California Department of Education Assessment Development and Administration Division

California Standards Tests Technical Report Spring 2014 Administration

Final Version Submitted February 13, 2015

Educational Testing Service

Contract No. 5417

Table of Contents (Summarized)

Chapter 1: Introduction 1
- Background 1
- Test Purpose 1
- Test Content 1
- Intended Population 2
- Intended Use and Purpose of Test Scores 2
- Testing Window 3
- Significant CAASPP Developments in 2014 3
- Limitations of the Assessment 3
- Groups and Organizations Involved with the CAASPP System 4
- Overview of the Technical Report 5

Chapter 2: An Overview of CST Processes 8
- Item Development 8
- Test Assembly 9
- Test Administration 10
- Universal Tools, Designated Supports, and Accommodations 12
- Scores 13
- Equating 14
- References 18

Chapter 3: Item Development 29
- Rules for Item Development 29
- Selection of Item Writers 30
- Item Review Process 31
- Field Testing 35
- CDE Data Review 36
- Item Banking 36
- References 37

Chapter 4: Test Assembly 38
- Test Length 38
- Rules for Item Selection 38
- Rules for Item Sequence and Layout 41
- Reference 42

Chapter 5: Test Administration 47
- Test Security and Confidentiality 47
- Procedures to Maintain Standardization 52
- Universal Tools, Designated Supports, and Accommodations for Students with Disabilities 54
- Testing Incidents 55
- Testing Improprieties 56
- References 57

Chapter 6: Performance Standards 58
- Background 58
- Standard-Setting Procedure 58
- Standard-Setting Methodologies 59
- Results 60
- References 62

Chapter 7: Scoring and Reporting 63
- Procedures for Maintaining and Retrieving Individual Scores 63
- Types of Scores and Subscores 64
- Score Verification Procedures 64
- Overview of Score Aggregation Procedures 65
- Reports Produced and Scores for Each Report 67
- Criteria for Interpreting Test Scores 69
- Criteria for Interpreting Score Reports 69
- References 70

Chapter 8: Analyses 81
- Background 81
- Samples Used for the Analyses 81
- Classical Item Analyses 82

Reliability Analyses 83
Decision Classification Analyses 85
Validity Evidence 86
IRT Analyses 93
Differential Item Functioning Analyses 95
References 98
Chapter 9: Quality Control Procedures 114
Quality Control of Item Development 114
Quality Control of the Item Bank 116
Quality Control of Test Form Development 116
Quality Control of Test Materials 117
Quality Control of Scanning 117
Quality Control of Image Editing 118
Quality Control of Answer Document Processing and Scoring 118
Quality Control of Psychometric Processes 119
Quality Control of Reporting 122
Reference 124
Chapter 10: Historical Comparisons 125
Base-year Comparisons 125
Examinee Performance 125
Test Characteristics 126

Table of Contents (Full)

Acronyms and Initialisms Used in the *CST Technical Report* xiii
Chapter 1: Introduction 1
Background 1
Test Purpose 1
Test Content 1
Intended Population 2
Intended Use and Purpose of Test Scores 2
Testing Window 3
Significant CAASPP Developments in 2014 3
Renamed the Program 3
Reduced the Number of Required Tests 3
Reduced the Number of Test Versions 3
Updated Universal Tools, Designated Supports, and Accommodations (formerly Modifications, Accommodations, and Variations) 3
Suspended Reporting of Adequate Yearly Progress and the Academic Performance Index 3
Limitations of the Assessment 3
Score Interpretation 3
Out-of-Level Testing 3
Score Comparison 4
Groups and Organizations Involved with the CAASPP System 4
State Board of Education 4
California Department of Education 4
Contractors 4
Overview of the Technical Report 5
References 7
Chapter 2: An Overview of CST Processes 8
Item Development 8
Item Formats 8
Item Specifications 8
Item Banking 8
Item Refresh Rate 9
Test Assembly 9
Test Length 9
Test Blueprints 9
Content Rules and Item Selection 9
Replacement Items 10
Psychometric Criteria 10
Test Administration 10
Test Security and Confidentiality 10
Procedures to Maintain Standardization 11
Universal Tools, Designated Supports, and Accommodations 12
Non-embedded Supports 12
Individualized Aids (Previously Called Modifications) 12
Special Services Summaries 12
Scores 13
Aggregation Procedures 13
Equating 14
Post-Equating 14
Pre-Equating 14
Equating Samples 17
Pre-Equating the Braille Versions of the CSTs 17
References 18
Appendix 2.A—CST Items and Estimated Time Chart 19
Appendix 2.B—Reporting Clusters 20
Science 20
Appendix 2.C—Universal Tools, Designated Supports, and Accommodations for the California Assessment of Student Performance and Progress 21
Appendix 2.D—Special Services Summary Tables 23

Chapter 3: Item Development **29**
Rules for Item Development **29**
Item Specifications 29
Expected Item Ratio 30
Selection of Item Writers **30**
Criteria for Selecting Item Writers 30
Item Review Process **31**
Contractor Review 31
Content Expert Reviews 32
Statewide Pupil Assessment Review Panel 35
Field Testing **35**
Stand-alone Field Testing 35
Embedded Field-test Items 35
CDE Data Review **36**
Item Banking **36**
References **37**
Chapter 4: Test Assembly **38**
Test Length **38**
Rules for Item Selection **38**
Test Blueprint 38
Content Rules and Item Selection 38
Replacement Items 39
Psychometric Criteria 39
Projected Psychometric Properties of the Assembled Tests 40
Rules for Item Sequence and Layout **41**
Reference **42**
Appendix 4.A—Technical Characteristics **43**
Appendix 4.B—Cluster Targets **44**
Chapter 5: Test Administration **47**
Test Security and Confidentiality **47**
ETS's Office of Testing Integrity 47
Test Development 47
Item and Data Review 47
Item Banking 48
Transfer of Forms and Items to the CDE 48
Security of Electronic Files Using a Firewall 48
Printing and Publishing 49
Test Administration 49
Test Delivery 49
Processing and Scoring 50
Data Management 50
Transfer of Scores via Secure Data Exchange 51
Statistical Analysis 51
Reporting and Posting Results 51
Student Confidentiality 51
Student Test Results 51
Procedures to Maintain Standardization **52**
Test Administrators 52
Directions for Administration 53
CAASPP LEA and Test Site Coordinator Manual 53
Test Management System Manuals 54
Test Booklets 54
Universal Tools, Designated Supports, and Accommodations for Students with Disabilities **54**
Identification 55
Scoring 55
Testing Incidents **55**
Social Media Security Breaches 55
Testing Improprieties **56**
References **57**
Chapter 6: Performance Standards **58**
Background **58**

Standard-Setting Procedure **58**
Standard-Setting Methodologies **59**
Modified Angoff Method 59
Bookmark Method 59
Results **60**
References **62**
Chapter 7: Scoring and Reporting **63**
Procedures for Maintaining and Retrieving Individual Scores **63**
Scoring and Reporting Specifications 63
Scanning and Scoring 63
Types of Scores and Subscores **64**
Raw Score 64
Subscore 64
Scale Score 64
Performance Levels 64
Score Verification Procedures **64**
Scoring Key Verification Process 65
Overview of Score Aggregation Procedures **65**
Individual Scores 65
Group Scores 66
Reports Produced and Scores for Each Report **67**
Types of Score Reports 67
Score Report Contents 68
Score Report Applications 69
Criteria for Interpreting Test Scores **69**
Criteria for Interpreting Score Reports **69**
References **70**
Appendix 7.A—Scale Score Distribution Tables **71**
Appendix 7.B—Demographic Summaries **72**
Appendix 7.C—Types of Score Reports **78**
Chapter 8: Analyses **81**
Background **81**
Samples Used for the Analyses **81**
Classical Item Analyses **82**
Multiple-Choice Items 82
Reliability Analyses **83**
Intercorrelations, Reliabilities, and SEMs for Reporting Clusters 84
Subgroup Reliabilities and SEMs 84
Conditional Standard Errors of Measurement 84
Decision Classification Analyses **85**
Validity Evidence **86**
Purpose of the CSTs 87
The Constructs to Be Measured 87
Interpretations and Uses of the Scores Generated 87
Intended Test Population(s) 88
Validity Evidence Collected 88
Evidence Based on Response Processes 91
Evidence Based on Internal Structure 91
Evidence Based on Consequences of Testing 93
IRT Analyses **93**
Post-Equating 93
Pre-Equating 94
Summaries of Scaled IRT *b*-values 94
Evaluation of Pre-Equating 95
Equating Results 95
Differential Item Functioning Analyses **95**
References **98**
Appendix 8.A—Classical Analyses **100**
Appendix 8.B—Reliability Analyses **101**
Appendix 8.C—IRT Analyses **109**
Scaling and Pre-Equating Results 110

Chapter 9: Quality Control Procedures **114**
Quality Control of Item Development **114**
Item Specifications 114
Item Writers 114
Internal Contractor Reviews 114
Assessment Review Panel Review 115
Statewide Pupil Assessment Review Panel Review 115
Data Review of Field-tested Items 115
Quality Control of the Item Bank **116**
Quality Control of Test Form Development **116**
Quality Control of Test Materials **117**
Collecting Test Materials 117
Processing Test Materials 117
Quality Control of Scanning **117**
Quality Control of Image Editing **118**
Quality Control of Answer Document Processing and Scoring **118**
Accountability of Answer Documents 118
Processing of Answer Documents 119
Scoring and Reporting Specifications 119
Storing Answer Documents 119
Quality Control of Psychometric Processes **119**
Score Key Verification Procedures 119
Quality Control of Item Analyses and the Equating Process 119
Score Verification Process 121
Year-to-Year Comparison Analyses 121
Offloads to Test Development 121
Quality Control of Reporting **122**
Electronic Reporting 122
Excluding Student Scores from Summary Reports 122
Reference **124**
Chapter 10: Historical Comparisons **125**
Base-year Comparisons **125**
Examinee Performance **125**
Test Characteristics **126**
Appendix 10.A—Historical Comparisons Tables, Examinee Performance **127**
Appendix 10.B—Historical Comparisons Tables, Test Characteristics **128**

Tables

Table 2.1 Scale-Score Ranges for Performance Levels 16
Table 2.C.1 Matrix One Part 2: Non-Embedded Supports for the CSTs 21
Table 2.D.1 Special Services Summary for Science, Grades Five, Eight, and Ten (Life Science) 23
Table 4.1 Statistical Targets for CST Test Assembly 40
Table 4.A.1 Summary of 2014 CST Projected Raw Score Statistics 43
Table 4.A.2 Summary of 2014 CST Projected Item Statistics 43
Table 7.1 Mean and Standard Deviation of Raw and Scale Scores for the CSTs 66
Table 7.2 Percentages of Examinees in Performance Levels for CSTs 66
Table 7.3 Subgroup Definitions 67
Table 7.4 Types of CST Reports 68
Table 7.A.1 Distribution of CST Scale Scores for Science 71
Table 7.B.1 Demographic Summary for Science, Grade Five 72
Table 7.B.2 Demographic Summary for Science, Grade Eight 74
Table 7.B.3 Demographic Summary for Grade Ten Life Science 76
Table 7.C.1 Score Reports Reflecting CST Results 78
Table 8.1 Mean and Median Proportion Correct and Point-Biserial by Test Form—Current Administration 82
Table 8.2 Reliabilities and SEMs for the CSTs 84
Table 8.3 Scale Score CSEM at Performance-level Cut Points 85
Table 8.4 Original Year of Administration for CSTs 91
Table 8.A.1 Item-by-item *p*-value and Point Biserial for Science, Grades Five, Eight, and Ten—Current Year (2014) and Original Year of Administration 100
Table 8.B.1 Subscore Reliabilities and Intercorrelations for Science 101
Table 8.B.2 Reliabilities and SEMs for the CSTs by Gender 101
Table 8.B.3 Reliabilities and SEMs for the CSTs by Economic Status 101

Table 8.B.4 Reliabilities and SEMs for the CSTs by Special Services 102
Table 8.B.5 Reliabilities and SEMs for the CSTs by English-Language Fluency 102
Table 8.B.6 Reliabilities and SEMs for the CSTs by English-Language Fluency 102
Table 8.B.7 Reliabilities and SEMs for the CSTs by Primary Ethnicity 102
Table 8.B.8 Reliabilities and SEMs for the CSTs by Primary Ethnicity 102
Table 8.B.9 Reliabilities and SEMs for the CSTs by Primary Ethnicity 103
Table 8.B.10 Reliabilities and SEMs for the CSTs by Primary Ethnicity-for-Not Economically Disadvantaged 103
Table 8.B.11 Reliabilities and SEMs for the CSTs by Primary Ethnicity-for-Not Economically Disadvantaged 103
Table 8.B.12 Reliabilities and SEMs for the CSTs by Primary Ethnicity-for-Not Economically Disadvantaged 103
Table 8.B.13 Reliabilities and SEMs for the CSTs by Primary Ethnicity-for-Economically Disadvantaged 103
Table 8.B.14 Reliabilities and SEMs for the CSTs by Primary Ethnicity-for-Economically Disadvantaged 104
Table 8.B.15 Reliabilities and SEMs for the CSTs by Primary Ethnicity-for-Economically Disadvantaged 104
Table 8.B.16 Subscore Reliabilities and SEM for Science by Gender/Economic Status 104
Table 8.B.17 Subscore Reliabilities and SEM for Science by Special Services/English Fluency 105
Table 8.B.18 Subscore Reliabilities and SEM for Science by Primary Ethnicity 105
Table 8.B.19 Subscore Reliabilities and SEM for Science by Primary Ethnicity-for-Not Economically Disadvantaged 106
Table 8.B.20 Subscore Reliabilities and SEM for Science by Primary Ethnicity-for-Economically Disadvantaged 107
Table 8.B.21 Reliability of Classification for Science, Grade Five 107
Table 8.B.22 Reliability of Classification for Science, Grade Eight 108
Table 8.B.23 Reliability of Classification for Life Science (Grade 10) 108
Table 8.C.1 IRT *b*-values for Life Science (Grade 10) 109
Table 8.C.2 Distribution of IRT *b*-values for Life Science (Grade 10) 109
Table 8.C.3 New Conversion for Science, Grade Five 110
Table 8.C.4 New Conversion for Science, Grade Eight (paper-pencil) 111
Table 8.C.5 New Conversion for Science, Grade Eight (Braille) 112
Table 8.C.6 New Conversion for Life Science (Grade 10) 113
Table 10.1 Base Years for CSTs 125
Table 10.A.1 Number of Examinees Tested, Scale Score Means, and Standard Deviations of CSTs Across Base Year, 2012, 2013, and 2014 127
Table 10.A.2 Percentage of Proficient and Above and Percentage of Advanced Across Base Year, 2012, 2013, and 2014 127
Table 10.A.3 Observed Score Distributions of CSTs Across Base Year, 2012, 2013, and 2014 for Science (Grades Five and Eight and Life Science [Grade Ten]) 127
Table 10.B.1 Mean Proportion Correct for Operational Test Items Across Base Year, 2012, 2013, and 2014 128
Table 10.B.2 Mean IRT *b*-values for Operational Test Items Across Base Year, 2012, 2013, and 2014 128
Table 10.B.3 Mean Point-Biserial Correlation for Operational Test Items Across Base Year, 2012, 2013, and 2014 128
Table 10.B.4 Score Reliabilities (Cronbach's Alpha) and SEM of CSTs Across Base Year, 2012, 2013, and 2014 128

Figures

Figure 3.1 The ETS Item Development Process for the CAASPP System 29
Figure 4.A.1 Plots of Target Information Function and Projected Information for Total Test for Science 43
Figure 4.B.1 Plots of Target Information Functions and Projected Information for Clusters for Science, Grade Five 44
Figure 4.B.2 Plots of Target Information Functions and Projected Information for Clusters for Science, Grade Eight 45
Figure 4.B.3 Plots of Target Information Functions and Projected Information for Clusters for Life Science, Grade Ten 46
Figure 6.1 Bookmark Standard-setting Process for the CSTs 60
Figure 8.1 Decision Accuracy for Achieving a Performance Level 86
Figure 8.2 Decision Consistency for Achieving a Performance Level 86

Acronyms and Initialisms Used in the *CST Technical Report*

ADA	Americans with Disabilities Act
AERA	American Educational Research Association
API	Academic Performance Index
ARP	Assessment Review Panel
ASL	American Sign Language
AYP	adequate yearly progress
CAASPP	California Assessment of Student Performance and Progress
CAHSEE	California High School Exit Examination
CalTAC	California Technical Assistance Center
CAPA	California Alternate Performance Assessment
CCR	*California Code of Regulations*
CDE	California Department of Education
CDS	county/district/school
CELDT	California English Language Development Test
CI	confidence interval
CMA	California Modified Assessment
CR	constructed response
CRL	California Reading List
CSEMs	conditional standard errors of measurement
CSTs	California Standards Tests
CSU	California State University
DFA	*Directions for Administration*
DIF	differential item functioning
DOK	depth of knowledge
DPLT	designated primary language test
DQS	Data Quality Services
d-study	decision study
EAP	Early Assessment Program
EC	*Education Code*
EL	English learner
ELA	English–language arts
EM	expectation maximization
EOC	end-of-course
ESEA	Elementary and Secondary Education Act
ETS	Educational Testing Service
FIA	final item analysis
GENASYS	Generalized Analysis System
g-study	generalizability study
HumRRO	Human Resource Research Organization
ICC	item characteristic curves intraclass correlation coefficient
IEP	individualized education program
I-FEP	initially fluent English proficient
IRT	item response theory
IT	Information Technology
LEA	local educational agency
MC	multiple choice
MCE	Manually Coded English
MH DIF	Mantel-Haenszel DIF
NCME	National Council on Measurement in Education
NPS	nonpublic, nonsectarian school
NRT	norm-referenced test
NSLP	National School Lunch Program
OIB	ordered item booklet
OTI	Office of Testing Integrity
p-value	item proportion correct
PSAA	Public School Accountability Act
Pt-Bis	point-biserial correlations
QC	quality control
QTR	Quick-turnaround Reporting
RACF	Random Access Control Facility
R-FEP	reclassified fluent English proficient
RTQs	released test questions
SBE	State Board of Education
SD	standard deviation
SEM	standard error of measurement
SFTP	secure file transfer protocol
SGID	School and Grade Identification sheet
SKM	score key management
SPAR	Statewide Pupil Assessment Review
STAR	Standardized Testing and Reporting
STS	Standards-based Tests in Spanish
TIF	test information function
USDOE	United States Department of Education
WRMSD	weighted root-mean-square difference

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 (ELA), 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 its 2014 administration, the CAASPP System had four components for the paper-pencil tests:

- CSTs, produced for California public schools to assess the California content standards for science in grades five, eight, and ten
- CMA, 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, produced for students with an IEP and who have significant cognitive disabilities in grades two through eleven and are not able to take the CSTs with accommodations and/or non-embedded accessibility supports or the CMA with accommodations
- STS, an assessment of students' achievement of California's content standards for Spanish-speaking English learners that is administered as the CAASPP System's designated primary language test (DPLT)

Test Purpose

The three grade-level CSTs for Science form the cornerstone of the CAASPP System. The CSTs, 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. These standards describe what students should know and be able to do at each grade level.

Test Content

The CST 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 CMA for Life Science administered in grade ten assesses science content standards in grades six, seven, eight, and 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 433,000 to 438,000 test-takers in 2014.

All students enrolled in grades five, eight, and ten in California public schools on the day testing begins are required to take a CST science assessment or, for eligible students, a CMA science assessment; or for students in grades two through eleven who meet the eligibility requirements, the CAPA. 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 the science CSTs, the science CMA, or the CAPA is made by their IEP team.

Parents 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 submit a written request may be exempted from taking the tests (*Education Code [EC]* Section 60615).

Intended Use and Purpose of Test Scores

The results for tests within the CAASPP System are used for three primary purposes, described as follows (excerpted from the *EC* Section 60602 Web page at http://www.leginfo.ca.gov/cgi-bin/displaycode?section=edc&group=60001-61000&file=60600-60603):

"60602. (a) (1) First and foremost, provide information on the academic status and progress of individual pupils to those pupils, their parents, and their teachers. This information should be designed to assist in the improvement of teaching and learning in California public classrooms. The Legislature recognizes that, in addition to statewide assessments that will occur as specified in this chapter, school districts will conduct additional ongoing pupil diagnostic assessment and provide information regarding pupil performance based on those assessments on a regular basis to parents or guardians and schools. The Legislature further recognizes that local diagnostic assessment is a primary mechanism through which academic strengths and weaknesses are identified."

"60602. (a) (4) Provide information to pupils, parents or guardians, teachers, schools, and school districts 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. (c) It is the intent of the Legislature that parents, classroom teachers, other educators, governing board members of school districts, and the public be involved, in an active and ongoing basis, in the design and implementation of the statewide pupil assessment program and the development of assessment instruments."

"60602. (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 Standardized Testing and Reporting Program become open and transparent to teachers, parents, and pupils, to assist all the 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 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]).

Significant CAASPP Developments in 2014

Renamed the Program

The paper-pencil tests administered in 2014 are a component of the CAASPP System.

Reduced the Number of Required Tests

Because California is in transition to the new assessment system, the number of non-computer-administered tests is reduced to only include grade-level science.

Reduced the Number of Test Versions

The number of CST for Science versions available for administration was reduced to three.

Updated Universal Tools, Designated Supports, and Accommodations (formerly Modifications, Accommodations, and Variations)

Students were permitted the use of universal tools, designated supports, and accommodations as outlined in 5 *CCR* Section 853.5(b), (d), (f), and (h).

Suspended Reporting of Adequate Yearly Progress and the Academic Performance Index

The Adequate Yearly Progress (AYP) report submitted to the U.S. Department of Education in 2014 does not include CST results. Reporting of Academic Performance Index (API) data has been suspended.

Limitations of the Assessment

Score Interpretation

An LEA may use CST 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 local assessments, classroom tests, student grades, classroom work, and teacher recommendations in addition to the child's CST results (CDE, 2013).

Out-of-Level Testing

Each CST 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, 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 2014 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

The SBE is responsible for assuring the compliance with programs that meet the requirement of the federal Elementary and Secondary Education Act (ESEA) and the state's Public School Accountability Act (PSAA) and was responsible for reporting CST results in terms of the AYP and API, which measure the academic performance and growth of schools on a variety of academic metrics. Data are not reported for either of these measures in 2014 for the CSTs.

California Department of Education

The CDE 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 CDE adopts textbooks for kindergarten through grade eight, adopts regulations to implement legislation, and has the authority to grant waivers of the *EC*.

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, and community partners which together, as a team, prepares students to live, work, and thrive in a highly connected world.

Contractors

Educational Testing Service

The CDE and the SBE contract with ETS to develop, administer, and report the CAASPP assessments. As the prime contractor, ETS has overall responsibility for working with the CDE to implement and maintain an effective assessment system and to coordinate the work of ETS and its subcontractor Pearson. 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 Management System for orders of materials and pre-identification services; and
- Completion of all psychometric activities.

