California Department of Education Assessment Development and Administration Division

California Alternate Performance Assessment

Technical Report

2014–15 Administration

Final Version Submitted March 11, 2016

Educational Testing Service

Contract No. 5417

Table of Contents

Acronyms and Initialisms Used in the *CAPA Technical Report* vi

Chapter 1: Introduction **1**
Background **1**
Test Purpose **1**
Test Content **2**
Intended Population **2**
Intended Use and Purpose of Test Scores **2**
Testing Window **3**
Significant CAASPP Developments in 2014–15 ..... **3**
Reduction in Paper Reporting 3
Origin of Demographic Data 3
Change in the Date for Students in Ungraded Programs 4
Reduced the Number of Required Tests 4
Suspended Reporting of Adequate Yearly Progress and the Academic Performance Index 4
Limitations of the Assessment **4**
Score Interpretation 4
Out-of-Level Testing 4
Score Comparison 4
Groups and Organizations Involved with the CAASPP System **4**
State Board of Education 4
California Department of Education 5
Contractor—Educational Testing Service 5
Overview of the Technical Report **5**
References **7**

Chapter 2: An Overview of CAPA for Science Processes **8**
Task (Item) Development **8**
Task Formats 8
Task (Item) Specifications 8
Item Banking 8
Task Refresh Rate 9
Test Assembly **9**
Test Length 9
Test Blueprints 9
Content Rules and Task Selection 9
Psychometric Criteria 10
Test Administration **10**
Test Security and Confidentiality 10
Procedures to Maintain Standardization 11
Universal Tools, Designated Supports, and Accommodations **12**
Scores **12**
Aggregation Procedures 12
Equating **13**
Post-Equating 13
Calibration 13
Scaling 13
Linear Transformation 14
References **16**

Chapter 3: Task (Item) Development **17**
Rules for Task Development **17**
Task Specifications 17
Expected Task Ratio 18
Selection of Task Writers **18**
Criteria for Selecting Task Writers 18
Task (Item) Review Process **19**
Contractor Review 19
Content Expert Reviews 21
Statewide Pupil Assessment Review Panel 22

Field Testing **23**
Stand-alone Field Testing 23
Embedded Field-test Tasks 23
CDE Data Review **23**
Item Banking **23**
References **25**
Chapter 4: Test Assembly **26**
Test Length **26**
Rules for Task Selection **26**
Test Blueprints 26
Content Rules and Task Selection 26
Psychometric Criteria 27
Projected Psychometric Properties of the Assembled Tests 28
Rules for Task Sequence and Layout 28
Chapter 5: Test Administration **29**
Test Security and Confidentiality **29**
ETS's Office of Testing Integrity 29
Test Development 29
Task and Data Review 30
Item Banking 30
Transfer of Forms and Tasks to the CDE 30
Security of Electronic Files Using a Firewall 31
Printing and Publishing 31
Test Administration 31
Test Delivery 31
Processing and Scoring 32
Data Management 32
Statistical Analysis 33
Reporting and Posting Results 33
Student Confidentiality 33
Student Test Results 33
Procedures to Maintain Standardization **34**
Test Administrators 34
CAPA Examiner's Manual 35
CAASPP Paper-Pencil Testing Test Administration Manual 36
Test Operations Management System Manuals 36
Universal Tools, Designated Supports, and Accommodations for Students with Disabilities **36**
Identification 36
Adaptations 37
Scoring 37
Testing Incidents **37**
Social Media Security Breaches 37
Testing Improprieties **38**
References **39**
Chapter 6: Performance Standards **40**
Background **40**
Standard-Setting Procedure **40**
Development of Competencies Lists 41
Standard-Setting Methodology **42**
Performance Profile Method 42
Results **42**
References **44**
Chapter 7: Scoring and Reporting **45**
Procedures for Maintaining and Retrieving Individual Scores **45**
Scoring and Reporting Specifications 46
Scanning and Scoring 46
Types of Scores **47**
Raw Score 47
Scale Score 47
Performance Levels 47
Score Verification Procedures **47**
Monitoring and Quality Control of Scoring 47
Score Verification Process 48

Overview of Score Aggregation Procedures **48**
Individual Scores 48
Reports Produced and Scores for Each Report **51**
Types of Score Reports 51
Student Score Report Contents 51
Student Score Report Applications 51
Criteria for Interpreting Test Scores **52**
Criteria for Interpreting Score Reports **52**
References **53**
Appendix 7.A—Scale Score Distribution Tables **54**
Appendix 7.B—Demographic Summaries **55**
Appendix 7.C—Types of Score Reports **56**
Chapter 8: Analyses **57**
Samples Used for the Analyses **57**
Classical Analyses **58**
Average Item Score 58
Polyserial Correlation of the Task Score with the Total Test Score 58
Reliability Analyses **59**
Subgroup Reliabilities and SEMs 60
Conditional Standard Errors of Measurement 61
Decision Classification Analyses **61**
Validity Evidence **62**
The Constructs to Be Measured 63
Interpretations and Uses of the Scores Generated 64
Intended Test Population(s) 64
Validity Evidence Collected 64
Evidence Based on Response Processes 66
Evidence of Interrater Agreement 67
Evidence Based on Internal Structure 67
Evidence Based on Consequences of Testing 67
IRT Analyses **68**
Post-Equating 68
Pre-Equating 68
Summaries of Scaled IRT *b*-values 69
Equating Results 69
Differential Item Functioning Analyses **69**
References **72**
Appendix 8.A—Classical Analyses: Task Statistics **74**
Appendix 8.B—Reliability Analyses **77**
Appendix 8.C—Validity Analyses **81**
Appendix 8.D—IRT Analyses **83**
Appendix 8.E—Disability Distributions **87**
Chapter 9: Quality Control Procedures **89**
Quality Control of Task Development **89**
Task Specifications 89
Task Writers 89
Internal Contractor Reviews 89
Assessment Review Panel Review 90
Statewide Pupil Assessment Review Panel Review 90
Data Review of Field-tested Tasks 90
Quality Control of the Item Bank **91**
Quality Control of Test Form Development **92**
Quality Control of Test Materials **92**
Collecting Test Materials 92
Processing Test Materials 92
Quality Control of Scanning **93**
Quality Control of Image Editing **93**
Quality Control of Answer Document Processing and Scoring **93**
Accountability of Answer Documents 93
Processing of Answer Documents 94
Scoring and Reporting Specifications 94
Storing Answer Documents 94

Quality Control of Psychometric Processes **94**
Quality Control of Task (Item) Analyses and the Scoring Process 94
Score Verification Process 95
Year-to-Year Comparison Analyses 95
Offloads to Test Development 96
Quality Control of Reporting **96**
Electronic Reporting 96
Excluding Student Scores from Summary Reports 97
Reference **98**
Chapter 10: Historical Comparisons **99**
Base Year Comparisons **99**
Examinee Performance **99**
Test Characteristics **100**
Appendix 10.A—Historical Comparisons Tables, Examinee Performance **101**
Appendix 10.B—Historical Comparisons Tables, Test Characteristics **104**

Tables

Table 1.1 Description of the CAPA for Science Assessment Levels 2
Table 2.1 CAPA for Science Items and Estimated Time Chart 9
Table 2.2 Scale Score Ranges for Performance Levels 15
Table 4.1 Statistical Targets for CAPA for Science Test Assembly 28
Table 4.2 Summary of 2012–13 CAPA for Science Projected Statistical Attributes 28
Table 7.1 Rubrics for CAPA for Science Scoring 46
Table 7.2 Summary Statistics Describing Student Scores: Science 48
Table 7.3 Percentages of Examinees in Each Performance Level 49
Table 7.4 Subgroup Definitions 50
Table 7.5 Types of CAPA for Science Reports 51
Table 7.A.1 Scale Score Frequency Distributions: Science, Levels I and III–V 54
Table 7.B.1 Demographic Percentage in Performance Level Summary for Science, All Examinees 55
Table 7.C.1 Score Reports Reflecting CAPA Results 56
Table 8.1 CAPA Raw Score Means and Standard Deviations: Tested cases with valid scores for 2012–13 and 2014–15 58
Table 8.2 Average Item Score and Polyserial Correlation 59
Table 8.3 Reliabilities and SEMs for the CAPA for Science 60
Table 8.4 DIF Flags Based on the ETS DIF Classification Scheme 71
Table 8.5 Subgroup Classification for DIF Analyses 71
Table 8.A.1 AIS and Polyserial Correlation: Level I, Science—Current Year (2015) and Original Year of Administration (2013) 74
Table 8.A.2 AIS and Polyserial Correlation: Level III, Science—Current Year (2015) and Original Year of Administration (2013) 74
Table 8.A.3 AIS and Polyserial Correlation: Level IV, Science—Current Year (2015) and Original Year of Administration (2013) 75
Table 8.A.4 AIS and Polyserial Correlation: Level V, Science—Current Year (2015) and Original Year of Administration (2013) 75
Table 8.A.5 Frequency of Operational Task Scores: Science 76
Table 8.B.1 Reliabilities and SEMs by Gender—Male 77
Table 8.B.2 Reliabilities and SEMs by Gender—Female 77
Table 8.B.3 Reliabilities and SEMs by Primary Ethnicity 77
Table 8.B.4 Reliabilities and SEMs by Primary Ethnicity for Economically Disadvantaged 78
Table 8.B.5 Reliabilities and SEMs by Primary Ethnicity for Not Economically Disadvantaged 78
Table 8.B.6 Reliabilities and SEMs by Disability 78
Table 8.B.7 Decision Accuracy and Decision Consistency: Level I, Science 79
Table 8.B.8 Decision Accuracy and Decision Consistency: Level III, Science 79
Table 8.B.9 Decision Accuracy and Decision Consistency: Level IV, Science 80
Table 8.B.10 Decision Accuracy and Decision Consistency: Level V, Science 80
Table 8.C.1 Interrater Agreement Analyses for Operational Tasks: Level I, Science 81
Table 8.C.2 Interrater Agreement Analyses for Operational Tasks: Level III, Science 81
Table 8.C.3 Interrater Agreement Analyses for Operational Tasks: Level IV, Science 81
Table 8.C.4 Interrater Agreement Analyses for Operational Tasks: Level V, Science 82
Table 8.D.1 Score Conversions: Level I, Science 83
Table 8.D.2 Score Conversions: Level III, Science 84
Table 8.D.3 Score Conversions: Level IV, Science 85
Table 8.D.4 Score Conversions: Level V, Science 86
Table 8.E.1 CAPA Primary Disability Distributions: Level I, Science 87

Table 8.E.2 CAPA Primary Disability Distributions: Level III, Science 87
Table 8.E.3 CAPA Primary Disability Distributions: Level IV, Science 88
Table 8.E.4 CAPA Primary Disability Distributions: Level V, Science 88
Table 10.A.1 Number of Examinees Tested, Scale Score Means, and Standard Deviations of CAPA for Science Across Base Year (2009), 2013, 2014, and 2015 101
Table 10.A.2 Percentage of Proficient and Above Across Base Year (2009), 2013, 2014, and 2015 101
Table 10.A.3 Percentage of Advanced Across Base Year (2009), 2013, 2014, and 2015 101
Table 10.A.4 Observed Score Distributions of CAPA for Science Across Base Year (2009), 2013, 2014, and 2015 for Science, Level I 102
Table 10.A.5 Observed Score Distributions of CAPA for Science Across Base Year (2009), 2013, 2014, and 2015 for Science, Level III 102
Table 10.A.6 Observed Score Distributions of CAPA for Science Across Base Year (2009), 2013, 2014, and 2015 for Science, Level IV 103
Table 10.A.7 Observed Score Distributions of CAPA for Science Across Base Year (2009), 2013, 2014, and 2015 for Science, Level V 103
Table 10.B.1 Average Item Score of CAPA for Science Operational Test Tasks Across Base Year (2009), 2013, 2014, and 2015 104
Table 10.B.2 Mean IRT *b*-values for Operational Test Tasks Across Base Year (2009), 2013, 2014, and 2015 104
Table 10.B.3 Mean Polyserial Correlation of CAPA for Science Operational Test Tasks Across Base Year (2009), 2013, 2014, and 2015 104
Table 10.B.4 Score Reliabilities of CAPA for Science Across Base Year (2009), 2013, 2014, and 2015 104
Table 10.B.5 SEM of CAPA for Science Across Base Year (2009), 2013, 2014, and 2015 104

Figures

Figure 3.1 The ETS Item Development Process for the California Assessment of Student Performance and Progress (CAASPP) System 17
Figure 8.1 Decision Accuracy for Achieving a Performance Level 62
Figure 8.2 Decision Consistency for Achieving a Performance Level 62

Acronyms and Initialisms Used in the *CAPA Technical Report*

1PPC	1-parameter partial credit	FIA	final item analysis
ADA	Americans with Disabilities Act	GENASYS	Generalized Analysis System
AERA	American Educational Research Association	HumRRO	Human Resource Research Organization
AIS	average task (item) score	IEP	individualized education program
APA	American Psychological Association	ICC	task (item) characteristic curve
API	Academic Performance Index	IRF	item response functions
ARP	Assessment Review Panel	IRT	task (item) response theory
AYP	Adequate Yearly Progress	IT	Information Technology
CAASPP	California Assessment of Performance and Progress	LEA	local educational agency
CALPADS	California Pupil Achievement Data System	MH	Mantel-Haenszel
CalTAC	California Technical Assistance Center	MR/ID	Mental retardation/intellectual disability
CAPA	California Alternate Performance Assessment	NCME	National Council on Measurement Education
CCR	*California Code of Regulations*	NPS	nonpublic, nonsectarian school
CDE	California Department of Education	QC	quality control
CDS	County-District-School	SBE	State Board of Education
CELDT	California English Language Development Test	SD	standard deviation
CI	confidence interval	SEM	standard error of measurement
CMA	California Modified Assessment	SFTP	secure file transfer protocol
CSEMs	conditional standard errors of measurement	SGID	School and Grade Identification sheet
CSTs	California Standards Tests	SMD	standardized mean difference
DIF	Differential Task (Item) Functioning	SPAR	Statewide Pupil Assessment Review
DPLT	designated primary language test	STAR	Standardized Testing and Reporting
DQS	Data Quality Services	STS	Standards-based Tests in Spanish
EC	*Education Code*	TBD	To Be Determined
EM	expectation maximization	TIF	test information function
ESEA	Elementary and Secondary Education Act	USDOE	United States Department of Education
ETS	Educational Testing Service	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 are designed to provide state-level input into instruction curricula and serve as a foundation for the state's school accountability programs.

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

During the 2014–15 administration, the CAASPP System had four components for the paper-pencil tests:

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

Test Purpose

The CAPA for Science outcomes are designed to show how well students in grades five, eight, and ten with significant cognitive disabilities are performing with respect to California's content standards in science that were adopted by the SBE in 1998. These standards describe what students should know and be able to do at each grade level; the CAPA for Science links directly to them at each grade level. IEP teams determine on a student-by-student basis whether a student takes the CSTs, CMA, or the CAPA for Science.

Test Content

The CAPA for Science are administered to students in one of four levels.

- Level I, for students with the most significant cognitive disabilities
- Level III, for students who are in grade five
- Level IV, for students who are in grade eight
- Level V, for students who are in grade ten

Table 1.1 displays CAPA for Science levels for tests administered in 2014–15 by grade and age ranges for ungraded programs.

Table 1.1 Description of the CAPA for Science Assessment Levels

Test Level	I	III	IV	V
Grades	5, 8, and 10	5	8	10
Age Ranges for Ungraded Programs *	10, 13, and15	10	13	15

* For students in ungraded programs and whose IEP teams designate that they take the CAPA for Science, their grade is determined by subtracting five from their chronological age on September 1, 2014.

Intended Population

Students with significant cognitive disabilities and an IEP take the CAPA for Science when they are unable to take the CSTs for Science with or without accommodations and/or non-embedded accessibility supports or the CMA for Science with accommodations. Most students eligible for the CAPA for Science take the assessment level that corresponds with their current school grade, but some students with complex and profound disabilities take the Level I assessment. Level I is administered to students in grades five, eight, and ten with the most significant cognitive disabilities who are receiving curriculum and instruction aligned to the CAPA for Science Level I blueprints.

The decision to place a student in CAPA for Science Level I must be made by the IEP team. Although it is possible that a student will take the CAPA for Science Level I throughout his or her education, the IEP team must reevaluate this decision each year. The decision to move a student from Level I to his or her grade-assigned CAPA for Science level is made on the basis of both the student's CAPA for Science performance from the previous year and on classroom assessments.

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

Intended Use and Purpose of Test Scores

The results for tests within the CAASPP System are used for two primary purposes, described in sections 60602.5 (a) and (a) (4). Sections 60602.5 (c) and (d) provide additional background on the tests. (Excerpted from the *EC* Section 60602 Web page at http://www.leginfo.ca.gov/cgi-bin/displaycode?section=edc&group=60001-61000&file=60600-60603 [outside source].)

"60602.5 (a) It is the intent of the Legislature in enacting this chapter to provide a system of assessments of pupils that has the primary purposes of assisting teachers, administrators,

and pupils and their parents; improving teaching and learning; and promoting high-quality teaching and learning using a variety of assessment approaches and item types. The assessments, where applicable and valid, will produce scores that can be aggregated and disaggregated for the purpose of holding schools and local educational agencies accountable for the achievement of all their pupils in learning the California academic content standards."

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

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

"60602.5 (d) It is the intent of the Legislature, insofar as is practically feasible and following the completion of annual testing, that the content, test structure, and test items in the assessments that are part of the statewide pupil assessment system become open and transparent to teachers, parents, and pupils, to assist stakeholders in working together to demonstrate improvement in pupil academic achievement. A planned change in annual test content, format, or design, should be made available to educators and the public well before the beginning of the school year in which the change will be implemented."

Testing Window

The CAPA for Science are administered within a 25-day window, which begins 12 days before and ends 12 days after the day on which 85 percent of the instructional year is completed.

The CAPA for Science are untimed. This assessment is administered individually and the testing time varies from one student to another, based on factors such as the student's response time and attention span. A student may be tested with the CAPA for Science over as many days as required within the LEA's testing window (*California Code of Regulations [CCR], Title 5,* Education, Division 1, Chapter 2, Subchapter 3.75, Article 2, § 855[a][2]; please note this section of *5 CCR* has been updated since the 2014–15 CAASPP administration).

Significant CAASPP Developments in 2014–15

Reduction in Paper Reporting

The Student Score Reports were the only printed reports received after test administration. LEAs were able to download preliminary and final aggregate and individual student data at the LEA and school levels. Student Score Reports were also available as downloadable PDFs.

Origin of Demographic Data

All student demographic data were derived from the California Longitudinal Pupil Achievement Data System (CALPADS) which caused some demographic fields used for data collection, such as those for student ethnicity/race and primary disability code, to be removed from answer documents. The fields remaining on answer documents are related to student identification and test conditions.

Change in the Date for Students in Ungraded Programs

The date used for determining the testing grade of a student in an ungraded program has changed; for 2014–15, it is September 1, 2014 (*EC* Section 48000[a][4]).

Reduced the Number of Required Tests

Because California is in transition to a new assessment for students with cognitive disabilities (the California Alternate Assessments), the number of non-computer-administered tests is reduced to include only grade-level science.

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 2015 does not include CAPA for Science results; AYP is calculated based on participation in ELA and mathematics only. Reporting of Academic Performance Index (API) data has been suspended.

Limitations of the Assessment

Score Interpretation

Teachers and administrators should not use CAASPP results in isolation to make inferences about instructional needs. In addition, 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 CAPA for Science results (CDE, 2013).

Out-of-Level Testing

With the exception of Level I, each CAPA for Science is designed to measure the content corresponding to a specific grade and is appropriate for students in the specific grade. Testing below a student's grade is not allowed for the CAPA for Science or any test in the CAASPP System; all students are required to take the 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 results for the CAPA for Science, the reviewer is limited to comparing results only within the same content area and CAPA level. For example, it is appropriate to compare scores obtained by students and/or schools on the 2013–14 CAPA for Science Level III test. Similarly, it is appropriate to compare scores obtained on the 2011–12 CAPA for Science Level III test with those obtained on the 2014–15 CAPA for Science test administered in 2014–15. It is not appropriate to compare scores obtained on Levels I and IV of the CAPA for Science. Since new score scales and cut scores were used for the 2009 CAPA for Science, results from tests administered in 2009 and later cannot meaningfully be compared to results obtained in previous years.

