

**California Department of Education
Assessment Development and
Administration Division**



**California Modified Assessment
Technical Report
Spring 2014–15 Administration**

**Submitted March 23, 2016
Educational Testing Service
Contract No. 5417**

Table of Contents

Acronyms and Initialisms Used in the <i>CMA Technical Report</i>	vii
Chapter 1: Introduction	1
Background	1
Test Purpose	1
Test Content	1
Intended Population	2
Intended Use and Purpose of Test Scores	2
Testing Window	3
Significant CAASPP Developments in 2014–15	3
Reduction in Paper Reporting.....	3
Reporting Cluster Data	3
Origin of Demographic Data.....	3
Updated Accessibility Supports.....	3
Individualized Aid Option	3
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 CMA for Science Processes	8
Item Development	8
Item Formats.....	8
Item Specifications.....	8
Item Banking.....	8
Item Refresh Rate.....	9
Test Assembly	9
Test Length.....	9
Test Blueprints.....	9
Content Rules and Item Selection.....	9
Psychometric Criteria.....	10
Test Administration	10
Test Security and Confidentiality.....	10
Procedures to Maintain Standardization	11
Universal Tools, Designated Supports, and Accommodations	12
Non-embedded Supports.....	12
Individualized Aids (now “Unlisted Resources”).....	12
Special Services Summaries	12
Scores	13
Aggregation Procedures	13
Equating	14
Post-Equating	14
Pre-Equating.....	14
References	18
Appendix 2.A—CMA for Science Items and Estimated Time Chart	19
Appendix 2.B—Reporting Clusters for Science	20
Science Modified Standards Assessment (Grade Five).....	20
Science Modified Standards Assessment (Grade Eight)	20
Science Modified Standards Assessment (Grade Ten)	20
Appendix 2.C—Universal Tools, Designated Supports, and Accommodations for the California Assessment of Student Performance and Progress	21
Appendix 2.D—Special Service Summary Tables	22
Chapter 3: Item Development	27
Rules for Item Development	27
Item Specifications.....	27
Expected Item Ratio.....	28

Selection of Item Writers	28
Criteria for Selecting Item Writers	28
Item Review Process	29
Contractor Review	29
Content Expert Reviews	30
Statewide Pupil Assessment Review Panel	33
Field Testing	33
Stand-alone Field Testing	33
Embedded Field-test Items	34
CDE Data Review	34
Item Banking	34
References	36
Chapter 4: Test Assembly	37
Test Length	37
Rules for Item Selection	37
Test Blueprint	37
Content Rules and Item Selection	37
Psychometric Criteria	38
Projected Psychometric Properties of the Assembled Tests	39
Rules for Item Sequence and Layout	40
Reference	41
Appendix 4.A—Technical Characteristics	42
Appendix 4.B—Cluster Targets	43
Chapter 5: Test Administration	46
Test Security and Confidentiality	46
ETS’s Office of Testing Integrity	46
Test Development	46
Item and Data Review	47
Item Banking	47
Transfer of Forms and Items to the CDE	47
Security of Electronic Files Using a Firewall	48
Printing and Publishing	48
Test Administration	48
Test Delivery	48
Processing and Scoring	49
Data Management	49
Statistical Analysis	50
Reporting and Posting Results	50
Student Confidentiality	50
Student Test Results	50
Procedures to Maintain Standardization	51
Test Administrators	51
Directions for Administration	52
CAASPP Paper-Pencil Testing Test Administration Manual	52
Test Operations Management System Manuals	53
Test Booklets	53
Universal Tools, Designated Supports, and Accommodations	53
Identification	53
Scoring	54
Testing Incidents	54
Social Media Security Breaches	54
Testing Improprieties	54
References	55
Chapter 6: Performance Standards	56
Background	56
Standard-Setting Procedure	56
Development of Competencies Lists	57
Standard-Setting Methodology	58
Bookmark Method	58
Results	59
References	60
Chapter 7: Scoring and Reporting	61
Procedures for Maintaining and Retrieving Individual Scores	61

Scoring and Reporting Specifications	61
Scanning and Scoring.....	61
Types of Scores and Subscores	62
Raw Score	62
Subscore.....	62
Scale Score.....	62
Performance Levels	62
Score Verification Procedures	62
Scoring Key Verification Process.....	62
Overview of Score Aggregation Procedures	63
Individual Scores.....	63
Reports Produced and Scores for Each Report	65
Types of Score Reports	65
Student Score Report Contents	66
Student Score Report Applications	66
Criteria for Interpreting Test Scores	67
Criteria for Interpreting Score Reports.....	67
Reference.....	68
Appendix 7.A—Scale Score Distribution Tables	69
Appendix 7.B—Demographic Summaries.....	70
Appendix 7.C—Types of Score Reports.....	76
Chapter 8: Analyses.....	78
Background	78
Samples Used for the Analyses	78
Classical Item Analyses.....	79
Multiple-Choice Items	79
Reliability Analyses.....	79
Intercorrelations, Reliabilities, and SEMs for Reporting Clusters	81
Subgroup Reliabilities and SEMs.....	81
Conditional Standard Errors of Measurement.....	81
Decision Classification Analyses	82
Validity Evidence.....	83
Purpose of the CMA for Science.....	84
The Constructs to Be Measured	84
Interpretations and Uses of the Scores Generated.....	84
Intended Test Population(s).....	85
Validity Evidence Collected.....	85
Evidence Based on Response Processes	87
Evidence Based on Internal Structure.....	87
Evidence Based on Consequences of Testing.....	89
IRT Analyses.....	89
Post-Equating	89
Pre-Equating.....	89
Summaries of Scaled IRT <i>b</i> -values.....	90
Evaluation of Pre-Equating	90
Equating Results.....	90
Differential Item Functioning Analyses	91
References	93
Appendix 8.A—Classical Analyses.....	95
Appendix 8.B—Reliability Analyses	99
Appendix 8.C—IRT Analyses	113
Chapter 9: Quality Control Procedures.....	116
Quality Control of Item Development	116
Item Specifications.....	116
Item Writers.....	116
Internal Contractor Reviews.....	116
Assessment Review Panel Review.....	117
Statewide Pupil Assessment Review Panel Review	117
Data Review of Field-tested Items	117
Quality Control of the Item Bank.....	118
Quality Control of Test Form Development	119
Quality Control of Test Materials	119
Collecting Test Materials.....	119

Processing Test Materials.....	119
Quality Control of Scanning	120
Quality Control of Image Editing	120
Quality Control of Answer Document Processing and Scoring	120
Accountability of Answer Documents.....	120
Processing of Answer Documents.....	121
Scoring and Reporting Specifications.....	121
Storing Answer Documents.....	121
Quality Control of Psychometric Processes	121
Score Key Verification Procedures.....	121
Quality Control of Item Analyses and the Equating Process.....	121
Score Verification Process.....	123
Year-to-Year Comparison Analyses.....	123
Offloads to Test Development.....	123
Quality Control of Reporting	123
Electronic Reporting.....	124
Excluding Student Scores from Summary Reports.....	124
Reference	125
Chapter 10: Historical Comparisons	126
Base Year Comparisons	126
Examinee Performance	127
Test Characteristics	127
Appendix 10.A—Historical Comparisons Tables, Examinee Performance	128
Appendix 10.B—Historical Comparisons Tables, Test Characteristics	131

Tables

Table 2.1 Scale-Score Ranges for Performance Levels.....	17
Table 2.C.1 Matrix One Part 2: Non-Embedded Supports for the CMA for Science.....	21
Table 2.D.1 Special Service Summary for Science, Grades Five, Eight, and Ten (Life Science)—All Tested.....	22
Table 2.D.2 Special Service Summary for Science, Grades Five, Eight, and Ten (Life Science)—English-Only Students.....	23
Table 2.D.3 Special Service Summary for Science, Grades Five, Eight, and Ten (Life Science)—Initially Fluent English Proficient (I-FEP) Students.....	24
Table 2.D.4 Special Service Summary for Science, Grades Five, Eight, and Ten (Life Science)—English Learner (EL) Students.....	25
Table 2.D.5 Special Service Summary for Science, Grades Five, Eight, and Ten (Life Science)—Reclassified Fluent English Proficient (R-FEP) Students.....	26
Table 3.1 Stand-alone Field-testing Timeline for the CMA for Science.....	33
Table 4.1 Statistical Targets for CMA for Science Test Assembly.....	39
Table 4.A.1 Summary of 2015 CMA for Science Projected Raw Score Statistics.....	42
Table 4.A.2 Summary of 2015 CMA for Science Projected Item Statistics.....	42
Table 7.1 Mean and Standard Deviation of Raw and Scale Scores for the CMA for Science.....	63
Table 7.2 Percentages of Examinees in Each Performance Level.....	64
Table 7.3 Subgroup Definitions.....	64
Table 7.4 Types of CMA for Science Reports.....	66
Table 7.A.1 Distribution of CMA for Science Scale Scores for Science.....	69
Table 7.B.1 Demographic Summary for Science, Grade Five.....	70
Table 7.B.2 Demographic Summary for Science, Grade Eight.....	72
Table 7.B.3 Demographic Summary for Life Science (Grade 10).....	74
Table 7.C.1 Score Reports Reflecting CMA for Science Results.....	76
Table 8.1 Mean and Median Proportion Correct and Point-Biserial by Test Form—Current Administration.....	79
Table 8.2 Reliabilities and SEMs for the CMA for Science.....	80
Table 8.3 Scale Score CSEM at Performance-level Cut Points.....	82
Table 8.4 Original Year of Administration for the CMA for Science.....	87
Table 8.A.1 Item-by-item p -value and Point Biserial for Science, Grade Five—Current Year (2015) and Original Year of Administration.....	95
Table 8.A.2 Item-by-item p -value and Point Biserial for Science, Grade Eight—Current Year (2015) and Original Year of Administration.....	96
Table 8.A.3 Item-by-item p -value and Point Biserial for Science, Grade Ten—Current Year (2015) and Original Year of Administration.....	97
Table 8.B.1 Subscore Reliabilities and Intercorrelations for Science.....	99
Table 8.B.2 Reliabilities and SEMs for the CMA for Science by Gender (Male).....	99
Table 8.B.3 Reliabilities and SEMs for the CMA for Science by Gender (Female).....	99

Table 8.B.4 Reliabilities and SEMs for the CMA for Science by Economic Status (Economically Disadvantaged)	99
Table 8.B.5 Reliabilities and SEMs for the CMA for Science by Economic Status (Not Economically Disadvantaged)....	100
Table 8.B.6 Reliabilities and SEMs for the CMA for Science by English-language Fluency (English Only).....	100
Table 8.B.7 Reliabilities and SEMs for the CMA for Science by English-language Fluency (Initially Fluent English Proficient).....	100
Table 8.B.8 Reliabilities and SEMs for the CMA for Science by English-language Fluency (English Learner).....	100
Table 8.B.9 Reliabilities and SEMs for the CMA for Science by English-language Fluency (Reclassified Fluent English Proficient)	100
Table 8.B.10 Reliabilities and SEMs for the CMA Science by English-language Fluency (Unknown)	100
Table 8.B.11 Reliabilities and SEMs for the CMA for Science by Primary Ethnicity (African American).....	101
Table 8.B.12 Reliabilities and SEMs for the CMA for Science by Primary Ethnicity (American Indian)	101
Table 8.B.13 Reliabilities and SEMs for the CMA for Science by Primary Ethnicity (Asian)	101
Table 8.B.14 Reliabilities and SEMs for the CMA for Science by Primary Ethnicity (Filipino).....	101
Table 8.B.15 Reliabilities and SEMs for the CMA for Science by Primary Ethnicity (Hispanic).....	101
Table 8.B.16 Reliabilities and SEMs for the CMA for Science by Primary Ethnicity (Pacific Islander).....	101
Table 8.B.17 Reliabilities and SEMs for the CMA for Science by Primary Ethnicity (White).....	101
Table 8.B.18 Reliabilities and SEMs for the CMA for Science by Primary Ethnicity (Unknown)	102
Table 8.B.19 Reliabilities and SEMs for the CMA for Science by Primary Ethnicity-for-Not-Economically-Disadvantaged (African American).....	102
Table 8.B.20 Reliabilities and SEMs for the CMA for Science by Primary Ethnicity-for-Not-Economically-Disadvantaged (American Indian).....	102
Table 8.B.21 Reliabilities and SEMs for the CMA for Science by Primary Ethnicity-for-Not-Economically-Disadvantaged (Asian).....	102
Table 8.B.22 Reliabilities and SEMs for the CMA for Science by Primary Ethnicity-for-Not-Economically-Disadvantaged (Filipino)	102
Table 8.B.23 Reliabilities and SEMs for the CMA for Science by Primary Ethnicity-for-Not-Economically-Disadvantaged (Hispanic)	102
Table 8.B.24 Reliabilities and SEMs for the CMA for Science by Primary Ethnicity-for-Not-Economically-Disadvantaged (Pacific Islander).....	103
Table 8.B.25 Reliabilities and SEMs for the CMA for Science by Primary Ethnicity-for-Not-Economically-Disadvantaged (White).....	103
Table 8.B.26 Reliabilities and SEMs for the CMA for Science by Primary Ethnicity-for-Not-Economically-Disadvantaged (Unknown).....	103
Table 8.B.27 Reliabilities and SEMs for the CMA for Science by Primary Ethnicity-for-Economically-Disadvantaged (African American).....	103
Table 8.B.28 Reliabilities and SEMs for the CMA for Science by Primary Ethnicity-for-Economically-Disadvantaged (American Indian).....	103
Table 8.B.29 Reliabilities and SEMs for the CMA for Science by Primary Ethnicity-for-Economically-Disadvantaged (Asian).....	103
Table 8.B.30 Reliabilities and SEMs for the CMA for Science by Primary Ethnicity-for-Economically-Disadvantaged (Filipino)	104
Table 8.B.31 Reliabilities and SEMs for the CMA for Science by Primary Ethnicity-for-Economically-Disadvantaged (Hispanic)	104
Table 8.B.32 Reliabilities and SEMs for the CMA for Science by Primary Ethnicity-for-Economically-Disadvantaged (Pacific Islander).....	104
Table 8.B.33 Reliabilities and SEMs for the CMA for Science by Primary Ethnicity-for-Economically-Disadvantaged (White).....	104
Table 8.B.34 Reliabilities and SEMs for the CMA for Science by Primary Ethnicity-for-Economically-Disadvantaged (Unknown).....	104
Table 8.B.36 Reliabilities and SEMs for the CMA for Science by Gender by Economic Status (Not Economically Disadvantaged).....	105
Table 8.B.37 Reliabilities and SEMs for the CMA for Science by Primary Disability (Autism).....	105
Table 8.B.38 Reliabilities and SEMs for the CMA for Science by Primary Disability (Deaf-Blindness)	105
Table 8.B.39 Reliabilities and SEMs for the CMA for Science by Primary Disability (Emotional Disturbance).....	105
Table 8.B.40 Reliabilities and SEMs for the CMA for Science by Primary Disability (Hearing Impairment)	105
Table 8.B.41 Reliabilities and SEMs for the CMA for Science by Primary Disability (Mental Retardation)	105
Table 8.B.42 Reliabilities and SEMs for the CMA for Science by Primary Disability (Multiple Disabilities)	106
Table 8.B.43 Reliabilities and SEMs for the CMA for Science by Primary Disability (Orthopedic Impairment)	106
Table 8.B.44 Reliabilities and SEMs for the CMA for Science by Primary Disability (Other Health Impairment)	106
Table 8.B.45 Reliabilities and SEMs for the CMA for Science by Primary Disability (Specific Learning Disability).....	106
Table 8.B.46 Reliabilities and SEMs for the CMA for Science by Primary Disability (Speech or Language Impairment) ..	106
Table 8.B.47 Reliabilities and SEMs for the CMA for Science by Primary Disability (Traumatic Brain Injury).....	106
Table 8.B.48 Reliabilities and SEMs for the CMA for Science by Primary Disability (Visual Impairment).....	107
Table 8.B.49 Reliabilities and SEMs for the CMA for Science by Primary Disability (Unknown).....	107
Table 8.B.50 Overall Subgroup Reliabilities.....	107

Table 8.B.51 Overall Subgroup Reliabilities—Primary Ethnicity	107
Table 8.B.52 Overall Subgroup Reliabilities by Primary Ethnicity—Not Economically Disadvantaged	107
Table 8.B.53 Overall Subgroup Reliabilities by Primary Ethnicity—Economically Disadvantaged	107
Table 8.B.54 Overall Subgroup Reliabilities by Gender/Economic Status	108
Table 8.B.55 Overall Subgroup Reliabilities by Primary Disability	108
Table 8.B.56 Overall Subgroup Reliabilities by Primary Disability (continued).....	108
Table 8.B.57 Subscore Reliabilities and SEM for Science by Gender/Economic Status	108
Table 8.B.58 Subscore Reliabilities and SEM for Science by English-language Fluency	109
Table 8.B.59 Subscore Reliabilities and SEM for Science by Primary Ethnicity	109
Table 8.B.60 Subscore Reliabilities and SEM for Science by Primary Ethnicity-for-Not Economically Disadvantaged	110
Table 8.B.61 Subscore Reliabilities and SEM for Science by Primary Ethnicity-for-Economically Disadvantaged	110
Table 8.B.62 Subscore Reliabilities and SEM for Science by Disability	111
Table 8.B.63 Subscore Reliabilities and SEM for Science by Disability (continued)	111
Table 8.B.64 Reliability of Classification for Science, Grade Five	111
Table 8.B.65 Reliability of Classification for Science, Grade Eight	112
Table 8.B.66 Reliability of Classification for Life Science (Grade 10)	112
Table 8.C.1 Conversions for Science, Grade Five	113
Table 8.C.2 Conversions for Science, Grade Eight.....	114
Table 8.C.3 Conversions for Life Science, Grade Ten	115
Table 10.1 Base Years for the CMA for Science.....	126
Table 10.A.1 Number of Examinees Tested Across Base Year, 2013, 2014, and 2015	128
Table 10.A.2 Scale Score Means and Standard Deviations of CMA for Science Across Base Year, 2013, 2014, and 2015	128
Table 10.A.3 Percentage of Proficient and Above Across Base Year, 2013, 2014, and 2015	128
Table 10.A.4 Percentage of Advanced Across Base Year, 2013, 2014, and 2015	128
Table 10.A.5 Observed Score Distributions of CMA for Science Across Base Year, 2013, 2014, and 2015 for Science, Grade Five	129
Table 10.A.6 Observed Score Distributions of CMA for Science Across Base Year, 2013, 2014, and 2015 for Science, Grade Eight.....	129
Table 10.A.7 Observed Score Distributions of CMA for Science Across Base Year, 2013, 2014, and 2015 for Life Science (Grade Ten).....	130
Table 10.B.1 Mean Proportion Correct for Operational Test Items Across Base Year, 2013, 2014, and 2015.....	131
Table 10.B.2 Mean IRT <i>b</i> -values for Operational Test Items Across Base Year, 2013, 2014, and 2015.....	131
Table 10.B.3 Mean Point-Biserial Correlation for Operational Test Items Across Base Year, 2013, 2014, and 2015.....	131
Table 10.B.4 Score Reliabilities (Cronbach’s Alpha) Across Base Year, 2013, 2014, and 2015	131
Table 10.B.5 SEM Across Base Year, 2013, 2014, and 2015.....	131