Pearson

ETS also monitors and manages the work of Pearson, subcontractor to ETS for the CAASPP System. Activities conducted by Pearson include the following:

- 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 CSTs administered in spring 2014. The technical report contains nine additional chapters as follows:

- Chapter 2 presents a conceptual overview of processes involved in a testing cycle for a CST 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 items before the 2014 administration—in 2014, 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 forms reused in 2014.
- Chapter 5 presents the processes involved in the actual administration of the 2014 CSTs 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 CST for Science tests.
- Chapter 7 details the types of scores and score reports that are produced at the end of each administration of the CSTs and includes a discussion of quick-turnaround reporting.
- Chapter 8 summarizes the results of the test- and item-level analyses performed during the spring 2014 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 performance-level classifications, and the procedures designed to ensure the validity of CST score uses and interpretations. Also discussed in this chapter are item response theory (IRT), CST 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.
- 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

Chapter 2: An Overview of CST Processes

This chapter provides an overview of the processes involved in a typical test development and administration cycle for a CST. Also described are the specifications maintained by ETS to implement each of those processes. In 2014, three CSTs in science were administered. For the tests in grades five and eight, intact forms from 2010 and 2011 were administered respectively. For the test in Life Science, a modified test form from 2010 was administered because two of the original items were replaced. All three tests 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 in science administered in 2014 contain four-option multiple-choice items.

Item Specifications

There was no new item development in 2014. Prior to the 2013 administration, the CST 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 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 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 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.

The items selected for each CST 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

Before newly developed items were placed in the item bank, 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 32; and the Statewide Pupil Assessment Review (SPAR) panel, described in Chapter 3, starting on page 35.

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 form reuse using intact forms and modified intact forms in the 2014 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.

Test Assembly

Test Length

The CST grade-level science tests are composed of 60 operational items each. The considerations used in deciding the test length are described on page 38 in Chapter 4.

Each CST also includes six field-test items in addition to the operational items. Although there was no new item development for the 2014 administration, the field-test items were included as part of the intact forms. The total number of items, including operational and field-test items, in each CST 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 items to conform to the SBE-approved California content standards and test blueprints. The test blueprints for the CSTs for Science, adopted in 2002 by the SBE, are linked on the CDE CAASPP Science Assessments Web page at http://www.cde.ca.gov/ta/tg/ca/caasppscience.asp.

Although the test blueprints specify the number of items at the individual standard level, scores for the CST items are grouped into subcontent areas referred to as "reporting clusters." For each CST reporting cluster, the percentage of questions correctly answered is reported on a student's score report. A description of the CST 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 2010 and 2011 and modified intact forms from 2011 were used during the 2014 administration. (See Table 8.4 on page 91 for administration years.) Prior to the 2013 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 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 38.

Replacement Items

During the 2013 STAR administration, certain test items were exposed on social media Web sites. Although there was no evidence that these exposures had any measurable effect on 2013 test results, the most cautious approach called for replacing these items for the test forms in the 2014 administration. Specifically, the CST for Life Science (Grade 10) contained two replacement items.

Before replacing items, ETS considered whether the form with an exposed item could be exchanged with a form from a different previous test administration that contained no exposures. If there were no appropriate forms available, ETS sought to replace exposed items with previously used or field-tested items that best met the psychometric criteria and statistical targets described in the next sections. As CSTs would be pre-equated, ETS assessment development staff also considered minimizing the position difference between the replacement items and the exposed items in order to reduce the context effect.

Psychometric Criteria

The staff assessed the projected test characteristics during the preliminary review of the assembled forms. The statistical targets used to develop the 2014 forms and the projected characteristics of the assembled forms were presented starting from page 39 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 41 in Chapter 4.

All the forms in the 2014 CST administration were used in prior operational test administrations. In the CST for Life Science, two items were replaced due to social media exposure in 2013. See Table 8.4 on page 91 for the list containing the administration in which each CST was originally administered. Finally, see page 39 in Chapter 4 for a description of the item replacement process.

Test Administration

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

Test Security and Confidentiality

All tests within the CAASPP System are secure documents. For the CST 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 47.

In the pursuit of enforcing secure practices, ETS and the 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 47.

- 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
- Transfer of scores via secure data exchange
- Statistical analysis
- Reporting and posting results
- Student confidentiality
- Student test results

Procedures to Maintain Standardization

The CST 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, as described in this section.

Test Administrators

The CSTs 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 examiners, proctors, and scribes. The responsibilities of each of the staff members are included in the *CAASPP LEA and Test Site Coordinator Manual* (CDE, 2014); see page 52 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 examiners to administer the CSTs to students to be followed exactly so that all students have an equal opportunity to demonstrate their academic achievement (See page 53 in Chapter 5 for more information.)

***CAASPP LEA and Test Site Coordinator Manual*—**Test administration procedures for LEA CAASPP Coordinators and CAASPP Test Site Coordinators (See page 53 in Chapter 5 for more information.)

Test Management System manuals—Instructions for the Web-based modules that allow LEA CAASPP Coordinators to set up test administrations, order materials, and submit and correct student Pre-ID data; every module has its own user manual with detailed

instructions on how to use the Test Management System (See page 54 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 IEPs and most ELs take the CSTs under standard conditions. However, some students with IEPs and some ELs may need assistance when taking the CSTs. 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." 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 in 2014 and were mapped to CST answer documents so had data that could be collected. Table 2.C.1 also shows the answer document options in section A3 that are reported in Appendix 2.D and were defined but 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 for the CSTs is to enable the students to take the assessments, rather than give them an advantage over other students or artificially inflate their scores.

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.

Individualized Aids (Previously Called Modifications)

Individualized aids, previously called modifications, fundamentally change what is being tested and may interfere with the construct being measured. All individualized aids must be identified, approved, and listed in the student's IEP or Section 504 plan.

Special Services Summaries

The percentage of students using various universal tools, designated supports, and accommodations during the 2014 administration of the CSTs is presented in Appendix 2.D, which starts on page 23. 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.

Scores

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

Total test raw scores on each CST are converted to three-digit scale scores using the pre-equating process described starting on page 14. CST 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.

In addition to scale scores for the total content-area test, performance on the associated reporting clusters is reported. 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 scores are found in Chapter 7, which starts on page 63.

Aggregation Procedures

In order to provide meaningful results to the stakeholders, CST 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 scores aggregated at the state level.

Please note that aggregation is performed on valid scores only, which are cases where examinees 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. 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 student performance by grade are provided in Table 7.B.1 on page 72 in Appendix 7.B.

Demographic Subgroup Scores

In Table 7.B.1 through Table 7.B.3 starting on page 72, 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

Prior to the 2013 administration, the CSTs were equated to a reference form 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 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 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.

Pre-Equating

During the 2014 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.

The original conversion tables from the previous administrations of the CSTs for Science in grades five and eight (which used intact forms without any edits or replacement to items) are directly applied to the current administration. For the CST for Life Science (Grade 10), in which replacement or edited items are used, conversion tables are newly generated using the true-score equating method described below. The item parameters used for true-score equating are post-equating item parameters (from the reused forms) for the unchanged

items and the post-equated parameters from the most recent administration for the replacement or edited items.

Table 8.4 on page 91 shows the years the forms were introduced for each test. Also note that in 2014, two items were replaced in the CST for Life Science (Grade 10) because of exposure on social media Web sites.

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 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 (EM) cycle with no constraints. For both stages, the metric of the scale was controlled by the constant discrimination parameters.

Scaling

Prior to the 2013 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 (ICCs) based on the old and new difficulty estimates were removed from the linking set. The differences were calculated using the following formula:

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

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. In the 2014 administration, the same true-score equating process was conducted to develop the conversion tables for the CST for Life Science (Grade 10) test form, which contained replacement items.

The true-score equating procedure is based on the relationship between raw scores and ability (theta). For the CSTs, 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 scale score scale using the reference form CST raw-score-to-scale-score conversion tables and linear interpolation. Complete raw-to-scale-score conversion tables for the 2014 CSTs are presented in Table 8.C.3 through Table 8.C.6 in Appendix 8.C, starting on page 110. The raw scores and corresponding transformed scale scores are listed in those tables.

For all of the CSTs, 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

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

* CSTs named by number only are grade-level tests.

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 in subsequent years.

Equating Samples

To establish the 2002 reference scales, ETS staff used data based on samples of students selected from the 2002 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 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 are included in the equating samples.

Due to the implementation of the pre-equating, no equating sample is necessary for the 2014 administration.

Pre-Equating the Braille Versions of the CSTs

In some cases, it is not possible to translate all of the operational items contained in a CST 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.

In 2014, only the CST for Science (Grade 8) had a braille version that was equated using a shortened test because of an item that could not be translated into braille.

References

California Department of Education. (2014). *2014 CAASPP LEA and test site coordinator manual.* Sacramento, CA. Retrieved from http://caaspp.org/rsc/pdfs/CAASPP.coord_man.2014.pdf

California Department of Education. (2014). 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 Items and Estimated Time Chart

California Standards Tests	Grade 5		Grade 8		Grade 10	
	Total No. of Items	Time	Total No. of Items	Time	Total No. of Items	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

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

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 (previously known as "Essay responses dictated orally, in Manually Coded English, or in American Sign Language to a scribe, audio recorder, or speech-to-text converter" or "Student marks responses in test booklet and responses are transferred to a scorable answer document by an employee of the school, district, or nonpublic school" or "Student dictates multiple-choice question responses orally, or in Manually Coded English to a scribe, audio recorder, or speech-to-text converter for selected-response items")	A
F	Alternate Response Options includes adapted keyboards, large keyboards, StickyKeys, MouseKeys, FilterKeys, adapted mouse, touch screen, head wand, and switches. (previously known as "Assistive device that does not interfere with the independent work of the student on the multiple choice and/or essay responses [writing portion of the test]")	–
G	Braille (paper-pencil tests)	A
H	Large-print versions of a paper-pencil test (as available)	A
J, K	Breaks (previously known as "Extended Time" or "Test over more than one day for a test or test part to be administered in a single sitting" or "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	Separate Setting (previously known as "Test individual student separately, provided that a test examiner directly supervises the student" or "Test student in a small group setting" or "Test administered at home or in hospital by test examiner")	A
N	Dictionary, thesaurus	N/A
O	American Sign Language	A
Q	Calculator	N/A
R	Arithmetic table or formulas (not provided) on the science tests	Unmapped
S	Math Tools (I.e., non-embedded ruler, non-embedded protractor)	–
V	Assistive device that interferes with the independent work of the student on the multiple-choice and/or essay responses	Unmapped
W	Unmapped	Unmapped
X	Abacus	A
Y	Leave blank	Unmapped
Z	Read Aloud (previously known as "Test questions and answer options read aloud to pupil or used audio CD presentation – excluding passages")	A
Mark nothing	Color Overlay (previously known as "Colored overlay, mask, or other means to maintain visual attention")	U
Mark nothing	Magnification (previously known as "Visual magnifying equipment")	D

Option	(U) Universal Tool | (D) Designated Support | (A) Accommodation	
Mark nothing	Noise buffers (e.g., individual carrel or study enclosure or noise-cancelling headphones)	D
Mark nothing	Scratch Paper	U
Mark nothing	Simplified or clarified test administration directions (does not apply to test questions)	U
Mark nothing	Special lighting or acoustics, assistive devices (specific devices may require CAASPP contractor certification), and/or special or adaptive furniture	D
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 (STAR) 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	English learners (ELs) may have the opportunity to be tested separately with other ELs provided that the student is directly supervised by an employee of the school who has signed the test security affidavit and the student has been provided such a flexible setting as part of his/her regular instruction or assessment.	Unmapped
D	Translations (Glossary) (previously known as "Access to translation glossaries/word lists (English-to-primary language). Glossaries/Word lists shall not include definitions 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" in Table 2.D.1 presents the number of students using various special services on the CST for Science in grade five. The column with the heading of "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 accessibility supports are not presented. 5 *CCR*, Section 853.5(h) states that the use of "accessibility supports that change the construct being measured by a CAASPP test invalidate the test score and results in a score that cannot be compared with other CAASPP results. Scores for pupils' tests with accessibility supports that change the construct being measured by a CAASPP test will not be counted as participating in statewide testing (and impacts the accountability participation rate indicator) but pupils will receive an individual score report with their actual score (raw score)."

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

Special Services Summary for Science, Grades Five, Eight, and Ten (Life Science)						
All Tested	**Grade 5**	**Pct. of Total**	**Grade 8**	**Pct. of Total**	**Grade 10 Life Sci.**	**Pct. of Total**
B: Marked responses in test booklet	300	0.07%	125	0.03%	126	0.03%
C: Scribe	20	0.00%	20	0.00%	17	0.00%
F: Alternative response options	18	0.00%	14	0.00%	14	0.00%
G: Braille	8	0.00%	16	0.00%	13	0.00%
H: Large-print versions of a paper-pencil test	62	0.01%	76	0.02%	55	0.01%
J: Breaks (Tested over more than one day)	198	0.05%	175	0.04%	112	0.03%
K: Breaks (Had supervised breaks)	1,523	0.35%	1,008	0.23%	1,333	0.30%
L: Most beneficial time of day	577	0.13%	257	0.06%	280	0.06%
M: Separate setting	50	0.01%	61	0.01%	49	0.01%
N: Dictionary, thesaurus	N/A	N/A	N/A	N/A	N/A	N/A
O: American Sign Language	10	0.00%	5	0.00%	8	0.00%
Q: Calculator	N/A	N/A	N/A	N/A	N/A	N/A
R: Used an arithmetic table	4	0.00%	3	0.00%	11	0.00%
S: Math Tools	1	0.00%	1	0.00%	8	0.00%
V: Used interfering assistive device	1	0.00%	0	0.00%	0	0.00%
W: Unmapped	22	0.01%	17	0.00%	83	0.02%
X: Abacus	330	0.08%	381	0.09%	360	0.08%
Y: Leave blank	198	0.05%	150	0.03%	186	0.04%
Z: Read aloud	2,000	0.46%	777	0.18%	464	0.11%
Univ. tool, desig. support, or acc. is in Section 504 plan	212	0.05%	128	0.03%	68	0.02%
Univ. tool, desig. support, or acc. is in IEP	3,413	0.79%	2,210	0.51%	2,386	0.55%

Special Services Summary for Science, Grades Five, Eight, and Ten (Life Science)						
EL Test Variation A	69	0.02%	276	0.06%	126	0.03%
EL Test Variation B	11	0.00%	7	0.00%	123	0.03%
EL Test Variation C	51	0.01%	139	0.03%	450	0.10%
EL Test Variation D	18	0.00%	73	0.02%	323	0.07%
Any universal tool, desig. support, or accommodation or EL variation	3,914	0.90%	2,821	0.65%	3,193	0.73%
No universal tool, desig. support, or accommodation or EL variation	429,529	99.10%	434,125	99.35%	434,484	99.27%
Students Not in Special Education	**Grade 5**	**Pct. of Total**	**Grade 8**	**Pct. of Total**	**Grade 10 Life Sci.**	**Pct. of Total**
B: Marked responses in test booklet	41	0.01%	22	0.01%	17	0.00%
C: Scribe	5	0.00%	5	0.00%	3	0.00%
F: Alternative response options	4	0.00%	2	0.00%	1	0.00%
G: Braille	0	0.00%	0	0.00%	0	0.00%
H: Large-print versions of a paper-pencil test	9	0.00%	10	0.00%	4	0.00%
J: Breaks (Tested over more than one day)	40	0.01%	21	0.01%	6	0.00%
K: Breaks (Had supervised breaks)	261	0.06%	87	0.02%	112	0.03%
L: Most beneficial time of day	104	0.03%	21	0.01%	26	0.01%
M: Separate setting	17	0.00%	37	0.01%	22	0.01%
N: Dictionary, thesaurus	N/A	N/A	N/A	N/A	N/A	N/A
O: American Sign Language	0	0.00%	1	0.00%	0	0.00%
Q: Calculator	N/A	N/A	N/A	N/A	N/A	N/A
R: Used an arithmetic table	3	0.00%	1	0.00%	3	0.00%
S: Math Tools	1	0.00%	0	0.00%	2	0.00%
V: Used interfering assistive device	1	0.00%	0	0.00%	0	0.00%
W: Unmapped	1	0.00%	1	0.00%	6	0.00%
X: Abacus	66	0.02%	50	0.01%	32	0.01%
Y: Leave blank	43	0.01%	29	0.01%	19	0.00%
Z: Read aloud	248	0.06%	75	0.02%	41	0.01%
Univ. tool, desig. support, or acc. is in Section 504 plan	197	0.05%	119	0.03%	62	0.02%
Univ. tool, desig. support, or acc. is in IEP	369	0.09%	137	0.03%	164	0.04%
EL Test Variation A	65	0.02%	274	0.07%	126	0.03%
EL Test Variation B	8	0.00%	6	0.00%	120	0.03%
EL Test Variation C	34	0.01%	133	0.03%	443	0.11%
EL Test Variation D	18	0.00%	73	0.02%	295	0.07%
Any universal tool, desig. support, or accommodation or EL variation	716	0.18%	700	0.17%	901	0.22%
No universal tool, desig. support, or accommodation or EL variation	407,648	99.82%	413,833	99.83%	410,935	99.78%
Students in Special Education	**Grade 5**	**Pct. of Total**	**Grade 8**	**Pct. of Total**	**Grade 10 Life Sci.**	**Pct. of Total**
B: Marked responses in test booklet	259	1.10%	103	0.46%	109	0.42%
C: Scribe	15	0.06%	15	0.07%	14	0.05%
F: Alternative response options	14	0.06%	12	0.05%	13	0.05%
G: Braille	8	0.03%	16	0.07%	13	0.05%
H: Large-print versions of a paper-pencil test	53	0.22%	66	0.29%	51	0.20%

Special Services Summary for Science, Grades Five, Eight, and Ten (Life Science)						
J: Breaks (Tested over more than one day)	157	0.66%	154	0.69%	106	0.41%
K: Breaks (Had supervised breaks)	1,259	5.33%	912	4.08%	1,212	4.70%
L: Most beneficial time of day	473	2.00%	232	1.04%	254	0.98%
M: Separate setting	33	0.14%	24	0.11%	26	0.10%
N: Dictionary, thesaurus	N/A	N/A	N/A	N/A	N/A	N/A
O: American Sign Language	10	0.04%	4	0.02%	8	0.03%
Q: Calculator	N/A	N/A	N/A	N/A	N/A	N/A
R: Used an arithmetic table	1	0.00%	2	0.01%	8	0.03%
S: Math Tools	0	0.00%	1	0.00%	6	0.02%
V: Used interfering assistive device	0	0.00%	0	0.00%	0	0.00%
W: Unmapped	21	0.09%	16	0.07%	76	0.29%
X: Abacus	263	1.11%	331	1.48%	328	1.27%
Y: Leave blank	155	0.66%	121	0.54%	166	0.64%
Z: Read aloud	1,751	7.41%	701	3.13%	420	1.63%
Univ. tool, desig. support, or acc. is in Section 504 plan	12	0.05%	9	0.04%	6	0.02%
Univ. tool, desig. support, or acc. is in IEP	3,043	12.88%	2,062	9.22%	2,207	8.56%
EL Test Variation A	4	0.02%	2	0.01%	0	0.00%
EL Test Variation B	3	0.01%	1	0.00%	3	0.01%
EL Test Variation C	17	0.07%	6	0.03%	7	0.03%
EL Test Variation D	0	0.00%	0	0.00%	28	0.11%
Any universal tool, desig. support, or accommodation or EL variation	3,194	13.52%	2,110	9.43%	2,277	8.83%
No universal tool, desig. support, or accommodation or EL variation	20,435	86.48%	20,266	90.57%	23,512	91.17%
English-Only Students	**Grade 5**	**Pct. of Total**	**Grade 8**	**Pct. of Total**	**Grade 10 Life Sci.**	**Pct. of Total**
B: Marked responses in test booklet	237	0.10%	109	0.05%	102	0.04%
C: Scribe	17	0.01%	13	0.01%	12	0.00%
F: Alternative response options	14	0.01%	8	0.00%	12	0.00%
G: Braille	6	0.00%	7	0.00%	10	0.00%
H: Large-print versions of a paper-pencil test	48	0.02%	57	0.02%	36	0.01%
J: Breaks (Tested over more than one day)	133	0.05%	104	0.04%	72	0.03%
K: Breaks (Had supervised breaks)	1,055	0.43%	674	0.28%	868	0.36%
L: Most beneficial time of day	389	0.16%	175	0.07%	195	0.08%
M: Separate setting	45	0.02%	51	0.02%	38	0.02%
N: Dictionary, thesaurus	N/A	N/A	N/A	N/A	N/A	N/A
O: American Sign Language	8	0.00%	2	0.00%	6	0.00%
Q: Calculator	N/A	N/A	N/A	N/A	N/A	N/A
R: Used an arithmetic table	1	0.00%	2	0.00%	7	0.00%
S: Math Tools	1	0.00%	1	0.00%	6	0.00%
V: Used interfering assistive device	0	0.00%	0	0.00%	0	0.00%
W: Unmapped	16	0.01%	5	0.00%	64	0.03%
X: Abacus	244	0.10%	277	0.12%	237	0.10%
Y: Leave blank	157	0.06%	106	0.04%	114	0.05%
Z: Read aloud	1,324	0.55%	455	0.19%	255	0.10%

