. of Total	Grade 8 Number	Grade 8 Pct. of Total	Grade 10 Life Sci. Number	Grade 10 Pct. of Total
B: Marked responses in test booklet	N/A	N/A	N/A	N/A	N/A	N/A
C: Scribe	23	0.07%	20	0.07%	26	0.07%
G: Braille	13	0.04%	15	0.05%	17	0.05%
H: Large-print versions of a paper-pencil test	54	0.17%	48	0.16%	53	0.15%
J: Breaks (Tested over more than one day)	N/A	N/A	N/A	N/A	N/A	N/A
K: Breaks (Had supervised breaks)	N/A	N/A	N/A	N/A	N/A	N/A
L: Most beneficial time of day	1,125	3.50%	680	2.29%	513	1.46%
M: Administered at home or in a hospital	25	0.08%	25	0.08%	37	0.11%
O: American Sign Language	N/A	N/A	N/A	N/A	N/A	N/A
X: Abacus	26	0.08%	22	0.07%	58	0.17%
Y: Unlisted resource	27	0.08%	11	0.04%	12	0.03%
Z: Read aloud	2,862	8.90%	872	2.94%	599	1.71%
Univ. tool, desig. support, or acc. is in Section 504 plan	27	0.08%	19	0.06%	11	0.03%
Univ. tool, desig. support, or acc. is in IEP	3,693	11.49%	1,574	5.31%	1,252	3.57%
EL Test Variation A	7	0.02%	4	0.01%	34	0.10%
EL Test Variation B	N/A	N/A	N/A	N/A	N/A	N/A
EL Test Variation C	N/A	N/A	N/A	N/A	N/A	N/A
EL Test Variation D	1	0.00%	2	0.01%	99	0.28%
Any universal tool, desig. support, or accommodation or EL variation	3,697	11.50%	1,579	5.33%	1,366	3.89%
No universal tool, desig. support, or accommodation or EL variation	28,457	88.50%	28,069	94.67%	33,718	96.11%

Table 2.D.4 Special Services Summary for Science, Grades Five, Eight, and Ten (Life Science)—English-Only Students

Answer Document Option	Grade 5 Number	Grade 5 Pct. of Total	Grade 8 Number	Grade 8 Pct. of Total	Grade 10 Life Sci. Number	Grade 10 Pct. of Total
B: Marked responses in test booklet	N/A	N/A	N/A	N/A	N/A	N/A
C: Scribe	26	0.01%	17	0.01%	15	0.01%
G: Braille	5	0.00%	6	0.00%	6	0.00%
H: Large-print versions of a paper-pencil test	42	0.02%	33	0.01%	26	0.01%
J: Breaks (Tested over more than one day)	N/A	N/A	N/A	N/A	N/A	N/A
K: Breaks (Had supervised breaks)	N/A	N/A	N/A	N/A	N/A	N/A
L: Most beneficial time of day	782	0.31%	492	0.21%	418	0.17%
M: Administered at home or in a hospital	23	0.01%	32	0.01%	47	0.02%
O: American Sign Language	N/A	N/A	N/A	N/A	N/A	N/A
X: Abacus	25	0.01%	17	0.01%	31	0.01%
Y: Unlisted resource	21	0.01%	6	0.00%	7	0.00%
Z: Read aloud	1,786	0.71%	549	0.23%	395	0.16%
Univ. tool, desig. support, or acc. is in Section 504 plan	105	0.04%	55	0.02%	56	0.02%
Univ. tool, desig. support, or acc. is in IEP	2,229	0.89%	966	0.41%	784	0.31%
EL Test Variation A	3	0.00%	4	0.00%	4	0.00%
EL Test Variation B	N/A	N/A	N/A	N/A	N/A	N/A
EL Test Variation C	N/A	N/A	N/A	N/A	N/A	N/A
EL Test Variation D	1	0.00%	1	0.00%	2	0.00%
Any universal tool, desig. support, or accommodation or EL variation	2,407	0.96%	1,083	0.45%	911	0.37%
No universal tool, desig. support, or accommodation or EL variation	247,811	99.04%	237,136	99.55%	248,002	99.63%

Table 2.D.5 Special Services Summary for Science, Grades Five, Eight, and Ten (Life Science)—Initially Fluent English Proficient (I-FEP) Students

Answer Document Option	Grade 5 Number	Grade 5 Pct. of Total	Grade 8 Number	Grade 8 Pct. of Total	Grade 10 Life Sci. Number	Grade 10 Pct. of Total
B: Marked responses in test booklet	N/A	N/A	N/A	N/A	N/A	N/A
C: Scribe	0	0.00%	0	0.00%	3	0.01%
G: Braille	2	0.01%	1	0.00%	3	0.01%
H: Large-print versions of a paper-pencil test	1	0.01%	0	0.00%	10	0.03%
J: Breaks (Tested over more than one day)	N/A	N/A	N/A	N/A	N/A	N/A
K: Breaks (Had supervised breaks)	N/A	N/A	N/A	N/A	N/A	N/A
L: Most beneficial time of day	12	0.06%	12	0.06%	15	0.04%
M: Administered at home or in a hospital	0	0.00%	4	0.02%	3	0.01%
O: American Sign Language	N/A	N/A	N/A	N/A	N/A	N/A
X: Abacus	1	0.01%	1	0.00%	0	0.00%
Y: Unlisted resource	0	0.00%	0	0.00%	0	0.00%
Z: Read aloud	22	0.12%	10	0.05%	8	0.02%
Univ. tool, desig. support, or acc. is in Section 504 plan	3	0.02%	0	0.00%	4	0.01%
Univ. tool, desig. support, or acc. is in IEP	30	0.16%	26	0.12%	31	0.09%
EL Test Variation A	1	0.01%	0	0.00%	3	0.01%
EL Test Variation B	N/A	N/A	N/A	N/A	N/A	N/A
EL Test Variation C	N/A	N/A	N/A	N/A	N/A	N/A
EL Test Variation D	0	0.00%	0	0.00%	1	0.00%
Any universal tool, desig. support, or accommodation or EL variation	36	0.19%	27	0.13%	41	0.12%
No universal tool, desig. support, or accommodation or EL variation	18,484	99.81%	20,960	99.87%	33,532	99.88%

Table 2.D.6 Special Services Summary for Science, Grades Five, Eight, and Ten (Life Science)—English Learner (EL) Students

Answer Document Option	Grade 5 Number	Grade 5 Pct. of Total	Grade 8 Number	Grade 8 Pct. of Total	Grade 10 Life Sci. Number	Grade 10 Pct. of Total
B: Marked responses in test booklet	N/A	N/A	N/A	N/A	N/A	N/A
C: Scribe	4	0.00%	5	0.01%	14	0.03%
G: Braille	4	0.00%	3	0.01%	3	0.01%
H: Large-print versions of a paper-pencil test	12	0.01%	10	0.02%	11	0.02%
J: Breaks (Tested over more than one day)	N/A	N/A	N/A	N/A	N/A	N/A
K: Breaks (Had supervised breaks)	N/A	N/A	N/A	N/A	N/A	N/A
L: Most beneficial time of day	405	0.46%	164	0.33%	108	0.21%
M: Administered at home or in a hospital	8	0.01%	8	0.02%	10	0.02%
O: American Sign Language	N/A	N/A	N/A	N/A	N/A	N/A
X: Abacus	5	0.01%	8	0.02%	29	0.06%
Y: Unlisted resource	6	0.01%	3	0.01%	4	0.01%
Z: Read aloud	1,082	1.22%	262	0.52%	164	0.33%
Univ. tool, desig. support, or acc. is in Section 504 plan	13	0.01%	5	0.01%	1	0.00%
Univ. tool, desig. support, or acc. is in IEP	1,298	1.46%	395	0.79%	301	0.60%
EL Test Variation A	99	0.11%	144	0.29%	671	1.34%
EL Test Variation B	N/A	N/A	N/A	N/A	N/A	N/A
EL Test Variation C	N/A	N/A	N/A	N/A	N/A	N/A
EL Test Variation D	17	0.02%	42	0.08%	750	1.49%
Any universal tool, desig. support, or accommodation or EL variation	1,460	1.64%	603	1.21%	1,582	3.15%
No universal tool, desig. support, or accommodation or EL variation	87,306	98.36%	49,387	98.79%	48,679	96.85%

Table 2.D.7 Special Services Summary for Science, Grades Five, Eight, and Ten (Life Science)—Reclassified Fluent English Proficient (R-FEP) Students

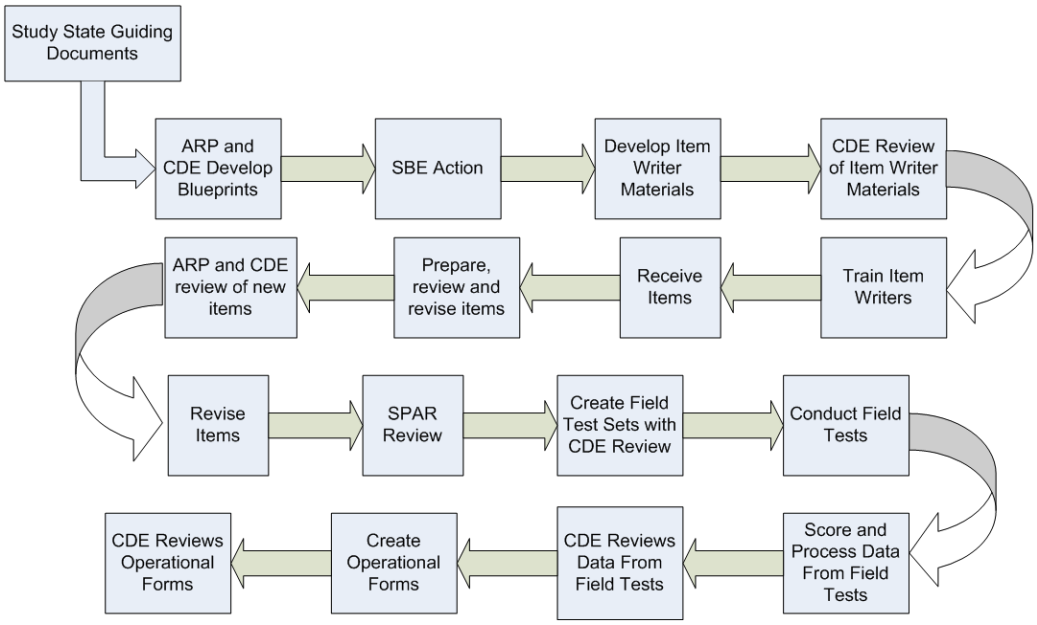
Answer Document Option	Grade 5 Number	Grade 5 Pct. of Total	Grade 8 Number	Grade 8 Pct. of Total	Grade 10 Life Sci. Number	Grade 10 Pct. of Total
B: Marked responses in test booklet	N/A	N/A	N/A	N/A	N/A	N/A
C: Scribe	2	0.00%	2	0.00%	6	0.00%
G: Braille	4	0.00%	6	0.00%	7	0.01%
H: Large-print versions of a paper-pencil test	6	0.01%	13	0.01%	11	0.01%
J: Breaks (Tested over more than one day)	N/A	N/A	N/A	N/A	N/A	N/A
K: Breaks (Had supervised breaks)	N/A	N/A	N/A	N/A	N/A	N/A
L: Most beneficial time of day	48	0.06%	82	0.06%	66	0.05%
M: Administered at home or in a hospital	1	0.00%	4	0.00%	15	0.01%
O: American Sign Language	N/A	N/A	N/A	N/A	N/A	N/A
X: Abacus	1	0.00%	2	0.00%	6	0.00%
Y: Unlisted resource	1	0.00%	2	0.00%	1	0.00%
Z: Read aloud	94	0.11%	103	0.08%	64	0.05%
Univ. tool, desig. support, or acc. is in Section 504 plan	6	0.01%	3	0.00%	6	0.00%
Univ. tool, desig. support, or acc. is in IEP	131	0.15%	186	0.15%	135	0.11%
EL Test Variation A	2	0.00%	8	0.01%	26	0.02%
EL Test Variation B	N/A	N/A	N/A	N/A	N/A	N/A
EL Test Variation C	N/A	N/A	N/A	N/A	N/A	N/A
EL Test Variation D	1	0.00%	2	0.00%	35	0.03%
Any universal tool, desig. support, or accommodation or EL variation	152	0.18%	206	0.16%	220	0.17%
No universal tool, desig. support, or accommodation or EL variation	86,331	99.82%	126,997	99.84%	127,058	99.83%

Chapter 3: Item Development

The intact test forms from the 2011–12 test administration were reused during the 2015–16 administration. Using an intact form permits the original score conversion tables from the previous administration to be used to look up student scores and performance levels. There was no new item development for the 2015–16 forms.

The California Standards Test (CST) for Science items were developed to measure California's content standards and designed to conform to principles of item writing defined by Educational Testing Service (ETS) (ETS, 2002). Each CST for Science item on the intact forms used in the 2015–16 administration went through a comprehensive development cycle as is described in Figure 3.1 below.

Figure 3.1 The ETS Item Development Process for the CAASPP System



Rules for Item Development

Educational Testing Service (ETS) maintained item development specifications for each CST for Science and developed an item utilization plan to guide the development of the items for each content area. Item writing emphasis was determined in consultation with the California Department of Education (CDE).

Item Specifications

The item specifications described the characteristics of the items that should be written to measure each content standard; items of the same type should consistently measure the content standards in the same way. To achieve this, the item specifications provided detailed information to item writers who developed items for the CSTs for Science. The specifications included the following:

- A full statement of each academic content standard, as defined by the State Board of Education (SBE) in 1998 (CDE, 2009)
- A description of each content strand

- The expected depth of knowledge (DOK) measured by items written for each standard (coded as 1, 2, 3, or 4; items assigned a DOK of 1 are the least cognitively complex, items assigned a DOK of 3 are the most cognitively complex, and the code of 4 would apply only to some writing tasks)
- The homogeneity of the construct measured by each standard
- A description of the kinds of item stems appropriate for multiple-choice items used to assess each standard
- A description of the kinds of distractors that are appropriate for multiple-choice items assessing each standard
- A description of appropriate data representations (such as charts, tables, graphs, or other illustrations) for mathematics, science, and history–social science items
- The content limits for the standard (such as one or two variables, maximum place values of numbers) for mathematics, science, and history–social science items
- A description of appropriate reading passages, where applicable, for English–language arts (ELA) items
- A description of specific kinds of items to be avoided, if any (for example, items with any negative expressions in the stem, e.g., "Which of the following is NOT. . .")

Expected Item Ratio

ETS prepared the item utilization plan for the development of CST for Science items. The plan included strategies for developing items that permitted coverage of all appropriate standards for all tests in each content area and at each grade level. ETS test development staff used this plan to determine the number of items to develop for each content area. Because item development has been halted, the item utilization plan is no longer necessary.

The item utilization plan assumed that each year, 35 percent of items on an operational form would be refreshed (replaced); these items would remain in the item bank for future use. The plan also declared that an additional five percent of the operational items were likely to become unusable because of normal attrition and noted a need to focus development on "critical" standards, which are those that were difficult to measure well or for which there were few usable items.

It was assumed that at least 60 percent of all field-tested science items were expected to have acceptable field-test statistics and become candidates for use in operational tests.

For the 2015–16 CST for Science administration, field-test items were repeated as a part of the reuse of the intact forms.

Selection of Item Writers

Criteria for Selecting Item Writers

The items for each CST for Science were written by individual item writers with a thorough understanding of the California content standards adopted in 1998. Applicants for item writing were screened by senior ETS content staff. Only those with strong content and teaching backgrounds were approved for inclusion in the training program for item writers. Because most of the participants were current or former California educators, they were particularly knowledgeable about the standards assessed by the CSTs for Science. All item writers met the following minimum qualifications:

- Possession of a Bachelor's degree in the relevant content area or in the field of education with special focus on a particular content of interest; an advanced degree in the relevant content area is desirable
- Previous experience in writing items for standards-based assessments, including knowledge of the many considerations that are important when developing items to match state-specific standards
- Previous experience in writing items in the content areas covered by CST for Science grades and/or courses
- Familiarity, understanding, and support of the California content standards
- Current or previous teaching experience in California, when possible

Item Review Process

The items selected for each CST for Science underwent an extensive item review process that was designed to provide the best standards-based tests possible. This section summarizes the various reviews performed that ensure the quality of the CST for Science items and test forms—currently being reused—at the time the items and forms were developed. See Table 8.4 on page 90 for the dates of the previous administrations. It should also be noted that two items on the CST for Life Science (Grade 10) were replaced in the form due to security breaches on social media Web sites.

Contractor Review

Once the items were written, ETS employed a series of internal reviews. The reviews established the criteria used to judge the quality of the item content and were designed to ensure that each item measured what it was intended to measure. The internal reviews also examined the overall quality of the test items before they were prepared for presentation to the CDE and the Assessment Review Panels (ARPs). Because of the complexities involved in producing defensible items for high-stakes programs such as the California Assessment of Student Performance and Progress (CAASPP) System, it was essential that many experienced individuals reviewed each item before it was brought to the CDE, the ARPs, and Statewide Pupil Assessment Review (SPAR) panels.

The ETS review process for the CSTs for Science included the following:

1. Internal content review
2. Internal editorial review
3. Internal sensitivity review

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

1. Internal Content Review

Test items and materials underwent two reviews by the content-area assessment specialists. These assessment specialists made sure that the test items and related materials were in compliance with ETS's written guidelines for clarity, style, accuracy, and appropriateness for California students as well as in compliance with the approved item specifications. 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 item specifications, including 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 a code for the standard it was intended to measure. The assessment specialists checked all items against their classification codes, both to evaluate the correctness of the classification and to ensure that the task posed by the item was relevant to the outcome it was intended to measure. The reviewers could 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.

2. Internal Editorial Review

After the content-area assessment specialists reviewed each item, a group of specially trained editors also reviewed each item in preparation for consideration by the CDE and the ARPs. 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. Internal Sensitivity Review

ETS assessment specialists who are specially trained to identify and eliminate questions that contain content or wording that could be construed to be offensive to or biased against members of specific ethnic, racial, or gender groups conducted the next level of review. These trained staff members reviewed every item before the CDE and ARP 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

Content Expert Reviews

Assessment Review Panels

ETS was responsible for working with ARPs as items were developed for the CSTs for Science. The ARPs are advisory panels to the CDE and ETS and provided guidance on matters related to item development for the CSTs for Science. The ARPs were responsible for reviewing all newly developed items for alignment to the California content standards;

these tests use the content standards for science adopted by the SBE in 1998. The ARPs also reviewed the items for accuracy of content, clarity of phrasing, and quality. In their examination of test items, the ARPs could raise concerns related to age/grade appropriateness and gender, racial, ethnic, and/or socioeconomic bias.

Composition of ARPs

The ARPs comprised current and former California teachers, resource specialists, administrators, curricular experts, and other education professionals. Current school staff members met minimum qualifications to serve on the CST ARPs, including:

- Three or more years of general teaching experience in grades kindergarten through twelve and in the relevant content areas (ELA, history–social science, mathematics, or science);
- Bachelor's or higher degree in a grade or content area related to ELA, history–social science, mathematics, or science; and
- Knowledge and experience with the California content standards in ELA, history–social science, mathematics, or science that are current at the time.

School administrators, local educational agency (LEA)/county content/program specialists, or university educators serving on the CST ARPs met the following qualifications:

- 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 ELA, history–social science, mathematics, or science;
- Bachelor's or higher degree in a grade-specific or subject area related to ELA, history–social science, mathematics, or science; and
- Knowledge of and experience with the California content standards in ELA, history–social science, mathematics, or science that are current at the time.

Every effort was made to ensure that ARP committees included representation of genders and of the geographic regions and ethnic groups in California. Efforts were also made to ensure representation by members with experience serving California's diverse special education population.

ARP members were recruited through an application process. Recommendations were solicited from LEAs and county offices of education as well as from CDE and SBE staff. Applications were reviewed by the ETS assessment directors, who 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 ARP membership.

ARP members were employed as teachers, program specialists, university personnel, and LEA personnel, had a minimum of a bachelor's degree, and had experience teaching students, whether in a classroom setting or one-on-one.

ARP Meetings for Review of CST for Science Items

ETS content-area assessment specialists facilitated the CST for Science ARP meetings. Each meeting began with a brief training session on how to review items. ETS provided this training, which consisted of the following topics:

- Overview of the purpose and scope of the CSTs for Science
- Overview of the test design specifications and blueprints for the CSTs for Science
- Analysis of the item specifications for the CSTs for Science

- Overview of criteria for evaluating multiple-choice test items and for reviewing constructed response writing tasks
- Review and evaluation of items for bias and sensitivity issues

The criteria for evaluating multiple-choice items included the following:

- Overall technical quality
- Match to the California content standards (For the CSTs for Science, these are the content standards for science adopted by the SBE in 1998.)
- Match to the construct being assessed by the standard
- Difficulty range
- Clarity
- Correctness of the answer
- Plausibility of the distractors
- Bias and sensitivity factors

Criteria also included more global factors, including—for ELA—the appropriateness, difficulty, and readability of reading passages. The ARPs also were trained on how to make recommendations for revising items.

Guidelines for reviewing items were provided by ETS and approved by the CDE. The set of guidelines for reviewing items is summarized below.

Does the item:

- Have one and only one clearly correct answer?
- Measure the content standard?
- Match the test item specifications?
- Align with the construct being measured?
- Test worthwhile concepts or information?
- Reflect good and current teaching practices?
- Have a stem that gives the student a full sense of what the item is asking?
- Avoid unnecessary wordiness?
- Use response options that relate to the stem in the same way?
- Use response options that are plausible and have reasonable misconceptions and errors?
- Avoid having one response option that is markedly different from the others?
- Avoid clues to students, such as absolutes or words repeated in both the stem and options?
- Reflect content that is free of bias against any person or group?

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?

As the first step of the item review process, ARP members reviewed a set of items independently and recorded their individual comments. The next step in the review process was for the group to discuss each item; the content-area assessment specialists facilitated the discussion and recorded all recommendations in a master item review booklet. Item review binders and other item evaluation materials also identified potential bias and sensitivity factors for the ARP to consider as a part of its item reviews.

Depending on CDE approval and the numbers of items still to be reviewed, some ARPs were divided further into smaller groups. The science ARP, for example, divided into content-area and grade-level groups. These smaller groups were also facilitated by the content-area assessment specialists.

ETS staff maintained the minutes summarizing the review process and then forwarded copies of the minutes to the CDE, emphasizing in particular the recommendations of the panel members.

Statewide Pupil Assessment Review (SPAR) Panel

The SPAR panel is responsible for reviewing and approving all achievement test items to be used statewide for the testing of students in California public schools, grades two through eleven. At the SPAR panel meetings, all new items were presented in binders for review. The SPAR panel representatives ensured that the test items conformed to the requirements of *Education Code* Section 60602 (a) (3). If the SPAR panel rejected specific items, the items were marked for rejection in the item bank and excluded from use on field tests. For the SPAR panel meeting, the item development coordinator was available by telephone to respond to any questions during the course of the meeting.

SPAR panelists were selected by the CDE and/or the office of the State Superintendent of Public Instruction.

Field Testing

The primary purposes of field testing are to obtain information about item performance and to obtain statistics that can be used to assemble operational forms. However, because the intact forms were used with the original field-test items intact for the 2015–16 CAASPP administration, data were not analyzed for current field-test items.

Stand-alone Field Testing

For each new CST for Science launched, a pool of items was initially constructed by administering the newly developed items in a stand-alone field test. In stand-alone field testing, students are recruited to take tests outside of the usual testing circumstances, and the test results are typically not used for instructional or accountability purposes (Schmeiser & Welch, 2006).

Embedded Field-test Items

Although a stand-alone field test is useful for developing a new test because it can produce a large pool of quality items, embedded field testing is generally preferred because the items being field-tested are seeded throughout the operational test. Variables such as test-taker motivation and test security are the same in embedded field testing as they will be when the field-tested items are later administered operationally.

Such field testing involves distributing the items being field-tested within an operational test form. Different forms contain the same operational items and different field-test items. For the 2015–16 administration, the original field-test items remained in their original positions in

the intact forms. Data from the 2015–16 administration of the CST for Science were not analyzed for these field-test items. The numbers of embedded field test items for the CSTs for Science are not presented in this report because for the 2015–16 administration, because field-test items were repeated as a part of the intact forms and there was no new item development.

Allocation of Students to Forms

The test forms for a given CST for Science were spiraled among students in the state so that a large representative sample of test-takers responded to the field-test items embedded in these forms. The spiraling design ensured that a diverse sample of students took each field-test item. The students did not know which items were field-test items and which items were operational items; therefore, their motivation was not expected to vary over the two types of items (Patrick & Way, 2008).

CDE Data Review

Once items were field-tested, ETS prepared the items that failed to meet the desired statistical criteria and the associated statistics for review by the CDE. ETS provided items with their statistical data, along with annotated comment sheets, for the CDE's use. ETS conducted an introductory training to highlight any new issues and serve as a statistical refresher. CDE consultants then made decisions about which items should be included for operational use in the item bank. ETS psychometric and content staff were available to CDE consultants throughout this process.

Item Banking

Once the ARP new item review was complete, the items were placed in the item bank along with their corresponding review information. Items that were accepted by the ARP, SPAR, and CDE were updated to a "field-test ready" status; items that were rejected were updated to a "rejected before use" status. ETS then delivered the items to the CDE by means of a delivery of the California electronic item bank. Subsequent updates to items were based on field-test and operational use of the items. However, only the latest content of the item is in the bank at any given time, along with the administration data from every administration that included the item.

After field-test or operational use, items that did not meet statistical specifications might be rejected; such items were updated with a status of "rejected for statistical reasons" and remain unavailable in the bank. These statistics were obtained by the psychometrics group at ETS, which carefully evaluated each item for its level of difficulty and discrimination as well as conformance to the item response theory Rasch model. Psychometricians also determined if the item functioned similarly for various subgroups of interest.

Items that were released were marked "released." They are not available for further use and remain unavailable in the bank. All unavailable items were marked with an availability indicator of "Unavailable," a reason for rejection as described above, and cause alerts so they are not inadvertently included on subsequent test forms. Statuses and availability were updated programmatically as items were presented for review, accepted or rejected, placed on a form for field-testing, presented for statistical review, used operationally, and released. All rejection and release indications were monitored and controlled through ETS's assessment development processes.

ETS currently provides and maintains the electronic item banks for several of the California assessments, including the California High School Exit Examination (CAHSEE), the

California English Language Development Test (CELDT), and CAASPP (CSTs for Science, California Modified Assessment for Science, California Alternate Performance Assessment for Science, and Standards-based Tests in Spanish). CAHSEE and CAASPP are currently consolidated in the California item banking system. ETS works with the CDE to obtain the data for assessments, such as the CELDT, under contract with other vendors for inclusion into the item bank. ETS provides the item banking application using the local area network architecture and the relational database management system, SQL 2008, already deployed. ETS provides updated versions of the item bank to the CDE on an ongoing basis and works with the CDE to determine the optimum process if a change in databases is desired.

References

California Department of Education. (2009). *California content standards*. Sacramento, CA. Retrieved from http://www.cde.ca.gov/be/st/ss/

Educational Testing Service. (2002). *ETS standards for quality and fairness*. Princeton, NJ: Author.

Patrick, R., & Way, D. (March, 2008). *Field testing and equating designs for state educational assessments*. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Schmeiser, C.B., & Welch, C.J. (2006). Test development. In R.L. Brennan (Ed.), *Educational measurement* (4th ed.). Westport, CT: American Council on Education and Praeger Publishers.

Chapter 4: Test Assembly

The California Standards Tests (CSTs) for Science were developed to measure students' performance relative to California's content standards approved by the State Board of Education (SBE) in 1998. They were also constructed to meet professional standards for validity and reliability. For each CST for Science, the content standards and desired psychometric attributes were used as the basis for assembling the test forms.

Test Length

The number of items in each CST for Science blueprint, adopted by the SBE in 2002, was determined by considering the construct that the test is intended to measure and the level of psychometric quality desired. Test length is closely related to the complexity of content to be measured by each test; this content is defined by the California content standards for each grade level and content area. Also considered is the goal that the test be short enough so that most of the students complete it in a reasonable amount of time.

The CST for Science grade-level science tests comprise 60 operational items and a total of 66 items on each test.

In addition to operational items, six items on each test are field-test items. For more details on the distribution of items, see Appendix 2.A—CST for Science Items and Estimated Time Chart on page 19.

Rules for Item Selection

Test Blueprint

All test items on CST for Science forms were selected to conform to the SBE-approved California content standards and test blueprints. The content blueprints for the CSTs for Science can be found on the California Department of Education (CDE) STAR CST Blueprints Web page at http://www.cde.ca.gov/ta/tg/sr/blueprints.asp.

Although the test blueprints call for the number of items at the individual standard level, scores for the CST for Science items are grouped into subcontent areas (reporting clusters). Although only the total test scale score are reported and cluster scores are no longer included in the score report, a list of the CST for Science reporting clusters by test and the number of items in the cluster that appear in each test are provided in Appendix 2.B—Reporting Clusters, which starts on page 20.

Content Rules and Item Selection

The intact forms from the 2011–12 administration were used during the 2015–16 administration of the CSTs for Science. Prior to the 2012–13 administration, test developers followed a number of rules when developing a new test form for a given grade and content area. First and foremost, they selected items that met the blueprint for that grade level and content area. Using an electronic item bank, assessment specialists began by identifying a number of linking items. These are items that appeared in a previous year's operational administration and were used to equate the administered test forms. Linking items were selected to proportionally represent the full blueprint. For example, if 25 percent of all of the items in a test were in the first reporting cluster, then 25 percent of the linking items should come from that cluster. The selected linking items were also reviewed by psychometricians to ensure that specific psychometric criteria were met.

After the linking items were approved, assessment specialists populated the rest of the test form. Their first consideration was the strength of the content and the match of each item to a specified content standard. In selecting items, team members also tried to ensure that they included a variety of formats and content and that at least some of the items included graphics for visual interest.

Another consideration was the difficulty of each item. Test developers strived to ensure that there were some easy and some hard items and that there were a number of items in the middle range of difficulty. If items did not meet all content and psychometric criteria, staff reviewed the other available items to determine if there were other selections that could improve the match of the test to all of the requirements. If such a match was not attainable, the content team worked in conjunction with psychometricians and the CDE to determine which combination of items would best serve the needs of the students taking the test. Chapter 3, starting on page 29, contains further information about this process.