Figures

Figure 3.1 The ETS Item Development Process for the CAASPP System	27
Figure 4.A.1 Plots of Target Information Function and Projected Information for Total Test for Science	42
Figure 4.B.1 Plots of Target Information Functions and Projected Information for Clusters for Science, Grade Five	43
Figure 4.B.2 Plots of Target Information Functions and Projected Information for Clusters for Science, Grade Eight.....	44
Figure 4.B.3 Plots of Target Information Functions and Projected Information for Clusters for Life Science, Grade Ten ...	45
Figure 6.1 Bookmark Standard-setting Process for the CMA.....	58
Figure 8.1 Decision Accuracy for Achieving a Performance Level.....	83
Figure 8.2 Decision Consistency for Achieving a Performance Level	83

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

ADA	Americans with Disabilities Act	IRT	item response theory
AERA	American Educational Research Association	IT	Information Technology
APA	American Psychological Association	LEA	local educational agency
ARP	Assessment Review Panel	MC	multiple choice
CAASPP	California Assessment of Student Performance and Progress	MCE	Manually Coded English
CAHSEE	California High School Exit Examination	MH DIF	Mantel-Haenszel DIF
CalTAC	California Technical Assistance Center	NCME	National Council on Measurement in Education
CAPA	California Alternate Performance Assessment	NPS	nonpublic, nonsectarian school
CCR	<i>California Code of Regulations</i>	OIB	ordered item booklet
CDE	California Department of Education	OTI	Office of Testing Integrity
CDS	county/district/school	<i>p</i> -value	item proportion correct
CELDT	California English Language Development Test	Pt-Bis	point-biserial correlations
CI	confidence interval	QC	quality control
CMA	California Modified Assessment	R-FEP	reclassified fluent English proficient
CSEMs	conditional standard errors of measurement	SBE	State Board of Education
CSTs	California Standards Tests	SD	standard deviation
DFA	<i>Directions for Administration</i>	SEM	standard error of measurement
DIF	differential item functioning	SFTP	secure file transfer protocol
DOK	depth of knowledge	SGID	School and Grade Identification sheet
EC	<i>Education Code</i>	SKM	score key management
EL	English learner	SPAR	Statewide Pupil Assessment Review
ELA	English–language arts	STAR	Standardized Testing and Reporting
ETS	Educational Testing Service	STS	Standards-based Tests in Spanish
FIA	final item analysis	TBD	To Be Determined
GENASYS	Generalized Analysis System	TIF	test information function
ICC	item characteristic curve	TOMS	Test Operations Management System
IEP	individualized education program	USDOE	United States Department of Education
I-FEP	initially fluent English proficient	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, mathematics, history–social science, and science. These standards were designed to provide state-level input into instruction curricula and serve as a foundation for the state’s school accountability programs.

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

During the 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

Test Purpose

The purpose of the three CMA for Science is to allow students with disabilities in grades five, eight, and ten greater access to an assessment that helps measure their achievement with respect to California’s content standards in science that were adopted by the SBE in 1998. These standards describe what students should know and be able to do at each grade level.

Test Content

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

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

Intended Population

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

The CMA for Science are designed for students with an IEP who meet eligibility criteria adopted by the SBE. The decision to administer the CMA for Science is made by a student's IEP team. The student's IEP team makes the decision annually by evaluating the student's progress on multiple measures. The IEP team must specify annually the CMA for Science the student is assigned to take.

Parents may submit a written request to have their child exempted from taking any or all parts of the tests within the CAASPP System. Only students whose parents submit a written request may be exempted from taking the tests (*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>*.)

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

Reporting Cluster Data

Reporting cluster data were not used in reporting student results to LEAs or test sites, or in Student Score Reports. However, reporting cluster results are included in this technical report.

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.

Updated Accessibility Supports

The following non-embedded accessibility supports are no longer denoted in Matrix One of the California Department of Education (CDE's) “Universal Tools, Designated Supports, and Accommodations for the California Assessment of Student Performance and Progress” for use on the science tests:

- The student used an assistive device that did not interfere with the independent work of the student
- The student used math manipulatives on the science tests

Individualized Aid Option

The option to note that an individualized aid was used by the student was included on the answer document.

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 local assessments, classroom tests, student grades, classroom work, and teacher recommendations in addition to the child's CMA for Science results (CDE, 2013).

Out-of-Level Testing

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

Score Comparison

When comparing scale score results for the CMA for Science, the reviewer is limited to comparing results only within the same content area and grade. For example, it is appropriate to compare scores obtained by students and/or schools on the 2014–15 grade five science test; it would not be appropriate to compare scores obtained on the grade five science test with those obtained on the grade ten science test. The reviewer may compare results for the same content area and grade, within a school, between schools, or between a school and its district, its county, or the state within the same year or to previous years.

Finally, it is inappropriate to conduct any type of score comparisons (including raw score, percent correct, scale score, or performance level comparisons) between CST for Science and CMA for Science tests. The CMA for Science are designed for students with an IEP who meet eligibility criteria adopted by the SBE. The CMA for Science were created using an independent procedure for test development and test blueprints developed for students eligible to take the CMA for Science, using CMA for Science blueprints. Performance levels specific to the CMA for Science were established. Therefore, comparison between CMA for Science and CST for Science results is discouraged.

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; 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 CMA for Science administered in spring 2014–15. The technical report contains nine additional chapters as follows:

- Chapter 2 presents a conceptual overview of processes involved in a testing cycle for a CMA 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 references to various chapters that detail the processes briefly discussed in this chapter.
- Chapter 3 describes the procedures followed during the development of valid CMA for Science items before the 2014–15 administration—in 2014–15, intact test forms from previous test administrations were used and there was no new item development. The chapter also explains the process of field-testing new items and the review of items by contractors and content experts.

- Chapter 4 details the content and psychometric criteria that guided the construction of the CMA for Science forms reused in 2014–15.
- Chapter 5 presents the processes involved in the actual administration of the 2014–15 CMA for Science with an emphasis on efforts made to ensure standardization of the tests. It also includes a detailed section that describes the procedures that were followed by ETS to ensure test security.
- Chapter 6 describes the standard-setting process previously conducted to establish cut scores for the CMA for Science.
- Chapter 7 details the types of scores and score reports that are produced at the end of each administration of the CMA for Science.
- Chapter 8 summarizes the results of the test- and item-level analyses performed during the 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 CMA for Science performance-level classifications, and the procedures designed to ensure the validity of CMA for Science score uses and interpretations. Also discussed in this chapter are item response theory, CMA for Science conversion tables, and the considerations and processes involved in pre-equating.
- Chapter 9 highlights the importance of controlling and maintaining the quality of the CMA for Science.
- Chapter 10 presents historical comparisons of various item- and test-level results for the past three years and for the base year, which vary according to test.

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

References

- California Code of Regulations, Title 5, Education, Division 1, Chapter 2, Subchapter 3.75, Article 2, §§ 853.5 and 855.* Retrieved from <http://www.cde.ca.gov/re/lr/rr/caaspp.asp>
- California Department of Education. (2013). *STAR Program information packet for school district and school staff* (p. 15). Sacramento, CA.
- California Department of Education, EdSource, & the Fiscal Crisis Management Assistance Team. (2014). *Fiscal, demographic, and performance data on California's K–12 schools*. Sacramento, CA: Ed-Data. http://www.ed-data.k12.ca.us/App_Resx/EdDataClassic/fsTwoPanel.aspx?#!bottom=/_layouts/EdDataClassic/profile.asp?Tab=1&level=04&report Number=16
- California Department of Education. (2015, May). *State Board of Education Responsibilities*. Retrieved from <http://www.cde.ca.gov/be/ms/po/sberesponsibilities.asp>
- California Department of Education. (2016, January). *Organization*. Retrieved from <http://www.cde.ca.gov/be/ms/po/sberesponsibilities.asp>

Chapter 2: An Overview of CMA for Science Processes

This chapter provides an overview of the processes involved in a typical test development and administration cycle for the California Modified Assessment (CMA) for Science. Also described are the specifications maintained by Educational Testing Service (ETS) to implement each of those processes. In 2014–15, three CMA for Science were administered; intact forms—i.e., test forms from previous administrations—from different years were used. Table 8.4 on page 87 lists the forms and their original year of use. All three tests are considered pre-equated.

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

Item Development

Item Formats

All CMA for Science contain three-option multiple-choice items.

Item Specifications

There were no new items developed in 2014–15. Prior to the 2012–13 administration, the CMA for Science items were developed to measure California content standards adopted by the state in 1998 and designed to conform to principles of item writing defined by ETS (ETS, 2002). ETS maintained and updated an item specifications document, otherwise known as “item writer guidelines,” for each CMA and used an item utilization plan to guide the development of the items for each content area. Item writing emphasis was determined in consultation with the California Department of Education (CDE).

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

The items selected for the CMA for Science underwent an extensive item review process that is designed to provide the best standards-based tests possible. Details about the item specifications, the item review process, and the item utilization plan are presented in Chapter 3, starting on page 27.

Item Banking

Before newly developed items 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 30; and the Statewide Pupil Assessment Review panel, described in Chapter 3, starting on page 33.

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

subsequently field-tested to obtain information about item performance and item statistics that could be used to assemble operational forms.

The CDE then reviewed those items with their statistical data flagged to determine whether they should be used operationally (see page 34 for more information about the CDE's data review). Any additional updates to item content and statistics were based on data collected from the operational use of the items. However, only the latest content of the item is retained in the bank at any time, along with the administration data from every administration that has included the item.

Further details on item banking are presented on page 34 in Chapter 3.

Item Refresh Rate

Prior to use intact forms in the 2014–15 administration, the item utilization plan required that each year, 30 percent of items on an operational form were refreshed (replaced); these items remained in the item bank for future use. Because the forms were reused, there were no items refreshed in the 2014–15 administration.

Test Assembly

Test Length

The number of operational items in each CMA for Science varies by content area and grade. There are 48 operational items on the CMA for Science in grade five, 54 operational items on the CMA for Science in grade eight, and 60 operational items on the CMA for Life Science in grade ten. The considerations used in deciding the test length are described on page 37 in Chapter 4.

Each CMA for Science also includes a various number of field-test items in addition to the operational items. Although there was no new item development for the 2014–15 administration, the field-test items were included as part of the reused forms but did not contribute to students' scores. The total number of items, including field-test items, in each CMA for Science and the estimated time to complete a test form are presented in Appendix 2.A on page 19.

Test Blueprints

ETS selected all CMA for Science items to conform to the State Board of Education-approved California content standards and test blueprints. The test blueprints for the CMA for Science can be found on the CDE California Assessment of Student Performance and Progress (CAASPP) Science Assessments Web page at <http://www.cde.ca.gov/ta/tg/ca/caasppscience.asp>.

The test blueprints specify the number of items at the individual standard level. In previous administrations, scores for the CMA for Science items were grouped into subcontent areas referred to as "reporting clusters." For each CMA for Science reporting cluster, the percentage of questions correctly answered was reported on a student's score report. Although only the total test scale scores are reported and cluster scores are no longer included in the score report in 2014–15, a description of the CMA for Science reporting clusters and the standards that comprise each cluster are provided in Appendix 2.B, which starts on page 20.

Content Rules and Item Selection

Intact test forms from different years were used during the 2014–15 administration. (See Table 8.4 on page 87 for administration years.) In a typical development cycle prior to using

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

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

Another consideration was the difficulty of each item. Test developers strived to ensure that there were some easy and some hard items and that there were a number of items in the middle range of difficulty. The detailed rules are presented in Chapter 4, which begins on page 37.

Psychometric Criteria

The staff assessed the projected test characteristics during the preliminary review of the assembled forms. The statistical targets used to develop the intact forms for 2014–15 administration and the projected characteristics of the forms are presented starting from page 38 in Chapter 4.

The items in test forms were organized and sequenced differently according to the requirements of the content area. Further details on the arrangement of items during test assembly are also described on page 40 in Chapter 4.

All the forms in the 2014–15 CMA for Science administration were used in prior operational test administrations. See Table 8.4 on page 87 for the list containing the administration in which each CMA for Science was originally administered.

Test Administration

It is of utmost priority to administer the CMA 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 CMA 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 46.

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 46.

- Test development
- Item and data review
- Item banking

- Transfer of forms and items to the CDE
- Security of electronic files using a firewall
- Printing and publishing
- Test administration
- Test delivery
- Processing and scoring
- Data management
- Transfer of scores via secure data exchange
- Statistical analysis
- Reporting and posting results
- Student confidentiality
- Student test results

Procedures to Maintain Standardization

The CMA 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 CMA for Science, as described in this section.

Test Administrators

The CMA 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 local educational agencies (LEAs) who are central to the processes include LEA CAASPP coordinators, CAASPP test site coordinators, test administrators, proctors, and scribes. The responsibilities of each of the staff members are included in the *CAASPP Paper-Pencil Testing Test Administration Manual* (CDE, 2015a); see page 52 in Chapter 5 for more information.

Test Directions

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

Directions for Administration (DFAs)—Manuals used by test administrators to administer the CMA for Science to students to be followed exactly so that all students have an equal opportunity to demonstrate their academic achievement (See page 52 in Chapter 5 for more information.)

CAASPP Paper-Pencil Testing Test Administration Manual—Test administration procedures for LEA CAASPP coordinators and CAASPP test site coordinators (See page 52 in Chapter 5 for more information.)

Test Operations Management System (TOMS) manuals—Instructions for the Web-based modules that allow LEA CAASPP coordinators to set up test administrations, assign tests, and assign student test settings; every module has its own user manual with detailed instructions on how to use TOMS (See page 53 in Chapter 5 for more information.)

Universal Tools, Designated Supports, and Accommodations

All public school students participate in the CAASPP Program, including students with disabilities and English learners. Most students with individualized education programs (IEPs) and most English learners (ELs) take the CMA for Science under standard conditions. However, some students with IEPs and some ELs may need assistance when taking the CMA for Science. This assistance takes the form of universal tools, designated supports, and accommodations. All students in these categories may have test administration directions simplified or clarified.

Appendix 2.C on page 21 presents an adaptation of Matrix One of the “Universal Tools, Designated Supports, and Accommodations for the California Assessment of Student Performance and Progress” (CDE, 2015b). Part 2 of Matrix One, found in Table 2.C.1, includes the non-embedded supports; Appendix 2.C shows only the supports that were allowed for the CMA for Science in 2014–15 and were mapped to CMA for Science answer documents so had data that could be collected. Table 2.C.1 also shows the answer document options in section A3 that are reported in Appendix 2.D and were defined but did not map to a specific universal tool, designated support, or accommodation, as well as the reported answer document options in section A4 that are unmapped.

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

Non-embedded Supports

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

Individualized Aids (now “Unlisted Resources”)

Individualized aids are resources that fundamentally change what is being tested and may interfere with the construct being measured. All individualized aids must be identified, approved, and listed in the student’s IEP or Section 504 plan. Individualized aids, when approved, are marked as option Y in Appendix 2.D. (Note that individualized aids have been subsequently renamed to “unlisted resources.”)

Special Services Summaries

The percentage of students using various universal tools, designated supports, and accommodations during the 2014–15 administration of the CMA for Science is presented in Appendix 2.D, which starts on page 22. The data are organized into two sections within each table. The first section presents the percentages of students using each accommodation in the total testing population. The second section presents the results for students in various categories based on the following levels of English-language fluency:

- **English only (EO)**—A student for whom there is a report of English as the primary language (i.e., language first learned, most frequently used at home, or most frequently spoken by the parents or adults in the home) on the “Home Language Survey”
- **Initially fluent English proficient (I-FEP)**—A student whose primary language is a language other than English who initially met the LEA criteria for determining proficiency in English
- **English learner (EL)**—A student who first learned or has a home language other than English who was determined to lack sufficient fluency in English on the basis of state oral language (K–12) and literacy (3–12) assessments to succeed in the school’s regular instructional program (For students tested for initial classification prior to May 2001, this determination is made on the basis of the state-approved instrument the LEA was using. For students tested after May 2001, use the California English Language Development Test [CELDT] results.)
- **Reclassified fluent English proficient (R-FEP)**—A student whose primary language is a language other than English who was reclassified from English learner to fluent-English proficient

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

Scores

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

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

Detailed descriptions of CMA for Science scores are found in Chapter 7, which starts on page 61.

Aggregation Procedures

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

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

1. Met attemptedness criteria
2. Did not have a parental exemption
3. Did not miss any part of the test due to illness or medical emergency
4. Did not test out of level (grade inappropriate)

Individual Scores

Table 7.1 and Table 7.2, starting on page 63 in Chapter 7, provide summary statistics for individual scores aggregated at the state level, describing overall student performance on each CMA for Science. Included in the tables are the means and standard deviations of student scores expressed in terms of both raw scores and scale scores; the raw score means and standard deviations expressed as percentages of the total raw score points in each test; and the percentages of students in each performance level.

Statistics summarizing CMA for Science student performance by grade are provided in Table 7.A.1 on page 69 in Appendix 7.A.

Demographic Subgroup Scores

In Table 7.B.1 through Table 7.B.3, starting on page 70 in Appendix 7.B, students are grouped by demographic characteristics, including gender, ethnicity, English-language fluency, economic status, and primary disability. The tables show the numbers of students with valid scores in each group, scale score means and standard deviations, and percent in a performance level, as well as percent correct for each reporting cluster for each demographic group. Table 7.3 on page 64 provides definitions for the demographic groups included in the tables.

Equating

Post-Equating

In the years when the new forms were developed prior to the 2012–13 administration, the CMA for Science were equated to a reference form using a linking items nonequivalent groups data collection design and post-equating methods based on item response theory (IRT) (Hambleton & Swaminathan, 1985). The “base” or “reference” calibrations for the CMA for Science were established by calibrating samples of item response data from a specific administration, through which item parameter estimates for the items in the reused forms were placed on the reference scale using a set of linking items selected from the previous year. Doing so established a scale to which subsequent item calibrations could be linked. For science in grade five, grade eight, and Life Science in grade ten, the reference scales were established in 2009, 2010, and 2011 respectively.