Special Services Summary for Science, Grades Five, Eight, and Ten (Life Science)						
Univ. tool, desig. support, or acc. is in Section 504 plan	170	0.07%	103	0.04%	57	0.02%
Univ. tool, desig. support, or acc. is in IEP	2,357	0.97%	1,448	0.60%	1,532	0.63%
EL Test Variation A	6	0.00%	11	0.00%	3	0.00%
EL Test Variation B	2	0.00%	0	0.00%	1	0.00%
EL Test Variation C	9	0.00%	2	0.00%	6	0.00%
EL Test Variation D	3	0.00%	0	0.00%	3	0.00%
Any universal tool, desig. support, or accommodation or EL variation	2,646	1.09%	1,617	0.67%	1,635	0.67%
No universal tool, desig. support, or accommodation or EL variation	240,022	98.91%	238,929	99.33%	242,243	99.33%
Initially Fluent English Proficient (I-FEP) Students	**Grade 5**	**Pct. of Total**	**Grade 8**	**Pct. of Total**	**Grade 10 Life Sci.**	**Pct. of Total**
B: Marked responses in test booklet	5	0.02%	4	0.01%	6	0.02%
C: Scribe	0	0.00%	0	0.00%	0	0.00%
F: Alternative response options	0	0.00%	0	0.00%	1	0.00%
G: Braille	1	0.00%	3	0.01%	0	0.00%
H: Large-print versions of a paper-pencil test	2	0.01%	8	0.03%	3	0.01%
J: Breaks (Tested over more than one day)	3	0.01%	4	0.01%	3	0.01%
K: Breaks (Had supervised breaks)	24	0.12%	36	0.12%	42	0.12%
L: Most beneficial time of day	11	0.05%	12	0.04%	3	0.01%
M: Separate setting	2	0.01%	3	0.01%	1	0.00%
N: Dictionary, thesaurus	N/A	N/A	N/A	N/A	N/A	N/A
O: American Sign Language	0	0.00%	0	0.00%	0	0.00%
Q: Calculator	N/A	N/A	N/A	N/A	N/A	N/A
R: Used an arithmetic table	0	0.00%	0	0.00%	0	0.00%
S: Math Tools	0	0.00%	0	0.00%	0	0.00%
V: Used interfering assistive device	0	0.00%	0	0.00%	0	0.00%
W: Unmapped	0	0.00%	2	0.01%	5	0.01%
X: Abacus	1	0.00%	6	0.02%	13	0.04%
Y: Leave blank	0	0.00%	5	0.02%	3	0.01%
Z: Read aloud	23	0.11%	17	0.05%	9	0.02%
Univ. tool, desig. support, or acc. is in Section 504 plan	4	0.02%	3	0.01%	2	0.01%
Univ. tool, desig. support, or acc. is in IEP	45	0.22%	68	0.22%	72	0.20%
EL Test Variation A	0	0.00%	0	0.00%	1	0.00%
EL Test Variation B	0	0.00%	0	0.00%	1	0.00%
EL Test Variation C	0	0.00%	1	0.00%	3	0.01%
EL Test Variation D	0	0.00%	0	0.00%	3	0.01%
Any universal tool, desig. support, or accommodation or EL variation	50	0.24%	78	0.25%	82	0.23%
No universal tool, desig. support, or accommodation or EL variation	20,569	99.76%	30,846	99.75%	36,313	99.77%
English Learner (EL) Students	**Grade 5**	**Pct. of Total**	**Grade 8**	**Pct. of Total**	**Grade 10 Life Sci.**	**Pct. of Total**
B: Marked responses in test booklet	45	0.05%	4	0.01%	6	0.01%
C: Scribe	3	0.00%	2	0.00%	2	0.00%
F: Alternative response options	2	0.00%	1	0.00%	0	0.00%

Special Services Summary for Science, Grades Five, Eight, and Ten (Life Science)						
G: Braille	0	0.00%	3	0.01%	0	0.00%
H: Large-print versions of a paper-pencil test	9	0.01%	5	0.01%	3	0.01%
J: Breaks (Tested over more than one day)	49	0.06%	53	0.10%	23	0.05%
K: Breaks (Had supervised breaks)	377	0.42%	205	0.38%	281	0.60%
L: Most beneficial time of day	154	0.17%	51	0.10%	61	0.13%
M: Separate setting	2	0.00%	3	0.01%	5	0.01%
N: Dictionary, thesaurus	N/A	N/A	N/A	N/A	N/A	N/A
O: American Sign Language	2	0.00%	2	0.00%	1	0.00%
Q: Calculator	N/A	N/A	N/A	N/A	N/A	N/A
R: Used an arithmetic table	2	0.00%	1	0.00%	2	0.00%
S: Math Tools	0	0.00%	0	0.00%	2	0.00%
V: Used interfering assistive device	1	0.00%	0	0.00%	0	0.00%
W: Unmapped	6	0.01%	8	0.02%	4	0.01%
X: Abacus	76	0.09%	56	0.11%	79	0.17%
Y: Leave blank	31	0.03%	21	0.04%	45	0.10%
Z: Read aloud	589	0.66%	218	0.41%	146	0.31%
Univ. tool, desig. support, or acc. is in Section 504 plan	28	0.03%	9	0.02%	5	0.01%
Univ. tool, desig. support, or acc. is in IEP	893	1.01%	478	0.90%	521	1.12%
EL Test Variation A	50	0.06%	255	0.48%	122	0.26%
EL Test Variation B	9	0.01%	7	0.01%	121	0.26%
EL Test Variation C	39	0.04%	130	0.24%	440	0.94%
EL Test Variation D	15	0.02%	68	0.13%	313	0.67%
Any universal tool, desig. support, or accommodation or EL variation	1,060	1.19%	859	1.61%	1,195	2.56%
No universal tool, desig. support, or accommodation or EL variation	87,703	98.81%	52,435	98.39%	45,483	97.44%
Reclassified Fluent English Proficient (R-FEP) Students	**Grade 5**	**Pct. of Total**	**Grade 8**	**Pct. of Total**	**Grade 10 Life Sci.**	**Pct. of Total**
B: Marked responses in test booklet	12	0.02%	7	0.01%	12	0.01%
C: Scribe	0	0.00%	5	0.00%	3	0.00%
F: Alternative response options	1	0.00%	4	0.00%	1	0.00%
G: Braille	1	0.00%	2	0.00%	3	0.00%
H: Large-print versions of a paper-pencil test	3	0.00%	6	0.01%	13	0.01%
J: Breaks (Tested over more than one day)	10	0.01%	11	0.01%	13	0.01%
K: Breaks (Had supervised breaks)	56	0.07%	77	0.07%	121	0.11%
L: Most beneficial time of day	21	0.03%	18	0.02%	16	0.01%
M: Separate setting	0	0.00%	3	0.00%	4	0.00%
N: Dictionary, thesaurus	N/A	N/A	N/A	N/A	N/A	N/A
O: American Sign Language	0	0.00%	1	0.00%	1	0.00%
Q: Calculator	NA	NA	NA	0.00%	NA	0.01%
R: Used an arithmetic table	1	0.00%	0	0.00%	2	0.00%
S: Math Tools	0	0.00%	0	0.00%	0	0.00%
V: Used interfering assistive device	0	0.00%	0	0.00%	0	0.00%
W: Unmapped	0	0.00%	2	0.00%	9	0.01%

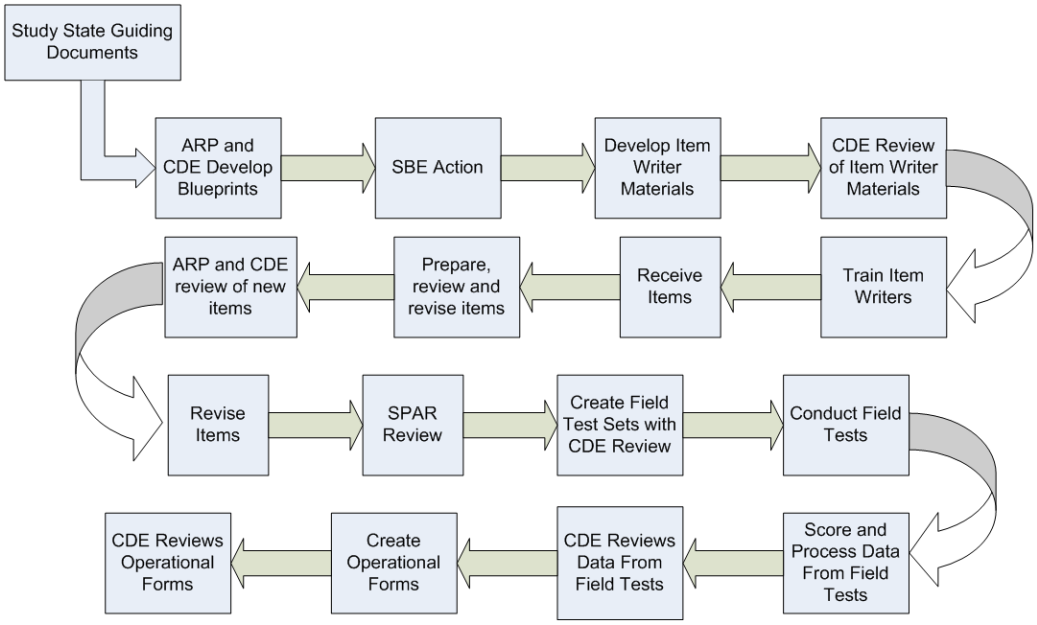
Special Services Summary for Science, Grades Five, Eight, and Ten (Life Science)						
X: Abacus	6	0.01%	40	0.04%	26	0.02%
Y: Leave blank	8	0.01%	14	0.01%	20	0.02%
Z: Read aloud	46	0.06%	66	0.06%	43	0.04%
Univ. tool, desig. support, or acc. is in Section 504 plan	7	0.01%	12	0.01%	4	0.00%
Univ. tool, desig. support, or acc. is in IEP	93	0.12%	186	0.17%	224	0.21%
EL Test Variation A	11	0.01%	9	0.01%	0	0.00%
EL Test Variation B	0	0.00%	0	0.00%	0	0.00%
EL Test Variation C	3	0.00%	1	0.00%	1	0.00%
EL Test Variation D	0	0.00%	1	0.00%	4	0.00%
Any universal tool, desig. support, or accommodation or EL variation	124	0.16%	220	0.20%	242	0.22%
No universal tool, desig. support, or accommodation or EL variation	79,587	99.84%	109,493	99.80%	107,730	99.78%

Chapter 3: Item Development

The intact or modified intact test forms from previous test administrations from different years were used during the 2014 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 2014 forms.

The CST items were developed to measure California's content standards and designed to conform to principles of item writing defined by ETS (ETS, 2002). Each CST item on the intact or modified intact forms used in 2014 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

ETS maintained item development specifications for each CST 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 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. The specifications included the following:

- A full statement of each academic content standard, as defined by the SBE (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, if applicable, for 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 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 2014 CST administration, field-test items were repeated as a part of the intact or modified intact forms.

Selection of Item Writers

Criteria for Selecting Item Writers

The items for each CST were written by individual item writers with a thorough understanding of the California content standards. 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. 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 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 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 items and test forms—currently being reused—at the time the items and forms were developed. See Table 8.4 on page 91 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 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 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. The ARPs are advisory panels to the CDE and ETS and provided guidance on matters related to item development for the CSTs. 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. 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 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.

School administrators, 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.

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 Items

ETS content-area assessment specialists facilitated the CST 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
- Overview of the CSTs' test design specifications and blueprints
- Analysis of the CSTs' item specifications
- 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
- 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 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 *EC* Section 60602. 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.

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 or modified intact forms are being used with the field-test items intact for the 2014 CAASPP administration, data were not analyzed for current field-test items.

Stand-alone Field Testing

For each new CST 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, examinees 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 2014 administration, the original field-test items remained in their original positions in the intact or modified intact forms. Data were not analyzed for field-test items. The numbers of embedded field test items for the CSTs are not presented in this report because for the 2014 administration, field-test items were repeated as a part of the intact or modified intact forms and there was no new item development.

Allocation of Students to Forms

The test forms for a given CST 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 IRT 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, CMA, CAPA, and STS). 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 CSTs were constructed to measure students' performance relative to California's content standards approved by the SBE. They were also constructed to meet professional standards for validity and reliability. For each CST, 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 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 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 Items and Estimated Time Chart on page 19.

Rules for Item Selection

Test Blueprint

All test items on CST forms were selected to conform to the SBE-approved California content standards and test blueprints. The content blueprints for the CSTs can be found on the 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 items are grouped into subcontent areas (reporting clusters). For each CST reporting cluster, the percentage of questions correctly answered is reported on a student's score report. A list of the CST 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 or modified intact test forms from different years of administration were used during the 2014 administration. Prior to the 2013 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.

Replacement Items

During the 2013 STAR administration, certain test items on the grade ten CST for Life Science were exposed on social media Web sites. Although there was no evidence these exposures had any measurable effect on 2013 test results, the most cautious approach called for replacing these two items on the Life Science test forms in the 2014 CAASPP administration.

Before replacing items, ETS considered whether the form with an exposed item could be exchanged with a form from a different previous administration that contained no exposures. If there were no appropriate forms available, ETS sought to replace exposed items with previously used or field-tested items that best met the psychometric criteria and statistical targets described in the next sections. As CSTs would be pre-equated, ETS assessment development staff also considered minimizing the position difference between the replacement items and the exposed items in order to reduce the context effect.

Psychometric Criteria

The three goals of CST 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. These rules are referred to as statistical targets.

Two types of assembly targets were developed for each CST: 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 forms for the intact or modified intact forms used in the 2014 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 the total test are presented in Figure 4.A.1 on page 43. These curves present the target TIF and the projected TIF for the total test.

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 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 Test Assembly

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

* CSTs named by number only are grade-level tests.

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 44 present the target and projected information curves for the reporting clusters in the administered tests.

Projected Psychometric Properties of the Assembled Tests

Prior to the 2013 administration, ETS psychometricians performed a preliminary review of the technical characteristics of the assembled tests. The expected or projected performance of examinees and the overall score reliability were estimated using the item-level statistics available in the California item bank for the selected items. The test reliability was based on Gulliksen's formula (Gulliksen, 1987) for estimating test reliability (r_{xx}) from item *p*-values and item point-biserial correlations:

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

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 were calculated by summing the item *p*-values, and estimated test raw score standard deviations were calculated by summing the item reliability indices. Table 4.A.1 on page 43 presents these summary values by grade.

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 2014 CST administration.

Table 4.A.2 on page 43 shows the mean observed statistics of the items for each CST 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 91 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 2014 CST Projected Raw Score Statistics

Content Area	CST *	Number of Op. Items	Mean Raw Score	Std. Dev. of Raw Scores	Reliability
Science	5	60	40.09	11.60	0.92
	8	60	39.24	11.27	0.92
	10 Life Science	60	36.32	12.16	0.93

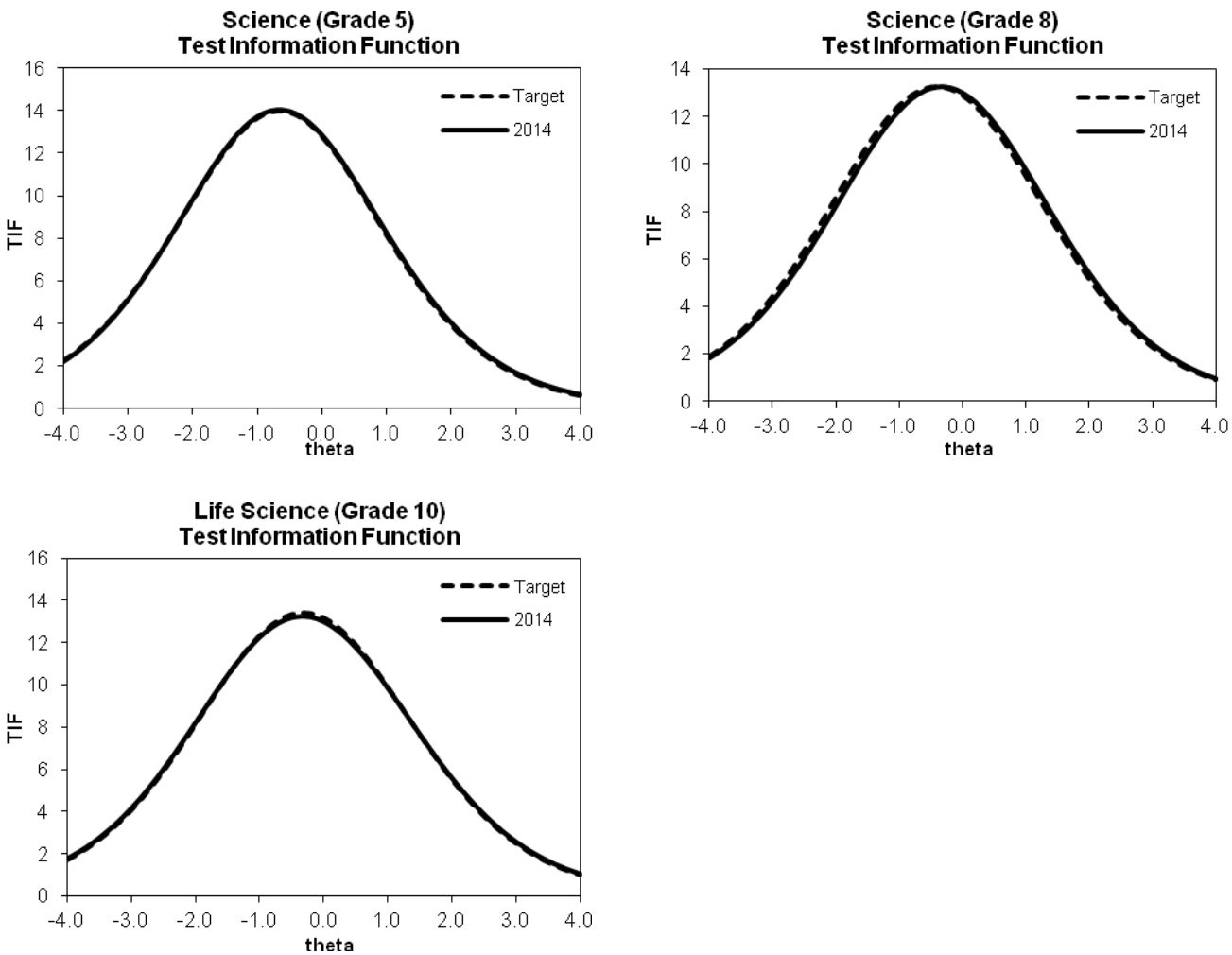
* CSTs named by number only are grade-level tests.

Table 4.A.2 Summary of 2014 CST Projected Item Statistics

Content Area	CST *	Mean b	SD b	Mean *p*–value	Min *p*–value	Max *p*–value	Mean Point Biserial	Min Point Biserial
Science	5	–0.64	0.53	0.70	0.46	0.85	0.43	0.28
	8	–0.34	0.75	0.70	0.43	0.95	0.43	0.08
	10 Life Science	–0.29	0.73	0.63	0.31	0.85	0.44	0.27

* CSTs named by number only are grade-level tests.

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

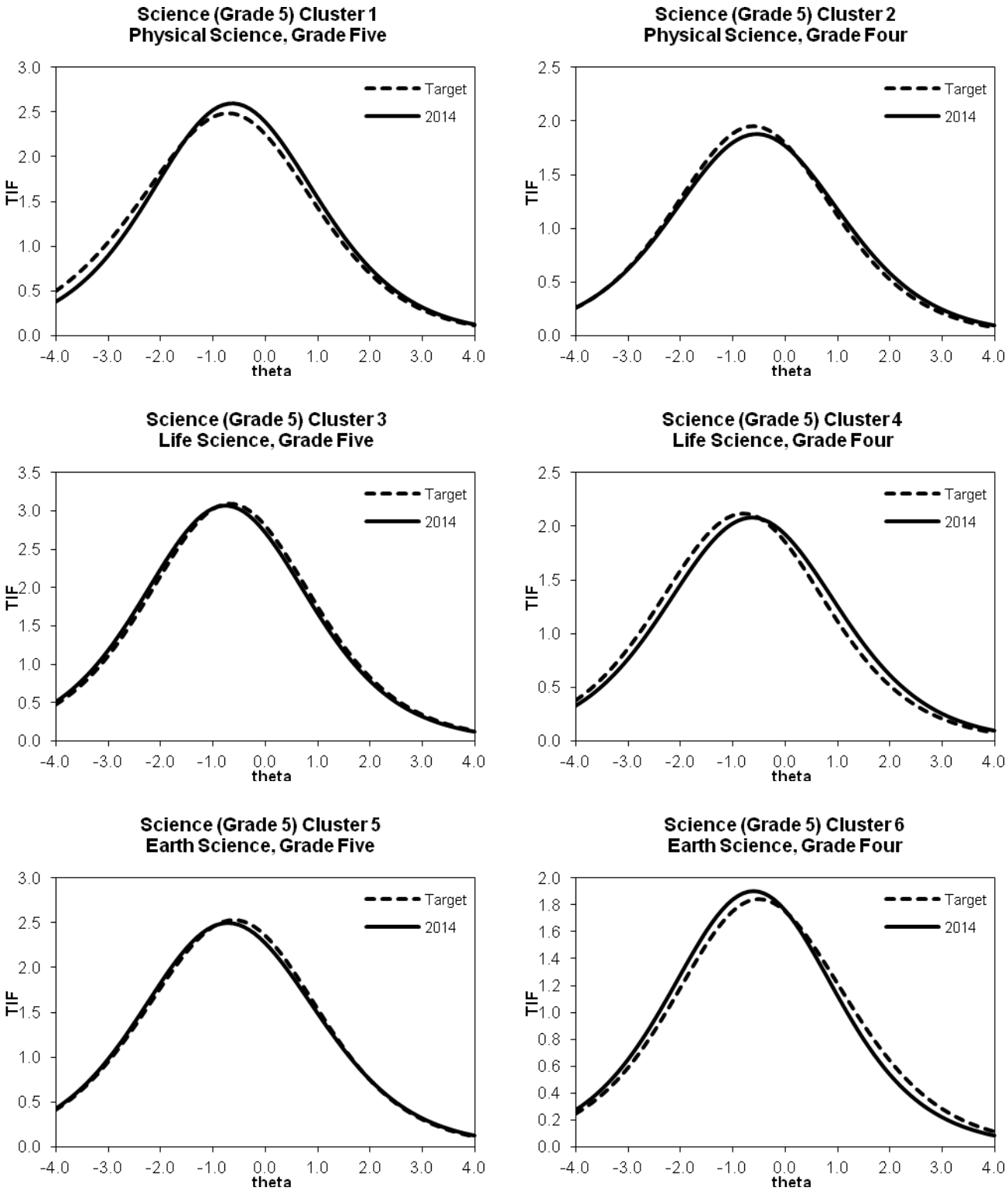


Figure 4.B.2 Plots of Target Information Functions and Projected Information for Clusters for Science, Grade Eight