Groups and Organizations Involved with the CAASPP System

State Board of Education

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

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

California Department of Education

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.

Within the CDE, it is the District, School & Innovation Branch that oversees programs promoting innovation and improved student achievement. Programs include oversight of statewide assessments and the collection and reporting of educational data. (CDE, 2016)

Contractor—Educational Testing Service

The CDE and the SBE contract with Educational Testing Service (ETS) to develop, administer, and report the CAASPP assessments. ETS has overall responsibility for working with the CDE to implement and maintain an effective assessment system as well as having responsibility for producing and distributing materials, processing the tests, and producing reports. Activities directly conducted by ETS include the following:

- Overall management of the program activities;
- Development of all test items;
- Construction and production of test booklets and related test materials;
- Support and training provided to counties, LEAs, and independently testing charter schools;
- Implementation and maintenance of the Test Operations Management System for orders of materials and pre-identification services; and
- Completion of all psychometric activities.
- Production of all scannable test materials;
- Packaging, distribution, and collection of testing materials to LEAs and independently testing charter schools;
- Scanning and scoring of all responses, including performance scoring of the writing 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 2014–15 CAPA for Science. The technical report contains nine additional chapters as follows:

- Chapter 2 presents a conceptual overview of processes involved in a testing cycle for a CAPA for Science form. This includes test construction, test administration, generation of test scores, and dissemination of score reports. Information about the distributions of scores aggregated by subgroups based on demographics and the use of special

services is included, as are the references to various chapters that detail the processes briefly discussed in this chapter.

- Chapter 3 describes the procedures followed during the development of valid CAPA for Science tasks before the 2014–15 administration—in 2014–15, intact test forms (form 1 of each CAPA for Science level) from the 2012–13 administration were reused and there was no new item development. The chapter also explains the process of field-testing new tasks and the review of tasks by contractors and content experts.
- Chapter 4 details the content and psychometric criteria that guided the construction of the CAPA for Science forms reused in 2014–15.
- Chapter 5 presents the processes involved in the actual administration of the 2014–15 CAPA for Science with an emphasis on efforts made to ensure standardization of the tests. It also includes a detailed section that describes the procedures that were followed by ETS to ensure test security.
- Chapter 6 describes the standard-setting process previously conducted to establish new cut scores.
- Chapter 7 details the types of scores and score reports that are produced at the end of each administration of the CAPA for Science.
- Chapter 8 summarizes the results of the task (item)-level analyses performed during the spring 2014–15 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 CAPA for Science performance-level classifications, and the procedures designed to ensure the validity of CAPA for Science score uses and interpretations.
- Chapter 9 highlights the importance of controlling and maintaining the quality of the CAPA for Science.
- Chapter 10 presents historical comparisons of various task (item)- and test-level results for the past three years and for the 2009 base year.

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, § 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/Pages/Home.aspx.

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

California Department of Education. (2016b, January). *Organization.* Retrieved from http://www.cde.ca.gov/be/ms/po/sberesponsibilities.asp

Chapter 2: An Overview of CAPA for Science Processes

This chapter provides an overview of the processes involved in a typical test development and administration cycle for the California Alternate Performance Assessment (CAPA) for Science. Also described are the specifications maintained by ETS to implement each of those processes. In 2014–15, due to the use of intact test forms from the 2012–13 California Assessment of Student Performance and Progress (CAASPP) administration, neither test development nor equating activities occurred.

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.

Task (Item) Development

Task Formats

Each CAPA for Science task involves a prompt that asks a student to perform a task or a series of tasks. Each CAPA for Science task consists of the Task Preparation, the Cue/ Direction, and the Scoring Rubrics. These rubrics define the rules for scoring a student's response to each task.

Task (Item) Specifications

The CAPA for Science tasks were developed to measure California content standards adopted by the state in 1998 and designed to conform to principles of task writing defined by Educational Testing Service (ETS) (ETS, 2002). ETS maintained and updated a task specifications document, otherwise known as "task writer guidelines," for each CAPA for Science and used an item utilization plan to guide the development of the tasks for each content area. Task writing emphasis was determined in consultation with the California Department of Education (CDE).

The task specifications described the characteristics of the tasks that should be written to measure each content standard; tasks of the same type should consistently measure the content standards in the same way. To achieve this, the task specifications provided detailed information to CAPA for Science task writers.

The tasks selected for each CAPA for Science underwent an extensive review process that is designed to provide the best standards-based tests possible. Details about the task specifications, the task review process, and the item utilization plan are presented in Chapter 3, starting on page 17.

Item Banking

Before newly developed tasks were placed in the item bank, ETS prepared them for review by content experts and various external review organizations such as the Assessment Review Panels (ARPs), which are described in Chapter 3, starting on page 21; and the Statewide Pupil Assessment Review (SPAR) panel, described in Chapter 3, starting on page 22.

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

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

The CDE then reviewed those tasks with their statistical data flagged to determine whether they should be used operationally (see page 23 for more information about the CDE's data review). Any additional updates to task content and statistics were based on data collected from the operational use of the tasks. However, only the latest content of the task is retained in the bank at any time, along with the administration data from every administration that has included the task.

Further details on item banking are presented on page 23 in Chapter 3.

Task Refresh Rate

Prior to form reuse in the 2014–15 administration, the item utilization plan required that each year, 25 percent of tasks on an operational form were refreshed (replaced); these tasks remained in the item bank for future use. Intact forms from the 2012–13 administration were used in both the 2013–14 and 2014–15 administrations.

Test Assembly

Test Length

Each CAPA for Science consists of twelve tasks, including eight operational tasks and four field-test tasks. The number of tasks in each CAPA for Science and the expected time to complete each test is presented in Table 2.1. Testing times for the CAPA for Science are approximate. The CAPA for Science are administered individually, and the testing time varies from one student to another based on factors such as the student's response time and attention span. A student may be tested with the CAPA for Science over as many days as necessary within the LEA's selected testing window.

Table 2.1 CAPA for Science Items and Estimated Time Chart

CAPA	Items	Times
Level I Science	12	45 minutes
Level III Science	12	45 minutes
Level IV Science	12	45 minutes
Level V Science	12	45 minutes

Test Blueprints

ETS selected all CAPA for Science tasks to conform to the State Board of Education (SBE)-approved California content standards adopted in 1998 and test blueprints. The revised blueprints for the CAPA for Science were approved by the SBE in 2006 for implementation beginning in 2008. The test blueprints for the CAPA for Science are linked on the CDE CAASPP Science Assessments Web page at http://www.cde.ca.gov/ta/tg/ca/caasppscience.asp.

Content Rules and Task Selection

Intact test forms for the CAPA for Science from the 2012–13 administration were reused during the 2014–15 administration. Prior to the 2012–13 administration, test developers followed a number of rules when developing a new test form for a given CAPA for Science level. First and foremost, they selected tasks that met the blueprint for that level. Using the electronic item bank, assessment specialists began by identifying a number of linking tasks.

These were tasks that appeared in previous operational test administrations and were then used to equate the subsequent (new) test forms. After the linking tasks were approved, assessment specialists populated the rest of the test form.

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

Another consideration was the difficulty of each task. Test developers strived to ensure that there were some easy and some hard tasks and that there were a number of tasks in the middle range of difficulty. The detailed rules are presented in Chapter 4, which begins on page 26.

Psychometric Criteria

The staff assessed the projected test characteristics during the preliminary review of the assembled forms. The statistical targets used to develop the 2012–13 forms and the projected characteristics of the assembled forms are presented starting from page 27 in Chapter 4.

The tasks in test forms were organized and sequenced to meet the requirements of the content area. Further details on the arrangement of tasks during test assembly are described on page 28 in Chapter 4.

Test Administration

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

Test Security and Confidentiality

All tests within the California Assessment of Student Performance and Progress (CAASPP) System are secure documents. For the CAPA for Science administration, every person having access to test materials maintains the security and confidentiality of the tests. ETS's Code of Ethics requires that all test information, including tangible materials (such as test booklets, test questions, test results), confidential files, processes, and activities are kept secure. To ensure security for all tests that ETS develops or handles, ETS maintains an Office of Testing Integrity (OTI). A detailed description of the OTI and its mission is presented in Chapter 5 on page 29.

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 29.

- Test development
- Task and data review
- Item banking
- Transfer of forms and tasks 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 CAPA for Science processes are designed so that the tests are administered and scored in a standardized manner. ETS takes all necessary measures to ensure the standardization of the CAPA for Science, as described in this section.

Test Administrators

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

Staff at LEAs who are central to the processes include LEA CAASPP coordinators, test site coordinators, test examiners, proctors, and observers. The responsibilities for each of the staff members are included in the *CAASPP Paper-Pencil Testing Test Administration Manual* (CDE, 2015a); see page 36 in Chapter 5 for more information.

Test Directions

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

***CAPA Examiner's Manual*—**The manual used by test examiners to administer and score the CAPA for Science, to be followed exactly so that all students have an equal opportunity to demonstrate their academic achievement (See page 35 in Chapter 5 for more information.)

***CAASPP Paper-Pencil Testing Test Administration Manual*—**Test administration procedures for LEA CAASPP coordinators and test site coordinators (See page 36 in Chapter 5 for more information.)

Test Operations Management System (TOMS) manuals—Instructions for the Web-based modules that allow LEA CAASPP coordinators to set up test administrations, assign tests, and assign test settings; every module has its own user manual with detailed instructions on how to use TOMS (See page 36 in Chapter 5 for more information.)

Training in the form of "CAPA for Science Train-the-Trainer" workshops is available each January and is presented in live workshops and a Webcast, which is later archived. An LEA representative who takes the training can then train test site staff to train CAPA for Science examiners and observers. Video segments that model CAPA for Science task administration are made available during the school year; sample materials that support the training are available all year on the caaspp.org Web site, at http://www.caaspp.org/training/capa/.

Universal Tools, Designated Supports, and Accommodations

All public school students participate in the CAASPP System, including students with disabilities and English learners. Students with an individualized education program (IEP) and who have significant cognitive disabilities may take the CAPA for Science when they are unable to take a California Standards Test or California Modified Assessment with or without universal tools, designated supports, and/or accommodations.

Examiners may adapt the CAPA for Science in light of a student's instructional mode as specified in each student's IEP or Section 504 plan in one of two ways: (1) suggested adaptations for particular tasks, as specified in the task preparation; and (2) core adaptations that are applicable for many of the tasks. Details of the adaptations are presented in the core adaptations of the *CAPA Examiner's Manual* (CDE, 2015b).

As noted on the CDE CAPA Participation Criteria Web page, "Since examiners may adapt the CAPA for Science based on students' instruction mode, accommodations and modifications do not apply to CAPA." (CDE, 2015c)

Scores

The CAPA for Science total test raw scores equal the sum of examinees' scores on the operational tasks. The total raw scores differ in the score range across different CAPA for Science levels.

Raw scores for Level I range from 0 to 40; for the other CAPA for Science levels, the raw-score range is from 0 to 32. Total test raw scores are transformed to two-digit scale scores using the scaling process described starting on page 13. CAPA for Science results are reported through the use of these scale scores; the scores range from 15 to 60 for each test. Also reported are performance levels obtained by categorizing the scale scores into the following levels: far below basic, below basic, basic, proficient, and advanced. The state's target is for all students to score at the proficient or advanced level.

Detailed descriptions of CAPA for Science scores are found in Chapter 7, which starts on page 45.

Aggregation Procedures

In order to provide meaningful results to the stakeholders, CAPA for Science scores for a given grade and level 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 types of individual and demographic subgroup CAPA for Science scores aggregated at the state level.

Please note that aggregation is performed on valid scores only, which are cases where examinees met one or more of the following criteria:

1. Met attemptedness criteria
2. Had a valid combination of grade and CAPA level
3. Did not have a parental exemption

Individual Scores

Table 7.2 and Table 7.3 starting on page 48 in Chapter 7 provide summary statistics for individual scores aggregated at the state level, describing overall student performance on each CAPA for Science. Included in the tables are the possible and actual ranges and the

means and standard deviations of student scores, expressed in terms of both raw scores and scale scores. The tables also present statistical information about the CAPA for Science tasks.

Demographic Subgroup Scores

Statistics summarizing CAPA for Science student performance by test and for selected groups of students are provided in Table 7.B.1 on page 55 in Appendix 7.B. In these tables, students are grouped by demographic characteristics, including gender, ethnicity, English-language fluency, primary disability, and economic status. The tables show the numbers of students with valid scores in each group, scale score means and standard deviations, as well as percentage in performance level for each demographic group. Table 7.4 on page 50 provides definitions for the demographic subgroups included in the tables.

Equating

Post-Equating

In the years when the new forms were developed prior to the 2013–14 administration, each CAPA for Science form was equated to a reference form using a linking items nonequivalent groups data collection design and methods based on item response theory (IRT) (Hambleton & Swaminathan, 1985). The “base” or “reference” calibrations for the CAPA for Science were established by calibrating samples of data from the 2008–09 administration. Doing so established a scale to which subsequent item calibrations could be linked.

The procedure used for post-equating the CAPA for Science forms involves three steps: calibration, scaling, and linear transformation. Each of those procedures, as described below, is applied to all of the grade-level CAPA for Science tests during the tests’ original years of administration. Results were not post-equated for the 2014–15 administration.

During the 2014–15 administration, because the intact test forms were used from the 2012–13 CAPA for Science administration, the raw-to-scale-score conversion tables from the 2012–13 CAPA administration are directly applied to the 2014–15 administration.

Calibration

To obtain item calibrations, a proprietary version of the PARSCALE program and the Rasch partial credit model were used. The estimation process was constrained by setting a common discrimination value for all tasks equal to 1.0 / 1.7 (or 0.588). This approach was in keeping with previous CAPA for Science calibration procedures accomplished using the WINSTEPS program (Linacre, 2000).

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 is controlled by the constant discrimination parameters.

Scaling

In the years when the new forms were developed prior to the 2012–13 administration, calibration of the tasks were linked to the previously obtained reference scale estimates using linking tasks 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 task 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 tasks in a current test form that also appeared in last 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 tasks and removing tasks for which the difficulty estimates changed significantly. Tasks 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 a given score for the transformed new form item at ability θ_j, and

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

Based on established procedures, any linking items for which the WRMSD was greater than 0.625 for Level I and 0.500 for Levels III through V 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.

Linear Transformation

Once the new task calibrations for each test were transformed to the base scale, raw-score-to-theta scoring tables were generated. The thetas in these tables were then linearly transformed to a two-digit score scale that ranged from 15 to 60. Because the basic and proficiency cut scores were required to be equal to 30 and 35, respectively, the following formula was used to make this transformation:

$$\text{Scale Score} = (35 - \theta_{proficient} \times \left(\frac{35-30}{\theta_{proficient} - \theta_{basic}} \right)) + \left(\frac{35-30}{\theta_{proficient} - \theta_{basic}} \right) \times \theta \tag{2.2}$$

where,

θ represents the student ability,

$\theta_{proficient}$ represents the theta cut score for proficient on the 2008–09 base scale, and

θ_{basic} represents the theta cut score for basic on the 2008–09 base scale.

Complete raw-score-to-scale-score conversion tables for the 2014–15 CAPA for Science are presented in Table 8.D.1 through Table 8.D.4 in Appendix 8.D, starting on page 83. The raw scores and corresponding transformed scale scores are listed in those tables.

The scale scores defining the various performance levels are presented in Table 2.2.

Table 2.2 Scale Score Ranges for Performance Levels

CAPA Level	Far Below Basic	Below Basic	Basic	Proficient	Advanced
Level I Science	15	16 – 29	30 – 34	35 – 38	39 – 60
Level III Science	15 – 21	22 – 29	30 – 34	35 – 39	40 – 60
Level IV Science	15 – 19	20 – 29	30 – 34	35 – 39	40 – 60
Level V Science	15 – 20	21 – 29	30 – 34	35 – 38	39 – 60

References

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

California Department of Education. (2015b). *2015 California Alternate Performance Assessment (CAPA) examiner's manual*. Sacramento, CA. Retrieved from http://www.caaspp.org/rsc/pdfs/CAPA.examiners_manual.nonsecure.2015.pdf

California Department of Education. (2015c). *CAPA participation criteria.* Retrieved from http://www.cde.ca.gov/ta/tg/sr/participcritria.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. (2013). *WINSTEPS: Rasch measurement* (Version 3.23). Chicago, IL: MESA Press.

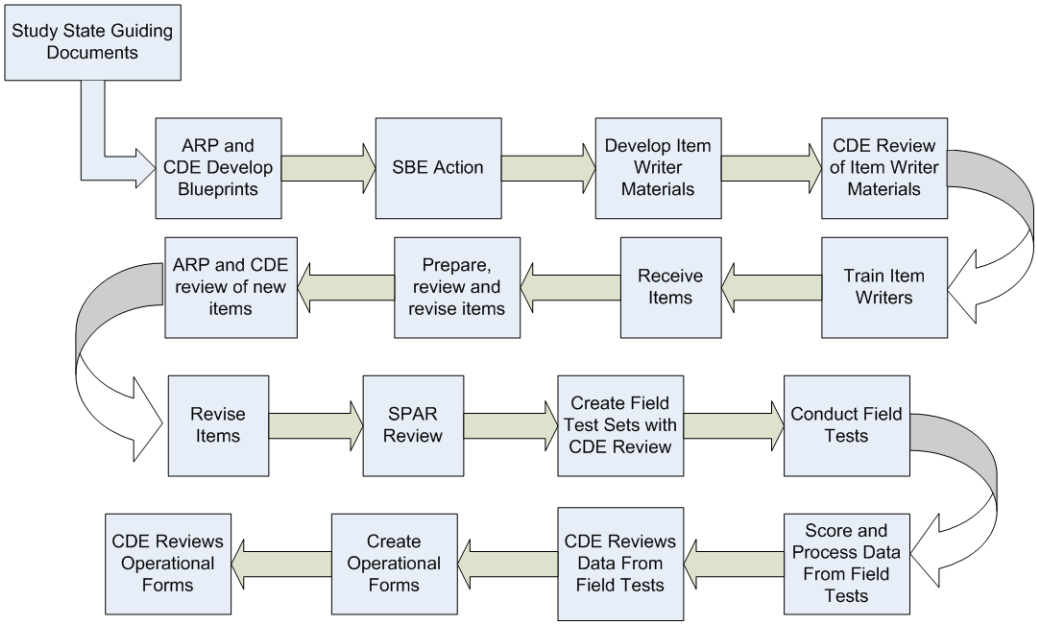
Stocking, M. L., and Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, pp. 201–10.

Chapter 3: Task (Item) Development

Intact test forms from the 2012–13 test administration were reused during the 2014–15 administration of the California Alternate Performance Assessment (CAPA). This reuse permitted score conversion tables from the previous administration to be used to look up student scores and performance levels. There was no new item (task) development for the 2014–15 forms.

The CAPA for Science tasks were developed to measure California's 1998 content standards for science and designed to conform to principles of item writing defined by Educational Testing Service (ETS) (ETS, 2002). Each CAPA for Science task went through a comprehensive development cycle as is described in Figure 3.1 below.

Figure 3.1 The ETS Item Development Process for the California Assessment of Student Performance and Progress (CAASPP) System



Rules for Task Development

The development of CAPA for Science tasks followed guidelines for task writing approved by the California Department of Education (CDE). These guidelines directed a task writer to assess a task for the relevance of the information being assessed, its relevance to the California content standards for science adopted in 1998, its match to the test and task specifications, and its appropriateness to the population being assessed. As described below, tasks were eliminated early in a rigorous task review process when they were only peripherally related to the test and task specifications, did not measure core outcomes reflected in the California content standards, or were not developmentally appropriate.

Task Specifications

ETS senior content staff led the task writers in the task development and review process. In addition, experienced ETS content specialists and assessment editors reviewed each task during the forms-construction process. The lead assessment specialists for each content area worked directly with the other ETS assessment specialists to carefully review and edit each task for such technical characteristics as quality, match to content standards, and conformity with California-approved task-writing practices. ETS followed the State Board of

Education (SBE)–approved item utilization plan to guide the development of the tasks for each content area.

Task specification documents included a description of the constructs to be measured and the California content standards for science adopted in 1998; tasks of the same type should consistently measure the content standards in the same way each year. The task specifications also provided specific and important guidance to task writers.