The procedure used for post-equating the CMA for Science involved three steps: item calibration, item scaling, and production of scoring tables. Each of those steps, as described below, was applied to all of the grade-level CMA for Science during the tests’ original years of administration. Results were not post-equated for the 2014–15 administration.

Pre-Equating

During the 2014–15 administration, because all the test forms were used in previous operational administrations, pre-equating was conducted prior to administration of the tests. Based on the sample invariant property of IRT, all the item parameter estimates were placed on the reference scale in their previous administrations through the post-equating procedure described previously. Item parameters derived in such a manner can be used to create raw-score-to-scale-score conversion tables prior to test administration. Neither calibration nor scaling was implemented in the pre-equating process.

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

Table 8.4 on page 87 shows the years the forms were introduced for each test.

Calibration

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

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

Scaling

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

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

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

where,

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

n_g is the number of intervals/groups,

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

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

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

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

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

Scoring Table Production

Once the new item calibrations for each test were transformed to the base scale after items' initial administration, IRT procedures were used to transform the new form number-correct scores (raw scores) to their corresponding ability (theta). The ability estimates were then transformed to scale scores through linear transformation.

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

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

where,

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

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

For each integer score ξ_n on the form after its original use, the procedure was used to first solve for the corresponding ability estimate using equation 2.2. The ability estimates were then expressed in the reporting scale metric by applying linear transformation with the appropriate slope and intercept, using equation 2.4:

$$\text{ScaleScore} = \text{Intercept} + \text{Slope} \times \theta \quad (2.4)$$

where,

θ represents student ability.

The slope and intercept for each CMA for Science were developed from the base forms using equations 2.5 and 2.6 because the basic and proficiency cut scores were required to be equal to 300 and 350, respectively.

$$\text{Slope} = \frac{350 - 300}{\theta_{\text{proficient}} - \theta_{\text{basic}}} \quad (2.5)$$

$$\text{Intercept} = 350 - \theta_{\text{proficient}} \times \left(\frac{350 - 300}{\theta_{\text{proficient}} - \theta_{\text{basic}}} \right) \quad (2.6)$$

where,

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

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

Complete raw-score-to-scale-score conversion tables for the CMA for Science are presented in Table 8.C.1 through Table 8.C.3 in Appendix 8.C, starting on page 113. The raw scores and corresponding rounded, transformed scale scores are also listed in those tables. Data used are from the forms' original administration.

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

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

Table 2.1 Scale-Score Ranges for Performance Levels

CMA	Far Below Basic	Below Basic	Basic	Proficient	Advanced
Grade 5 Science	150 – 242	243 – 299	300 – 349	350 – 400	401 – 600
Grade 8 Science	150 – 263	264 – 299	300 – 349	350 – 405	406 – 600
Grade 10 Life Science	150 – 250	251 – 299	300 – 349	350 – 409	410 – 600

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). *Universal Tools, Designated Supports, and Accommodations for the California Assessment of Student Performance and Progress*. Sacramento, CA. <http://www.cde.ca.gov/ta/tg/ai/caasppmatrix1.asp>

Educational Testing Service. (2014). *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.

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

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

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

California Modified Assessment	Grade 5		Grade 8		Grade 10	
	Total No. of Items	Time	Total No. of Items	Time	Total No. of Items	Time
Science		120		135		150
Part 1	57	40	63	45	66	50
Part 2		40		45		50
Part 3		40		45		50

Appendix 2.B—Reporting Clusters for Science

Science Modified Standards Assessment (Grade Five)

Physical Sciences

Life Sciences

Earth Sciences

Science Modified Standards Assessment (Grade Eight)

Motion

Matter

Earth Science

Investigation and Experimentation

Science Modified Standards Assessment (Grade Ten)

Cell Biology and Genetics

Evolution and Ecology

Physiology

Investigation and Experimentation

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

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

Option	(U) Universal Tool (D) Designated Support (A) Accommodation	
Answer Document Section A3—Accommodations and Modifications		
B	Pupil marks in paper-pencil test booklet (other than responses including highlighting)	U
C	Scribe	A
G	Braille (paper-pencil tests)	A
H	Large-print versions of a paper-pencil test (as available)	A
J	Breaks (Tested over more than one day)	U
K	Breaks (Supervised breaks within a section of the test)	U
L	Administration of the test to the pupil at the most beneficial time of day	D
M	Separate setting	D
O	American Sign Language	A
X	Abacus	A
Y	Individualized aid	–
Z	Read aloud	A
Option	(U) Universal Tool (D) Designated Support (A) Accommodation	
Answer Document Section A4—English Learner (EL) Test Variations		
A	Translated Test Directions	D
B	Additional supervised breaks within a testing day or following each section within a test part provided that the test section is completed within a testing day. A test section is identified by a “STOP” at the end of it.	Unmapped
C	English learners (ELs) may have the opportunity to be tested separately with other ELs provided that the student is directly supervised by an employee of the school who has signed the test security affidavit and the student has been provided such a flexible setting as part of his/her regular instruction or assessment.	Unmapped
D	Translations (Glossary) (previously known as “Access to translation glossaries/word lists (English-to-primary language). Glossaries/Word lists shall not include definitions or formulas.)	D

Universal Tools (U) Are available for all pupils. Pupils may turn the support(s) on/off when embedded as part of the technology platform for the computer-administered CAASPP tests or may choose to use it/them when provided as part of a paper-pencil test.

Designated Supports (D) Are features that are available for use by any pupil for whom the need has been indicated prior to the assessment, by an educator or group of educators.

Accommodations (A) For the CAASPP assessment system, eligible pupils shall be permitted to take the tests with accommodations if specified in the pupil’s individualized educational program (IEP) or Section 504 plan.

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

Appendix 2.D—Special Service Summary Tables

Notes:

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

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

Answer Document Option	Grade 5 Number	Grade 5 Pct. of Total	Grade 8 Number	Grade 8 Pct. of Total	Grade 10 Life Sci. Number	Grade 10 Pct. of Total
B: Marked in test booklet	N/A	N/A	N/A	N/A	N/A	N/A
C: Scribe	465	1.98%	150	0.77%	52	0.53%
G: Braille	0	0.00%	9	0.05%	2	0.02%
H: Large-print versions of a paper-pencil test	68	0.29%	47	0.24%	15	0.15%
J: Breaks (Tested over more than one day)	N/A	N/A	N/A	N/A	N/A	N/A
K: Breaks (Had supervised breaks)	N/A	N/A	N/A	N/A	N/A	N/A
L: Most beneficial time of day	2,673	11.36%	1,281	6.57%	360	3.67%
M: Separate setting	63	0.27%	58	0.30%	23	0.23%
O: American Sign Language	45	0.19%	28	0.14%	40	0.41%
X: Abacus	18	0.08%	55	0.28%	17	0.17%
Y: Individualized aid	66	0.28%	36	0.18%	11	0.11%
Z: Read aloud	7,826	33.25%	2,729	14.00%	492	5.02%
Univ. tool, desig. sup., and acc. is in Section 504 plan	40	0.17%	3	0.02%	2	0.02%
Univ. tool, desig. sup., and acc. is in IEP	8,695	36.94%	3,644	18.69%	896	9.14%
English Learner Test Variation A	0	0.00%	0	0.00%	0	0.00%
English Learner Test Variation B	N/A	N/A	N/A	N/A	N/A	N/A
English Learner Test Variation C	N/A	N/A	N/A	N/A	N/A	N/A
English Learner Test Variation D	0	0.00%	0	0.00%	0	0.00%
Any Universal tool, desig. support, and acc or Additional univ. tool, design. sup. for EL	9,208	39.12%	3,807	19.53%	944	9.63%
No Universal tool, desig. support, and acc or Additional univ. tool, design. sup. for EL	14,327	60.88%	15,690	80.47%	8,862	90.37%

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

Answer Document Option	Grade 5		Grade 8		Grade 10	
	Grade 5 Number	Pct. of Total	Grade 8 Number	Pct. of Total	Grade 10 Life Sci. Number	Grade 10 Pct. of Total
B: Marked in test booklet	N/A	N/A	N/A	N/A	N/A	N/A
C: Scribe	288	2.31%	78	0.77%	30	0.60%
G: Braille test	0	0.00%	5	0.05%	1	0.02%
H: Large-print versions of a paper-pencil test	34	0.27%	18	0.18%	10	0.20%
J: Breaks (Tested over more than one day)	N/A	N/A	N/A	N/A	N/A	N/A
K: Breaks (Had supervised breaks)	N/A	N/A	N/A	N/A	N/A	N/A
L: Most beneficial time of day	1,418	11.40%	711	7.02%	190	3.79%
M: Separate setting	30	0.24%	33	0.33%	8	0.16%
O: American Sign Language	28	0.23%	16	0.16%	29	0.58%
X: Abacus	10	0.08%	34	0.34%	10	0.20%
Y: Individualized aid	40	0.32%	17	0.17%	9	0.18%
Z: Read aloud	3,907	31.40%	1,334	13.16%	258	5.15%
Univ. tool, desig. sup., and acc. is in Section 504 plan	32	0.26%	3	0.03%	2	0.04%
Univ. tool, desig. sup., and acc. is in IEP	4,401	35.37%	1,855	18.30%	465	9.29%
English Learner Test Variation A	0	0.00%	0	0.00%	0	0.00%
English Learner Test Variation B	N/A	N/A	N/A	N/A	N/A	N/A
English Learner Test Variation C	N/A	N/A	N/A	N/A	N/A	N/A
English Learner Test Variation D	0	0.00%	0	0.00%	0	0.00%
Any Universal tool, desig. support, and acc or additional univ. tool, design. sup. for EL	4,692	37.71%	1,952	19.26%	499	9.96%
No Universal tool, desig. support, and acc or additional univ. tool, design. sup. for EL	7,751	62.29%	8,183	80.74%	4,509	90.04%

Table 2.D.3 Special Service Summary for Science, Grades Five, Eight, and Ten (Life Science)—Initially Fluent English Proficient (I-FEP) Students

Answer Document Option	Grade 5 Number	Grade 5 Pct. of Total	Grade 8 Number	Grade 8 Pct. of Total	Grade 10 Life Sci. Number	Grade 10 Pct. of Total
B: Marked in test booklet	N/A	N/A	N/A	N/A	N/A	N/A
C: Scribe	7	3.41%	3	1.35%	1	0.50%
G: Braille	0	0.00%	1	0.45%	0	0.00%
H: Large-print versions of a paper-pencil test	1	0.49%	1	0.45%	0	0.00%
J: Breaks (Tested over more than one day)	N/A	N/A	N/A	N/A	N/A	N/A
K: Breaks (Had supervised breaks)	N/A	N/A	N/A	N/A	N/A	N/A
L: Most beneficial time of day	15	7.32%	14	6.31%	3	1.49%
M: Separate setting	1	0.49%	1	0.45%	0	0.00%
O: American Sign Language	1	0.49%	1	0.45%	0	0.00%
X: Abacus	0	0.00%	0	0.00%	1	0.50%
Y: Individualized aid	1	0.49%	0	0.00%	1	0.50%
Z: Examiner read test questions aloud	69	33.66%	30	13.51%	14	6.93%
Univ. tool, desig. sup., and acc. is in Section 504 plan	2	0.98%	0	0.00%	0	0.00%
Univ. tool, desig. sup., and acc. is in IEP	75	36.59%	39	17.57%	17	8.42%
English Learner Test Variation A	0	0.00%	0	0.00%	0	0.00%
English Learner Test Variation B	N/A	N/A	N/A	N/A	N/A	N/A
English Learner Test Variation C	N/A	N/A	N/A	N/A	N/A	N/A
English Learner Test Variation D	0	0.00%	0	0.00%	0	0.00%
<i>Any</i> Universal tool, desig. support, and acc. or additional univ. tool, design. sup. for EL	79	38.54%	41	18.47%	19	9.41%
<i>No</i> Universal tool, desig. support, and acc. or additional univ. tool, design. sup. for EL	126	61.46%	181	81.53%	183	90.59%

**Table 2.D.4 Special Service Summary for Science, Grades Five, Eight, and Ten (Life Science)—
English Learner (EL) Students**

Answer Document Option	Grade 5 Number	Grade 5 Pct. of Total	Grade 8 Number	Grade 8 Pct. of Total	Grade 10 Life Sci. Number	Grade 10 Pct. of Total
B: Marked in test booklet	N/A	N/A	N/A	N/A	N/A	N/A
C: Scribe	162	1.55%	64	0.84%	19	0.52%
G: Braille	0	0.00%	3	0.04%	1	0.03%
H: Large-print versions of a paper-pencil test	32	0.31%	19	0.25%	5	0.14%
J: Breaks (Tested over more than one day)	N/A	N/A	N/A	N/A	N/A	N/A
K: Breaks (Had supervised breaks)	N/A	N/A	N/A	N/A	N/A	N/A
L: Most beneficial time of day	1,196	11.42%	489	6.41%	156	4.24%
M: Separate setting	31	0.30%	20	0.26%	14	0.38%
O: American Sign Language	14	0.13%	11	0.14%	10	0.27%
X: Abacus	8	0.08%	18	0.24%	6	0.16%
Y: Individualized aid	23	0.22%	14	0.18%	1	0.03%
Z: Examiner read test questions aloud	3,704	35.37%	1,169	15.32%	186	5.06%
Univ. tool, desig. sup., and acc. is in Section 504 plan	6	0.06%	0	0.00%	0	0.00%
Univ. tool, desig. sup., and acc. is in IEP	4,070	38.86%	1,497	19.61%	367	9.99%
English Learner Test Variation A	0	0.00%	0	0.00%	0	0.00%
English Learner Test Variation B	N/A	N/A	N/A	N/A	N/A	N/A
English Learner Test Variation C	N/A	N/A	N/A	N/A	N/A	N/A
English Learner Test Variation D	0	0.00%	0	0.00%	0	0.00%
Any Universal tool, desig. support, and acc. or additional univ. tool, design. sup. for EL	4,272	40.79%	1,556	20.39%	378	10.29%
No Universal tool, desig. support, and acc. or additional univ. tool, design. sup. for EL	6,201	59.21%	6,077	79.61%	3,297	89.71%

**Table 2.D.5 Special Service Summary for Science, Grades Five, Eight, and Ten (Life Science)—
Reclassified Fluent English Proficient (R-FEP) Students**

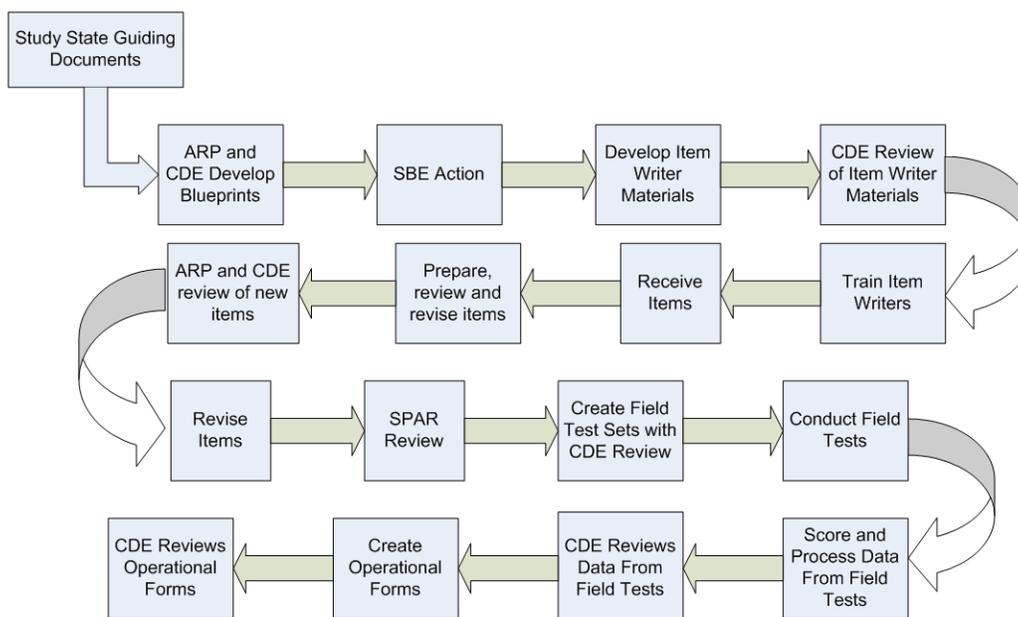
Answer Document Option	Grade 5 Number	Grade 5 Pct. of Total	Grade 8 Number	Grade 8 Pct. of Total	Grade 10 Life Sci. Number	Grade 10 Pct. of Total
B: Marked in test booklet	N/A	N/A	N/A	N/A	N/A	N/A
C: Scribe	8	1.98%	5	0.33%	2	0.22%
G: Braille	0	0.00%	0	0.00%	0	0.00%
H: Large-print versions of a paper-pencil test	1	0.25%	9	0.60%	0	0.00%
J: Breaks (Tested over more than one day)	N/A	N/A	N/A	N/A	N/A	N/A
K: Breaks (Had supervised breaks)	N/A	N/A	N/A	N/A	N/A	N/A
L: Most beneficial time of day	43	10.64%	67	4.47%	11	1.21%
M: Separate setting	1	0.25%	4	0.27%	1	0.11%
O: American Sign Language	2	0.50%	0	0.00%	1	0.11%
X: Abacus	0	0.00%	3	0.20%	0	0.00%
Y: Individualized aid	2	0.50%	5	0.33%	0	0.00%
Z: Examiner read test questions aloud	141	34.90%	196	13.08%	33	3.63%
Univ. tool, desig. sup., and acc. is in Section 504 plan	0	0.00%	0	0.00%	0	0.00%
Univ. tool, desig. sup., and acc. is in IEP	148	36.63%	253	16.88%	47	5.16%
English Learner Test Variation A	0	0.00%	0	0.00%	0	0.00%
English Learner Test Variation B	N/A	N/A	N/A	N/A	N/A	N/A
English Learner Test Variation C	N/A	N/A	N/A	N/A	N/A	N/A
English Learner Test Variation D	0	0.00%	0	0.00%	0	0.00%
Any Universal tool, desig. support, and acc. or additional univ. tool, design. sup. for EL	160	39.60%	258	17.21%	47	5.16%
No Universal tool, desig. support, and acc. or additional univ. tool, design. sup. for EL	244	60.40%	1,241	82.79%	863	94.84%

Chapter 3: Item Development

Intact test forms from previous test administrations from different years were reused during the 2014–15 administration. Using an intact form permits the original score conversion tables from the previous administration to be used to look up student scores and performance levels. There was no new item development for the 2014–15 forms.