Psychometric Criteria

The three goals of CST for Science test development were as follows:

1. The test must have desired precision of measurement at all ability levels.
2. The test score must be valid and reliable for the intended population and for the various subgroups of test-takers.
3. The test forms must be comparable across years of administration to ensure the generalizability of scores over time.

In order to achieve these goals, a set of rules was developed that outlines the desired psychometric properties of each CST for Science. These rules are referred to as statistical targets.

Two types of assembly targets were developed for each CST for Science: the total test target and (reporting) cluster targets. These targets were provided to test developers before a test construction cycle began. The test developers and psychometricians worked together to design the tests to these targets.

Primary Statistical Targets

The total test targets, or primary statistical targets, used for assembling the CST for Science forms for the intact forms used in the 2015–16 administration were the test information function (TIF) and an average point-biserial correlation.

The TIF is the sum of the item information function based on the item response theory (IRT) item parameters. When using an IRT model, the target TIF makes it possible to choose items to produce a test that has the desired precision of measurement at all ability levels.

The graphs for each total test are presented in Figure 4.A.1 on page 44. These curves present the target TIF and the projected TIF for the total test at each grade level.

Due to the unique characteristics of the Rasch IRT model, the information curve conditional on each ability level is determined by item difficulty (*b*-values) alone. In this case, the TIF would, therefore, suffice as the target for conditional test difficulty. Although additional item difficulty targets are not imperative when the target TIF is used for form construction, the target mean and standard deviation (SD) of item difficulty (*b*-values) consistent with the TIF were still provided to test development staff to help with the test construction process. The target *b*-value range approximates a minimum proportion-correct value (*p*-value) of 0.20 and a maximum *p*-value of 0.95 for each test.

The point-biserial correlation describes the relationship between student performance on a dichotomously scored item and student performance on the test as a whole. It is used as a measure of how well an item discriminates among test-takers who differ in their ability, and it is related to the overall reliability of the test.

The minimum target value for an item point biserial was set at 0.14 for each test. This value approximates a biserial correlation of 0.20.

Assembly Targets

The target values for the CSTs for Science are presented in Table 4.1. These specifications were developed from the analyses of test forms in their original year of administration.

Table 4.1 Statistical Targets for CST for Science Test Assembly

CST	Target Mean *b*	Target SD *b*	Min *p*-value	Max *p*-value	Mean Point Biserial	Min Point Biserial
Grade 5 Science	–0.67	0.57	0.20	0.95	> 0.34	0.14
Grade 8 Science	–0.40	0.74	0.20	0.95	> 0.34	0.14
Grade 10 Life Science	–0.29	0.72	0.20	0.95	> 0.34	0.14

Target information functions are also used to evaluate the items selected to measure each subscore in the interest of maintaining some consistency in the accuracy of cluster scores across years. Because the clusters include fewer items than the total test, there is always more variability between the target and the information curves constructed for the new form clusters than there is for the total test.

Figure 4.B.1 through Figure 4.B.3 starting on page 45 present the target and projected information curves for the reporting clusters in the administered tests.

Projected Psychometric Properties of the Assembled Tests

In the years when new forms are developed, Educational Testing Service psychometricians performed a preliminary review of the technical characteristics of the assembled tests. The expected or projected performance of students and the overall score reliability are estimated using the item-level statistics available in the California item bank for the selected items. The test reliability is based on Gulliksen's formula (Gulliksen, 1987) for estimating test reliability (r_{xx}) from item *p*-values and item point-biserial correlations:

$$r_{xx} = \left(\frac{K}{K-1}\right)\left(1 - \frac{\sum_{g=1}^{K} s_g^{\,2}}{\left(\sum_{g=1}^{K} r_{xg} s_g\right)^2}\right), \qquad (4.1)$$

where,

K is the number of items in the test,

$s_g^{\,2}$ is the estimated item variances, i.e., $p_g(1-p_g)$, where p_g is the item *p*-value for item *g,*

r_{xg} is the item point-biserial correlation for item *g,* and

$r_{xg}s_g$ is the item reliability index.

In addition, estimated test raw score means are calculated by summing the item *p*-values, and estimated test raw score standard deviations (SDs) are calculated by summing the item

reliability indices. Table 4.A.1 on page 44 presents these summary values by grade based on the item-level statistics from the year the form was previously administered.

It should be noted that the projected reliabilities in Table 4.A.1 were based on item *p*-values and point-biserial correlations that, for some of the items, were based on external field-testing using samples of students that were not fully representative of the state. Chapter 8 presents item *p*-values, point-biserial correlations, and test reliability estimates based on the data from the 2015–16 CST for Science administration.

Table 4.A.2 on page 44 shows the mean observed statistics of the items for each CST for Science based on the item-level statistics from the year the form was previously administered except for the replacement items—for these, the item bank values from the most recent administration were used. See Table 8.4 on page 90 for the dates of the original administrations. These values can be compared to the target values in Table 4.1.

Rules for Item Sequence and Layout

The items on the science test forms were sequenced according to reporting cluster; that is, all items from a single reporting cluster were presented together and then all of the items from the next reporting cluster were presented. Items from the Investigation and Experimentation reporting cluster were the exception to this rule: these items assess aspects of practical knowledge in various clusters; they were presented with their associated clusters and then aggregated for reporting purposes as an Investigation and Experimentation cluster.

Reference

Gulliksen, H. (1987). *Theory of mental tests*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Appendix 4.A—Technical Characteristics

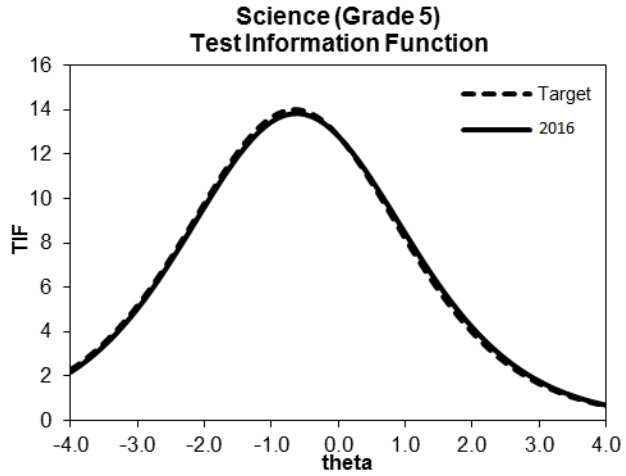
Table 4.A.1 Summary of 2015–16 CSTs for Science Projected Raw Score Statistics

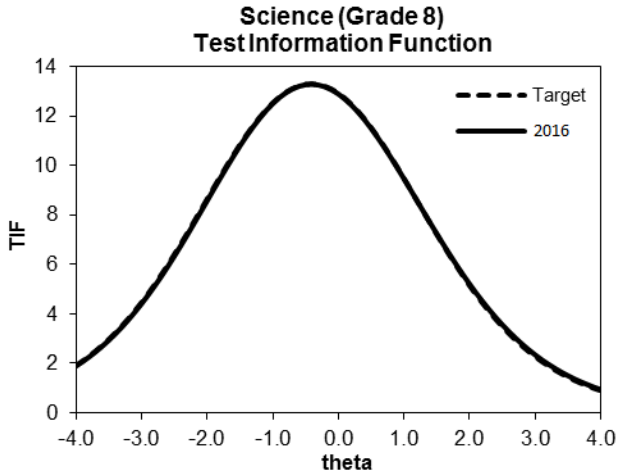
CST	Number of Op. Items	Mean Raw Score	Std. Dev. of Raw Scores	Reliability
Grade 5 Science	60	41.41	11.20	0.92
Grade 8 Science	60	41.60	11.18	0.92
Grade 10 Life Science	60	37.68	12.49	0.93

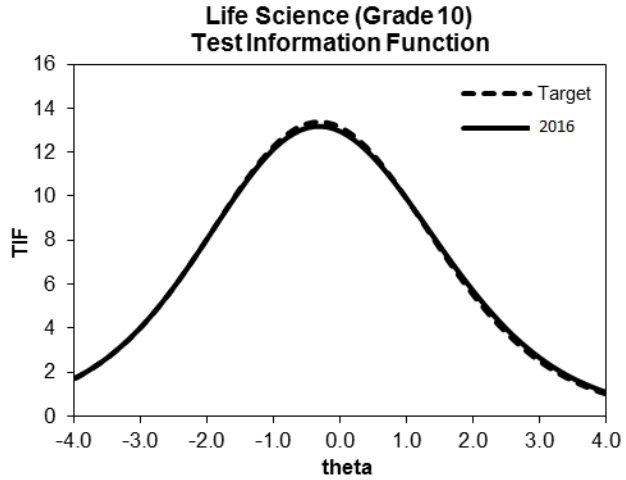
Table 4.A.2 Summary of 2015–16 CSTs for Science Projected Item Statistics

CST	Mean b	SD b	Mean *p*–value	Min *p*–value	Max *p*–value	Mean Point Biserial	Min Point Biserial
Grade 5 Science	-0.62	0.60	0.71	0.46	0.88	0.42	0.29
Grade 8 Science	-0.39	0.74	0.72	0.36	0.89	0.43	0.26
Grade 10 Life Science	-0.26	0.77	0.66	0.31	0.86	0.45	0.27

Figure 4.A.1 Plots of Target Information Function and Projected Information for Total Test for Science

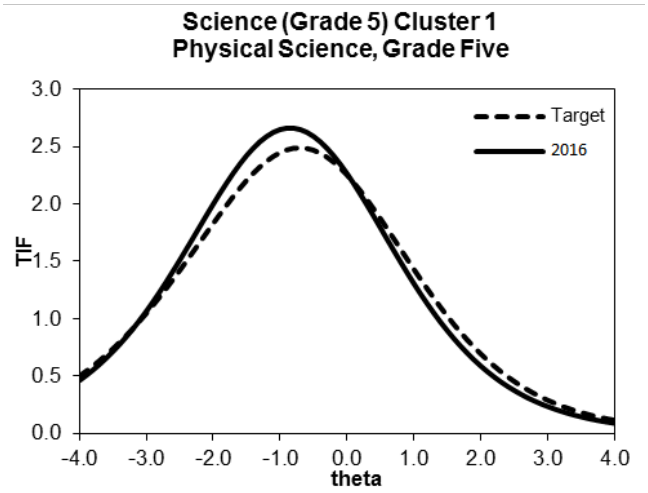


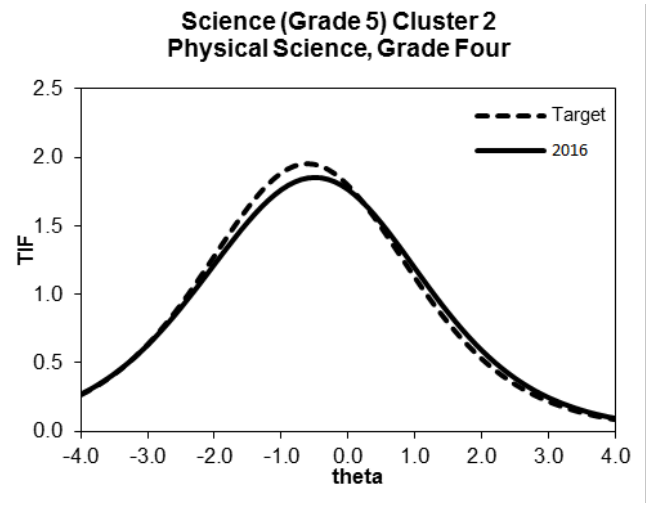


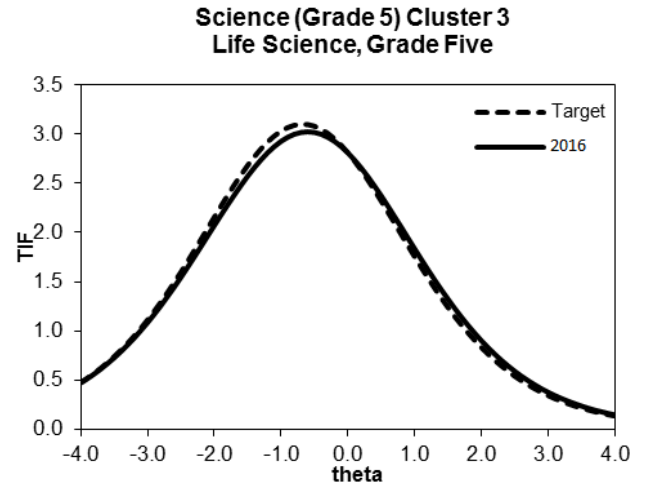


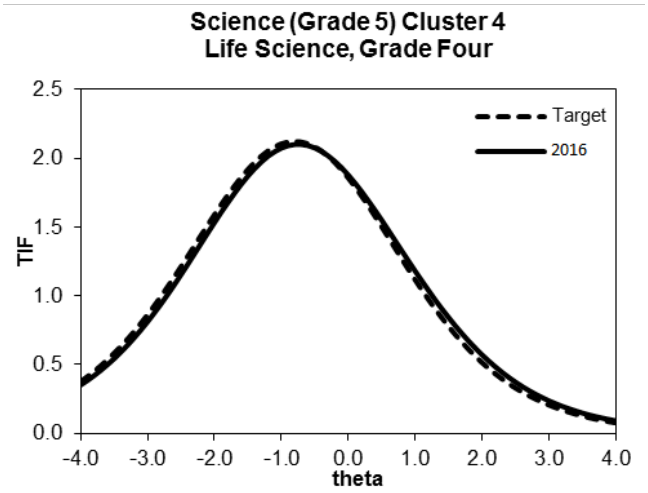
Appendix 4.B—Cluster Targets

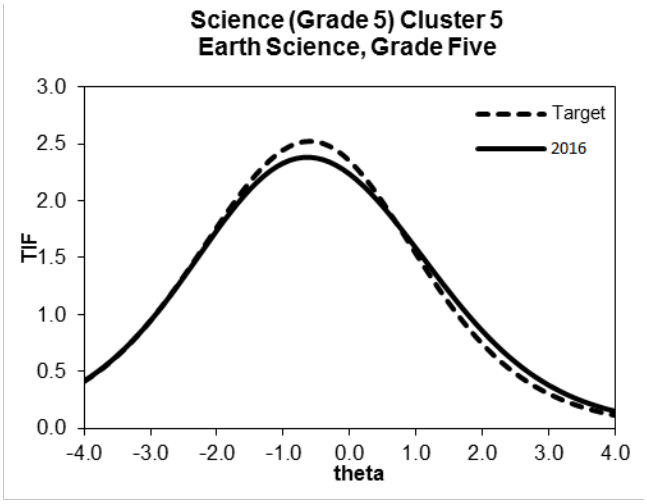
Figure 4.B.1 Plots of Target Information Functions and Projected Information for Clusters for Science, Grade Five











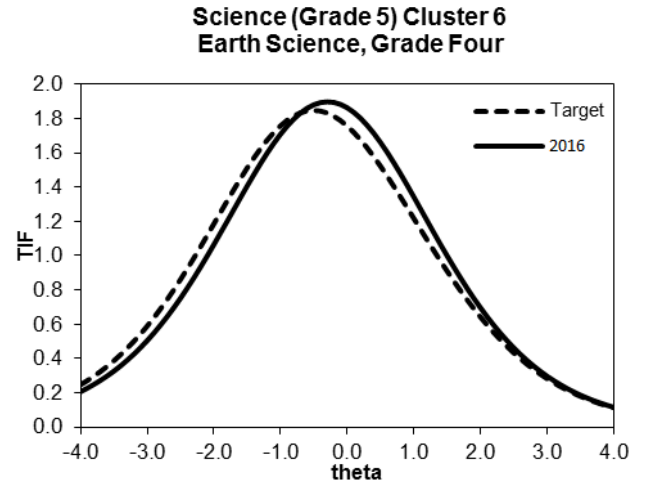


Figure 4.B.2 Plots of Target Information Functions and Projected Information for Clusters for Science, Grade Eight

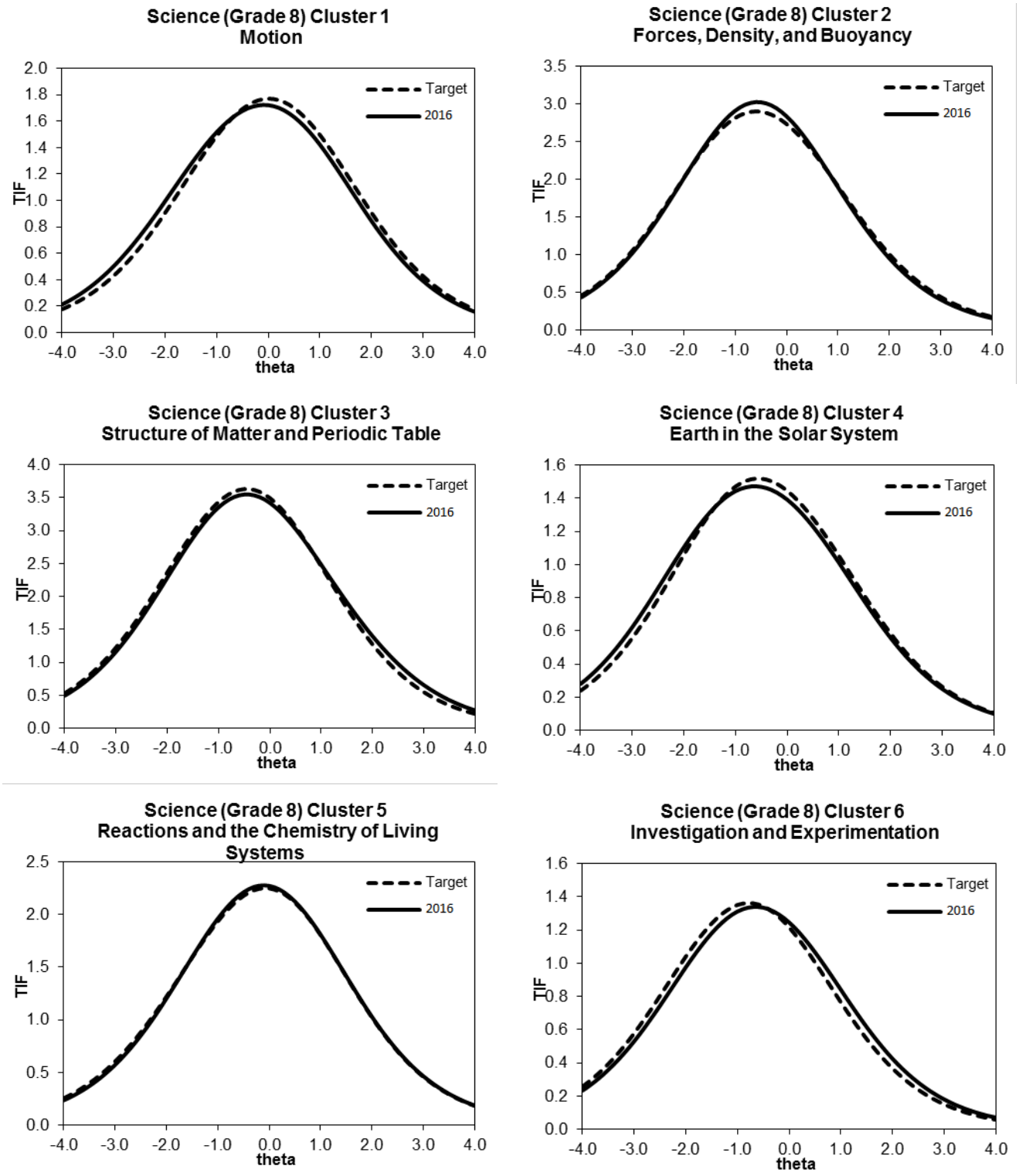
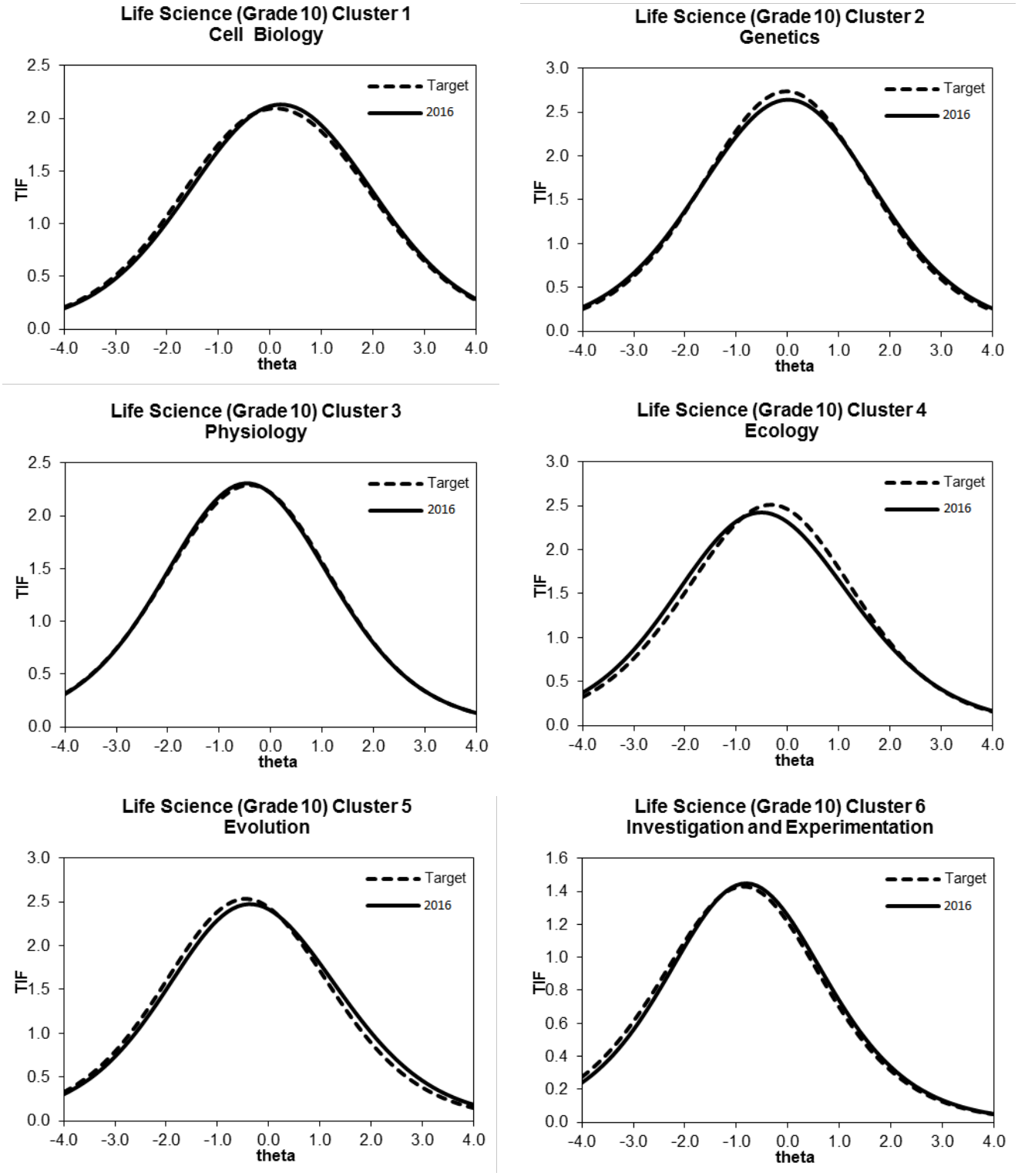


Figure 4.B.3 Plots of Target Information Functions and Projected Information for Clusters for Life Science, Grade Ten



Chapter 5: Test Administration

Test Security and Confidentiality

All tests within the California Assessment of Student Performance and Progress (CAASPP) System are secure documents. For the California Standards Tests (CSTs) for Science administration, every person having access to testing materials maintains the security and confidentiality of the tests. Educational Testing Service's (ETS's) Code of Ethics requires that all test information, including tangible materials (such as test booklets), confidential files, processes, and activities are kept secure. ETS has systems in place that maintain tight security for test questions and test results, as well as for student data. To ensure security for all the tests that ETS develops or handles, ETS maintains an Office of Testing Integrity (OTI), which is described in the next section.

ETS's Office of Testing Integrity

The OTI is a division of ETS that provides quality assurance services for all testing programs administered by ETS and resides in the ETS legal department. The Office of Professional Standards Compliance at ETS publishes and maintains *ETS Standards for Quality and Fairness,* which supports the OTI's goals and activities. The purposes of the *ETS Standards for Quality and Fairness* are to help ETS design, develop, and deliver technically sound, fair, and useful products and services, and to 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
- Report on security activities

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

Test Development

There was no new item development for the 2015–16 forms. For newly developed forms, during the test development process, ETS staff members consistently adhere to the following established security procedures:

- Only authorized individuals have access to test content at any step during the test development, item review, and data analysis processes.
- Test developers keep all hard-copy test content, computer disk copies, art, film, proofs, and plates in locked storage when not in use.
- ETS shreds working copies of secure content as soon as they are no longer needed during the test development process.
- Test developers take further security measures when test materials are to be shared outside of ETS; this is achieved by using registered and/or secure mail, using express delivery methods, and actively tracking records of dispatch and receipt of the materials.

Item and Data Review

As mentioned in Chapter 3, Assessment Review Panel (ARP) meetings were not held for the 2015–16 administration because there was no new item development for the 2015–16 CST for Science forms. However, for administrations when new forms were developed, ETS facilitated ARP meetings every year to review all newly developed CST for Science items and associated statistics. ETS enforced security measures at ARP meetings to protect the integrity of meeting materials using the following guidelines:

- Individuals who participated in the ARPs signed a confidentiality agreement.
- Meeting materials were strictly managed before, during, and after the review meetings.
- Meeting participants were supervised at all times during the meetings.
- Use of electronic devices was prohibited in the meeting rooms.

Item Banking

Once the ARP review was complete, the items were placed in the item bank. ETS then delivered the items to the California Department of Education (CDE) through the California electronic item bank. Subsequent updates to content and statistics associated with items were based on data collected from field testing and the operational use of the items. The latest version of the item is retained in the bank along with the data from every administration that had included the item.

Security of the electronic item banking system is of critical importance. The measures that ETS takes for assuring the security of electronic files include the following:

- Electronic forms of test content, documentation, and item banks are backed up electronically, with the backups kept off site, to prevent loss from a system breakdown or a natural disaster.
- The offsite backup files are kept in secure storage with access limited to authorized personnel only.
- To prevent unauthorized electronic access to the item bank, state-of-the-art network security measures are used.

ETS routinely maintains many secure electronic systems for both internal and external access. The current electronic item banking application includes a login/password system to provide authorized access to the database or designated portions of the database. In addition, only users authorized to access the specific system query language database are able to use the electronic item banking system. Designated administrators at the CDE and at ETS authorize users to access these electronic systems.

Transfer of Forms and Items to the CDE

ETS shares a secure file transfer protocol (SFTP) site with the CDE. SFTP is a method for reliable and exclusive routing of files. Files reside on a password-protected server that only authorized users may access. On that site, ETS posts Microsoft Word and Excel, Adobe Acrobat PDF, or other document files for the CDE to review. ETS sends a notification e-mail to the CDE to announce that files are posted. Item data are always transmitted in an encrypted format to the SFTP site; 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 not stored permanently on the shared SFTP server.

Security of Electronic Files Using a Firewall

A firewall is software that prevents unauthorized entry to files, e-mail, and other organization-specific programs. ETS data exchange 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 included in the Test Operations Management System (TOMS) (CDE, 2016a) 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 an assurance of confidentiality in the users of this information.

Printing and Publishing

After items and test forms are approved, the files are sent for printing on a CD using a secure courier system. According to the established procedures, the OTI preapproves all printing vendors before they can work on secured confidential and proprietary testing materials. The printing vendor must submit a completed ETS Printing Plan and a Typesetting Facility Security Plan; both plans document security procedures, access to testing materials, a log of work in progress, personnel procedures, and access to the facilities by the employees and visitors. After reviewing the completed plans, representatives of the OTI visit the printing vendor to conduct an onsite inspection. The printing vendor ships printed test booklets to ETS, which distributes the booklets to local educational agencies (LEAs) in securely packaged boxes.

Test Administration

ETS receives testing materials from printers, packages them, and sends them to LEAs. After testing, the LEAs return materials to ETS for scoring. During these events, ETS takes extraordinary measures to protect the testing materials. ETS uses customized business applications to verify that inventory controls are in place, from materials receipt to packaging. The reputable carriers used by ETS provide a specialized handling and delivery service that maintains test security and meets the CAASPP System schedule. The carriers provide inside delivery directly to the LEA CAASPP coordinators or authorized recipients of the assessment materials.

Test Delivery

Test security requires accounting for all secure materials before, during, and after each test administration. The LEA CAASPP coordinators are, therefore, required to keep all testing materials in central locked storage except during actual test administration times. CAASPP test site coordinators are responsible for accounting for and returning all secure materials to the LEA CAASPP coordinator, who is responsible for returning them to the Scoring and Processing Center. The following measures are in place to ensure security of CAASPP testing materials:

- LEA CAASPP coordinators are required to sign and submit a “CAASPP Test Security Agreement for LEA CAASPP Coordinators and CAASPP Test Site Coordinators (For all CAASPP assessments, including field tests)” form to the California Technical Assistance Center before ETS can ship any testing materials to the LEA.
- CAASPP test site coordinators have to sign and submit a “CAASPP Test Security Agreement for LEA CAASPP Coordinators and CAASPP Test Site Coordinators (For all CAASPP assessments, including field tests)” form to the LEA CAASPP coordinator before any testing materials can be delivered to the school/test site.

- Anyone having access to the testing materials must sign and submit a "CAASPP Test Security Affidavit for Test Examiners, Proctors, Scribes, and Any Other Persons Having Access to CAASPP Tests (For all CAASPP assessments, including field tests)" form to the CAASPP test site coordinator before receiving access to any testing materials.
- It is the responsibility of each person participating in the CAASPP System to report immediately any violation or suspected violation of test security or confidentiality. The CAASPP test site coordinator is responsible for immediately reporting any security violation to the LEA CAASPP coordinator. The LEA CAASPP coordinator must contact the CDE immediately; the coordinator will be asked to follow up with a written explanation of the violation or suspected violation.

Processing and Scoring

An environment that promotes the security of the test prompts, student responses, data, and employees throughout a project is of utmost concern to ETS. ETS requires the following standard safeguards for security at its sites:

- There is controlled access to the facility.
- No test materials may leave the facility during the project without the permission of a person or persons designated by the CDE.
- All scoring personnel must sign a nondisclosure and confidentiality form in which they agree not to use or divulge any information concerning tests, scoring guides, or individual student responses.
- All staff must wear ETS identification badges at all times in ETS facilities.