The task specifications described the general characteristics of the tasks for each content standard, indicated task types or content to be avoided, and defined the content limits for the tasks. More specifically, the specifications included the following:

- A statement of the strand or topic for the standard
- A full statement of the academic content standard, as found in each CAPA for Science blueprint
- The construct(s) appropriately measured by the standard
- A description of specific kinds of tasks to be avoided, if any (such as tasks about insignificant details)
- A description of appropriate data representations (such as charts, tables, graphs, or other artwork) for science tasks
- The content limits for the standard (such as one or two variables, maximum place values of numbers) for science tasks
- A description of appropriate stimulus cards (if applicable) for tasks

Expected Task Ratio

ETS developed the item utilization plan for the development of CAPA for Science tasks. The plan included strategies for developing tasks that permitted coverage of all appropriate standards for all tests at each grade level. ETS test development staff used this plan to determine the number of tasks to develop for each test.

The item utilization plan assumed that each year, 25 percent of items on an operational form would be refreshed (replaced); these items would remain in the item bank for future use. The item utilization 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, those that were difficult to measure well or for which there were few usable items.

For the 2014–15 CAPA for Science administration, field-test items were repeated as a part of the intact reused form. Detailed information about field testing was presented in the *2013 CAPA Technical Report*.

Selection of Task Writers

Criteria for Selecting Task Writers

The tasks for each CAPA for Science were written by individual task writers with a thorough understanding of the California content standards adopted in 1998. Applicants for task 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 task writers. Because most of the participants were current or former California educators, they were particularly knowledgeable about the standards assessed by the CAPA for Science. All task 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 tasks for standards-based assessments, including knowledge of the many considerations that are important when developing tasks to measure state-specific standards
- Previous experience in writing tasks in the content areas covered by CAPA for Science levels
- Familiarity, understanding, and support of the California content standards
- Current or previous teaching experience in California, when possible
- Knowledge about the abilities of the students taking the tests

Task (Item) Review Process

The tasks selected for the CAPA for Science underwent an extensive task review process that was designed to provide the best standards-based tests possible. This section summarizes the various reviews performed to ensure the quality of the CAPA for Science tasks and test forms—currently being reused—at the time the tasks and forms were developed.

Contractor Review

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

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

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

Throughout this multistep task review process, the lead content-area assessment specialists and development team members continually evaluated the relevance of the information being assessed by the task, its relevance to the California content standards for science adopted in 1998, its match to the test and task specifications, and its appropriateness to the population being assessed. Tasks that were only peripherally related to the test and task specifications, did not measure core outcomes reflected in the California content standards of 1998, or were not developmentally appropriate were eliminated early in this rigorous review process.

1. Internal Content Review

Test tasks and materials underwent two reviews by the content-area assessment specialists. These assessment specialists made sure that the test tasks 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 task

specifications. Assessment specialists reviewed each task on the basis of the following characteristics:

- Relevance of each task as the task relates to the purpose of the test
- Match of each task to the task specifications, including cognitive level
- Match of each task to the principles of quality task development
- Match of each task to the identified standard or standards
- Difficulty of the task
- Accuracy of the content of the task
- Readability of the task or stimulus card
- CAPA-level appropriateness of the task
- Appropriateness of any illustrations, graphs, or figures

Each task was classified with a code for the standard it was intended to measure. The assessment specialists checked all tasks against their classification codes, both to evaluate the correctness of the classification and to ensure that a given task was of a type appropriate to the outcome it was intended to measure. The reviewers could accept the task and classification as written, suggest revisions, or recommend that the task be discarded. These steps occurred prior to the CDE's review.

2. Internal Editorial Review

After the content-area assessment specialists reviewed each task, a group of specially trained editors also reviewed each task in preparation for consideration by the CDE and the ARPs. The editors checked tasks for clarity, correctness of language, appropriateness of language for the grade level assessed, adherence to the style guidelines, and conformity with accepted task-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 task 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
- Task accessibility for English-language learners

Content Expert Reviews

Assessment Review Panels

ETS was responsible for working with ARPs as tasks were developed for the CAPA. The ARPs are advisory panels to the CDE and ETS, and provided guidance on matters related to task development for the CAPA. The ARPs were responsible for reviewing all newly developed tasks for alignment to the California content standards for science adopted by the SBE in 1998. The ARPs also reviewed the tasks for accuracy of content, clarity of phrasing, and quality. In their examination of test tasks, the ARPs could raise concerns related to age/level 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 CAPA ARPs, including:

- Three or more years of general teaching experience in grades kindergarten through twelve and in the content areas (English–language arts [ELA], mathematics, or science);
- Bachelor's or higher degree in a grade or content area related to ELA, mathematics, or science;
- Knowledge and experience with the California content standards for ELA, mathematics, or science;
- Special education credential;
- Experience with more than one type of disability; and
- Three to five years as a teacher or school administrator with a special education credential.

Every effort is made to ensure that ARP committees include representation of genders and of the geographic regions and ethnic groups in California. Efforts are 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 local educational agencies (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 faculty members, 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. Due to the use of intact forms in 2014–15, no field test items were developed. Consequently, no ARP meetings were convened.

ARP Meetings for Review of CAPA for Science Tasks

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

- Overview of the purpose and scope of the CAPA for Science
- Overview of the CAPA for Science test design specifications and blueprints

- Analysis of the CAPA for Science task specifications
- Overview of criteria for reviewing constructed-response tasks
- Review and evaluation of tasks for bias and sensitivity issues

Criteria also involved more global factors. The ARPs also were trained on how to make recommendations for revising tasks.

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

Does the task:

- Measure the content standard?
- Match the test task specifications?
- Align with the construct being measured?
- Test worthwhile concepts or information?
- Reflect good and current teaching practices?
- Have wording that gives the student a full sense of what the task is asking?
- Avoid unnecessary wordiness?
- Reflect content that is free of bias against any person or group?

Is the stimulus, if any, for the task:

- Required in order to respond to the task?
- Likely to be interesting to students?
- Clearly and correctly labeled?
- Providing all the information needed to respond to the task?

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

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 tasks to be used statewide for the testing of students in California public schools, grades two through eleven. At the SPAR panel meetings, all new tasks were presented in binders for review. The SPAR panel representatives ensured that the test tasks conformed to the requirements of *Education Code* Section 60602. If the SPAR panel rejected specific tasks, the tasks were marked for rejection in the item bank and excluded from use on field tests. For the SPAR panel meeting, the task 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 task performance and to obtain statistics that can be used to assemble operational forms. However, because intact 2012–13 test forms were administered in 2014–15 and field-test items included in the test forms were analyzed in the 2012–13 administration cycle, no additional field test analyses were conducted.

Stand-alone Field Testing

In 2008, a pool of tasks was initially constructed for the CAPA aligned to the 2006 blueprints by administering the newly developed tasks 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 Tasks

Although a stand-alone field test is useful for developing a new test because it can produce a large pool of quality tasks, embedded field testing is generally preferred because the tasks 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 tasks are later administered operationally.

Such field testing involves distributing the tasks being field-tested within an operational test form. Different forms contain the same core set of operational tasks and different sets of field-test tasks.

Allocation of Students to Forms

The test forms for a given CAPA for Science were distributed by random assignment to LEAs so that a large representative sample of test takers responded to the field-test items embedded in these forms. The random assignment of specific forms ensured that a diverse sample of students took each field-test task. The students did not know which tasks were field-test tasks and which tasks were operational tasks; therefore, their motivation was not expected to vary over the two types of tasks (Patrick & Way, 2008).

CDE Data Review

From 2008 through 2013, once tasks were field-tested, ETS prepared tasks that failed to meet the desired statistical criteria and the associated statistics for review by the CDE. ETS provided tasks 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 tasks 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 (task) review was complete, the tasks were placed in the item bank along with their corresponding review information. Tasks that were accepted by the ARP, SPAR, and CDE were updated to a "field-test ready" status; tasks that were rejected were updated to a "rejected before use" status. ETS then delivered the tasks to the CDE by means of a delivery of the California electronic item bank. Subsequent updates to tasks were based on field-test and operational use of the tasks. However, only the latest content

of the task is in the bank at any given time, along with the administration data from every administration that has included the task.

After field-test or operational use, tasks that did not meet statistical specifications might be rejected; such tasks 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 task for its level of difficulty and discrimination as well as conformance to the Rasch partial credit model. Psychometricians also determined if the task functioned similarly for various subgroups of interest.

All unavailable tasks 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. Status and availability of a task were updated programmatically as tasks were presented for review, accepted or rejected, placed on a form for field-testing, presented for statistical review, and used operationally. All rejection 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 (California Standards Tests, California Modified Assessment, CAPA, and Standards-based Tests in Spanish. CAHSEE and CAASPP are currently consolidated in the California item banking system. ETS works with the CDE to obtain the data for assessments such as the CELDT, under contract with other vendors for inclusion into the item bank. ETS provides the item banking application using the local area network architecture and the relational database management system, SQL 2008, already deployed. ETS provides updated versions of the item bank to the CDE on an ongoing basis and works with the CDE to determine the optimum process if a change in databases is desired.

References

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

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

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

Chapter 4: Test Assembly

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

Test Length

The number of tasks in each CAPA for Science blueprint, adopted in 2006, 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 science, adopted in 1998, for each CAPA level. Also considered is the goal that the tests be short enough so that most of the students complete it in a reasonable amount of time.

Each CAPA for Science consists of 12 tasks, including eight operational tasks and four field-test tasks. Since intact forms from 2012–13 were used in the 2014–15 administration, see the 2013 *CAPA Technical Report* for more details on the distribution of items at each level.

Rules for Task Selection

Test Blueprints

Educational Testing Service (ETS) developed all CAPA for Science tasks to conform to the SBE-approved California content standards and test blueprints. The CAPA for Science blueprints were revised and approved by the SBE in 2006 for implementation beginning in 2008.

The California content standards approved by the SBE in 1998 were used as the basis for choosing tasks for the tests. The blueprints for the CAPA for Science can be found on the California Department of Education (CDE) Standardized Testing and Reporting CAPA Blueprints Web page at http://www.cde.ca.gov/ta/tg/sr/capablueprints.asp.

Content Rules and Task Selection

Intact test forms from the 2012–13 CAPA for Science administration were reused during the 2014–15 administration, when 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 tasks 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 tasks. These are tasks that appeared in a previous year's operational administration and were used to equate the administered test forms. Linking tasks were selected to proportionally represent the full blueprint. The selected linking tasks were also reviewed by psychometricians to ensure that the specific psychometric criteria were met.

After the linking tasks 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 task to a specified content standard. In selecting tasks, team members also tried to ensure that they included a variety of formats and content and that at least some of them included graphics for visual interest.

Another consideration was the difficulty of each task. Test developers strived to ensure that the tasks were spread evenly from easy to hard, with some easy and some hard tasks, and

a number of tasks in the middle range of difficulty. If tasks did not meet all content and psychometric criteria, staff reviewed the other available tasks 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 tasks would best serve the needs of the students taking the test. Chapter 3, starting on page 17, contains further information about this process.

Psychometric Criteria

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

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

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

Total test assembly targets were developed for each CAPA for Science. These targets were provided to test developers before a test construction cycle began.

Primary Statistical Targets

The total test targets, or primary statistical targets, used for assembling the CAPA for Science forms for the 2012–13 administration were the average and standard deviation of item difficulty based on the item response theory (IRT) *b*-parameters, average item score (AIS), and average polyserial correlation.

Due to the unique characteristics of the Rasch model, the information curve conditional on each ability level is determined by item difficulty (*b*-values) alone. In this case, the test information function (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 polyserial correlation describes the relationship between student performance on a polytomously 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 is related to the overall reliability of the test.

Assembly Targets

The target values for the CAPA for Science, presented in Table 4.1, were used in the 2014–15 test forms, which are intact test forms developed and used in the 2012–13 administration. These specifications were developed from the analyses of test forms administered in 2008–09, the base year in which test results were reported using new scales and new cut scores for the five performance levels: far below basic, below basic, basic, proficient, and advanced.

Table 4.1 Statistical Targets for CAPA for Science Test Assembly

CAPA Level	Target Mean *b*	Target SD *b*	Mean AIS	Mean Polyserial
Level I Science	–0.27	0.50	2.75	0.80
Level III Science	–0.76	0.50	2.20	0.80
Level IV Science	–0.61	0.50	2.20	0.80
Level V Science	–0.31	0.50	2.20	0.80

Projected Psychometric Properties of the Assembled Tests

In the years when the new forms were developed prior to the 2012–13 administration, ETS psychometricians performed a preliminary review of the technical characteristics of the assembled tests. Table 4.1 shows the projected statistical attributes of each CAPA for Science based on the most recent banked item statistics. These values can be compared to the target values in Table 4.2.

Table 4.2 Summary of 2012–13 CAPA for Science Projected Statistical Attributes

CAPA Level	Mean *b*	SD *b*	Mean AIS	Min AIS	Max AIS	Mean Polyserial
Level I Science	–0.29	0.12	2.90	2.37	3.11	0.78
Level III Science	–1.09	0.42	2.63	2.24	3.04	0.72
Level IV Science	–1.10	0.37	2.69	2.17	3.03	0.68
Level V Science	–0.51	0.62	2.57	1.97	3.42	0.70

Rules for Task Sequence and Layout

Linking tasks typically were placed in each form first; the sequence of the linking tasks was kept consistent from form to form. The initial tasks on a form and in each session are relatively easier than those tasks that follow so that many students can experience success early in each testing session. The remaining tasks were sequenced within a form and within a session by alternating easier and more difficult tasks.

Chapter 5: Test Administration

Test Security and Confidentiality

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

ETS's Office of Testing Integrity

The OTI is a division of ETS that provides quality assurance services for all testing programs administered by ETS and resides in the ETS Legal Department. The Office of Professional Standards Compliance of 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.

OTI's mission is to:

- Minimize any testing security violations that can impact the fairness of testing
- Minimize and investigate any security breach
- Report on security activities

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

Test Development

There was no new item development for the 2014–15 forms. Prior to 2012–13 administration, 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 development, 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 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.

Task and Data Review

As mentioned in Chapter 3, Assessment Review Panel (ARP) meetings were not held for the 2014–15 administration because there was no new task development for the 2014–15 CAPA for Science forms. However, before the 2013–14 administration, ETS facilitated ARP meetings every year to review all newly developed CAPA for Science tasks 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

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

Security of the electronic item banking system is of critical importance. The measures that ETS takes for ensuring 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 will be 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 Tasks 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 can 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. Task 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. All ETS data exchange and internal e-mail remain within the ETS firewall at all ETS locations, ranging from Princeton, New Jersey, to San Antonio, Texas, to Concord and Sacramento, California.

All electronic applications included in the Test Operations Management System (TOMS) (CDE, 2015a) remain protected by the ETS firewall software at all times. Due to the sensitive nature of the student information processed by TOMS, the firewall plays a significant role in maintaining an assurance of confidentiality in the users of this information.

Printing and Publishing

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

Test Administration

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

Test Delivery

Test security requires accounting for all secure materials before, during, and after each test administration. The LEA CAASPP coordinators are, therefore, required to keep all testing materials in central, locked storage except during actual test administration times. CAASPP test site coordinators are responsible for accounting for and returning all secure materials to the LEA CAASPP coordinator, who is responsible for returning them to the Scoring and Processing Centers. 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 may 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 may 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 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 test site coordinator is responsible for immediately reporting any security violation to the LEA CAASPP coordinator. The LEA CAASPP coordinator must contact the CDE immediately; the coordinator will be asked to follow up with a written explanation of the violation or suspected violation.

Processing and Scoring

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

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

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

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

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

Data Management

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

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

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

Statistical Analysis

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

Reporting and Posting Results

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

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

Student Confidentiality

To meet Elementary and Secondary Education Act and state requirements, LEAs must collect demographic data about students. This includes information about students' ethnicity, parent education, disabilities, whether the student qualifies for the National School Lunch Program, and so forth (CDE, 2015b). ETS takes precautions to prevent any of this information from becoming public or being used for anything other than testing purposes. These procedures are applied to all documents in which these student demographic data may appear.

Student Test Results

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

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

represent points of vulnerability, particularly to unauthorized access or denial of service. Routers, switches, firewalls, and gateways may possess little in the way of logical access.

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

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

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

- Scoring
- Transfer of scores by means of secure data exchange
- Reporting
- 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 be administered 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 CAPA for Science processes are designed so that the tests are administered and scored in a standardized manner.

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

Test Administrators

The CAPA for Science are administered in conjunction with the other tests that comprise the CAASPP Assessment System. The responsibilities for LEA and test site staff members are included in the *CAASPP Paper-Pencil Testing Test Administration Manual* (CDE, 2015b). This manual is described in the next section.

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

LEA CAASPP Coordinator

Each LEA designates an LEA CAASPP coordinator who is responsible for ensuring the proper and consistent administration of the CAASPP tests. LEAs include public school districts, statewide benefit charter schools, state board–authorized charter schools, county office of education programs, and charter schools testing independently from their home district.

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

CAASPP Test Site Coordinator

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

CAASPP test site coordinators are responsible for making sure that the school has the proper testing materials, distributing testing materials within a school, securing materials before, during, and after the administration period, answering questions from test examiners, preparing and packaging materials to be returned to the LEA after testing, and returning the materials to the LEA (CDE, 2015b).

Test Examiner

The CAPA for Science are administered to students individually 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. For the CAPA for Science, the test examiner must be a certificated or licensed school staff member (*5 CCR* Section 850[w]). Test examiners must follow the directions in the *CAPA Examiner's Manual* (CDE, 2015c) 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 Assessment System (5 *CCR* Section 850[y]). Test proctors must sign CAASPP Test Security Affidavits (5 *CCR* Section 859 [c]).

Observer

To establish scoring reliability, the test site coordinator and principal of the school should objectively and randomly select 10 percent of the students who will take the CAPA for Science at each level at each site to receive a second rating. The observer is a certificated or licensed employee (5 *CCR* Section 850[w]) who observes the administration of each task and completes a separate answer document for those students who are second-rated.

CAPA Examiner's Manual

The *CAPA Examiner's Manual* describes the CAPA for Science administrative procedures and scoring rubrics and contains the manipulative lists and all the tasks for all the CAPA for

Science tests at each level. Examiners must follow task preparation guidelines exactly (CDE, 2015c).

CAASPP Paper-Pencil Testing Test Administration Manual

Test administration procedures are to be followed exactly so that all students have an equal opportunity to demonstrate their academic achievement. The *CAASPP Paper-Pencil Testing Test Administration Manual* contributes to this goal by providing information about the responsibilities of LEA and test site coordinators, as well as those of the other staff involved in the administration cycle (CDE, 2015b). 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 Operations Management System Manuals

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

- **Test Administration Setup—**This module allows LEAs to determine and calculate dates for scheduling test administrations for LEAs, verify contact information for those LEAs, and request Pre-ID labels. (CDE, 2015d)
- **Student Paper-Pencil Test Registration—**This module allows LEAs to assign paper-pencil science tests to students in grades five, eight, and ten. (CDE, 2015e)
- **Set Condition Codes—**This module allows LEA CAASPP coordinators and CAASPP test site coordinators to apply condition codes (to note that a student was absent during testing, for example) to student records.

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

All public school students participate in the CAASPP Assessment System, including students with disabilities and English learners. ETS policy states that reasonable testing accommodations be provided to students 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 English learners take the Smarter Balanced for ELA and mathematics and the CST for Science under standard conditions. However, some students with disabilities and some English learners may need assistance when taking tests. This assistance takes the form of universal tools, designated supports, and accommodations. The "Universal Tools, Designated Supports, and Accommodations for the California Assessment of Student Performance and Progress" for administration of California statewide assessments are available on the CDE's Web site (CDE, 2015f). Because examiners may adapt the CAPA for Science in light of a student's instructional mode, universal tools, designated supports, and accommodations do not apply to the CAPA for Science.

Adaptations

Students eligible for the CAPA for Science represent a diverse population. Without compromising the comparability of scores, adaptations are allowed on the CAPA for Science to ensure the student's optimal performance. These adaptations are regularly used for the student in the classroom throughout the year. The CAPA for Science include two types of adaptations:

1. Suggested adaptations for particular tasks, as specified in the task preparation instructions; and
2. Core adaptations, which are applicable for many of the tasks.

The core adaptations may be appropriate for students across many of the CAPA for Science tasks, and are provided on page 23 of the nonsecure *CAPA Examiner's Manual* (CDE, 2015c).