The California Modified Assessment (CMA) for Science items were developed to measure California’s content standards and designed to conform to principles of item writing defined by Educational Testing Service (ETS) (ETS, 2002). Each CMA for Science item on the intact forms used in 2014–15, went through a comprehensive development cycle as is described in Figure 3.1 below.

Figure 3.1 The ETS Item Development Process for the CAASPP System



Rules for Item Development

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

Item Specifications

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

- A full statement of each academic content standard, as defined by the State Board of Education (SBE) in 1998 (CDE, 2009)
- A description of each content strand
- The expected depth of knowledge (DOK) measured by items written for each standard (coded as 1, 2, 3, or 4; items assigned a DOK of 1 are the least cognitively complex,

items assigned a DOK of 3 are the most cognitively complex, and the code of 4 would apply only to some writing tasks)

- The homogeneity of the construct measured by each standard
- A description of the kinds of item stems appropriate for multiple-choice items used to assess each standard
- A description of the kinds of distractors that are appropriate for multiple-choice items assessing each standard
- A description of appropriate data representations (such as charts, tables, graphs, or other illustrations) for mathematics and science items
- The content limits for the standard (such as one or two variables, maximum place values of numbers) for mathematics and science items
- A description of appropriate reading passages, where applicable, for English–Language Arts (ELA) items
- A description of specific kinds of items to be avoided, if any (for example, items with any negative expressions in the stem, e.g., “Which of the following is NOT. . .”)

Expected Item Ratio

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

The item utilization plan assumed that after the first two operational administrations, 30 percent of items on an operational form would be refreshed (replaced) each year; these items would remain in the item bank for future use. The plan also declared that an additional five percent of the operational items were likely to become unusable because of normal attrition and noted a need to focus development on “critical” standards, which are those that were difficult to measure well or for which there were few usable items.

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

For the 2014–15 CMA for Science administration, field-test items were repeated as a part of the reuse of the intact form.

Selection of Item Writers

Criteria for Selecting Item Writers

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

- Possession of a Bachelor’s degree in the relevant content area or in the field of education with special focus on a particular content of interest; an advanced degree in the relevant content area is desirable

- Previous experience in writing items for standards-based assessments, including knowledge of the many considerations that are important when developing items to match state-specific standards
- Previous experience in writing items in the content areas covered by CMA grades and/or courses
- Familiarity, understanding, and support of the California content standards
- Current or previous teaching experience in California, when possible

Item Review Process

The items selected for each CMA for Science underwent an extensive item review process that was designed to provide the best standards-based tests possible. This section summarizes the various reviews performed that ensure the quality of the CMA for Science items and test forms—currently being reused—at the time the items and forms were developed. See Table 8.4 on page 87 for the dates of the previous administrations.

Contractor Review

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

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

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

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

1. Internal Content Review

Test items and materials underwent two reviews by the content-area assessment specialists. These assessment specialists made sure that the test items and related materials were in compliance with ETS’s written guidelines for clarity, style, accuracy, and appropriateness for California students as well as in compliance with the approved item specifications. Assessment specialists reviewed each item in terms of the following characteristics:

- Relevance of each item to the purpose of the test
- Match of each item to the item specifications, including DOK
- Match of each item to the principles of quality item writing
- Match of each item to the identified standard or standards
- Difficulty of the item
- Accuracy of the content of the item

- Readability of the item or passage
- Grade-level appropriateness of the item
- Appropriateness of any illustrations, graphs, or figures

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

2. Internal Editorial Review

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

3. Internal Sensitivity Review

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

The review process promoted a general awareness of and responsiveness to the following:

- Cultural diversity
- Diversity of background, cultural tradition, and viewpoints to be found in the test-taking populations
- Changing roles and attitudes toward various groups
- Role of language in setting and changing attitudes toward various groups
- Contributions of diverse groups (including ethnic and minority groups, individuals with disabilities, and women) to the history and culture of the United States and the achievements of individuals within these groups
- Item accessibility for English-language learners

Content Expert Reviews

Assessment Review Panels

ETS was responsible for working with ARPs as items were developed for the CMA for Science. The ARPs are advisory panels to the CDE and ETS and provided guidance on matters related to item development for the CMA for Science. The ARPs were responsible for reviewing all newly developed items for alignment to the California content standards; these tests use the content standards for science adopted by the SBE in 1998. The ARPs also reviewed the items for accuracy of content, clarity of phrasing, and quality. In their examination of test items, the ARPs could raise concerns related to age/grade appropriateness and gender, racial, ethnic, and/or socioeconomic bias.

Composition of ARPs

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

- Three or more years of general teaching experience in grades kindergarten through twelve and in the relevant content areas (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 in ELA, mathematics, or science that are current at the time;
- Special education credential;
- Experience with more than one type of disability; and
- Three to five years of experience as a teacher or school administrator with a special education credential.

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

- Three or more years of experience as a school administrator, LEA/county content/program specialist, or university instructor in a grade-specific area or area related to science;
- Bachelor’s or higher degree in a grade-specific or content area related to science; and
- Knowledge of and experience with the California content standards in ELA, mathematics, or science that are current at the time.

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

ARP members were recruited through an application process. Recommendations were solicited from LEAs and county offices of education as well as from CDE and SBE staff. Applications were reviewed by the ETS assessment directors, who confirmed that the applicant’s qualifications met the specified criteria. Applications that met the criteria were forwarded to CDE and SBE staff for further review and agreement on ARP membership.

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

ARP Meetings for Review of CMA for Science Items

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

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

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

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

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

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

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

Does the item:

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

Is the stimulus, if any, for the item:

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

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

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

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

Statewide Pupil Assessment Review Panel

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

Field Testing

The primary purposes of field testing are to obtain information about item performance and to obtain statistics that can be used to assemble operational forms. However, because the intact forms were used with the original field-test items for the 2014–15 CAASPP administration, data were not analyzed for current field-test items.

Stand-alone Field Testing

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

CMA for Science stand-alone field testing for each new test occurred in the fall before the test became operational in the following spring.

The stand-alone field-testing timeline for the CMA for Science is presented in Table 3.1.

Table 3.1 Stand-alone Field-testing Timeline for the CMA for Science

CMA	Field-test Year
Grade 5 Science	2007
Grade 8 Science	2008
Grade 10 Life Science	2009

Embedded Field-test Items

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

Such field testing involves distributing the items being field-tested within an operational test form. Different forms contain the same core set of operational items and different sets of field-test items. For the 2014–15 administration, the original field-test items remained in their original positions in the intact forms. Data were not analyzed for field-test items. The numbers of embedded field-test items for the CMA for Science are not presented in this report, because for the 2014–15 administration, field-test items were repeated as a part of the intact forms and there was no new item development.

Allocation of Students to Forms

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

CDE Data Review

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

Item Banking

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

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

All unavailable items were marked with an availability indicator of “Unavailable,” a reason for rejection as described above, and cause alerts so they are not inadvertently included on

subsequent test forms. Statuses and availability were updated programmatically as items were presented for review, accepted or rejected, placed on a form for field-testing, presented for statistical review, 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 for Science, CMA for Science, California Alternate Performance Assessment for Science, and Standards-based Tests in Spanish). CAHSEE and CAASPP are currently consolidated in the California item banking system. ETS works with the CDE to obtain the data for assessments such as the CELDT, under contract with other vendors for inclusion into the item bank. ETS provides the item banking application using the LAN architecture and the relational database management system, SQL 2008, already deployed. ETS provides updated versions of the item bank to the CDE on an ongoing basis and works with the CDE to determine the optimum process if a change in databases is desired.

References

- California Department of Education. (2009). *California content standards*. Sacramento, CA. Retrieved from <http://www.cde.ca.gov/be/st/ss/>
- Educational Testing Service (2002). *ETS standards for quality and fairness*. Princeton, NJ: Author.
- Patrick, R., & Way, D. (March, 2008). *Field testing and equating designs for state educational assessments*. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.
- Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R.L. Brennan (Ed.), *Educational measurement* (4th ed.). Westport, CT: American Council on Education and Praeger Publishers.

Chapter 4: Test Assembly

The California Modified Assessment (CMA) 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 CMA for Science, the content standards and desired psychometric attributes were used as the basis for assembling the test forms.

Test Length

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

The number of operational items on each CMA for Science varies across grades. There are 48 operational items on the CMA for Science in grade five. There are 54 operational items on the CMA for Science in grade eight. There are 60 operational items on the CMA for Life Science in grade ten.

The total number of items also varies. There are a total of 57 items on the CMA for Science in grade five. There are a total of 63 items on the CMA for Science in grade eight. There are a total of 66 items on the CMA for Life Science in grade ten.

In addition to operational items, a certain number of the items on each test are field-test items—nine on the grade-level tests in grades five and eight and six on the CMA for Life Science in grade ten. For more details on the distribution of items, see Appendix 2.A—CMA Items and Estimated Time Chart, starting on page 19.

Rules for Item Selection

Test Blueprint

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

Although the test blueprints called for the number of items at the individual standard level, scores for the CMA for Science items are grouped into subcontent areas (reporting clusters). A list of the CMA for Science reporting clusters by test is provided in Appendix 2.B—Reporting Clusters for Science, which starts on page 20.

Content Rules and Item Selection

Intact test forms from previous testing administrations from different years were used 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 grade and content area. First and foremost, they selected items that met the blueprint for that grade level and content area. Using an electronic item bank, assessment specialists began by identifying a number of linking items. These are items that appeared in a previous year's operational administration and were used to equate the administered test forms. Linking items were selected to proportionally represent the full blueprint. For example, if 25 percent of all of the

items in a test are in the first reporting cluster, then 25 percent of the linking items should come from that cluster. The selected linking items were also reviewed by psychometricians to ensure that specific psychometric criteria were met.

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

Another consideration was the difficulty of each item. Test developers strived to ensure that there were some easy and some hard items and that there were a number of items in the middle range of difficulty. If items did not meet all content and psychometric criteria, staff reviewed the other available items to determine if there were other selections that could improve the match of the test to all of the requirements. If such a match was not attainable, the content team worked in conjunction with psychometricians and the CDE to determine which combination of items would best serve the needs of the students taking the test. Chapter 3, starting on page 27, contains further information about this process.

Psychometric Criteria

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

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

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

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

Primary Statistical Targets

The total test targets, or primary statistical targets, used for assembling the intact CMA for Science forms used in the 2014–15 administration were the test information function (TIF) and an average point-biserial correlation.

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

The graphs for each total test are presented in Figure 4.A.1 on page 42 for the science tests. These curves present the target TIF and the projected TIF for the total test at each grade level.

Due to the unique characteristics of the Rasch IRT model, the information curve conditional on each ability level is determined by item difficulty (*b*-values) alone. In this case, the TIF would, therefore, suffice as the target for conditional test difficulty. Although additional item difficulty targets are not imperative when the target TIF is used for form construction, the

target mean and standard deviation (SD) of item difficulty consistent with the TIF were still provided to test development staff to help with the test construction process. The target *b*-value range approximates a minimum proportion-correct value (*p*-value) of 0.33 and a maximum *p*-value of 0.95 for each test.

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

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

Assembly Targets

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

Table 4.1 Statistical Targets for CMA for Science Test Assembly

CMA	Pt-Bis Mean	Pt-Bis Minimum	<i>b</i>-value Mean	<i>b</i>-value St. Dev.	<i>p</i>-value Minimum	<i>p</i>-value Maximum
Grade 5 Science	0.37	0.14	-0.43	0.74	0.33	0.95
Grade 8 Science	0.33	0.14	-0.46	0.50	0.33	0.95
Grade 10 Life Science	0.30	0.14	-0.22	0.50	0.33	0.95

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

Figure 4.B.1 through Figure 4.B.3, starting on page 43, present the target and projected information curves for the clusters in the administered tests.

Projected Psychometric Properties of the Assembled Tests

In the years when the new forms were developed prior to the 2012–13 administration, Educational Testing Service psychometricians performed a preliminary review of the technical characteristics of the assembled tests. The expected or projected performance of examinees and the overall score reliability were estimated using the item-level statistics available in the California item bank for the selected items. The test reliability was based on Gulliksen's formula (Gulliksen, 1987) for estimating test reliability (r_{xx}) from item *p*-values and item point-biserial correlations:

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

where,

K is the number of items in the test,

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

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

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

In addition, estimated test raw score means were calculated by summing the item p -values, and estimated test raw score standard deviations were calculated by summing the item reliability indices. Figure 4.A.1 on page 42 presents these summary values by content area and grade.

It should be noted that the projected reliabilities in Table 4.A.1 were based on item p -values and point-biserial correlations that, for some of the items, were based on external field-testing using samples of students that were not fully representative of the state. Chapter 8 presents item p -values, point-biserial correlations, and test reliability estimates based on the data from the 2014–15 CMA for Science administration.

Table 4.A.2 on page 42 shows the mean observed statistics of the items on each CMA for Science based on the item-level statistics from the year the form was previously administered. See Table 8.4 on page 87 for the dates of the original administrations. These values can be compared to the target values in Table 4.1.

Rules for Item Sequence and Layout

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

Reference

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

Appendix 4.A—Technical Characteristics

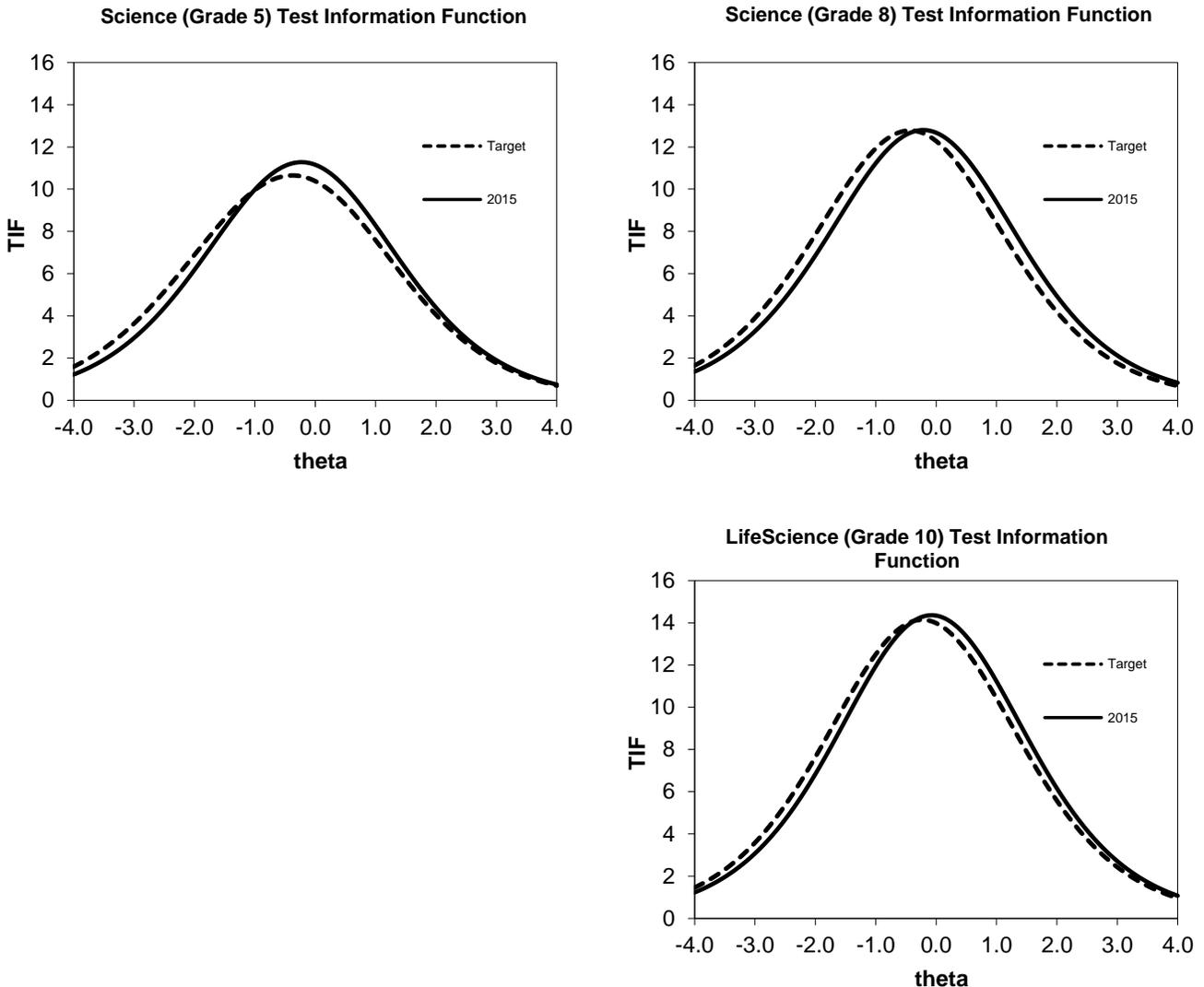
Table 4.A.1 Summary of 2015 CMA for Science Projected Raw Score Statistics

CMA	Number of Items	Mean Raw Score	Std. Dev. of Raw Scores	Reliability
Grade 5 Science	48	26.88	7.21	0.80
Grade 8 Science	54	29.63	7.53	0.79
Grade 10 Life Science	60	30.97	8.81	0.83

Table 4.A.2 Summary of 2015 CMA for Science Projected Item Statistics

CMA	Mean b	SD b	Mean ρ -value	Min ρ -value	Max ρ -value	Mean Point Biserial	Min Point Biserial
Grade 5 Science	-0.25	0.52	0.56	0.33	0.84	0.31	0.05
Grade 8 Science	-0.24	0.50	0.55	0.35	0.88	0.29	0.15
Grade 10 Life Science	-0.07	0.43	0.52	0.32	0.72	0.30	-0.09