No recording or photographic equipment is allowed in the scoring area without the consent of the CDE.

The completed and scored answer documents are stored in secure warehouses. After they are stored, they will not be handled again. School and LEA personnel are not allowed to look at a completed answer document unless required for transcription or to investigate irregular cases.

All answer documents, test booklets, and other secure testing materials are destroyed after October 31 each year.

Data Management

ETS provides overall security for assessment materials through its limited-access facilities and through its secure data processing capabilities. ETS enforces stringent procedures to prevent unauthorized attempts to access its facilities. Entrances are monitored by security personnel and a computerized badge-reading system is utilized. Upon entering a facility, all ETS employees are required to display identification badges that must be worn at all times while in the facility. Visitors must sign in and out. While they are at the facility, they are assigned a visitor badge and escorted by ETS personnel. Access to the Data Center is further controlled by the computerized badge-reading system that allows entrance only to those employees who possess the proper authorization.

Data, electronic files, test files, programs (source and object), and all associated tables and parameters are maintained in secure network libraries for all systems developed and maintained in a client-server environment. Only authorized software development employees are given access as needed for development, testing, and implementation in a strictly controlled configuration management environment.

For mainframe processes, ETS limits and controls access to all data files (test and production), source code, object code, databases, and tables, regulating who is authorized to alter, update, or even read the files. All attempts to access files on the mainframe by unauthorized users are logged and monitored. In addition, ETS controls versions of the software and data files. Unapproved changes are not implemented without prior review and approval.

Statistical Analysis

The Information Technology (IT) department at ETS loads data files into a database. The Data Quality Services group at ETS extracts the data from the database and performs quality control procedures before passing files to the ETS Statistical Analysis group. The Statistical Analysis group keeps the files on secure servers and adheres to the ETS Code of Ethics and the ETS Information Protection Policies to prevent any unauthorized access.

Reporting and Posting Results

After statistical analysis has been completed, the following deliverables are produced:

- Printed Student Score Reports are produced and shipped to the designated LEA for distribution
- PDFs of Student Score Reports available through TOMS
- A file of individual student results—available for download from TOMS—that shows students' scale scores and performance levels
- A file of aggregated student results available for download through TOMS
- Encrypted files of summary results (sent to the CDE by means of SFTP) (Any summary results that have fewer than 11 students are not reported.)
- Item-level statistics based on the results, which are entered into the item bank

Student Confidentiality

To meet Elementary and Secondary Education Act and state requirements, LEAs must collect demographic data about students. This includes information about students' ethnicity, parent education, disabilities, whether the student qualifies for the National School Lunch Program, and so forth (CDE, 2016b). ETS takes precautions to prevent any of this information from becoming public or being used for anything other than testing purposes. These procedures are applied to all documents in which these student demographic data may appear.

Student Test Results

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

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

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

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

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

- Scoring
- Transfer of scores by means of secure data exchange
- Reporting
- Analysis and reporting of erasure marks
- Posting of aggregate data
- Storage

In addition to protecting the confidentiality of testing materials, ETS's Code of Ethics further prohibits ETS employees from financial misuse, conflicts of interest, and unauthorized appropriation of ETS's property and resources. Specific rules are also given to ETS employees and their immediate families who may take a test developed by ETS, such as a CAASPP examination. The ETS Office of Testing Integrity verifies that these standards are followed throughout ETS. It does this, in part, by conducting periodic onsite security audits of departments, with follow-up reports containing recommendations for improvement.

Procedures to Maintain Standardization

The CST for Science processes are designed so that the tests are administered and scored in a standardized manner.

ETS employs personnel who facilitate various processes involved in the standardization of an administration cycle and takes all necessary measures to ensure the standardization of the CSTs for Science, as described in this section.

Test Administrators

The CSTs for Science are administered in conjunction with the other tests that comprise the CAASPP System. The responsibilities for LEA and test site staff members are included in the *CAASPP Paper-Pencil Testing Test Administration Manual* (CDE, 2016b). This manual is described in the next section.

The staff members centrally involved in the test administration are as follows:

LEA CAASPP Coordinator

Each LEA designates an LEA CAASPP coordinator who is responsible for ensuring the proper and consistent administration of the CAASPP tests. LEAs include public school districts, statewide benefit charter schools, state board–authorized charter schools, county

office of education programs, and charter schools testing independently from their home district.

LEA CAASPP coordinators are also responsible for securing testing materials upon receipt, distributing testing materials to schools, tracking the materials, training and answering questions from LEA staff and CAASPP test site coordinators, reporting any testing irregularities or security breaches to the CDE, receiving scorable and nonscorable materials from schools after an administration, and returning the materials to the CAASPP contractor for processing.

CAASPP Test Site Coordinator

The superintendent of the school district or the LEA CAASPP coordinator designates a CAASPP test site coordinator at each test site from among the employees of the LEA. (*California Code of Regulations,* Title 5 *[5 CCR],* Section 858 [a])

CAASPP test site coordinators are responsible for making sure that the school has the proper testing materials, distributing testing materials within a school, securing materials before, during, and after the administration period, answering questions from test administrators, preparing and packaging materials to be returned to the LEA after testing, and returning the materials to the LEA. (CDE, 2016b)

Test Administrator

CSTs for Science are administered by test administrators who may be assisted by test proctors and scribes. A test administrator is an employee of an LEA or an employee of a nonpublic, nonsectarian school (NPS) who has been trained to administer the tests and has signed a CAASPP Test Security Affidavit. Test administrators must follow the directions in the *California Standards Tests Directions for Administration* (*DFA*) (CDE, 2016c) exactly.

Test Proctor

A test proctor is an employee of an LEA or a person, assigned by an NPS to implement the individualized education program (IEP) of a student, who has received training designed to prepare the proctor to assist the test administrator in the administration of tests within the CAASPP System (5 *CCR* Section 850 [y]). Test proctors must sign CAASPP Test Security Affidavits (5 *CCR* Section 859 [c]).

Scribe

A scribe is an employee of an LEA or a person, assigned by an NPS to implement the IEP of a student, who is required to transcribe a student's responses to the format required by the test. A student's parent or guardian is not eligible to serve as the student's scribe (5 *CCR* Section 850 [s]). Scribes must sign CAASPP Test Security Affidavits (5 *CCR* Section 859 [c]).

Directions for Administration (DFAs)

CST for Science *DFA*s are manuals used by test administrators to administer the CSTs for Science to students (CDE, 2016c). Test administrators must follow all directions and guidelines and read, word-for-word, the instructions to students in "SAY" boxes to ensure test standardization.

CAASPP Paper-Pencil Testing Test Administration Manual

Test administration procedures are to be followed exactly so that all students have an equal opportunity to demonstrate their academic achievement. The *CAASPP Paper-Pencil Testing Test Administration Manual* contributes to this goal by providing information about the responsibilities of LEA CAASPP and CAASPP test site coordinators, as well as those of the

other staff involved in the administration cycle (CDE, 2016b). However, the manual is not intended as a substitute for 5 *CCR* or to detail all of the coordinator's responsibilities.

Test Operations Management System Manuals

TOMS is a series of secure, Web-based modules that allow LEA CAASPP coordinators to set up test administrations and ensure test sites order materials. Every module has its own user manual with detailed instructions on how to use TOMS. The TOMS modules used to manage paper-pencil test processes are as follows:

- **Test Administration Setup—**This module allows LEAs to determine and calculate dates for scheduling test administrations for LEAs, verify contact information for those LEAs, and request Pre-ID labels. (CDE, 2016d)
- **Student Paper-Pencil Test Registration—**This modules allows LEAs to assign paper-pencil science tests to students in grades five, eight, and ten. (CDE, 2016e)
- **Set Condition Codes—**This module allows LEA CAASPP coordinators and CAASPP test site coordinators to apply condition codes (to note that a student was absent during testing, for example) to student records.

Test Booklets

For each grade-level test, multiple versions of test booklets are administered. The versions differ only in terms of the field-test items they contain. These versions are spiraled—comingled—and packaged consecutively and are distributed at the student level; that is, each classroom or group of test-takers receives at least one of each version of the test.

The test booklets, along with answer documents and other supporting materials, are packaged by school or group. All materials are sent to the LEA CAASPP coordinator for proper distribution within the LEA. Special formats of test booklets are also available for test-takers who require accommodations to participate in testing. These special formats include large-print and braille testing materials.

Universal Tools, Designated Supports, and Accommodations for Students with Disabilities

All public school students participate in the CAASPP System, including students with disabilities and ELs. ETS policy states that reasonable testing accommodations be provided to candidates with documented disabilities that are identified in the Americans with Disabilities Act (ADA). The ADA mandates that test accommodations be individualized, meaning that no single type of test accommodation may be adequate or appropriate for all individuals with any given type of disability. The ADA authorizes that test-takers with disabilities may be tested under standard conditions if ETS determines that only minor adjustments to the testing environment are required (e.g., wheelchair access, large-print test book, a sign language interpreter for spoken directions).

Identification

Most students with disabilities and most ELs take the CSTs for Science under standard conditions. However, some students with disabilities and some ELs may need assistance when taking the CSTs for Science. This assistance takes the form of universal tools, designated supports, and accommodations (see Appendix 2.C on page 21 in Chapter 2 for details). During the test, these students may use the special services specified in their IEP or Section 504 plan. If students use universal tools, designated supports, and/or accommodations for the CSTs for Science, test administrators are responsible for marking

the universal tools, designated supports, and/or accommodations used on the students' answer documents.

Scoring

The purpose of universal tools, designated supports, and accommodations in testing is to allow *all* students the opportunity to demonstrate what they know and what they are able to do, rather than give students using them an advantage over other students or artificially inflate their scores. Universal tools, designated supports, and accommodations minimize or remove the barriers that could otherwise prevent students from generating results that reflect their achievement in the content area.

Scores for students tested with non-embedded accessibility supports are counted as far below basic for aggregate reporting; universal tools, designated supports, or accommodations do not result in changes to students' scores.

Testing Incidents

Testing incidents—breaches and irregularities—are circumstances that may compromise the reliability and validity of test results

The LEA CAASPP coordinator is responsible for immediately notifying the CDE of any irregularities or breaches that occur before, during, or after testing. The test administrator is responsible for immediately notifying the LEA CAASPP coordinator of any security breaches or testing irregularities that occur in the administration of the test. Once the LEA CAASPP coordinator has determined that an irregularity or breach has occurred, the LEA CAASPP coordinator or CAASPP test site coordinator reports the incident using the secure Security and Test Administration Incident Reporting System. The information and procedures to assist in identifying incidents and notifying the CDE are provided in the *CAASPP Paper-Pencil Testing Test Administration Manual* (CDE, 2016b).

Social Media Security Breaches

Social media security breaches are exposures of test questions and testing materials through social media Web sites. These security breaches raise serious concerns that require comprehensive investigation and additional statistical analyses. In recognizing the importance of and the need to provide valid and reliable results to the state, LEAs, and schools, both the CDE and ETS take every precaution necessary, including extensive statistical analyses, to ensure that all test results maintain the highest levels of psychometric integrity.

There were no high-risk social media security breaches associated with the CSTs for Science during the 2015–16 that required any item to be withheld from scoring.

Testing Improprieties

A testing impropriety is any event that occurs before, during, or after test administrations that does not conform to the instructions stated in the *DFAs* (CDE, 2016c) and the *CAASPP Paper-Pencil Testing Test Administration Manual* (CDE, 2016b). These events include test administration errors, disruptions, and student cheating. Testing improprieties generally do not affect test results and are not reported to the CDE or the CAASPP System testing contractor. The CAASPP test site coordinator should immediately notify the LEA CAASPP coordinator of any testing improprieties that occur. It is recommended by the CDE that LEAs and schools maintain records of testing improprieties.

References

California Department of Education. (2016a). *CAASPP Test Operations Management System* [Web site]. Sacramento, CA. http://www.caaspp.org/administration/toms/

California Department of Education. (2016b). *2015–16 CAASPP paper-pencil testing test administration manual*. Sacramento, CA. Retrieved from http://www.caaspp.org/rsc/pdfs/CAASPP.ppt_tam.2016.pdf

California Department of Education. (2016c). *2016 California Standards Tests directions for administration.* Sacramento, CA. Retrieved from http://www.caaspp.org/rsc/pdfs/CST.grade-8_dfa.2016.pdf

California Department of Education. (2016d). *California Assessment of Student Performance and Progress Test Operations Management System: 2015–16 Test administration setup guide.* Sacramento, CA. Retrieved from http://www.caaspp.org/rsc/pdfs/CAASPP.test_admin_setup.2016.pdf

California Department of Education. (2016e). *California Assessment of Student Performance and Progress Test Operations Management System: 2015–16 Student paper-pencil test registration user guide.* Sacramento, CA. Retrieved from http://www.caaspp.org/rsc/pdfs/CAASPP.student-test-registration-guide.2016.pdf

Chapter 6: Performance Standards

Background

The California Standards Tests (CSTs) for English–Language Arts (ELA) and Mathematics became part of California's standardized testing program in 1999; however, they are no longer administered starting with the 2013–14 administration. Five performance standards for the ELA tests were developed in 2000 and adopted by the State Board of Education (SBE) for the 2000–01 administration of those tests.

Also in 2001, the CSTs for history–social science and end-of-course science were introduced in grades nine through eleven; these tests also are no longer administered starting in 2014. The performance standards for those tests were established in the same year and were adopted in their first operational administration in 2001–02. The performance standards for mathematics tests were established in 2001 and adopted in the 2001–02 operational administration of those CSTs.

In 2003, performance standards were adopted for the CST for Science (Grade 5) and were reported operationally starting in 2004. In 2005, performance standards were adopted for the science CSTs for grades eight and ten and were reported operationally starting in 2006. The performance standards for the CSTs for Science were defined by the SBE as far below basic, below basic, basic, proficient, and advanced.

In 2015–16, the CSTs for Science in grades five and eight and Life Science in grade ten were administered to eligible students. Consequently, the performance standards for the grades and subjects were applied to the scores of students.

A review of the standard-setting literature supports the need for attention to best practices (Brandon, 2004; Hambleton & Pitoniak, 2006, Tannenbaum & Katz, 2013), which include the following

- careful selection of panel members;
- a sufficient number of panel members to represent varying perspectives;
- sufficient time devoted to develop a common understanding of the assessment domain;
- adequate training of panel members;
- development of a description of each performance level;
- multiple rounds of judgments; and
- the inclusion of data, where appropriate, to inform judgments.

California employed carefully designed standard-setting procedures to facilitate the development of performance standards for each CST for Science. These processes are described in the sections that follow.

Standard-Setting Procedure

The process of standard setting is designed to identify a "cut score" or minimum test score that is required to qualify a student for each performance level. The process generally requires that a panel of subject-matter experts and others with relevant perspectives (for example, teachers, school administrators) be assembled. The panelists for the CST for Science standard settings were selected based on the following characteristics:

- Familiarity with the subject matter assessed
- Familiarity with students in the respective grade levels
- An understanding of large-scale assessments
- An appreciation of the consequences of setting these cut scores

In recruiting panelists, the goal was to include a representative sample of California educators with experience in the education of students who take the CSTs and who are familiar with the California content standards for science adopted by the SBE in 1998. Invited panelists included teachers, administrators, and/or curriculum specialists. The final selection of panelists for the workshops was made by the CDE.

Also, in the interest of equity, representatives from diverse geographic regions, and from different gender and major racial/ethnic subgroups were requested to participate (Educational Testing Service [ETS], 2004, 2006).

The standard-setting processes implemented for CSTs for Science required panelists to follow these steps, which include training and practice prior to making judgments:

1. At the start of the workshop, panelists received training that included the purpose of standard setting and their role in the work, the meaning of a “cut score” and “impact data,” and specific training and practice in the method being used. Impact data included the percentage of students assessed in a previous administration of the test that would fall into each level, given the panelists’ judgments of cut scores.
2. Panelists looked at the content standards upon which the test items are based and discussed the expectations in the content area. This allowed the panelists to understand how their perception of item difficulty may relate to the complexity of content standards.
3. Panelists became familiar with the difficulty level of the items by taking the actual test and then assessing and discussing the demands of the test items.
4. Panelists discussed the meaning of the performance standard descriptions and visualized the knowledge and skills of students who would belong in each performance level.
5. Panelists identified characteristics of a “borderline” examinee. The borderline examinee is defined as a test-taker who possesses just enough knowledge of the content to move over the border separating a performance level from the performance level below it.
6. Panelists made individual judgments and discussed feedback related to other panelists’ judgments and feedback based on student performance data (impact data). Panelists could revise their judgments during the process if they wished.
7. The final recommended cut scores were based on the median of panelists’ judgment scores. For the CSTs for Science, the cut scores recommended by the panelists and the recommendation of the State Superintendent of Public Instruction were presented for public comment at regional public hearings. Comments and recommendations were then presented to the SBE for approval.

Standard-Setting Methodologies

Several methodologies exist to collect panelists’ judgments and to translate their results appropriately into cut scores. For the ELA CSTs, the modified Angoff method was used for standard setting (Hurtz & Auerbach, 2003), while the Bookmark method was used to set the

performance standards for the history–social science, mathematics, and science CSTs (Mitzel, Lewis, Patz, & Green, 2001). Both methods represent an appropriate balance between statistical rigor and informed opinion, as explained in the following sections.

Modified Angoff Method

A modified Angoff approach is widely used for recommending cut scores (Brandon, 2004; Hurtz & Auerbach, 2003; Norcini & Shea, 1997). This approach utilizes panelists' estimates of the percentage of borderline examinees that would answer each item correctly. The percentages are summed across the set of test items for each panelist and then the average is computed across panelists to arrive at the full panel's recommended cut score.

Bookmark Method

The Bookmark method for setting cut scores was introduced in 1999 and has been used widely across the United States (Lewis, et al., 1999; Mitzel, et al., 2001). In California, the Bookmark method was used in standard settings for most of the CAASPP paper-pencil tests.

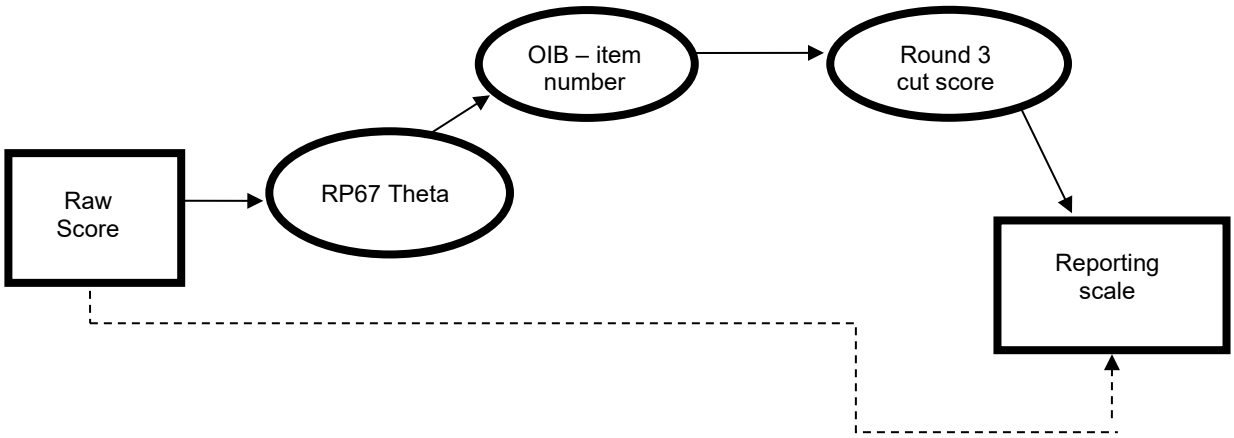
The Bookmark method is an item-mapping procedure in which panelists consider content covered by items in a specially constructed book where items are ordered from easiest to hardest, based on operational student performance data from a previous test administration. The "item map," which accompanies the ordered item booklet (OIB), includes information on the content measured by each operational test question, information about each question's difficulty, the correct answer for each question, and where each question was located in the test booklet before the questions were reordered by difficulty.

Panelists are asked to place a bookmark in the OIB to demarcate each performance level. The bookmarks are placed with the assumption that the borderline students will perform successfully at a given performance level with a probability of at least 0.67. Conversely, these students are expected to perform successfully on the items after the bookmark with a probability of less than 0.67 (Huynh, 1998).

In this method, the panelists' cut-score recommendations are presented in the metric of the OIB and are derived by obtaining the median of the corresponding bookmarks placed for each performance level across panelists.

Each item location corresponds to a value of theta, based on a response probability of 0.67 (RP67 Theta), which maps back to a raw score on this test form. Figure 6.1 below may best illustrate the relationship among the various metrics used when the Bookmark method is applied. The solid lines represent steps in the standard-setting process described above; the dotted line represents the scaling described in the next section.

Figure 6.1 Bookmark Standard-setting Process for the CSTs



Results

The cut scores obtained as a result of the standard-setting process were on the number-correct or raw-score scale; the scores were then translated to a score scale that ranges between 150 and 600.

The cut score for the basic performance level was set to 300 for every grade and content area; this means that a student must earn a score of 300 or higher to achieve a basic classification. The cut score for the proficient performance level was set to 350 for every grade and content area; this means that a student must earn a score of 350 or higher to achieve a proficient classification.

The cut scores for the other performance levels were derived using procedures based on item response theory (IRT) and usually vary by grade and subject area. Each raw cut score for a given test was mapped to an IRT *theta* (θ) using the test characteristic function or curve and then transformed to the scale-score metric using the following equation:

$$\text{Scale Cut Score}=(350-\theta_{proficient}\times\left(\frac{350-300}{\theta_{proficient}-\theta_{basic}}\right))+\left(\frac{350-300}{\theta_{proficient}-\theta_{basic}}\right)\times\theta_{cut\text{-}score} \quad (6.1)$$

where,

$\theta_{cut-score}$ represents the student ability at cut scores for performance levels other than proficient or basic, e.g., below basic or advanced,

$\theta_{proficient}$ represents the theta corresponding to the cut score for proficient, and

θ_{basic} represents the theta corresponding to the cut score for basic.

Please note that an IRT test characteristic function or curve is the sum of item characteristic curves (ICC), where an ICC represents the probability of correctly responding to an item conditioned on examinee ability.

The scale-score ranges for each performance level are presented in Table 2.1 on page 16. The cut score for each performance level is the lower bound of each scale-score range. The scale-score ranges do not change from year to year. Once established, they remain

unchanged from administration to administration until such time that new performance standards are adopted.

Table 7.2 on page 67 in Chapter 7 presents the percentages of students meeting each performance level for the 2015–16 administration.

References

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

Educational Testing Service. (2004). *STAR 5th grade science California Standards Test standard setting technical report. March 5, 2004* (California Department of Education Contract Number 5417). Princeton, NJ: Author.

Educational Testing Service. (2006). *California STAR grades 8 and 10 science California Standards Tests (CSTs) standard setting results March 9, 2006* (California Department of Education Contract Number 5417). Princeton, NJ: Author.

Hurtz, G.M., & Auerbach, M.A. (2003). A meta-analysis of the effects of modifications to the Angoff method on cutoff scores and judgment consensus. *Educational and Psychological Measurement*, *63*, 584–601.

Huynh, H. (1998). On score locations of binary and partial credit items and their applications to item mapping and criterion-referenced interpretation. *Journal of Educational and Behavioral Statistics, 23*(19), 35–56.

Lewis, D. M., Green, D. R., Mitzel, H. C., Baum, K., & Patz, R. J. (1999). *The bookmark standard setting procedure: Methodology and recent implications*. Manuscript under review.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249–81). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Norcini, J. J., & Shea, J. A. (1997). The credibility and comparability of standards. *Applied Measurement in Education, 10,* 39–59.

Tannenbaum, R. J. & Katz, I. R. (2013). Standard setting. In K. F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology:* (Vol. 3, pp. 455–477). Washington, DC: American Psychological Association.

Chapter 7: Scoring and Reporting

Educational Testing Service (ETS) conforms to high standards of quality and fairness (ETS, 2002) when scoring tests and reporting scores. These standards dictate that ETS provides accurate and understandable assessment results to the intended recipients. It is also ETS's mission to provide appropriate guidelines for score interpretation and cautions about the limitations in the meaning and use of the test scores. Finally, ETS conducts analyses needed to ensure that the assessments are equitable for various groups of test-takers.

Procedures for Maintaining and Retrieving Individual Scores

Items for all California Standards Tests (CSTs) for Science are multiple choice. Students are presented with a question and asked to select the correct answer from among four possible choices; students mark their answer choices in an answer document. All multiple-choice questions are machine scored.

In the 2015–16 administration, because the raw-score-to-scale-score conversion tables were developed before tests were administered using pre-equating, preliminary individual student results were available for download prior to the printing of paper reports. This electronic reporting was made possible through the Online Reporting System (ORS).

In order to score and report CST for Science results, ETS follows an established set of written procedures. The specifications for these procedures are presented in the next sections.

Scoring and Reporting Specifications

ETS develops standardized scoring procedures and specifications so that test materials are processed and scored accurately. These documents include the following:

- Scoring Rules—Describes the following:
  - the rules for how and when scores are reported, including whether or not the student data will be part of the CST for Science reporting and how performance levels are reported for students who used an individualized aid, and how scores are reported under certain conditions (for example, when a student was not tested)
  - CST for Science reporting cluster names
  - General reporting descriptions such as how to calculate number tested
- Include Indicators—Defines the appropriate codes to use when a student does not take or complete a test or when a score will not be reported

The scoring specifications are reviewed and revised by the California Department of Education (CDE) and ETS each year. After a version agreeable to all parties is finalized, the CDE issues a formal approval of the scoring and reporting specifications.

Scanning and Scoring

Answer documents are scanned and scored by ETS in accordance with the scoring specifications that have been approved by the CDE. Answer documents are designed to produce a single complete record for each student. This record includes demographic data and scanned responses for each student; once computed, the scored responses and the total test scores for a student are also merged into the same record. All scores, including those available via electronic reporting, must comply with the ETS scoring specifications. ETS has quality control checks in place to ensure the quality and accuracy of scanning and the transfer of scores into the database of student records.

Each local educational agency (LEA) must return scorable and nonscorable materials within five working days after the selected last day of testing for each test administration period.

Types of Scores and Subscores

Raw Score

For all of the tests, the total test raw score equals the number of multiple-choice test items correctly answered.

Subscore

The items in each CST for Science are aggregated into groups of related content standards to form reporting clusters. A subscore is a measure of a student's performance on the items in each reporting cluster. These results are provided only in this technical report. A description of the CST for Science reporting clusters is provided in Appendix 2.B of Chapter 2, starting on page 20.

Scale Score

Raw scores obtained on each CST for Science are transformed to three-digit scale scores using the equating process described in Chapter 2 on page 8. Scale scores range from 150 to 600 on each CST for Science. The scale scores of students who have been tested in different years at a given grade level and content area can be compared. However, the raw scores of these students cannot be meaningfully compared, because these scores are affected by the relative difficulty of the test taken as well as the ability of the student.

Performance Levels

The performance of each student on each CST for Science is categorized into one of the following performance levels:

- far below basic
- below basic
- basic
- proficient
- advanced

For all CSTs for Science, the cut score for the basic performance level is 300 for every test; this means that a student must earn a score of 300 or higher to achieve a basic classification. The cut score for the proficient performance level is 350; this means that a student must earn a score of 350 or higher to achieve a proficient classification. The cut scores for the other performance levels usually vary by grade.

Score Verification Procedures

Various necessary measures are taken to ascertain that the scoring keys are applied to the student responses as intended and that the student scores are computed accurately. In 2015–16, every regular and special-version multiple-choice test is certified by ETS prior to being included in electronic reporting. To certify a test, psychometricians gather a certain number of test cases and verify the accurate application of scoring keys and scoring tables.

Scoring Key Verification Process

Scoring keys, provided in the form planners, are produced by ETS and verified by performing multiple quality-control checks. The form planners contain the information about an assembled test form, including scoring keys, test name, administration year, subscore

identification, and the standards and statistics associated with each item. The quality control checks that are performed before keys are finalized are listed below:

1. Keys in the form planners are checked against their matching test booklets to ensure that the correct keys are listed.
2. The form planners are checked for accuracy against the Form Planner Specification document and the Score Key and Score Conversion document before the keys are loaded into the score key management (SKM) system at ETS.
3. The printed lists of the scoring keys are checked again once the keys have been loaded into the SKM system.
4. The demarcations of various sections in the actual test booklets are checked against the list of demarcations provided by ETS test development staff.
5. Scoring is verified internally at ETS, which generates scores and verifies the scoring of the data by comparing the two results. Any discrepancies are then resolved.
6. The entire scoring system is tested using a test deck that includes typical and extremely atypical response vectors.
7. Classical item analyses are computed on an early sample of data to provide an additional check of the keys. Although rare, if an item is found to be problematic, a follow-up process is carried out for it to be excluded from further analyses.

Overview of Score Aggregation Procedures

In order to provide meaningful results to the stakeholders, CST for Science scores for a given grade are aggregated at the school, independently testing charter school, district, county, and state levels. The aggregated scores are generated both for individual scores and group scores. The next section contains a description of the types of aggregation performed on CST for Science scores.

Individual Scores

The tables in this section provide state-level summary statistics describing student performance on each CST for Science.

Score Distributions and Summary Statistics

Summary statistics that describe student performance on each CST for Science are presented in Table 7.1.

Included in the table are the number of items in each test, the number of students taking each test, and the means and standard deviations of student scores expressed in terms of both raw scores and scale scores. The last two columns in the table list the raw score means and standard deviations as percentages of the total raw score points in each test.

Table 7.1 Mean and Standard Deviation of Raw and Scale Scores for the CSTs for Science

			Scale Score		Raw Score		Raw Score Percent Correct	
CST	No. of Items	No. of Students	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Grade 5 Science	60	442,498	357	64	40.83	11.31	68.06	18.85
Grade 8 Science	60	433,015	381	95	41.45	11.02	69.09	18.37
Grade 10 Life Science	60	449,114	354	65	38.45	11.69	64.08	19.49

The percentages of students in each performance level are presented in Table 7.2. The last column of the table presents the overall percentage of students who were classified at the proficient level or higher.

The numbers in the summary tables may not match exactly the results reported on the CDE's Web site because of slight differences in the samples used to compute the statistics. The P2 data file was used for the analyses in this chapter. This file contained the entire test-taking population and all the student records used as of September 15, 2016. This file contained data collected from all LEAs but did not include corrections of demographic data through the California Pupil Achievement Data System. In addition, students with invalid scores were excluded from the tables.