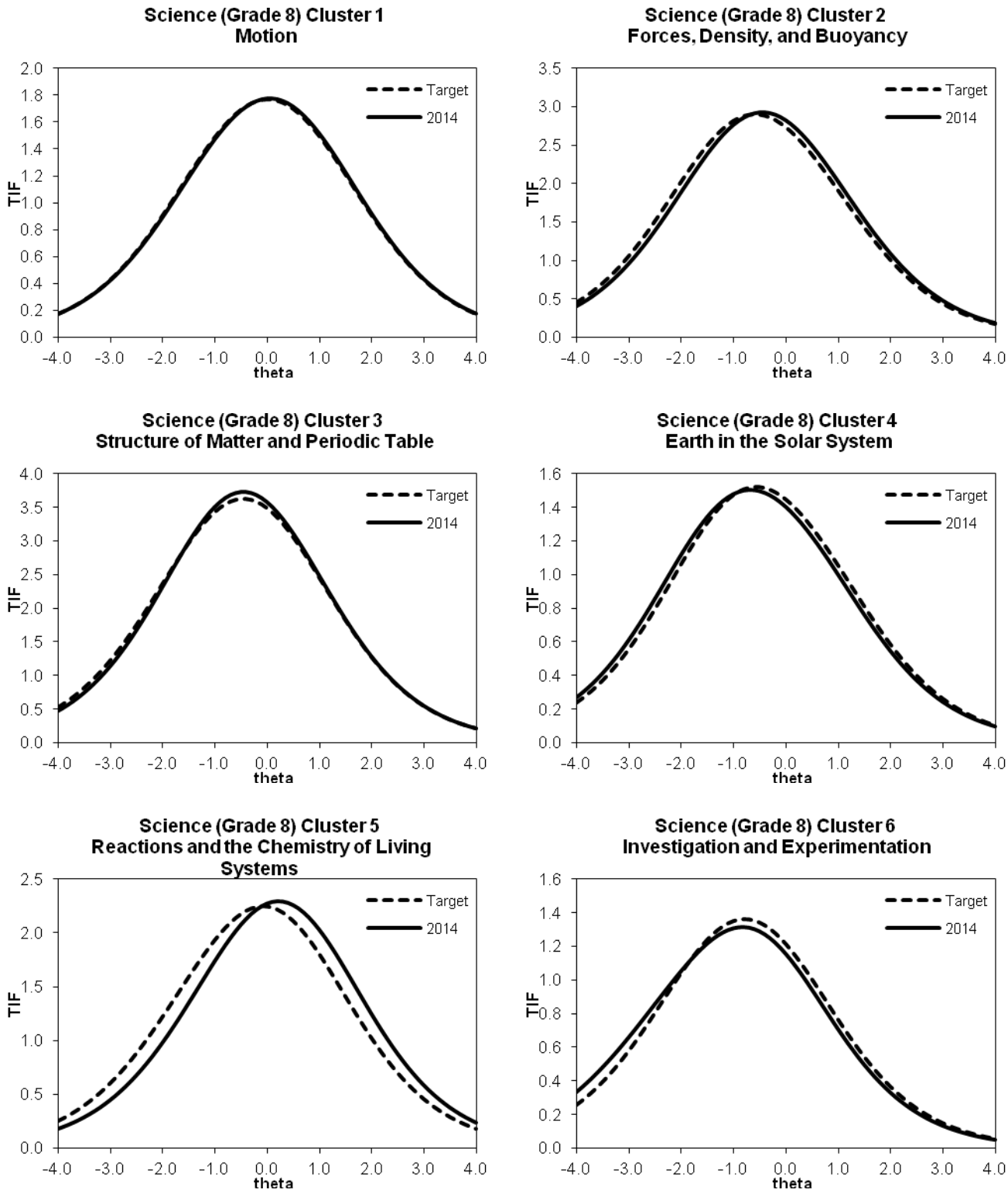
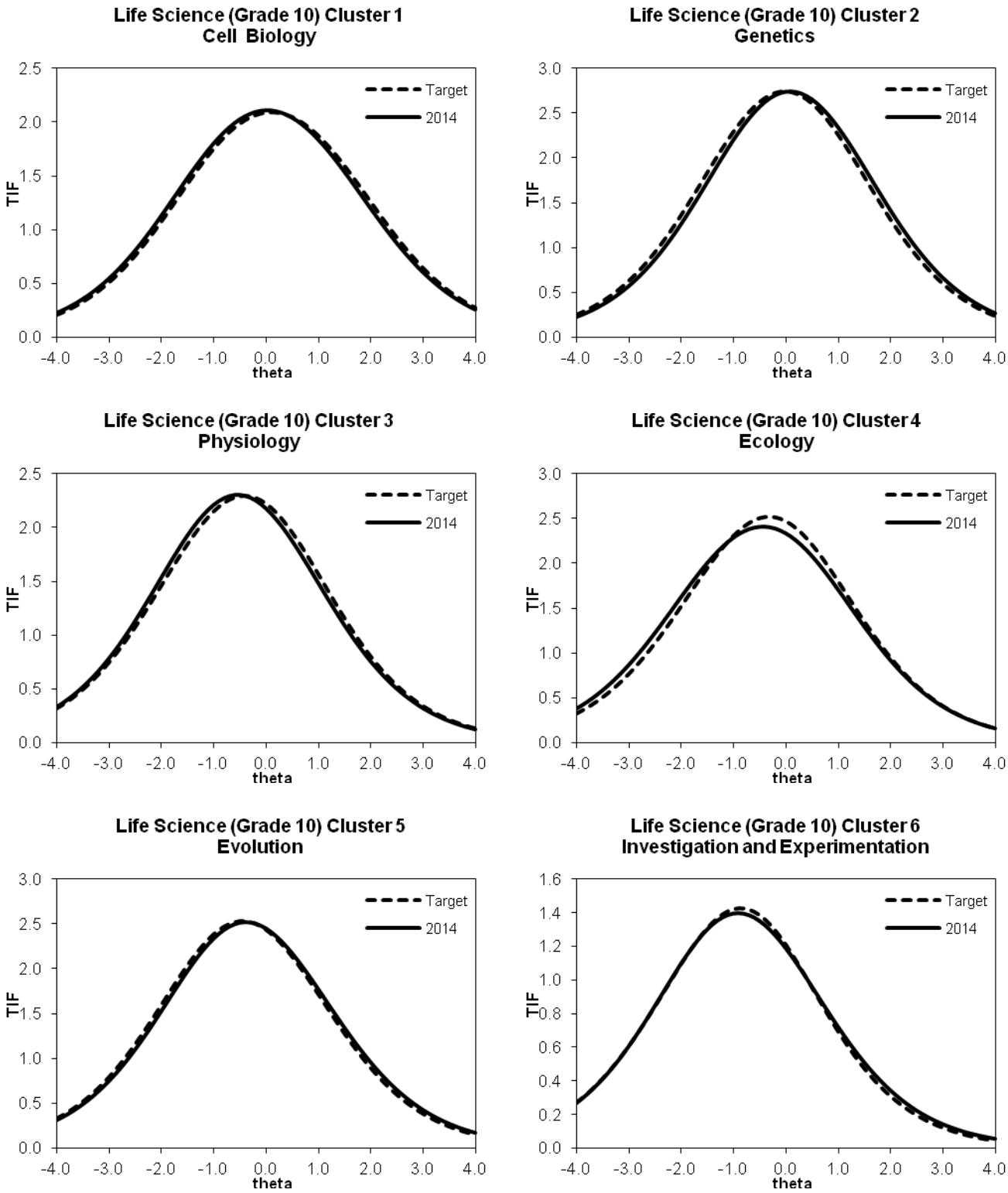


Figure 4.B.3 Plots of Target Information Functions and Projected Information for Clusters for Life Science, Grade Ten



Chapter 5: Test Administration

Test Security and Confidentiality

All tests within the CAASPP System are secure documents. For the CST administration, every person having access to testing 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), 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

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, ARP meetings were not held in 2014 because there was no new item development for the 2014 CST forms. However, before the 2014 administration,

ETS facilitated ARP meetings every year to review all newly developed CST 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 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 Management System (CDE, 2014a) remain protected by the ETS firewall software at all times. Due to the sensitive nature of the student

information processed by the Test Management System, 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 Pearson and other authorized locations. Pearson distributes the booklets to LEAs in securely packaged boxes.

Test Administration

Pearson receives testing materials from printers, packages them, and sends them to LEAs. After testing, the LEAs return materials to Pearson for scoring. During these events, Pearson takes extraordinary measures to protect the testing materials. Pearson's customized Oracle business applications verify that inventory controls are in place, from materials receipt to packaging. The reputable carriers used by Pearson 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 Pearson. Pearson 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 Pearson identification badges at all times in Pearson 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

Pearson provides overall security for assessment materials through its limited-access facilities and through its secure data processing capabilities. Pearson 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 Pearson 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 Pearson 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, Pearson utilizes Random Access Control Facility (RACF) to limit and control access to all data files (test and production), source code, object code, databases, and tables. RACF controls 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, Pearson uses ChangeMan, a mainframe configuration management tool, to control versions of the software and data files. ChangeMan provides another level of security, combined with RACF, to place the correct tested version of code into production. Unapproved changes are not implemented without prior review and approval.

Transfer of Scores via Secure Data Exchange

After scoring is completed, Pearson sends scored data files to ETS using secure data exchange procedures. ETS and Pearson have implemented procedures and systems to provide efficient coordination of secure data exchange. This includes the established SFTP site that is used for secure data transfers between ETS and Pearson. These well-established procedures provide timely, efficient, and secure transfer of data. Access to the CAASPP data files is limited to appropriate personnel with direct project responsibilities.

Statistical Analysis

The Information Technology (IT) area at ETS retrieves the Pearson data files from the SFTP site and loads them into a database. The Data Quality Services (DQS) area 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 on student data, the following deliverables are produced:

- Paper reports, some with individual student results and others with summary results
- A file of individual student results—available for download through the electronic reporting function of the Test Management System's Quick-turnaround Reporting (QTR) module—that shows students' scale scores and performance levels
- 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 ESEA 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 (NSLP), and so forth (CDE, 2014b). 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, including the Pre-ID files and reports.

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 Computer Processing Center is controlled by employee and visitor identification badges. The Center is secured by doors that can only be unlocked by the badges of personnel who have functional responsibilities within its secure perimeter. Authorized personnel accompany visitors to the Data 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 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, as described in this section.

Test Administrators

The CSTs 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 LEA and Test Site Coordinator Manual* (CDE, 2014c). 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. (*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 examiners, preparing and packaging materials to be returned to the LEA after testing, and returning the materials to the LEA. (CDE, 2014c)

Test Examiner

CSTs are administered by test examiners who may be assisted by test proctors and scribes. A test examiner 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 examiners must follow the directions in the *California Standards Tests Directions for Administration* (*DFA*) (CDE, 2014d) exactly.

Test Proctor

A test proctor is an employee of an LEA or a person, assigned by an NPS to implement the IEP of a student, who has received training designed to prepare the proctor to assist the test examiner 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

CST *DFA*s are manuals used by test examiners to administer the CSTs to students (CDE, 2014d). Test examiners must follow all directions and guidelines and read, word-for-word, the instructions to students in "SAY" boxes to ensure test standardization.

CAASPP LEA and Test Site Coordinator Manual

Test administration procedures are to be followed exactly so that all students have an equal opportunity to demonstrate their academic achievement. The *CAASPP LEA and Test Site Coordinator 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, 2014c). However, the manual is

not intended as a substitute for the *CCR*, Title 5, Education (*5 CCR*), or to detail all of the coordinator's responsibilities.

Test Management System Manuals

The Test Management System is a series of secure, Web-based modules that allow LEA CAASPP Coordinators to set up test administrations, order materials, and submit and correct student Pre-ID data. Every module has its own user manual with detailed instructions on how to use the Test Management System. The modules of the Test Management System are as follows:

- **Test Administration Setup—**This module allows LEAs to determine and calculate dates for scheduling test administrations for LEAs, to verify contact information for those LEAs, and to update the LEA's shipping information. (CDE, 2014e)
- **Order Management—**This module allows LEAs to enter quantities of testing materials for schools. Its manual includes guidelines for determining which materials to order. (CDE, 2014f)
- **Pre-ID—**This module allows LEAs to enter or upload student information, including some demographics, and to identify the test(s) the student will take. This information is printed on student answer documents or on labels that can be affixed to answer documents. Its manual includes the CDE's Pre-ID layout. (CDE, 2014b)
- **Extended Pre-ID Data Corrections—**This module allows LEAs to correct the data that were submitted during Pre-ID prior to the last day of the LEA's selected testing window. (CDE, 2014b)

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, depending on how the LEA CAASPP Coordinator ordered the materials. 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 under standard conditions. However, some students with disabilities and some ELs may need assistance when taking the CSTs. 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, test examiners 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 is to enable students to take the CSTs, not to give them an advantage over other students or to inflate their scores artificially. 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 examiner 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 and the CDE have determined that an irregularity or breach has occurred, the CDE instructs the LEA CAASPP Coordinator on how and where to identify the irregularity or breach on the student answer document. The information and procedures to assist in identifying incidents and notifying the CDE are provided in the *CAASPP LEA and Test Site Coordinator Manual* (CDE, 2014c).

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 in 2014 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, 2014d) and the *CAASPP LEA and Test Site Coordinator Manual* (CDE, 2014c). 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. (2014a). *2014 Test Management System.* Sacramento, CA. http://caaspp.org/administration/tms/

California Department of Education. (2014b). *2014 CAASPP Pre-ID and Extended Pre-ID Data Corrections instructions manual*. Sacramento, CA. Retrieved from http://www.startest.org/pdfs/calif-tac.pre-id_xdc_manual.2014.pdf

California Department of Education. (2014c). *2014 CAASPP LEA and test site coordinator manual*. Sacramento, CA. Retrieved from http://caaspp.org/rsc/pdfs/CAASPP.coord_man.2014.pdf

California Department of Education. (2014d). *2014 California Standardized Testing and Reporting directions for administration.* Sacramento, CA. Retrieved from http://caaspp.org/rsc/pdfs/CST.grade-8_dfa.2014.pdf

California Department of Education. (2014e). *2014 CAASPP Test Administration Setup manual*. Sacramento, CA. Retrieved from http://www.startest.org/pdfs/STAR.test_admin_setup.2014.pdf

California Department of Education. (2014f). *2014 CAASPP Order Management manual*. Sacramento, CA. Retrieved from http://www.startest.org/pdfs/calif-tac.order_mgmt.2014.pdf

Chapter 6: Performance Standards

Background

The CSTs for ELA and mathematics became part of California's standardized testing program in 1999; however, they are no longer administered starting in 2014. Five performance standards for the ELA tests were developed in 2000 and adopted by the SBE for the 2001 administration of those tests.

Also in 2001, the CSTs for history–social science and end-of-course (EOC) 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 2002. The performance standards for mathematics tests were established in 2001 and adopted in the 2002 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 science tests are all administered in 2014. The performance standards for the CSTs were defined by the SBE as far below basic, below basic, basic, proficient, and advanced.

The state target is to have all students achieve the proficient or advanced level by 2014. Schools and LEAs are expected to provide additional assistance to students scoring at or below the basic level.

California employed carefully designed standard-setting procedures to facilitate the development of performance standards for each CST. 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 standard setting 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

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

The standard-setting processes implemented for CSTs 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 examinees 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, 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, while the Bookmark method was used to set the performance standards for the history–social science, mathematics, and science CSTs. 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 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

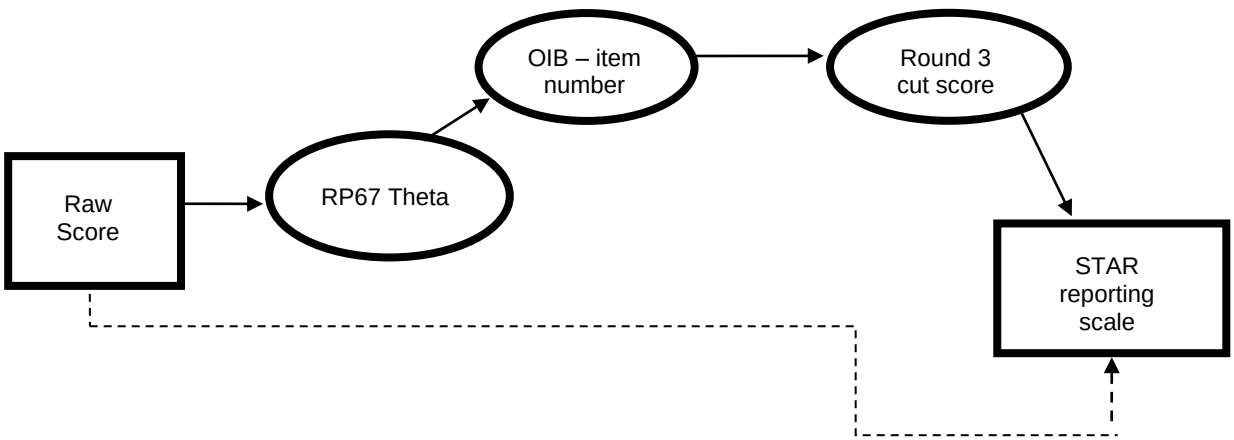
difficulty, the correct answer for each question, and where each question was located in the test booklet before the questions were reordered by difficulty.

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 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 66 in Chapter 7 presents the percentages of examinees meeting each performance level in 2014.

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.

Chapter 7: Scoring and Reporting

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 CSTs 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 2014 administration, because the raw-score-to-scale-score conversion tables were developed before tests were administered using pre-equating, individual student results were available for download prior to the printing of paper reports. This electronic reporting was made possible through the QTR module in the Test Management System.

In order to score and report CST 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:

- General Reporting Specifications—Provides the calculation rules for the information presented on CAASPP summary reports and defines the appropriate codes to use when a student does not take or complete a test or when a score will not be reported
- Score Key and Score Conversion—Defines file formats and information that is provided for scoring and the process of converting raw scores to scale scores
- Form Planner Specifications—Describes, in detail, the contents of files that contain keys required for scoring
- Aggregation Rules—Describes how and when a school's results are aggregated at the school, district, county, and state levels
- "What If" List—Provides a variety of anomalous scenarios that may occur when test materials are returned by LEAs to Pearson and defines the action(s) to be taken in response
- Edit Specifications—Describes edits, defaults, and solutions to errors encountered while data are being captured as answer documents are processed
- Reporting Cluster Names and Item Numbers—Identifies the reporting clusters for each test and the number of items in each cluster

The scoring specifications are reviewed and revised by the CDE, ETS, and Pearson 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 Pearson 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. Pearson 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 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 answered correctly.

Subscore

The items in each CST are aggregated into groups of related content standards to form reporting clusters. A subscore is a measure of an examinee's performance on the items in each reporting cluster. These results are reported both as raw scores and percent of items answered correctly. A description of the CST reporting clusters is provided in Appendix 2.B of Chapter 2, starting on page 20.

Scale Score

Raw scores obtained on each CST are transformed to three-digit scale scores using the equating process described in Chapter 2 on page 14. Scale scores range from 150 to 600 on each CST. The scale scores of examinees that have been tested in different years at a given grade level and content area can be compared. However, the raw scores of these examinees cannot be meaningfully compared, because these scores are affected by the relative difficulty of the test taken as well as the ability of the examinee.

Performance Levels

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

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

For all CSTs, 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 2014, 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 Pearson. ETS independently generates scores and verifies Pearson's 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 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 scores.

Individual Scores

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

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 examinees 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

				Scale Score		Raw Score		Raw Score Percent Correct	
Content Area	CST *	No. of Items	No. of Examinees	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Science	5	60	432,439	368	70	42.65	11.29	71.08	18.82
	8	60	435,173	392	94	42.23	10.81	70.38	18.01
	10 Life Science	60	435,705	360	65	40.04	11.41	66.73	19.01

* CSTs named by number only are grade-level tests.

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 examinees that 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 P1 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 July 29, 2014.

Table 7.2 Percentages of Examinees in Performance Levels for CSTs

Content Area	CST *	Far Below Basic	Below Basic	Basic	Proficient	Advanced	Proficient/ Advanced †
Science	5	6%	11%	24%	33%	27%	59%
	8	6%	9%	18%	22%	44%	66%
	10 Life Science	7%	11%	26%	28%	28%	56%

* CSTs named by number only are grade-level tests.

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

Table 7.A.1 in Appendix 7.A starting on page 71 shows the distributions of scale scores for each CST.

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.

Group Scores

Statistics summarizing student performance by each grade-level test for selected groups of students are provided starting on page 72 in Table 7.B.1 through Table 7.B.3 for the CSTs.

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. Students' economic status was determined by considering the education level of their parents and whether or not they participated in the National School Lunch Program (NSLP).

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 – 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 • EL • Reclassified fluent English proficient
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 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 reports. These categories and the specific reports in each category are given in Table 7.4.

Table 7.4 Types of CST Reports

1. Summary Reports	▪ CAASPP Student Master List Summary ▪ CAASPP Subgroup Summary (including Ethnicity for Economic Status)
2. Individual Reports	▪ CAASPP Student Record Label ▪ CAASPP Student Master List ▪ CAASPP Student Report for CSTs
3. Internet Reports	▪ CST Scores (state, county, LEA, school)

These reports are sent to the independently testing charter schools, counties, or school districts; the LEA forwards the appropriate reports to test sites or, in the case of the CAASPP Student Report, sends the report(s) to the child's parent or guardian and forwards a copy to the student's school or test site. Reports such as the CAASPP Student Report, Student Record Label, and Student Master List 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 summary reports, student record labels, and the master lists via electronic reporting, accessed using the QTR module to the Test Management System. This module permits LEAs to download a file containing student data that includes scale scores and performance levels for all tests taken.

Score Report Contents

The CAASPP Student Report provides scale scores, performance levels, and reporting cluster (subscore) 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. These same scale scores and performance levels are available in the LEA's electronic reporting file. In addition, percent-correct scores are provided at the cluster level.

Also given for each cluster is the average percent-correct for proficient students, which is presented as a range from the percent-correct score associated with the lowest proficient score on the total test to the percent-correct score associated with the lowest advanced score on the total test, less one percent. The average percent-correct estimates associated with the lowest proficient and advanced scores were obtained empirically for the tests that have sample sizes of 25 or more examinees at both the minimum proficient and the minimum advanced score levels. In cases where the available sample sizes were less than 25, "data smoothing" was conducted before obtaining the averages (Lu & Smith, 2009).

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 scores for these students are counted as far below basic, regardless of the scale score obtained.

Further information about the CAASPP Student Report and the other reports is provided in Appendix 7.C on page 78.

Score Report Applications

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

CST results help LEAs and schools identify strengths and weaknesses in their instructional programs. Each year, LEAs and school staffs examine CST 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 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 results (CDE, 2014a). 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, 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 or cluster 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 *2014 CAASPP Post-Test Guide* (CDE, 2014b).

References

California Department of Education. (2014a). *2014 CAASPP CST/CMA, CAPA, and STS printed reports*. Sacramento, CA. Retrieved from http://caaspp.org/rsc/pdfs/CAASPP.reports.2014.pdf

California Department of Education. (2014b). *2014 CAASPP post-test guide.* Sacramento, CA. Retrieved from http://caaspp.org/rsc/pdfs/CAASPP.post-test_guide.2014.pdf

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

Lu, Y., & Smith, R. L. (2009, April). *An alternative method to estimate cluster performance of proficient students on a large scale state assessment.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

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.