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



Appendix 4.B—Cluster Targets

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

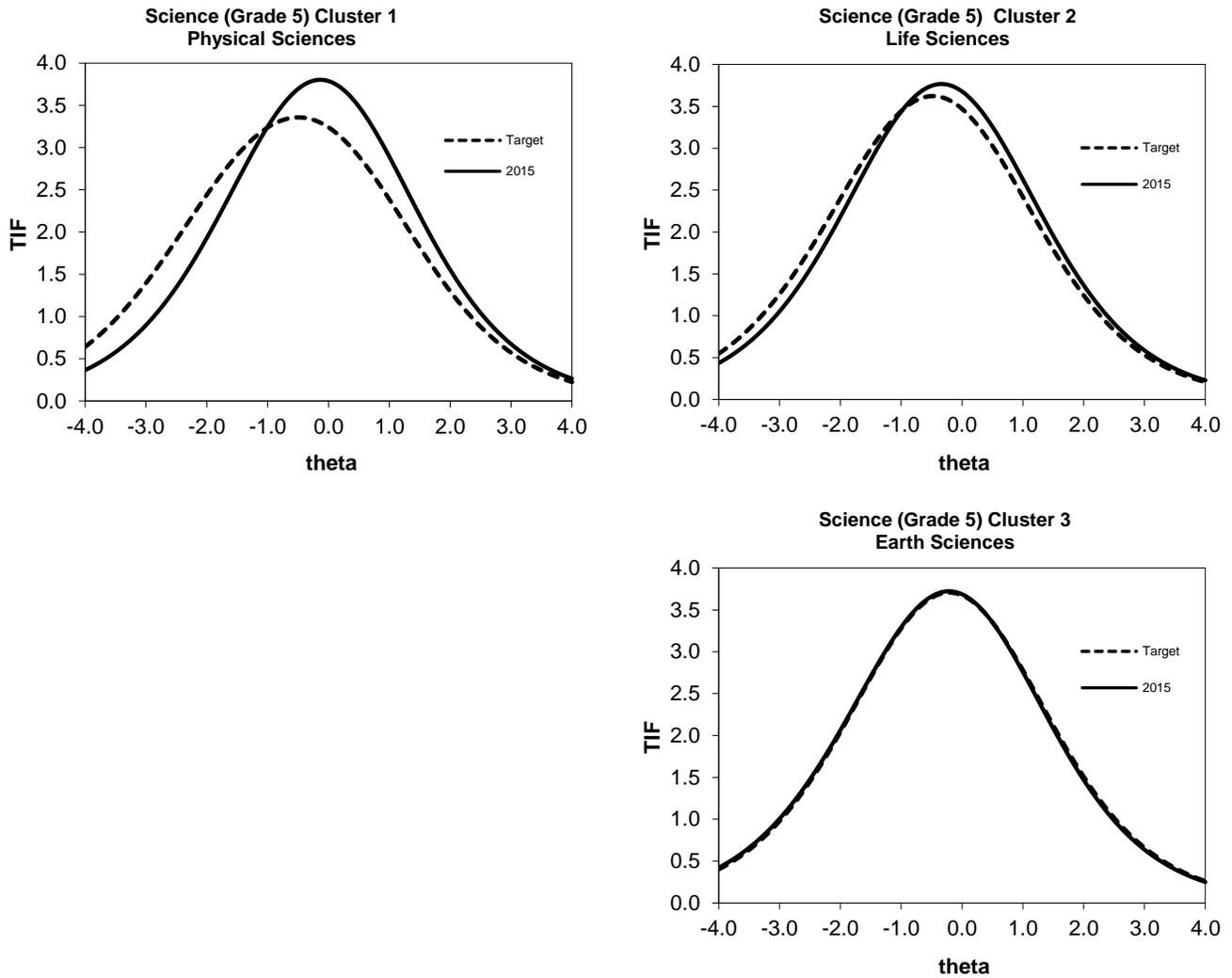


Figure 4.B.2 Plots of Target Information Functions and Projected Information for Clusters for Science, Grade Eight

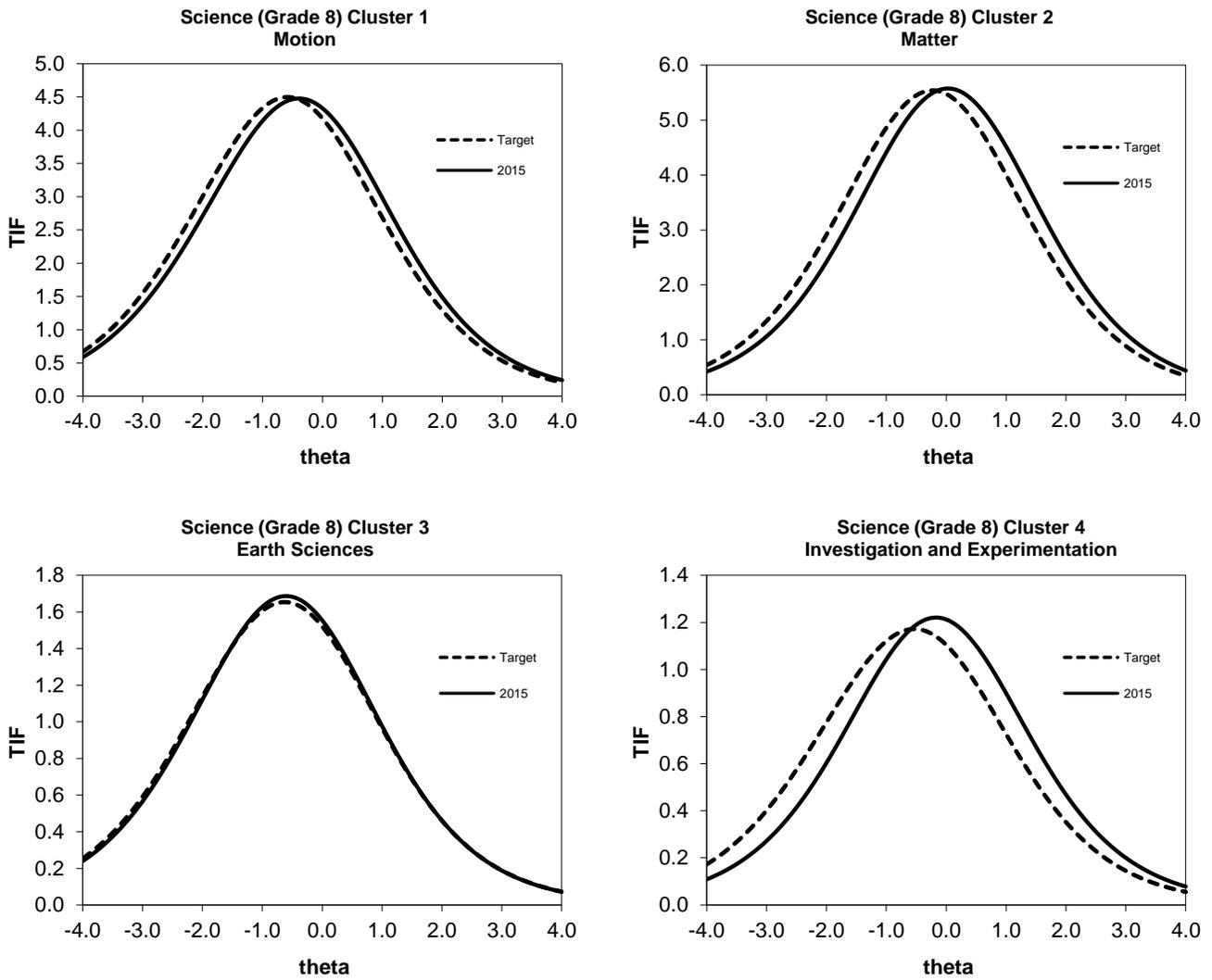
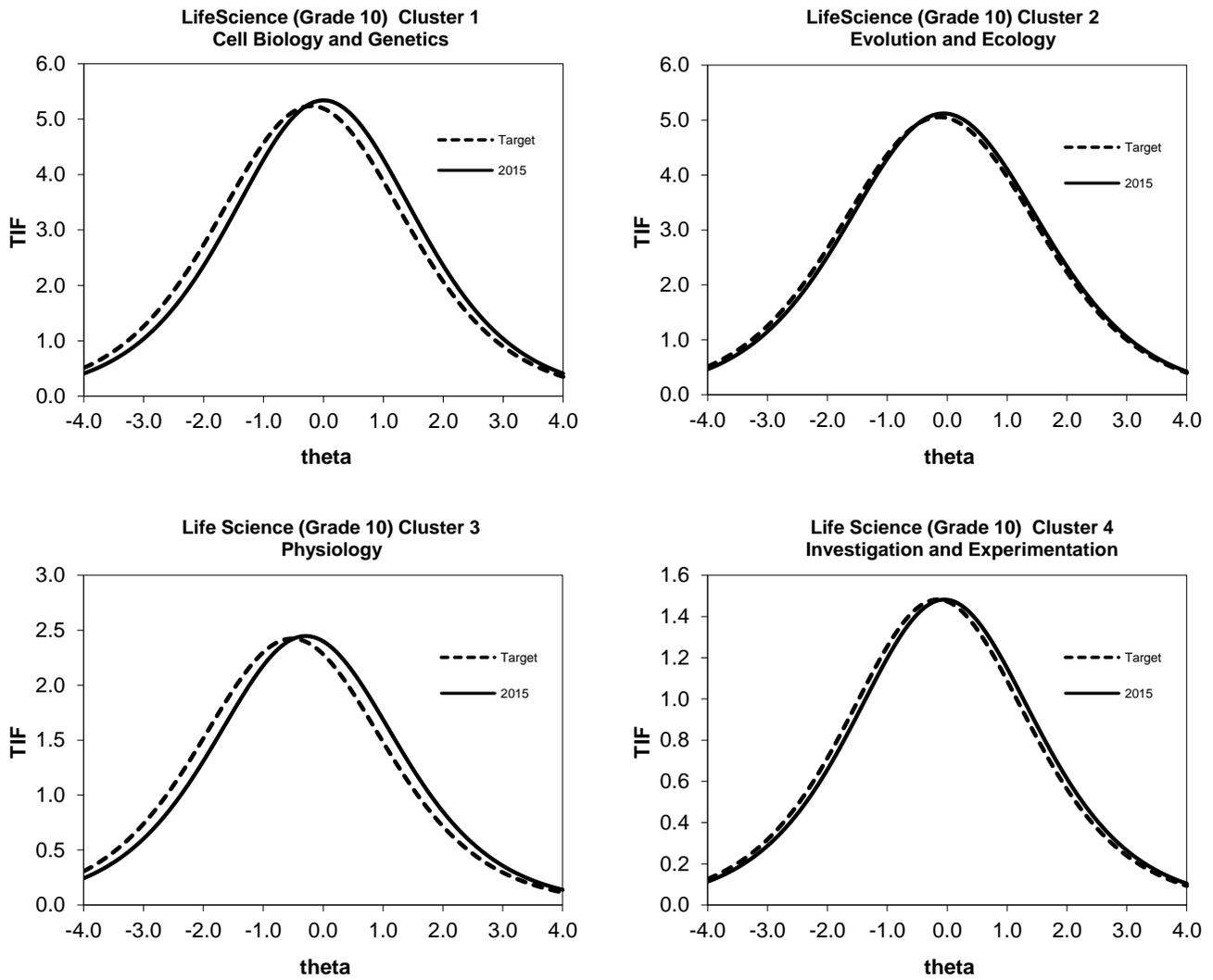


Figure 4.B.3 Plots of Target Information Functions and Projected Information for Clusters for Life Science, Grade Ten



Chapter 5: Test Administration

Test Security and Confidentiality

All tests within the California Assessment of Student Performance and Progress (CAASPP) Program are secure documents. For the California Modified Assessment (CMA) 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.

The OTI's mission is to

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

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

Test Development

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

Item and Data Review

As mentioned in Chapter 3, Assessment Review Panel (ARP) meetings were not held in 2014–15 because there was no new item development for the 2014–15 CMA for Science forms. However, before the 2014–15 administration, ETS facilitated ARP meetings every year to review all newly developed CMA for Science items and associated statistics. ETS enforced security measures at ARP meetings to protect the integrity of meeting materials using the following guidelines:

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

Item Banking

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

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

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

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

Transfer of Forms and Items to the CDE

ETS shares a secure file transfer protocol (SFTP) site with the CDE. SFTP is a method for reliable and exclusive routing of files. Files reside on a password-protected server that only authorized users may access. On that site, ETS posts Microsoft Word and Excel, Adobe Acrobat PDF, or other document files for the CDE to review. ETS sends a notification e-mail to the CDE to announce that files are posted. Item data are always transmitted in an encrypted format to the SFTP site; test data are never sent via e-mail. The SFTP server is used as a conduit for the transfer of files; secure test data are not stored permanently on the shared SFTP server.

Security of Electronic Files Using a Firewall

A firewall is software that prevents unauthorized entry to files, e-mail, and other organization-specific programs. ETS data exchange and internal e-mail remain within the ETS firewall at all ETS locations, ranging from Princeton, New Jersey, to San Antonio, Texas, to Concord and Sacramento, California.

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

Test Administration

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

Test Delivery

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

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

- Anyone having access to the testing materials must sign and submit a “CAASPP Test Security Affidavit for Test Examiners, Proctors, Scribes, and Any Other Persons Having Access to CAASPP Tests (For all CAASPP assessments, including field tests)” form to the test site coordinator before receiving access to any testing materials.
- It is the responsibility of each person participating in the CAASPP Program 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 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, including Pre-ID files and reports.

Student Test Results

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

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

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

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

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

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

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

Procedures to Maintain Standardization

The CMA 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 CMA for Science, as described in this section.

Test Administrators

The CMA for Science are administered in conjunction with the other tests that comprise the CAASPP Program. 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 Administrator

The CMA for Science are administered by test administrators who may be assisted by test proctors and scribes. A test administrator is an employee of an LEA or an employee of a nonpublic, nonsectarian school (NPS) who has been trained to administer the tests and has signed a CAASPP Test Security Affidavit. Test administrators must follow the directions in the *California Modified Assessment Directions for Administration (DFA)* (CDE, 2015c) exactly.

Test Proctor

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

Scribe

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

Directions for Administration

CMA for Science DFAs are manuals used by test administrators to administer the CMA for Science to students (CDE, 2015c). Test administrators must follow all directions and guidelines and read, word-for-word, the instructions to students in "SAY" boxes to ensure test standardization.

CAASPP Paper-Pencil Testing Test Administration Manual

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

Test Booklets

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

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

Universal Tools, Designated Supports, and Accommodations

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

Identification

Most students with disabilities and most English learners take the CMA for Science under standard conditions. However, some students with disabilities and some English learners may need assistance when taking the CMA for Science. This assistance takes the form of universal tools, designated supports, and accommodations (see Appendix 2.D on page 22 in Chapter 2 for details). During the test, these students may use the special services specified in their IEP or Section 504 plan. If students use universal tools, designated supports, and/or accommodations for the CMA for Science, test examiners are responsible for marking the universal tools, designated supports, and/or accommodations used on the students’ answer documents. Because the CMA for Science were developed with

modifications built into the test, non-embedded accessibility supports are not allowed. Students who require additional modifications take the California Standards Tests for Science with non-embedded accessibility supports.

Scoring

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

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 CMA for Science 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 *DFAs* (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 Program 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://caaspp.org/administration/toms/>
- California Department of Education. (2014b). *2015 CAASPP paper-pencil testing test administration manual*. Sacramento, CA. Retrieved from http://caaspp.org/rsc/pdfs/CAASPP.coord_man.2015.pdf
- California Department of Education. (2015c). *2015 California Modified Assessment directions for administration*. Sacramento, CA. Retrieved from http://www.caaspp.org/rsc/pdfs/CMA.grade-5_dfa.2015.pdf
- California Department of Education. (2015d). *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>

Chapter 6: Performance Standards

Background

The California Modified Assessment (CMA) were introduced to California’s standardized testing program in stages, starting with the lower grades in 2008. Performance standards for each new test were developed after the introductory year for operational use in subsequent administrations. The CMA for English–Language Arts (ELA) and Mathematics in grades three through five and science in grade five were established in spring 2008. For each of these tests, the performance standards were developed in September and October 2008 and adopted by the State Board of Education (SBE) for their 2009 operational administration. In spring 2009, the CMA for ELA in grades six through eight, mathematics in grades six and seven, and science in grade eight were introduced. The performance standards for those tests were developed in August 2009 and adopted by the SBE for the 2010 operational administration of those CMA.

The CMA for high school phase 1 (ELA in grade nine, Life Science in grade ten, and end-of-course Algebra I) were introduced in spring 2010. The performance standards for those tests were developed in August 2010 and adopted by the SBE for use starting in the 2011 operational administration. Finally, the CMA for high school phase 2 (ELA in grades ten and eleven and end-of-course Geometry) were introduced in spring 2011, for these tests were established in fall 2011 and adopted by the SBE for use starting in the 2012 operational administration.

The performance standards for the CMA were defined by the SBE as far below basic, below basic, basic, proficient, and advanced. Performance standards are developed from a general description of the performance level (policy-level descriptors) and competencies lists, which operationally define each level. Cut scores numerically define the performance levels.

In 2014–15, the CMA for Science in grades five and eight and Life Science in grade ten were administered to eligible students. Consequently, the performance standards for the grades and subjects were applied to the scores of students.

California employed carefully designed standard-setting procedures to facilitate the development of performance standards for each CMA. The standard-setting method used for the CMA for Science is the Bookmark method (Mitzel, et al., 2001). These processes are described in the sections that follow.

Standard-Setting Procedure

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

- Familiarity with the subject matter assessed
- Familiarity with students in the respective grade levels
- Experience with English learners
- Experience in special education and general education classrooms as well as integrated classrooms

- Familiarity with the California content standards
- An understanding of the CMA
- An appreciation of the consequences of setting these cut scores

Panelists were recruited from diverse geographic regions and from different gender and major racial/ethnic subgroups to be representative of the educators of the state’s CMA-eligible students (Educational Testing Service [ETS], 2009a, 2009b, 2010, 2011).

For each test, three cut scores were developed in order to differentiate four of the five performance levels: below basic, basic, proficient, and advanced. Far below basic was defined as chance-level performance.

The standard-setting processes implemented for the CMA required panelists to follow these steps, which include 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 test items are based and take notes on their own expectations in the content area. This allowed the panelists to understand how their perceptions may relate to the complexity of the content standards.
2. At the start of the workshop, panelists received training, which 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 Bookmark method. Impact data included the percentage of examinees assessed in a previous administration of the test that would fall into each level, given the panelists’ judgments of cut scores.
3. Panelists became familiar with the difficulty level of the items by taking the actual test and then assessing and discussing the demands of the test items.
4. Panelists reviewed the draft list of competencies as a group, noting the increasing demands of each subsequent level. In this step, they began to visualize the knowledge and skills of students in each performance level.
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). Panelists could revise their judgments during the process if they wished.
7. The final recommended cut scores were based on the median of panelists’ judgment scores at the end of three rounds (in the Bookmark method, the panel recommendation is calculated by taking the median of the small group [table] medians). For the CMA for Science, the cut scores recommended by the panelists and the recommendation of the State Superintendent of Public Instruction were presented for public comment at regional public hearings. Comments and recommendations were then presented to the SBE for adoption.

Development of Competencies Lists

Prior to the CMA 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. For each content area, one panel of educators was assembled for each grade to identify and discuss the competencies required of students taking the CMA for each performance level (below basic, basic, proficient, and advanced). The lists were used to facilitate the discussion and construction of the target student definitions during the standard-setting workshop.