Table 7.2 Percentages of Students in Performance Levels for CSTs for Science

CST	Far Below Basic	Below Basic	Basic	Proficient	Advanced	Proficient/ Advanced *
Grade 5 Science	8%	10%	28%	34%	20%	53%
Grade 8 Science	9%	11%	20%	23%	38%	61%
Grade 10 Life Science	9%	12%	30%	27%	23%	50%

* May not exactly match the sum of percent proficient and percent advanced due to rounding.

Table 7.A.1 in Appendix 7.A on page 72 shows the distributions of scale scores for each CST for Science.

The results are reported in terms of 15 score intervals, each of which contains 30 scale score points. A cell value of "N/A" indicates that there are no obtainable scale scores within that scale-score range for the particular CST for Science.

Group Scores

Statistics summarizing student performance by each grade-level test for selected groups of students are provided starting on page 73 in Table 7.B.1 through Table 7.B.3 for the CSTs for Science.

In these tables, students are grouped by demographic characteristics, including gender, ethnicity, English-language fluency, need for special education services, and economic status. The tables show, for each demographic group, the numbers of valid cases, scale score means and standard deviations, the percentages of students in each performance level, as well as the mean percent correct in each reporting cluster.

Table 7.3 provides definitions of the demographic groups included in the tables. To protect privacy when the number of students in a subgroup is 10 or fewer, the summary statistics at the test- and reporting-cluster–level are not reported and are presented as hyphens. Percentages in these tables may not sum up to 100 due to rounding.

Table 7.3 Subgroup Definitions

Subgroup	Definition
Gender	• Male • Female
Ethnicity	• American Indian or Alaska Native • Asian – Asian Indian – Cambodian – Chinese

Subgroup	Definition
	– Hmong – Japanese – Korean – Laotian – Vietnamese – Other Asian • Pacific Islander – Guamanian – Native Hawaiian – Samoan – Tahitian – Other Pacific Islander • Filipino • Hispanic or Latino • African American • White (not Hispanic)
English-language Fluency	• English only • Initially fluent English proficient • English learner (EL) • Reclassified fluent English proficient • To be determined (TBD)
Economic Status	• Not economically disadvantaged • Economically disadvantaged
Special Services	• No special services • Special services

Reports Produced and Scores for Each Report

The tests that make up the California Assessment of Student Performance and Progress (CAASPP) System provide results or score summaries that are reported for different purposes. The three major purposes are:

1. Communicating with parents and guardians;
2. Informing decisions needed to support student achievement; and
3. Evaluating school programs.

A detailed description of the uses and applications of CAASPP reports is presented in the next section.

Types of Score Reports

There are three categories of CST for Science reports. These categories and the specific reports in each category are given in Table 7.4.

Table 7.4 Types of CST for Science Reports

1. Reports in the ORS	▪ Home Page Dashboard ▪ Listing (Group, Roster, Student) scale scores and performance levels by grade and content area ▪ Student detail
2. Student Results Report in TOMS	▪ Student Score Data Extract

3. Student Score Reports These reports are printed and available as downloadable PDFs from the Test Operations Management System (TOMS).	▪ Student Score Report for Smarter Balanced Summative Assessments for English language arts/literacy and mathematics and CST for Science—Grades five and eight ▪ Student Score Report for CST for Science—Grade ten
4. Aggregated Internet Reports (Internet reporting) These reports are available at the Public Reporting Web site at http://caaspp.cde.ca.gov/.	▪ CST for Science Scores

The CAASPP aggregate reports and student data files for the LEA are available for the LEA CAASPP coordinator to download from the Test Operations Management System (TOMS). The LEA forwards the appropriate reports to test sites or, in the case of the CAASPP Student Score Report, sends the report(s) to the child's parent or guardian and forwards a copy to the student's school or test site. CAASPP Student Score Reports that include individual student results are not distributed beyond the student's school. Internet reports are described on the CDE Web site and are accessible to the public online at http://caaspp.cde.ca.gov/.

Because results were pre-equated, individual student scores were also available to LEAs prior to the release of final reports via electronic reporting, accessed using the ORS. This application permits LEAs to view preliminary results data for all tests taken.

Student Score Report Contents

The CAASPP Student Score Report provides scale scores and performance level results for the CST for Science taken. Scale scores are reported on a scale ranging from 150 to 600. The performance levels reported are: far below basic, below basic, basic, proficient, and advanced.

Reports for students with disabilities and ELs who use universal tools, designated supports, or accommodations include a notation that indicates the student used non-embedded supports (accommodations) or was tested with non-embedded accessibility supports (modifications).

Scores for students who use non-embedded supports are reported in the same way as they are for nonaccommodated students. Non-embedded accessibility supports (modifications), however, change what is being tested and, therefore, change scores. If students use non-embedded accessibility supports (modifications), their scores are counted differently from nonmodified test scores on summary reports—CST for Science scores for these students are counted as far below basic, regardless of the scale score obtained.

Further information about the CAASPP Student Score Report and the other reports is provided in Appendix 7.C on page 79.

Student Score Report Applications

CST for Science results provide parents and guardians with information about their child's progress. The results are a tool for increasing communication and collaboration between parents or guardians and teachers. Along with report cards from teachers and information from school and classroom tests, the CAASPP Student Score Report can be used by parents and guardians while talking with teachers about ways to improve their child's achievement of the California content standards.

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

CST for Science results help LEAs and schools identify strengths and weaknesses in their instructional programs. Each year, LEAs and school staffs examine CST for Science results for each test administered. Their findings are used to help determine:

- The extent to which students are learning the academic standards,
- Instructional areas that can be improved,
- Teaching strategies that can be developed to address needs of students, and
- Decisions about how to use funds to ensure that students achieve the standards.

Criteria for Interpreting Test Scores

An LEA may use CST for Science results to help make decisions about student placement, promotion, retention, or other considerations related to student achievement. However, it is important to remember that a single test can provide only limited information. Other relevant information should be considered as well. It is advisable for parents to evaluate their child's strengths and weaknesses in the relevant topics by reviewing classroom work and progress reports in addition to the child's CST for Science results (CDE, 2016). It is also important to note that a student's score in a content area contains measurement error and could vary somewhat if the student were retested.

Criteria for Interpreting Score Reports

The information presented in various reports must be interpreted with caution when making performance comparisons. When comparing scale score and performance-level results for the CSTs for Science, the user is limited to comparisons within the same content area and grade. This is because the score scales are different for each content area and grade. The user may compare scale scores for the same content area and grade, within a school, between schools, or between a school and its district, its county, or the state. The user can also make comparisons within the same grade and content area across years. Comparing scores obtained in different grades or content areas should be avoided because the results are not on the same scale. Comparisons between raw scores should be limited to comparisons within not only content area and grade but also test year. For more details on the criteria for interpreting information provided on the score reports, see the *2015–16 CAASPP Post-Test Guide* (CDE, 2016).

References

California Department of Education. (2016). *2015–16 CAASPP post-test guide.* Sacramento, CA. Retrieved from http://www.caaspp.org/rsc/pdfs/CAASPP.post-test_guide.2016.pdf

Educational Testing Service. (2002). *ETS standards for quality and fairness*. Princeton, NJ: Author.

Appendix 7.A—Scale Score Distribution Tables

A cell value of "N/A" indicates that there are no obtainable scale scores within that scale-score range for the particular CST for Science.

Table 7.A.1. Distribution of CST for Science Scale Scores

Scale Score	Grade 5 Science	Grade 8 Science	Grade 10 Life Science
570 – 600	1,178	15,527	1,176
540 – 569	3,458	9,324	2,733
510 – 539	N/A	22,570	4,726
480 – 509	6,061	12,907	6,505
450 – 479	19,593	42,835	17,702
420 – 449	41,607	45,654	33,650
390 – 419	62,960	45,692	51,445
360 – 389	74,350	56,548	80,606
330 – 359	75,547	48,980	76,245
300 – 329	76,260	48,594	83,618
270 – 299	46,078	34,924	45,773
240 – 269	25,222	20,127	33,029
210 – 239	8,979	17,930	10,858
180 – 209	1,127	6,862	984
150 – 179	78	4,541	64

A frequency count of "N/A" indicates that there are no obtainable scale scores within that scale-score range.

Appendix 7.B—Demographic Summaries

To protect privacy when the number of students in a subgroup is 10 or fewer, the summary statistics at the test- and reporting-cluster–level are not reported and are presented as hyphens in the tables in Appendix 7.B. Percentages in these tables may not sum up to 100 due to rounding.

Table 7.B.1 Demographic Summary for Science, Grade Five

	Number Tested	Mean Scale Score	Std. Dev. of Scale Scores	Percent in Far Below Basic	Percent in Below Basic	Percent in Basic	Percent in Proficient	Percent in Advanced	Mean Percent Correct in Physical Science, Grade Five	Mean Percent Correct in Physical Science, Grade Four	Mean Percent Correct in Life Science, Grade Five	Mean Percent Correct in Life Science, Grade Four	Mean Percent Correct in Earth Science, Grade Five	Mean Percent Correct in Earth Science, Grade Four
All valid scores	442,498	357	64	8%	10%	28%	34%	20%	73%	65%	68%	71%	66%	62%
Male	221,927	359	67	9%	10%	26%	33%	21%	73%	66%	68%	72%	67%	63%
Female	220,571	355	62	7%	11%	30%	34%	18%	74%	63%	69%	71%	66%	62%
Gender unknown	0	--	--	--	--	--	--	--	--	--	--	--	--	--
American Indian	2,254	344	60	10%	13%	32%	32%	13%	70%	61%	64%	68%	63%	58%
Asian American	41,537	400	68	3%	4%	15%	34%	44%	84%	77%	79%	82%	75%	75%
Pacific Islander	2,200	344	57	9%	13%	34%	33%	11%	71%	60%	65%	67%	63%	58%
Filipino	11,023	382	57	2%	4%	22%	42%	29%	81%	73%	76%	78%	72%	72%
Hispanic	239,292	338	57	11%	14%	35%	30%	10%	68%	59%	63%	66%	61%	55%
African American	23,559	331	58	15%	16%	33%	27%	9%	67%	56%	61%	63%	59%	53%
White	105,907	386	60	3%	4%	19%	41%	32%	81%	73%	76%	80%	74%	72%
Two or more races	16,726	380	65	5%	6%	20%	38%	31%	80%	71%	74%	78%	72%	70%
English only	248,515	369	63	5%	8%	25%	37%	25%	77%	68%	72%	75%	70%	67%
Initially fluent English prof.	18,450	401	63	1%	3%	16%	37%	42%	85%	77%	80%	83%	77%	75%
EL	88,249	305	48	23%	24%	36%	14%	2%	57%	49%	52%	54%	51%	44%
Reclassified fluent Eng. prof.	86,288	367	53	2%	6%	32%	42%	19%	77%	69%	73%	76%	70%	65%
To Be Determined (TBD)	304	310	73	38%	13%	18%	19%	11%	60%	49%	51%	53%	51%	44%
English prof. unknown	692	324	73	29%	13%	19%	27%	12%	63%	57%	56%	60%	55%	51%
No special ed. services	410,770	360	63	7%	10%	28%	35%	21%	74%	66%	69%	72%	67%	63%
Special ed. services	31,728	321	65	23%	17%	28%	22%	10%	61%	53%	57%	60%	55%	50%
Special ed. unknown	0	--	--	--	--	--	--	--	--	--	--	--	--	--
Not econ. disadvantaged	167,724	392	61	2%	4%	17%	40%	36%	83%	75%	78%	81%	75%	74%
Economically disadvantaged	274,774	336	57	11%	14%	35%	30%	10%	68%	59%	63%	65%	61%	55%
Unknown economic status	0	--	--	--	--	--	--	--	--	--	--	--	--	--
Primary Ethnicity—Not Economically Disadvantaged														
American Indian	730	373	61	4%	6%	24%	40%	25%	78%	69%	73%	77%	71%	68%
Asian American	26,413	419	63	1%	2%	9%	32%	55%	89%	82%	83%	87%	80%	81%
Pacific Islander	695	369	59	4%	5%	27%	42%	21%	78%	67%	72%	75%	71%	68%
Filipino	6,991	391	56	1%	3%	17%	44%	35%	84%	76%	78%	81%	75%	75%
Hispanic	41,846	368	58	4%	7%	26%	40%	22%	77%	68%	72%	75%	70%	66%
African American	5,548	360	60	7%	9%	29%	37%	19%	75%	64%	69%	73%	68%	63%
White	75,049	398	57	1%	2%	15%	42%	39%	84%	77%	79%	83%	77%	76%
Two or more races	10,452	398	62	2%	3%	14%	39%	41%	85%	77%	79%	83%	76%	76%

	Number Tested	Mean Scale Score	Std. Dev. of Scale Scores	Percent in Far Below Basic	Percent in Below Basic	Percent in Basic	Percent in Proficient	Percent in Advanced	Mean Percent Correct in Physical Science, Grade Five	Mean Percent Correct in Physical Science, Grade Four	Mean Percent Correct in Life Science, Grade Five	Mean Percent Correct in Life Science, Grade Four	Mean Percent Correct in Earth Science, Grade Five	Mean Percent Correct in Earth Science, Grade Four
Primary Ethnicity—Economically Disadvantaged														
American Indian	1,524	330	55	13%	17%	35%	28%	8%	66%	56%	60%	64%	59%	54%
Asian American	15,124	367	64	6%	8%	25%	36%	24%	77%	68%	71%	74%	68%	65%
Pacific Islander	1,505	332	52	11%	17%	37%	29%	7%	67%	57%	62%	64%	60%	53%
Filipino	4,032	364	56	4%	7%	29%	40%	19%	77%	68%	71%	73%	69%	65%
Hispanic	197,446	331	54	12%	16%	36%	28%	7%	66%	57%	61%	64%	60%	53%
African American	18,011	322	55	17%	18%	35%	24%	6%	65%	53%	58%	60%	56%	50%
White	30,858	357	57	7%	8%	29%	39%	16%	74%	65%	69%	72%	67%	62%
Two or more races	6,274	350	59	9%	10%	31%	36%	14%	72%	62%	67%	70%	65%	59%
Primary Ethnicity—Unknown Economic Status														
American Indian	0	--	--	--	--	--	--	--	--	--	--	--	--	--
Asian American	0	--	--	--	--	--	--	--	--	--	--	--	--	--
Pacific Islander	0	--	--	--	--	--	--	--	--	--	--	--	--	--
Filipino	0	--	--	--	--	--	--	--	--	--	--	--	--	--
Hispanic	0	--	--	--	--	--	--	--	--	--	--	--	--	--
African American	0	--	--	--	--	--	--	--	--	--	--	--	--	--
White	0	--	--	--	--	--	--	--	--	--	--	--	--	--
Two or more races	0	--	--	--	--	--	--	--	--	--	--	--	--	--

Table 7.B.2 Demographic Summary for Science, Grade Eight

	Number Tested	Mean Scale Score	Std. Dev. of Scale Scores	Percent in Far Below Basic	Percent in Below Basic	Percent in Basic	Percent in Proficient	Percent in Advanced	Mean Percent Correct in Motion	Mean Percent Correct in Forces, Density, and Buoyancy	Mean Percent Correct in Structure of Matter and Periodic Table	Mean Percent Correct in Earth in the Solar System	Mean Percent Correct in Reactions and the Chemistry of Living Systems	Mean Percent Correct in Investigation and Experimentation
All valid scores	433,015	381	95	9%	11%	20%	23%	38%	66%	72%	68%	70%	65%	75%
Male	218,152	385	100	10%	10%	18%	22%	40%	66%	73%	68%	71%	65%	74%
Female	214,863	378	89	8%	11%	21%	24%	35%	66%	71%	68%	69%	65%	76%
Gender unknown	0	--	--	--	--	--	--	--	--	--	--	--	--	--
American Indian	2,425	356	89	13%	14%	21%	24%	28%	62%	67%	63%	65%	61%	70%
Asian American	40,991	453	96	3%	4%	9%	16%	69%	79%	84%	81%	80%	78%	89%
Pacific Islander	2,256	362	86	11%	13%	23%	25%	28%	64%	68%	65%	66%	61%	73%
Filipino	12,281	418	84	3%	5%	14%	25%	54%	72%	79%	76%	76%	73%	84%
Hispanic	229,650	353	84	12%	14%	24%	24%	25%	61%	67%	62%	66%	60%	69%
African American	24,692	339	85	17%	16%	25%	22%	21%	59%	63%	59%	62%	58%	66%
White	106,882	418	90	4%	5%	13%	23%	55%	72%	79%	75%	76%	73%	83%
Two or more races	13,838	406	96	7%	7%	15%	22%	50%	70%	77%	72%	74%	70%	80%
English only	235,392	395	94	7%	8%	17%	23%	44%	68%	75%	71%	72%	68%	78%
Initially fluent English prof.	20,854	432	95	3%	5%	12%	20%	59%	75%	81%	77%	78%	75%	85%
EL	49,203	293	71	31%	25%	25%	12%	6%	50%	54%	48%	54%	46%	52%
Reclassified fluent Eng. prof.	126,507	383	83	5%	10%	23%	27%	35%	67%	73%	68%	71%	66%	77%
To Be Determined (TBD)	303	312	107	38%	18%	15%	10%	19%	55%	57%	50%	54%	48%	59%
English prof. unknown	756	324	99	30%	13%	19%	18%	19%	57%	60%	54%	58%	53%	61%
No special ed. services	404,072	386	93	8%	10%	19%	24%	40%	67%	73%	69%	71%	66%	77%
Special ed. services	28,943	311	88	29%	20%	21%	15%	14%	52%	56%	53%	59%	51%	55%
Special ed. unknown	0	--	--	--	--	--	--	--	--	--	--	--	--	--
Not econ. disadvantaged	171,057	426	92	3%	5%	12%	22%	58%	74%	80%	76%	77%	74%	84%
Economically disadvantaged	261,958	352	85	13%	14%	24%	24%	25%	61%	67%	62%	66%	60%	69%
Unknown economic status	0	--	--	--	--	--	--	--	--	--	--	--	--	--
Primary Ethnicity—Not Economically Disadvantaged														
American Indian	872	391	89	7%	9%	16%	24%	43%	68%	74%	70%	71%	69%	78%
Asian American	25,567	478	87	1%	2%	5%	13%	79%	83%	87%	85%	83%	82%	92%
Pacific Islander	804	391	88	6%	9%	17%	27%	41%	69%	74%	71%	71%	67%	79%
Filipino	7,929	430	83	2%	4%	11%	24%	59%	74%	81%	78%	78%	75%	86%
Hispanic	43,268	388	87	6%	9%	19%	26%	40%	67%	74%	70%	72%	67%	78%
African American	6,737	369	89	11%	11%	21%	24%	33%	64%	69%	66%	68%	64%	74%
White	77,261	434	86	2%	3%	10%	22%	62%	75%	82%	78%	79%	76%	87%
Two or more races	8,619	430	93	4%	4%	11%	20%	61%	74%	81%	77%	78%	75%	85%

	Number Tested	Mean Scale Score	Std. Dev. of Scale Scores	Percent in Far Below Basic	Percent in Below Basic	Percent in Basic	Percent in Proficient	Percent in Advanced	Mean Percent Correct in Motion	Mean Percent Correct in Forces, Density, and Buoyancy	Mean Percent Correct in Structure of Matter and Periodic Table	Mean Percent Correct in Earth in the Solar System	Mean Percent Correct in Reactions and the Chemistry of Living Systems	Mean Percent Correct in Investigation and Experimentation
Primary Ethnicity—Economically Disadvantaged														
American Indian	1,553	337	83	17%	17%	24%	23%	19%	58%	64%	59%	62%	57%	66%
Asian American	15,424	412	95	5%	7%	15%	22%	52%	72%	77%	74%	74%	71%	82%
Pacific Islander	1,452	346	81	13%	14%	26%	25%	21%	61%	65%	61%	63%	58%	69%
Filipino	4,352	396	82	4%	7%	18%	27%	44%	69%	76%	72%	73%	69%	80%
Hispanic	186,382	345	81	13%	15%	26%	24%	21%	60%	66%	60%	65%	58%	67%
African American	17,955	328	80	19%	18%	26%	21%	16%	57%	61%	57%	60%	55%	62%
White	29,621	377	86	8%	10%	21%	26%	35%	65%	72%	67%	70%	65%	75%
Two or more races	5,219	367	88	11%	11%	21%	25%	32%	63%	69%	65%	68%	63%	73%
Primary Ethnicity—Unknown Economic Status														
American Indian	0	--	--	--	--	--	--	--	--	--	--	--	--	--
Asian American	0	--	--	--	--	--	--	--	--	--	--	--	--	--
Pacific Islander	0	--	--	--	--	--	--	--	--	--	--	--	--	--
Filipino	0	--	--	--	--	--	--	--	--	--	--	--	--	--
Hispanic	0	--	--	--	--	--	--	--	--	--	--	--	--	--
African American	0	--	--	--	--	--	--	--	--	--	--	--	--	--
White	0	--	--	--	--	--	--	--	--	--	--	--	--	--
Two or more races	0	--	--	--	--	--	--	--	--	--	--	--	--	--

Table 7.B.3 Demographic Summary for Grade Ten Life Science

	Number Tested	Mean Scale Score	Std. Dev. of Scale Scores	Percent in Far Below Basic	Percent in Below Basic	Percent in Basic	Percent in Proficient	Percent in Advanced	Mean Percent Correct in Cell Biology	Mean Percent Correct in Genetics	Mean Percent Correct in Physiology	Mean Percent Correct in Ecology	Mean Percent Correct in Evolution	Mean Percent Correct in Investigation and Experimentation
All valid scores	449,114	354	65	9%	12%	30%	27%	23%	56%	58%	68%	69%	64%	75%
Male	227,760	355	68	10%	11%	27%	26%	25%	58%	58%	68%	70%	64%	73%
Female	221,354	352	61	7%	12%	32%	28%	21%	54%	59%	67%	68%	64%	77%
Gender unknown	0	--	--	--	--	--	--	--	--	--	--	--	--	--
American Indian	2,558	343	60	10%	14%	32%	27%	17%	53%	54%	65%	67%	61%	71%
Asian American	43,942	397	70	3%	5%	17%	27%	49%	70%	72%	78%	79%	76%	87%
Pacific Islander	2,283	338	57	11%	15%	34%	27%	14%	52%	54%	63%	65%	60%	69%
Filipino	13,577	375	57	3%	6%	25%	34%	33%	62%	66%	75%	76%	72%	83%
Hispanic	235,203	334	56	11%	15%	36%	25%	12%	49%	52%	62%	63%	59%	69%
African American	25,447	325	57	16%	18%	34%	22%	10%	48%	49%	60%	60%	55%	63%
White	112,169	380	64	4%	6%	22%	31%	38%	64%	66%	75%	77%	72%	83%
Two or more races	13,935	373	67	6%	8%	24%	29%	34%	62%	64%	73%	74%	70%	80%
English only	242,138	364	65	6%	9%	27%	29%	29%	59%	61%	71%	72%	68%	78%
Initially fluent English prof.	32,952	381	66	3%	6%	24%	31%	37%	64%	67%	75%	76%	72%	83%
EL	47,921	289	42	32%	30%	29%	7%	1%	37%	39%	47%	46%	43%	48%
Reclassified fluent Eng. prof.	124,972	351	55	5%	11%	37%	29%	17%	54%	58%	67%	69%	64%	76%
To Be Determined (TBD)	314	307	58	32%	20%	26%	14%	9%	40%	44%	53%	54%	49%	57%
English prof. unknown	817	310	62	29%	19%	27%	16%	10%	41%	44%	54%	55%	52%	57%
No special ed. services	415,939	358	64	7%	11%	30%	28%	25%	57%	60%	69%	70%	66%	76%
Special ed. services	33,175	304	56	29%	24%	29%	12%	6%	43%	42%	52%	52%	47%	53%
Special ed. unknown	0	--	--	--	--	--	--	--	--	--	--	--	--	--
Not econ. disadvantaged	188,949	381	65	4%	6%	22%	30%	38%	64%	67%	75%	77%	72%	83%
Economically disadvantaged	260,165	333	57	12%	16%	36%	24%	12%	50%	52%	62%	63%	59%	69%
Unknown economic status	0	--	--	--	--	--	--	--	--	--	--	--	--	--
Primary Ethnicity—Not Economically Disadvantaged														
American Indian	1,054	364	62	5%	9%	28%	31%	27%	59%	61%	72%	74%	67%	78%
Asian American	26,742	416	67	1%	2%	11%	25%	60%	75%	77%	82%	83%	80%	90%
Pacific Islander	900	358	59	7%	8%	29%	34%	22%	58%	60%	69%	71%	66%	76%
Filipino	9,002	384	57	2%	4%	21%	35%	38%	65%	68%	77%	78%	74%	86%
Hispanic	50,952	355	59	7%	10%	31%	30%	22%	56%	59%	69%	70%	65%	76%
African American	7,981	343	59	10%	13%	32%	28%	17%	53%	55%	66%	66%	62%	70%
White	83,462	389	62	3%	4%	19%	32%	43%	67%	69%	78%	79%	75%	86%
Two or more races	8,856	389	65	3%	4%	19%	30%	43%	67%	69%	78%	79%	75%	85%
Primary Ethnicity—Economically Disadvantaged														
American Indian	1,504	327	54	14%	17%	35%	24%	10%	48%	49%	60%	63%	57%	65%
Asian American	17,200	369	65	6%	8%	25%	30%	31%	61%	64%	71%	73%	68%	81%
Pacific Islander	1,383	325	52	13%	19%	37%	22%	8%	48%	50%	59%	60%	56%	64%
Filipino	4,575	358	55	4%	9%	31%	33%	22%	57%	60%	70%	71%	67%	78%
Hispanic	184,251	328	53	13%	17%	38%	23%	9%	48%	51%	60%	61%	57%	67%
African American	17,466	316	54	19%	21%	34%	19%	7%	45%	46%	58%	57%	52%	59%
White	28,707	351	60	8%	11%	31%	29%	21%	55%	57%	68%	69%	64%	75%
Two or more races	5,079	344	60	10%	13%	31%	28%	17%	53%	55%	66%	67%	62%	71%

	Number Tested	Mean Scale Score	Std. Dev. of Scale Scores	Percent in Far Below Basic	Percent in Below Basic	Percent in Basic	Percent in Proficient	Percent in Advanced	Mean Percent Correct in Cell Biology	Mean Percent Correct in Genetics	Mean Percent Correct in Physiology	Mean Percent Correct in Ecology	Mean Percent Correct in Evolution	Mean Percent Correct in Investigation and Experimentation
Primary Ethnicity—Unknown Economic Status														
American Indian	0	--	--	--	--	--	--	--	--	--	--	--	--	--
Asian American	0	--	--	--	--	--	--	--	--	--	--	--	--	--
Pacific Islander	0	--	--	--	--	--	--	--	--	--	--	--	--	--
Filipino	0	--	--	--	--	--	--	--	--	--	--	--	--	--
Hispanic	0	--	--	--	--	--	--	--	--	--	--	--	--	--
African American	0	--	--	--	--	--	--	--	--	--	--	--	--	--
White	0	--	--	--	--	--	--	--	--	--	--	--	--	--
Two or more races	0	--	--	--	--	--	--	--	--	--	--	--	--	--

Appendix 7.C—Types of Score Reports

Table 7.C.1 Score Reports Reflecting CST for Science Results

2015–16 CAASPP Student Score Reports	
Description	**Use and Distribution**
The CAASPP Student Score Report—CST for Science A report for the Smarter Balanced Summative Assessments for English Language Arts/Literacy and Mathematics and the CST for Science at the student's grade level (grade five or eight) or Life Science	
This report provides parents/guardians and teachers with the student's results, presented in tables and graphs. Data presented for the science assessment taken include the following: • Scale scores • Performance levels	This report includes individual student results and is not distributed beyond parents/guardians and the student's school. Two copies of this report are provided for each student. One is for the student's current teacher and one is to be distributed by the local educational agency (LEA) to parents/guardians.
Subgroup Summary	
This set of reports disaggregates and reports results by the following subgroups: • All students • Disability status • Economic status • Gender • English proficiency • Primary ethnicity • Economic status These reports contain no individual student-identifying information and are aggregated at the school, LEA, county, and state levels. For each subgroup within a report and for the total number of students, the following data are included for each test: • Total number tested in the subgroup • Percent of enrollment tested in the subgroup • Number and percent of valid scores • Number tested who received scores • Mean scale score • Standard deviation of scale score • Number and percent of students scoring at each performance level	This report is a resource for evaluators, researchers, teachers, parents/guardians, community members, and administrators. Each LEA can download this report for the whole LEA and the schools within it from the Test Operations Management System. ***Note:*** The data on this report may be shared with parents/guardians, community members, and the media only if the data are for 11 or more students.

Chapter 8: Analyses

Background

This chapter summarizes the item- and test-level statistics obtained for the California Standards Tests (CSTs) for Science administered during the 2015–16 test administration.

The statistics presented in this chapter are divided into three sections in the following order:

1. Classical Item Analyses
2. Reliability Analyses
3. Item Response Theory (IRT) Analyses

Prior to the 2012–13 administration, differential item functioning (DIF) analyses were performed based on the final item analysis (FIA) sample for all operational and field-test items to assess differences in the item performance of groups of students that differ in their demographic characteristics. In 2015–16, because the intact forms were used, DIF analyses were not performed.

Each of the sets of analyses on data from the 2015–16 administration is presented in the body of the text and in the appendixes as listed below.

1. Appendix 8.A on page 98 presents the classical item analyses, including proportion-correct value (*p*-value) and point-biserial correlation (Pt-Bis) for each item in each operational test. Because all forms were intact, *p*-values and Pt-Bis are shown for both the original and the current administration of the tests. In addition, the average and median *p*-value and Pt-Bis for the operational test forms based on their current administration are presented in Table 8.1 on page 81.
2. Appendix 8.B on page 99 presents results of the reliability analyses of total test scores and subscores for the population as a whole and for selected subgroups. Also presented are results of the analyses of the accuracy and consistency of the performance classifications.
3. Appendix 8.C on page 111 presents the scoring tables obtained as a result of the IRT equating process. (For the summaries of *b*-values for the three CSTs for Science, refer to Appendix D of the *CST Technical Report* in the year each grade-level science form was administered originally; see Table 8.4 on page 90 for administration years.)