Scoring

CAPA for Science tasks are scored using a 5-point holistic rubric (Level I) or a 4-point holistic rubric (Levels III, IV, and V) approved by the CDE. The rubrics include specific behavioral descriptors for each score point to minimize subjectivity in the rating process and facilitate score comparability and reliability. Student performance on each task is scored by one primary examiner, usually the child's teacher, or by another licensed or certificated staff member who is familiar to the student and who has completed the CAPA training. To establish scoring reliability, approximately 10 percent of students receive a second independent rating by a trained observer who is also a licensed or certificated staff member and has completed the CAPA training. The answer document indicates whether the test was scored by the examiner or the observer.

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 Paper-Pencil Testing Test Administration Manual* (CDE, 2015b).

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 social media security breaches associated with the CAPA for Science assessments in 2014–15.

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 *CAPA Examiner's Manual* (CDE, 2015c) and the *CAASPP Paper-Pencil Testing Test Administration Manual* (CDE, 2015b). 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. (2015a). *2015 Test Operations Management System.* Sacramento, CA. http://www.caaspp.org/administration/tms/

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

California Department of Education (2015c). *2015 California Alternate Performance Assessment (CAPA) examiner's manual.* Sacramento, CA. Retrieved from http://www.caaspp.org/rsc/pdfs/CAPA.examiners_manual.nonsecure.2015.pdf

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

California Department of Education. (2015e). *2015 California Assessment of Student Performance and Progress Test Operations Management System: Student Paper-Pencil Test Registration User Guide.* Sacramento, CA. Retrieved from http://www.caaspp.org/rsc/pdfs/CAASPP.ppt-registration.2015.pdf

California Department of Education. (2015f). Universal Tools, Designated Supports, and Accommodations for the California Assessment of Student Performance and Progress. Sacramento, CA. Retrieved from http://www.cde.ca.gov/ta/tg/ai/caasppmatrix1.asp

Chapter 6: Performance Standards

Background

The California Alternate Performance Assessment (CAPA) for English–Language Arts (ELA), Mathematics, and Science were first administered as a part California's standardized testing program in 2003. Subsequently, the CAPA were revised to better link these tests to the grade-level California content standards adopted by the State Board of Education (SBE) in 1998. The revised blueprints for the CAPA were approved by the SBE in 2006 for implementation beginning in 2008; new tasks were developed to meet the revised blueprints and field-tested.

From September 16 to 18, 2008, Educational Testing Service (ETS) conducted a standard-setting workshop in Sacramento, California, to recommend cut scores that delineated the revised performance standards for the CAPA for ELA and mathematics levels I through V and the CAPA for science levels I and III through V (the CAPA for Science is not assessed in Level II). The performance standards were defined by the SBE as far below basic, below basic, basic, proficient, and advanced.

Performance standards are developed from a general description of each performance level (policy-level descriptors) and the associated competencies lists, which operationally define each level. Cut scores numerically define the performance levels. This chapter describes the process of developing performance standards, which were first applied to the CAPA operational tests in the 2008–09 administration.

California employed carefully designed standard-setting procedures to facilitate the development of performance standards for each CAPA. The standard-setting method used for the CAPA was the Performance Profile Method, a holistic judgment approach based on profiles of student test performance for the areas of ELA and mathematics at all five test levels and for science at levels I, III, IV, and V (ETS, 2003). Four panels of educators were convened to recommend cut scores; one panel for each content area focused on all levels above Level I and a separate panel focused on Level I. After the standard setting, ETS met with representatives of the California Department of Education (CDE) to review the preliminary results and provided an executive summary of the procedure and tables that showed the panel-recommended cut scores and impact data. The final cut scores were adopted by the SBE in November 2008. An overview of the standard setting workshop and final results are provided below; see the technical report for the standard setting (Educational Testing Service [ETS], 2008a) for more detailed information.

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 CAPA standard setting were selected based on the following characteristics:

- Familiarity with the California content standards
- Direct experience in the education of students who take the CAPA
- Experience administering the CAPA

Panelists were recruited to be representative of the educators of the state's CAPA-eligible students (ETS, 2008b). Panelists were assigned to one of four panels (Level I, ELA, mathematics, or science) such that the educators on each panel had experience administering CAPA across the levels in the content area(s) to which they were assigned.

As with other standard setting processes, panelists participating in the CAPA workshop followed these steps, which included training and practice prior to making judgments:

1. Prior to attending the workshop, all panelists received a pre-workshop assignment. The task was to review, on their own, the content standards upon which the CAPA tasks are based and take notes on their own expectations for students at each performance level. This allowed the panelists to understand how their perceptions may relate to the complexity of content standards.
2. 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. Impact data included the percentage of students assessed in a previous test administration of the test who would fall into each performance level, given the panelists' judgments of cut scores.
3. Panelists became familiar with the tasks by reviewing the actual test and the rubrics and then assessing and discussing the demands of the tasks.
4. Panelists reviewed the draft list of competencies as a group, noting the increasing demands of each subsequent level. The competencies lists were developed by a subset of the standard-setting panelists based on the California content standards and policy-level descriptors (see the next section). In this step, they began to visualize the knowledge and skills of students in each performance level and the differences between levels.
5. Panelists identified characteristics of a "borderline" test-taker or "target student." This student is defined as one who possesses just enough knowledge of the content to move over the border separating a performance level from the performance level below it.
6. After training in the method was complete and confirmed through an evaluation questionnaire, panelists made individual judgments. Working in small groups, they discussed feedback related to other panelists' judgments and feedback based on student performance data (impact data). Note that no impact data were presented to the Level I panel due to the change in the Level I rubric. Panelists could revise their judgments during the process if they wished.
7. The final recommended cut scores were based on an average of panelists' judgment scores at the end of three rounds. For the CAPA, 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 adoption.

Development of Competencies Lists

Prior to the CAPA standard-setting workshop, ETS facilitated a meeting in which a subset of the standard-setting panelists was assembled to develop lists of competencies based on the California content standards and policy-level descriptors. Four panels of educators were assembled to identify and discuss the competencies required of students in the CAPA levels and content areas for each performance level (below basic, basic, proficient, and advanced). Panels consisted of educators with experience working with students who take

the CAPA. Panelists were assigned to one of four panels (Level I, ELA, mathematics, or science) based on experience working with students and administering the CAPA. At the conclusion of the meeting, the CDE reviewed the draft lists and delivered the final lists for use in standard setting. The lists were used to facilitate the discussion and construction of the target student definitions during the standard-setting workshop.

Standard-Setting Methodology

Performance Profile Method

Because of the small number of tasks and the fact that all CAPA tasks are constructed response items, ETS applied a procedure that combined the Policy Capturing Method (Plake & Hambleton, 2001; Jaeger, 1995a; Jaeger, 1995b) and the Dominant Profile Method (Plake & Hambleton, 2001; Plake, Hambleton, & Jaeger, 1997; Putnam, Pence, & Jaeger, 1995). Both methods are holistic methods in that they ask panelists to make decisions based on an examinee's score profile or performance rather than on each separate item.

The combined procedure that was used in 2008 is called the Performance Profile Method in this report. The procedure was a modification to the Performance Profile Method used for the CAPA standard setting in 2003 (ETS, 2003). The task for panelists was to mark the raw score representing the competencies a student should have for the basic, proficient, and advanced performance levels; cut scores for below basic and far below basic performance levels were set statistically.

For each test, materials were developed so that panelists could review score patterns, or performance profiles, for the eight CAPA tasks; panelists used the profiles and corresponding raw scores to make cut-score judgments. Profiles for Levels II–V were selected using 2008 student performance data. Profiles for Level I were informed by 2008 student performance data; however, due to a change in the Level I rubric after the 2008 test administration, the selection of Level I profiles also relied on verification by CAPA assessment experts, taking into account the changes in the Level I rubric (see Chapter 7 for more information on the rubric change).

The student profiles were presented at selected raw score points in an increasing order. For most raw score points, two to three profiles were presented; but in the portion of the score range where total scores were achieved by a large group of students as indicated by the operational data, up to five profiles were presented. While it is recognized that any number of combinations of item ratings may result in the same total raw scores, the intent in the Performance Profile Method is to use a cut score that is compensatory in nature. Therefore, profiles within the same total raw score are ordered randomly. Panelists were instructed that it is permissible to select total raw scores "between" the presented raw score profiles as their recommended cut score judgment for any level.

More details regarding the process implemented for the CAPA standard setting and results summary can be found in the standard-setting technical report (ETS, 2008a).

Results

The cut scores obtained as a result of the standard setting process were expressed in terms of raw scores; the panel median score after three rounds of judgments was the cut score recommendation for each level. These scores were transformed to scale scores that range between 15 and 60.

The cut score for the basic performance level was set equal to a scale score of 30 for every test level and content area; this means that a student must earn a score of 30 or higher to achieve a basic classification. The cut score for the proficient level was set equal to 35 for each test level and content area; this means that a student must earn a score of 35 or higher to achieve a proficient classification.

The cut scores for the other performance levels usually vary by test level and content area. They were derived using procedures based on item response theory (IRT). Please note that in the case of polytomously scored items, the IRT test characteristic function is the sum of the item response functions (IRF), where the IRF of an item is the weighted sum of the response functions for each score category (weighted by the scores of the categories).

Each raw cut score for a given test was mapped to an IRT *theta* (θ) using the test characteristic function and then transformed to the scale score metric using the following equation:

$$\text{Scale Cut Score}=(35-\theta_{proficient}\times\left(\frac{35-30}{\theta_{proficient}-\theta_{basic}}\right))+\left(\frac{35-30}{\theta_{proficient}-\theta_{basic}}\right)\times\theta_{cut\text{-}score} \tag{6.1}$$

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.

The scale-score ranges for each performance level are presented in Table 2.2 on page 15. 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 7.3 on page 49 in Chapter 7 presents the percentages of examinees meeting each performance level in 2014–15.

References

Educational Testing Service. (2003). *CAPA standard setting technical report* (California Department of Education Contract Number 5417). Princeton, NJ: Author.

Educational Testing Service. (2008a). *Technical report on the standard setting workshop for the California Alternate Performance Assessment. December 29, 2008* (California Department of Education Contract Number 5417). Princeton, NJ: Author.

Educational Testing Service K–12 Statistical Analysis Group. (2008b). *A study to examine the effects of changes to the CAPA Level I rubric involving the hand-over-hand prompt*, Unpublished memorandum. Princeton, NJ: Author.

Jaeger, R. M. (1995a). Setting performance standards through two-stage judgmental policy capturing. *Applied Measurement in Education,* 8, pp. 15–40.

Jaeger, R. M. (1995b). Setting standards for complex performances: An iterative, judgmental policy-capturing strategy. *Educational Measurement: Issues and Practice,* 14 (4), pp. 16–20.

Plake, B. S., & Hamilton, R.K. (2001). The analytic judgment method for setting standards on complex performance assessments. In G.J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives,* (pp. 283–312). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Plake, B., Hamilton, R., & Jaeger, R. M. (1997). A new standard setting method for performance assessments: The dominant profile judgment method and some field-test results. *Educational and Psychological Measurement,* 57, pp. 400–11.

Putnam, S.E., Pence, P., & Jaeger, R. M. (1995). A multi-stage dominant profile method for setting standards on complex performance assessments. *Applied Measurement in Education,* 8, pp. 57–83.

Chapter 7: Scoring and Reporting

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

Procedures for Maintaining and Retrieving Individual Scores

Each California Alternate Performance Assessment (CAPA) for Science is composed of performance tasks; each test includes eight performance tasks that are scored by a trained examiner using a rubric that depends on the test level being assessed. After the student has responded to a task, the examiner marks the score using the corresponding circle on the student's answer document.

In the 2014–15 administration, preliminary individual student results were available for download prior to the printing of paper reports. This electronic reporting was made possible through the Online Reporting System.

Scoring Rubric

The scoring rubric represents the guideline for scoring the task. The rubric varies according to the CAPA for Science level. The rubric for CAPA for Science in Level I has a range of 0–5, with 5 being the maximum score. The rubric for CAPA for Science in Levels III, IV, and V has a range of 0–4, with 4 being the maximum score.

Beginning with the administration of the 2008–09 CAPA for Science, the Level I rubric was changed to take into account issues related to scoring students who required a hand-over-hand prompt (ETS, 2008). ETS believed there was a significant difference between levels of prompting when dealing with this special population of students as evidenced by the amount of special education research that deals exclusively with prompting hierarchies. A child with significant cognitive disabilities who is able to complete a task successfully at one level of prompting may take weeks or months to increase his or her proficiency in that task in order to be able to complete the task successfully at a less intrusive level of prompting. The differences within prompting levels are the reason why ETS supported a rubric that differentiates between levels of prompting and scores the responses accordingly. For Level I science, all tasks are scored using the same rubric. For all other levels, the rubric is specific to the task. Both rubrics are presented in Table 7.1. Note that a score of zero in Level I indicates that the student did not orient toward a task after multiple prompts had been utilized. In Levels III–V, a score of zero implies that the student did not attempt the task. In both cases, the score is defined as "No Response" for the purpose of scoring the task.

Table 7.1 Rubrics for CAPA for Science Scoring

Level I		Levels III–V	
Score Points	**Description**	**Score Points**	**Description**
5	Correct with no prompting		
4	Correct with verbal or gestural prompt	4	Completes task with 100 percent accuracy
3	Correct with modeled prompt	3	Partially completes task (as defined for each task)
2	Correct with hand-over-hand prompt (student completes task independently)	2	Minimally completes task (as defined for each task)
1	Orients to task or incorrect response after attempting the task independently	1	Attempts task
0	No response	0	Does not attempt task

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

Scoring and Reporting Specifications

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

- Scoring Rules—Describes the rules for how and when scores are reported, including whether or not the student data will be part of the CAPA for Science reporting and how scores are reported under certain conditions (for example, when a student was not tested)
- Include Indicators—Defines the appropriate codes to use when a student does not take or complete a test or when a score will not be reported

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

Scanning and Scoring

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

Each 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

Raw Score

There are four test levels and eight operational tasks per level for the CAPA for Science. Performance scoring for Level I is based on a rubric with a range of 0–5 with a maximum score of 5. Performance scoring for Levels III–V is based on a rubric with a range of 0–4 with a maximum score of 4. For all CAPA for Science tests, the total test raw score equals the sum of the eight operational task scores. The raw scores for Level I range from 0 to 40; for the other CAPA for Science levels, the raw score range is from 0 to 32.

Scale Score

Raw scores obtained on each CAPA for Science test are converted to two-digit scale scores using the calibration process described in Chapter 2 on page 13. Scale scores range from 15 to 60 on each CAPA for Science. The scale scores of examinees that have been tested in different years at a given CAPA for Science test level 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

For the CAPA for Science content-area tests, the performance of each student is categorized into one of the following performance levels:

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

For all CAPA for Science tests, the cut score for the basic performance level is 30; this means that a student must earn a scale score of 30 or higher to achieve a basic classification. The cut score for the proficient performance level is 35; this means that a student must earn a scale score of 35 or higher to achieve a proficient classification. The cut scores for the other performance levels usually vary by level and content area.

Score Verification Procedures

Various necessary measures are taken to ascertain that the student scores are computed accurately.

Monitoring and Quality Control of Scoring

Scorer Selection

Careful consideration is given to the selection of examiners for proper administration and scoring of the CAPA for Science. It is preferred that the special education teacher or case carrier who regularly works with the student being tested administer and score the test. The examiner is required to be certificated or licensed and have successfully completed comprehensive training on CAPA for Science administration.

If the examiner or case carrier is not available to administer the test, it may be administered and scored by another CAPA for Science-trained staff member such as a school psychologist; speech, physical, or occupational therapist; program specialist; or certified teacher, principal, or assistant principal. This individual should have experience working with students with significant cognitive disabilities and must be trained to administer the CAPA for Science (CDE, 2015a).

Quality Control

Each student's responses to the CAPA for Science tasks are rated by a single examiner; the total score is based on that rater's ratings. In addition, approximately 10 percent of students at each test site are also rated by an observer to provide data that can be used to assess the accuracy and reliability of the scores. The observer, who is expected to meet the same qualification requirements as an examiner, scores the test at the same time as the test is being administered, but independently of the examiner. The score from the observer does not count toward the student's CAPA for Science score.

Score Verification Process

After ETS applies the scoring tables to generate scale scores for each student, ETS verifies scale scores by conducting QC and reasonableness checks, which are described in Chapter 9 on page 96.

Overview of Score Aggregation Procedures

In order to provide meaningful results to the stakeholders, CAPA for Science scores for a given content area 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 CAPA for Science scores.

Individual Scores

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

Score Distributions and Summary Statistics

Summary statistics that describe student performance on each CAPA for Science are presented in Table 7.2. Included in this table is the number of tasks 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. In addition, summary statistics for the operational tasks on each test are provided.

Table 7.2 Summary Statistics Describing Student Scores: Science

Level	I	III	IV	V
		Scale Score Information		
Number of examinees	3,706	3,323	3,315	3,267
Mean score	37.55	35.99	36.01	35.89
SD *	10.84	5.43	5.53	4.83
Possible range	15–60	15–60	15–60	15–60
Obtained range	15–60	15–60	15–60	15–60
Median	37.00	36.00	36.00	36.00
Reliability	0.89	0.86	0.85	0.84
SEM †	3.90	2.28	2.36	2.35
		Raw Score Information		
Mean score	24.45	20.71	21.73	20.37
SD *	11.67	6.13	6.16	5.86
Possible range	0–40	0–32	0–32	0–32
Obtained range	0–40	0–32	0–32	0–32
Median	26.00	21.00	22.00	21.00
Reliability	0.89	0.86	0.85	0.84
SEM †	3.62	2.02	2.12	1.94

Level	I	III	IV	V
		Task Information		
Number of tasks	8	8	8	8
Mean AIS ‡	3.07	2.6	2.73	2.56
SD AIS ‡	0.21	0.29	0.26	0.55
Min. AIS	2.6	2.25	2.29	1.94
Max. AIS	3.32	3.12	2.99	3.36
Possible range	0-5	0-4	0-4	0-4
Mean polyserial	0.8	0.76	0.74	0.75
SD polyserial	0.03	0.04	0.04	0.06
Min. polyserial	0.76	0.69	0.67	0.63
Max. polyserial	0.83	0.81	0.79	0.82

* Standard Deviation | † Standard Error of Measurement | ‡ Average Item (Task) Score

The percentages of students in each performance level are presented in Table 7.3.

The numbers in the summary tables may not match exactly the results reported on the CDE Web site because of slight differences in the samples used to compute the statistics. The P2 data file was used for the analyses in this chapter. This file contained the entire test-taking population but did not include corrections of demographic data through the California Longitudinal Pupil Assessment Data System (CALPADS). In addition, students with invalid scores were excluded from the tabled results.

Table 7.3 Percentages of Examinees in Each Performance Level

CAPA Level	Far Below Basic	Below Basic	Basic	Proficient	Advanced
Level I Science	8%	7%	18%	27%	40%
Level III Science	2%	4%	25%	53%	16%
Level IV Science	1%	6%	25%	49%	18%
Level V Science	1%	4%	27%	43%	24%

Table 7.A.1 in Appendix 7.A on page 54 shows the distributions of scale scores for each CAPA for Science. The results are reported in terms of three score intervals. A cell value of “N/A” indicates that there are no obtainable scale scores within that scale-score range for the particular CAPA for Science.

Group Scores

Statistics summarizing student performance by content area for selected groups of students are provided on page 55 in Table 7.B.1 for the CAPA for Science. In this table, students are grouped by demographic characteristics, including gender, ethnicity, English-language fluency, economic status, and primary disability. The tables show, for each demographic group, the numbers of valid cases and percentages of students in each performance level by demographic group. Table 7.4 provides definitions of the demographic groups included in the tables.

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

Table 7.4 Subgroup Definitions

Subgroup	Definition
Gender	• Male • Female
Ethnicity	• African American • American Indian or Alaska Native • Asian – Asian Indian – Cambodian – Chinese – Hmong – Japanese – Korean – Laotian – Vietnamese – Other Asian • Hispanic or Latino • Pacific Islander – Guamanian – Native Hawaiian – Samoan – Tahitian – Other Pacific Islander • Filipino • White (not Hispanic)
English-language Fluency	• English only • Initially fluent English proficient • English learner • Reclassified fluent English proficient • To Be Determined (TBD)
Economic Status	• Not economically disadvantaged • Economically disadvantaged
Primary Disability	• Intellectual disability • Hearing impairment • Speech or language impairment • Visual impairment • Emotional disturbance • Orthopedic impairment • Other health impairment • Specific learning impairment • Deaf-blindness • Multiple disabilities • Autism • Traumatic brain injury

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.