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

Scale Score	Grade 5 Science	Grade 8 Science	Grade 10 Life Science
570 – 600	3,656	17,821	1,640
540 – 569	7,901	11,011	4,072
510 – 539	N/A	25,833	N/A
480 – 509	11,155	28,913	14,567
450 – 479	28,096	30,647	20,625
420 – 449	48,017	46,840	37,886
390 – 419	47,322	59,862	55,253
360 – 389	84,592	54,544	69,031
330 – 359	68,712	46,363	85,438
300 – 329	60,070	44,819	67,436
270 – 299	41,427	25,886	47,836
240 – 269	24,938	21,981	24,268
210 – 239	5,902	11,605	6,915
180 – 209	603	6,710	680
150 – 179	48	2,338	58

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

				Percent in Performance Level					Mean Percent Correct in Reporting Cluster					
	Number Tested	Mean Scale Score	Std. Dev. of Scale Scores	Far Below Basic	Below Basic	Basic	Proficient	Advanced	Physical Science, Grade Five	Physical Science, Grade Four	Life Science, Grade Five	Life Science, Grade Four	Earth Science, Grade Five	Earth Science, Grade Four
All valid scores	432,439	368	70	6%	11%	24%	33%	27%	71%	69%	73%	71%	71%	70%
Male	216,610	370	72	7%	10%	23%	32%	29%	72%	69%	73%	72%	71%	71%
Female	215,513	365	67	6%	11%	26%	33%	25%	71%	69%	72%	71%	70%	69%
Gender unknown	316	327	65	20%	16%	29%	22%	12%	61%	57%	60%	59%	58%	56%
American Indian	2,354	352	67	9%	13%	27%	32%	19%	69%	64%	69%	68%	66%	65%
Asian American	41,018	414	77	2%	4%	12%	28%	53%	82%	80%	83%	82%	80%	83%
Pacific Islander	2,267	355	62	7%	12%	30%	33%	18%	70%	66%	69%	68%	68%	67%
Filipino	11,288	390	63	2%	5%	18%	39%	36%	79%	75%	79%	77%	77%	79%
Hispanic	227,446	346	60	9%	15%	30%	32%	15%	65%	63%	68%	66%	66%	63%
African American	24,822	342	62	11%	16%	29%	30%	14%	66%	63%	65%	64%	64%	62%
White	109,823	398	67	2%	4%	15%	36%	42%	80%	77%	81%	80%	78%	79%
Two or more races	13,421	390	72	4%	6%	17%	34%	39%	78%	75%	78%	78%	76%	77%
English only	242,022	380	69	4%	8%	21%	35%	33%	76%	72%	76%	75%	74%	74%
Initially fluent English prof.	20,590	409	72	1%	4%	14%	33%	48%	82%	80%	83%	82%	79%	81%
EL	88,537	313	50	17%	25%	35%	18%	4%	55%	53%	57%	55%	56%	52%
Reclassified fluent Eng. prof.	79,627	380	58	1%	5%	25%	42%	28%	75%	73%	78%	76%	75%	75%
English prof. unknown	1,663	337	64	14%	15%	29%	29%	13%	63%	59%	65%	64%	62%	60%
No special ed. services	407,504	370	70	6%	10%	24%	33%	27%	72%	70%	74%	72%	71%	71%
Special ed. services	23,490	338	70	16%	17%	26%	25%	16%	62%	60%	63%	63%	62%	60%
Special ed. unknown	1,445	340	63	13%	15%	29%	29%	14%	64%	60%	65%	64%	65%	62%
Not econ. disadvantaged	170,853	403	69	2%	4%	14%	35%	45%	81%	78%	82%	81%	79%	80%
Economically disadvantaged	258,335	345	61	9%	15%	30%	31%	14%	65%	63%	67%	65%	65%	63%
Unknown economic status	3,251	345	64	12%	13%	27%	32%	16%	66%	62%	66%	66%	64%	63%
Primary Ethnicity—Not Economically Disadvantaged														
American Indian	809	382	68	4%	6%	20%	39%	31%	77%	73%	77%	76%	74%	74%
Asian American	26,277	435	72	1%	2%	7%	25%	65%	87%	84%	87%	87%	84%	88%
Pacific Islander	838	377	64	4%	6%	24%	36%	29%	75%	72%	76%	74%	74%	74%
Filipino	7,028	400	62	1%	3%	15%	39%	42%	81%	78%	82%	80%	79%	82%
Hispanic	40,593	375	64	4%	8%	22%	38%	29%	74%	72%	76%	74%	73%	73%
African American	6,942	366	64	6%	10%	25%	35%	25%	73%	69%	72%	71%	71%	70%
White	79,895	409	64	1%	3%	12%	35%	49%	83%	80%	83%	83%	81%	82%
Two or more races	8,471	411	67	2%	3%	12%	34%	50%	83%	80%	83%	83%	81%	83%

				Percent in Performance Level					Mean Percent Correct in Reporting Cluster					
	Number Tested	Mean Scale Score	Std. Dev. of Scale Scores	Far Below Basic	Below Basic	Basic	Proficient	Advanced	Physical Science, Grade Five	Physical Science, Grade Four	Life Science, Grade Five	Life Science, Grade Four	Earth Science, Grade Five	Earth Science, Grade Four
Primary Ethnicity—Economically Disadvantaged														
American Indian	1,513	337	60	12%	17%	30%	28%	12%	64%	60%	64%	64%	62%	60%
Asian American	14,546	378	70	5%	8%	22%	34%	31%	74%	71%	76%	74%	73%	74%
Pacific Islander	1,413	342	58	8%	15%	34%	31%	12%	67%	63%	65%	64%	64%	63%
Filipino	4,204	373	61	3%	8%	24%	39%	26%	75%	70%	75%	73%	73%	74%
Hispanic	185,691	339	58	10%	16%	32%	30%	12%	63%	62%	66%	64%	64%	61%
African American	17,501	333	59	13%	18%	31%	28%	10%	63%	60%	63%	61%	61%	59%
White	28,981	366	62	5%	9%	25%	38%	23%	73%	69%	73%	72%	71%	69%
Two or more races	4,486	357	64	7%	12%	26%	35%	20%	70%	66%	70%	69%	68%	66%
Primary Ethnicity—Unknown Economic Status														
American Indian	32	341	48	6%	16%	34%	41%	3%	67%	62%	67%	64%	64%	64%
Asian American	195	365	74	12%	10%	17%	28%	33%	72%	69%	73%	72%	67%	68%
Pacific Islander	16	335	50	13%	6%	38%	38%	6%	65%	55%	70%	67%	60%	59%
Filipino	56	380	58	2%	7%	18%	43%	30%	75%	74%	77%	78%	73%	77%
Hispanic	1,162	328	57	16%	17%	31%	28%	8%	60%	57%	62%	61%	59%	58%
African American	379	328	56	15%	18%	31%	28%	7%	63%	59%	60%	62%	58%	58%
White	947	370	64	6%	8%	20%	40%	26%	74%	69%	73%	74%	72%	71%
Two or more races	464	339	63	13%	15%	30%	28%	14%	64%	61%	65%	64%	64%	60%

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

				Percent in Performance Level					Mean Percent Correct in Reporting Cluster					
	Number Tested	Mean Scale Score	Std. Dev. of Scale Scores	Far Below Basic	Below Basic	Basic	Proficient	Advanced	Motion	Forces, Density, and Buoyancy	Structure of Matter and Periodic Table	Earth in the Solar System	Reactions and the Chemistry of Living Systems	Investigation and Experimentation
All valid scores	435,173	392	94	6%	9%	18%	22%	44%	64%	71%	72%	74%	62%	80%
Male	218,535	398	99	7%	8%	16%	21%	47%	66%	73%	73%	76%	62%	80%
Female	216,268	386	89	6%	10%	20%	24%	41%	63%	70%	72%	72%	62%	81%
Gender unknown	370	339	90	18%	15%	25%	21%	22%	57%	60%	59%	65%	53%	70%
American Indian	2,604	369	90	10%	12%	21%	24%	33%	60%	67%	68%	71%	57%	76%
Asian American	42,016	461	94	2%	3%	8%	14%	73%	75%	83%	84%	83%	76%	91%
Pacific Islander	2,484	376	84	6%	10%	22%	26%	36%	62%	69%	70%	72%	59%	78%
Filipino	12,264	426	85	2%	4%	12%	22%	60%	69%	78%	80%	81%	69%	87%
Hispanic	225,170	364	83	9%	13%	23%	25%	31%	60%	66%	67%	70%	56%	76%
African American	26,215	352	85	12%	15%	24%	22%	27%	58%	63%	64%	66%	55%	73%
White	111,777	427	90	3%	5%	12%	20%	60%	70%	78%	79%	80%	69%	87%
Two or more races	12,643	414	97	5%	6%	14%	21%	54%	68%	75%	76%	78%	67%	84%
English only	239,481	405	94	5%	7%	16%	22%	50%	66%	74%	75%	77%	65%	83%
Initially fluent English prof.	30,837	432	94	3%	4%	12%	20%	61%	71%	78%	80%	80%	70%	87%
EL	52,985	306	70	23%	25%	28%	16%	9%	51%	53%	52%	58%	44%	62%
Reclassified fluent Eng. prof.	109,439	396	82	3%	7%	20%	27%	44%	65%	73%	74%	76%	63%	83%
English prof. unknown	2,431	353	90	14%	14%	21%	24%	27%	59%	63%	64%	68%	54%	73%
No special ed. services	413,046	396	93	6%	9%	18%	22%	45%	65%	72%	73%	75%	63%	81%
Special ed. services	22,091	330	91	21%	19%	23%	18%	20%	54%	57%	58%	64%	49%	65%
Special ed. unknown	36	302	75	31%	25%	11%	19%	14%	49%	52%	51%	60%	46%	56%
Not econ. disadvantaged	179,277	433	92	3%	4%	11%	19%	62%	71%	79%	80%	81%	70%	87%
Economically disadvantaged	252,421	364	85	9%	13%	23%	24%	31%	60%	66%	67%	70%	57%	76%
Unknown economic status	3,475	349	88	14%	15%	22%	23%	26%	58%	62%	63%	67%	53%	72%
Primary Ethnicity—Not Economically Disadvantaged														
American Indian	1,002	402	91	5%	7%	16%	25%	47%	65%	74%	75%	77%	63%	83%
Asian American	26,319	485	87	1%	1%	5%	11%	82%	79%	87%	88%	86%	80%	93%
Pacific Islander	985	403	86	4%	6%	16%	24%	50%	66%	74%	76%	76%	64%	83%
Filipino	7,740	439	83	1%	3%	9%	20%	67%	71%	81%	83%	82%	72%	89%
Hispanic	44,342	395	86	5%	8%	18%	24%	46%	65%	72%	74%	76%	63%	82%
African American	8,373	378	88	8%	10%	20%	23%	39%	62%	69%	70%	72%	60%	78%
White	82,637	441	87	2%	3%	10%	19%	67%	72%	81%	82%	82%	72%	89%
Two or more races	7,879	441	91	2%	3%	10%	18%	66%	72%	81%	82%	82%	72%	89%
Primary Ethnicity—Economically Disadvantaged														
American Indian	1,581	349	83	13%	14%	24%	24%	25%	58%	62%	63%	68%	54%	72%
Asian American	15,557	422	93	4%	6%	13%	20%	58%	69%	77%	78%	78%	69%	86%
Pacific Islander	1,487	359	79	8%	13%	26%	27%	26%	58%	65%	66%	69%	56%	75%
Filipino	4,475	404	83	3%	6%	17%	25%	49%	66%	75%	76%	77%	65%	83%
Hispanic	179,454	357	81	10%	14%	24%	25%	28%	59%	65%	65%	69%	55%	75%
African American	17,507	340	81	14%	17%	26%	22%	22%	56%	60%	62%	64%	53%	70%
White	28,322	386	87	6%	9%	19%	25%	42%	64%	71%	72%	74%	61%	80%
Two or more races	4,038	374	89	9%	10%	20%	25%	36%	61%	67%	69%	72%	59%	77%

				Percent in Performance Level					Mean Percent Correct in Reporting Cluster					
	Number Tested	Mean Scale Score	Std. Dev. of Scale Scores	Far Below Basic	Below Basic	Basic	Proficient	Advanced	Motion	Forces, Density, and Buoyancy	Structure of Matter and Periodic Table	Earth in the Solar System	Reactions and the Chemistry of Living Systems	Investigation and Experimentation
Primary Ethnicity—Unknown Economic Status														
American Indian	21	339	70	14%	10%	38%	24%	14%	56%	59%	63%	68%	49%	73%
Asian American	140	389	102	10%	7%	16%	29%	39%	66%	72%	70%	70%	62%	79%
Pacific Islander	12	343	58	8%	8%	42%	33%	8%	58%	63%	68%	61%	49%	69%
Filipino	49	402	99	6%	8%	18%	16%	51%	67%	72%	74%	78%	64%	80%
Hispanic	1,374	339	83	15%	17%	24%	23%	21%	57%	60%	60%	65%	51%	70%
African American	335	325	85	22%	17%	23%	21%	16%	52%	57%	58%	62%	48%	66%
White	818	372	86	8%	11%	20%	25%	36%	62%	68%	68%	72%	58%	78%
Two or more races	726	344	92	17%	16%	20%	23%	24%	57%	60%	61%	67%	52%	69%

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

				Percent in Performance Level					Mean Percent Correct in Reporting Cluster					
	Number Tested	Mean Scale Score	Std. Dev. of Scale Scores	Far Below Basic	Below Basic	Basic	Proficient	Advanced	Cell Biology	Genetics	Physiology	Ecology	Evolution	Investigation and Experimentation
All valid scores	435,705	360	65	7%	11%	26%	28%	28%	60%	60%	70%	69%	69%	78%
Male	219,630	363	69	9%	11%	24%	27%	30%	62%	60%	71%	70%	69%	78%
Female	215,594	357	61	6%	11%	29%	29%	25%	58%	61%	69%	68%	69%	79%
Gender unknown	481	320	58	20%	21%	30%	18%	11%	47%	47%	59%	56%	55%	67%
American Indian	3,026	348	62	10%	13%	28%	27%	21%	56%	56%	67%	66%	65%	76%
Asian American	39,488	401	70	3%	5%	15%	26%	51%	73%	71%	80%	79%	79%	85%
Pacific Islander	2,469	344	60	10%	14%	31%	28%	18%	56%	56%	65%	64%	63%	75%
Filipino	13,312	379	58	2%	6%	22%	33%	37%	66%	66%	77%	75%	76%	83%
Hispanic	223,128	341	57	10%	15%	32%	27%	16%	54%	55%	64%	63%	63%	75%
African American	26,483	334	57	13%	16%	32%	25%	14%	52%	53%	63%	60%	60%	71%
White	114,981	385	64	4%	5%	18%	31%	42%	67%	67%	78%	77%	77%	84%
Two or more races	12,818	375	67	6%	8%	20%	30%	36%	64%	64%	75%	74%	74%	81%
English only	242,707	370	65	6%	8%	23%	30%	33%	62%	63%	74%	72%	72%	80%
Initially fluent English prof.	36,294	381	66	3%	6%	22%	31%	38%	66%	66%	77%	75%	76%	83%
EL	46,333	297	42	27%	31%	31%	9%	2%	43%	42%	48%	46%	44%	62%
Reclassified fluent Eng. prof.	107,643	357	55	4%	10%	33%	32%	22%	59%	61%	70%	68%	69%	80%
English prof. unknown	2,728	333	64	16%	17%	28%	22%	17%	52%	51%	62%	60%	59%	71%
No special ed. services	410,298	363	64	6%	10%	26%	29%	29%	61%	61%	71%	70%	70%	79%
Special ed. services	25,357	311	58	26%	23%	28%	14%	9%	47%	44%	54%	52%	51%	64%
Special ed. unknown	50	297	50	34%	26%	26%	10%	4%	44%	41%	45%	49%	44%	60%
Not econ. disadvantaged	191,637	385	66	4%	6%	19%	30%	42%	67%	67%	78%	77%	77%	83%
Economically disadvantaged	240,183	340	58	10%	15%	32%	26%	16%	54%	55%	64%	63%	63%	75%
Unknown economic status	3,885	332	64	17%	17%	27%	21%	17%	52%	51%	63%	60%	58%	71%
Primary Ethnicity—Not Economically Disadvantaged														
American Indian	1,243	369	64	5%	9%	22%	31%	33%	62%	62%	74%	73%	72%	80%
Asian American	24,087	418	68	1%	3%	10%	24%	62%	78%	75%	85%	84%	84%	87%
Pacific Islander	1,100	359	62	7%	10%	28%	30%	26%	59%	61%	70%	69%	69%	78%
Filipino	8,738	387	57	1%	4%	19%	34%	42%	69%	69%	79%	77%	79%	84%
Hispanic	51,650	359	60	6%	10%	27%	31%	26%	59%	60%	71%	70%	70%	79%
African American	9,762	349	59	9%	12%	29%	29%	21%	57%	58%	68%	66%	66%	75%
White	86,993	394	63	2%	4%	15%	31%	48%	69%	69%	81%	80%	80%	85%
Two or more races	8,064	392	64	3%	4%	16%	31%	46%	69%	68%	80%	79%	79%	84%
Primary Ethnicity—Economically Disadvantaged														
American Indian	1,755	333	56	14%	16%	32%	25%	13%	52%	51%	63%	61%	60%	72%
Asian American	15,255	373	64	4%	8%	23%	30%	35%	66%	65%	73%	72%	72%	81%
Pacific Islander	1,349	333	56	12%	17%	34%	26%	12%	54%	53%	61%	60%	59%	72%
Filipino	4,525	364	57	4%	9%	28%	32%	27%	62%	62%	72%	69%	71%	80%
Hispanic	169,819	336	55	11%	16%	34%	26%	14%	53%	54%	63%	61%	61%	74%
African American	16,369	325	54	15%	19%	34%	22%	10%	49%	51%	60%	57%	57%	69%
White	27,050	357	61	7%	10%	26%	30%	25%	59%	59%	70%	69%	69%	78%
Two or more races	4,061	349	60	10%	12%	28%	29%	21%	57%	57%	67%	66%	66%	76%

				Percent in Performance Level					Mean Percent Correct in Reporting Cluster					
	Number Tested	Mean Scale Score	Std. Dev. of Scale Scores	Far Below Basic	Below Basic	Basic	Proficient	Advanced	Cell Biology	Genetics	Physiology	Ecology	Evolution	Investigation and Experimentation
Primary Ethnicity—Unknown Economic Status														
American Indian	28	312	59	21%	25%	29%	14%	11%	42%	44%	55%	56%	51%	66%
Asian American	146	344	68	15%	14%	25%	23%	23%	56%	54%	66%	64%	61%	74%
Pacific Islander	20	307	49	30%	15%	25%	30%	0%	47%	38%	54%	54%	51%	65%
Filipino	49	357	62	6%	12%	20%	37%	24%	59%	56%	71%	70%	69%	78%
Hispanic	1,659	320	59	21%	20%	30%	18%	12%	48%	48%	58%	55%	54%	68%
African American	352	316	56	23%	20%	30%	18%	10%	48%	48%	57%	53%	53%	63%
White	938	359	66	9%	11%	23%	27%	30%	59%	57%	73%	70%	68%	78%
Two or more races	693	331	62	15%	20%	27%	23%	15%	51%	50%	62%	60%	59%	70%

Appendix 7.C—Types of Score Reports

Table 7.C.1 Score Reports Reflecting CST Results

2014 CAASPP CST Printed Reports	
Description	**Distribution**
The CST Student Report	
This report provides parents/guardians and teachers with the student's results, presented in tables and graphs. Data presented include the following: • Scale scores • Performance levels (advanced, proficient, basic, below basic, and far below basic) • Number and percent correct in each reporting cluster • Comparison of the student's scores on specific reporting clusters to the range of scores of students statewide who scored proficient on the total test	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 distributed by the LEA to parents/guardians.
Student Record Label	
These reports are printed on adhesive labels to be affixed to the student's permanent school records. Each student shall have an individual record of accomplishment that includes CAASPP testing results (see California *EC* Section 60607[a]). Data presented include the following for each content area tested: • Scale scores • Performance levels	This report includes individual student results and is not distributed beyond the student's school.
Student Master List	
This report is an alphabetical roster that presents individual student results. Data include the following: • Percent correct for each reporting cluster within each content area tested • A scale score and performance level for each for the grade-level test or tests taken	This report provides administrators and teachers with all students' results within each grade or within each grade and year-round schedule at a school. Because this report includes individual student results, it is not distributed beyond the student's school. It is recommended that Student Master List reports be retained until the grade level exits the school.

2014 CAASPP CST Printed Reports	
Description	**Distribution**
Student Master List Summary	
This report summarizes student results at the school, LEA, county, and state levels for each grade. It does not include any individual student information. For each CST, the following data are summarized: • By content area tested: – Number of students enrolled – Number and percent of students tested – Number and percent of valid scores – Number tested with scores – Mean percent correct • Mean scale score • Scale score standard deviation • Number and percent of students scoring at each performance level • The number of items for each reporting cluster and the mean percent correct	This report is a resource for evaluators, researchers, teachers, parents/guardians, community members, and administrators. One copy is packaged for the school and one for the LEA. This report is also produced for LEAs, counties, and the state. ***Note:*** The data in this report may be shared with parents/guardians, community members, and the media only if the data are for 11 or more students. It is recommended that summary reports be retained for at least five years.
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 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. One copy is packaged for the school and one for the LEA. This report is also produced for LEAs, counties, and the state. ***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. It is recommended that summary reports be retained for at least five years.

2014 CAASPP CST Printed Reports	
Description	**Distribution**
Subgroup Summary—Ethnicity for Economic Status	
This report, a part of the Subgroup Summary, disaggregates and reports results by cross-referencing each ethnicity with economic status. The economic status for each student is “economically disadvantaged,” “not economically disadvantaged,” or “economic status unknown.” A student is defined as “economically disadvantaged” if the most educated parent of the student, as indicated in CALPADS, has not received a high school diploma or the student is eligible to participate in the free or reduced-price lunch program also known as the National School Lunch Program (NSLP). As with the standard Subgroup Summary, this disaggregation contains no individual student-identifying information and is 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: • 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. One copy is packaged for the school, and one for the LEA. This report is also produced for LEAs, counties, and the state. ***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. It is recommended that summary reports be retained for at least five years.

Chapter 8: Analyses

Background

This chapter summarizes the item- and test-level statistics obtained for the CSTs administered during the spring of 2014 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 2014, 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 2014, because the intact or modified intact forms were used, DIF analyses were not performed.

Each of the sets of analyses is presented in the body of the text and in the appendixes as listed below.

1. Appendix 8.A on page 100 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 either intact or modified with replacement items, *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 82.
2. Appendix 8.B on page 101 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 109 presents the summaries of Rasch item difficulty statistics (*b*-values) for the operational items for the CST for Life Science (Grade 10) form, which contains replacement items. (For the summaries of *b*-values for the three science CSTs, 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 91 for administration years.) In addition, the appendix presents the scoring tables obtained as a result of the IRT equating process.

Samples Used for the Analyses

CST analyses were conducted at different times after test administration and involved varying proportions of the full CST 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 July 29, 2014, which comprised approximately 90–99 percent of the full CST data.

During the 2014 administration, neither IRT calibrations nor scaling are implemented because of the intact forms and pre-equating. The IRT results of the tests with items replaced due to security breaches are presented in Appendix 8.C. These IRT results are

derived based on FIA samples (the data which comprise approximately 90–99 percent of the full CST data) of the most recent administration when the items were used.

For the intact forms used without any replacement or edited items, 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 91 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 examinees 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 examinees that answered item i correctly, and

N_i is the total number of examinees that 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 percents, and mean scores, are presented in Table 8.A.1 on page 100. Each set of values is presented for both the current and the original presentation of each CST.