Samples Used for the Analyses

CST for Science analyses were conducted at different times after test administration and involved varying proportions of the full CST for Science data. The classical item analyses presented in Appendix 8.A and the reliability statistics included in Appendix 8.B were calculated using the sample of student data that was the last “daily feed” data received on September 15, 2016, which comprised approximately 100 percent of the full CST for Science data. This file contained data collected from all local educational agencies (LEAs) but did not include corrections of demographic data through the California Longitudinal Pupil Achievement Data System. In addition, students with invalid scores were excluded.

During the 2015–16 administration, neither IRT calibrations nor scaling are implemented because the intact forms from 2011–12 were reused and results were pre-equated. For the reused intact forms, the IRT results were derived based on the equating sample of the previous administration which can be found in Appendix D of the *CST Technical Report* in

the year each grade-level science form was administered originally; see Table 8.4 on page 90 for administration years.

Classical Item Analyses

Multiple-Choice Items

The classical item statistics that included overall and item-by-item proportion-correct indices and the point-biserial correlation indices were computed for the operational items. The *p*-value of an item represents the proportion of students in the sample that answered an item correctly. The formula for *p*-value is:

$$p\text{-}value_i = \frac{N_{ic}}{N_i} \tag{8.1}$$

where,

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

N_i is the total number of students who attempted the item.

The point-biserial correlation is a special case of the Pearson product-moment correlation used to measure the strength of the relationship between two variables, one dichotomously and one continuously measured—in this case, the item score (right/wrong) and the total test score. The formula for the Pearson product-moment correlation is:

$$r_{X_iT} = \frac{\mathrm{cov}(X_i, T)}{s_{X_i} s_T} \tag{8.2}$$

where,

$cov(X_i, T)$ is the covariance between the score of item *i* and total score T,

s_{X_i} is the standard deviation for the score of item i, and

s_T is the standard deviation for total score T.

The classical statistics for the current administration of the overall test are presented in Table 8.1. The item-by-item values for the classical statistics, including *p*-values, point-biserial correlations, distributional percentages, and mean scores, are presented in Table 8.A.1 on page 98. Each set of values is presented for both the current and the original presentation of each CST for Science item.

Table 8.1 Mean and Median Proportion Correct and Point-Biserial by Test Form—Current Administration

CST*	No. of Items	No. of Students	Mean *p*-value	Mean Pt-Bis	Median *p*-value	Median Pt-Bis
Grade 5 Science	60	442,498	0.68	0.42	0.67	0.42
Grade 8 Science	60	433,015	0.69	0.42	0.71	0.42
Grade 10 Life Science	60	449,114	0.64	0.43	0.65	0.44

Reliability Analyses

Reliability focuses on the extent to which differences in test scores reflect true differences in the knowledge, ability, or skill being tested, rather than fluctuations due to chance or random factors. The variance in the distribution of test scores—essentially, the differences

among individuals—is partly due to real differences in the knowledge, skill, or ability being tested (true-score variance) and partly due to random unsystematic errors in the measurement process (error variance).

The number used to describe reliability is an estimate of the proportion of the total variance that is true-score variance. Several different ways of estimating this proportion exist. The estimates of reliability reported here are internal-consistency measures, which are derived from analysis of the consistency of the performance of individuals on items within a test (internal-consistency reliability). They reflect the proportion of variance associated with the true score were students administered forms containing different exemplars of the content found in the current test form. Therefore, they apply only to the test form being analyzed. They do not take into account form-to-form variation due to equating limitations or lack of parallelism, nor are they responsive to day-to-day variation due, for example, to students' state of health or testing environment.

Reliability coefficients can range from 0 to 1. The higher the reliability coefficient for a set of scores, the more likely individuals would be to obtain very similar scores if they were retested. The formula for the internal-consistency reliability as measured by Cronbach's Alpha (Cronbach, 1951) is defined by equation 8.3:

$$\alpha = \frac{n}{n-1}\left[1 - \frac{\sum_{i=1}^{n} s_i^{\,2}}{s_t^2}\right] \tag{8.3}$$

where,

n is the number of items,

s_i^2 is the variance of scores on item i, and

s_t^2 is the variance of the total score.

The standard error of measurement (SEM) provides a measure of score instability in the score metric. The SEM was computed as shown in equation 8.4:

$$s_e = s_t\sqrt{1-\alpha} \tag{8.4}$$

where,

α is the reliability estimated in equation 8.3, and

s_t is the standard deviation of the total score (either the total raw score or scale score).

The SEM is particularly useful in determining the confidence interval (CI) that captures a student's true score. Assuming that measurement error is normally distributed, it can be said that upon infinite replications of the testing occasion, approximately 95 percent of the CIs of $\pm$1.96 SEM around the observed score would contain a student's true score (Crocker & Algina, 1986). For example, if a student's observed score on a given test equals 15 points, and the SEM equals 1.92, one can be 95 percent confident that the student's true score lies between 11 and 19 points (15 $\pm$ 3.76 rounded to the nearest integer).

Table 8.2 shows the reliability and SEM for each of the CSTs for Science, along with the number of items and students upon whom those analyses were performed.

Table 8.2 Reliabilities and SEMs for the CSTs for Science

CST	No. of Items	No. of Students	Reliab.	Mean Scale Score	Scale Score Std. Dev.	Scale Score SEM	Mean Raw Score	Raw Score Std. Dev.	Raw Score SEM
Grade 5 Science	60	442,498	0.92	357	64	18.38	40.83	11.31	3.23
Grade 8 Science	60	433,015	0.92	381	95	27.09	41.45	11.02	3.15
Grade 10 Life Science	60	449,114	0.92	354	65	17.95	38.45	11.69	3.23

Intercorrelations, Reliabilities, and SEMs for Reporting Clusters

For each grade-level CST for Science, number-correct scores are computed for six reporting clusters. The number of items within each reporting cluster is limited, and cluster scores alone should not be used in making inferences about individual students.

Intercorrelations and reliability estimates for the reporting clusters are presented in Table 8.B.1 on page 99. Consistent with results from previous years, the reliabilities across reporting clusters vary significantly according to the number of items in each cluster.

Subgroup Reliabilities and SEMs

The reliabilities of the CSTs for Science were examined for various subgroups of the student population. The subgroups included in these analyses were defined by their gender, ethnicity, economic status, provision of special services, and English-language fluency. The reliability analyses are also presented by primary ethnicity within economic status.

Reliabilities and SEM information for the total test scores and the reporting cluster scores are reported for each subgroup analysis. Table 8.B.2 through Table 8.B.32 present the overall test reliabilities for the various subgroups. Table 8.B.33 through Table 8.B.37 present the cluster-level reliabilities for the subgroups. Table 8.B.33 presents the cluster-level reliabilities for the subgroups based on gender and economic status.

The next table, Table 8.B.34, shows the same analyses for the subgroups based on provision of special services and English-language fluency. Table 8.B.35 presents results for the subgroups based on primary ethnicity of the students. The last set of tables, Table 8.B.36 and Table 8.B.37, present the cluster-level reliabilities for the subgroups based on primary ethnicity within economic status.

Reliability values are estimates that approach the true reliability as the number of student whose scores contribute to the estimates increases. Reliabilities are reported only for samples that comprise 11 or more students. Results based on samples that contain 50 or fewer students should be interpreted with caution because these estimate may meaningfully deviate from the true reliability. Also, in some cases, score reliabilities were not estimable and are presented in the tables as hyphens.

Conditional Standard Errors of Measurement

As part of the IRT-based equating procedures, scale-score conversion tables and conditional standard errors of measurement (CSEMs) are produced. CSEMs for CST scale scores are based on IRT and are calculated by the IRTEQUATE module in a computer system called the Generalized Analysis System (GENASYS).

The CSEM is estimated as a function of measured ability. It is typically smaller in scale score units toward the center of the scale in the test metric, where more items are located, and larger at the extremes, where there are fewer items. A student's CSEM under the IRT framework is equal to the inverse of the square root of the test information function:

$$\mathrm{CSEM}(\hat{\theta}) = \frac{1}{\sqrt{I\left(\hat{\theta}\right)}} a \quad (8.5)$$

where,

$\mathrm{CSEM}(\hat{\theta})$ is the standard error of measurement, and

$\mathrm{I}(\hat{\theta})$ is the test information function at ability level $\hat{\theta}$.

The statistic is multiplied by a, where a is the original scaling factor needed to transform theta to the scale-score metric. The value of a varies by grade level.

CSEMs vary across the scale. When a test has cut scores, it is important to provide CSEMs at the cut scores.

Table 8.3 presents the scale score CSEMs at the lowest score required for a student to be classified in the below basic, basic, proficient, and advanced performance levels for each CST for Science.

The CSEMs tend to be higher at the advanced cut points for all tests. The pattern of lower values of CSEMs at the basic and proficient levels are expected since (1) more items tend to be of middle difficulty; and (2) items at the extremes still provide information toward the middle of the scale. This results in more precise scores in the middle of the scale and less precise scores at the extremes of the scale.

Table 8.3 Scale Score CSEM at Performance-level Cut Points

CST	Below Basic Min SS	Below Basic CSEM	Basic Min SS	Basic CSEM	Proficient Min SS	Proficient CSEM	Advanced Min SS	Advanced CSEM
Grade 5 Science	268	16	300	15	350	17	410	22
Grade 8 Science	253	23	300	23	350	24	403	27
Grade 10 Life Science	269	16	300	15	350	16	399	19

Decision Classification Analyses

The methodology used for estimating the reliability of classification decisions is described in Livingston and Lewis (1995) and is implemented using the Educational Testing Service (ETS)-proprietary computer program RELCLASS-COMP (Version 4.14).

Decision accuracy describes the extent to which students are classified in the same way as they would be on the basis of the average of all possible forms of a test. Decision accuracy answers the following question: How does the actual classification of test-takers, based on their single-form scores, agree with the classification that would be made on the basis of their true scores, if their true scores were somehow known? RELCLASS-COMP estimates decision accuracy using an estimated multivariate distribution of reported classifications on the current form of the exam and the classifications based on an all-forms average (true score).

Decision consistency describes the extent to which students are classified in the same way as they would be on the basis of a single form of a test other than the one for which data are available. Decision consistency answers the following question: What is the agreement between the classifications based on two nonoverlapping, equally difficult forms of the test? RELCLASS-COMP also estimates decision consistency using an estimated multivariate

distribution of reported classifications on the current form of the exam and classifications on a hypothetical alternate form using the reliability of the test and strong true-score theory.

In each case, the proportion of classifications with exact agreement is the sum of the entries in the diagonal of the contingency table representing the multivariate distribution. Reliability of classification at a cut score is estimated by collapsing the multivariate distribution at the passing score boundary into an *n* by *n* table (where *n* is the number of performance levels) and summing the entries in the diagonal. Figure 8.1 and Figure 8.2 present the two scenarios graphically.

Figure 8.1 Decision Accuracy for Achieving a Performance Level

		Decision made on the all-forms average	
		Does not achieve a performance level	Achieves a performance level
True status on the form taken	Does not achieve a performance level	Correct classification	Misclassification
	Achieves a performance level	Misclassification	Correct classification

Figure 8.2 Decision Consistency for Achieving a Performance Level

		Decision made on a hypothetical alternate form	
		Does not achieve a performance level	Achieves a performance level
Decision made on the form taken	Does not achieve a performance level	Correct classification	Misclassification
	Achieves a performance level	Misclassification	Correct classification

The results of these analyses are presented in Table 8.B.38 through Table 8.B.40 in Appendix 8.B, starting on page 99.

Each table includes the contingency tables for both accuracy and consistency of the various performance-level classifications. The proportion of students being accurately classified is determined by summing across the diagonals of the upper tables. The proportion of consistently classified students is determined by summing the diagonals of the lower tables.

The classifications are collapsed to below-proficient versus proficient and above.

Validity Evidence

Validity refers to the degree to which each interpretation or use of a test score is supported by evidence that is gathered (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014; ETS, 2002). It is a central concern underlying the development, administration, and scoring of a test and the uses and interpretations of test scores.

Validation is the process of accumulating evidence to support each proposed score interpretation or use. It involves more than a single study or gathering of one particular kind of evidence. Validation involves multiple investigations and various kinds of evidence (AERA, APA, & NCME, 2014; Cronbach, 1971; ETS, 2002; Kane, 2006). The process begins with test design and continues through the entire assessment process, including item development and field testing, analyses of item and test data, test scaling, scoring, and score reporting.

This section presents the evidence gathered to support the intended uses and interpretations of scores for the CST for Science testing program. The description is organized according to the kinds of evidence included in *The Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014). These standards require a clear definition of the purpose of the test, which includes a description of the qualities—called constructs—that are to be assessed by a test, the population to be assessed, as well as how the scores are to be interpreted and used.

In addition, the *Standards* identify five kinds of evidence that can provide support for score interpretations and uses, which are as follows:

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

These kinds of evidence are also defined as important elements of validity information in documents developed by the U.S. Department of Education (USDOE) for the peer review of testing programs administered by states in response to the Elementary and Secondary Education Act (USDOE, 2001).

The next section defines the purpose of the CSTs for Science, followed by a description and discussion of the kinds of validity evidence that have been gathered.

Purpose of the CSTs for Science

As mentioned in Chapter 1, the CSTs for Science comprise the California Assessment of Student Performance and Progress (CAASPP) System's implementation of the remaining paper-pencil tests. The CSTs for Science are designed to show how well students in grades five, eight, and ten are performing with respect to California's content standards in science that were adopted by the SBE in 1998. These content standards were approved in 1998 by the State Board of Education (SBE); they describe what students should know and be able to do at each grade level.

The Constructs to Be Measured

Test blueprints and specifications written to define the procedures used to measure the content standards provide an operational definition of the construct to which each set of standards refers—that is, they define, for each content area to be assessed, the tasks to be presented, the administration instructions to be given, and the rules used to score student responses. They control as many aspects of the measurement procedure as possible so that the testing conditions will remain the same over test administrations (Cronbach, 1971; Cronbach, Gleser, Nanda, & Rajaratnam, 1972) to minimize construct-irrelevant score variance (Messick, 1989). The content blueprints for the CSTs for Science can be found on the CDE STAR CST Blueprints Web page at http://www.cde.ca.gov/ta/tg/sr/blueprints.asp. ETS developed all CST for Science test items to conform to the SBE-approved content standards and test blueprints.

Interpretations and Uses of the Scores Generated

Total test scores expressed as scale scores and student performance levels are generated for each grade-level and content-area test. The total test scale score is used to draw inferences about a student's achievement in the content area and to classify the

achievement into one of five performance levels: advanced, proficient, basic, below basic, and far below basic.

Reporting cluster scores, also called subscores, are used to draw inferences about a student's achievement in each of several specific knowledge or skill areas covered by each test. In past years, when cluster results were reported, the results compared an individual student's percent-correct score to the average percent-correct for the state as a whole. The range of scores for students who scored proficient on the total test was also provided for each cluster using a percent-correct metric. The reference points for this range were: (1) the average percent-correct for students who received the lowest score qualifying for the proficient performance level; and (2) the average percent-correct for students who received the lowest score qualifying for the advanced performance level, minus one percent. A detailed description of the uses and applications of CST for Science scores as used in past years is presented in the Student Score Reports Applications section on page 69 of Chapter 7. Note that these were not used in reporting student results to LEAs or test sites, or in Student Score Reports.

The tests that make up the CAASPP System in science, along with other assessments, provide results or score summaries that are used for different purposes. The three major purposes are:

1. Communicating with parents and guardians;
2. Informing decisions needed to support student achievement; and
3. Evaluating school programs.

These are the only uses and interpretations of scores for which validity evidence has been gathered. If the user wishes to interpret or use the scores in other ways, the user is cautioned that the validity of doing so has not been established (AERA, APA, & NCME, 2014, Standard 1.3). The user is advised to gather evidence to support these additional interpretations or uses (AERA, APA, & NCME, 2014, Standard 1.4).

Intended Test Population(s)

California public school students in grades five, eight, and ten are the intended test population for the CSTs for Science. Only those students whose parents/guardians have submitted written requests to exempt them from CAASPP System testing do not take a grade-level science test. See the subsection "Intended Population" on page 2 for a more detailed description of the intended test population.

Validity Evidence Collected

Evidence Based on Content

According to *The Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014), analyses that demonstrate a strong relationship between a test's content and the construct that the test was designed to measure can provide important evidence of validity. In current K–12 testing, the construct of interest usually is operationally defined by state content standards and the test blueprints that specify the content, format, and scoring of items that are admissible measures of the knowledge and skills described in the content standards. Evidence that the items meet these specifications and represent the domain of knowledge and skills referenced by the standards supports the inference that students' scores on these items can appropriately be regarded as measures of the intended construct.

As noted in the AERA, APA, and NCME *Standards* (2014), evidence based on test content may involve logical analyses of test content in which experts judge the adequacy with which the test content conforms to the test specifications and represents the intended domain of content. Such reviews can also be used to determine whether the test content contains material that is not relevant to the construct of interest. Analyses of test content may also involve the use of empirical evidence of item quality.

Also to be considered in evaluating test content are the procedures used for test administration and test scoring. As Kane (2006, p. 29) has noted, although evidence that appropriate administration and scoring procedures have been used does not provide compelling evidence to support a particular score interpretation or use, such evidence may prove useful in refuting rival explanations of test results. Evidence based on content includes the following:

Description of the state standards—As was noted in Chapter 1, the SBE adopted rigorous content standards in 1997 and 1998 in four major content areas: English–language arts, history–social science, mathematics, and science. These standards were designed to guide instruction and learning for all students in the state and to bring California students to world-class levels of achievement. The content standards for science adopted in 1998 guided the development of the CSTs for Science.

Specifications and blueprints—ETS maintains item specifications for each CST for Science. Item specifications describe the characteristics of items that should be written to measure each content standard. A thorough description of the specifications can be found in Chapter 3, starting on page 29. Once items were developed and field-tested, ETS selected all CST for Science test items to conform to the SBE-approved California content standards and test blueprints. Test blueprints for the CSTs for Science were proposed by ETS and reviewed and approved by the Assessment Review Panels (ARPs), which are advisory panels to the CDE and ETS on areas related to item development for the CSTs. Test blueprints were also reviewed and approved by the CDE and presented to the SBE for adoption. There have been no recent changes in the blueprints for the CSTs. The test blueprints for the CSTs for Science can be found on the CDE STAR CST Blueprints Web page at http://www.cde.ca.gov/ta/tg/sr/blueprints.asp.

Item development process—A detailed description of the item development process for the CSTs for Science is presented in Chapter 3, starting on page 29.

Item review process—Chapter 3 explains in detail the extensive item review process applied to items that were written for use in the CSTs for Science. In brief, items written for the CSTs for Science underwent multiple review cycles and involved multiple groups of reviewers. One of the reviews was carried out by an external reviewer, that is, the ARPs. The ARPs were responsible for reviewing all newly developed items for alignment to the California content standards.

Form construction process—For each test, the content standards, blueprints, and test specifications were used as the basis for choosing items. Additional targets for item difficulty and discrimination that were used for test construction were defined in light of what are desirable statistical characteristics in test items and statistical evaluations of the CST for Science items.

Guidelines for test construction were established with the goal of maintaining parallel forms to the greatest extent possible from year to year. Details can be found in Chapter 4, starting on page 39.

Additionally, an external review panel, the Statewide Pupil Assessment Review (SPAR), was responsible for reviewing and approving the achievement tests to be used statewide for the testing of students in California public schools, grades two through eleven. More information about the SPAR is given in Chapter 3, starting on page 35.

Alignment study—Strong alignment between standards and assessments is fundamental to meaningful measurement of student achievement and instructional effectiveness. Alignment results should demonstrate that the assessments represent the full range of the content standards and that these assessments measure student knowledge in the same manner and at the same level of complexity as expected in the content standards.

Human Resource Research Organization (HumRRO) performed an alignment study for the CSTs in April 2007. HumRRO utilized the Webb alignment method to evaluate the alignment of the 2006 CSTs to the California content standards. The Webb method requires a set of raters to evaluate each test item on two different dimensions: (1) the standard(s) targeted by items, and (2) the depth of knowledge required of students to respond to items. These ratings form the basis of the four separate Webb alignment analyses: categorical concurrence, depth-of-knowledge consistency, range-of-knowledge correspondence, and balance-of-knowledge representation. Good alignment was found for the CSTs in English–language arts, mathematics, science, and history–social science.

Participation of California Educators—California educators, including teachers and school administrators, participate in deciding the cut score, which is the minimum test score that is required to qualify a student for each performance level, for each test. They are invited as panelists to the standard setting meetings and provide their individual judgments of cut scores. The final recommended cut scores are based on the median of panelists' judgment scores. See Chapter 6 Standard Setting which starts on page 58 for more details about this procedure.

California educators who have strong content and teaching backgrounds were included and trained to be item writers. See Chapter 3 Item Development which starts on page 29 for more details on selection and training of item writers.

Evidence Based on Relations to Other Variables

Empirical results concerning the relationships between the score on a test and measures of other variables external to the test can also provide evidence of validity when these relationships are found to be consistent with the definition of the construct that the test is intended to measure. As indicated in the *Standards* (AERA, APA, & NCME, 2014), the variables investigated can include other tests that measure the same construct and different constructs, criterion measures that scores on the test are expected to predict, as well as demographic characteristics of students that are expected to be related and unrelated to test performance.

Correlations Between Scores on the CSTs and Scores on the CAT/6 Survey

Convergent validity evidence was collected in 2004 by examining the relationship between CSTs and their CAT/6 Survey (TerraNova, 2nd Edition, 2000) counterparts. The CAT/6 Survey is a norm-referenced test that assesses students in reading, language, spelling, mathematics, and science and evaluates student achievement in terms of norms. The CSTs were expected to relate closely to their counterparts in the CAT/6 Survey programs when they measured similar constructs, and to correlate less well when they measured different constructs. A full description of the study can be found in the *California Standardized*

Testing Program Technical Report, Spring 2005 Administration (CDE, 2005). A summary of findings follows:

Correlations Between Scores on the CST for English–Language Arts (ELA) and Scores on the CAT/6 Survey Reading/Language/Spelling—The study showed that, as expected, CST for ELA scores in all grades correlated highly with scores on both the CAT/6 Survey Reading Language tests, because these tests assessed similar skills. The correlation coefficients between the CST for ELA and CAT/6 Survey Spelling tests were somewhat lower, which is to be expected because these tests measured somewhat different skills.

Correlations Between Scores on the CST for Mathematics and Scores on the CAT/6 Survey Mathematics—In grades two through seven, student scores on the CST Mathematics tests correlated highly with their scores on CAT/6 Survey Mathematics test. This was expected because these tests assessed similar skills. In general, more moderate results were found in the upper grades when students' CAT/6 scores were correlated with the end-of-course CSTs and the integrated tests. This was expected since the CSTs at the upper grade levels were designed to measure more specific content defined by the state's content standards, whereas the CAT/6 tests were designed to assess content that was most commonly taught across the nation at the time that the CAT/6 tests were published.

Correlations Between Scores on the CST for Science and Scores on the CAT/6 Survey Science—All end-of-course (EOC) science CSTs correlated moderately high with the CAT/6 Survey Science tests across grades. This was expected since the EOC tests were designed to assess a narrower range of course-related content than were the CAT/6 Survey Science tests.

Differential Item Functioning Analyses

Analyses of DIF can provide evidence of the degree to which a score interpretation or use is valid for individuals who differ in particular demographic characteristics. For the CSTs for Science, DIF analyses were performed after the test forms' original administration on all operational items and all field-test items for which sufficient student samples were available.

The results of the DIF analyses are presented in Appendix 8.E of the *CST Technical Report* produced for the year the form was administered originally. The report is linked on the CDE Technical Reports and Studies Web page at http://www.cde.ca.gov/ta/tg/sr/technicalrpts.asp. The year of original administration for each CST for Science is shown in Table 8.4.

Table 8.4 Original Year of Administration for CSTs for Science

CST	Year
Grade 5 Science	2011–12
Grade 8 Science	2011–12
Grade 10 Life Science	2011–12

Evidence Based on Response Processes

As noted in the AERA, APA, and NCME *Standards* (2014), additional support for a particular score interpretation or use can be provided by theoretical and empirical evidence indicating that students are using the intended response processes when responding to the items in a test. This evidence may be gathered from interacting with students in order to understand what processes underlie their item responses.

Evidence Based on Internal Structure

As suggested by the *Standards* (AERA, APA, & NCME, 2014), evidence of validity can also be obtained from studies of the properties of the item scores and the relationship between these scores and scores on components of the test. To the extent that the score properties and relationships found are consistent with the definition of the construct measured by the test, support is gained for interpreting these scores as measures of the construct.

For the CSTs for Science, it is assumed that a single construct underlies the total scores obtained on each test. Evidence to support this assumption can be gathered from the results of item analyses, evaluations of internal consistency, and studies of dimensionality and reliability.

With respect to the subscores that are reported, these scores are intended to reflect students' knowledge and/or skill in an area that is part of the construct underlying the total test. Analyses of the intercorrelations among the subscores themselves and between the subscores and total test score can be used for studying this aspect of the construct. Information about the internal consistency of the items on which each subscore is based is also useful to provide.

Classical Statistics

Point-biserial correlations calculated for the items in a test show the degree to which the items discriminate between students with low and high scores on a test. To the degree that the correlations are high, evidence that the items assess the same construct is provided. As shown in Table 8.1, the mean point biserial was between 0.42 and 0.43. The point biserials for the individual items in the CSTs for Science of the 2015–16 administration and their previous administrations are presented in Table 8.A.1.

Also germane to the validity of a score interpretation are the ranges of item difficulty for the items on which a test score will be based. The finding that items have difficulties that span the range of student ability provides evidence that students at all levels of ability are adequately measured by the items. Information on average item *p*-values is given in Table 8.1; individual *p*-values are presented in Table 8.A.1 side by side with the *p*-values of these items obtained when the intact forms were used originally.

The summaries of *b*-values can be found in Appendix D of the *CST Technical Report* in the year the form was administered originally; see Table 8.4 on page 90 for administration years.)

The data in Table 8.1 indicate that all of the CSTs for Science had *p*-values with means ranging from 0.64 to 0.69.

Reliability

Reliability is a prerequisite for validity. The finding of reliability in student scores supports the validity of the inference that the scores reflect a stable construct. This section will describe briefly findings concerning the total test level, as well as reliability results for the reporting clusters.

Overall reliability—The reliability analyses on each of the operational CSTs for Science are presented in Table 8.2. The results indicate that the reliabilities for the grade-level CSTs for Science were very high, all 0.92.

Reporting cluster reliabilities—For each CST for Science, number-correct scores are computed for the reporting clusters. The reliabilities of these scores are presented in

Table 8.B.1. The reliabilities of reporting clusters invariably are lower than those for the total tests because they are based on very few items. Consistent with the findings of previous years, the cluster reliabilities also are affected by the number of items in each cluster, with cluster scores based on fewer items having somewhat lower reliabilities than cluster scores based on more items.

Because the reliabilities of scores at the cluster level are lower, schools supplement the score results with other information when interpreting the results.

Subgroup reliabilities—The reliabilities of the operational CSTs for Science are also examined for various subgroups of the student population that differed in their demographic characteristics. The characteristics considered are gender, ethnicity, economic status, provision of special services, English-language fluency, and ethnicity-for-economic status. The results of these analyses can be found in Table 8.B.2 through Table 8.B.32.

Reliability of performance classifications—The methodology used for estimating the reliability of classification decisions is described in the section "Decision Classification Analyses" on page 84. The results of these analyses are presented in Table 8.B.38 through Table 8.B.40 in Appendix 8.B; these tables start on page 109. When the classifications are collapsed to below-proficient versus proficient and above, the proportion of students that were classified accurately was 0.92 across all CSTs for Science. Similarly, the proportion of students that were classified consistently ranged from 0.88 to 0.89 for students classified into below-proficient versus proficient and advanced.

These levels of accuracy and consistency are high, and they are consistent with levels seen in previous years.

Evidence Based on Consequences of Testing

As observed in the *Standards*, tests are usually administered "in the expectation that some benefit will be realized from the interpretation and use intended by the test developers" (AERA, APA, & NCME, 2014, p. 19). When this is the case, evidence that the expected benefits accrue will provide support for the intended use of the scores. The CDE and ETS are in the process of determining what kinds of information can be gathered to assess the consequences of administration of the CSTs for Science.

IRT Analyses

Post-Equating

Prior to the 2012–13 administration, the CSTs for Science were equated to a reference form using a common-item nonequivalent groups design and post-equating methods based on IRT. The "base" or "reference" calibrations for the CSTs for Science were established by calibrating samples of data from a specific administration. Doing so established a scale to which subsequent item calibrations could be linked.

The procedures used for post-equating the CSTs for Science prior to 2013 involved three steps: item calibration, item parameter scaling, and true-score equating. ETS used GENASYS for the IRT item calibration and equating work. As part of this system, a proprietary version of the PARSCALE computer program (Muraki & Bock, 1995) was used and parameterized to result in one-parameter calibrations. Research at ETS has suggested that PARSCALE calibrations done in this manner produce results that are virtually identical to results based on WINSTEPS (Way, Kubiak, Henderson, & Julian, 2002). The post-equating procedures were applied to all of CSTs for Science.

Pre-Equating

During the 2015–16 administration, because intact forms were used from the 2011–12 administration without any edits or replacement of items, conversion tables from the original administration are directly applied to the current administration.

Descriptions of IRT analyses such as the model-data fit analyses can be found in Chapter 8 of the original-year technical report; the results of the IRT analyses are presented in Appendix 8.D of the *CST Technical Report* for the year each grade-level science form was administered originally. Reports are linked on the CDE Technical Reports and Studies Web page at http://www.cde.ca.gov/ta/tg/sr/technicalrpts.asp. The year of original administration for each multiple-choice CST for Science is shown in Table 8.4.

The details on all equating procedures are given in Chapter 2, starting on page 14.

Summaries of Scaled IRT *b*-values

For the post-equating procedures implemented prior to the 2012–13 administration, once the IRT *b*-values were placed on the item bank scale, analyses were performed to assess the overall test difficulty, the difficulty level of reporting clusters, and the distribution of items in a particular range of item difficulty.

During the 2015–16 administration, for all CSTs for Science the raw-to-scale score tables from the 2011–12 administration are directly applied. Neither IRT calibration nor scaling is implemented, but banked *b*-value parameters derived through the post-equating procedure from their previous administrations are used for pre-equating.

The summaries of *b*-values of the operational items of the three CST for Science tests can be found in Appendix D of the *CST Technical Report* in the year each grade-level science form was administered originally; see Table 8.4 on page 90 for administration years.