Types of Score Reports

There are three categories of CAPA for Science reports. These categories and the specific reports in each category are given in Table 7.5.

Table 7.5 Types of CAPA for Science Reports

1. Electronic Summary Report	▪ CAASPP Aggregate Report (includes subgroups)
2. Individual Reports for Students (Paper and Online)	▪ CAASPP Student Data File ▪ CAASPP Student Score Report for CAPA
3. Internet Reports	▪ CAPA for Science Summary Scores (state, county, LEA, school)

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

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

Student Score Report Contents

The CAASPP Student Score Report provides scale scores and performance level results for the CAPA for Science taken. Scale scores are reported on a scale ranging from 15 to 60. The performance levels reported are: far below basic, below basic, basic, proficient, and advanced.

Further information about the CAASPP Student Score Report and the other reports is provided in Appendix 7.C on page 56.

Student Score Report Applications

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

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

CAPA results help LEAs and schools identify strengths and weaknesses in their instructional programs. Each year, LEAs and school staff examine CAPA for Science results at each level tested. Their findings are used to help determine:

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

Criteria for Interpreting Test Scores

An LEA may use CAPA for Science results to help make decisions about student placement, promotion, retention, or other considerations related to student achievement. Since the adoption of new science standards in California in 2014, school curricula are no longer aligned to the 1998 standards; therefore, users should be careful when interpreting student scores. It is also 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 CAPA for Science results (CDE, 2015b). 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 CAPA for Science, the user is limited to comparisons within the same content area and level. This is because the score scales are different for each content area and level. The user may compare scale scores for the same content area and level, 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 level and content area across years. Comparing scores obtained in different levels or content areas should be avoided because the results are not on the same scale. Comparisons between raw scores should be limited to comparisons within not only content area and level but also test year. Since new score scales and cut scores were applied beginning with the 2008–09 test results, results from this and subsequent years cannot meaningfully be compared to results obtained in prior years. For more details on the criteria for interpreting information provided on the score reports, see the *201-15 CAASPP Post-Test Guide* (CDE, 2015b).

References

California Department of Education. (2015a), *2015 CAPA examiner's manual.* Sacramento, CA. Retrieved from http://www.caaspp.org/rsc/pdfs/CAPA.examiners_manual.nonsecure.2015.pdf

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

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

Educational Testing Service. (2008) *A study to examine the effects of changes to the CAPA Level I rubric involving the hand-over-hand prompt,* Unpublished memorandum, Princeton, NJ: Author.

Appendix 7.A—Scale Score Distribution Tables

In Appendix 7.A, a cell value of "N/A" indicates that there are no obtainable scale scores within that scale-score range for the particular CAPA.

Table 7.A.1 Scale Score Frequency Distributions: Science, Levels I and III–V

Scale Score	Level I Freq.	Level I Pct.	Level III Freq.	Level III Pct.	Level IV Freq.	Level IV Pct.	Level V Freq.	Level V Pct.
60	374	10.09	48	1.44	45	1.36	27	0.83
57–59	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
54–56	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
51–53	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
48–50	127	3.43	N/A	N/A	N/A	N/A	N/A	N/A
45–47	81	2.19	50	1.50	88	2.65	39	1.19
42–44	397	10.71	176	5.30	132	3.98	146	4.47
39–41	495	13.36	441	13.27	548	16.53	581	17.78
36–38	799	21.56	1156	34.79	1016	30.65	1170	35.81
33–35	624	16.84	888	26.72	902	27.21	841	25.74
30–32	251	6.77	359	10.80	347	10.47	270	8.26
27–29	103	2.78	95	2.86	126	3.80	106	3.24
24–26	102	2.75	38	1.14	33	1.00	19	0.58
21–23	31	0.84	18	0.54	17	0.51	20	0.61
18–20	34	0.92	6	0.18	18	0.54	12	0.37
15–17	288	7.77	48	1.44	43	1.30	36	1.10

Appendix 7.B—Demographic Summaries

Table 7.B.1 Demographic Percentage in Performance Level Summary for Science, All Examinees

	Number Tested	Far Below Basic	Below Basic	Basic	Proficient	Advanced
All valid scores	13,611	3%	5%	24%	43%	25%
Male	9,001	3%	5%	23%	43%	26%
Female	4,610	4%	6%	25%	41%	24%
Gender Unknown	0	-	-	-	-	-
American Indian	99	0%	2%	14%	48%	35%
Asian American	1,051	5%	7%	28%	41%	19%
Pacific Islander	65	6%	12%	15%	52%	14%
Filipino	439	3%	5%	28%	49%	15%
Hispanic	7,433	3%	5%	22%	43%	26%
African American	1,183	2%	5%	22%	45%	25%
White	2,985	3%	6%	26%	40%	25%
Two or More Races	356	4%	5%	24%	42%	24%
English only	8,175	3%	6%	24%	42%	25%
Initially fluent English proficient	234	5%	5%	25%	44%	21%
English learner	758	3%	4%	20%	45%	28%
Reclassified fluent English proficient	4,400	4%	5%	23%	43%	25%
TBD	0	-	-	-	-	-
English proficiency unknown	44	5%	18%	27%	36%	14%
Intellectual disability	5,373	2%	5%	25%	43%	25%
Hearing impairment	158	2%	3%	20%	54%	22%
Speech or language impairment	308	0%	1%	10%	57%	32%
Visual impairment	97	20%	11%	22%	29%	19%
Emotional disturbance	83	0%	0%	6%	45%	49%
Orthopedic impairment	816	11%	10%	23%	38%	19%
Other health impairment	594	2%	2%	18%	44%	34%
Specific learning impairment	714	0%	0%	3%	50%	46%
Deaf-blindness	7	-	-	-	-	-
Multiple disabilities	658	13%	12%	29%	31%	15%
Autism	4,205	2%	6%	26%	42%	23%
Traumatic brain injury	84	10%	4%	17%	40%	30%
Unknown	514	4%	8%	22%	43%	24%
Not economically disadvantaged	5,076	4%	7%	27%	40%	22%
Economically disadvantaged	8,535	3%	5%	22%	44%	27%
Primary Ethnicity—Not Economically Disadvantaged						
American Indian	30	0%	7%	20%	57%	17%
Asian American	600	6%	8%	28%	41%	19%
Pacific Islander	31	13%	10%	16%	48%	13%
Filipino	270	4%	6%	30%	46%	14%
Hispanic	1,704	5%	7%	25%	39%	23%
African American	389	3%	8%	29%	42%	17%
White	1,857	3%	7%	27%	40%	23%
Two or More Races	195	6%	5%	27%	40%	23%
Primary Ethnicity—Economically Disadvantaged						
American Indian	69	0%	0%	12%	45%	43%
Asian American	451	4%	7%	28%	41%	20%
Pacific Islander	34	0%	15%	15%	56%	15%
Filipino	169	3%	3%	24%	53%	17%
Hispanic	5,729	3%	4%	22%	44%	27%
African American	794	2%	4%	18%	47%	29%
White	1,128	2%	5%	23%	41%	29%
Two or More Races	161	3%	6%	20%	45%	26%

* Results for groups with 10 or fewer members are not reported.

Appendix 7.C—Types of Score Reports

Table 7.C.1 Score Reports Reflecting CAPA Results

2014–15 CAASPP Student Score Reports	
Description	**Use and Distribution**
The CAASPP Student Report—CAPA for Science A report for the CAPA for Science	
This report provides parents/guardians and teachers with the student's results, presented in tables and graphs. Data presented for the CAPA for Science taken include the following: • Scale scores • Performance levels	This report includes individual student results and is not distributed beyond parents/guardians and the student's school. Two copies of this report are provided for each student. One is for the student's current teacher and one is to be distributed by the LEA to parents/ guardians.
Subgroup Summary	
This set of reports disaggregates and reports results by the following subgroups: • All students • Disability status • Economic status • Gender • English proficiency • Primary ethnicity • Economic status These reports contain no individual student-identifying information and are aggregated at the school, LEA, county, and state levels. For each subgroup within a report and for the total number of students, the following data are included for each test: • Total number tested in the subgroup • Percent of enrollment tested in the subgroup • Number and percent of valid scores • Number tested who received scores • Mean scale score • Standard deviation of scale score • Number and percent of students scoring at each performance level	This report is a resource for evaluators, researchers, teachers, parents/guardians, community members, and administrators. Each LEA can download this report for the whole LEA and the schools within in from TOMS. ***Note:*** The data on this report may be shared with parents/guardians, community members, and the media only if the data are for 11 or more students.

Chapter 8: Analyses

This chapter summarizes the task (item)- and test-level statistics obtained for the California Alternate Performance Assessment (CAPA) for Science administered during the 2014–15 test administration.

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

1. Classical Item Analyses
2. Reliability Analyses
3. Analyses in Support of Validity Evidence
4. Item Response Theory (IRT) Analyses

Prior to 2013–14, 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 2014–15, because forms were reused, 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 74 presents the classical item analyses, including average item score (AIS) and polyserial correlation coefficient, and associated flags, for the operational tasks of each test. Also presented in this appendix is information about the distribution of scores for the operational tasks. In addition, the mean, minimum, and maximum of AIS and polyserial correlation for each operational task are presented in Table 8.A.1 through Table 8.A.4, which start on page 74.
2. Appendix 8.B on page 77 presents results of the reliability analyses of total test scores 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 81 presents tables showing the results of the rater agreement for each operational task.
4. Appendix 8.D on page 74 presents the scoring tables obtained as a result of the IRT equating process after the 2012–13 administration.
5. Appendix 8.E on page 87 shows the distribution of primary disabilities for students who took each CAPA for Science level.

Samples Used for the Analyses

CAPA for Science analyses were conducted using a data file which comprised of the entire test-taking population. CAPA for Science analyses were conducted at different times after test administration and involved varying proportions of the full CAPA for Science data.

During the 2014–15 administration, neither IRT calibrations nor scaling analyses are implemented because intact forms from the 2012–13 administration were used. The summary statistics describing the samples for 2012–13 and 2014–15 are presented in Table 8.1.

For the intact forms without any replacement or edited items, the IRT results for calibration and scaling based on the equating sample of the previous administration can be found in Appendix D of the *2013 CAPA Technical Report,* which is the report for the year each CAPA for Science form was administered originally.

Table 8.1 CAPA Raw Score Means and Standard Deviations: Tested cases with valid scores for 2012–13 and 2014–15

Level	2013 N	2013 Mean	2013 SD	2015 N	2015 Mean	2015 SD
Level I Science	3,724	24.39	11.36	3,706	24.45	11.67
Level III Science	3,446	21.02	5.84	3,323	20.71	6.13
Level IV Science	3,275	21.51	5.99	3,315	21.73	6.16
Level V Science	3,435	20.20	5.94	3,267	20.37	5.86

Classical Analyses

Average Item Score

The Average Item Score (AIS) indicates the average score that students obtained on a task. Desired values generally fall within the range of 30 percent to 80 percent of the maximum obtainable task score. Occasionally, a task that falls outside this range is included in a test form because of the quality and educational importance of the task content or because it is the best available measure for students with very high or low achievement.

CAPA for Science task scores range from 0 to 5 for Level I and 0 to 4 for Levels III, IV, and V. For tasks scored using a 0–4 point rubric, 30 percent is represented by the value 1.20 and 80 percent is represented by the value 3.20. For tasks scored using a 0–5 point rubric, 30 percent is represented by the value 1.50 and 80 percent is represented by the value 4.00.

Polyserial Correlation of the Task Score with the Total Test Score

This statistic describes the relationship between students' scores on a specific task and their total test scores. The polyserial correlation is used when an interval variable is correlated with an ordinal variable that is assumed to reflect an underlying continuous latent variable.

Polyserial correlations are based on a polyserial regression model (Drasgow, 1988). The ETS proprietary software Generalized Analysis System (GENASYS) estimates the value of β for each item using maximum likelihood. In turn, it uses this estimate of β to compute the polyserial correlation from the following formula:

$$r_{polyreg} = \frac{\hat{\beta} s_{tot}}{\sqrt{\hat{\beta}^2 s^2_{tot} + 1}} \tag{8.1}$$

where,

s_{tot} is the standard deviation of the students' total scores; and

β is the item parameter to be estimated from the data, with the estimate denoted as $\hat{\beta}$, using maximum likelihood.

β is a regression coefficient (slope) for predicting the continuous version of a binary item score onto the continuous version of the total score. There are as many regressions as there are boundaries between scores with all sharing a common slope, β. For a polytomously scored item, there are k-1 regressions, where k is the number of score points on the item. Beta (β) is the slope for all k-1 regressions.

The polyserial correlation is sometimes referred to as a discrimination index because it is an indicator of the degree to which students who do well on the total test also do well on a given task. A task is considered discriminating if high-ability students tend to receive higher scores and low-ability students tend to receive lower scores on the task.

Tasks with negative or extremely low correlations can indicate serious problems with the task itself or can indicate that students have not been taught the content. Based on the range of polyserials produced in field-test analyses, an indicator of poor discrimination was set to less than 0.60.

A descriptive summary of the classical item statistics for the overall test are presented in Table 8.2. The task-by-task values are presented in Table 8.A.1 through Table 8.A.4. Some tasks were flagged for unusual statistics; these flags are shown in the tables. Although the flag definition appears in the heading of each table, the flags are displayed in the body of the tables only where applicable for the specific CAPA for Science presented. The flag classifications are as follows:

- Difficulty flags
  - A: Low average task score (below 1.5 at Level I; below 1.2 at Levels III–V)
  - H: High average task score (above 4.0 at Level I; above 3.2 at Levels III–V)
- Discrimination flag
  - R: Polyserial correlation less than 0.60
- Omit/nonresponse/flag
  - O: Omit/nonresponse rates greater than 5 percent

Table 8.2 Average Item Score and Polyserial Correlation

Level	No. of Tasks	No. of Examinees	Mean AIS	Mean Polyserial	Min. AIS	Min. Polyserial	Max. AIS	Max. Polyserial
Level I Science	8	3,706	3.07	0.80	2.60	0.76	3.32	0.83
Level III Science	8	3,323	2.60	0.76	2.25	0.69	3.12	0.81
Level IV Science	8	3,315	2.73	0.74	2.29	0.67	2.99	0.79
Level V Science	8	3,267	2.56	0.75	1.94	0.63	3.36	0.82

As noted previously, the score distributions for individual operational tasks comprising each CAPA for Science test are provided by test and level in Table 8.A.5.

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 can range from 0 to 1. The higher the reliability coefficient for a set of scores, the more likely individuals would be to obtain very similar scores if they were retested.

The formula for the internal-consistency reliability as measured by Cronbach's Alpha (Cronbach, 1951) is defined by equation 8.2:

$$\alpha = \frac{n}{n-1}\left[1 - \frac{\sum_{i=1}^{n} s_i^{\,2}}{s_t^2}\right] \tag{8.2}$$

where,

n is the number of tasks,

s_i^2 is the variance of scores on the task 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 reported in equation 8.3. The mathematic form of the SEM is as follows:

$$s_e = s_t\sqrt{1-\alpha} \tag{8.3}$$

where,

α is the reliability estimated using equation 8.2, 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.A.3 gives the reliability and SEM for the CAPA for Science, along with the number of tasks and examinees upon which those analyses were performed.

Table 8.3 Reliabilities and SEMs for the CAPA for Science

Level	No. of Items	No. of Examinees	Reliab.	Mean Scale Score	Scale Score S.D.	Scale Score SEM	Mean Raw Score	Raw Score S.D.	Raw Score SEM
Level I Science	8	3,706	0.89	37.55	10.84	3.90	24.45	11.67	3.62
Level III Science	8	3,323	0.86	35.99	5.43	2.28	20.71	6.13	2.02
Level IV Science	8	3,315	0.85	36.01	5.53	2.36	21.73	6.16	2.12
Level V Science	8	3,267	0.84	35.89	4.83	2.35	20.37	5.86	1.94

Subgroup Reliabilities and SEMs

The reliabilities of the CAPA 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, disability group, and English-language fluency. The reliability analyses are also presented by primary ethnicity within economic status.

Table 8.B.1 through Table 8.B.6 present the reliabilities and SEM information for the total test scores for each subgroup. Note that the reliabilities are reported only for samples that

are comprised of 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 CAPA for Science 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:

$$\mathrm{CSEM}(\hat{\theta}) = \frac{1}{\sqrt{I\left(\hat{\theta}\right)}} a \tag{8.4}$$

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 level and content area.

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

Table 8.D.1 through Table 8.D.4 in Appendix 8.D present the scale score CSEMs at the score required for a student to be classified in the below basic, basic, proficient, and advanced performance levels for the CAPA for Science. 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.

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 and Figure 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.7 through Table 8.B.10 in Appendix 8.B, starting on page 79.

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

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

Validity Evidence

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

Validation is the process of accumulating evidence to support each proposed score interpretation or use. It involves more than a single study or gathering of one particular kind of evidence. Validation involves multiple investigations and various kinds of evidence (AERA, APA, & NCME, 2014; Cronbach, 1971; ETS, 2002; Kane, 2006). The process begins with test design and continues through the entire assessment process, including task 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 CAPA for Science. The description is organized in the manner prescribed by *The Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014). These standards require a clear definition of the purpose of the test, which includes a description of the qualities—called constructs—that are to be assessed by a test, the population to be assessed, as well as how the scores are to be interpreted and used.

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

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

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

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

The Constructs to Be Measured

The CAPA for Science are designed to show how well students with an individualized education program (IEP) and who have significant cognitive disabilities perform relative to the California content standards. These content standards were approved by the SBE in 1998; they describe what students should know and be able to do at each 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 test blueprints for the CAPA for Science can be found on the California Department of Education (CDE) Standardized Testing and Reporting (STAR) CAPA Blueprints Web page at http://www.cde.ca.gov/ta/tg/sr/capablueprints.asp. ETS developed all CAPA for Science tasks to conform to the SBE-approved content standards and test blueprints.

Interpretations and Uses of the Scores Generated

Total test scores expressed as scale scores and student performance levels are generated for each student for each grade-level 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.

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

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

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

Intended Test Population(s)

Students with an IEP and who have significant cognitive disabilities in grades two through eleven take the CAPA for Science when they are unable to take the Smarter Balanced for English Language Arts/Literacy and Mathematics and the California Standards Test or California Modified Assessment for Science with or without universal tools, designated supports, and accommodations. Participation in the CAPA for Science and eligibility are determined by a student's IEP team. Only those students whose parents/guardians have submitted written requests to exempt them from California Assessment of Student Performance and Progress System testing do not take the tests. See the subsection "Intended Population" on page 2 for a more detailed description of the intended test population.

Validity Evidence Collected

Evidence Based on Content

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

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

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

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

Specifications and blueprints—ETS maintains task specifications for the CAPA for Science. The task specifications describe the characteristics of the tasks that should be written to measure each content standard. A thorough description of the specifications can be found in Chapter 3, starting on page 17. Once the tasks were developed and field-tested, ETS selected all CAPA for Science test tasks to conform to the SBE-approved California content standards and test blueprints. Test blueprints for the CAPA for Science were proposed by ETS and reviewed and approved by the Assessment Review Panels (ARPs), which are advisory panels to the CDE and ETS on areas related to task development for the CAPA for Science. 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 CAPA for Science; the blueprints were most recently revised and adopted by the SBE in 2006 for implementation beginning in 2008. The test blueprints for the CAPA for Science can be found on the CDE STAR CAPA Blueprints Web page at http://www.cde.ca.gov/ta/tg/sr/capablueprints.asp.

Task development process—A detailed description of the task development process for the CAPA for Science is presented in Chapter 3, starting on page 17.

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

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 26.

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 22.

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 CAPA in April 2007 (HumRRo, 2007). HumRRO utilized the Webb alignment method to evaluate the alignment of the performance tasks field-tested in the 2007 CAPA 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. The results indicated that the performance tasks assess the majority of CAPA standards well across levels.