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

					Mean		Median	
Content Area	**CST***	**Admin. Year**	**No. of Items**	**No. of Examinees**	***p*-value**	**Pt-Bis**	***p*-value**	**Pt-Bis**
Science	5	2014	60	432,439	0.71	0.43	0.72	0.42
	8	2014	60	435,173	0.70	0.42	0.72	0.42
	10 Life Science	2014	60	435,705	0.67	0.42	0.68	0.42

* CSTs named by number only are grade-level tests.

T

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). 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 may 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 the 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 an examinee'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 an examinee's true score (Crocker & Algina, 1986). For example, if an examinee's observed score on a given test equals 15 points, and the SEM equals 1.92, one can be 95 percent confident that the examinee'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 examinees upon which those analyses were performed.

Table 8.2 Reliabilities and SEMs for the CSTs

					Scale Score			Raw Score		
Content Area	CST *	No. of Items	No. of Examinees	Reliab.	Mean	Std. Dev.	SEM	Mean	Std. Dev.	SEM
	5	60	432,439	0.92	368	70	19.48	42.65	11.29	3.14
Science	8	60	435,173	0.92	392	94	27.28	42.23	10.81	3.13
	10 Life Science	60	435,705	0.92	360	65	18.25	40.04	11.41	3.19

* CSTs named by number only are grade-level tests.

Intercorrelations, Reliabilities, and SEMs for Reporting Clusters

For each grade-level science CST, 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 101. 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 examinee 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.15 present the overall test reliabilities for the various subgroups. Table 8.B.16 through Table 8.B.20 present the cluster-level reliabilities for the subgroups. Table 8.B.16 presents the cluster-level reliabilities for the subgroups based on gender and economic status.

The next table, Table 8.B.17, shows the same analyses for the subgroups based on provision of special services and English-language fluency. Table 8.B.18 presents results for the subgroups based on primary ethnicity of the examinees. The last set of tables, Table 8.B.19 through Table 8.B.20, present the cluster-level reliabilities for the subgroups based on primary ethnicity within economic status.

Note that the reliabilities are reported only for samples that comprise 11 or more examinees. Also, in some cases, score reliabilities were not estimable and are presented in the tables as hyphens. Finally, results based on samples that contain 50 or fewer examinees should be interpreted with caution due to small sample sizes.

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. An examinee's CSEM under the IRT framework is equal to the inverse of the square root of the test information function:

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

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.

SEMs 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.

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

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

* CSTs named by number only are grade-level tests.

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 ETS-proprietary computer program RELCLASS-COMP (Version 4.14).

Decision accuracy describes the extent to which examinees 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 examinees 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 a form actually taken	
		Does not achieve a performance level	Achieves a performance level
True status on all-forms average	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 the alternate form taken	
		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.21 through Table 8.B.23 in Appendix 8.B, starting on page 107.

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], 1999; 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, 1999; 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 testing program. The description is organized in the manner prescribed by *The Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999). 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, followed by a description and discussion of the kinds of validity evidence that have been gathered.

Purpose of the CSTs

As mentioned in Chapter 1, the CSTs for Science comprise the CAASPP System's implementation of the remaining paper-pencil tests.

The Constructs to Be Measured

The CSTs for Science, given in English, are designed to show how well students in grades five, eight, and ten perform relative to the California content standards in science. These content standards were approved by the SBE; they describe what students should know and be able to do at each grade level.

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 examinee 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 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 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, student performance levels, and subscores for each reporting cluster 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. Reporting cluster results compare 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 is also provided for each cluster using a percent-correct metric. The reference points for this range are: (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 scores is presented in Chapter 7, which starts on page 63.

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, 1999, Standard 1.3). The user is advised to gather evidence to support these additional interpretations or uses (AERA, APA, & NCME, 1999, 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 in 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.

Validity Evidence Collected

Evidence Based on Content

According to *The Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999), 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* (1999), 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: ELA, 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.

Specifications and blueprints—ETS maintains item specifications for each CST. The item specifications describe the characteristics of the 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 the items were developed and field-tested, ETS selected all CST test items to conform to the SBE-approved California content standards and test blueprints. Test blueprints for the CSTs 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 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 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. In brief, items written for the CSTs 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 (refer to “Test Assembly” on page 10 in Chapter 2 for the replacement item selection rules). 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 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 38.

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.

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, 1999), 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 examinees 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 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 EOC 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 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, 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's 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

Content Area	CST *	Year
Science	5	2010
	8	2011
	10 Life Science	2010

* CSTs named by number only are grade-level tests.

Evidence Based on Response Processes

As noted in the AERA, APA, and NCME *Standards* (1999), additional support for a particular score interpretation or use can be provided by theoretical and empirical evidence indicating that examinees are using the intended response processes when responding to the items in a test. This evidence may be gathered from interacting with examinees in order to understand what processes underlie their item responses.

Evidence Based on Internal Structure

As suggested by the *Standards* (AERA, APA, & NCME, 1999), 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, 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 examinees' 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 of the 2014 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 examinee ability provides evidence that examinees 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 or modified intact forms were used originally. (The summaries of the IRT difficulty indicator *b*-values for operational items in the CST for Life Science, which included replacement items, can be found in Table 8.C.1. The distributions of the *b*-values for Life Science that contained replacement items are given in Table 8.C.2. For operational items in the three science tests, 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 91 for administration years.)

The data in Table 8.1 indicate that all of the CSTs had *p*-values with means ranging from 0.67 to 0.71.

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 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, 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 are also examined for various subgroups of the examinee 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.15.

Reliability of performance classifications—The methodology used for estimating the reliability of classification decisions is described in the section "Decision Classification Analyses" on page 85. The results of these analyses are presented in Table 8.B.21 through Table 8.B.23 in Appendix 8.B; these tables start on page 107. 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. Similarly, the proportion of students that were classified consistently was 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.

Dimensionality

Dimensionality analyses were conducted by a CDE psychometrics team (Gaffney et al., 2010; Gaffney & Perryman, 2009). The study investigated the factor structures of the CSTs in grades three and five as part of peer review for the Elementary and Secondary Education Act (ESEA).

Two factors corresponding to the ELA and mathematics domain were found for the CSTs in these grades, as would be expected, since these tests were designed to measure different constructs.

Evidence Based on Consequences of Testing

As observed in the *Standards*, tests are usually administered "with the expectation that some benefit will be realized from the intended use of the scores" (AERA, APA, & NCME, 1999, p. 18). 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.

IRT Analyses

Post-Equating

Prior to the 2013 administration, the CSTs 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 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 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 CSTs.

Pre-Equating

During the 2014 administration, because all the test forms were used in a previous operational administration, pre-equating was conducted prior to administration of the tests. Based on the sample invariant property of item response theory (IRT), all the item parameter estimates were placed on the reference scale in their previous administrations through the post-equating procedure described above. For all CST intact forms without any edits or replacement of items, the conversion tables from the previous administration when the forms were used originally are directly applied to the current administration. Otherwise, in the cases where replacement or edited items are used, conversion tables are generated using the true-score equating method described in Chapter 2, page 16. The item parameters used for true-score equating are post-equating item parameters from the intact forms for the unchanged items and the post-equating item parameters from the most recent administration for the replacement or edited items.

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's 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 is shown in Table 8.4.

The CST for Life Science has IRT *b*-value and distribution tables available in Appendix 8.C starting on page 109 because it contains two replacement items.

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 procedure prior to the 2013 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 2014 administration, for the CSTs in science for grades five and eight, the raw-to-scale score tables from the intact forms are directly applied. For the form for Life Science, due to two replaced items, the scaled *b*-values of these items from the most recent administration are adopted to develop the pre-equated score conversion tables. 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.

Table 8.C.1 presents univariate statistics (mean, standard deviation, minimum, and maximum) for the scaled IRT *b*-values for the tests containing replacement or edited items. The results for the overall test are presented for the operational items in each reporting cluster. (For operational items of the other tests, the summaries of *b*-values 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 91 for administration years.)

Table 8.C.2 shows the distributions of operational items for tests containing replacement items across 16 intervals of *b*-values. In these tables, the intervals range from "greater than or equal to 3.5" to "less than –3.5" points within each interval. (For operational items of the other tests, the distributions of *b*-values 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 91 for administration years.)

Evaluation of Pre-Equating

Pre-equating is performed on the basis of the assumption of item response theory (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 2014 administration, for all CSTs intact forms without any edits, the conversion tables from their original administrations (listed in Table 8.4 on page 91) are directly applied to the current administration. For operational CSTs with replacement items or edits—that is, the CST for Life Science (Grade 10)—the new conversion table was developed based on the item parameters that were used to develop the original conversion tables for the unchanged items, and the item parameters based on FIA samples of the most recent administration for the replacement. Specifically, IRT true-score equating procedures were used to transform the new form number-correct scores to their respective reference form number-correct scale. These converted raw scores were then transformed to scale scores through table lookup and linear interpolation. The reference conversions for the forms used in 2014 were taken from one of the 2004 or 2006 forms (see Chapter 10 starting on page 125 for specifics on the base years).

Complete raw-score-to-scale-score conversion tables for the CSTs administered in 2014 are presented in Table 8.C.3 through Table 8.C.5 starting on page 110. 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 2013 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 during the 2014 administration because forms are either intact or intact with some modification, 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 91 for administration years.

The statistical procedure of DIF analysis that was conducted prior to the 2013 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. Note, however, that analyses of ELs on the CST for ELA were presented for readers’ interest. Differential performance on an ELA item

because of the language difficulties of nonnative speakers did not indicate that an item was unfair or biased.

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.

Tables showed the distributions of operational and field-test items across the DIF category classifications for the CSTs. In these tables, classifications of B- or C- indicated DIF against a focal group; classifications of B+ and C+ indicate DIF in favor of a focal group. The last two columns of each table showed the total number of items flagged for DIF in one or more comparisons.

References

AERA, APA, & NCME. 1999. *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

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.

Gaffney, T., & Perryman, C. (2009, July). *A longitudinal look at the factor structure of educational achievement tests*. Paper presented at the meeting of the Psychometric Society, Cambridge, England.

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 (2014) and Original Year of Administration

Item-by-item *p*-value and Point Biserial for Science, Grades Five, Eight, and Ten												
CSTs	**Grade 5**				**Grade 8**				**Life Science (Grade 10)**			
Years	**2014**		**2010**		**2014**		**2011**		**2014**		**2010**	
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	0.81	0.30	0.80	0.33	0.44	0.06	0.43	0.08	0.79	0.26	0.73	0.33
2	0.76	0.37	0.72	0.39	0.69	0.36	0.67	0.37	0.66	0.41	0.63	0.41
3	0.69	0.36	0.67	0.36	0.83	0.26	0.82	0.32	0.78	0.41	0.75	0.45
4	0.73	0.45	0.71	0.47	0.80	0.40	0.79	0.43	0.66	0.34	0.64	0.36
5	0.76	0.33	0.73	0.34	0.54	0.41	0.52	0.42	0.70	0.41	0.65	0.45
6	0.81	0.37	0.79	0.41	0.57	0.33	0.56	0.35	0.43	0.41	0.42	0.39
7	0.78	0.41	0.77	0.41	0.72	0.44	0.69	0.45	0.47	0.38	0.43	0.36
8	0.65	0.39	0.62	0.40	0.56	0.34	0.56	0.35	0.73	0.46	0.67	0.49
9	0.61	0.44	0.57	0.44	0.52	0.50	0.51	0.49	0.62	0.34	0.59	0.37
10	0.82	0.40	0.80	0.42	0.74	0.48	0.73	0.51	0.33	0.42	0.31	0.37
11	0.56	0.47	0.56	0.47	0.86	0.49	0.85	0.52	0.77	0.37	0.70	0.44
12	0.69	0.44	0.66	0.45	0.63	0.42	0.63	0.43	0.43	0.35	0.41	0.35
13	0.50	0.44	0.48	0.43	0.96	0.35	0.94	0.42	0.58	0.40	0.53	0.40
14	0.74	0.41	0.73	0.43	0.70	0.43	0.69	0.45	0.84	0.28	0.81	0.32
15	0.67	0.45	0.64	0.47	0.72	0.33	0.71	0.35	0.57	0.43	0.54	0.46
16	0.71	0.36	0.72	0.38	0.79	0.33	0.77	0.37	0.43	0.27	0.41	0.27
17	0.78	0.51	0.76	0.53	0.66	0.44	0.67	0.47	0.51	0.55	0.47	0.52
18	0.58	0.37	0.58	0.38	0.83	0.33	0.82	0.36	0.58	0.35	0.55	0.37
19	0.73	0.52	0.69	0.54	0.87	0.45	0.85	0.47	0.76	0.41	0.70	0.44
20	0.67	0.51	0.65	0.51	0.70	0.55	0.67	0.57	0.66	0.50	0.65	0.51
21	0.69	0.46	0.66	0.46	0.82	0.40	0.80	0.44	0.71	0.33	0.68	0.37
22	0.89	0.46	0.84	0.52	0.79	0.48	0.77	0.50	0.83	0.43	0.78	0.46
23	0.80	0.43	0.78	0.44	0.44	0.42	0.47	0.43	0.42	0.36	0.38	0.36
24	0.77	0.43	0.73	0.43	0.74	0.43	0.72	0.46	0.64	0.45	0.58	0.45
25	0.76	0.52	0.76	0.52	0.69	0.44	0.66	0.46	0.45	0.39	0.45	0.39
26	0.64	0.41	0.61	0.43	0.80	0.45	0.79	0.48	0.46	0.37	0.45	0.39
27	0.82	0.37	0.79	0.39	0.76	0.49	0.76	0.50	0.67	0.41	0.64	0.43
28	0.63	0.33	0.62	0.35	0.74	0.50	0.71	0.51	0.75	0.41	0.74	0.45
29	0.76	0.46	0.74	0.47	0.48	0.44	0.49	0.43	0.72	0.54	0.69	0.55
30	0.65	0.40	0.60	0.40	0.66	0.50	0.65	0.51	0.79	0.40	0.75	0.44
31	0.79	0.48	0.78	0.47	0.86	0.45	0.85	0.48	0.82	0.46	0.79	0.51
32	0.52	0.38	0.52	0.35	0.70	0.40	0.69	0.40	0.83	0.50	0.81	0.54
33	0.80	0.52	0.79	0.52	0.78	0.41	0.76	0.44	0.88	0.38	0.83	0.44
34	0.69	0.39	0.70	0.38	0.65	0.49	0.63	0.46	0.64	0.33	0.60	0.35
35	0.85	0.41	0.83	0.40	0.55	0.44	0.54	0.42	0.64	0.45	0.62	0.45
36	0.71	0.46	0.69	0.47	0.78	0.48	0.76	0.50	0.76*	0.49*	0.75	0.51
37	0.64	0.40	0.63	0.38	0.72	0.44	0.71	0.46	0.82	0.44	0.76	0.50
38	0.57	0.53	0.56	0.52	0.88	0.49	0.85	0.52	0.52	0.35	0.51	0.36
39	0.72	0.52	0.67	0.53	0.78	0.40	0.78	0.42	0.76	0.46	0.72	0.48
40	0.80	0.41	0.79	0.43	0.80	0.49	0.79	0.51	0.50	0.44	0.47	0.41
41	0.77	0.45	0.74	0.46	0.55	0.45	0.53	0.46	0.85	0.35	0.83	0.42
42	0.59	0.41	0.60	0.44	0.56	0.38	0.55	0.41	0.72	0.55	0.66	0.56
43	0.70	0.37	0.71	0.37	0.81	0.51	0.80	0.52	0.65	0.41	0.61	0.43
44	0.81	0.50	0.81	0.49	0.54	0.33	0.51	0.32	0.87	0.45	0.85	0.48
45	0.60	0.33	0.60	0.33	0.79	0.40	0.76	0.40	0.61	0.46	0.56	0.47
46	0.79	0.43	0.78	0.45	0.88	0.42	0.86	0.44	0.77	0.45	0.73	0.50
47	0.59	0.28	0.57	0.29	0.68	0.43	0.67	0.43	0.80	0.36	0.75	0.42
48	0.81	0.42	0.83	0.45	0.83	0.45	0.84	0.47	0.63	0.46	0.57	0.46
49	0.65	0.40	0.62	0.41	0.86	0.42	0.85	0.42	0.83	0.49	0.77	0.54
50	0.66	0.44	0.60	0.42	0.64	0.39	0.62	0.40	0.78	0.49	0.72	0.53
51	0.83	0.37	0.83	0.37	0.55	0.38	0.54	0.40	0.44	0.42	0.44	0.41
52	0.74	0.41	0.73	0.41	0.85	0.49	0.83	0.51	0.58	0.35	0.53	0.35
53	0.74	0.44	0.75	0.43	0.64	0.43	0.63	0.44	0.74	0.58	0.67	0.59
54	0.76	0.52	0.77	0.52	0.78	0.40	0.73	0.41	0.72*	0.56*	0.65	0.55
55	0.70	0.53	0.69	0.52	0.51	0.32	0.50	0.30	0.82	0.53	0.76	0.57
56	0.67	0.44	0.64	0.44	0.59	0.37	0.57	0.37	0.78	0.39	0.76	0.44
57	0.47	0.40	0.46	0.38	0.72	0.40	0.69	0.39	0.65	0.56	0.59	0.57
58	0.63	0.54	0.63	0.53	0.73	0.41	0.72	0.44	0.72	0.54	0.66	0.54
59	0.76	0.35	0.76	0.42	0.72	0.42	0.70	0.44	0.63	0.46	0.60	0.46
60	0.81	0.49	0.78	0.51	0.44	0.39	0.43	0.37	0.51	0.47	0.47	0.44

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 Reliabilities and Intercorrelations for Science									
Subscore Area	**N of Items**							**Reliab.**	**SEM**
Grade 5 Science									
1. Physical Science (Grade 5)	11	1.00	.	.	.	.	.	0.68	1.35
2. Physical Science (Grade 4)	8	0.63	1.00	.	.	.	.	0.60	1.18
3. Life Science (Grade 5)	13	0.68	0.64	1.00	.	.	.	0.75	1.42
4. Life Science (Grade 4)	9	0.63	0.60	0.68	1.00	.	.	0.66	1.20
5. Earth Science (Grade 5)	11	0.62	0.58	0.67	0.63	1.00	.	0.65	1.36
6. Earth Science (Grade 4)	8	0.63	0.61	0.68	0.64	0.65	1.00	0.68	1.14
Grade 8 Science		**1**	**2**	**3**	**4**	**5**	**6**		
1. Motion	8	1.00	.	.	.	.	.	0.48	1.24
2. Forces, Density, and Buoyancy	13	0.55	1.00	.	.	.	.	0.73	1.42
3. Structure of Matter and Periodic Table	16	0.54	0.73	1.00	.	.	.	0.80	1.56
4. Earth in the Solar System	7	0.42	0.58	0.61	1.00	.	.	0.58	1.01
5. Reactions and the Chemistry of Living Systems	10	0.49	0.64	0.68	0.55	1.00	.	0.64	1.38
6. Investigation and Experimentation	6	0.47	0.63	0.64	0.52	0.58	1.00	0.58	0.85
Grade 10 Life Science		**1**	**2**	**3**	**4**	**5**	**6**		
1. Cell Biology	10	1.00	.	.	.	.	.	0.64	1.35
2. Genetics	12	0.62	1.00	.	.	.	.	0.70	1.50
3. Physiology	10	0.60	0.61	1.00	.	.	.	0.70	1.27
4. Ecology	11	0.61	0.62	0.69	1.00	.	.	0.72	1.33
5. Evolution	11	0.59	0.62	0.68	0.73	1.00	.	0.76	1.32
6. Investigation and Experimentation	6	0.45	0.49	0.54	0.58	0.60	1.00	0.52	0.91

Table 8.B.2 Reliabilities and SEMs for the CSTs by Gender

		Male			Female		
Content Area	**CST ***	**N**	**Rel**	**SEM**	**N**	**Rel**	**SEM**
Science	5	216,610	0.93	3.11	215,513	0.92	3.17
	8	218,535	0.92	3.07	216,268	0.91	3.17
	10 Life Science	219,630	0.93	3.16	215,594	0.91	3.22

* CSTs named by number only are grade-level tests.

Table 8.B.3 Reliabilities and SEMs for the CSTs by Economic Status

		Not Econ. Disadvantaged			Econ. Disadvantaged		
Content Area	**CST ***	**N**	**Rel**	**SEM**	**N**	**Rel**	**SEM**
Science	5	170,853	0.91	2.80	258,335	0.91	3.34
	8	179,277	0.91	2.86	252,421	0.90	3.29
	10 Life Science	191,637	0.92	2.99	240,183	0.91	3.34

* CSTs named by number only are grade-level tests.

Table 8.B.4 Reliabilities and SEMs for the CSTs by Special Services

		No Special Services			Special Services		
Content Area	**CST ***	**N**	**Rel**	**SEM**	**N**	**Rel**	**SEM**
Science	5	407,504	0.92	3.13	23,490	0.93	3.34
	8	413,046	0.91	3.11	22,091	0.91	3.41
	10 Life Science	410,298	0.92	3.18	25,357	0.91	3.47

* CSTs named by number only are grade-level tests.

Table 8.B.5 Reliabilities and SEMs for the CSTs by English-Language Fluency

		English Only			Initially Desig. Fluent		
Content Area	**CST ***	**N**	**Rel**	**SEM**	**N**	**Rel**	**SEM**
Science	5	242,022	0.92	3.03	20,590	0.91	2.75
	8	239,481	0.91	3.05	30,837	0.91	2.88
	10 Life Science	242,707	0.92	3.11	36,294	0.91	3.03

* CSTs named by number only are grade-level tests.

Table 8.B.6 Reliabilities and SEMs for the CSTs by English-Language Fluency

		EL			Redesigned Fluent		
Content Area	**CST ***	**N**	**Rel**	**SEM**	**N**	**Rel**	**SEM**
Science	5	88,537	0.88	3.56	79,627	0.87	3.07
	8	52,985	0.87	3.54	109,439	0.88	3.13
	10 Life Science	46,333	0.84	3.56	107,643	0.89	3.25

* CSTs named by number only are grade-level tests.

Table 8.B.7 Reliabilities and SEMs for the CSTs by Primary Ethnicity

		American Indian			Asian		
Content Area	**CST ***	**N**	**Rel**	**SEM**	**N**	**Rel**	**SEM**
Science	5	2,354	0.92	3.26	41,018	0.92	2.68
	8	2,604	0.91	3.26	42,016	0.91	2.64
	10 Life Science	3,026	0.92	3.28	39,488	0.92	2.85

* CSTs named by number only are grade-level tests.

Table 8.B.8 Reliabilities and SEMs for the CSTs by Primary Ethnicity

		Pacific Islander			Filipino		
Content Area	**CST ***	**N**	**Rel**	**SEM**	**N**	**Rel**	**SEM**
Science	5	2,267	0.91	3.27	11,288	0.90	2.94
	8	2,484	0.90	3.24	12,264	0.89	2.92
	10 Life Science	2,469	0.91	3.32	13,312	0.90	3.05

* CSTs named by number only are grade-level tests.