The distributions of *b*-values of the operational items of the three CST for Science tests can be found in Appendix D of the *CST Technical Report* in the year each grade-level science form was administered originally; see Table 8.4 on page 90 for administration years.

Evaluation of Pre-Equating

Pre-equating is performed on the basis of the assumption of IRT models that item parameters remain invariant across samples given a similar ability distribution. To produce results that are sufficiently accurate for high-stakes decisions, the intact forms were used so that item parameters were obtained from large, representative samples, and factors that may affect item parameter estimations, such as context effects (e.g., item positions) and speededness, were well controlled.

To ensure that items performed similarly in the current administration as in the year they were administered originally in the intact forms, comparisons of classical statistics such as *p*-values and point-biserial correlations are made between the current administration and the item bank values in the year of the original administration.

Equating Results

During the 2015–16 administration, for all CSTs for Science intact forms without any edits, the conversion tables from their original administration in 2011–12 (listed in Table 8.4 on page 90) are directly applied to the current administration. For braille CST for Science forms in which the nonbraillable items were removed—that is, the CST for Science in grades five and eight—the conversion table was developed by equating the shortened test to the full-length test from the original administration as is described in Chapter 2

Complete raw-score-to-scale-score conversion tables for the CSTs for Science administered in 2015–16 are presented in Table 8.C.1 through Table 8.C.5 starting on page 111. The raw scores and corresponding transformed scale scores are listed in those tables. The scale scores were truncated at both ends of the scale so that the minimum reported scale score was 150 and the maximum reported scale score was 600. The scale scores defining the various performance-level cut points are presented in Table 2.1, which is in Chapter 2 on page 16.

Differential Item Functioning Analyses

Analyses of DIF assess differences in the item performance of groups of students who differ in their demographic characteristics.

Prior to the 2012–13 administration, DIF analyses were performed based on the FIA sample and were performed on all operational items and on all field-test items for which sufficient student samples were available. DIF analyses are not implemented for the 2015–16 administration because intact forms were used and all items were evaluated for DIF during the previous administration when the forms were used originally. These DIF results can be found in Appendix E of the *CST Technical Report* in the year each grade-level science form was administered originally; see Table 8.4 on page 90 for administration years.

The statistical procedure of DIF analysis that was conducted prior to the 2012–13 administration is described in this section.

The sample size requirements for the DIF analyses were 100 in the focal group and 400 in the combined focal and reference groups. These sample sizes were based on standard operating procedures with respect to DIF analyses at ETS. The DIF analyses utilized the Mantel-Haenszel (MH) DIF statistic (Mantel & Haenszel, 1959; Holland & Thayer, 1985). This statistic is based on the estimate of constant odds ratio and is described as the following:

The α_{MH} is the constant odds ratio taken from Dorans and Holland (1993, equation 7) and computed as the following:

$$\alpha_{MH} = \frac{\left(\sum_m R_{rm} \frac{W_{fm}}{N_{tm}}\right)}{\left(\sum_m R_{fm} \frac{W_{rm}}{N_{tm}}\right)} \tag{8.6}$$

$$MH\ D\text{-}DIF = -2.35 \ln[\alpha_{MH}] \tag{8.7}$$

where,

R = number right,

W = number wrong,

N = total in:

fm = focal group at ability m,

rm = reference group at ability m, and

tm = total group at ability m.

Items analyzed for DIF at ETS are classified into one of three categories: A, B, or C. Category A contains items with negligible DIF. Category B contains items with slight to moderate DIF. Category C contains items with moderate to large values of DIF.

These categories have been used by ETS testing programs for more than 15 years. The definitions of the categories based on evaluations of the item-level MH D-DIF statistics are as follows:

DIF Category	Definition
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-.”

The factors considered in the DIF analyses included gender, ethnicity, level of English-language fluency, and primary disability.

Tables also listed the operational and field-test items exhibiting significant DIF (C-DIF). Test developers were instructed to avoid selecting field-test items flagged as having shown DIF that disadvantages a focal group (C-DIF) for future operational test forms unless their inclusion was deemed essential to meeting test-content specifications.

References

AERA, APA, & NCME. 2014. *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

California Department of Education. (2005). *California Standardized Testing Program technical report, spring 2005 administration*. Sacramento, CA. Retrieved from http://www.cde.ca.gov/ta/tg/sr/documents/startechrpt05.pdf

California Department of Education. (2010). *California Standardized Testing Program technical report, spring 2010 administration*. Sacramento, CA. Retrieved from http://www.cde.ca.gov/ta/tg/sr/documents/csttechrpt2010.pdf

California Department of Education. (2011). *California Standardized Testing Program technical report, spring 2011 administration*. Sacramento, CA. Retrieved from http://www.cde.ca.gov/ta/tg/sr/documents/csttechrpt2011.pdf

California Department of Education. (2012). *California Standardized Testing Program technical report, spring 2012 administration*. Sacramento, CA. Retrieved from http://www.cde.ca.gov/ta/tg/sr/documents/cst12techrpt.pdf

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

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16,* 292–334.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York, NY: Wiley.

CTB/McGraw-Hill. (2000). *TerraNova, The Second Edition, CAT Technical Report.* Monterey, CA: Author.

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.

Educational Testing Service. (2002). *ETS standards for quality and fairness*. Princeton, NJ: Author.

Gaffney, T., Cudeck, R., Ferrer, E., & Widaman, K. F. (2010). On the factor structure of standardized educational achievement tests. *Journal of Applied Measurement, 11(4),* 384–408.

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

HumRRO. (2007). *Independent evaluation of the alignment of the California Standards Tests (CSTs) and the California Alternate Performance Assessment (CAPA).* Alexandria, VA: Author. Retrieved from http://www.cde.ca.gov/ta/tg/sr/documents/alignmentreport.pdf

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Washington, DC: American Council on Education and National Council on Measurement in Education.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classification based on test scores. *Journal of Educational Measurement*, *32,* 179–97.

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

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp.13–103). New York, NY: Macmillan.

Muraki, E., & Bock, R. D. (1995). *PARSCALE: Parameter scaling of rating data* (Computer software, Version 2.2). Chicago, IL: Scientific Software.

United States Department of Education (2001). Elementary and Secondary Education Act (Public Law 107–11), Title VI, Chapter B, § 4, Section 6162. Retrieved from http://www2.ed.gov/policy/elsec/leg/esea02/index.html

Way, W. D., Kubiak, A. T., Henderson, D., & Julian, M. W. (2002, April). *Accuracy and stability of calibrations for mixed-item-format tests using the 1-parameter and generalized partial credit models.* Paper presented at the meeting of the National Council on Measurement in Education, New Orleans, LA.

Appendix 8.A—Classical Analyses

Table 8.A.1 Item-by-item *p*-value and Point Biserial for Science, Grades Five, Eight, and Ten—Current Year (2016) and Original Year of Administration

Years	Grade 5 2016		Grade 5 2012		Grade 8 2016		Grade 8 2012		Grade 10 2016		Grade 10 2012	
Items	*p*-value	Pt-Bis	*p*-value	Pt-Bis	*p*-value	Pt-Bis	*p*-value	Pt-Bis	*p*-value	Pt-Bis	*p*-value	Pt-Bis
1	0.77	0.47	0.77	0.49	0.85	0.39	0.87	0.41	0.69	0.33	0.71	0.35
2	0.73	0.36	0.74	0.39	0.52	0.25	0.51	0.26	0.82	0.33	0.81	0.38
3	0.83	0.37	0.85	0.38	0.77	0.28	0.81	0.32	0.76	0.40	0.77	0.43
4	0.67	0.45	0.71	0.47	0.79	0.43	0.80	0.44	0.62	0.36	0.60	0.38
5	0.83	0.44	0.84	0.45	0.51	0.38	0.52	0.41	0.70	0.41	0.66	0.43
6	0.67	0.39	0.68	0.37	0.58	0.32	0.57	0.34	0.32	0.37	0.36	0.38
7	0.66	0.40	0.69	0.39	0.71	0.45	0.71	0.44	0.54	0.39	0.64	0.40
8	0.58	0.43	0.60	0.43	0.55	0.37	0.59	0.36	0.60	0.35	0.58	0.34
9	0.66	0.49	0.69	0.51	0.49	0.49	0.53	0.50	0.41	0.38	0.43	0.38
10	0.51	0.46	0.61	0.48	0.71	0.49	0.75	0.50	0.31	0.40	0.31	0.39
11	0.65	0.39	0.68	0.41	0.84	0.50	0.86	0.52	0.60	0.58	0.63	0.59
12	0.76	0.37	0.77	0.37	0.62	0.41	0.63	0.42	0.58	0.37	0.63	0.39
13	0.70	0.37	0.72	0.40	0.86	0.52	0.87	0.52	0.62	0.46	0.62	0.48
14	0.57	0.33	0.61	0.34	0.88	0.39	0.89	0.43	0.79	0.40	0.80	0.44
15	0.85	0.39	0.88	0.39	0.73	0.33	0.72	0.32	0.40	0.53	0.43	0.51
16	0.69	0.32	0.66	0.30	0.80	0.33	0.79	0.36	0.42	0.26	0.40	0.28
17	0.60	0.44	0.69	0.45	0.67	0.43	0.69	0.46	0.56	0.51	0.57	0.52
18	0.80	0.36	0.82	0.36	0.71	0.38	0.76	0.43	0.41	0.24	0.43	0.27
19	0.73	0.33	0.76	0.33	0.86	0.47	0.85	0.49	0.57	0.49	0.58	0.51
20	0.73	0.50	0.76	0.51	0.54	0.32	0.57	0.35	0.53	0.41	0.58	0.43
21	0.66	0.47	0.70	0.47	0.78	0.46	0.80	0.48	0.81	0.38	0.82	0.43
22	0.64	0.31	0.70	0.33	0.77	0.48	0.78	0.49	0.71	0.52	0.77	0.52
23	0.79	0.42	0.80	0.43	0.83	0.39	0.84	0.40	0.80	0.45	0.78	0.48
24	0.67	0.31	0.68	0.33	0.66	0.41	0.71	0.43	0.62	0.44	0.63	0.47
25	0.42	0.24	0.47	0.29	0.79	0.52	0.81	0.53	0.62	0.44	0.62	0.46
26	0.59	0.36	0.61	0.34	0.78	0.45	0.80	0.48	0.60	0.32	0.61	0.35
27	0.84	0.36	0.84	0.40	0.74	0.49	0.77	0.49	0.53	0.41	0.49	0.40
28	0.61	0.32	0.60	0.35	0.74	0.52	0.72	0.52	0.82	0.41	0.80	0.42
29	0.66	0.50	0.67	0.49	0.45	0.45	0.52	0.44	0.72	0.48	0.75	0.51
30	0.76	0.40	0.74	0.39	0.61	0.46	0.67	0.50	0.57	0.37	0.64	0.37
31	0.80	0.47	0.83	0.45	0.81	0.45	0.85	0.47	0.80	0.48	0.82	0.51
32	0.82	0.35	0.84	0.34	0.67	0.38	0.70	0.40	0.50	0.51	0.49	0.49
33	0.75	0.54	0.80	0.52	0.73	0.37	0.75	0.36	0.85	0.48	0.84	0.52
34	0.54	0.41	0.52	0.37	0.62	0.53	0.69	0.56	0.65	0.44	0.64	0.46
35	0.71	0.47	0.73	0.46	0.30	0.29	0.36	0.30	0.74	0.51	0.71	0.54
36	0.67	0.49	0.71	0.51	0.78	0.49	0.79	0.51	0.70	0.41	0.67	0.42
37	0.77	0.52	0.81	0.50	0.69	0.44	0.72	0.47	0.68	0.57	0.68	0.59
38	0.76	0.44	0.79	0.44	0.87	0.48	0.87	0.51	0.80	0.44	0.80	0.49
39	0.61	0.57	0.73	0.54	0.72	0.41	0.75	0.43	0.50	0.34	0.55	0.39
40	0.72	0.50	0.72	0.48	0.78	0.50	0.81	0.50	0.62	0.33	0.64	0.36
41	0.45	0.37	0.51	0.39	0.86	0.43	0.88	0.43	0.78	0.55	0.75	0.57
42	0.58	0.43	0.62	0.43	0.49	0.35	0.57	0.41	0.36	0.31	0.37	0.28
43	0.58	0.41	0.58	0.41	0.69	0.41	0.73	0.45	0.87	0.46	0.85	0.50
44	0.66	0.43	0.70	0.43	0.47	0.31	0.52	0.35	0.58	0.36	0.60	0.42
45	0.43	0.30	0.46	0.30	0.66	0.44	0.71	0.46	0.75	0.50	0.76	0.53
46	0.74	0.38	0.78	0.40	0.67	0.53	0.73	0.57	0.66	0.48	0.67	0.50
47	0.80	0.48	0.82	0.49	0.65	0.46	0.65	0.47	0.75	0.54	0.75	0.56
48	0.80	0.48	0.80	0.48	0.81	0.43	0.83	0.47	0.40	0.31	0.44	0.34
49	0.67	0.40	0.70	0.42	0.87	0.39	0.88	0.41	0.64	0.32	0.64	0.36
50	0.54	0.30	0.58	0.30	0.62	0.38	0.63	0.39	0.37	0.37	0.41	0.38
51	0.56	0.48	0.63	0.50	0.49	0.36	0.55	0.41	0.75	0.48	0.73	0.50
52	0.85	0.46	0.87	0.47	0.84	0.49	0.85	0.51	0.83	0.48	0.83	0.52
53	0.66	0.39	0.73	0.42	0.63	0.43	0.64	0.43	0.81	0.53	0.81	0.56
54	0.76	0.52	0.79	0.52	0.79	0.40	0.80	0.41	0.79	0.52	0.77	0.56
55	0.72	0.44	0.75	0.44	0.45	0.37	0.49	0.38	0.75	0.44	0.75	0.47
56	0.64	0.44	0.69	0.44	0.60	0.36	0.60	0.36	0.70	0.53	0.72	0.55
57	0.62	0.41	0.63	0.41	0.74	0.42	0.74	0.41	0.63	0.45	0.65	0.48
58	0.82	0.48	0.82	0.49	0.69	0.40	0.75	0.44	0.66	0.55	0.65	0.56
59	0.65	0.53	0.68	0.54	0.61	0.45	0.61	0.45	0.70	0.46	0.73	0.48
60	0.53	0.42	0.52	0.39	0.73	0.42	0.72	0.41	0.79	0.51	0.78	0.52

Appendix 8.B—Reliability Analyses

The reliabilities are reported only for samples that comprise 11 or more examinees. Also, in some cases in Appendix 8.B, score reliabilities were not estimable and are presented in the tables as hyphens.

Table 8.B.1 Subscore Reliabilities and Intercorrelations for Science

Subscore Area	N of Items							Reliab.	SEM
Grade 5 Science									
1. Physical Science (Grade 5)	11	1.00	-	-	-	-	-	0.69	1.33
2. Physical Science (Grade 4)	8	0.61	1.00	-	-	-	-	0.64	1.19
3. Life Science (Grade 5)	13	0.67	0.62	1.00	-	-	-	0.70	1.52
4. Life Science (Grade 4)	9	0.64	0.61	0.67	1.00	-	-	0.67	1.20
5. Earth Science (Grade 5)	11	0.61	0.57	0.65	0.63	1.00	-	0.64	1.38
6. Earth Science (Grade 4)	8	0.61	0.59	0.65	0.64	0.63	1.00	0.64	1.22
Grade 8 Science									
1. Motion	8	1.00	-	-	-	-	-	0.52	1.21
2. Forces, Density, and Buoyancy	13	0.59	1.00	-	-	-	-	0.72	1.44
3. Structure of Matter and Periodic Table	16	0.58	0.71	1.00	-	-	-	0.78	1.61
4. Earth in the Solar System	7	0.46	0.58	0.60	1.00	-	-	0.54	1.05
5. Reactions and the Chemistry of Living Systems	10	0.53	0.65	0.68	0.56	1.00	-	0.66	1.33
6. Investigation and Experimentation	6	0.55	0.67	0.67	0.51	0.63	1.00	0.64	0.91
Grade 10 Life Science									
1. Cell Biology	10	1.00	-	-	-	-	-	0.65	1.36
2. Genetics	12	0.64	1.00	-	-	-	-	0.74	1.48
3. Physiology	10	0.60	0.61	1.00	-	-	-	0.66	1.32
4. Ecology	11	0.59	0.61	0.65	1.00	-	-	0.72	1.32
5. Evolution	11	0.59	0.65	0.64	0.69	1.00	-	0.70	1.38
6. Investigation and Experimentation	6	0.50	0.57	0.58	0.64	0.67	1.00	0.73	0.88

Table 8.B.2 Reliabilities and SEMs for the CSTs for Science by Gender (Female)

CST	N	Rel	SEM
Grade 5 Science	220,571	0.91	3.25
Grade 8 Science	214,863	0.91	3.19
Grade 10 Life Science	221,354	0.91	3.25

Table 8.B.3 Reliabilities and SEMs for the CSTs for Science by Gender (Male)

CST	N	Rel	SEM
Grade 5 Science	221,927	0.92	3.20
Grade 8 Science	218,152	0.93	3.12
Grade 10 Life Science	227,760	0.93	3.20

Table 8.B.4 Reliabilities and SEMs for the CSTs for Science by Economic Status (Not Economically Disadvantaged)

CST	N	Rel	SEM
Grade 5 Science	167,724	0.90	2.91
Grade 8 Science	171,057	0.91	2.87
Grade 10 Life Science	188,949	0.92	3.02

Table 8.B.5 Reliabilities and SEMs for the CSTs for Science by Economic Status (Economically Disadvantaged)

CST	N	Rel	SEM
Grade 5 Science	274,774	0.90	3.40
Grade 8 Science	261,958	0.90	3.32
Grade 10 Life Science	260,165	0.91	3.37

Table 8.B.6 Reliabilities and SEMs for the CSTs for Science by Special Services (No Special Services)

CST	N	Rel	SEM
Grade 5 Science	410,770	0.91	3.21
Grade 8 Science	404,072	0.91	3.13
Grade 10 Life Science	415,939	0.92	3.21

Table 8.B.7 Reliabilities and SEMs for the CSTs for Science by Special Services (Special Services)

CST	N	Rel	SEM
Grade 5 Science	31,728	0.93	3.42
Grade 8 Science	28,943	0.91	3.46
Grade 10 Life Science	33,175	0.90	3.48

Table 8.B.8 Reliabilities and SEMs for the CSTs for Science by English-Language Fluency (English Only)

CST	N	Rel	SEM
Grade 5 Science	248,515	0.91	3.13
Grade 8 Science	235,392	0.92	3.07
Grade 10 Life Science	242,138	0.92	3.15

Table 8.B.9 Reliabilities and SEMs for the CSTs for Science by English-Language Fluency (Initially Designated Fluent)

CST	N	Rel	SEM
Grade 5 Science	18,450	0.89	2.83
Grade 8 Science	20,854	0.91	2.84
Grade 10 Life Science	32,952	0.92	3.03

Table 8.B.10 Reliabilities and SEMs for the CSTs for Science by English-Language Fluency (EL)

CST	N	Rel	SEM
Grade 5 Science	88,249	0.88	3.57
Grade 8 Science	49,203	0.87	3.57
Grade 10 Life Science	47,921	0.84	3.56

Table 8.B.11 Reliabilities and SEMs for the CSTs for Science by English-Language Fluency (Redesigned Fluent)

CST	N	Rel	SEM
Grade 5 Science	86,288	0.87	3.20
Grade 8 Science	126,507	0.89	3.18
Grade 10 Life Science	124,972	0.90	3.28

Table 8.B.12 Reliabilities and SEMs for the CSTs for Science by Primary Ethnicity (American Indian)

CST	N	Rel	SEM
Grade 5 Science	2,254	0.91	3.33
Grade 8 Science	2,425	0.91	3.28
Grade 10 Life Science	2,558	0.92	3.31

Table 8.B.13 Reliabilities and SEMs for the CSTs for Science by Primary Ethnicity (Asian)

CST	N	Rel	SEM
Grade 5 Science	41,537	0.92	2.82
Grade 8 Science	40,991	0.91	2.67
Grade 10 Life Science	43,942	0.92	2.88

Table 8.B.14 Reliabilities and SEMs for the CSTs for Science by Primary Ethnicity (Pacific Islander)

CST	N	Rel	SEM
Grade 5 Science	2,200	0.90	3.35
Grade 8 Science	2,256	0.91	3.27
Grade 10 Life Science	2,283	0.91	3.36

Table 8.B.15 Reliabilities and SEMs for the CSTs for Science by Primary Ethnicity (Filipino)

CST	N	Rel	SEM
Grade 5 Science	11,023	0.89	3.03
Grade 8 Science	12,281	0.89	2.94
Grade 10 Life Science	13,577	0.90	3.09

Table 8.B.16 Reliabilities and SEMs for the CSTs for Science by Primary Ethnicity (Hispanic)

CST	N	Rel	SEM
Grade 5 Science	239,292	0.90	3.39
Grade 8 Science	229,650	0.90	3.32
Grade 10 Life Science	235,203	0.91	3.37

Table 8.B.17 Reliabilities and SEMs for the CSTs for Science by Primary Ethnicity (African American)

CST	N	Rel	SEM
Grade 5 Science	23,559	0.91	3.41
Grade 8 Science	24,692	0.91	3.37
Grade 10 Life Science	25,447	0.91	3.41

Table 8.B.18 Reliabilities and SEMs for the CSTs for Science by Primary Ethnicity (White)

CST	N	Rel	SEM
Grade 5 Science	105,907	0.90	2.97
Grade 8 Science	106,882	0.91	2.92
Grade 10 Life Science	112,169	0.92	3.02

Table 8.B.19 Reliabilities and SEMs for the CSTs for Science by Primary Ethnicity-for-Not Economically Disadvantaged (American Indian)

CST	N	Rel	SEM
Grade 5 Science	730	0.91	3.09
Grade 8 Science	872	0.91	3.09
Grade 10 Life Science	1,054	0.92	3.16

Table 8.B.20 Reliabilities and SEMs for the CSTs for Science by Primary Ethnicity-for-Not Economically Disadvantaged (Asian)

CST	N	Rel	SEM
Grade 5 Science	26,413	0.89	2.60
Grade 8 Science	25,567	0.89	2.47
Grade 10 Life Science	26,742	0.91	2.71

Table 8.B.21 Reliabilities and SEMs for the CSTs for Science by Primary Ethnicity-for-Not Economically Disadvantaged (Pacific Islander)

CST	N	Rel	SEM
Grade 5 Science	695	0.90	3.15
Grade 8 Science	804	0.90	3.11
Grade 10 Life Science	900	0.91	3.22

Table 8.B.22 Reliabilities and SEMs for the CSTs for Science by Primary Ethnicity-for-Not Economically Disadvantaged (Filipino)

CST	N	Rel	SEM
Grade 5 Science	6,991	0.88	2.92
Grade 8 Science	7,929	0.88	2.86
Grade 10 Life Science	9,002	0.89	3.01

Table 8.B.23 Reliabilities and SEMs for the CSTs for Science by Primary Ethnicity-for-Not Economically Disadvantaged (Hispanic)

CST	N	Rel	SEM
Grade 5 Science	41,846	0.90	3.16
Grade 8 Science	43,268	0.90	3.12
Grade 10 Life Science	50,952	0.91	3.23

Table 8.B.24 Reliabilities and SEMs for the CSTs for Science by Primary Ethnicity-for-Not Economically Disadvantaged (African American)

CST	N	Rel	SEM
Grade 5 Science	5,548	0.91	3.22
Grade 8 Science	6,737	0.91	3.22
Grade 10 Life Science	7,981	0.91	3.31

Table 8.B.25 Reliabilities and SEMs for the CSTs for Science by Primary Ethnicity-for-Not Economically Disadvantaged (White)

CST	N	Rel	SEM
Grade 5 Science	75,049	0.88	2.85
Grade 8 Science	77,261	0.89	2.81
Grade 10 Life Science	83,462	0.91	2.94

Table 8.B.26 Reliabilities and SEMs for the CSTs for Science by Primary Ethnicity-for-Economically Disadvantaged (American Indian)

CST	N	Rel	SEM
Grade 5 Science	1,524	0.90	3.43
Grade 8 Science	1,553	0.90	3.38
Grade 10 Life Science	1,504	0.90	3.40

Table 8.B.27 Reliabilities and SEMs for the CSTs for Science by Primary Ethnicity-for-Economically Disadvantaged (Asian)

CST	N	Rel	SEM
Grade 5 Science	15,124	0.92	3.14
Grade 8 Science	15,424	0.91	2.97
Grade 10 Life Science	17,200	0.92	3.13

Table 8.B.28 Reliabilities and SEMs for the CSTs for Science by Primary Ethnicity-for-Economically Disadvantaged (Pacific Islander)

CST	N	Rel	SEM
Grade 5 Science	1,505	0.89	3.44
Grade 8 Science	1,452	0.90	3.36
Grade 10 Life Science	1,383	0.90	3.44

Table 8.B.29 Reliabilities and SEMs for the CSTs for Science by Primary Ethnicity-for-Economically Disadvantaged (Filipino)

CST	N	Rel	SEM
Grade 5 Science	4,032	0.89	3.20
Grade 8 Science	4,352	0.89	3.09
Grade 10 Life Science	4,575	0.90	3.22

Table 8.B.30 Reliabilities and SEMs for the CSTs for Science by Primary Ethnicity-for-Economically Disadvantaged (Hispanic)

CST	N	Rel	SEM
Grade 5 Science	197,446	0.90	3.44
Grade 8 Science	186,382	0.90	3.36
Grade 10 Life Science	184,251	0.90	3.41

Table 8.B.31 Reliabilities and SEMs for the CSTs for Science by Primary Ethnicity-for-Economically Disadvantaged (African American)

CST	N	Rel	SEM
Grade 5 Science	18,011	0.90	3.47
Grade 8 Science	17,955	0.90	3.43
Grade 10 Life Science	17,466	0.90	3.45

Table 8.B.32 Reliabilities and SEMs for the CSTs for Science by Primary Ethnicity-for-Economically Disadvantaged (White)

CST	N	Rel	SEM
Grade 5 Science	30,858	0.90	3.24
Grade 8 Science	29,621	0.91	3.18
Grade 10 Life Science	28,707	0.92	3.25

Table 8.B.33 Subscore Reliabilities and SEM for CSTs for Science by Gender/Economic Status

	Subscore Area	N of Items	Male		Female		Not Econ. Dis.		Econ. Dis.	
			Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM
	Grade 5 Science									
1.	Physical Science (Grade 5)	11	0.70	1.32	0.67	1.33	0.63	1.15	0.65	1.42
2.	Physical Science (Grade 4)	8	0.66	1.17	0.62	1.21	0.60	1.10	0.60	1.24
3.	Life Science (Grade 5)	13	0.72	1.51	0.69	1.52	0.65	1.38	0.68	1.59
4.	Life Science (Grade 4)	9	0.68	1.19	0.66	1.21	0.61	1.06	0.63	1.28
5.	Earth Science (Grade 5)	11	0.67	1.37	0.62	1.39	0.58	1.27	0.61	1.44
6.	Earth Science (Grade 4)	8	0.66	1.20	0.62	1.23	0.61	1.11	0.58	1.28
	Grade 8 Science									
1.	Motion	8	0.55	1.20	0.49	1.22	0.52	1.13	0.46	1.26
2.	Forces, Density, and Buoyancy	13	0.75	1.41	0.69	1.47	0.69	1.30	0.69	1.53
3.	Structure of Matter and Periodic Table	16	0.80	1.60	0.76	1.61	0.76	1.46	0.75	1.70
4.	Earth in the Solar System	7	0.58	1.03	0.50	1.08	0.52	0.97	0.51	1.10
5.	Reactions and the Chemistry of Living Systems	10	0.70	1.31	0.62	1.35	0.64	1.24	0.62	1.39
6.	Investigation and Experimentation	6	0.66	0.91	0.62	0.90	0.61	0.77	0.61	0.98

Subscore Area		N of Items	Male		Female		Not Econ. Dis.		Econ. Dis.	
			Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM
Grade 10 Life Science										
1.	Cell Biology	10	0.67	1.34	0.62	1.37	0.66	1.31	0.57	1.39
2.	Genetics	12	0.75	1.47	0.72	1.49	0.74	1.41	0.69	1.53
3.	Physiology	10	0.69	1.30	0.62	1.34	0.64	1.23	0.62	1.39
4.	Ecology	11	0.74	1.29	0.69	1.34	0.68	1.21	0.69	1.39
5.	Evolution	11	0.72	1.37	0.67	1.39	0.68	1.29	0.66	1.45
6.	Investigation and Experimentation	6	0.75	0.89	0.70	0.87	0.72	0.77	0.71	0.95

Table 8.B.34 Subscore Reliabilities and SEM for CSTs for Science by Special Services/English Fluency

Subscore Area	N of Items	No Spec. Serv.		Spec. Serv.		Eng. Only		Ini. Fluent		Learner		Red. Fluent	
		Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM
Grade 5 Science													
1. Physical Science (Grade 5)	11	0.67	1.32	0.72	1.44	0.66	1.27	0.62	1.10	0.60	1.52	0.57	1.30
2. Physical Science (Grade 4)	8	0.63	1.19	0.66	1.24	0.63	1.17	0.59	1.07	0.54	1.28	0.55	1.19
3. Life Science (Grade 5)	13	0.69	1.51	0.73	1.60	0.69	1.47	0.63	1.34	0.61	1.67	0.59	1.50
4. Life Science (Grade 4)	9	0.66	1.19	0.68	1.30	0.65	1.16	0.59	1.02	0.56	1.37	0.55	1.18
5. Earth Science (Grade 5)	11	0.63	1.37	0.69	1.46	0.63	1.34	0.56	1.24	0.57	1.51	0.47	1.36
6. Earth Science (Grade 4)	8	0.64	1.22	0.64	1.26	0.64	1.19	0.62	1.09	0.46	1.31	0.53	1.23
Grade 8 Science													
1. Motion	8	0.51	1.20	0.48	1.30	0.52	1.19	0.52	1.11	0.38	1.32	0.45	1.21
2. Forces, Density, and Buoyancy	13	0.71	1.43	0.72	1.60	0.72	1.40	0.69	1.29	0.64	1.65	0.64	1.46
3. Structure of Matter and Periodic Table	16	0.77	1.59	0.76	1.76	0.78	1.56	0.77	1.44	0.68	1.82	0.73	1.63
4. Earth in the Solar System	7	0.53	1.05	0.54	1.15	0.54	1.03	0.52	0.96	0.43	1.20	0.47	1.05
5. Reactions and the Chemistry of Living Systems	10	0.65	1.33	0.65	1.42	0.66	1.30	0.65	1.22	0.53	1.47	0.58	1.35
6. Investigation and Experimentation	6	0.63	0.89	0.59	1.08	0.65	0.86	0.61	0.76	0.47	1.11	0.56	0.91
Grade 10 Life Science													
1. Cell Biology	10	0.65	1.35	0.56	1.39	0.65	1.35	0.67	1.31	0.41	1.38	0.59	1.37
2. Genetics	12	0.73	1.47	0.64	1.54	0.75	1.45	0.74	1.41	0.52	1.56	0.69	1.51
3. Physiology	10	0.65	1.31	0.63	1.44	0.66	1.28	0.62	1.23	0.49	1.49	0.57	1.35
4. Ecology	11	0.70	1.30	0.71	1.46	0.71	1.27	0.66	1.22	0.59	1.52	0.64	1.34
5. Evolution	11	0.69	1.37	0.63	1.51	0.70	1.34	0.67	1.29	0.48	1.55	0.63	1.40
6. Investigation and Experimentation	6	0.72	0.86	0.68	1.04	0.74	0.83	0.70	0.77	0.58	1.09	0.66	0.89