Evidence Based on Relations to Other Variables

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

Differential Item Functioning Analyses

Analyses of DIF provided evidence of the degree to which a score interpretation or use was valid for individuals who differ in particular demographic characteristics. For the CAPA for Science, DIF analyses were performed after the test forms' original administration in 2012-13 on all operational tasks and field-test tasks for which sufficient student samples were available.

The results of the DIF analyses are presented in Appendix 8.E of the *2013 CAPA Technical Report,* which is the report for the year each 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 vast majority of the tasks exhibited little or no significant DIF, suggesting that, in general, scores based on the CAPA for Science tasks would have the same meaning for individuals who differed in their demographic characteristics.

Evidence Based on Response Processes

As noted in the APA, AERA, and NCME *Standards* (2014), 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. Finally, evidence may also be derived from feedback provided by observers or judges involved in the scoring of examinee responses.

Evidence of Interrater Agreement

Rater consistency is critical to the scores of CAPA for Science tasks and their interpretations. These findings provide evidence of the degree to which raters agree in their observations about the qualities evident in students' responses. In order to monitor and evaluate the accuracy of rating, approximately 10 percent of students' test responses were scored twice. They were scored once by the primary examiner (rater 1) and a second time by an independent, trained observer (rater 2). Evidence that the raters' scores are consistent helps to support the inference that the scores have the intended meaning. The data collected were used to evaluate interrater agreement.

Interrater Agreement

As noted previously, approximately 10 percent of the test population's responses to the tasks were scored by two raters. Across all CAPA for Science levels, the percentage of students for whom the raters were in exact agreement ranged from 92 percent to 99 percent. The results are presented in Table 8.C.1 through Table 8.C.4.

Evidence Based on Internal Structure

As suggested by the *Standards* (AERA, APA, & NCME, 2014), evidence of validity can also be obtained from studies of the properties of the item (task) 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 CAPA for Science, it is assumed that a single construct underlies the total scores obtained on each test. Evidence to support this assumption can be gathered from the results of task analyses, evaluations of internal consistency, and studies of reliability.

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.

Overall reliability—The reliability analyses are presented in Table 8.3. The results indicate that the reliabilities for all CAPA for Science levels tended to be high, ranging from 0.84 to 0.89.

Subgroup reliabilities—The reliabilities of the operational CAPA for Science are also examined for various subgroups of the examinee population that differed in their demographic characteristics. The characteristics considered were gender, ethnicity, economic status, disability group, English-language fluency, and ethnicity-by-economic status. The results of these analyses can be found in Table 8.B.1 through Table 8.B.6.

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, 2014, 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 the administration of the CAPA for Science.

IRT Analyses

Post-Equating

Prior to the 2013–14 administration, the CAPA for Science were equated to a reference form using a common-item nonequivalent groups design and post-equating methods based on IRT. The “base” or “reference” calibrations for the CAPA for Science were established by calibrating samples of data from a specific administration. Doing so established a scale to which subsequent item (task) calibrations could be linked.

The procedures used for post-equating the CAPA for Science prior to 2013–14 involved three steps: task calibration, task parameter scaling, and production of raw-score-to-scale-score conversions using the scaled task parameters. ETS used GENASYS for the IRT item calibration and equating work. The IRT model used to calibrate the CAPA for Science test tasks was the one-parameter partial credit (1PPC) model, a more restrictive version of the generalized partial-credit model (Muraki, 1992), in which all tasks were assumed to be equally discriminating. This model stated that the probability that an examinee with ability θ will perform in the *kth* category of m_j ordered score categories of task *j* can be expressed as:

$$P_{jk}(\theta) = \frac{\exp\left[\sum_{v=1}^{k} 1.7 a_j(\theta - b_j + d_{jv})\right]}{\sum_{c=1}^{m_j} \exp\left[\sum_{v=1}^{c} 1.7 a_j(\theta - b_j + d_{jv})\right]} \tag{8.5}$$

where,

m_j is the number of possible score categories (c=1…m_j) for task *j,*

a_j is the slope parameter (equal to 0.588) for task *j*,

b_j is the difficulty of task *j*, and

d_{jv} is the threshold parameter for category v of task *j*.

For the task calibrations, the PARSCALE program (Muraki & Bock, 1995) was constrained by setting a common discrimination value for all tasks equal to 1.0 / 1.7 (or 0.588) and by setting the lower asymptote for all tasks to zero. The resulting estimation is equivalent to the Rasch partial credit model for polytomously scored tasks.

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.

Pre-Equating

During the 2014–15 administration, because intact test forms from the 2012–13 administration were reused, the conversion tables from the previous administration when the forms were originally used are directly applied to the 2014–15 operational scoring.

Descriptions of IRT analyses such as the model-data fit analyses can be found in Chapter 8 of the original-year (2013) technical report; the results of the IRT analyses are presented in Appendix 8.D of the original-year (2013) technical report. The *2013 CAPA Technical Report,* which is the report for the year each form was administered originally, is linked on the CDE’s

Technical Reports and Studies Web page at http://www.cde.ca.gov/ta/tg/sr/technicalrpts.asp.

Summaries of Scaled IRT *b*-values

For the post-equating procedure prior to the 2013–14 administration, once the IRT *b*-values were placed on the item bank scale, analyses were performed to assess the overall test difficulty and the distribution of tasks in a particular range of item difficulty.

During the 2014–15 administration, neither IRT calibrations nor scaling are implemented. The summaries of *b*-values can be found in Appendix D of the *2013 CAPA Technical Report,* which is the report for the year each form was administered originally.

Equating Results

During the 2014–15 administration, for the reused intact forms, the conversion tables from their original administrations (2013) are directly applied to the current administration.

Complete raw-score-to-scale-score conversion tables for the CAPA for Science administered in 2014–15 based on P2 data—the entire test-taking population but not corrections of demographic data through the California Longitudinal Pupil Assessment Data System or students with invalid scores were excluded from the tabled results—are presented in Table 8.D.1 through Table 8.D.4, starting on page 83. The raw scores and corresponding transformed scale scores are listed in those tables. For all of the 2014–15 CAPA for Science, scale scores were truncated at both ends of the scale so that the minimum reported scale score was 15 and the maximum reported scale score was 60. The scale scores defining the cut scores for all performance levels are presented in Table 2.2, which is on page 15 in Chapter 2.

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 2013–14 administration, DIF analyses were performed based on the final item analyses (FIA) sample and were performed on all operational tasks and all field-test tasks for which sufficient student samples were available. DIF analyses are not implemented during the 2014–15 administration because forms are reused and all tasks were evaluated for DIF during the previous administration when the intact forms were originally used. These DIF results can be found in Appendix E of the *2013 CAPA Technical Report,* which is the report for the year the form was administered originally.

The statistical procedure of DIF analysis that was conducted prior to the 2013–14 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.

DIF analyses of the polytomously scored CAPA for Science tasks were completed using two procedures. The first was the Mantel-Haenszel (MH) ordinal procedure, which is based on the Mantel procedure (Mantel, 1963; Mantel & Haenszel, 1959). The MH ordinal procedure compares the proportion of examinees in the reference and focal groups obtaining each task score after matching the examinees on their total test score. As with dichotomously scored tasks, the common odds ratio is estimated across the matched score groups. The

resulting estimate was interpreted as the relative likelihood of obtaining a given task score for members of two groups that are matched on ability.

As such, the common odds ratio provides an estimated effect size; a value of one indicates equal odds and thus no DIF (Dorans & Holland, 1993). The corresponding statistical test is H_0: $\alpha = 1$, where α is a common odds ratio assumed equal for all matched score categories $s = 1$ to S. Values of less than one indicate DIF in favor of the focal group; a value of one indicates the null condition; and a value greater than one indicates DIF in favor of the reference group. The associated ($MH\chi^2$) is distributed as a Chi-square random variable with one degree of freedom.

The $MH\chi^2$ Mantel Chi-square statistic was used in conjunction with a second procedure, the standardization procedure (Dorans & Schmitt, 1993). This procedure produces a DIF statistic based on the standardized mean difference (SMD) in average task scores between members of two groups that have been matched on their overall test score. The SMD compares the task means of the two studied groups after adjusting for differences in the distribution of members across the values of the matching variable (total test score).

The standardized mean difference is computed as the following:

$$SMD = \sum_m w_m \left(E_{fm} - E_{rm}\right) / \sum_m w_m \tag{8.5}$$

where,

$w_m / \sum w_m$ is the weighting factor at score level m supplied by the standardization group to weight differences in item performance between a focal group (E_{fm}) and a reference group (E_{rm}) (Doran & Kulick, 2006).

A negative SMD value means that, conditional on the matching variable, the focal group has a lower mean task score than the reference group. In contrast, a positive SMD value means that, conditional on the matching variable, the reference group has a lower mean task score than the focal group. The SMD is divided by the standard deviation (SD) of the total group task score in its original metric to produce an effect-size measure of differential performance.

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.

The ETS classification system assigns tasks to one of the three DIF categories on the basis of a combination of statistical significance of the Mantel Chi-square statistic and the magnitude of the SMD effect-size:

DIF Category	Definition
A (negligible)	• The Mantel Chi-square statistic is not statistically significant (at the 0.05 level) or |SMD/SD| < 0.17.
B (moderate)	• The Mantel Chi-square statistic is statistically significant (at the 0.05 level) and 0.17 ≤ |SMD/SD| < 0.25.
C (large)	• The Mantel Chi-square statistic is statistically significant (at the 0.05 level) and |SMD/SD| > 0.25.

In addition, the categories identify which group is being advantaged; categories are displayed in Table 8.4. The categories have been used by all ETS testing programs for more than 15 years.

Table 8.4 DIF Flags Based on the ETS DIF Classification Scheme

Flag	Descriptor
A–	Negligible favoring members of the reference group
B–	Moderate favoring members of the reference group
C–	Large favoring members of the reference group
A+	Negligible favoring members of the focal group
B+	Moderate favoring members of the focal group
C+	Large favoring members of the focal group

Category C contains tasks with large values of DIF. As shown in Table 8.4, tasks classified as C+ tend to be easier for members of the focal group than for members of the reference group with comparable total scores. Tasks classified as C– tend to be more difficult for members of the focal group than for members of the reference group whose total scores on the test are like those of the focal group.

Table lists specific subgroups that were used for DIF analyses for the CAPA for Science including primary disability. Table 8.D.1 to Table 8.D.4, starting on page 83 in Appendix 8.D, show the sample size for disability groups within CAPA for Science test level.

Table 8.5 Subgroup Classification for DIF Analyses

DIF Type	Reference Group	Focal Group
Gender	Male	Female
Race/Ethnicity	White	• African American • American Indian • Asian • Combined Asian Group (Asian/Pacific Islander/Filipino) • Filipino • Hispanic/Latin American • Pacific Islander
Disability	Intellectual Disability (ID)	• Autism • Deaf-blindness • Emotional Disturbance • Hearing Impairment • Multiple Disabilities • Orthopedic Impairment • Other Health Impairment • Specific Learning Disability • Speech or Language Impairment • Traumatic Brain Injury • Visual Impairment

References

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

California Department of Education. (2013). *California Alternate Performance Assessment technical report, spring 2013 administration*. Sacramento, CA. Retrieved from http://www.cde.ca.gov/ta/tg/sr/documents/capa13techrpt.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, D. C.: 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.

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

Dorans, N. J., & Kulick, E. (2006). Differential item functioning on the mini-mental state examination: An application of the Mantel-Haenszel and standardization procedures. *Medical Care*, *44,* 107–14.

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

Drasgow F. (1988). Polychoric and polyserial correlations. In L. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences* (Vol. 7, pp. 69–74). New York: Wiley.

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

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

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

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.

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

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

Appendix 8.A—Classical Analyses: Task Statistics

Table 8.A.1 AIS and Polyserial Correlation: Level I, Science—Current Year (2015) and Original Year of Administration (2013)

Flag values are as follows:
A = low average task score
R = low correlation with criterion
O = high percent of omits/not responding
H = high average task score

Form	2013 Task	2013 AIS	2013 Polyserial	2013 Flag	2015 Task	2015 AIS	2015 Polyserial	2015 Flag
Operational	1	3.16	0.82		1	3.09	0.83	
Operational	2	3.11	0.79		2	3.12	0.79	
Operational	3	3.05	0.76		3	3.02	0.77	
Operational	4	3.01	0.80		4	3.03	0.81	
Operational	5	3.26	0.82		5	3.32	0.83	
Operational	6	2.58	0.77		6	2.60	0.78	
Operational	7	3.07	0.73		7	3.12	0.76	
Operational	8	3.10	0.80		8	3.24	0.82	

Table 8.A.2 AIS and Polyserial Correlation: Level III, Science—Current Year (2015) and Original Year of Administration (2013)

Flag values are as follows:
A = low average task score
R = low correlation with criterion
O = high percent of omits/not responding
H = high average task score

Form	2013 Task	2013 AIS	2013 Polyserial	2013 Flag	2015 Task	2015 AIS	2015 Polyserial	2015 Flag
Operational	1	2.52	0.81		1	2.51	0.80	
Operational	2	2.55	0.64		2	2.54	0.69	
Operational	3	2.22	0.70		3	2.25	0.72	
Operational	4	2.50	0.78		4	2.41	0.81	
Operational	5	2.78	0.74		5	2.73	0.75	
Operational	6	3.00	0.75		6	2.87	0.75	
Operational	7	2.36	0.74		7	2.36	0.74	
Operational	8	3.12	0.76		8	3.12	0.79	

Table 8.A.3 AIS and Polyserial Correlation: Level IV, Science—Current Year (2015) and Original Year of Administration (2013)

Flag values are as follows:
A = low average task score
R = low correlation with criterion
O = high percent of omits/not responding
H = high average task score

Form	2013 Task	2013 AIS	2013 Polyserial	2013 Flag	2015 Task	2015 AIS	2015 Polyserial	2015 Flag
Operational	1	2.52	0.71		1	2.50	0.72	
Operational	2	2.93	0.76		2	2.91	0.77	
Operational	3	2.51	0.67		3	2.58	0.67	
Operational	4	2.88	0.70		4	2.94	0.69	
Operational	5	2.17	0.78		5	2.29	0.79	
Operational	6	2.92	0.74		6	2.93	0.77	
Operational	7	2.97	0.80		7	2.99	0.78	
Operational	8	2.60	0.73		8	2.67	0.74	

Table 8.A.4 AIS and Polyserial Correlation: Level V, Science—Current Year (2015) and Original Year of Administration (2013)

Flag values are as follows:
A = low average task score
R = low correlation with criterion
O = high percent of omits/not responding
H = high average task score

Form	2013 Task	2013 AIS	2013 Polyserial	2013 Flag	2015 Task	2015 AIS	2015 Polyserial	2015 Flag
Operational	1	1.96	0.67		1	1.94	0.63	
Operational	2	3.24	0.79	H	2	3.24	0.79	H
Operational	3	2.16	0.78		3	2.17	0.78	
Operational	4	2.16	0.73		4	2.18	0.72	
Operational	5	2.48	0.74		5	2.52	0.72	
Operational	6	2.87	0.80		6	2.92	0.78	
Operational	7	3.38	0.79	H	7	3.36	0.82	H
Operational	8	1.99	0.73		8	2.12	0.73	

Table 8.A.5 Frequency of Operational Task Scores: Science

Science Level	Score on Task	1		2		3		4		5		6		7		8	
		Count	Percent	Count	Percent	Count	Percent	Count	Percent	Count	Percent	Count	Percent	Count	Percent	Count	Percent
I	0	385	10.32	365	9.79	371	9.95	353	9.47	346	9.28	455	12.20	326	8.74	389	10.43
	1	997	26.74	989	26.52	1,104	29.61	1,118	29.98	924	24.78	1,343	36.02	1,068	28.64	861	23.09
	2	184	4.93	186	4.99	153	4.10	168	4.51	132	3.54	199	5.34	148	3.97	218	5.85
	3	163	4.37	195	5.23	189	5.07	171	4.59	124	3.33	227	6.09	205	5.50	209	5.60
	4	400	10.73	390	10.46	376	10.08	357	9.57	310	8.31	342	9.17	363	9.73	312	8.37
	5	1,600	42.91	1,604	43.01	1,536	41.19	1,562	41.89	1,893	50.76	1,163	31.19	1,619	43.42	1,740	46.66
III	0	117	3.50	103	3.08	144	4.31	164	4.91	116	3.47	104	3.11	135	4.04	119	3.56
	1	596	17.85	397	11.89	721	21.59	474	14.20	385	11.53	372	11.14	631	18.90	222	6.65
	2	892	26.71	1,348	40.37	1,114	33.36	1,213	36.33	862	25.82	700	20.96	973	29.14	357	10.69
	3	987	29.56	644	19.29	938	28.09	886	26.53	954	28.57	928	27.79	1,170	35.04	1,217	36.45
	4	747	22.37	847	25.37	422	12.64	602	18.03	1,022	30.61	1,235	36.99	430	12.88	1,424	42.65
IV	0	61	1.83	86	2.59	104	3.13	100	3.01	188	5.65	92	2.77	87	2.62	106	3.19
	1	420	12.63	416	12.51	424	12.75	359	10.80	974	29.29	173	5.20	301	9.05	459	13.80
	2	1,255	37.74	747	22.47	945	28.42	588	17.68	769	23.13	624	18.77	690	20.75	871	26.20
	3	1,002	30.14	577	17.35	1,194	35.91	952	28.63	500	15.04	1,493	44.90	802	24.12	978	29.41
	4	587	17.65	1,499	45.08	658	19.79	1,326	39.88	894	26.89	943	28.36	1,445	43.46	911	27.40
V	0	87	2.65	76	2.31	146	4.45	120	3.66	85	2.59	92	2.80	89	2.71	160	4.87
	1	1,149	35.00	195	5.94	990	30.16	959	29.21	572	17.42	268	8.16	196	5.97	1,009	30.73
	2	1,187	36.16	363	11.06	946	28.82	1,242	37.83	999	30.43	624	19.01	329	10.02	1,023	31.16
	3	632	19.25	943	28.72	611	18.61	211	6.43	858	26.13	1,199	36.52	603	18.37	585	17.82
	4	228	6.94	1,706	51.96	590	17.97	751	22.88	769	23.42	1,100	33.51	2,066	62.93	506	15.41

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. Finally, results based on samples that contain 50 or fewer examinees should be interpreted with caution due to small sample sizes.