Table 8.B.9 Reliabilities and SEMs for the CSTs by Primary Ethnicity

		Hispanic			African American			White		
Content Area	**CST ***	**N**	**Rel**	**SEM**	**N**	**Rel**	**SEM**	**N**	**Rel**	**SEM**
Science	5	227,446	0.91	3.34	24,822	0.91	3.35	109,823	0.91	2.85
	8	225,170	0.90	3.29	26,215	0.91	3.35	111,777	0.91	2.90
	10 Life Science	223,128	0.91	3.34	26,483	0.91	3.37	114,981	0.92	2.98

* CSTs named by number only are grade-level tests.

Table 8.B.10 Reliabilities and SEMs for the CSTs by Primary Ethnicity-for-Not Economically Disadvantaged

		American Indian			Asian		
Content Area	**CST ***	**N**	**Rel**	**SEM**	**N**	**Rel**	**SEM**
Science	5	809	0.91	3.02	26,277	0.90	2.44
	8	1,002	0.91	3.08	26,319	0.90	2.44
	10 Life Science	1,243	0.92	3.12	24,087	0.91	2.67

* CSTs named by number only are grade-level tests.

Table 8.B.11 Reliabilities and SEMs for the CSTs by Primary Ethnicity-for-Not Economically Disadvantaged

		Pacific Islander			Filipino		
Content Area	**CST ***	**N**	**Rel**	**SEM**	**N**	**Rel**	**SEM**
Science	5	838	0.91	3.07	7,028	0.88	2.83
	8	985	0.90	3.08	7,740	0.88	2.82
	10 Life Science	1,100	0.92	3.21	8,738	0.89	2.98

* CSTs named by number only are grade-level tests.

Table 8.B.12 Reliabilities and SEMs for the CSTs by Primary Ethnicity-for-Not Economically Disadvantaged

		Hispanic			African American			White		
Content Area	**CST ***	**N**	**Rel**	**SEM**	**N**	**Rel**	**SEM**	**N**	**Rel**	**SEM**
Science	5	40,593	0.91	3.08	6,942	0.91	3.16	79,895	0.89	2.73
	8	44,342	0.90	3.12	8,373	0.91	3.22	82,637	0.89	2.80
	10 Life Science	51,650	0.91	3.21	9,762	0.91	3.28	86,993	0.91	2.89

* CSTs named by number only are grade-level tests.

Table 8.B.13 Reliabilities and SEMs for the CSTs by Primary Ethnicity-for-Economically Disadvantaged

		American Indian			Asian		
Content Area	**CST ***	**N**	**Rel**	**SEM**	**N**	**Rel**	**SEM**
Science	5	1,513	0.91	3.38	14,546	0.92	3.06
	8	1,581	0.90	3.36	15,557	0.91	2.94
	10 Life Science	1,755	0.91	3.38	15,255	0.92	3.10

* CSTs named by number only are grade-level tests.

Table 8.B.14 Reliabilities and SEMs for the CSTs by Primary Ethnicity-for-Economically Disadvantaged

		Pacific Islander			Filipino		
Content Area	**CST ***	**N**	**Rel**	**SEM**	**N**	**Rel**	**SEM**
Science	5	1,413	0.90	3.38	4,204	0.90	3.11
	8	1,487	0.89	3.34	4,475	0.89	3.07
	10 Life Science	1,349	0.90	3.39	4,525	0.90	3.19

* CSTs named by number only are grade-level tests.

Table 8.B.15 Reliabilities and SEMs for the CSTs by Primary Ethnicity-for-Economically Disadvantaged

		Hispanic			African American			White		
Content Area	**CST ***	**N**	**Rel**	**SEM**	**N**	**Rel**	**SEM**	**N**	**Rel**	**SEM**
Science	5	185,691	0.90	3.39	17,501	0.91	3.42	28,981	0.91	3.16
	8	179,454	0.90	3.33	17,507	0.90	3.41	28,322	0.91	3.17
	10 Life Science	169,819	0.90	3.38	16,369	0.90	3.43	27,050	0.92	3.21

* CSTs named by number only are grade-level tests.

Table 8.B.16 Subscore Reliabilities and SEM for Science by Gender/Economic Status

Subscore Reliabilities and SEM 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	0.69	1.34	0.67	1.35	0.63	1.19	0.63	1.44
2. Physical Science (Grade 4)	8	0.63	1.17	0.58	1.18	0.57	1.06	0.56	1.24
3. Life Science (Grade 5)	13	0.77	1.40	0.73	1.44	0.71	1.26	0.73	1.51
4. Life Science (Grade 4)	9	0.68	1.19	0.65	1.21	0.63	1.06	0.62	1.28
5. Earth Science (Grade 5)	11	0.67	1.34	0.63	1.37	0.60	1.23	0.61	1.43
6. Earth Science (Grade 4)	8	0.70	1.12	0.66	1.16	0.65	1.00	0.64	1.22
Grade 8 Science									
1. Motion	8	0.50	1.22	0.46	1.25	0.47	1.17	0.43	1.28
2. Forces, Density, and Buoyancy	13	0.76	1.39	0.71	1.46	0.72	1.29	0.70	1.51
3. Structure of Matter and Periodic Table	16	0.82	1.54	0.78	1.58	0.78	1.40	0.78	1.66
4. Earth in the Solar System	7	0.62	0.97	0.52	1.05	0.53	0.92	0.56	1.06
5. Reactions and the Chemistry of Living Systems	10	0.67	1.36	0.60	1.39	0.63	1.30	0.59	1.42
6. Investigation and Experimentation	6	0.60	0.86	0.55	0.85	0.55	0.73	0.55	0.93
Grade 10 Life Science									
1. Cell Biology	10	0.67	1.33	0.61	1.37	0.65	1.29	0.58	1.40
2. Genetics	12	0.72	1.49	0.68	1.51	0.69	1.44	0.67	1.54
3. Physiology	10	0.73	1.24	0.67	1.29	0.69	1.15	0.66	1.34
4. Ecology	11	0.75	1.30	0.70	1.35	0.70	1.22	0.69	1.40
5. Evolution	11	0.78	1.31	0.73	1.33	0.75	1.21	0.73	1.40
6. Investigation and Experimentation	6	0.56	0.91	0.47	0.91	0.47	0.83	0.51	0.96

Table 8.B.17 Subscore Reliabilities and SEM for Science by Special Services/English Fluency

Subscore Reliabilities and SEM for Science by Special Services/English Fluency													
Subscore Area	N of	No Spec. Serv.		Spec. Serv.		Eng. Only		Ini. Fluent		Learner		Red. Fluent	
	Items	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM
Grade 5 Science													
1. Physical Science (Grade 5)	11	0.67	1.34	0.70	1.44	0.66	1.29	0.65	1.17	0.55	1.53	0.56	1.33
2. Physical Science (Grade 4)	8	0.60	1.17	0.61	1.24	0.59	1.14	0.56	1.04	0.47	1.31	0.49	1.15
3. Life Science (Grade 5)	13	0.75	1.41	0.77	1.52	0.74	1.36	0.72	1.22	0.68	1.62	0.65	1.36
4. Life Science (Grade 4)	9	0.66	1.19	0.69	1.27	0.65	1.15	0.62	1.04	0.54	1.37	0.55	1.17
5. Earth Science (Grade 5)	11	0.64	1.35	0.70	1.43	0.64	1.31	0.57	1.22	0.55	1.51	0.49	1.32
6. Earth Science (Grade 4)	8	0.68	1.14	0.70	1.20	0.67	1.09	0.64	1.00	0.57	1.29	0.55	1.12
Grade 8 Science													
1. Motion	8	0.47	1.23	0.47	1.30	0.47	1.22	0.48	1.16	0.37	1.33	0.42	1.24
2. Forces, Density, and Buoyancy	13	0.73	1.42	0.72	1.57	0.74	1.38	0.72	1.31	0.62	1.62	0.66	1.43
3. Structure of Matter and Periodic Table	16	0.80	1.55	0.80	1.73	0.80	1.51	0.78	1.42	0.71	1.82	0.75	1.57
4. Earth in the Solar System	7	0.57	1.00	0.63	1.10	0.57	0.98	0.53	0.92	0.52	1.17	0.48	1.00
5. Reactions and the Chemistry of Living Systems	10	0.63	1.37	0.62	1.43	0.64	1.35	0.63	1.30	0.48	1.47	0.56	1.39
6. Investigation and Experimentation	6	0.57	0.84	0.56	1.02	0.58	0.82	0.55	0.74	0.46	1.07	0.46	0.84
Grade 10 Life Science													
1. Cell Biology	10	0.64	1.35	0.57	1.42	0.65	1.33	0.66	1.31	0.45	1.43	0.58	1.37
2. Genetics	12	0.69	1.49	0.63	1.56	0.70	1.48	0.69	1.45	0.53	1.57	0.64	1.51
3. Physiology	10	0.69	1.26	0.68	1.41	0.70	1.21	0.67	1.18	0.53	1.47	0.61	1.30
4. Ecology	11	0.72	1.32	0.70	1.46	0.72	1.28	0.69	1.25	0.55	1.51	0.64	1.36
5. Evolution	11	0.75	1.31	0.71	1.47	0.76	1.27	0.73	1.23	0.55	1.53	0.69	1.35
6. Investigation and Experimentation	6	0.50	0.90	0.57	1.04	0.52	0.88	0.44	0.85	0.47	1.08	0.40	0.92

Table 8.B.18 Subscore Reliabilities and SEM for Science by Primary Ethnicity

Subscore Reliabilities and SEM for Science by Primary Ethnicity															
Subscore Area	N of	Am. Ind.		Asian		Pac. Island		Filipino		Hispanic		African Am.		White	
	Items	Rel.	SEM	Rel.	SEM	Rel.	SEM	Rel.	SEM	Rel.	SEM	Rel.	SEM	Rel.	SEM
Grade 5 Science															
1. Physical Science (Grade 5)	11	0.66	1.39	0.69	1.14	0.62	1.40	0.60	1.25	0.63	1.44	0.64	1.43	0.63	1.21
2. Physical Science (Grade 4)	8	0.58	1.23	0.62	1.02	0.56	1.22	0.55	1.12	0.56	1.24	0.57	1.24	0.56	1.08
3. Life Science (Grade 5)	13	0.76	1.47	0.75	1.20	0.72	1.49	0.69	1.32	0.73	1.51	0.73	1.53	0.71	1.28
4. Life Science (Grade 4)	9	0.64	1.25	0.67	1.01	0.62	1.25	0.60	1.13	0.62	1.28	0.64	1.28	0.62	1.08
5. Earth Science (Grade 5)	11	0.66	1.41	0.63	1.20	0.61	1.41	0.55	1.29	0.61	1.43	0.65	1.44	0.60	1.25
6. Earth Science (Grade 4)	8	0.68	1.19	0.69	0.94	0.63	1.19	0.60	1.05	0.63	1.22	0.66	1.22	0.65	1.03
Grade 8 Science															
1. Motion	8	0.47	1.27	0.51	1.08	0.42	1.27	0.45	1.19	0.43	1.28	0.41	1.29	0.45	1.18
2. Forces, Density, and Buoyancy	13	0.73	1.48	0.73	1.19	0.69	1.48	0.68	1.32	0.69	1.50	0.70	1.54	0.72	1.31
3. Structure of Matter and Periodic Table	16	0.79	1.64	0.79	1.28	0.77	1.62	0.75	1.43	0.78	1.67	0.79	1.69	0.78	1.43
4. Earth in the Solar System	7	0.58	1.05	0.54	0.88	0.54	1.06	0.45	0.95	0.55	1.06	0.57	1.10	0.54	0.93
5. Reactions and the Chemistry of Living Systems	10	0.62	1.41	0.66	1.22	0.59	1.41	0.57	1.33	0.58	1.43	0.60	1.43	0.62	1.32
6. Investigation and Experimentation	6	0.58	0.91	0.54	0.65	0.52	0.90	0.49	0.75	0.54	0.92	0.57	0.95	0.55	0.74

Subscore Reliabilities and SEM for Science by Primary Ethnicity															
Subscore Area	N of	Am. Ind.		Asian		Pac. Island		Filipino		Hispanic		African Am.		White	
	Items	Rel.	SEM	Rel.	SEM	Rel.	SEM	Rel.	SEM	Rel.	SEM	Rel.	SEM	Rel.	SEM
Grade 10 Life Science															
1. Cell Biology	10	0.61	1.38	0.69	1.22	0.59	1.39	0.61	1.31	0.57	1.40	0.56	1.41	0.64	1.30
2. Genetics	12	0.69	1.53	0.70	1.38	0.65	1.54	0.65	1.46	0.66	1.54	0.67	1.54	0.69	1.45
3. Physiology	10	0.70	1.30	0.71	1.10	0.68	1.33	0.64	1.19	0.66	1.34	0.67	1.35	0.69	1.14
4. Ecology	11	0.72	1.36	0.70	1.18	0.70	1.38	0.65	1.27	0.69	1.40	0.69	1.41	0.71	1.21
5. Evolution	11	0.76	1.36	0.75	1.16	0.75	1.38	0.70	1.25	0.73	1.39	0.73	1.42	0.75	1.20
6. Investigation and Experimentation	6	0.54	0.95	0.41	0.80	0.54	0.95	0.37	0.85	0.50	0.96	0.54	0.99	0.49	0.82

Table 8.B.19 Subscore Reliabilities and SEM for Science by Primary Ethnicity-for-Not Economically Disadvantaged

Subscore Reliabilities and SEM for Science by Primary Ethnicity-for-Not Economically Disadvantaged															
Subscore Area	N of	Am. Ind.		Asian		Pac. Island		Filipino		Hispanic		African Am.		White	
	Items	Rel.	SEM	Rel.	SEM	Rel.	SEM	Rel.	SEM	Rel.	SEM	Rel.	SEM	Rel.	SEM
Grade 5 Science															
1. Physical Science (Grade 5)	11	0.64	1.28	0.64	1.03	0.64	1.31	0.56	1.20	0.63	1.32	0.64	1.34	0.58	1.16
2. Physical Science (Grade 4)	8	0.59	1.14	0.56	0.94	0.56	1.15	0.53	1.08	0.57	1.16	0.58	1.18	0.52	1.04
3. Life Science (Grade 5)	13	0.71	1.36	0.69	1.09	0.71	1.39	0.67	1.26	0.72	1.38	0.73	1.43	0.67	1.23
4. Life Science (Grade 4)	9	0.59	1.15	0.62	0.92	0.61	1.18	0.58	1.09	0.63	1.17	0.65	1.21	0.58	1.02
5. Earth Science (Grade 5)	11	0.63	1.31	0.57	1.12	0.58	1.34	0.52	1.25	0.60	1.33	0.64	1.36	0.55	1.20
6. Earth Science (Grade 4)	8	0.69	1.09	0.65	0.83	0.63	1.11	0.56	1.00	0.63	1.12	0.67	1.15	0.61	0.97
Grade 8 Science															
1. Motion	8	0.49	1.23	0.47	1.01	0.40	1.23	0.44	1.16	0.44	1.25	0.43	1.27	0.44	1.15
2. Forces, Density, and Buoyancy	13	0.71	1.40	0.69	1.09	0.70	1.40	0.66	1.28	0.70	1.42	0.71	1.47	0.69	1.26
3. Structure of Matter and Periodic Table	16	0.77	1.54	0.75	1.16	0.77	1.52	0.73	1.36	0.77	1.56	0.79	1.61	0.75	1.37
4. Earth in the Solar System	7	0.56	0.98	0.48	0.82	0.52	1.00	0.43	0.91	0.54	1.00	0.57	1.05	0.50	0.90
5. Reactions and the Chemistry of Living Systems	10	0.63	1.37	0.63	1.15	0.59	1.38	0.56	1.30	0.60	1.38	0.60	1.40	0.60	1.29
6. Investigation and Experimentation	6	0.53	0.83	0.48	0.56	0.51	0.82	0.46	0.71	0.54	0.84	0.59	0.88	0.51	0.69
Grade 10 Life Science															
1. Cell Biology	10	0.62	1.35	0.67	1.15	0.63	1.36	0.61	1.29	0.60	1.36	0.59	1.38	0.63	1.27
2. Genetics	12	0.71	1.48	0.68	1.32	0.65	1.52	0.63	1.44	0.67	1.51	0.68	1.52	0.67	1.42
3. Physiology	10	0.68	1.21	0.67	1.00	0.69	1.27	0.63	1.15	0.67	1.27	0.67	1.30	0.66	1.10
4. Ecology	11	0.71	1.28	0.65	1.09	0.71	1.33	0.63	1.23	0.70	1.33	0.69	1.36	0.68	1.17
5. Evolution	11	0.75	1.28	0.72	1.07	0.75	1.32	0.67	1.21	0.74	1.32	0.74	1.36	0.73	1.16
6. Investigation and Experimentation	6	0.51	0.88	0.35	0.75	0.52	0.91	0.35	0.83	0.49	0.91	0.52	0.95	0.44	0.79

Table 8.B.20 Subscore Reliabilities and SEM for Science by Primary Ethnicity-for-Economically Disadvantaged

Subscore Reliabilities and SEM 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	0.64	1.44	0.68	1.31	0.58	1.45	0.61	1.33	0.62	1.46	0.62	1.46	0.64	1.35
2. Physical Science (Grade 4)	8	0.54	1.27	0.60	1.15	0.54	1.25	0.56	1.18	0.54	1.26	0.55	1.26	0.56	1.19
3. Life Science (Grade 5)	13	0.75	1.53	0.75	1.37	0.71	1.54	0.70	1.40	0.72	1.53	0.72	1.56	0.73	1.42
4. Life Science (Grade 4)	9	0.62	1.30	0.66	1.17	0.59	1.30	0.61	1.20	0.61	1.30	0.62	1.31	0.63	1.20
5. Earth Science (Grade 5)	11	0.64	1.46	0.62	1.33	0.60	1.44	0.57	1.34	0.60	1.44	0.64	1.46	0.63	1.36
6. Earth Science (Grade 4)	8	0.64	1.23	0.67	1.10	0.61	1.23	0.61	1.12	0.62	1.24	0.64	1.25	0.64	1.16
Grade 8 Science															
1. Motion	8	0.44	1.29	0.50	1.17	0.40	1.29	0.45	1.23	0.42	1.29	0.39	1.31	0.42	1.25
2. Forces, Density, and Buoyancy	13	0.71	1.53	0.73	1.33	0.65	1.53	0.67	1.40	0.68	1.52	0.68	1.56	0.72	1.44
3. Structure of Matter and Periodic Table	16	0.78	1.70	0.79	1.45	0.75	1.68	0.75	1.53	0.77	1.69	0.77	1.72	0.78	1.58
4. Earth in the Solar System	7	0.56	1.09	0.56	0.97	0.53	1.09	0.46	1.00	0.55	1.08	0.56	1.12	0.57	1.02
5. Reactions and the Chemistry of Living Systems	10	0.58	1.43	0.63	1.32	0.57	1.43	0.57	1.37	0.56	1.44	0.58	1.44	0.60	1.39
6. Investigation and Experimentation	6	0.58	0.96	0.54	0.77	0.51	0.95	0.51	0.82	0.54	0.94	0.55	0.99	0.56	0.86
Grade 10 Life Science															
1. Cell Biology	10	0.56	1.41	0.66	1.31	0.53	1.41	0.60	1.35	0.55	1.41	0.52	1.42	0.61	1.37
2. Genetics	12	0.64	1.56	0.69	1.46	0.63	1.55	0.65	1.49	0.65	1.55	0.65	1.56	0.68	1.52
3. Physiology	10	0.67	1.35	0.69	1.23	0.65	1.37	0.64	1.27	0.64	1.37	0.65	1.38	0.70	1.26
4. Ecology	11	0.69	1.42	0.69	1.29	0.68	1.42	0.66	1.33	0.68	1.42	0.67	1.44	0.72	1.33
5. Evolution	11	0.74	1.41	0.74	1.29	0.73	1.42	0.70	1.32	0.72	1.41	0.71	1.44	0.75	1.32
6. Investigation and Experimentation	6	0.53	0.99	0.42	0.88	0.54	0.98	0.39	0.90	0.50	0.98	0.53	1.02	0.54	0.91

Table 8.B.21 Reliability of Classification for Science, Grade Five

	Placement Score	Far Below Basic	Below Basic	Basic	Proficient	Advanced	Category Total
Decision Accuracy	0–22	0.04	0.02	0.00	0.00	0.00	0.06
	23–30	0.01	0.07	0.03	0.00	0.00	0.11
	31–41	0.00	0.02	0.18	0.04	0.00	0.24
All-forms Average	42–51	0.00	0.00	0.04	0.25	0.04	0.33
	52–60	0.00	0.00	0.00	0.04	0.22	0.27
	Estimated Proportion Correctly Classified: Total = 0.76, Proficient & Above = 0.92						
Decision Consistency	0–22	0.04	0.02	0.00	0.00	0.00	0.06
	23–30	0.02	0.05	0.03	0.00	0.00	0.11
	31–41	0.00	0.03	0.15	0.05	0.00	0.24
Alternate Form	42–51	0.00	0.00	0.06	0.21	0.06	0.33
	52–60	0.00	0.00	0.00	0.05	0.21	0.27
	Estimated Proportion Consistently Classified: Total = 0.67, Proficient & Above = 0.89						

Table 8.B.22 Reliability of Classification for Science, Grade Eight

	Placement Score	Far Below Basic	Below Basic	Basic	Proficient	Advanced	Category Total
Decision Accuracy	0–23	0.05	0.02	0.00	0.00	0.00	0.07
	24–30	0.01	0.06	0.03	0.00	0.00	0.09
	31–38	0.00	0.02	0.12	0.04	0.00	0.18
All-forms Average	39–45	0.00	0.00	0.04	0.14	0.04	0.22
	46–60	0.00	0.00	0.00	0.04	0.40	0.44
	Estimated Proportion Correctly Classified: Total = 0.76, Proficient & Above = 0.92						
Decision Consistency	0–23	0.04	0.02	0.00	0.00	0.00	0.07
	24–30	0.02	0.04	0.03	0.00	0.00	0.09
	31–38	0.00	0.03	0.10	0.05	0.00	0.18
Alternate Form	39–45	0.00	0.00	0.05	0.11	0.06	0.22
	46–60	0.00	0.00	0.00	0.06	0.38	0.44
	Estimated Proportion Consistently Classified: Total = 0.68, Proficient & Above = 0.89						

Table 8.B.23 Reliability of Classification for Life Science (Grade 10)