Table 8.B.35 Subscore Reliabilities and SEM for CSTs for Science by Primary Ethnicity

Subscore Area	N of Items	Am. Ind.		Asian		Pac. Island		Filipino		Hispanic		African Am.		White	
		Rel.	SEM	Rel.	SEM	Rel.	SEM	Rel.	SEM	Rel.	SEM	Rel.	SEM	Rel.	SEM
Grade 5 Science															
1. Physical Science (Grade 5)	11	0.67	1.38	0.70	1.08	0.65	1.38	0.62	1.19	0.65	1.42	0.65	1.43	0.63	1.19
2. Physical Science (Grade 4)	8	0.62	1.22	0.64	1.06	0.60	1.23	0.59	1.13	0.60	1.24	0.61	1.25	0.60	1.12
3. Life Science (Grade 5)	13	0.69	1.57	0.70	1.34	0.66	1.58	0.62	1.43	0.67	1.59	0.69	1.60	0.65	1.40
4. Life Science (Grade 4)	9	0.64	1.25	0.67	1.03	0.62	1.27	0.60	1.12	0.63	1.27	0.65	1.30	0.61	1.08
5. Earth Science (Grade 5)	11	0.64	1.42	0.63	1.25	0.61	1.43	0.55	1.32	0.61	1.44	0.66	1.44	0.59	1.29
6. Earth Science (Grade 4)	8	0.63	1.24	0.66	1.07	0.59	1.25	0.58	1.15	0.58	1.28	0.60	1.27	0.61	1.14
Grade 8 Science															
1. Motion	8	0.48	1.26	0.55	1.04	0.48	1.23	0.47	1.15	0.46	1.26	0.45	1.28	0.50	1.15
2. Forces, Density, and Buoyancy	13	0.71	1.51	0.70	1.21	0.69	1.51	0.64	1.34	0.69	1.53	0.70	1.56	0.69	1.32
3. Structure of Matter and Periodic Table	16	0.77	1.67	0.78	1.35	0.74	1.67	0.72	1.49	0.75	1.70	0.76	1.71	0.76	1.48
4. Earth in the Solar System	7	0.53	1.10	0.55	0.93	0.49	1.11	0.48	0.99	0.50	1.10	0.51	1.13	0.52	0.98
5. Reactions and the Chemistry of Living Systems	10	0.65	1.36	0.65	1.17	0.64	1.38	0.58	1.27	0.62	1.39	0.64	1.39	0.63	1.25
6. Investigation and Experimentation	6	0.64	0.96	0.60	0.68	0.62	0.94	0.57	0.79	0.60	0.98	0.62	1.00	0.62	0.79
Grade 10 Life Science															
1. Cell Biology	10	0.60	1.39	0.70	1.25	0.58	1.40	0.62	1.34	0.56	1.39	0.56	1.39	0.64	1.32
2. Genetics	12	0.72	1.51	0.75	1.35	0.69	1.53	0.70	1.44	0.69	1.53	0.69	1.54	0.74	1.41
3. Physiology	10	0.66	1.35	0.65	1.18	0.61	1.38	0.58	1.25	0.61	1.39	0.62	1.40	0.65	1.22
4. Ecology	11	0.71	1.34	0.69	1.16	0.69	1.37	0.62	1.24	0.69	1.39	0.71	1.41	0.69	1.20
5. Evolution	11	0.68	1.42	0.69	1.24	0.67	1.44	0.62	1.31	0.65	1.45	0.67	1.46	0.68	1.29
6. Investigation and Experimentation	6	0.72	0.92	0.69	0.71	0.72	0.94	0.63	0.80	0.70	0.95	0.72	0.98	0.73	0.76

Table 8.B.36 Subscore Reliabilities and SEM for CSTs for Science by Primary Ethnicity-for-Not Economically Disadvantaged

Subscore Area	N of Items	Am. Ind.		Asian		Pac. Island		Filipino		Hispanic		African Am.		White	
		Rel.	SEM	Rel.	SEM	Rel.	SEM	Rel.	SEM	Rel.	SEM	Rel.	SEM	Rel.	SEM
Grade 5 Science															
1. Physical Science (Grade 5)	11	0.68	1.24	0.64	0.97	0.62	1.27	0.59	1.13	0.63	1.29	0.64	1.32	0.57	1.12
2. Physical Science (Grade 4)	8	0.63	1.15	0.58	0.99	0.61	1.18	0.58	1.09	0.60	1.18	0.62	1.20	0.56	1.08
3. Life Science (Grade 5)	13	0.69	1.46	0.64	1.25	0.63	1.49	0.59	1.39	0.66	1.48	0.69	1.51	0.60	1.35
4. Life Science (Grade 4)	9	0.64	1.13	0.60	0.93	0.64	1.17	0.57	1.08	0.63	1.16	0.65	1.20	0.55	1.03
5. Earth Science (Grade 5)	11	0.61	1.33	0.56	1.18	0.57	1.35	0.52	1.29	0.58	1.35	0.64	1.36	0.53	1.24
6. Earth Science (Grade 4)	8	0.63	1.17	0.60	0.99	0.57	1.19	0.56	1.10	0.60	1.20	0.62	1.22	0.56	1.10
Grade 8 Science															
1. Motion	8	0.46	1.21	0.50	0.97	0.48	1.18	0.46	1.12	0.48	1.21	0.47	1.24	0.48	1.11
2. Forces, Density, and Buoyancy	13	0.70	1.42	0.65	1.11	0.66	1.44	0.62	1.30	0.69	1.43	0.71	1.49	0.66	1.26
3. Structure of Matter and Periodic Table	16	0.77	1.57	0.74	1.24	0.75	1.58	0.72	1.44	0.75	1.59	0.77	1.63	0.73	1.43
4. Earth in the Solar System	7	0.47	1.05	0.50	0.88	0.47	1.06	0.47	0.96	0.51	1.04	0.52	1.08	0.49	0.95
5. Reactions and the Chemistry of Living Systems	10	0.65	1.30	0.60	1.09	0.63	1.32	0.57	1.24	0.62	1.32	0.66	1.34	0.60	1.22
6. Investigation and Experimentation	6	0.66	0.87	0.53	0.59	0.63	0.86	0.55	0.75	0.60	0.89	0.64	0.92	0.58	0.74
Grade 10 Life Science															
1. Cell Biology	10	0.62	1.36	0.67	1.19	0.59	1.37	0.62	1.32	0.60	1.37	0.59	1.38	0.63	1.30
2. Genetics	12	0.73	1.46	0.73	1.28	0.72	1.48	0.70	1.42	0.72	1.49	0.72	1.51	0.73	1.38
3. Physiology	10	0.65	1.27	0.61	1.10	0.57	1.33	0.57	1.22	0.63	1.32	0.63	1.36	0.62	1.19
4. Ecology	11	0.67	1.26	0.62	1.09	0.67	1.30	0.58	1.21	0.69	1.31	0.70	1.35	0.65	1.16
5. Evolution	11	0.69	1.35	0.64	1.17	0.66	1.38	0.60	1.28	0.67	1.38	0.68	1.41	0.66	1.25
6. Investigation and Experimentation	6	0.72	0.84	0.64	0.63	0.71	0.86	0.61	0.75	0.71	0.87	0.73	0.92	0.71	0.72

Table 8.B.37 Subscore Reliabilities and SEM for CSTs for Science by Primary Ethnicity-for-Economically Disadvantaged

Subscore Area	N of Items	Am. Ind.		Asian		Pac. Island		Filipino		Hispanic		African Am.		White	
		Rel.	SEM	Rel.	SEM	Rel.	SEM	Rel.	SEM	Rel.	SEM	Rel.	SEM	Rel.	SEM
Grade 5 Science															
1. Physical Science (Grade 5)	11	0.63	1.45	0.70	1.26	0.63	1.43	0.63	1.28	0.64	1.45	0.63	1.46	0.64	1.33
2. Physical Science (Grade 4)	8	0.58	1.25	0.63	1.18	0.57	1.26	0.58	1.19	0.59	1.25	0.59	1.26	0.60	1.20
3. Life Science (Grade 5)	13	0.66	1.62	0.71	1.48	0.65	1.61	0.63	1.51	0.66	1.61	0.67	1.63	0.67	1.52
4. Life Science (Grade 4)	9	0.61	1.30	0.66	1.18	0.59	1.31	0.61	1.20	0.62	1.30	0.62	1.33	0.62	1.21
5. Earth Science (Grade 5)	11	0.62	1.46	0.64	1.35	0.59	1.47	0.56	1.38	0.60	1.46	0.64	1.46	0.62	1.38
6. Earth Science (Grade 4)	8	0.59	1.27	0.64	1.19	0.55	1.28	0.57	1.21	0.55	1.29	0.57	1.29	0.60	1.23
Grade 8 Science															
1. Motion	8	0.46	1.28	0.52	1.15	0.45	1.25	0.47	1.19	0.44	1.27	0.43	1.29	0.48	1.23
2. Forces, Density, and Buoyancy	13	0.69	1.56	0.71	1.36	0.69	1.55	0.65	1.41	0.68	1.55	0.68	1.59	0.70	1.45
3. Structure of Matter and Periodic Table	16	0.74	1.72	0.78	1.51	0.72	1.72	0.71	1.57	0.74	1.72	0.74	1.74	0.75	1.62
4. Earth in the Solar System	7	0.53	1.13	0.55	1.01	0.47	1.14	0.47	1.03	0.49	1.12	0.49	1.15	0.52	1.06
5. Reactions and the Chemistry of Living Systems	10	0.62	1.40	0.64	1.28	0.63	1.40	0.58	1.32	0.60	1.41	0.62	1.41	0.63	1.34
6. Investigation and Experimentation	6	0.60	1.01	0.61	0.81	0.59	0.99	0.58	0.86	0.59	1.00	0.60	1.03	0.62	0.91
Grade 10 Life Science															
1. Cell Biology	10	0.53	1.41	0.68	1.33	0.52	1.41	0.60	1.37	0.53	1.39	0.53	1.40	0.61	1.38
2. Genetics	12	0.68	1.53	0.74	1.45	0.65	1.56	0.69	1.49	0.67	1.54	0.66	1.55	0.72	1.49
3. Physiology	10	0.62	1.40	0.64	1.29	0.60	1.41	0.58	1.32	0.59	1.41	0.60	1.42	0.66	1.32
4. Ecology	11	0.70	1.39	0.71	1.27	0.67	1.42	0.64	1.30	0.68	1.41	0.69	1.43	0.71	1.31
5. Evolution	11	0.64	1.46	0.69	1.35	0.64	1.47	0.63	1.38	0.64	1.46	0.65	1.49	0.68	1.39
6. Investigation and Experimentation	6	0.71	0.97	0.69	0.82	0.70	0.98	0.65	0.87	0.69	0.97	0.71	1.00	0.73	0.87

Table 8.B.38 Reliability of Classification for CSTs for Science, Grade Five

	Placement Score	Far Below Basic	Below Basic	Basic	Proficient	Advanced	Category Total
Decision Accuracy	0 – 22	0.06	0.02	0.00	0.00	0.00	0.08
	23 – 29	0.01	0.06	0.03	0.00	0.00	0.10
	30 – 41	0.00	0.02	0.22	0.04	0.00	0.28
	42 – 51	0.00	0.00	0.04	0.26	0.04	0.34
All-forms Average	52 – 60	0.00	0.00	0.00	0.04	0.16	0.20
	Estimated Proportion Correctly Classified: Total = 0.75, Proficient & Above = 0.92						
Decision Consistency	0 – 22	0.05	0.02	0.00	0.00	0.00	0.08
	23 – 29	0.02	0.05	0.03	0.00	0.00	0.10
	30 – 41	0.00	0.03	0.19	0.06	0.00	0.28
	42 – 51	0.00	0.00	0.06	0.22	0.06	0.34
Alternate Form	52 – 60	0.00	0.00	0.00	0.05	0.15	0.20
	Estimated Proportion Consistently Classified: Total = 0.66, Proficient & Above = 0.88						

Table 8.B.39 Reliability of Classification for CSTs for Science, Grade Eight

	Placement Score	Far Below Basic	Below Basic	Basic	Proficient	Advanced	Category Total
Decision Accuracy	0 – 24	0.07	0.02	0.00	0.00	0.00	0.09
	25 – 31	0.01	0.06	0.03	0.00	0.00	0.11
	32 – 39	0.00	0.02	0.13	0.04	0.00	0.20
All-forms Average	40 – 46	0.00	0.00	0.04	0.15	0.04	0.23
	47 – 60	0.00	0.00	0.00	0.04	0.34	0.38
	Estimated Proportion Correctly Classified: Total = 0.74, Proficient & Above = 0.92						
Decision Consistency	0 – 24	0.06	0.02	0.00	0.00	0.00	0.09
	25 – 31	0.02	0.05	0.03	0.00	0.00	0.11
	32 – 39	0.00	0.03	0.10	0.05	0.00	0.20
Alternate Form	40 – 46	0.00	0.00	0.05	0.12	0.06	0.23
	47 – 60	0.00	0.00	0.00	0.05	0.32	0.38
	Estimated Proportion Consistently Classified: Total = 0.66, Proficient & Above = 0.89						

Table 8.B.40 Reliability of Classification for CSTs for Life Science (Grade 10)

	Placement Score	Far Below Basic	Below Basic	Basic	Proficient	Advanced	Category Total
Decision Accuracy	0 – 20	0.06	0.02	0.00	0.00	0.00	0.09
	21 – 27	0.01	0.07	0.03	0.00	0.00	0.12
	28 – 39	0.00	0.03	0.23	0.04	0.00	0.30
All-forms Average	40 – 48	0.00	0.00	0.04	0.19	0.03	0.27
	49 – 60	0.00	0.00	0.00	0.04	0.20	0.23
	Estimated Proportion Correctly Classified: Total = 0.75, Proficient & Above = 0.92						
Decision Consistency	0 – 20	0.05	0.03	0.01	0.00	0.00	0.09
	21 – 27	0.02	0.06	0.04	0.00	0.00	0.12
	28 – 39	0.00	0.04	0.20	0.05	0.00	0.30
Alternate Form	40 – 48	0.00	0.00	0.06	0.16	0.05	0.27
	49 – 60	0.00	0.00	0.00	0.05	0.19	0.23
	Estimated Proportion Consistently Classified: Total = 0.66, Proficient & Above = 0.88						

Appendix 8.C—IRT Analysis

Table 8.C.1 Conversion for the CST for Science, Grade Five (paper-pencil)

Raw Scr.	Freq. Distrib.	Theta	Scale Score	Rprtd Score	Raw Scr.	Freq. Distrib.	Theta	Scale Score	Rprtd Score
0	-	N/A	47.3185	150	**31**	8,394	–0.5500	303.9477	304
1	-	–4.8587	68.8887	150	**32**	8,930	–0.4776	308.1231	308
2	-	–4.1432	93.5010	150	**33**	9,201	–0.4050	312.3101	312
3	-	–3.7149	121.3956	150	**34**	9,834	–0.3319	316.5286	317
4	-	–3.4040	138.3815	150	**35**	10,048	–0.2581	320.7858	321
5	14	–3.1573	153.0218	153	**36**	10,661	–0.1836	325.0823	325
6	15	–2.9511	165.3858	165	**37**	11,107	–0.1080	329.4426	329
7	38	–2.7727	175.5296	176	**38**	11,447	–0.0313	333.8646	334
8	87	–2.6146	184.6489	185	**39**	11,955	0.0468	338.3698	338
9	192	–2.4718	192.9998	193	**40**	12,310	0.1266	342.9730	343
10	314	–2.3411	200.6301	201	**41**	12,772	0.2082	347.6840	348
11	534	–2.2201	207.5338	208	**42**	13,284	0.2921	352.5196	353
12	871	–2.1069	214.0656	214	**43**	13,777	0.3785	357.5069	358
13	1,271	–2.0003	220.2549	220	**44**	14,079	0.4679	362.6578	363
14	1,751	–1.8992	226.1396	226	**45**	14,573	0.5607	368.0084	368
15	2,265	–1.8028	231.6609	232	**46**	14,882	0.6575	373.6000	374
16	2,820	–1.7103	236.9874	237	**47**	15,142	0.7590	379.4614	379
17	3,229	–1.6212	242.1349	242	**48**	15,671	0.8660	385.6314	386
18	3,703	–1.5350	247.1212	247	**49**	15,837	0.9796	392.1674	392
19	3,955	–1.4514	251.9648	252	**50**	15,871	1.1011	399.1943	399
20	4,422	–1.3699	256.6522	257	**51**	15,658	1.2324	406.7809	407
21	4,768	–1.2904	261.2308	261	**52**	15,593	1.3757	415.0352	415
22	5,142	–1.2124	265.7252	266	**53**	14,856	1.5344	424.1931	424
23	5,528	–1.1358	270.1434	270	**54**	14,063	1.7134	434.5784	435
24	5,846	–1.0604	274.4968	274	**55**	12,687	1.9203	446.4795	446
25	6,226	–0.9860	278.7972	279	**56**	10,967	2.1678	460.8485	461
26	6,484	–0.9123	283.0468	283	**57**	8,626	2.4795	478.9971	479
27	7,068	–0.8393	287.2672	287	**58**	6,061	2.9086	503.6774	504
28	7,252	–0.7666	291.4553	291	**59**	3,458	3.6249	548.1332	548
29	7,671	–0.6943	295.6267	296	**60**	1,178	N/A	624.0193	600
30	8,084	–0.6222	299.7892	300					

Note: Performance-level cut scores are highlighted. To protect student privacy, the frequency distribution is not shown if based on 10 or fewer student records.

Table 8.C.2 Conversion for the CST for Science, Grade Five (Braille)

Raw Scr.	Freq. Distrib.	Theta	Scale Score	Rprtd Score	Raw Scr.	Freq. Distrib.	Theta	Scale Score	Rprtd Score
0	–	N/A	47.3185	150	**31**	–	–0.5219	305.5647	306
1	–	–4.8502	69.0684	150	**32**	–	–0.4482	309.8189	310
2	–	–4.1344	93.9671	150	**33**	–	–0.3740	314.0934	314
3	–	–3.7058	121.8793	150	**34**	–	–0.2993	318.4060	318
4	–	–3.3946	138.9711	150	**35**	–	–0.2238	322.7605	323
5	–	–3.1475	153.5721	154	**36**	–	–0.1474	327.1691	327
6	–	–2.9409	165.9245	166	**37**	–	–0.0698	331.6434	332
7	–	–2.7621	176.1753	176	**38**	–	0.0091	336.1930	336
8	–	–2.6036	185.2831	185	**39**	–	0.0896	340.8392	341
9	–	–2.4605	193.6342	194	**40**	–	0.1720	345.5926	346
10	–	–2.3293	201.3257	201	**41**	–	0.2566	350.4706	350
11	–	–2.2078	208.2595	208	**42**	–	0.3438	355.5015	356
12	–	–2.0942	214.7996	215	**43**	–	0.4338	360.7007	361
13	–	–1.9871	221.0029	221	**44**	–	0.5273	366.0905	366
14	–	–1.8855	226.9073	227	**45**	–	0.6248	371.7050	372
15	–	–1.7885	232.5048	233	**46**	–	0.7269	377.5967	378
16	–	–1.6954	237.8518	238	**47**	–	0.8346	383.8201	384
17	–	–1.6058	243.0232	243	**48**	–	0.9488	390.4088	390
18	–	–1.5190	248.0372	248	**49**	–	1.0709	397.4329	397
19	–	–1.4347	252.9115	253	**50**	–	1.2027	405.0581	405
20	–	–1.3526	257.6614	258	**51**	–	1.3466	413.3780	413
21	–	–1.2723	262.2888	262	**52**	–	1.5059	422.5325	423
22	–	–1.1935	266.8224	267	**53**	–	1.6855	432.9511	433
23	–	–1.1162	271.2810	271	**54**	–	1.8929	444.9411	445
24	–	–1.0399	275.6809	276	**55**	–	2.1410	459.2177	459
25	–	–0.9646	280.0271	280	**56**	–	2.4533	477.5065	478
26	–	–0.8900	284.3325	284	**57**	–	2.8830	501.9452	502
27	–	–0.8159	288.6075	289	**58**	–	3.5998	546.6354	547
28	–	–0.7422	292.8570	293	**59**	–	N/A	624.0193	600
29	–	–0.6687	297.0959	297					
30	–	–0.5954	301.3278	301					

Note: Performance-level cut scores are highlighted. To protect student privacy, the frequency distribution is not shown if based on 10 or fewer student records.

Table 8.C.3 Conversion for the CST for Science, Grade Eight (paper-pencil)

Raw Scr.	Freq. Distrib.	Theta	Scale Score	Rprtd Score	Raw Scr.	Freq. Distrib.	Theta	Scale Score	Rprtd Score
0	–	N/A	–96.6361	150	**31**	8,087	–0.3173	296.1385	296
1	–	–4.7084	–55.7441	150	**32**	8,565	–0.2417	302.4793	302
2	–	–3.9906	–15.7140	150	**33**	9,098	–0.1657	308.8467	309
3	–	–3.5600	21.7100	150	**34**	9,702	–0.0893	315.2603	315
4	–	–3.2468	48.7047	150	**35**	10,344	–0.0122	321.7304	322
5	–	–2.9977	70.9407	150	**36**	10,884	0.0657	328.2713	328
6	14	–2.7891	88.2331	150	**37**	11,268	0.1447	334.8953	335
7	33	–2.6083	103.4604	150	**38**	12,022	0.2249	341.6235	342
8	101	–2.4477	117.2834	150	**39**	12,679	0.3065	348.4704	348
9	196	–2.3024	129.4457	150	**40**	13,011	0.3898	355.4531	355
10	299	–2.1692	140.5642	150	**41**	13,448	0.4751	362.6077	363
11	501	–2.0456	151.0311	151	**42**	14,012	0.5626	369.9504	370
12	792	–1.9298	160.8902	161	**43**	14,248	0.6527	377.5089	378
13	1,088	–1.8206	169.9466	170	**44**	14,837	0.7458	385.3241	385
14	1,493	–1.7168	178.6441	179	**45**	14,970	0.8424	393.4246	393
15	1,916	–1.6176	186.9994	187	**46**	15,448	0.9430	401.8634	402
16	2,275	–1.5224	195.0393	195	**47**	15,272	1.0484	410.7026	411
17	2,670	–1.4305	202.7327	203	**48**	15,368	1.1593	420.0282	420
18	3,005	–1.3415	210.1749	210	**49**	15,179	1.2769	429.8981	430
19	3,324	–1.2551	217.4281	217	**50**	15,104	1.4025	440.4052	440
20	3,524	–1.1708	224.5127	225	**51**	14,897	1.5379	451.8007	452
21	3,835	–1.0883	231.4495	231	**52**	14,242	1.6855	464.2053	464
22	4,242	–1.0074	238.2463	238	**53**	13,696	1.8485	477.8274	478
23	4,606	–0.9279	244.9157	245	**54**	12,906	2.0321	493.3250	493
24	4,897	–0.8495	251.4847	251	**55**	11,833	2.2436	511.0988	511
25	5,082	–0.7720	257.9826	258	**56**	10,736	2.4957	532.2446	532
26	5,541	–0.6953	264.4170	264	**57**	9,324	2.8122	559.1550	559
27	6,112	–0.6192	270.8018	271	**58**	7,591	3.2466	595.2122	595
28	6,519	–0.5435	277.1555	277	**59**	5,404	3.9683	660.5382	600
29	6,826	–0.4680	283.4902	283	**60**	2,532	N/A	742.1487	600
30	7,379	–0.3927	289.8127	290					

Note: Performance-level cut scores are highlighted. To protect student privacy, the frequency distribution is not shown if based on 10 or fewer student records.

Table 8.C.4 Conversion for the CST for Science, Grade Eight (Braille)

Raw Scr.	Freq. Distrib.	Theta	Scale Score	Rprtd Score	Raw Scr.	Freq. Distrib.	Theta	Scale Score	Rprtd Score
0	–	N/A	–96.6361	150	**31**	–	–0.2208	304.2270	304
1	–	–4.6542	–53.5008	150	**32**	–	–0.1423	310.8140	311
2	–	–3.9357	–11.7381	150	**33**	–	–0.0632	317.4480	317
3	–	–3.5044	27.7292	150	**34**	–	0.0166	324.1483	324
4	–	–3.1904	53.5572	150	**35**	–	0.0973	330.9201	331
5	–	–2.9405	75.3854	150	**36**	–	0.1792	337.7924	338
6	–	–2.7310	93.6114	150	**37**	–	0.2625	344.7770	345
7	–	–2.5493	108.4891	150	**38**	–	0.3473	351.8965	352
8	–	–2.3878	122.1308	150	**39**	–	0.4341	359.1768	359
9	–	–2.2417	134.6004	150	**40**	–	0.5230	366.6407	367
10	–	–2.1075	145.8933	150	**41**	–	0.6145	374.3084	374
11	–	–1.9829	156.2765	156	**42**	–	0.7090	382.2238	382
12	–	–1.8661	166.0954	166	**43**	–	0.8069	390.4480	390
13	–	–1.7557	175.4069	175	**44**	–	0.9088	399.0025	399
14	–	–1.6508	184.2792	184	**45**	–	1.0154	407.9404	408
15	–	–1.5505	192.6510	193	**46**	–	1.1275	417.3329	417
16	–	–1.4541	200.7236	201	**47**	–	1.2462	427.3160	427
17	–	–1.3609	208.5402	209	**48**	–	1.3730	437.9572	438
18	–	–1.2706	216.1283	216	**49**	–	1.5094	449.3653	449
19	–	–1.1828	223.5144	224	**50**	–	1.6580	461.8962	462
20	–	–1.0970	230.7313	231	**51**	–	1.8221	475.6651	476
21	–	–1.0130	237.7851	238	**52**	–	2.0066	491.1367	491
22	–	–0.9305	244.6960	245	**53**	–	2.2191	509.0775	509
23	–	–0.8492	251.5061	252	**54**	–	2.4723	530.1378	530
24	–	–0.7691	258.2330	258	**55**	–	2.7897	557.2680	557
25	–	–0.6897	264.8899	265	**56**	–	3.2250	593.3916	593
26	–	–0.6110	271.4914	271	**57**	–	3.9476	658.7221	600
27	–	–0.5327	278.0593	278	**58**	–	N/A	742.1487	600
28	–	–0.4547	284.6047	285					
29	–	–0.3769	291.1370	291					
30	–	–0.2990	297.6753	298					

Note: Performance-level cut scores are highlighted. To protect student privacy, the frequency distribution is not shown if based on 10 or fewer student records.

Table 8.C.5 Conversion for the CST for Life Science (Grade 10)

Raw Scr.	Freq. Distrib.	Theta	Scale Score	Rprtd Score	Raw Scr.	Freq. Distrib.	Theta	Scale Score	Rprtd Score
0	–	N/A	44.7476	150	**31**	10,367	–0.2019	312.2895	312
1	–	–4.5955	73.6178	150	**32**	10,559	–0.1257	316.5395	317
2	–	–3.8780	104.1178	150	**33**	10,996	–0.0491	320.8067	321
3	–	–3.4477	130.2494	150	**34**	11,378	0.0280	325.1139	325
4	–	–3.1347	148.4820	150	**35**	11,919	0.1059	329.4566	329
5	14	–2.8859	162.4965	162	**36**	12,213	0.1847	333.8472	334
6	29	–2.6775	174.0023	174	**37**	12,335	0.2645	338.3001	338
7	39	–2.4968	184.1029	184	**38**	12,498	0.3456	342.8279	343
8	132	–2.3364	193.1127	193	**39**	12,779	0.4283	347.4348	347
9	293	–2.1913	201.2699	201	**40**	13,125	0.5126	352.1429	352
10	520	–2.0582	208.7498	209	**41**	13,295	0.5990	356.9591	357
11	949	–1.9346	215.6594	216	**42**	13,296	0.6877	361.9066	362
12	1,448	–1.8189	222.0822	222	**43**	13,512	0.7791	366.9972	367
13	2,080	–1.7097	228.1646	228	**44**	13,443	0.8735	372.2605	372
14	2,829	–1.6060	233.9504	234	**45**	13,476	0.9715	377.7190	378
15	3,552	–1.5068	239.4865	239	**46**	13,449	1.0736	383.4204	383
16	4,229	–1.4115	244.8089	245	**47**	13,430	1.1805	389.3915	389
17	4,970	–1.3196	249.9435	250	**48**	13,389	1.2930	395.6803	396
18	5,447	–1.2305	254.9152	255	**49**	12,819	1.4123	402.3417	402
19	5,747	–1.1440	259.7466	260	**50**	12,718	1.5396	409.4487	409
20	6,191	–1.0595	264.4630	264	**51**	12,519	1.6768	417.0954	417
21	6,445	–0.9768	269.0776	269	**52**	12,038	1.8261	425.4127	425
22	6,803	–0.8957	273.6006	274	**53**	11,176	1.9910	434.5658	435
23	7,128	–0.8160	278.0494	278	**54**	10,436	2.1765	444.9797	445
24	7,313	–0.7373	282.4364	282	**55**	9,385	2.3899	456.9842	457
25	7,796	–0.6595	286.7739	287	**56**	8,317	2.6440	471.3061	471
26	8,178	–0.5824	291.0700	291	**57**	6,505	2.9625	489.3082	489
27	8,555	–0.5059	295.3383	295	**58**	4,726	3.3986	514.0652	514
28	9,036	–0.4297	299.5833	300	**59**	2,733	4.1220	555.4196	555
29	9,529	–0.3538	303.8191	304	**60**	1,176	N/A	602.3473	600
30	9,834	–0.2779	308.0512	308					

Note: Performance-level cut scores are highlighted. To protect student privacy, the frequency distribution is not shown if based on 10 or fewer student records.