Standard-Setting Methodology

Bookmark Method

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

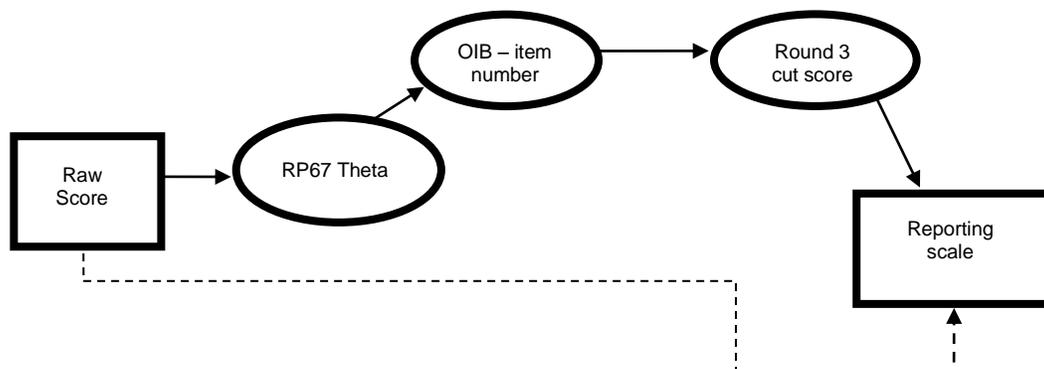
The Bookmark method is an item-mapping procedure in which panelists consider content covered by items in a specially constructed book where items are ordered from easiest to hardest based on operational student performance data from a previous test administration. The “item map,” which accompanies the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

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

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

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

Figure 6.1 Bookmark Standard-setting Process for the CMA



Results

The cut scores obtained as a result of the standard-setting process are on the item response theory (IRT) scale; each recommended cut score was associated with a theta value in the OIB. This RP67 Theta has a corresponding number-correct or raw score for the test form upon which standards were set; the scores were then translated to a score scale that ranges between 150 and 600.

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

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

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

where,

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

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

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

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

The scale-score ranges for each performance level are presented in Table 2.1 on page 17. The cut score for each performance level is the lower bound of each scale-score range. The scale-score ranges do not change from year to year. Once established, they remain unchanged from administration to administration until such time that new performance standards are adopted.

Table 7.2 on page 64 in Chapter 7 presents the percentages of examinees meeting each performance level for the 2014–15 administration.

References

- Educational Testing Service. (2009a). *Technical report on the standard setting workshop for the California Modified Assessment: ELA grades three through five, mathematics grades three through five, and science grade five. February 6, 2009* (California Department of Education Contract Number 5417). Princeton, NJ: Author.
- Educational Testing Service. (2009b). *Technical report on the standard setting workshop for the California Modified Assessment: ELA grades six through eight, mathematics grades six and seven, and science grade eight. November 5, 2009* (California Department of Education Contract Number 5417). Princeton, NJ: Author.
- Educational Testing Service. (2010). *Technical report on the standard setting workshop for the California Modified Assessment: ELA grade nine, Algebra I, and Life Science grade ten. November 9, 2010* (California Department of Education Contract Number 5417). Princeton, NJ: Author.
- Educational Testing Service. (2011). *Technical report on the standard setting workshop for the California Modified Assessment: ELA grades ten and eleven and Geometry. November 1, 2011* (California Department of Education Contract Number 5417). Princeton, NJ: Author.
- Huynh, H. (1998). On score locations of binary and partial credit items and their applications to item mapping and criterion-referenced interpretation. *Journal of Educational and Behavioral Statistics*, 23(19), 35–56.
- Lewis, D. M., Green, D. R., Mitzel, H. C., Baum, K., & Patz, R. J. (1999). *The bookmark standard setting procedure: Methodology and recent implications*. Manuscript under review.
- Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249–81). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Chapter 7: Scoring and Reporting

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

Procedures for Maintaining and Retrieving Individual Scores

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

In the 2014–15 administration, because the raw-score-to-scale-score conversion tables were developed before tests were administered using pre-equating, preliminary individual student results were available for download prior to the printing of paper reports. This electronic reporting was made possible through the Online Reporting System.

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

Scoring and Reporting Specifications

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

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

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

Scanning and Scoring

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

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

Types of Scores and Subscores

Raw Score

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

Subscore

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

Scale Score

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

Performance Levels

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

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

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

Score Verification Procedures

Various necessary measures are taken to ascertain that the scoring keys are applied to the student responses as intended and that the student scores are computed accurately. In 2014–15, every regular and special-version multiple-choice test is certified by ETS prior to being included in electronic reporting. To certify a test, psychometricians gather a certain number of test cases and verify the accurate application of scoring keys and scoring tables.

Scoring Key Verification Process

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

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

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

Overview of Score Aggregation Procedures

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

Individual Scores

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

Score Distributions and Summary Statistics

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

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

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

CMA	No. of Items	No. of Examinees	Scale Score		Raw Score		Raw Score	Raw Score
			Mean	Std. Dev.	Mean	Std. Dev.	% Correct	% Correct
Grade 5 Science	48	23,236	343	57	28.26	7.39	58.87	15.40
Grade 8 Science	54	19,212	330	64	30.76	7.97	56.97	14.75
Grade 10 Life Science	60	9,601	312	63	33.51	9.18	55.85	15.30

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

The numbers in the summary tables may not match exactly the results reported on the CDE’s Web site because of slight differences in the samples used to compute the statistics. The P2 data file was used for the analyses in this chapter. This file contained the entire test-taking population and all the student records used as of October 28, 2015. This file contained data collected from all LEAs but did not include corrections of demographic data through the California Longitudinal Pupil Assessment Data System. In addition, students with invalid scores were excluded from the tables.

Table 7.2 Percentages of Examinees in Each Performance Level

CMA *	Far Below Basic	Below Basic	Basic	Proficient	Advanced	Proficient/Advanced *
Grade 5 Science	3%	21%	31%	30%	14%	45%
Grade 8 Science	13%	19%	31%	24%	13%	37%
Grade 10 Life Science	15%	28%	28%	20%	7%	28%

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

Table 7.A.1 in Appendix 7.A on page 69 shows the distributions of scale scores for each CMA for Science.

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

Group Scores

Statistics summarizing student performance by each grade-level test for selected groups of students are provided starting on page 70 in Table 7.B.1 through Table 7.B.3 for the CMA for Science.

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

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

Table 7.3 Subgroup Definitions

Subgroup	Definition
Gender	<ul style="list-style-type: none"> • Male • Female
Ethnicity	<ul style="list-style-type: none"> • American Indian or Alaska Native • Asian <ul style="list-style-type: none"> – Asian Indian – Cambodian – Chinese – Hmong – Japanese – Korean – Laotian

Subgroup	Definition
	<ul style="list-style-type: none"> – Vietnamese – Other Asian • Pacific Islander <ul style="list-style-type: none"> – Guamanian – Native Hawaiian – Samoan – Tahitian – Other Pacific Islander • Filipino • Hispanic or Latino • African American • White (not Hispanic)
English-language Fluency	<ul style="list-style-type: none"> • English only • Initially fluent English proficient • English learner • Reclassified fluent English proficient • To be determined (TBD)
Economic Status	<ul style="list-style-type: none"> • Not economically disadvantaged • Economically disadvantaged
Primary Disability	<ul style="list-style-type: none"> • Intellectual disability (ID) • 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 California Assessment of Student Performance and Progress (CAASPP) System provide results or score summaries that are reported for different purposes. The three major purposes are:

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

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

Types of Score Reports

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

Table 7.4 Types of CMA for Science Reports

1. Electronic Summary Report	▪ CAASPP Aggregate Report (includes subgroups)
2. Individual Reports	<ul style="list-style-type: none"> ▪ CAASPP Student Data File ▪ CAASPP Student Score Report for Smarter Balanced English Language Arts/Literacy (ELA) and Mathematics and CST/CMA in Grades Five and Eight ▪ CAASPP Student Score Report for CST/CMA in Grade Ten
3. Internet Reports	▪ CMA 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 the Test Operations Management System (TOMS). The LEA forwards the appropriate reports to test sites or, in the case of the CAASPP Student Score Report, sends the report(s) to the child's parent or guardian and forwards a copy to the student's school or test site. CAASPP Student Score Reports that include individual student results are not distributed beyond the student's school. Internet reports are described on the CDE Web site and are accessible to the public online at <http://caaspp.cde.ca.gov/>.

Because results were pre-equated, individual student scores were also available to LEAs prior to the release of final reports via electronic reporting, accessed using the 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 CMA for Science taken. Scale scores are reported on a scale ranging from 150 to 600. The performance levels reported are: far below basic, below basic, basic, proficient, and advanced.

Reports for students with disabilities and English learners who use universal tools, designated supports, and accommodations include a notation that indicates that the student used non-embedded supports (accommodations). Scores for students who use non-embedded supports (accommodations) are reported in the same way as they are for nonaccommodated students.

Further information about the CAASPP Student Score Report is provided in Appendix 7.C on page 76.

Student Score Report Applications

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

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

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

Criteria for Interpreting Test Scores

An LEA may use CMA for Science results to help make decisions about student placement, promotion, retention, or other considerations related to student achievement. However, it is important to remember that a single test can provide only limited information. Other relevant information should be considered as well. It is advisable for parents to evaluate their child's strengths and weaknesses in the relevant topics by reviewing classroom work and progress reports in addition to the child's CMA for Science results (CDE, 2015). 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 CMA for Science, the user is limited to comparisons within the same content area and grade. This is because the score scales are different for each content area and grade. The user may compare scale scores for the same content area and grade, within a school, between schools, or between a school and its district, its county, or the state. The user can also make comparisons within the same grade and content area across years. Comparing scores obtained in different grades or content areas should be avoided because the results are not on the same scale. Comparisons between raw scores should be limited to comparisons within not only content area and grade but also test year. For more details on the criteria for interpreting information provided on the score reports, see the *2015 CAASPP Post-Test Guide* (CDE, 2015).

Reference

California Department of Education. (2015). *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.

Appendix 7.A—Scale Score Distribution Tables

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

Scale Score	Grade 5	Grade 8	Grade 10
570 – 600	13	22	6
540 – 569	26	48	13
510 – 539	77	135	17
480 – 509	97	106	52
450 – 479	796	604	128
420 – 449	970	664	321
390 – 419	2,235	1,968	562
360 – 389	4,059	1,898	944
330 – 359	5,499	3,227	1,595
300 – 329	3,918	4,330	1,749
270 – 299	3,675	3,058	1,729
240 – 269	1,478	2,158	1,287
210 – 239	329	697	968
180 – 209	59	272	204
150 – 179	5	25	26

Appendix 7.B—Demographic Summaries

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

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

	No. Tested	Mean Scale Scores	Std. Dev. of Scale Scores	Percent in Far Below Basic	Percent in Below Basic	Percent in Basic	Percent in Proficient	Percent in Advanced	Mean Percent Correct in Physical Sciences	Mean Percent Correct in Life Sciences	Mean Percent Correct in Earth Sciences
All valid scores	23,236	343	57	3%	21%	31%	30%	14%	56%	62%	59%
Male	15,312	345	58	3%	21%	30%	31%	16%	57%	62%	59%
Female	7,924	338	54	3%	22%	34%	29%	12%	54%	61%	58%
Gender unknown	0	–	–	–	–	–	–	–	–	–	–
American Indian	181	348	54	1%	18%	30%	38%	13%	58%	63%	60%
Asian American	747	341	60	4%	22%	30%	29%	15%	55%	60%	59%
Pacific Islander	99	343	52	2%	24%	31%	30%	12%	56%	62%	59%
Filipino	200	346	52	2%	18%	37%	29%	15%	58%	62%	59%
Hispanic	15,330	338	54	3%	23%	33%	29%	12%	55%	60%	58%
African American	2,182	331	53	3%	26%	34%	28%	9%	54%	59%	54%
White	3,962	366	60	2%	13%	23%	35%	27%	61%	69%	65%
Two or more Races	535	357	62	2%	18%	26%	34%	21%	59%	67%	61%
English only	12,266	350	58	2%	18%	29%	32%	18%	58%	65%	60%
Initially fluent English prof.	197	365	55	1%	13%	25%	36%	25%	61%	70%	64%
English learner	10,361	333	52	3%	25%	34%	28%	9%	54%	58%	56%
Reclassified fluent English prof.	402	360	59	2%	13%	27%	35%	22%	60%	67%	64%
TBD	1	–	–	–	–	–	–	–	–	–	–
English prof. unknown	9	–	–	–	–	–	–	–	–	–	–
Autism	1,762	340	62	3%	26%	28%	27%	15%	55%	60%	59%
Deaf-blindness	2	–	–	–	–	–	–	–	–	–	–
Emotional disturbance	356	343	60	4%	22%	26%	29%	19%	57%	63%	57%
Hearing impairment	244	322	54	7%	30%	32%	24%	7%	51%	54%	55%
ID	405	294	43	12%	47%	31%	9%	1%	45%	46%	45%
Multiple disabilities	26	325	55	4%	42%	19%	27%	8%	48%	58%	56%
Orthopedic impairment	129	331	58	4%	28%	36%	20%	12%	51%	61%	55%
Other health impairment	2,629	352	58	2%	17%	29%	33%	19%	58%	65%	61%
Specific learning disability	14,209	344	55	2%	20%	32%	31%	14%	56%	62%	59%
Speech or language impairment	2,036	335	53	3%	24%	33%	29%	11%	54%	59%	57%
Traumatic brain injury	39	349	54	0%	21%	31%	33%	15%	58%	64%	60%
Visual impairment	44	328	55	7%	25%	32%	25%	11%	48%	61%	56%
Disability unknown	1,355	343	57	3%	20%	31%	31%	15%	56%	62%	58%

	No. Tested	Mean Scale Scores	Std. Dev. of Scale Scores	Percent in Far Below Basic	Percent in Below Basic	Percent in Basic	Percent in Proficient	Percent in Advanced	Mean Percent Correct in Physical Sciences	Mean Percent Correct in Life Sciences	Mean Percent Correct in Earth Sciences
Not economically disadvantaged	5,070	360	61	2%	15%	26%	34%	24%	60%	67%	63%
Economically disadvantaged	18,166	338	54	3%	23%	33%	29%	12%	55%	60%	57%
Primary Ethnicity—Not Economically Disadvantaged											
American Indian	35	359	59	0%	17%	34%	31%	17%	61%	66%	62%
Asian American	337	344	65	4%	20%	29%	28%	18%	56%	61%	60%
Pacific Islander	24	355	52	4%	17%	25%	38%	17%	60%	66%	61%
Filipino	125	352	51	1%	14%	34%	32%	18%	58%	64%	62%
Hispanic	2,006	353	57	2%	16%	29%	34%	19%	58%	65%	62%
African American	418	335	53	4%	21%	36%	29%	11%	53%	61%	56%
White	1,941	375	62	1%	11%	21%	36%	32%	63%	72%	67%
Two or more Races	184	365	67	3%	16%	22%	32%	27%	62%	68%	63%
Primary Ethnicity—Economically Disadvantaged											
American Indian	146	346	52	1%	18%	29%	40%	12%	57%	63%	59%
Asian American	410	338	57	4%	23%	31%	30%	12%	55%	60%	58%
Pacific Islander	75	340	52	1%	27%	33%	28%	11%	55%	61%	58%
Filipino	75	336	51	3%	23%	41%	24%	9%	58%	59%	55%
Hispanic	13,324	336	53	3%	24%	34%	29%	11%	54%	60%	57%
African American	1,764	330	53	3%	27%	34%	27%	8%	54%	59%	54%
White	2,021	358	57	2%	15%	26%	34%	23%	59%	67%	62%
Two or more Races	351	353	59	1%	18%	28%	34%	18%	58%	66%	60%

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

	No. Tested	Mean Scale Scores	Std. Dev. of Scale Scores	Percent in Far Below Basic	Percent in Below Basic	Percent in Basic	Percent in Proficient	Percent in Advanced	Mean Percent Correct in Motion	Mean Percent Correct in Matter	Mean Percent Correct in Earth Science	Mean Percent Correct in Investigation and Experimentation
All valid scores	19,212	330	64	13%	19%	31%	24%	13%	61%	51%	67%	55%
Male	12,788	332	66	13%	19%	29%	24%	14%	62%	51%	68%	55%
Female	6,424	327	59	12%	19%	34%	24%	10%	60%	50%	66%	55%
Gender unknown	0	—	—	—	—	—	—	—	—	—	—	—
American Indian	164	340	60	9%	16%	35%	25%	16%	63%	53%	70%	60%
Asian American	527	340	76	13%	18%	27%	24%	19%	63%	54%	66%	58%
Pacific Islander	79	329	69	14%	22%	29%	22%	14%	60%	51%	66%	54%
Filipino	216	333	60	11%	13%	40%	25%	12%	61%	52%	66%	59%
Hispanic	12,703	327	61	13%	20%	32%	24%	11%	61%	50%	67%	54%
African American	2,025	313	59	19%	24%	31%	19%	8%	58%	47%	62%	50%
White	3,109	352	71	9%	14%	26%	28%	23%	66%	56%	72%	60%
Two or more Races	389	338	65	10%	16%	32%	26%	15%	64%	52%	69%	59%
English Only	9,979	335	66	13%	18%	30%	25%	15%	62%	52%	68%	56%
Initially fluent English prof.	218	345	71	11%	15%	27%	28%	20%	65%	54%	70%	57%
English learner	7,527	317	56	15%	23%	34%	21%	7%	59%	47%	64%	52%
Reclassified fluent English prof.	1,481	362	68	6%	11%	25%	33%	25%	68%	57%	75%	65%
TBD	0	—	—	—	—	—	—	—	—	—	—	—
English prof. unknown	7	—	—	—	—	—	—	—	—	—	—	—
Autism	1,309	344	76	12%	18%	26%	22%	21%	62%	55%	71%	58%
Deaf-blindness	4	—	—	—	—	—	—	—	—	—	—	—
Emotional disturbance	456	325	66	14%	22%	29%	22%	13%	60%	49%	66%	54%
Hearing impairment	239	319	58	15%	25%	27%	24%	8%	58%	49%	64%	54%
ID	409	278	48	38%	30%	25%	6%	1%	48%	41%	51%	40%
Multiple disabilities	25	299	60	32%	12%	36%	12%	8%	53%	45%	55%	52%
Orthopedic impairment	124	320	62	16%	19%	35%	18%	12%	59%	48%	66%	54%
Other health impairment	2,079	335	66	12%	19%	29%	25%	16%	62%	52%	68%	56%
Specific learning disability	12,668	331	62	12%	19%	32%	25%	12%	62%	50%	68%	56%
Speech or language impairment	965	328	60	14%	17%	34%	24%	11%	60%	51%	67%	55%
Traumatic brain injury	38	314	52	16%	29%	21%	32%	3%	58%	49%	59%	51%
Visual impairment	47	316	59	19%	23%	26%	21%	11%	57%	49%	64%	48%
Disability unknown	849	324	62	15%	20%	32%	23%	10%	61%	49%	63%	53%
Not economically disadvantaged	4,404	345	69	11%	15%	28%	27%	19%	64%	54%	70%	59%
Economically disadvantaged	14,808	326	62	14%	20%	32%	23%	11%	61%	49%	66%	54%
Primary Ethnicity—Not Economically Disadvantaged												
American Indian	41	343	62	10%	15%	32%	20%	24%	64%	53%	71%	62%
Asian American	208	348	76	10%	15%	29%	25%	21%	64%	56%	68%	61%
Pacific Islander	20	366	85	15%	5%	30%	20%	30%	67%	60%	74%	61%
Filipino	131	331	58	11%	15%	41%	21%	12%	60%	52%	66%	60%