Table 8.B.1 Reliabilities and SEMs by Gender—Male

Level	N	Reliab.	SEM
I Science	2,353	0.88	3.94
III Science	2,274	0.86	2.28
IV Science	2,182	0.86	2.36
V Science	2,192	0.84	2.36

Table 8.B.2 Reliabilities and SEMs by Gender—Female

Level	N	Reliab.	SEM
I Science	1,353	0.90	3.83
III Science	1,049	0.85	2.28
IV Science	1,133	0.84	2.37
V Science	1,075	0.84	2.33

Table 8.B.3 Reliabilities and SEMs by Primary Ethnicity

	American Indian			Asian			Pacific Islander			Filipino		
Level	**N**	**Reliab.**	**SEM**	**N**	**Reliab.**	**SEM**	**N**	**Reliab.**	**SEM**	**N**	**Reliab.**	**SEM**
I Science	19	0.62	4.48	316	0.87	4.10	19	0.89	4.16	111	0.84	4.19
III Science	26	0.86	2.29	237	0.85	2.38	14	0.86	2.01	116	0.85	2.31
IV Science	30	0.77	2.43	255	0.87	2.43	20	0.84	2.53	103	0.79	2.45
V Science	24	0.79	2.05	243	0.85	2.34	12	0.88	2.44	109	0.85	2.38
	Hispanic			**African American**			**White**			**Unknown Ethnicity**		
Level	**N**	**Reliab.**	**SEM**	**N**	**Reliab.**	**SEM**	**N**	**Reliab.**	**SEM**	**N**	**Reliab.**	**SEM**
I Science	2,062	0.90	3.76	273	0.89	3.84	799	0.87	4.09	107	0.86	4.21
III Science	1,923	0.86	2.26	259	0.86	2.30	675	0.87	2.26	73	0.86	2.60
IV Science	1,793	0.85	2.33	325	0.86	2.30	702	0.85	2.40	87	0.86	2.44
V Science	1,655	0.83	2.34	326	0.81	2.35	809	0.85	2.38	89	0.86	2.30

Table 8.B.4 Reliabilities and SEMs by Primary Ethnicity for Economically Disadvantaged

	American Indian			Asian			Pacific Islander			Filipino		
Level	**N**	**Reliab.**	**SEM**	**N**	**Reliab.**	**SEM**	**N**	**Reliab.**	**SEM**	**N**	**Reliab.**	**SEM**
I Science	11	0.73	4.33	129	0.89	3.95	7	–	–	47	0.80	4.06
III Science	19	0.83	2.33	113	0.82	2.48	10	–	–	46	0.81	2.43
IV Science	21	0.59	2.28	109	0.82	2.45	11	0.79	2.74	31	0.61	2.42
V Science	18	0.80	1.94	100	0.82	2.42	6	-	-	45	0.88	2.40
	Hispanic			**African American**			**White**			**Unknown Ethnicity**		
Level	**N**	**Reliab.**	**SEM**	**N**	**Reliab.**	**SEM**	**N**	**Reliab.**	**SEM**	**N**	**Reliab.**	**SEM**
I Science	1,491	0.90	3.73	168	0.88	3.78	274	0.86	4.08	43	0.86	4.25
III Science	1,508	0.85	2.24	178	0.86	2.29	259	0.86	2.25	32	0.81	2.98
IV Science	1,445	0.84	2.33	217	0.83	2.26	285	0.84	2.36	44	0.85	2.41
V Science	1,285	0.82	2.34	231	0.81	2.31	310	0.84	2.43	42	0.89	2.20

Table 8.B.5 Reliabilities and SEMs by Primary Ethnicity for Not Economically Disadvantaged

	American Indian			Asian			Pacific Islander			Filipino		
Level	**N**	**Reliab.**	**SEM**	**N**	**Reliab.**	**SEM**	**N**	**Reliab.**	**SEM**	**N**	**Reliab.**	**SEM**
I Science	8	–	–	187	0.85	4.18	12	0.90	4.00	64	0.85	4.27
III Science	7	–	–	124	0.88	2.30	4	–	–	70	0.87	2.22
IV Science	9	–	–	146	0.90	2.40	9	–	–	72	0.82	2.47
V Science	6	–	–	143	0.87	2.30	6	–	–	64	0.81	2.37
	Hispanic			**African American**			**White**			**Unknown Ethnicity**		
Level	**N**	**Reliab.**	**SEM**	**N**	**Reliab.**	**SEM**	**N**	**Reliab.**	**SEM**	**N**	**Reliab.**	**SEM**
I Science	571	0.90	3.84	105	0.88	3.93	525	0.87	4.11	64	0.87	4.20
III Science	415	0.87	2.30	81	0.85	2.33	416	0.88	2.27	70	0.89	2.36
IV Science	348	0.89	2.32	108	0.88	2.34	417	0.85	2.43	72	0.84	2.51
V Science	370	0.85	2.35	95	0.81	2.44	499	0.85	2.36	64	0.84	2.37

Table 8.B.6 Reliabilities and SEMs by Disability

	MR/ID			Hard of Hearing			Deafness			Speech Impairment		
Level	**N**	**Reliab.**	**SEM**	**N**	**Reliab.**	**SEM**	**N**	**Reliab.**	**SEM**	**N**	**Reliab.**	**SEM**
I Science	1,437	0.88	3.81	24	0.92	3.67	0	–	–	19	0.95	2.06
III Science	1,153	0.83	2.26	50	0.74	2.09	0	–	–	153	0.72	2.21
IV Science	1,382	0.84	2.36	39	0.66	2.59	0	–	–	86	0.72	2.29
V Science	1,401	0.80	2.38	45	0.85	2.24	0	–	–	50	0.75	2.20
	Visual Impairment			**Emotional Disturbance**			**Orthopedic Impairment**			**Other Health Impairment**		
Level	**N**	**Reliab.**	**SEM**	**N**	**Reliab.**	**SEM**	**N**	**Reliab.**	**SEM**	**N**	**Reliab.**	**SEM**
I Science	44	0.93	3.69	5	–	–	440	0.89	3.88	97	0.94	3.35
III Science	20	0.97	1.81	21	0.78	1.95	105	0.90	2.22	196	0.78	2.17
IV Science	13	0.89	2.68	21	0.57	2.08	134	0.86	2.42	144	0.79	2.29
V Science	20	0.91	2.39	36	0.67	2.27	137	0.86	2.46	157	0.83	2.26
	Specific Learning Disability			**Deaf-Blindness**			**Multiple Disabilities**			**Autism**		
Level	**N**	**Reliab.**	**SEM**	**N**	**Reliab.**	**SEM**	**N**	**Reliab.**	**SEM**	**N**	**Reliab.**	**SEM**
I Science	23	0.91	2.59	3	–	–	395	0.88	3.91	1,039	0.81	4.17
III Science	241	0.62	2.15	2	–	–	80	0.89	2.48	1,176	0.87	2.32
IV Science	209	0.59	2.02	0	–	–	87	0.92	2.27	1,046	0.86	2.40
V Science	241	0.72	2.26	2	–	–	96	0.91	2.21	944	0.86	2.35
	Traumatic Brain Injury			**Unknown Disability**								
Level	**N**	**Reliab.**	**SEM**	**N**	**Reliab.**	**SEM**						
I Science	20	0.95	3.03	160	0.89	3.87						
III Science	14	0.72	2.09	112	0.85	2.36						
IV Science	28	0.84	2.38	126	0.88	2.27						
V Science	22	0.80	2.25	116	0.85	2.37						

Table 8.B.7 Decision Accuracy and Decision Consistency: Level I, Science

	Placement Score	Far Below Basic	Below Basic	Basic	Proficient	Advanced	Category Total †
Decision Accuracy	0–5	0.05	0.02	0.01	0.00	0.00	0.08
	6–10	0.01	0.03	0.03	0.00	0.00	0.07
	11–19	0.00	0.02	0.11	0.05	0.00	0.18
All-forms Average *	20–29	0.00	0.00	0.04	0.17	0.05	0.27
	30–40	0.00	0.00	0.00	0.05	0.35	0.40
	Estimated Proportion Correctly Classified: Total = 0.72, Proficient & Above = 0.91						
Decision Consistency	0–5	0.05	0.02	0.01	0.00	0.00	0.08
	6–10	0.02	0.03	0.03	0.00	0.00	0.07
	11–19	0.01	0.03	0.09	0.05	0.00	0.18
Alternate Form *	20–29	0.00	0.01	0.06	0.14	0.07	0.27
	30–40	0.00	0.00	0.00	0.06	0.34	0.40
	Estimated Proportion Consistently Classified: Total = 0.63, Proficient & Above = 0.87						

* Values in table are proportions of the total sample.

† Inconsistencies with category cell entries are due to rounding.

Table 8.B.8 Decision Accuracy and Decision Consistency: Level III, Science

	Placement Score	Far Below Basic	Below Basic	Basic	Proficient	Advanced	Category Total †
Decision Accuracy	0–3	0.00	0.01	0.01	0.00	0.00	0.02
	4–10	0.00	0.03	0.01	0.00	0.00	0.04
	11–18	0.00	0.00	0.20	0.04	0.00	0.25
All-forms Average *	19–26	0.00	0.01	0.06	0.41	0.05	0.53
	27–32	0.00	0.00	0.00	0.04	0.12	0.16
	Estimated Proportion Correctly Classified: Total = 0.76, Proficient & Above = 0.88						
Decision Consistency	0–3	0.00	0.01	0.01	0.00	0.00	0.02
	4–10	0.00	0.03	0.02	0.00	0.00	0.04
	11–18	0.00	0.02	0.17	0.06	0.00	0.25
Alternate Form *	19–26	0.00	0.01	0.09	0.35	0.08	0.53
	27–32	0.00	0.00	0.00	0.05	0.12	0.16
	Estimated Proportion Consistently Classified: Total = 0.67, Proficient & Above = 0.84						

* Values in table are proportions of the total sample.

† Inconsistencies with category cell entries are due to rounding.

Table 8.B.9 Decision Accuracy and Decision Consistency: Level IV, Science

	Placement Score	Far Below Basic	Below Basic	Basic	Proficient	Advanced	Category Total †
Decision Accuracy	0–3	0.00	0.01	0.01	0.00	0.00	0.01
	4–12	0.00	0.04	0.01	0.00	0.00	0.06
	13–19	0.00	0.02	0.18	0.06	0.00	0.25
All-forms Average *	20–27	0.00	0.00	0.06	0.39	0.04	0.49
	28–32	0.00	0.00	0.00	0.05	0.13	0.18
	Estimated Proportion Correctly Classified: Total = 0.74, Proficient & Above = 0.88						
Decision Consistency	0–3	0.00	0.01	0.01	0.00	0.00	0.01
	4–12	0.00	0.04	0.02	0.00	0.00	0.06
	13–19	0.00	0.03	0.15	0.07	0.00	0.25
Alternate Form *	20–27	0.00	0.01	0.08	0.33	0.07	0.49
	28–32	0.00	0.00	0.00	0.06	0.13	0.18
	Estimated Proportion Consistently Classified: Total = 0.65, Proficient & Above = 0.84						

* Values in table are proportions of the total sample.

† Inconsistencies with category cell entries are due to rounding.

Table 8.B.10 Decision Accuracy and Decision Consistency: Level V, Science

	Placement Score	Far Below Basic	Below Basic	Basic	Proficient	Advanced	Category Total †
Decision Accuracy	0–3	0.00	0.01	0.01	0.00	0.00	0.01
	4–10	0.00	0.03	0.02	0.00	0.00	0.04
	11–18	0.00	0.00	0.22	0.05	0.00	0.27
All-forms Average *	19–24	0.00	0.01	0.07	0.29	0.06	0.43
	25–32	0.00	0.00	0.00	0.05	0.19	0.24
	Estimated Proportion Correctly Classified: Total = 0.73, Proficient & Above = 0.87						
Decision Consistency	0–3	0.00	0.01	0.01	0.00	0.00	0.01
	4–10	0.00	0.03	0.02	0.00	0.00	0.04
	11–18	0.00	0.02	0.19	0.07	0.00	0.27
Alternate Form *	19–24	0.00	0.01	0.09	0.24	0.09	0.43
	25–32	0.00	0.00	0.00	0.06	0.18	0.24
	Estimated Proportion Consistently Classified: Total = 0.63, Proficient & Above = 0.83						

* Values in table are proportions of the total sample.

† Inconsistencies with category cell entries are due to rounding.

Appendix 8.C—Validity Analyses

Note that, while the correlations are reported only for samples that comprise 11 or more examinees, results based on samples that contain 50 or fewer examinees should be interpreted with caution due to small sample sizes. Correlations between scores on any two content-area tests where 10 or fewer examinees took the tests are expressed as hyphens. Correlations between scores on two content-area tests that cannot be administered to the same group of students are expressed as "N/A."

Table 8.C.1 Interrater Agreement Analyses for Operational Tasks: Level I, Science

Task	First Rating			Second Rating			% Agreement			MAD *	Corr. †
	N	Mean	SD	N	Mean	SD	Exact	Adjacent	Neither		
1	783	3.05	1.94	783	3.07	1.93	95.65	3.71	0.64	0.05	0.99
3	783	3.22	1.91	783	3.20	1.91	95.91	3.07	1.02	0.07	0.98
4	783	3.14	1.90	783	3.13	1.90	95.27	3.20	1.53	0.08	0.98
6	783	3.12	1.91	783	3.11	1.91	96.68	2.30	1.02	0.06	0.98
7	783	3.45	1.90	783	3.43	1.90	95.91	3.20	0.90	0.06	0.98
9	783	2.83	1.89	783	2.83	1.89	97.06	2.30	0.64	0.04	0.99
10	783	3.21	1.91	783	3.22	1.90	96.68	2.05	1.28	0.07	0.97
12	783	3.37	1.90	783	3.35	1.90	95.52	3.71	0.77	0.07	0.98

* Mean absolute difference between first and second ratings

† Pearson correlation between first and second ratings

Table 8.C.2 Interrater Agreement Analyses for Operational Tasks: Level III, Science

Task	First Rating			Second Rating			% Agreement			MAD *	Corr. †
	N	Mean	SD	N	Mean	SD	Exact	Adjacent	Neither		
1	929	2.47	1.06	929	2.46	1.07	96.21	3.79	0.00	0.04	0.98
3	929	2.50	1.02	929	2.50	1.02	94.91	3.57	1.52	0.07	0.95
4	929	2.23	1.00	929	2.24	0.99	95.02	4.33	0.65	0.06	0.96
6	929	2.43	0.99	929	2.45	0.99	94.26	4.87	0.87	0.07	0.96
7	929	2.81	1.03	929	2.81	1.04	96.54	3.14	0.32	0.04	0.97
9	929	2.94	1.04	929	2.94	1.05	96.32	3.46	0.22	0.04	0.97
10	929	2.37	1.00	929	2.37	1.01	96.00	3.57	0.43	0.05	0.97
12	929	3.18	0.94	929	3.18	0.94	97.94	1.62	0.43	0.03	0.97

* Mean absolute difference between first and second ratings

† Pearson correlation between first and second ratings

Table 8.C.3 Interrater Agreement Analyses for Operational Tasks: Level IV, Science

Task	First Rating			Second Rating			% Agreement			MAD *	Corr. †
	N	Mean	SD	N	Mean	SD	Exact	Adjacent	Neither		
1	613	2.49	0.93	613	2.50	0.94	97.04	2.63	0.33	0.03	0.98
3	613	2.96	1.12	613	2.95	1.13	96.05	3.78	0.16	0.04	0.98
4	613	2.64	0.99	613	2.62	1.01	94.41	4.28	1.32	0.07	0.94
6	613	3.05	1.05	613	3.03	1.07	95.07	4.28	0.66	0.06	0.95
7	613	2.33	1.29	613	2.31	1.29	94.41	4.77	0.82	0.07	0.97
9	613	3.02	0.90	613	3.02	0.91	96.71	2.80	0.49	0.04	0.96
10	613	3.09	1.05	613	3.07	1.07	95.39	4.11	0.49	0.05	0.96
12	613	2.74	1.08	613	2.74	1.09	95.39	3.95	0.66	0.06	0.97

* Mean absolute difference between first and second ratings

† Pearson correlation between first and second ratings

Table 8.C.4 Interrater Agreement Analyses for Operational Tasks: Level V, Science

Task	First Rating			Second Rating			% Agreement			MAD *	Corr. †
	N	Mean	SD	N	Mean	SD	Exact	Adjacent	Neither		
1	496	1.85	0.90	496	1.85	0.92	92.87	6.31	0.81	0.08	0.94
3	496	3.34	0.90	496	3.33	0.90	97.96	1.43	0.61	0.03	0.96
4	496	2.18	1.12	496	2.19	1.13	94.91	4.07	1.02	0.07	0.96
6	496	2.12	1.16	496	2.15	1.16	95.72	3.46	0.81	0.06	0.96
7	496	2.55	1.05	496	2.53	1.07	93.89	4.89	1.22	0.07	0.96
9	496	2.90	0.95	496	2.86	0.97	94.70	4.48	0.81	0.07	0.92
10	496	3.42	0.95	496	3.40	0.96	98.17	1.63	0.20	0.02	0.97
12	496	2.09	1.06	496	2.10	1.07	93.69	5.30	1.02	0.07	0.96

* Mean absolute difference between first and second ratings

† Pearson correlation between first and second ratings

Appendix 8.D—IRT Analyses

Table 8.D.1 Score Conversions: Level I, Science

Raw Score	Freq. Distrib.	Theta	Scale Score	CSEM	Performance Level
40	374	N/A	60	-	**Advanced**
39	127	1.5450	50	9	
38	81	1.1760	46	6	
37	72	0.9891	44	4	
36	214	0.8635	43	3	
35	111	0.7679	42	3	
34	72	0.6897	41	3	
33	55	0.6229	41	3	
32	194	0.5638	40	2	
31	103	0.5105	39	2	
30	71	0.4612	39	2	
29	68	0.4152	38	2	**Proficient**
28	166	0.3715	38	2	
27	114	0.3297	38	2	
26	67	0.2891	37	2	
25	62	0.2495	37	2	
24	167	0.2105	36	2	
23	88	0.1717	36	2	
22	67	0.1329	36	2	
21	52	0.0936	35	2	
20	158	0.0537	35	2	
19	82	0.0127	34	2	**Basic**
18	75	–0.0299	34	2	
17	54	–0.0745	34	2	
16	130	–0.1218	33	2	
15	73	–0.1726	33	2	
14	60	–0.2282	32	2	
13	60	–0.2900	31	3	
12	73	–0.3607	31	3	
11	58	–0.4440	30	3	
10	58	–0.5465	29	3	**Below Basic**
9	45	–0.6800	28	4	
8	102	–0.8675	26	5	
7	31	–1.1467	23	6	
6	34	–1.5357	19	6	
5	35	–1.9807	15	3	**Far Below Basic**
4	33	–2.4409	15	-	
3	25	–2.9317	15	-	
2	29	–3.5115	15	-	
1	31	–4.3565	15	-	
0	135	N/A	15	-	

Table 8.D.2 Score Conversions: Level III, Science

Raw Score	Freq. Distrib.	Theta	Scale Score	CSEM	Performance Level
32	48	N/A	60	-	**Advanced**
31	50	2.9363	46	8	
30	79	2.1811	44	3	
29	97	1.7168	42	2	
28	121	1.3718	41	2	
27	144	1.0914	40	2	
26	176	0.8507	39	2	**Proficient**
25	197	0.6363	38	2	
24	219	0.4398	38	2	
23	246	0.2559	37	1	
22	247	0.0809	36	1	
21	247	–0.0881	36	1	
20	200	–0.2532	35	1	
19	225	–0.4162	35	1	
18	176	–0.5787	34	1	**Basic**
17	148	–0.7422	34	1	
16	139	–0.9083	33	1	
15	119	–1.0785	32	1	
14	77	–1.2546	32	1	
13	75	–1.4385	31	2	
12	58	–1.6328	30	2	
11	30	–1.8400	30	2	
10	37	–2.0632	29	2	**Below Basic**
9	31	–2.3054	28	2	
8	27	–2.5688	27	2	
7	13	–2.8553	26	2	
6	11	–3.1662	25	2	
5	14	–3.5047	24	2	
4	10	–3.8796	23	2	
3	8	–4.3116	21	2	**Far Below Basic**
2	6	–4.8524	19	3	
1	8	–5.6714	16	2	
0	40	N/A	15	-	

Table 8.D.3 Score Conversions: Level IV, Science

Raw Score	Freq. Distrib.	Theta	Scale Score	CSEM	Performance Level
32	45	N/A	60	-	
31	88	2.6704	46	8	
30	132	1.9559	43	3	**Advanced**
29	157	1.5289	41	2	
28	190	1.2183	40	2	
27	201	0.9695	39	2	
26	206	0.7580	38	2	
25	198	0.5704	38	2	
24	200	0.3988	37	2	**Proficient**
23	204	0.2379	36	2	
22	208	0.0841	36	2	
21	213	–0.0653	35	1	
20	199	–0.2127	35	1	
19	182	–0.3601	34	1	
18	159	–0.5094	33	2	
17	149	–0.6625	33	2	
16	120	–0.8216	32	2	**Basic**
15	103	–0.9889	32	2	
14	75	–1.1666	31	2	
13	49	–1.3574	30	2	
12	41	–1.5635	29	2	
11	31	–1.7874	29	2	
10	30	–2.0306	28	2	
9	24	–2.2943	27	2	
8	24	–2.5787	25	2	**Below Basic**
7	9	–2.8840	24	2	
6	8	–3.2116	23	2	
5	9	–3.5661	22	2	
4	13	–3.9584	20	3	
3	5	–4.4120	18	3	
2	8	–4.9815	16	2	**Far Below Basic**
1	10	–5.8391	15	1	
0	25	N/A	15	-	

Table 8.D.4 Score Conversions: Level V, Science

Raw Score	Freq. Distrib.	Theta	Scale Score	CSEM	Performance Level
32	27	N/A	60	-	**Advanced**
31	39	3.3858	45	8	
30	52	2.6934	43	3	
29	94	2.2847	42	2	
28	116	1.9870	41	2	
27	134	1.7463	40	2	
26	170	1.5379	39	2	
25	161	1.3488	39	1	
24	197	1.1709	38	1	**Proficient**
23	232	0.9988	37	1	
22	233	0.8289	37	1	
21	267	0.6586	36	1	
20	241	0.4861	36	1	
19	235	0.3108	35	1	
18	188	0.1324	34	1	**Basic**
17	181	–0.0487	34	1	
16	121	–0.2323	33	1	
15	116	–0.4187	33	1	
14	92	–0.6095	32	1	
13	70	–0.8074	31	2	
12	60	–1.0167	31	2	
11	48	–1.2435	30	2	
10	43	–1.4956	29	2	**Below Basic**
9	32	–1.7809	28	2	
8	31	–2.1050	27	2	
7	12	–2.4657	26	2	
6	7	–2.8536	24	2	
5	9	–3.2611	23	2	
4	11	–3.6922	21	2	
3	7	–4.1685	20	2	**Far Below Basic**
2	5	–4.7442	18	3	
1	7	–5.5913	15	1	
0	29	N/A	15	-	

Appendix 8.E—Disability Distributions

Table 8.E.1 CAPA Primary Disability Distributions: Level I, Science

Disability	Frequency	Percent
Intellectual disability	1,437	38.8%
Hearing impairment	24	0.6%
Speech or language impairment	19	0.5%
Visual impairment	44	1.2%
Emotional disturbance*	–	–
Orthopedic impairment	440	11.9%
Other health impairment	97	2.6%
Specific learning disability	23	0.6%
Deaf–blindness*	–	–
Multiple disabilities	395	10.7%
Autism	1,039	28.0%
Traumatic brain injury	20	0.5%
Unknown	160	4.3%
TOTAL	3,706	100.0%

* Results for groups with fewer than 11 members are not reported.