Chapter 9: Quality Control Procedures

Rigorous quality control procedures were implemented throughout the test development, administration, scoring, and reporting processes. As part of this effort, Educational Testing Service (ETS) maintains an Office of Testing Integrity (OTI) that resides in the ETS legal department. The OTI provides quality assurance services for all testing programs administered by ETS. In addition, the Office of Professional Standards Compliance at ETS publishes and maintains the *ETS Standards for Quality and Fairness,* which supports the OTI's goals and activities. The purposes of the *ETS Standards for Quality and Fairness* are to help ETS design, develop, and deliver technically sound, fair, and useful products and services; and to help the public and auditors evaluate those products and services.

In addition, each department at ETS that is involved in the testing cycle designs and implements an independent set of procedures to ensure the quality of its products. In the next sections, these procedures are described.

Quality Control of Item Development

The item development process for the California Standards Tests (CSTs) for Science prior to the 2012–13 administration is described in detail in Chapter 3, starting on page 29; there was no new item development in 2015–16 because the forms were reused. The next sections highlight elements of the process devoted specifically to the quality control of the items that were previously developed and reused during the 2015–16 CST for Science administration.

Item Specifications

ETS maintained item specifications for each CST for Science and developed an item utilization plan to guide the development of the items for each content area. Item writing emphasis was determined in consultation with the California Department of Education (CDE). Adherence to the specifications ensured the maintenance of quality and consistency in the item development process.

Item Writers

The items for each CST for Science were written by item writers with a thorough understanding of the California content standards. CST for Science item writers were current California educations with at least three years of teaching experience and a bachelor's degree or teaching credential in science.

The item writers were carefully screened and selected by senior ETS content staff and approved by the CDE. Only those with strong content and teaching backgrounds were invited to participate in an extensive training program for item writers.

Internal Contractor Reviews

Once items were written, ETS assessment specialists made sure that each item underwent an intensive internal review process. Every step of this process is designed to produce items that exceed industry standards for quality. For the CSTs for Science, it included three rounds of content reviews, two rounds of editorial reviews, an internal fairness review, and a high-level review and approval by a content-area director. A carefully designed and monitored workflow and detailed checklists helped to ensure that all items met the specifications for the process.

Content Review

ETS assessment specialists made sure that the test items and related materials complied with ETS's written guidelines for clarity, style, accuracy, and appropriateness, and with approved item specifications.

The artwork and graphics for the items were created during the internal content review period so assessment specialists could evaluate the correctness and appropriateness of the art early in the item development process. ETS selected visuals that were relevant to the item content and that were easily understood so students would not struggle to determine the purpose or meaning of the questions.

Editorial Review

Another step in the ETS internal review process involved a team of specially trained editors who checked questions for clarity, correctness of language, grade-level appropriateness of language, adherence to style guidelines, and conformity to acceptable item-writing practices. The editorial review also included rounds of copyediting and proofreading. ETS strives for error-free items beginning with the initial rounds of review.

Fairness Review

One of the final steps in the ETS internal review process is to have all items and stimuli reviewed for fairness. Only ETS staff members who had participated in the ETS Fairness Training, a rigorous internal training course, conducted this bias and sensitivity review. These staff members had been trained to identify and eliminate test questions that contained content that could be construed as offensive to, or biased against, members of specific ethnic, racial, or gender groups.

Assessment Director Review

As a final quality control step, the content area's assessment director or another senior-level content reviewer read each item before it was presented to the CDE.

Assessment Review Panel Review

The Assessment Review Panels (ARPs) were committees that advised the CDE and ETS on areas related to item development for the CSTs for Science. The ARPs were responsible for reviewing all newly developed items for alignment to the California content standards. The ARPs also reviewed the items for accuracy of content, clarity of phrasing, and quality. See page 32 in Chapter 3 for additional information on the function of ARPs within the item-review process.

Statewide Pupil Assessment Review Panel Review

The Statewide Pupil Assessment Review (SPAR) panel was responsible for reviewing and approving the achievement tests that were used statewide for the testing of students in California public schools in grades five, eight, and ten. The SPAR panel representatives ensured that the test items conformed to the requirements of *Education Code* Section 60602. If the SPAR panel rejected specific items, the items were replaced with other items. See page 35 in Chapter 3 for additional information on the function of the SPAR panel within the item-review process.

Data Review of Field-tested Items

ETS field-tested newly developed items to obtain statistical information about item performance. This information was used to evaluate items that were candidates for use in operational test forms. These items that were flagged after field-test and operational use were examined carefully at data review meetings, where content experts discussed items

that had poor statistics and did not meet the psychometric criteria for item quality. The CDE defined the criteria for acceptable or unacceptable item statistics. These criteria ensured that the item (1) had an appropriate level of difficulty for the target population; (2) discriminated well between students who differ in ability; and (3) conformed well to the statistical model underlying the measurement of the intended constructs. The results of analyses for differential item functioning (DIF) were used to make judgments about the appropriateness of items for various subgroups when the items were first used.

The ETS content experts made recommendations about whether to accept or reject each item for inclusion in the California item bank. The CDE content experts reviewed the recommendations and made the final decision on each item.

The field-test items that appeared in the CSTs for Science administered in 2015–16 were statistically reviewed in data review meetings the year they were originally administered. There was no data review of field-test items in 2015–16. See Table 8.4 on page 90 for the list of the original administrations of each test administered in 2015–16.

Quality Control of the Item Bank

After the data review, items were placed in the item bank along with their statistics and reviewers' evaluations of their quality. ETS then delivered the items to the CDE through the California electronic item bank. The item bank database is maintained by a staff of application systems programmers, led by the Item Bank Manager, at ETS. All processes are logged, all change requests—including item bank updates for item availability status—are tracked, and all output and California item bank deliveries are quality-controlled for accuracy.

Quality of the item bank and secure transfer of the California item bank to the CDE are very important. The ETS internal item bank database resides on a server within the ETS firewall; access to the SQL Server database is strictly controlled by means of system administration. The electronic item banking application includes a login/password system to authorize access to the database or designated portions of the database. In addition, only users authorized to access the specific database are able to use the item bank. Users are authorized by a designated administrator at the CDE and at ETS.

ETS has extensive experience in accurate and secure data transfer of many types, including CDs, secure remote hosting, secure Web access, and secure file transfer protocol (SFTP), which is the current method used to deliver the California electronic item bank to the CDE. In addition, all files posted on the SFTP site by the item bank staff are encrypted with a password.

The measures taken for ensuring the accuracy, confidentiality, and security of electronic files are as follows:

- Electronic forms of test content, documentation, and item banks are backed up electronically, with the backup media kept off site, to prevent loss from system breakdown or a natural disaster.
- The offsite backup files are kept in secure storage, with access limited to authorized personnel only.
- Advanced network security measures are used to prevent unauthorized electronic access to the item bank.

Quality Control of Test Form Development

The ETS Assessment Development group is committed to providing the highest quality product to the students of California and has in place a number of quality control (QC) checks to ensure that outcome. During the item development process, there were multiple senior reviews of items and passages, including one by the assessment director. Test forms certification was a formal quality control process established as a final checkpoint prior to printing. In it, content, editorial, and senior development staff reviewed test forms for accuracy and clueing issues.

ETS also included quality checks throughout preparation of the form planners. A form planner specifications document was developed by the test development team lead with input from ETS's item bank and statistics groups; this document was then reviewed by all team members who built forms at a training session specific to form planners before the form-building process started. After trained content team members signed off on a form planner, a representative from the internal QC group reviewed each file for accuracy against the specifications document. Assessment directors reviewed and signed off on form planners prior to processing.

As processes are refined and enhanced, ETS implements further QC checks as appropriate.

Quality Control of Test Materials

Collecting Test Materials

Once the tests are administered, local educational agencies (LEAs) return scorable and nonscorable materials within five working days after the last selected testing day of each test administration period. The freight-return kits provided to the LEAs contain color-coded labels identifying scorable and nonscorable materials and labels with bar-coded information identifying the school and district. The LEAs apply the appropriate labels and number the cartons prior to returning the materials to the processing center by means of their assigned carrier. The use of the color-coded labels streamlines the return process.

All scorable and nonscorable materials are delivered to the ETS scanning and scoring facilities in Ewing, New Jersey. ETS closely monitor the return of materials. The California Technical Assistance Center (CalTAC) at ETS monitors returns and notifies LEAs that do not return their materials in a timely manner. CalTAC contacts the LEA California Assessment of Student Performance and Progress (CAASPP) coordinators and works with them to facilitate the return of the test materials.

Processing Test Materials

Upon receipt of the test materials, ETS uses precise inventory and test processing systems, in addition to quality assurance procedures, to maintain an up-to-date accounting of all the testing materials within its facilities. The materials are removed carefully from the shipping cartons and examined for a number of conditions, including physical damage, shipping errors, and omissions. A visual inspection to compare the number of students recorded on the School and Grade Identification (SGID) sheets with the number of answer documents in the stack is also conducted.

ETS's image scanning process captures security information electronically and compares scorable material quantities reported on the SGIDs to actual documents scanned. LEAs are contacted by phone if there are any missing shipments or the quantity of materials returned appears to be less than expected.

Quality Control of Scanning

Before any CAASPP documents are scanned, ETS conducts a complete check of the scanning system. ETS creates test decks for every test and form. Each test deck consists of approximately 700 answer documents marked to cover response ranges, demographic data, blanks, double marks, and other responses. Fictitious students are created to verify that each marking possibility is processed correctly by the scanning program. The output file generated as a result of this activity is thoroughly checked against each answer document after every stage to verify that the scanner is capturing marks correctly. When the program output is confirmed to match the expected results, a scan program release form is signed and the scan program is placed in the production environment under configuration management.

The intensity levels of each scanner are constantly monitored for quality control purposes. Intensity diagnostics sheets are run before and during each batch to verify that the scanner is working properly. In the event that a scanner fails to properly pick up items on the diagnostic sheets, the scanner is recalibrated to work properly before being allowed to continue processing student documents.

Documents received in poor condition (torn, folded, or water-stained) that could not be fed through the high-speed scanners are either scanned using a flat-bed scanner or keyed into the system manually.

Quality Control of Image Editing

Prior to submitting any CAASPP operational documents through the image editing process, ETS creates a mock set of documents to test all of the errors listed in the edit specifications. The set of test documents is used to verify that each image of the document is saved so that an editor will be able to review the documents through an interactive interface. The edits are confirmed to show the appropriate error, the correct image to edit the item, and the appropriate problem and resolution text that instructs the editor on the actions that should be taken.

Once the set of mock test documents is created, the image edit system completes the following procedures:

1. Scan the set of test documents.
2. Verify that the images from the documents are saved correctly.
3. Verify that the appropriate problem and resolution text displays for each type of error.
4. Submit the post-edit program to assure that all errors have been corrected.

ETS checks the post file against expected results to ensure the appropriate corrections are made. The post file will have all keyed corrections and any defaults from the edit specifications.

Quality Control of Answer Document Processing and Scoring

Accountability of Answer Documents

In addition to the quality control checks carried out in scanning and image editing, the following manual quality checks are conducted to verify that the answer documents are correctly attributed to the students, schools, LEAs, and subgroups, and document counts are compared to the SGIDs.

Any discrepancies identified in the steps outlined above are followed up by ETS staff with the LEAs for resolution.

Processing of Answer Documents

Prior to processing operational answer documents and executing subsequent data processing programs, ETS conducts an end-to-end test. As part of this test, ETS prepares approximately 700 test cases covering all tests and many scenarios designed to exercise particular business rule logic. ETS marks answer documents for those 700 test cases. They are then scanned, scored, and aggregated. The results at various inspection points are checked by psychometricians and Data Quality Services staff. Additionally, a post-scan test file of approximately 50,000 records across the CAASPP System is scored and aggregated to test a broader range of scoring and aggregation scenarios. These procedures assure that students and LEAs receive the correct scores when the actual scoring process is carried out. In 2015–16, end-to-end testing also included the inspection of results in electronic reporting.

Scoring and Reporting Specifications

ETS develops standardized scoring procedures and specifications so testing materials are processed and scored accurately. These documents include the Scoring Rules specifications and the Include Indicators specifications. Each is explained in detail in Chapter 7, starting on page 64. The scoring specifications are reviewed and revised by the CDE and ETS each year. After a version that all parties endorse is finalized, the CDE issues a formal approval of the scoring and reporting specifications.

Storing Answer Documents

After the answer documents have been scanned, edited, and scored, and have cleared the clean-post process, they are palletized and placed in the secure storage facilities at ETS. The materials are stored until October 31 of each year, after which ETS requests permission to destroy the materials. After receiving CDE approval, the materials are destroyed in a secure manner.

Quality Control of Psychometric Processes

Score Key Verification Procedures

ETS takes various necessary measures to ascertain that the scoring keys are applied to the student responses as expected and the student scores are computed accurately. Scoring keys, provided in the form planners, are produced by ETS and verified thoroughly by performing multiple quality control checks. The form planners contain the information about an assembled test form; other information in the form planner includes the test name, administration year, subscore identification, and standards and statistics associated with each item. The quality control checks that are performed before keys are finalized are listed on page 65 in Chapter 7.

Quality Control of Item Analyses and the Equating Process

When the forms were first administered, the psychometric analyses conducted at ETS underwent comprehensive quality checks by a team of psychometricians and data analysts. Detailed checklists were consulted by members of the team for each of the statistical procedures performed on each CST for Science following its original administration. Quality assurance checks also included a comparison of the current year's statistics to statistics from previous years. The results of preliminary classical item analyses that provided a check on scoring keys were also reviewed by a senior psychometrician. The items that were

flagged for questionable statistical attributes were sent to test development staff for their review; their comments were reviewed by the psychometricians before items were approved to be included in the equating process.

The results of the equating process were reviewed by a psychometric manager in addition to the aforementioned team of psychometricians and data analysts. If the senior psychometrician and the manager reached a consensus that an equating result did not conform to the norm, special binders were prepared for review by senior psychometric advisors at ETS, along with several pieces of informative analyses to facilitate the process.

When the forms were equated following their original administration, a few additional checks were performed for the calibration, scaling, and scoring table creation processes, as described below.

Calibrations

During the calibration that was conducted for the original administration of each form and that is described in more detail in Chapter 2 starting on page 15, checks were made to ascertain that the correct options for the analyses were selected. Checks were also made on the number of items, number of students with valid scores, item response theory (IRT) Rasch item difficulty estimates, standard errors for the Rasch item difficulty estimates, and the match of selected statistics to the results on the same statistics obtained during preliminary item analyses. Psychometricians also performed detailed reviews of plots and statistics to investigate if the model fit the data.

Scaling

During the scaling that was conducted for the original administration of each form, checks were made to ensure the following:

- The correct items were used for linking;
- The scaling evaluation process, including stability analysis and subsequent removal of items from the linking set (if any), was implemented according to specification (see details in the “Evaluation of Scaling” section in Chapter 8 of the original year’s technical report); and
- The resulting scaling constants were correctly applied to transform the new item difficulty estimates onto the item bank scale.

Scoring Tables

Once the equating activities were complete and raw-score-to-scale score conversion tables were generated after the original administration of each content-area test, the psychometricians carried out quality control checks on each scoring table. Scoring tables were checked to verify the following:

- All raw scores were included in the tables;
- Scale scores increased as raw scores increased;
- The minimum reported scale score was 150 and the maximum reported scale score was 600; and
- The cut points for the performance levels were correctly identified.

As a check on the reasonableness of the performance levels, when the tests were originally administered, psychometricians compared results from the current year with results from the past year at the cut points and the percentage of students in each performance level within the equating samples. After all quality control steps were completed and any differences

were resolved, a senior psychometrician inspected the scoring tables as the final step in quality control.

During the current administration, the data derived from prior item analyses are used to pre-equate the 2015–16 results. Key checks and classical item analyses as well as associated quality assurance checks are also conducted on the current data.

In addition, the scoring tables are reused and are checked against the scoring tables in the reuse-year technical report to ensure exact match. In addition, prior to reporting in 2013, every regular and special-version multiple-choice test was certified by ETS prior to being included in electronic reporting. To certify a test, psychometricians gathered a certain number of test cases and verified the accurate application of scoring keys and conversion tables.

Score Verification Process

ETS utilizes the raw-to-scale scoring tables to assign scale scores for each student and verifies scale scores by independently generating the scale scores for students in a small number of LEAs and comparing these scores. The selection of LEAs is based on the availability of data for all schools included in those LEAs, known as “pilot LEAs.”

Year-to-Year Comparison Analyses

Year-to-year comparison analyses are conducted each year for quality control of the scoring procedure in general and as reasonableness checks for the CST for Science results.

- The first set of year-to-year comparison analyses looks at the tendencies and trends for the schools and LEAs for which ETS has received complete or near-complete results by mid-June.
- The second set of year-to-year comparison analyses uses over 90 percent of the entire testing population to look at the tendencies and trends for the state as a whole, as well as a few large LEAs.

The results of the year-to-year comparison analyses are provided to the CDE, and their reasonableness is jointly discussed. Any anomalies in the results are investigated further, and scores are released only after explanations that satisfy both the CDE and ETS are obtained.

Offloads to Test Development

During the original administration of the CST for Science forms that are reused in 2015–16, the statistics based on classical item analyses and the IRT analyses were obtained at two different times in the testing cycle. The first time, the statistics were obtained on the equating samples to ensure the quality of equating and then on larger sample sizes to ensure the stability of the statistics that were to be used for future test assembly. The resulting statistics for all items were provided to test development staff in specially designed Excel spreadsheets called “statistical offloads.” The offloads were thoroughly checked by the psychometric staff before their release for test development review.

During the 2015–16 administration, only statistics based on classical item analyses of the operational items and obtained on larger samples are included in the statistical offloads.

Quality Control of Reporting

For the quality control of various CAASPP student and summary reports, the following four general areas are evaluated:

1. Comparing report formats to input sources from the CDE-approved samples
2. Validating and verifying the report data by querying the appropriate student data
3. Evaluating the production print execution performance by comparing the number of report copies, sequence of report order, and offset characteristics to the CDE's requirements
4. Proofreading reports by the CDE and ETS prior to any LEA mailings

All reports are required to include a single, accurate county/district/school (CDS) code, a charter school number (if applicable), an LEA name, and a school name. All elements conform to the CDE's official CDS code and naming records. From the start of processing through scoring and reporting, the CDS Master File is used to verify and confirm accurate codes and names. The CDS Master File is provided by the CDE to ETS throughout the year as updates are available.

After the reports are validated against the CDE's requirements, a set of reports for pilot LEAs is provided to the CDE and ETS for review and approval. ETS prepares paper score reports on the actual report forms, foldered as they are expected to look in production. The CDE and ETS review and sign off on the report package after a thorough examination.

Upon the CDE's approval of the reports generated from the pilot LEAs, ETS proceeds with the first production batch test. The first production batch is selected to validate a subset of LEAs that contains examples of key reporting characteristics representative of the state as a whole. The first production batch test incorporates CDE-selected LEAs and provides the last check prior to generating all reports and providing them to the LEAs.

Electronic Reporting

Because results were pre-equated, students' scale scores and performance levels for CST for Science multiple-choice tests were made available to LEAs prior to the printing of paper reports. The reporting module in the Test Operations Management System made it possible for LEAs to securely download an electronic reporting file containing these results.

In addition, TOMS communicates with the Online Reporting System (ORS) that provides authorized users with interactive and cumulative online reports at the student, school, and LEA levels. The ORS provides access to score reports, which provide preliminary score data for each administered test available in the reporting system (CDE, 2016).

The ORS provides the preliminary summative reports containing information outlining student knowledge and skills, as well as performance levels. The online aggregate reports provide functionality at the student, classroom, school, and LEA levels.

Before an LEA can download a student data file, ETS statisticians approved a QC file of test results data and ETS IT successfully processed it. Once the data were deemed reliable and ETS processed a scorable answer document for every student who took a CST for Science in that test administration for the LEA, the LEA was notified that these results were available.

Excluding Student Scores from Summary Reports

ETS provides specifications to the CDE that document when to exclude student scores from summary reports. These specifications include the logic for handling answer documents that, for example, indicate the student tested but marked no answers, was absent, was not tested due to parent/guardian request, or did not complete the test due to illness. The methods for handling other anomalies are also covered in the specifications.

Reference

California Department of Education. (2016). *CAASPP Online Reporting System User Guide for California.* Sacramento, CA. Retrieved from http://www.caaspp.org/rsc/pdfs/CAASPP.ORS-guide.2016.pdf

Educational Testing Service. (2002). *ETS standards for quality and fairness*. Princeton, NJ: Author.

Chapter 10: Historical Comparisons

Base-year Comparisons

Historical comparisons of the California Standards Test (CST) for Science results are routinely performed to identify the trends in student performance and test characteristics over time. Such comparisons were performed over a period of the three most recent years of administration—2014, 2015, and 2016—and the base year.

The indicators of student performance include the mean and standard deviation of scale scores, observed score ranges, and the percentage of students classified into proficient and advanced performance levels. Test characteristics are compared by looking at the mean proportion correct, overall reliability and standard errors of measurement (SEM), as well as the mean item response theory (IRT) *b*-value for each CST for Science.

The base year of each CST for Science refers to the year in which the base score scale was established. Operational forms administered in the years following the base year are linked to the base year score scale using procedures described in Chapter 2.

The base years for the CSTs for Science are presented in Table 10.1.

Table 10.1 Base Years for CSTs for Science

CST	Base Year
Grade 5 Science	2004
Grade 8 Science	2006
Grade 10 Life Science	2006

The base years differ over CSTs for Science because the grade-level science CSTs were introduced in grade five in 2004 and in grades eight and ten in 2006. Thus, 2004 is the base year for the grade five CST for Science, and 2006 is the base year for the CSTs for Science in grades eight and ten.

Student Performance

Table 10.A.1 on page 129 contains the number of students assessed and the means and standard deviations of students' scale scores in the base year and in 2014, 2015, and 2016 for each CST for Science. As noted in previous chapters, the CST for Science reporting scales range from 150 to 600 for all of the tests.

CST for Science scale scores are used to classify student results into one of five performance levels: far below basic, below basic, basic, proficient, and advanced. The percentages of students qualifying for the proficient and advanced levels are presented in Table 10.A.3 and Table 10.A.4 on page 129; please note that this information may differ slightly from information found on the California Department of Education (CDE) California Assessment of Student Performance and Progress (CAASPP) reporting Web page at http://caaspp.cde.ca.gov due to differing dates on which data were accessed. The goal is for all students to achieve at or above the proficient level by 2014.

Table 10.A.5 through Table 10.A.7 show, for each CST for Science, the distribution of scale scores observed in the base year and in 2014, 2015, and 2016. Frequency counts are provided for each scale score interval of 30. A frequency count of "N/A" indicates that there are no obtainable scale scores within that scale-score range. For all CSTs for Science, a minimum score of 300 is required for a student to reach the basic level of performance, and

a minimum score of 350 is required for a student to reach the proficient level of performance.

Test Characteristics

The item and test analysis results of the CSTs for Science over the past several years indicate that the CSTs for Science meet the technical criteria established in professional standards for high-stakes tests. In addition, every year efforts are made to improve the technical quality of each CST for Science. For example, in the years prior to 2013, efforts were made to field test more easy items for some CSTs for Science where previous field testing resulted in an overabundance of very difficult items.

Table 10.B.1 in Appendix 10.B, which starts on page 132, presents the average proportion-correct values for the operational items in each CST for Science based on the equating samples. The mean proportion correct is affected by both the difficulty of the items and the abilities of the students administered the items.

Table 10.B.2 shows the mean equated IRT *b*-values for the CST for Science operational items based on the equating samples. The mean equated IRT *b*-values reflect only average item difficulty. Please note that comparisons of mean *b*-values should be made only within a given test; they should not be compared across grade-level tests.

The average point-biserial correlations for the CSTs for Science are presented in Table 10.B.3. The reliabilities and SEM expressed in raw score units appear in Table 10.B.4 and Table 10.B.5. Like the average proportion correct, point-biserial correlations and reliabilities are affected by both item characteristics and student characteristics.

Appendix 10.A—Historical Comparisons Tables, Student Performance

Table 10.A.1 Number of Students Tested (with valid scores) of CSTs for Science Across Base Year, 2014, 2015, and 2016

CST	Base	2014	2015	2016
Grade 5 Science	483,931	432,439	437,432	442,498
Grade 8 Science	478,667	435,173	432,391	433,015
Grade 10 Life Science	461,634	435,705	434,766	449,114

Table 10.A.2 Scale Score Means and Standard Deviations of CSTs for Science Across Base Year, 2014, 2015, and 2016

CST	Base Mean	Base S.D.	2014 Mean	2014 S.D.	2015 Mean	2015 S.D.	2016 Mean	2016 S.D.
Grade 5 Science	318	44	368	70	360	64	357	64
Grade 8 Science	331	71	392	94	387	96	381	95
Grade 10 Life Science	327	58	360	65	358	65	354	65

Table 10.A.3 Percentage of Proficient and Above Across Base Year, 2014, 2015, and 2016

CST	Base	2014	2015	2016
Grade 5 Science	24%	59%	55%	53%
Grade 8 Science	38%	66%	63%	61%
Grade 10 Life Science	34%	56%	53%	50%

Table 10.A.4 Percentage of Advanced Across Base Year, 2014, 2015, and 2016

CST	Base	2014	2015	2016
Grade 5 Science	3%	27%	21%	20%
Grade 8 Science	17%	44%	41%	38%
Grade 10 Life Science	13%	28%	25%	23%

Table 10.A.5 Observed Score Distributions of CSTs for Science Across Base Year, 2014, 2015, and 2016, Grade Five

Observed Score Distributions	Base	2014	2015	2016
570–600	3	3,656	1,220	1,178
540–569	30	7,901	3,634	3,458
510–539	105	N/A	N/A	N/A
480–509	590	11,155	6,378	6,061
450–479	1,414	28,096	20,472	19,593
420–449	7,041	48,017	42,785	41,607
390–419	21,141	47,322	63,795	62,960
360–389	49,769	84,592	74,955	74,350
330–359	93,778	68,712	75,668	75,547
300–329	141,637	60,070	73,871	76,260
270–299	106,602	41,427	43,221	46,078
240–269	55,651	24,938	22,721	25,222
210–239	5,829	5,902	7,657	8,979
180–209	298	603	1,000	1,127
150–179	43	48	55	78

A frequency count of "N/A" indicates that there are no obtainable scale scores within that scale-score range.

Table 10.A.6 Observed Score Distributions of CSTs for Science Across Base Year, 2014, 2015, and 2016, Grade Eight

Observed Score Distributions	Base	2014	2015	2016
570–600	1,428	17,821	17,588	15,527
540–569	2,429	11,011	10,438	9,324
510–539	4,190	25,833	25,193	22,570
480–509	5,863	28,913	13,877	12,907
450–479	13,160	30,647	45,634	42,835
420–449	26,047	46,840	47,119	45,654
390–419	36,389	59,862	46,047	45,692
360–389	61,392	54,544	55,443	56,548
330–359	76,067	46,363	46,978	48,980
300–329	85,106	44,819	45,892	48,594
270–299	67,991	25,886	32,694	34,924
240–269	57,597	21,981	18,724	20,127
210–239	32,609	11,605	16,560	17,930
180–209	7,222	6,710	6,161	6,862
150–179	1,177	2,338	4,043	4,541

A frequency count of "N/A" indicates that there are no obtainable scale scores within that scale-score range.

Table 10.A.7 Observed Score Distributions of CSTs for Life Science Across Base Year, 2014, 2015, and 2016 (Grade Ten)

Observed Score Distributions	Base	2014	2015	2016
570–600	238	1,640	1,037	1,176
540–569	648	4,072	2,843	2,733
510–539	1,121	N/A	4,935	4,726
480–509	4,068	14,567	6,939	6,505
450–479	6,808	20,625	18,172	17,702
420–449	15,540	37,886	35,811	33,650
390–419	38,323	55,253	53,952	51,445
360–389	59,085	69,031	81,458	80,606
330–359	74,451	85,438	74,001	76,245
300–329	91,629	67,436	76,955	83,618
270–299	85,472	47,836	39,908	45,773
240–269	71,350	24,268	28,419	33,029
210–239	11,828	6,915	9,388	10,858
180–209	942	680	897	984
150–179	131	58	51	64

A frequency count of "N/A" indicates that there are no obtainable scale scores within that scale-score range.

Appendix 10.B—Historical Comparisons Tables, Test Characteristics

Table 10.B.1 Mean Proportion Correct for Operational Test Items Across Base Year, 2014, 2015, and 2016

CST	Base	2014	2015	2016
Grade 5 Science	0.47	0.71	0.69	0.68
Grade 8 Science	0.50	0.70	0.70	0.69
Grade 10 Life Science	0.51	0.67	0.65	0.64

Table 10.B.2 Mean IRT *b*-values for Operational Test Items Across Base Year, 2014, 2015, and 2016

CST	Base	2014	2015	2016
Grade 5 Science	-0.19	-0.64	-0.62	-0.62
Grade 8 Science	0.10	-0.34	-0.39	-0.39
Grade 10 Life Science	-0.04	-0.31	-0.26	-0.26

Table 10.B.3 Mean Point-Biserial Correlation for Operational Test Items Across Base Year, 2014, 2015, and 2016

CST	Base	2014	2015	2016
Grade 5 Science	0.33	0.43	0.41	0.42
Grade 8 Science	0.35	0.42	0.42	0.42
Grade 10 Life Science	0.41	0.42	0.43	0.43

Table 10.B.4 Score Reliabilities (Cronbach's Alpha) of CSTs for Science Across Base Year, 2014, 2015, and 2016

CST	Base	2014	2015	2016
Grade 5 Science	0.86	0.92	0.92	0.92
Grade 8 Science	0.88	0.92	0.92	0.92
Grade 10 Life Science	0.92	0.92	0.92	0.92

Table 10.B.5 SEM of CSTs for Science Across Base Year, 2014, 2015, and 2016

CST	Base	2014	2015	2016
Grade 5 Science	3.6	3.1	3.2	3.2
Grade 8 Science	3.5	3.1	3.1	3.2
Grade 10 Life Science	3.4	3.2	3.2	3.2