	No. Tested	Mean Scale Scores	Std. Dev. of Scale Scores	Percent in Far Below Basic	Percent in Below Basic	Percent in Basic	Percent in Proficient	Percent in Advanced	Mean Percent Correct in Motion	Mean Percent Correct in Matter	Mean Percent Correct in Earth Science	Mean Percent Correct in Investigation and Experimentation
Hispanic	1,847	337	64	11%	17%	30%	27%	15%	63%	53%	69%	57%
African American	450	317	66	20%	21%	28%	22%	9%	58%	49%	62%	50%
White	1,574	360	71	8%	11%	24%	29%	28%	67%	58%	74%	63%
Two or more Races	133	352	71	8%	14%	30%	26%	23%	66%	55%	72%	64%
Primary Ethnicity—Economically Disadvantaged												
American Indian	123	338	60	8%	16%	36%	27%	13%	63%	53%	70%	60%
Asian American	319	335	76	14%	19%	26%	23%	17%	62%	52%	65%	56%
Pacific Islander	59	317	58	14%	27%	29%	22%	8%	58%	48%	64%	52%
Filipino	85	335	62	11%	11%	38%	31%	11%	62%	53%	67%	56%
Hispanic	10,856	325	60	14%	21%	32%	23%	10%	61%	49%	66%	54%
African American	1,575	312	57	19%	24%	32%	18%	7%	58%	47%	62%	50%
White	1,535	343	70	10%	16%	28%	27%	19%	64%	53%	70%	58%
Two or more Races	256	331	61	12%	17%	34%	26%	12%	62%	50%	67%	57%

Table 7.B.3 Demographic Summary for Life Science (Grade 10)

	No. Tested	Mean Scale Scores	Std. Dev. of Scale Scores	Percent in Far Below Basic	Percent in Below Basic	Percent in Basic	Percent in Proficient	Percent in Advanced	Mean Percent Correct in Cell Biology and Genetics	Mean Percent Correct in Evolution and Ecology	Mean Percent Correct in Physiology	Mean Percent Correct in Investigation and Experimentation
All valid scores	9,601	312	63	15%	28%	28%	20%	7%	55%	55%	60%	56%
Male	6,386	313	65	17%	28%	26%	21%	8%	55%	55%	60%	56%
Female	3,215	311	56	13%	30%	33%	20%	5%	55%	55%	59%	55%
Gender unknown	0	–	–	–	–	–	–	–	–	–	–	–
American Indian	82	318	58	10%	30%	26%	28%	6%	55%	57%	63%	60%
Asian American	220	318	63	12%	30%	28%	21%	9%	55%	57%	62%	59%
Pacific Islander	28	321	78	21%	29%	14%	21%	14%	59%	55%	54%	64%
Filipino	92	330	61	7%	22%	36%	27%	9%	58%	61%	63%	61%
Hispanic	6,363	308	60	17%	29%	30%	19%	6%	54%	54%	59%	55%
African American	882	298	59	20%	35%	24%	16%	5%	52%	50%	57%	51%
White	1,725	333	68	10%	22%	27%	27%	14%	59%	60%	66%	61%
Two or more Races	209	328	68	12%	19%	32%	23%	13%	58%	59%	63%	61%
English Only	4,888	319	65	14%	26%	28%	23%	9%	56%	56%	62%	57%
Initially fluent English prof.	198	328	66	13%	22%	27%	26%	13%	59%	58%	63%	61%
English learner	3,608	295	53	20%	34%	29%	14%	2%	51%	51%	55%	51%
Reclassified fluent English prof.	898	342	64	7%	19%	28%	31%	15%	61%	62%	70%	65%
TBD	2	–	–	–	–	–	–	–	–	–	–	67%
English prof. unknown	7	–	–	–	–	–	–	–	–	–	–	69%
Autism	599	334	73	10%	23%	27%	24%	15%	58%	61%	66%	60%
Deaf-blindness	3	–	–	–	–	–	–	–	–	–	–	33%
Emotional disturbance	300	318	69	16%	24%	27%	24%	9%	56%	56%	62%	58%
Hearing impairment	175	303	54	15%	34%	32%	14%	5%	52%	53%	57%	54%
ID	235	264	42	37%	45%	15%	3%	0%	44%	43%	45%	40%
Multiple disabilities	20	290	58	20%	40%	20%	20%	0%	50%	51%	51%	53%
Orthopedic impairment	73	312	58	12%	27%	36%	21%	4%	53%	56%	61%	56%
Other health impairment	951	319	68	15%	27%	24%	22%	11%	57%	56%	62%	56%
Specific learning disability	6,402	312	60	15%	28%	30%	20%	6%	55%	55%	60%	56%
Speech or language impairment	328	310	57	11%	33%	29%	22%	5%	55%	55%	58%	55%
Traumatic brain injury	20	324	70	10%	35%	20%	20%	15%	56%	60%	60%	61%
Visual impairment	20	345	89	5%	35%	15%	30%	15%	64%	59%	66%	62%
Disability unknown	475	307	66	22%	26%	25%	21%	7%	53%	53%	59%	55%
Not economically disadvantaged	2,295	327	66	11%	24%	28%	24%	12%	58%	58%	64%	60%
Economically disadvantaged	7,306	308	61	17%	30%	28%	19%	6%	54%	54%	59%	55%
Primary Ethnicity—Not Economically Disadvantaged												
American Indian	27	322	56	4%	33%	30%	22%	11%	56%	56%	68%	59%
Asian American	83	325	67	11%	29%	29%	19%	12%	55%	59%	66%	60%
Pacific Islander	12	362	86	8%	17%	17%	33%	25%	69%	64%	63%	75%
Filipino	56	337	65	7%	16%	36%	30%	11%	59%	63%	64%	63%
Hispanic	978	319	63	12%	27%	31%	21%	9%	56%	56%	62%	57%

	No. Tested	Mean Scale Scores	Std. Dev. of Scale Scores	Percent in Far Below Basic	Percent in Below Basic	Percent in Basic	Percent in Proficient	Percent in Advanced	Mean Percent Correct in Cell Biology and Genetics	Mean Percent Correct in Evolution and Ecology	Mean Percent Correct in Physiology	Mean Percent Correct in Investigation and Experimentation
African American	194	308	61	16%	35%	21%	22%	6%	55%	52%	60%	55%
White	858	339	69	9%	20%	26%	29%	16%	61%	62%	67%	62%
Two or more Races	87	340	65	10%	13%	34%	26%	16%	61%	62%	67%	65%
Primary Ethnicity—Economically Disadvantaged												
American Indian	55	316	59	13%	29%	24%	31%	4%	55%	57%	61%	60%
Asian American	137	314	60	13%	31%	27%	23%	7%	55%	56%	60%	58%
Pacific Islander	16	290	57	31%	38%	13%	13%	6%	52%	49%	47%	56%
Filipino	36	320	55	6%	31%	36%	22%	6%	55%	59%	62%	58%
Hispanic	5,385	306	59	17%	30%	29%	18%	5%	53%	54%	58%	54%
African American	688	295	58	22%	35%	25%	14%	4%	52%	50%	56%	50%
White	867	327	67	11%	25%	27%	25%	11%	58%	58%	64%	59%
Two or more Races	122	320	70	14%	24%	30%	21%	11%	56%	57%	60%	58%

Appendix 7.C—Types of Score Reports

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

2014–15 CAASPP Student Score Reports	
Description	Use and Distribution
<p>The CAASPP Student Report—CMA for Science A report for the Smarter Balanced Summative Assessments for ELA and Mathematics and the CMA for Science at the student’s grade level (grade five or eight)</p>	
<p>This report provides parents/guardians and teachers with the student’s results, presented in tables and graphs.</p> <p>Data presented for the science assessment taken include the following:</p> <ul style="list-style-type: none"> • Scale scores • Performance levels <p>The report is formatted with the student’s mailing address positioned for use in windowed envelopes for mailing to parents/guardians if the LEA provided mailing addresses.</p>	<p>This report includes individual student results and is not distributed beyond parents/guardians and the student’s school.</p> <p>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.</p>
<p>The CAASPP Student Report—CMA for Science A report for the CMA for Science in grade ten</p>	
<p>This report provides parents/guardians and teachers with the student’s results, presented in tables and graphs.</p> <p>Data presented for the science assessment taken include the following:</p> <ul style="list-style-type: none"> • Scale scores • Performance levels <p>The report is formatted with the student’s mailing address positioned for use in windowed envelopes for mailing to parents/guardians if the LEA provided mailing addresses.</p>	<p>This report includes individual student results and is not distributed beyond parents/guardians and the student’s school.</p> <p>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.</p>
<p>Subgroup Summary</p>	
<p>This set of reports disaggregates and reports results by the following subgroups:</p> <ul style="list-style-type: none"> • All students • Disability status • Economic status • Gender • English proficiency • Primary ethnicity • Economic status <p>These reports contain no individual student-identifying information and are aggregated at the school, LEA, county, and state levels.</p> <p>For each subgroup within a report and for the total number of students, the following data are included for each test:</p> <ul style="list-style-type: none"> • Total number tested in the subgroup 	<p>This report is a resource for evaluators, researchers, teachers, parents/guardians, community members, and administrators.</p> <p>Each LEA can download this report for the whole LEA and the schools within it from TOMS.</p> <p>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.</p>

2014–15 CAASPP Student Score Reports	
Description	Use and Distribution
<ul style="list-style-type: none"> • 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 	

Chapter 8: Analyses

Background

This chapter summarizes the item- and test-level statistics obtained during the tests' original year of administration for the California Modified Assessment (CMA) for Science administered during the spring of 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 the 2012–13 administration, differential item functioning (DIF) analyses were performed based on the final item analysis (FIA) sample for all operational and field-test items to assess differences in the item performance of groups of students that differ in their demographic characteristics. In 2014–15, because intact forms were used, DIF analyses were not performed.

Each of the sets of analyses on data from the 2014–15 administration is presented in the body of the text and in the appendixes as listed below.

1. Appendix 8.A on page 95 presents the classical item analyses, including proportion-correct value (p -value) and point-biserial correlation (Pt-Bis) for each item in each operational test. Because intact forms were used, p -values and Pt-Bis are shown for both the original and the current administration of the tests. In addition, the mean and median p -value and Pt-Bis for the operational test forms based on their current administration are presented in Table 8.1 on page 79.
2. Appendix 8.B on page 99 presents results of the reliability analyses of total test scores and subscores for the population as a whole and for selected subgroups. Also presented are results of the analyses of the accuracy and consistency of the performance classifications.
3. Appendix 8.C on page 113 presents the scoring tables obtained as a result of the IRT equating process.

Samples Used for the Analyses

CMA for Science analyses were conducted at different times after test administration and involved varying proportions of the full CMA data. The classical item analyses presented in Appendix 8.A and the reliability statistics included in Appendix 8.B were calculated using the P2 data file, which contained the entire test-taking population and all the student records used as of October 28, 2015. This file contained data collected from all local educational agencies (LEAs) but did not include corrections of demographic data through the California Longitudinal Pupil Achievement Data System. In addition, students with invalid scores were excluded.

During the 2014–15 administration, neither IRT calibrations nor scaling are implemented because intact forms were reused and results were pre-equated. For the reused intact forms, the IRT results were derived based on the equating sample of the previous administration which can be found in Appendix D of the *CMA Technical Report* in the year the form was administered originally; see Table 8.4 on page 87 for administration years.

Classical Item Analyses

Multiple-Choice Items

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

$$p\text{-value}_i = \frac{N_{ic}}{N_i} \quad (8.1)$$

where,

N_{ic} is the number of examinees that answered item i correctly, and

N_i is the total number of examinees that attempted the item.

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

$$r_{X_i T} = \frac{\text{cov}(X_i, T)}{s_{X_i} s_T} \quad (8.2)$$

where,

$\text{cov}(X_i, T)$ is the covariance between the score of item i and total score T ,

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

s_T is the standard deviation for total score T .

The classical statistics for the current administration of the overall test are presented in Table 8.1. The item-by-item values for the classical statistics, including p -values, and point-biserial correlations are presented in Table 8.A.1 on page 95. Each set of values is presented for both the current and the original presentation of each CMA for Science.

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

CMA	No. of Items	No. of Examinees	Mean p -value	Mean Pt-Bis	Median p -value	Median Pt-Bis
Grade 5 Science	48	23,236	0.59	0.32	0.58	0.33
Grade 8 Science	54	19,212	0.57	0.31	0.56	0.32
Grade 10 Life Science	60	9,601	0.56	0.32	0.58	0.30

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 8.3:

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

where,

n is the number of items,

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

s_t^2 is the variance of the total score.

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

$$s_e = s_t \sqrt{1 - \alpha} \quad (8.4)$$

where,

α is the reliability estimated in equation 8.3, and

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

The SEM is particularly useful in determining the confidence interval (CI) that captures an examinee's true score. Assuming that measurement error is normally distributed, it can be said that upon infinite replications of the testing occasion, approximately 95 percent of the CIs of ± 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 ± 3.76 rounded to the nearest integer).

Table 8.2 gives the reliability and SEM for each of the CMA for Science, along with the number of items and examinees upon which those analyses were performed.

Table 8.2 Reliabilities and SEMs for the CMA for Science

CMA	No. of Items	No. of Examinees	Reliability	Mean Scale Score	Scale Std. Dev.	Scale Score SEM	Mean Raw Score	Raw Score Std. Dev.	Raw Score SEM
Grade 5 Science	48	23,236	0.82	343	57	24.25	28.26	7.39	3.17
Grade 8 Science	54	19,212	0.82	330	64	27.39	30.76	7.97	3.41
Grade 10 Life Science	60	9,601	0.85	312	63	24.39	33.51	9.18	3.58

Intercorrelations, Reliabilities, and SEMs for Reporting Clusters

For each grade-level science CMA, number-correct scores are computed for the three or four reporting clusters. The number of items within each reporting cluster is limited, and cluster scores alone should not be used in making inferences about individual students.

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

Subgroup Reliabilities and SEMs

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

Reliabilities and SEM information for the total test scores and the reporting cluster scores are reported for each subgroup analysis. Table 8.B.2 through Table 8.B.18 present the reliabilities for the subgroups based on gender, economic status, English-language fluency, and primary ethnicity. The next set of tables, Table 8.B.19 through Table 8.B.36, shows the same analyses for the subgroups based on primary ethnicity within economic status and gender within economic status. Table 8.B.37 through Table 8.B.49 present the reliabilities for subgroups based on primary disability.

Test-level reliabilities for the various subgroups are compiled in Table 8.B.50 through Table 8.B.56. The corresponding reporting cluster-level reliabilities are provided in Table 8.B.57 through Table 8.B.63.

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

Conditional Standard Errors of Measurement

As part of the IRT-based equating procedures, scale-score conversion tables and conditional standard errors of measurement (CSEMs) are produced. CSEMs for CMA 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:

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

where,

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

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

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

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

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

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

Table 8.3 Scale Score CSEM at Performance-level Cut Points

CMA	Below Basic		Basic		Proficient		Advanced	
	Min	CSEM	Min	CSEM	Min	CSEM	Min	CSEM
	SS		SS		SS		SS	
Grade 5 Science	243	24	300	22	350	23	401	26
Grade 8 Science	264	26	300	25	350	26	406	29
Grade 10 Life Science	251	23	300	23	350	24	410	28

Decision Classification Analyses

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

Decision accuracy describes the extent to which 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 test 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 test and classifications on a hypothetical alternate form using the reliability of the test and strong true-score theory.

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

Figure 8.1 Decision Accuracy for Achieving a Performance Level

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

Figure 8.2 Decision Consistency for Achieving a Performance Level

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

The results of these analyses are presented in Table 8.B.64 through Table 8.B.66 in Appendix 8.B, starting on page 111.

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

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

Validity Evidence

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

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

This section presents the evidence gathered to support the intended uses and interpretations of scores for the CMA for Science testing program. 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 CMA for Science, followed by a description and discussion of the kinds of validity evidence that have been gathered.

Purpose of the CMA for Science

As mentioned in Chapter 1, the CMA for Science comprise the California Assessment of Student Performance and Progress (CAASPP) System implementation of the remaining paper-pencil tests for students whose individualized education program or Section 504 plan require they take the CMA for Science. The purpose of the CMA for Science is to allow students with disabilities greater access to an assessment that helps measure their achievement with respect to California’s content standards that were approved in 1998 by the State Board of Education (SBE).

The Constructs to Be Measured

The CMA for Science, given in English, are designed to show how well students in grades five, eight, and ten perform relative to the California content standards in science. These content standards were approved in 1998 by the SBE; they describe what students should know and be able to do at each grade level.

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

Interpretations and Uses of the Scores Generated

Total test scores expressed as scale scores and student performance levels are generated for each grade. The total test scale score is used to draw inferences about a student’s achievement in the content area and to classify the achievement into one of five performance levels: advanced, proficient, basic, below basic, and far below basic.

Reporting cluster scores, also called subscores, are used to draw inferences about a student’s achievement in each of several specific knowledge or skill areas covered by each

test. In past years, when cluster results were reported, the results compared an individual student's percent-correct score to the average percent-correct for the state as a whole. The range of scores for students who scored proficient on the total test was also provided for each cluster using a percent-correct metric. The reference points for this range were: (1) the average percent-correct for students who received the lowest score qualifying for the proficient performance level; and (2) the average percent-correct for students who received the lowest score qualifying for the advanced performance level, minus one percent. A detailed description of the uses and applications of CMA for Science scores as used in past years is presented in Chapter 7, which starts on page 61. Note that these were not used in reporting student results to LEAs or test sites, or in Student Score Reports.