Table 8.E.2 CAPA Primary Disability Distributions: Level III, Science

Disability	Frequency	Percent
Intellectual disability	1,153	34.7%
Hearing impairment	50	1.5%
Speech or language impairment	153	4.6%
Visual impairment	20	0.6%
Emotional disturbance	21	0.6%
Orthopedic impairment	105	3.2%
Other health impairment	196	5.9%
Specific learning disability	241	7.3%
Deaf–blindness*	–	–
Multiple disabilities	80	2.4%
Autism	1,176	35.4%
Traumatic brain injury	14	0.4%
Unknown	112	3.4%
TOTAL	3,323	100.0%

* Results for groups with fewer than 11 members are not reported.

Table 8.E.3 CAPA Primary Disability Distributions: Level IV, Science

Disability	Frequency	Percent
Intellectual disability	1,382	41.7%
Hearing impairmetn	39	1.2%
Speech or language impairment	86	2.6%
Visual impairment	13	0.4%
Emotional disturbance	21	0.6%
Orthopedic impairment	134	4.0%
Other health impairment	144	4.3%
Specific learning disability	209	6.3%
Deaf–blindness*	–	–
Multiple disabilities	87	2.6%
Autism	1,046	31.6%
Traumatic brain injury	28	0.8%
Unknown	126	3.8%
TOTAL	3,315	100.0%

* Results for groups with fewer than 11 members are not reported.

Table 8.E.4 CAPA Primary Disability Distributions: Level V, Science

Disability	Frequency	Percent
Intellectual disability	1,401	42.9%
Hearing impairment	45	1.4%
Speech or language impairment	50	1.5%
Visual impairment	20	0.6%
Emotional disturbance	36	1.1%
Orthopedic impairment	137	4.2%
Other health impairment	157	4.8%
Specific learning disability	241	7.4%
Deaf–blindness*	–	–
Multiple disabilities	96	2.9%
Autism	944	28.9%
Traumatic brain injury	22	0.7%
Unknown	116	3.6%
TOTAL	3,267	100.0%

* Results for groups with fewer than 11 members are not reported.

Chapter 9: Quality Control Procedures

Rigorous quality control procedures were implemented throughout the test development, administration, scoring, and reporting processes. As part of this effort, Educational Testing Service (ETS) maintains an Office of Testing Integrity (OTI) that resides in the ETS legal department. The OTI provides quality assurance services for all testing programs administered by ETS. In addition, the Office of Professional Standards Compliance at ETS publishes and maintains the *ETS Standards for Quality and Fairness* (ETS, 2002), 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 Task Development

The task development process for the California Alternate Performance Assessment (CAPA) for Science prior to the 2012–13 administration is described in detail in Chapter 3, starting on page 17; there was no new item development for the 2014–15 because of the form reuse. The next sections highlight elements of the process devoted specifically to the quality control of the tasks that were previously developed and reused during the 2014–15 CAPA for Science administration.

Task Specifications

ETS maintained task specifications for each CAPA for Science and developed an item utilization plan to guide the development of the tasks for each test. Task writing emphasis was determined in consultation with the California Department of Education (CDE). Adherence to the specifications ensured the maintenance of quality and consistency in the task development process.

Task Writers

The tasks for the CAPA for Science were written by task writers with a thorough understanding of the California content standards. The task writers were carefully screened and selected by senior ETS content staff and approved by the CDE. Only those with strong content and teaching backgrounds, experienced with students who have severe cognitive disabilities, were invited to participate in an extensive training program for task writers.

Internal Contractor Reviews

Once tasks were written, ETS assessment specialists made sure that each task underwent an intensive internal review process. Every step of this process is designed to produce tasks that exceed industry standards for quality. 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 tasks met the specifications for the process.

Content Review

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

The artwork and graphics for the tasks were created during the internal content review period so assessment specialists could evaluate the correctness and appropriateness of the art early in the task development process. ETS selected visuals that were relevant to the task 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 tasks for clarity, correctness of language, grade-level appropriateness of language, adherence to style guidelines, and conformity to acceptable task-writing practices. The editorial review also included rounds of copyediting and proofreading. ETS strives for error-free tasks beginning with the initial rounds of review.

Fairness Review

One of the final steps in the ETS internal review process is to have all tasks 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 tasks 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 task before it was presented to the CDE.

Assessment Review Panel Review

The Assessment Review Panels (ARPs) were panels that advised the CDE and ETS on areas related to task development for the CAPA for Science. The ARPs were responsible for reviewing all newly developed tasks for alignment to the California content standards. The ARPs also reviewed the tasks for accuracy of content, clarity of phrasing, and quality. See page 21 in Chapter 3 for additional information on the function of ARPs within the task-review process.

Statewide Pupil Assessment Review Panel Review

The Statewide Pupil Assessment Review (SPAR) panel was responsible for reviewing and approving the achievement tests that were used statewide for the testing of students in California public schools in grades two through eleven. The SPAR panel representatives ensured that the CAPA for Science tasks conformed to the requirements of *Education Code* Section 60602. See page 22 in Chapter 3 for additional information on the function of the SPAR panel within the task-review process.

Data Review of Field-tested Tasks

ETS field-tested newly developed tasks to obtain statistical information about task performance. This information was used to evaluate tasks that were candidates for use in operational test forms. These tasks that were flagged after field-test and operational use were examined carefully at data review meetings, where content experts discussed tasks that had poor statistics and did not meet the psychometric criteria for task quality. The CDE

defined the criteria for acceptable or unacceptable task statistics. These criteria ensured that the task (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 task for inclusion in the California item bank. The CDE content experts reviewed the recommendations and made the final decision on each task.

The field-test items that appeared in the CAPA for Science administered in 2014–15 were statistically reviewed in data review meetings in 2013, the year they were originally administered. There was no data review of field-test items in 2014–15.

Quality Control of the Item Bank

After the data review, tasks were placed in the item bank along with their statistics and reviewers' evaluations of their quality. ETS then delivered the tasks 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—California item bank updates for task 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 task development process, there were multiple senior reviews of tasks, 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 review test forms for accuracy and clueing issues.

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

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

Quality Control of Test Materials

Collecting Test Materials

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

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

Processing Test Materials

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

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

Quality Control of Scanning

Before any CAASPP documents are scanned, ETS conducts a complete check of the scanning system. ETS creates test decks for every test and form. Each test deck consists of approximately 700 answer documents marked to cover response ranges, demographic data, blanks, double marks, and other responses. Fictitious students are created to verify that each marking possibility is processed correctly by the scanning program. The output file generated as a result of this activity is thoroughly checked against each answer document after each 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 tasks on the diagnostic sheets, the scanner is recalibrated to work properly before being allowed to continue processing student documents.

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

Quality Control of Image Editing

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

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

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

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

Quality Control of Answer Document Processing and Scoring

Accountability of Answer Documents

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

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

Processing of Answer Documents

Prior to processing operational answer documents and executing subsequent data processing programs, ETS conducts an end-to-end test. As part of this test, ETS prepares approximately 700 test cases covering all tests and many scenarios designed to exercise particular business rule logic. ETS marks answer documents for those 700 test cases. They are then scanned, scored, and aggregated. The results at various inspection points are checked by psychometricians and Data Quality Services staff. Additionally, a post-scan test file of approximately 50,000 records across the CAASPP System is scored and aggregated to test a broader range of scoring and aggregation scenarios. These procedures assure that students and LEAs receive the correct scores when the actual scoring process is carried out. In 2014–15, 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 that testing materials are processed and scored accurately. These documents include the Scoring Rules specifications and the Include Indicators specifications. Each is explained in detail in Chapter 7, starting on page 46. The scoring specifications are reviewed and revised by the CDE and ETS each year. After a version that all parties endorse is finalized, the CDE issues a formal approval of the scoring and reporting specifications.

Storing Answer Documents

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

Quality Control of Psychometric Processes

Quality Control of Task (Item) Analyses and the Scoring Process

When the forms were first administered in the 2012–13 administration, 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 CAPA for Science following its original administration. Quality assurance checks also included a comparison of the current year's statistics to statistics from previous years. The results of preliminary classical task analyses that provided a check on scoring reasonableness and the application of scoring rubrics were also reviewed by a senior psychometrician. The tasks that were flagged for questionable statistical attributes were sent to test development staff for their review; their comments were reviewed by the psychometricians before tasks 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 are performed for each process 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 13, checks were made to ascertain that the correct options for the analyses were selected. Checks were also made on the number of tasks, number of examinees with valid scores, IRT Rasch task difficulty estimates, standard errors for the Rasch task difficulty estimates, and the match of selected statistics to the results on the same statistics obtained during preliminary task 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 15 and the maximum reported scale score was 60; 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. 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.

Score Verification Process

ETS utilizes the raw-to-scale scoring tables to assign scale scores for each student and verifies scale scores by independently generating the scale scores for students in a small number of LEAs. 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 CAPA for Science results. Year-

to-year comparison analyses use 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 2012–13 administration of the CAPA for Science forms that were reused in 2014–15, the statistics based on classical task analyses were obtained to ensure the stability of the statistics. The resulting classical 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.

Quality Control of Reporting

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

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

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

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

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

Electronic Reporting

Because no equating was conducted during the 2014–15 administration, students’ scale scores and performance levels for the CAPA for Science were made available to LEAs prior to the printing of paper reports. The reporting module in the Test Operations Management System made it possible for LEAs to securely download an electronic reporting file containing these results.

Before an LEA can download a student data file, ETS statisticians approved a QC file of test results data and ETS IT successfully processed it. Once the data were deemed reliable and

ETS processed a scorable answer document for every student who took the CAPA for Science in that test administration for the LEA, the LEA was notified that these results were available.

Excluding Student Scores from Summary Reports

ETS provides specifications to the CDE that document when to exclude student scores from summary reports. These specifications include the logic for handling answer documents that, for example, indicate the student was absent, was not tested due to parent/guardian request, or did not complete the test due to illness.

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 California Alternate Performance Assessment (CAPA) for Science results are routinely performed to identify the trends in examinee performance and test characteristics over time. Such comparisons were performed for the three most recent years of administration—2013, 2014, and 2015—and for the 2009 base year.

The indicators of examinee performance include the mean and standard deviation of scale scores, observed scale 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 score reliability, and standard errors of measurement (SEM), as well as the mean item response theory (IRT) *b*-value for each CAPA for Science.

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

The CAPA were first administered in 2003. Subsequently, the CAPA were revised to better link them to the grade-level California content standards adopted in 1998. The revised blueprints for the CAPA were approved by the SBE in 2006 for implementation beginning in 2008; new tasks were developed to meet the revised blueprints and then field-tested.

A standard setting was held in the fall of 2008 to establish new cut scores for the below basic, basic, proficient, and advanced performance levels based on the revised test blueprint for Levels I and III through V in science. The 2008–09 administration was the first in which test results were reported using the new scales and cut scores for the four performance levels; thus, 2009 became the base year.

Examinee Performance

Table 10.A.1 on page 101 contains the number of examinees assessed and the means and standard deviations of examinees' scale scores in the base year (2009) and in 2013, 2014, and 2015 for each CAPA for Science. As noted in previous chapters, the CAPA for Science reporting scales range from 15 to 60 for all levels.

CAPA for Science scale scores are used to classify student results into one of five performance levels: far below basic, below basic, basic, proficient, and advanced. The percentages of students qualifying for the proficient and advanced levels are presented in Table 10.A.2 on page 101; please note that this information may differ slightly from information found on the California Department of Education California Assessment of Student Performance and Progress 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. The percentages of student receiving a score in the advanced performance levels are presented in Table 10.A.3.

Table 10.A.4 through Table 10.A.7 show for each CAPA for Science the distribution of scale scores observed in the base year, 2013, 2014, and 2015. Frequency counts are provided for each scale score interval of 3. A frequency count of "N/A" indicates that there are no obtainable scale scores within that scale-score range. For all CAPA for Science tests, a minimum score of 30 is required for a student to reach the basic level of performance, and a minimum score of 35 is required for a student to reach the proficient level of performance.

Test Characteristics

The item (task) and test analysis results of the CAPA for Science over the comparison years indicate that the CAPA for Science meet the technical criteria established in professional standards for high-stakes tests. In addition, every year, efforts were made to improve the technical quality of each CAPA for Science.

Table 10.B.1 and Table 10.B.2 in Appendix 10.B, which start on page 104, present, respectively, the average task scores and the equated item response theory (IRT) *b*-value means for the tasks in each CAPA for Science. The average task scores were affected by both the difficulty of the items and the abilities of the students administered the tasks.

The average polyserial correlations for the CAPA 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 and Table 10.B.5. Like the average item score, polyserial 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 CAPA for Science Across Base Year (2009), 2013, 2014, and 2015

Level	2009 Valid Scores	2013 Valid Scores	2014 Valid Scores	2015 Valid Scores	2009 Mean	2009 S.D.	2013 Mean	2013 S.D.	2014 Mean	2014 S.D.	2015 Mean	2015 S.D.
Level I Science	3,296	3,724	3,800	3,706	35.59	11.25	37.35	10.29	37.61	11.14	37.55	10.84
Level III Science	3,267	3,446	3,551	3,323	36.24	5.45	36.10	4.63	36.09	4.65	35.99	5.43
Level IV Science	3,190	3,275	3,290	3,315	35.56	5.53	35.91	5.37	35.73	5.69	36.01	5.53
Level V Science	3,396	3,435	3,450	3,267	35.35	5.34	35.84	4.98	35.86	5.12	35.89	4.83

Table 10.A.2 Percentage of Proficient and Above Across Base Year (2009), 2013, 2014, and 2015

Level	Base	2013	2014	2015
Level I Science	59%	68%	66%	67%
Level III Science	69%	71%	70%	69%
Level IV Science	58%	66%	66%	67%
Level V Science	61%	66%	65%	67%

Table 10.A.3 Percentage of Advanced Across Base Year (2009), 2013, 2014, and 2015

Level	Base	2013	2014	2015
Level I Science	33%	39%	41%	40%
Level III Science	19%	17%	16%	16%
Level IV Science	15%	17%	18%	18%
Level V Science	17%	24%	24%	24%

Table 10.A.4 Observed Score Distributions of CAPA for Science Across Base Year (2009), 2013, 2014, and 2015 for Science, Level I

Observed Score Distributions	Base	2013	2014	2015
60	280	322	414	374
57–59	N/A	N/A	N/A	N/A
54–56	N/A	N/A	N/A	N/A
51–53	N/A	N/A	N/A	N/A
48–50	81	123	117	127
45–47	69	80	87	81
42–44	267	403	439	397
39–41	394	518	514	495
36–38	588	846	767	799
33–35	611	609	568	624
30–32	271	272	276	251
27–29	108	92	86	103
24–26	207	131	153	102
21–23	N/A	48	35	31
18–20	49	32	45	34
15–17	371	248	299	288

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

Table 10.A.5 Observed Score Distributions of CAPA for Science Across Base Year (2009), 2013, 2014, and 2015 for Science, Level III

Observed Score Distribution	Base	2013	2014	2015
60	69	19	32	48
57–59	N/A	N/A	N/A	N/A
54–56	N/A	N/A	N/A	N/A
51–53	N/A	N/A	N/A	N/A
48–50	N/A	N/A	N/A	N/A
45–47	105	55	45	50
42–44	122	188	184	176
39–41	493	521	498	441
36–38	934	1,224	1,248	1156
33–35	1,093	885	980	888
30–32	268	363	393	359
27–29	104	105	88	95
24–26	29	37	42	38
21–23	20	22	21	18
18–20	10	1	4	6
15–17	20	26	16	48

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

Table 10.A.6 Observed Score Distributions of CAPA for Science Across Base Year (2009), 2013, 2014, and 2015 for Science, Level IV

Observed Score Distributions	Base	2013	2014	2015
60	46	50	50	45
57–59	N/A	N/A	N/A	N/A
54–56	N/A	N/A	N/A	N/A
51–53	N/A	N/A	N/A	N/A
48–50	N/A	N/A	N/A	N/A
45–47	44	79	75	88
42–44	157	107	110	132
39–41	393	496	551	548
36–38	1,010	1,003	960	1016
33–35	864	927	853	902
30–32	420	376	367	347
27–29	155	129	196	126
24–26	36	56	52	33
21–23	10	17	17	17
18–20	19	13	20	18
15–17	36	22	39	43

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

Table 10.A.7 Observed Score Distributions of CAPA for Science Across Base Year (2009), 2013, 2014, and 2015 for Science, Level V

Observed Score Distributions	Base	2013	2014	2015
60	33	38	46	27
57–59	N/A	N/A	N/A	N/A
54–56	N/A	N/A	N/A	N/A
51–53	N/A	N/A	N/A	N/A
48–50	N/A	N/A	N/A	N/A
45–47	46	50	50	39
42–44	129	137	133	146
39–41	373	588	589	581
36–38	1,288	1,217	1,163	1170
33–35	874	852	929	841
30–32	332	335	327	270
27–29	196	135	123	106
24–26	36	19	27	19
21–23	25	20	12	20
18–20	14	13	19	12
15–17	50	31	32	36

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 Average Item Score of CAPA for Science Operational Test Tasks Across Base Year (2009), 2013, 2014, and 2015

Level	Base	2013	2014	2015
Level I Science	2.75	3.04	3.07	3.07
Level III Science	2.71	2.63	2.63	2.60
Level IV Science	2.47	2.69	2.68	2.73
Level V Science	2.47	2.53	2.53	2.56

Table 10.B.2 Mean IRT *b*-values for Operational Test Tasks Across Base Year (2009), 2013, 2014, and 2015

Level	Base	2013	2014	2015
Level I Science	–0.23	–0.32	–0.32	–0.32
Level III Science	–1.29	–1.10	–1.10	–1.10
Level IV Science	–0.95	–1.14	–1.14	–1.14
Level V Science	–0.54	–0.57	–0.57	–0.57

Table 10.B.3 Mean Polyserial Correlation of CAPA for Science Operational Test Tasks Across Base Year (2009), 2013, 2014, and 2015

Level	Base	2013	2014	2015
Level I Science	0.82	0.79	0.81	0.80
Level III Science	0.75	0.74	0.73	0.76
Level IV Science	0.75	0.74	0.75	0.74
Level V Science	0.78	0.75	0.76	0.75

Table 10.B.4 Score Reliabilities of CAPA for Science Across Base Year (2009), 2013, 2014, and 2015

Level	Base	2013	2014	2015
Level I Science	0.91	0.88	0.90	0.89
Level III Science	0.85	0.85	0.84	0.86
Level IV Science	0.85	0.85	0.86	0.85
Level V Science	0.87	0.85	0.85	0.84

Table 10.B.5 SEM of CAPA for Science Across Base Year (2009), 2013, 2014, and 2015

Level	Base	2013	2014	2015
Level I Science	3.76	3.97	3.83	3.62
Level III Science	2.43	2.27	2.27	2.02
Level IV Science	2.46	2.35	2.35	2.12
Level V Science	2.30	2.32	2.33	1.94