	Placement Score	Far Below Basic	Below Basic	Basic	Proficient	Advanced	Category Total
Decision Accuracy	0–21	0.05	0.02	0.00	0.00	0.00	0.07
	22–28	0.01	0.07	0.03	0.00	0.00	0.11
	29–39	0.00	0.02	0.20	0.04	0.00	0.26
All-forms Average	40–48	0.00	0.00	0.04	0.20	0.03	0.28
	49–60	0.00	0.00	0.00	0.04	0.23	0.28
	Estimated Proportion Correctly Classified: Total = 0.75, Proficient & Above = 0.92						
Decision Consistency	0–21	0.05	0.02	0.00	0.00	0.00	0.07
	22–28	0.02	0.05	0.04	0.00	0.00	0.11
	29–39	0.00	0.04	0.17	0.05	0.00	0.26
Alternate Form	40–48	0.00	0.00	0.06	0.17	0.05	0.28
	49–60	0.00	0.00	0.00	0.05	0.22	0.28
	Estimated Proportion Consistently Classified: Total = 0.66, Proficient & Above = 0.89						

Appendix 8.C—IRT Analyses

Table 8.C.1 IRT *b*-values for Life Science (Grade 10)

Content Area	Number of items	Mean	Standard Deviation	Minimum	Maximum
Cell Biology	10	0.08	0.78	–0.96	1.44
Genetics	12	0.04	0.63	–0.96	1.03
Physiology	10	–0.51	0.60	–1.34	0.66
Ecology	11	–0.42	0.74	–1.66	0.56
Evolution	11	–0.34	0.53	–1.05	0.71
Investigation and Experimentation	6	–0.99	0.49	–1.57	–0.03
All operational items	60	–0.30	0.72	–1.66	1.44

Table 8.C.2 Distribution of IRT *b*-values for Life Science (Grade 10)

IRT *b*–value	
>=3.5	
3.0 – < 3.5	
2.5 – < 3.0	
2.0 – < 2.5	
1.5 – < 2.0	
1.0 – < 1.5	2
0.5 – < 1.0	10
0.0 – < 0.5	7
–0.5 – < 0.0	17
–1.0 – < –0.5	13
–1.5 – < –1.0	8
–2.0 – < –1.5	3
–2.5 – < –2.0	
–3.0 – < –2.5	
–3.5 – < –3.0	
< –3.5	
Total	60

Scaling and Pre-Equating Results

Table 8.C.3 New Conversion for Science, Grade Five

Grade 5									
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**	7,430	–0.5813	302.1382	302
1	–	–4.8561	68.9433	150	**32**	7,679	–0.5099	306.2595	306
2	–	–4.1421	93.5582	150	**33**	8,073	–0.4381	310.3994	310
3	–	–3.7153	121.3657	150	**34**	8,591	–0.3659	314.5646	315
4	–	–3.4058	138.2692	150	**35**	9,001	–0.2930	318.7702	319
5	–	–3.1605	152.8431	153	**36**	9,494	–0.2193	323.0195	323
6	–	–2.9556	165.1456	165	**37**	9,802	–0.1447	327.3259	327
7	17	–2.7786	175.1713	175	**38**	10,272	–0.0689	331.6993	332
8	43	–2.6217	184.2352	184	**39**	10,708	0.0083	336.1500	336
9	85	–2.4803	192.5306	193	**40**	11,148	0.0872	340.6987	341
10	162	–2.3508	200.0383	200	**41**	11,739	0.1679	345.3562	345
11	313	–2.2310	206.8915	207	**42**	12,166	0.2509	350.1357	350
12	549	–2.1190	213.3720	213	**43**	12,679	0.3364	355.0715	355
13	805	–2.0136	219.5089	220	**44**	13,073	0.4248	360.1775	360
14	1,165	–1.9137	225.2765	225	**45**	13,478	0.5166	365.4755	365
15	1,424	–1.8183	230.7487	231	**46**	13,802	0.6125	370.9999	371
16	1,959	–1.7269	236.0251	236	**47**	14,349	0.7130	376.7873	377
17	2,494	–1.6389	241.1222	241	**48**	14,694	0.8190	382.9139	383
18	2,834	–1.5538	246.0571	246	**49**	15,196	0.9315	389.4152	389
19	3,254	–1.4712	250.8045	251	**50**	15,588	1.0519	396.3589	396
20	3,490	–1.3908	255.4332	255	**51**	15,757	1.1821	403.8464	404
21	3,946	–1.3123	259.9630	260	**52**	15,977	1.3243	412.0894	412
22	4,254	–1.2353	264.4059	264	**53**	16,078	1.4819	421.1809	421
23	4,666	–1.1597	268.7747	269	**54**	16,094	1.6598	431.4239	431
24	4,887	–1.0853	273.0751	273	**55**	15,845	1.8654	443.3701	443
25	5,260	–1.0118	277.3188	277	**56**	14,811	2.1117	457.4900	457
26	5,412	–0.9390	281.5046	282	**57**	13,285	2.4221	475.6943	476
27	5,894	–0.8669	285.6652	286	**58**	11,155	2.8497	500.2448	500
28	6,189	–0.7952	289.7988	290	**59**	7,901	3.5650	544.4915	544
29	6,698	–0.7238	293.9165	294	**60**	3,656	N/A	624.0193	600
30	7,087	–0.6526	298.0285	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.4 New Conversion for Science, Grade Eight (paper-pencil)

Grade 8 (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**	7,762	–0.2623	300.7558	301
1	–	–4.6998	–55.3964	150	**32**	8,366	–0.1867	307.0871	307
2	–	–3.9761	–14.6840	150	**33**	8,919	–0.1109	313.4512	313
3	–	–3.5404	23.7939	150	**34**	9,540	–0.0346	319.8514	320
4	–	–3.2228	50.7369	150	**35**	10,230	0.0423	326.3089	326
5	–	–2.9699	73.0720	150	**36**	10,815	0.1200	332.8246	333
6	11	–2.7580	91.0841	150	**37**	11,310	0.1987	339.4283	339
7	30	–2.5742	106.3372	150	**38**	11,859	0.2786	346.1280	346
8	64	–2.4110	120.2245	150	**39**	12,376	0.3598	352.9423	353
9	136	–2.2635	132.8635	150	**40**	13,057	0.4427	359.8968	360
10	233	–2.1283	144.0709	150	**41**	13,446	0.5274	367.0081	367
11	373	–2.0029	154.5785	155	**42**	13,813	0.6144	374.2955	374
12	615	–1.8855	164.4838	164	**43**	14,225	0.7039	381.7952	382
13	861	–1.7749	173.8483	174	**44**	14,595	0.7963	389.5551	390
14	1,173	–1.6698	182.6473	183	**45**	14,765	0.8921	397.6013	398
15	1,491	–1.5695	191.0280	191	**46**	15,124	0.9919	405.9828	406
16	1,824	–1.4733	199.1031	199	**47**	15,377	1.0964	414.7444	415
17	2,222	–1.3805	206.9038	207	**48**	15,622	1.2063	423.9491	424
18	2,419	–1.2908	214.4589	214	**49**	15,667	1.3228	433.7363	434
19	2,772	–1.2036	221.7958	222	**50**	15,550	1.4472	444.1957	444
20	3,048	–1.1186	228.9100	229	**51**	15,279	1.5813	455.4467	455
21	3,365	–1.0356	235.8602	236	**52**	15,366	1.7274	467.6645	468
22	3,673	–0.9543	242.6815	243	**53**	14,844	1.8888	481.2722	481
23	3,957	–0.8743	249.3907	249	**54**	14,068	2.0706	496.5756	497
24	4,217	–0.7956	256.0023	256	**55**	13,480	2.2802	514.0628	514
25	4,812	–0.7178	262.5307	263	**56**	12,353	2.5302	535.2707	535
26	5,322	–0.6408	268.9887	269	**57**	11,011	2.8443	561.7836	562
27	5,722	–0.5645	275.3933	275	**58**	9,069	3.2760	598.1208	598
28	6,192	–0.4886	281.7635	282	**59**	6,319	3.9948	662.8192	600
29	6,659	–0.4131	288.1043	288	**60**	2,433	N/A	742.1487	600
30	7,312	–0.3377	294.4289	294					

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 New Conversion for Science, Grade Eight (Braille)

Grade 8 (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**	–	–0.2054	305.5257	306
1	–	–4.6652	–53.9658	150	**32**	–	–0.1285	311.9714	312
2	–	–3.9407	–12.1034	150	**33**	–	–0.0513	318.4476	318
3	–	–3.5042	27.7526	150	**34**	–	0.0264	324.9786	325
4	–	–3.1858	53.9611	150	**35**	–	0.1050	331.5590	332
5	–	–2.9322	76.0468	150	**36**	–	0.1844	338.2259	338
6	–	–2.7196	94.5931	150	**37**	–	0.2650	344.9846	345
7	–	–2.5352	109.7310	150	**38**	–	0.3469	351.8560	352
8	–	–2.3714	123.5016	150	**39**	–	0.4303	358.8600	359
9	–	–2.2233	136.0828	150	**40**	–	0.5157	366.0258	366
10	–	–2.0875	147.6508	150	**41**	–	0.6032	373.3626	373
11	–	–1.9615	158.1181	158	**42**	–	0.6932	380.9088	381
12	–	–1.8435	167.9877	168	**43**	–	0.7862	388.6991	389
13	–	–1.7323	177.3452	177	**44**	–	0.8825	396.7851	397
14	–	–1.6267	186.2526	186	**45**	–	0.9827	405.2095	405
15	–	–1.5258	194.7614	195	**46**	–	1.0876	414.0108	414
16	–	–1.4289	202.8681	203	**47**	–	1.1980	423.2565	423
17	–	–1.3355	210.6836	211	**48**	–	1.3149	433.0555	433
18	–	–1.2451	218.2630	218	**49**	–	1.4396	443.5521	444
19	–	–1.1573	225.6371	226	**50**	–	1.5740	454.8417	455
20	–	–1.0717	232.8289	233	**51**	–	1.7205	467.0981	467
21	–	–0.9880	239.8571	240	**52**	–	1.8822	480.7055	481
22	–	–0.9059	246.7527	247	**53**	–	2.0643	496.0491	496
23	–	–0.8252	253.5321	254	**54**	–	2.2741	513.5805	514
24	–	–0.7456	260.2133	260	**55**	–	2.5245	534.7716	535
25	–	–0.6670	266.8061	267	**56**	–	2.8388	561.3395	561
26	–	–0.5892	273.3294	273	**57**	–	3.2708	597.6100	598
27	–	–0.5119	279.8129	280	**58**	–	3.9898	662.3948	600
28	–	–0.4351	286.2610	286	**59**	–	N/A	742.1487	600
29	–	–0.3585	292.6853	293					
30	–	–0.2819	299.1064	299					

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.6 New Conversion for Life Science (Grade 10)

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**	9,143	–0.2365	310.3591	310
1	–	–4.6201	72.9274	150	**32**	9,758	–0.1612	314.5566	315
2	–	–3.9024	102.6817	150	**33**	9,899	–0.0857	318.7697	319
3	–	–3.4719	128.9831	150	**34**	10,458	–0.0096	323.0129	323
4	–	–3.1588	147.2769	150	**35**	10,867	0.0671	327.2948	327
5	12	–2.9099	161.0150	161	**36**	11,106	0.1447	331.6188	332
6	20	–2.7014	172.6109	173	**37**	11,316	0.2233	335.9980	336
7	49	–2.5207	182.7643	183	**38**	12,024	0.3030	340.4499	340
8	93	–2.3603	191.8046	192	**39**	12,148	0.3842	344.9761	345
9	195	–2.2153	199.9795	200	**40**	12,598	0.4671	349.5978	350
10	343	–2.0822	207.4139	207	**41**	12,952	0.5519	354.3243	354
11	579	–1.9588	214.2613	214	**42**	13,294	0.6390	359.1783	359
12	949	–1.8433	220.6965	221	**43**	13,496	0.7286	364.1828	364
13	1,256	–1.7343	226.7826	227	**44**	13,639	0.8212	369.3479	369
14	1,844	–1.6308	232.5681	233	**45**	13,940	0.9172	374.7117	375
15	2,287	–1.5319	238.0958	238	**46**	13,854	1.0173	380.2976	380
16	2,798	–1.4369	243.4075	243	**47**	14,102	1.1221	386.1428	386
17	3,400	–1.3453	248.5283	249	**48**	14,067	1.2324	392.2927	392
18	3,830	–1.2567	253.4728	253	**49**	13,809	1.3495	398.8098	399
19	4,337	–1.1705	258.2678	258	**50**	13,763	1.4744	405.7660	406
20	4,728	–1.0865	262.9503	263	**51**	13,614	1.6090	413.3049	413
21	5,175	–1.0043	267.5333	268	**52**	13,174	1.7558	421.5196	422
22	5,590	–0.9237	272.0259	272	**53**	12,680	1.9180	430.5908	431
23	5,910	–0.8445	276.4452	276	**54**	12,032	2.1007	440.7887	441
24	6,427	–0.7664	280.8018	281	**55**	10,880	2.3112	452.4979	452
25	6,771	–0.6893	285.1054	285	**56**	9,745	2.5624	466.5075	467
26	7,215	–0.6129	289.3658	289	**57**	8,306	2.8776	484.4417	484
27	7,843	–0.5371	293.5937	294	**58**	6,261	3.3108	509.2961	509
28	8,080	–0.4617	297.7979	298	**59**	4,072	4.0311	551.1557	551
29	8,618	–0.3865	301.9873	302	**60**	1,640	N/A	602.3473	600
30	8,693	–0.3115	306.1720	306					

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, 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 CSTs prior to the 2013 administration is described in detail in Chapter 3, starting on page 29; there was no new item development in 2013 because of the form reuse. 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 2014 CST administration.

Item Specifications

ETS maintained item specifications for each CST 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 CDE. Adherence to the specifications ensured the maintenance of quality and consistency in the item development process.

Item Writers

The items for each CST were written by item writers with a thorough understanding of the California content standards. 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 ARPs were committees that advised the CDE and ETS on areas related to item development for the CSTs. 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 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 *EC* 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 examinees 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 administered in 2014 were statistically reviewed in data review meetings the year they were originally administered. There was no data review of field-test items in 2014. See Table 8.4 on page 91 for the list of the original administrations of each test administered in 2014.

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, 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 materials are delivered to the Pearson scanning and scoring facilities in Iowa City, Iowa. The nonscorable materials, including test booklets, are returned to the Security Processing Department in Pearson's Cedar Rapids, Iowa, facility. ETS and Pearson 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 CAASPP Coordinators and works with them to facilitate the return of the test materials.

Processing Test Materials

Upon receipt of the test materials, Pearson 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.

Pearson'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, Pearson conducts a complete check of the scanning system. ETS and Pearson create test decks for every test and form. Each test deck consists of approximately 25 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, Pearson 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.

Pearson 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:

1. Grade counts are compared to the District Master File Sheets.
2. Document counts are compared to the School Master File Sheets.
3. Document counts are compared to the SGIDs.

Any discrepancies identified in the steps outlined above are followed up by Pearson 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 2014, 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:

- General Reporting Specifications
- Form Planner Specifications
- Aggregation Rules
- “What If” List
- Edit Specifications
- Reporting Cluster Names and Item Numbers

Each of these documents is explained in detail in Chapter 7, starting on page 63. The scoring specifications are reviewed and revised by the CDE, ETS, and Pearson 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 Pearson. 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 and Pearson take 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 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 examinees with valid scores, 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 before ETS delivered them to Pearson.

During the current administration, the data derived from prior item analyses are used to pre-equate the 2014 results. Key checks and classical item analyses as well as associated quality assurance checks are also conducted on the current data.

For tests with no item replacements, the scoring tables are reused and are checked against the scoring tables in the reuse-year technical report to ensure exact match. For tests with item replacements, the new scoring tables are checked against scoring tables in the reuse year conversion for reasonableness. 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

Pearson utilizes the raw-to-scale scoring tables to assign scale scores for each student. ETS verifies Pearson's scale scores by independently generating the scale scores for students in a small number of LEAs and comparing these scores with those generated by Pearson. 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 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 forms that are reused in 2014, 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 2014 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, ETS, and Pearson prior to any LEA mailings

All reports are required to include a single, accurate 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. Pearson sends paper 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 review.

Upon the CDE's approval of the reports generated from the pilot LEAs, Pearson 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 mailing them to the LEAs.

Electronic Reporting

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

Before an LEA could download a student data file, ETS statisticians approved a QC file of test results data and ETS IT successfully processed the QC file. Once the data were deemed reliable and Pearson processed a scorable answer document for every student who took a CST 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

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

Chapter 10: Historical Comparisons

Base-year Comparisons

Historical comparisons of the CST results are routinely performed to identify the trends in examinee performance and test characteristics over time. Such comparisons were performed over a period of the three most recent years of administration—2012, 2013, and 2014—and the base year.

The indicators of examinee performance include the mean and standard deviation of scale scores, observed score ranges, and the percentage of examinees classified into proficient and advanced performance levels. Test characteristics are compared by looking at the mean proportion correct, overall reliability and SEM, as well as the mean IRT *b*-value for each CST.

The base year of each CST 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 are presented in Table 10.1.

Table 10.1 Base Years for CSTs

Content Area	CST	Base Year
Science	5	2004
	8	2006
	10 Life Science	2006

The base years differ over CSTs 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 science CST, and 2006 is the base year for the science CSTs in grades eight and ten.

Examinee Performance

Table 10.A.1 on page 127 contains the number of examinees assessed and the means and standard deviations of examinees' scale scores in the base year and in 2012, 2013, and 2014 for each CST. As noted in previous chapters, the CST reporting scales range from 150 to 600 for all of the tests.

CST 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.2 on page 127; please note that this information may differ slightly from information found on the CDE's 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.3 shows for each CST the distribution of scale scores observed in the base year and in 2012, 2013, and 2014. 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, 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 over the past several years indicate that the CSTs 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 example, in the years prior to 2014, efforts were made to field test more easy items for some CSTs where previous field testing resulted in an overabundance of very difficult items.

Table 10.B.1 in Appendix 10.B, which starts on page 128, presents the average proportion-correct values for the operational items in each CST 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 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 standard errors of measurement (SEM) expressed in raw score units appear in Table 10.B.4. 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, Examinee Performance

Table 10.A.1 Number of Examinees Tested, Scale Score Means, and Standard Deviations of CSTs Across Base Year, 2012, 2013, and 2014

Content Area	CST *	Number of Examinees (valid scores)				Scale Score Mean and Standard Deviation							
						Base		2012		2013		2014	
		Base	2012	2013	2014	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Science	5	483,931	430,759	422,520	432,439	318	44	367	66	367	67	368	70
	8	478,667	435,529	430,581	435,173	331	71	395	99	391	97	392	94
	10 Life Science	461,634	449,842	442,013	435,705	327	58	357	69	360	70	360	65

* CSTs named by number only are grade-level tests.

Table 10.A.2 Percentage of Proficient and Above and Percentage of Advanced Across Base Year, 2012, 2013, and 2014

Content Area	CST *	% Proficient and Above				% Advanced			
		Base	2012	2013	2014	Base	2012	2013	2014
Science	5	24%	59%	57%	59%	3%	25%	24%	27%
	8	38%	66%	67%	66%	17%	45%	42%	44%
	10 Life Science	34%	53%	54%	56%	13%	27%	28%	28%

* CSTs named by number only are grade-level tests.

Table 10.A.3 Observed Score Distributions of CSTs Across Base Year, 2012, 2013, and 2014 for Science (Grades Five and Eight and Life Science [Grade Ten])

Observed Score Distributions	Grade 5				Grade 8				Grade 10 Life Science			
	Base	2012	2013	2014	Base	2012	2013	2014	Base	2012	2013	2014
570–600	3	1,774	2,278	3,656	1,428	22,032	21,402	17,821	238	1,454	2,570	1,640
540–569	30	4,991	5,668	7,901	2,429	12,514	11,638	11,011	648	3,716	5,408	4,072
510–539	105	N/A	N/A	N/A	4,190	29,333	13,034	25,833	1,121	5,990	N/A	N/A
480–509	590	8,458	20,302	11,155	5,863	15,868	27,968	28,913	4,068	8,010	17,027	14,567
450–479	1,414	25,460	13,240	28,096	13,160	49,726	45,062	30,647	6,808	21,099	22,393	20,625
420–449	7,041	49,343	44,031	48,017	26,047	49,896	45,778	46,840	15,540	38,420	38,447	37,886
390–419	21,141	67,092	60,538	47,322	36,389	45,423	44,016	59,862	38,323	55,525	53,587	55,253
360–389	49,769	73,394	69,402	84,592	61,392	52,213	53,651	54,544	59,085	79,972	76,444	69,031
330–359	93,778	69,546	80,486	68,712	76,067	42,860	56,194	46,363	74,451	69,836	77,072	85,438
300–329	141,637	65,237	60,445	60,070	85,106	40,763	34,839	44,819	91,629	73,234	62,371	67,436
270–299	106,602	37,686	40,821	41,427	67,991	29,021	31,785	25,886	85,472	41,751	49,088	47,836
240–269	55,651	20,195	19,342	24,938	57,597	17,485	21,855	21,981	71,350	35,293	28,673	24,268
210–239	5,829	6,750	5,379	5,902	32,609	16,577	12,402	11,605	11,828	13,827	8,127	6,915
180–209	298	790	551	603	7,222	6,957	7,977	6,710	942	1,565	651	680
150–179	43	43	37	48	1,177	4,861	2,980	2,338	131	150	155	58

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, 2012, 2013, and 2014

		Mean *p*-value			
Content Area	**CST ***	**Base**	**2012**	**2013**	**2014**
Science	**5**	0.47	0.71	0.70	0.71
	8	0.50	0.71	0.71	0.70
	10 Life Science	0.51	0.65	0.66	0.67

* CSTs named by number only are grade-level tests.

Table 10.B.2 Mean IRT *b*-values for Operational Test Items Across Base Year, 2012, 2013, and 2014

		Mean IRT *b*-value			
Content Area	**CST ***	**Base**	**2012**	**2013**	**2014**
Science	**5**	–0.19	–0.62	–0.61	–0.64
	8	0.10	–0.39	–0.36	–0.34
	10 Life Science	–0.04	–0.26	–0.27	–0.31

* CSTs named by number only are grade-level tests.

Table 10.B.3 Mean Point-Biserial Correlation for Operational Test Items Across Base Year, 2012, 2013, and 2014

		Mean Point-Biserial Correlation			
Content Area	**CST ***	**Base**	**2012**	**2013**	**2014**
Science	**5**	0.33	0.42	0.42	0.43
	8	0.35	0.44	0.43	0.42
	10 Life Science	0.41	0.45	0.45	0.42

* CSTs named by number only are grade-level tests.

Table 10.B.4 Score Reliabilities (Cronbach's Alpha) and SEM of CSTs Across Base Year, 2012, 2013, and 2014

		Reliability				SEM			
Content Area	**CST ***	**Base**	**2012**	**2013**	**2014**	**Base**	**2012**	**2013**	**2014**
Science	**5**	0.86	0.92	0.92	0.92	3.6	3.2	3.2	3.1
	8	0.88	0.93	0.92	0.92	3.5	3.1	3.1	3.1
	10 Life Science	0.92	0.93	0.93	0.92	3.4	3.2	3.2	3.2

* CSTs named by number only are grade-level tests.