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

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

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

Intended Test Population(s)

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

Validity Evidence Collected

Evidence Based on Content

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

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

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

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

Specifications and blueprints—ETS maintains item specifications for each CMA for Science. The item specifications describe the characteristics of the items that should be written to measure each content standard. A thorough description of the specifications can be found in Chapter 3, starting on page 27. Once the items were developed and field-tested, ETS selected all CMA for Science test items to conform to the SBE-approved California content standards and test blueprints. Test blueprints for the CMA for Science were proposed by ETS and reviewed and approved by the Assessment Review Panels (ARPs), which are advisory panels to the CDE and ETS on areas related to item development for the CMA 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 CMA for Science. The test blueprints for the CMA for Science can be found on the CDE STAR CMA Blueprints Web page at <http://www.cde.ca.gov/ta/tg/sr/cmablueprints.asp>.

Item development process—A detailed description of the item development process for the CMA for Science is presented in Chapter 3, starting on page 27.

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

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

Guidelines for test construction were established with the goal of maintaining parallel forms to the greatest extent possible from year to year. Details can be found in Chapter 4, starting on page 37.

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 33.

Evidence Based on Relations to Other Variables

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

Differential Item Functioning Analyses

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

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

Table 8.4 Original Year of Administration for the CMA for Science

CMA	Year
Grade 5 Science	2011
Grade 8 Science	2012
Grade 10 Life Science	2012

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.

Evidence Based on Internal Structure

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

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

With respect to the subscores that are reported, these scores are intended to reflect examinees' knowledge and/or skill in an area that is part of the construct underlying the total test. Analyses of the intercorrelations among the subscores themselves and between the subscores and total test score can be used for studying this aspect of the construct.

Information about the internal consistency of the items on which each subscore is based is also useful to provide.

Classical Statistics

Point-biserial correlations calculated for the items in a test show the degree to which the items discriminate between students with low and high scores on a test. To the degree that the correlations are high, evidence that the items assess the same construct is provided. As shown in Table 8.1, the mean point biserial was between 0.31 and 0.32. The point biserials for the individual items in the CMA for Science are presented in Table 8.1.

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

The summaries of b -values can be found in Appendix D of the *CMA Technical Report* for the year the form was administered originally; see Table 8.4 on page 87 for administration years.

The data in Table 8.1 indicate that CMA for Science had average p -values that range from 0.56 to 0.59.

Reliability

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

Overall reliability—The reliability analyses on each of the operational CMA for Science are presented in Table 8.2. The results indicate that the reliabilities of the CMA for Science tests were moderately high, ranging from 0.82 to 0.85.

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

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

Subgroup reliabilities—The reliabilities of the operational CMA for Science are also examined for various subgroups of the examinee population that differed in their demographic characteristics. The characteristics considered are gender, ethnicity, economic status, primary disabilities, English-language fluency, and ethnicity-for-economic status. The results of these analyses can be found in Table 8.B.2 through Table 8.B.49.

Reliability of performance classifications—The methodology used for estimating the reliability of classification decisions is described in the section “Decision Classification Analyses” on page 82. The results of these analyses are presented in Table 8.B.64

through Table 8.B.66 in Appendix 8.B; these tables start on page 111. When the classifications are collapsed to below proficient versus proficient and above, the proportion of students that were classified accurately ranged from 0.87 to 0.91 across all the CMA for Science. Similarly, the proportion of students that were classified consistently ranged from 0.82 to 0.87 for students classified into below proficient versus proficient and advanced.

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

Dimensionality

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

Two factors corresponding to the English–language arts and mathematics domain were found for the CMA in these grades, as would be expected, since these tests were designed to measure different constructs.

Evidence Based on Consequences of Testing

As observed in the *Standards*, tests are usually administered “with the expectation that some benefit will be realized from the intended use of the scores” (AERA, APA, & NCME, 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 administration of the CMA for Science.

IRT Analyses

Post-Equating

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

The procedures used for post-equating the CMA for Science prior to 2012–13 involved three steps: item calibration, item parameter scaling, and production of raw-score-to-scale-score conversions using the scaled item parameters. ETS used GENASYs for the IRT item calibration and equating work. As part of this system, a proprietary version of the PARSCALE computer program (Muraki & Bock, 1995) was used and parameterized to result in one-parameter calibrations. Research at ETS has suggested that PARSCALE calibrations done in this manner produce results that are virtually identical to results based on WINSTEPS (Way, Kubiak, Henderson, & Julian, 2002). The post-equating procedures were applied to all the CMA for Science tests.

Pre-Equating

During the 2014–15 administration, because intact test forms from previous operational administrations were used without any edits or replacement of items, pre-equating was conducted prior to administration of the tests. Based on the sample invariant property of IRT, all the item parameter estimates were placed on the reference scale in their previous administrations through the post-equating procedure described above. For all CMA for

Science, the conversion tables from previous administrations when the forms were originally used are directly applied to the current administration.

Descriptions of IRT analyses such as the model-data fit analyses can be found in Chapter 8 of the original-year technical report; the results of the IRT analyses are presented in Appendix 8.D of the original-year-technical report. *CMA Technical Reports* are linked on the CDE Technical Reports and Studies Web page at <http://www.cde.ca.gov/ta/tg/sr/technicalrpts.asp>.

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

Summaries of Scaled IRT b -values

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

During the 2014–15 administration, neither IRT calibrations nor scaling are implemented, but scaled b -value parameters derived through the post-equating procedure from their previous administrations are used for pre-equating the CMA for Science. The summaries of b -values can be found in Appendix D of the *CMA Technical Report* in the year the form was administered originally; see Table 8.4 on page 87 for administration years.

Evaluation of Pre-Equating

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

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

Equating Results

During the 2014–15 administration, for all CMA for Science using intact forms without any edits, the conversion tables from their original administrations (listed in Table 8.4 on page 87) are directly applied to the current administration.

Complete raw-score-to-scale-score conversion tables for the CMA for Science administered in 2014–15 are presented in Table 8.C.1 through Table 8.C.3 starting on page 113. The raw scores and corresponding transformed scale scores are listed in those tables. The scale scores were truncated at both ends of the scale so that the minimum reported scale score was 150 and the maximum reported scale score was 600. The scale scores defining the various performance-level cut points are presented in Table 2.1, which is in Chapter 2 on page 17.

Differential Item Functioning Analyses

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

Prior to the 2012–13 administration, DIF analyses were performed based on the FIA sample and were performed on all operational items and on all field-test items for which sufficient student samples were available. DIF analyses are not implemented during the 2014–15 administration because intact forms were used and all items were evaluated for DIF during the previous administration when the forms were originally administered. These DIF results, including the specific subgroups DIF analyses for the CMA for Science, can be found in Appendix E of the *CMA Technical Report* in the year the form was administered originally; see Table 8.1 on page 87 for administration years.

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

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

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

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

$$MH\ D - DIF = -2.35 \ln[\alpha_{MH}] \quad (8.7)$$

where,

R = number right,

W = number wrong,

N = total in:

fm = focal group at ability m ,

rm = reference group at ability m , and

tm = total group at ability m .

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

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

DIF Category	Definition
A (negligible)	<ul style="list-style-type: none"> • Absolute value of MH D-DIF is not significantly different from zero, or is less than one. • Positive values are classified as “A+” and negative values as “A-.”
B (moderate)	<ul style="list-style-type: none"> • Absolute value of MH D-DIF is significantly different from zero but not from one, and is at least one; OR • Absolute value of MH D-DIF is significantly different from one, but is less than 1.5. • Positive values are classified as “B+” and negative values as “B-.”
C (large)	<ul style="list-style-type: none"> • Absolute value of MH D-DIF is significantly different from one, and is at least 1.5. • Positive values are classified as “C+” and negative values as “C-.”

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

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

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

References

- AERA, APA, & NCME. 2014. *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- California Department of Education. (2011). *California Modified Assessment technical report, spring 2011 administration*. Sacramento, CA. Retrieved from <http://www.cde.ca.gov/ta/tg/sr/documents/cmatechrpt2011.pdf>
- California Department of Education. (2012). *California Modified Assessment technical report, spring 2012 administration*. Sacramento, CA. Retrieved from <http://www.cde.ca.gov/ta/tg/sr/documents/cma12techrpt.pdf>
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York, NY: Holt.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 292–334.
- Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York, NY: Wiley.
- Dorans, N. J., & Holland, P. W. (1993). *DIF detection and description: Mantel-Haenszel and standardization*. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Educational Testing Service. (2002). *ETS standards for quality and fairness*. Princeton, NJ: Author.
- Gaffney, T., Cudeck, R., Ferrer, E., & Widaman, K. F. (2010). On the factor structure of standardized educational achievement tests. *Journal of Applied Measurement*, 11(4), 384-408.
- Gaffney, T., & Perryman, C. (2009, July). *A longitudinal look at the factor structure of educational achievement tests*. Paper presented at the meeting of the Psychometric Society, Cambridge, England.
- Holland, P. W., & Thayer, D. T. (1985). *An alternative definition of the ETS delta scale of item difficulty* (Research Report 85–43). Princeton, NJ: Educational Testing Service.
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Washington, DC: American Council on Education and National Council on Measurement in Education.
- Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classification based on test scores. *Journal of Educational Measurement*, 32, 179–97.
- Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analyses of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719–48.

- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp.13–103). New York, NY: Macmillan.
- Muraki, E., & Bock, R. D. (1995). *PARSCALE: Parameter scaling of rating data* (Computer software, version 2.2). Chicago, IL: Scientific Software, Inc.
- United States Department of Education. (2001). Elementary and Secondary Education Act (Public Law 107-11), Title VI, Chapter B, § 4, Section 6162. Retrieved from <http://www2.ed.gov/policy/elsec/leg/esea02/index.html>
- Way, W. D., Kubiak, A. T., Henderson, D., & Julian, M. W. (2002, April). *Accuracy and stability of calibrations for mixed-item-format tests using the one-parameter and generalized partial credit models*. Paper presented at the meeting of the National Council on Measurement in Education, New Orleans, LA.

Appendix 8.A—Classical Analyses

Table 8.A.1 Item-by-item p -value and Point Biserial for Science, Grade Five—Current Year (2015) and Original Year of Administration

Items	2015 p -value	2015 Pt-Bis	2011 p -value	2011 Pt-Bis
1	0.71	0.31	0.70	0.34
2	0.55	0.24	0.59	0.27
3	0.44	0.26	0.44	0.29
4	0.57	0.35	0.61	0.34
5	0.55	0.26	0.55	0.30
6	0.46	0.26	0.49	0.27
7	0.49	0.20	0.52	0.27
8	0.75	0.44	0.72	0.45
9	0.84	0.23	0.83	0.28
10	0.31	0.09	0.29	0.08
11	0.52	0.33	0.50	0.35
12	0.55	0.38	0.56	0.40
13	0.46	0.35	0.47	0.34
14	0.61	0.23	0.63	0.26
15	0.51	0.13	0.48	0.08
16	0.77	0.39	0.76	0.38
17	0.60	0.48	0.60	0.47
18	0.69	0.30	0.70	0.31
19	0.57	0.33	0.57	0.32
20	0.48	0.30	0.45	0.28
21	0.60	0.36	0.60	0.36
22	0.48	0.34	0.47	0.32
23	0.44	0.31	0.42	0.28
24	0.60	0.29	0.63	0.31
25	0.64	0.32	0.64	0.32
26	0.66	0.47	0.63	0.47
27	0.83	0.35	0.82	0.38
28	0.59	0.38	0.54	0.38
29	0.71	0.45	0.71	0.44
30	0.65	0.39	0.63	0.39
31	0.60	0.44	0.60	0.42
32	0.63	0.30	0.62	0.33
33	0.70	0.42	0.71	0.40
34	0.44	0.28	0.42	0.29
35	0.66	0.26	0.69	0.26
36	0.51	0.34	0.53	0.35
37	0.52	0.33	0.51	0.34
38	0.72	0.41	0.70	0.41
39	0.52	0.30	0.50	0.30
40	0.55	0.19	0.55	0.19
41	0.63	0.39	0.63	0.41
42	0.41	0.23	0.43	0.25
43	0.54	0.34	0.55	0.36
44	0.81	0.38	0.84	0.39
45	0.58	0.38	0.62	0.39
46	0.55	0.31	0.56	0.33
47	0.51	0.34	0.54	0.33
48	0.75	0.36	0.75	0.35

Table 8.A.2 Item-by-item p -value and Point Biserial for Science, Grade Eight—Current Year (2015) and Original Year of Administration

Items	2015 p -value	2015 Pt-Bis	2012 p -value	2012 Pt-Bis
1	0.68	0.34	0.70	0.35
2	0.78	0.25	0.77	0.29
3	0.65	0.28	0.65	0.27
4	0.64	0.25	0.63	0.22
5	0.70	0.34	0.70	0.33
6	0.50	0.25	0.50	0.27
7	0.59	0.35	0.60	0.35
8	0.55	0.32	0.57	0.33
9	0.89	0.35	0.87	0.36
10	0.59	0.36	0.60	0.36
11	0.65	0.36	0.64	0.36
12	0.52	0.32	0.52	0.34
13	0.53	0.32	0.53	0.32
14	0.57	0.35	0.58	0.39
15	0.53	0.40	0.54	0.40
16	0.52	0.21	0.51	0.19
17	0.67	0.27	0.67	0.29
18	0.60	0.21	0.61	0.23
19	0.47	0.19	0.47	0.26
20	0.53	0.35	0.53	0.37
21	0.68	0.34	0.67	0.37
22	0.56	0.21	0.57	0.22
23	0.52	0.31	0.52	0.34
24	0.49	0.36	0.50	0.39
25	0.45	0.28	0.46	0.28
26	0.43	0.25	0.45	0.29
27	0.46	0.24	0.47	0.23
28	0.46	0.31	0.45	0.30
29	0.49	0.31	0.45	0.32
30	0.56	0.36	0.58	0.37
31	0.68	0.41	0.66	0.42
32	0.68	0.46	0.68	0.46
33	0.41	0.23	0.44	0.26
34	0.52	0.22	0.52	0.23
35	0.57	0.36	0.57	0.35
36	0.41	0.19	0.41	0.21
37	0.62	0.41	0.62	0.39
38	0.73	0.41	0.75	0.41
39	0.81	0.38	0.82	0.38
40	0.58	0.20	0.56	0.20
41	0.69	0.31	0.72	0.33
42	0.60	0.36	0.61	0.38
43	0.67	0.43	0.68	0.41
44	0.50	0.39	0.54	0.40
45	0.43	0.16	0.44	0.17
46	0.56	0.35	0.54	0.35
47	0.45	0.31	0.46	0.30
48	0.51	0.28	0.50	0.33
49	0.41	0.13	0.39	0.18
50	0.60	0.32	0.59	0.32
51	0.55	0.39	0.54	0.37
52	0.34	0.11	0.35	0.14
53	0.54	0.33	0.55	0.34
54	0.66	0.39	0.64	0.40

Table 8.A.3 Item-by-item p -value and Point Biserial for Science, Grade Ten—Current Year (2015) and Original Year of Administration

Items	2015 p -value	2015 Pt-Bis	2012 p -value	2012 Pt-Bis
1	0.74	0.24	0.69	0.29
2	0.42	0.29	0.43	0.27
3	0.62	0.22	0.59	0.24
4	0.72	0.22	0.69	0.28
5	0.49	0.21	0.47	0.22
6	0.42	0.27	0.38	0.24
7	0.31	-0.08	0.31	-0.05
8	0.48	0.28	0.46	0.25
9	0.45	0.20	0.42	0.19
10	0.58	0.27	0.55	0.30
11	0.43	0.24	0.42	0.19
12	0.65	0.25	0.60	0.29
13	0.62	0.21	0.59	0.25
14	0.50	0.26	0.49	0.26
15	0.55	0.29	0.52	0.26
16	0.46	0.31	0.44	0.29
17	0.59	0.35	0.58	0.36
18	0.49	0.30	0.47	0.31
19	0.49	0.30	0.48	0.30
20	0.57	0.35	0.54	0.33
21	0.66	0.43	0.62	0.40
22	0.53	0.27	0.50	0.26
23	0.67	0.39	0.63	0.38
24	0.65	0.35	0.61	0.36
25	0.59	0.42	0.51	0.39
26	0.67	0.44	0.61	0.43
27	0.53	0.30	0.50	0.32
28	0.72	0.41	0.69	0.42
29	0.71	0.46	0.65	0.46
30	0.64	0.42	0.61	0.38
31	0.54	0.25	0.51	0.27
32	0.51	0.35	0.51	0.35
33	0.63	0.49	0.61	0.49
34	0.59	0.44	0.55	0.41
35	0.61	0.39	0.58	0.39
36	0.47	0.30	0.44	0.26
37	0.63	0.43	0.59	0.44
38	0.49	0.24	0.47	0.25
39	0.67	0.41	0.65	0.42
40	0.68	0.53	0.63	0.54
41	0.58	0.41	0.56	0.39
42	0.54	0.30	0.51	0.30
43	0.66	0.44	0.65	0.41
44	0.47	0.29	0.45	0.27
45	0.32	0.13	0.33	0.14
46	0.72	0.42	0.67	0.43
47	0.47	0.24	0.46	0.23
48	0.45	0.29	0.42	0.25
49	0.73	0.37	0.70	0.37
50	0.28	0.14	0.26	0.08
51	0.64	0.36	0.62	0.34
52	0.56	0.31	0.47	0.34
53	0.36	0.17	0.36	0.15
54	0.62	0.40	0.56	0.38

Items	2015 <i>p</i>-value	2015 Pt-Bis	2012 <i>p</i>-value	2012 Pt-Bis
55	0.42	0.28	0.39	0.26
56	0.72	0.47	0.66	0.47
57	0.43	0.23	0.41	0.21
58	0.42	0.36	0.42	0.30
59	0.67	0.42	0.62	0.41
60	0.63	0.32	0.60	0.29

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 Subscore Reliabilities and Intercorrelations for Science

Subscore Area	No. of items	Intercorrelations				Reliab.	SEM
		1.	2.	3.	4.		
Grade 5							
1. Physical Sciences	16	1.00	–	–		0.51	1.86
2. Life Sciences	16	0.54	1.00	–		0.69	1.78
3. Earth Sciences	16	0.49	0.60	1.00		0.62	1.83
Grade 8							
1. Motion	19	1.00	–	–	–	0.60	2.00
2. Matter	23	0.53	1.00	–	–	0.65	2.27
3. Earth Science	7	0.46	0.49	1.00	–	0.54	1.13
4. Investigation and Experimentation	5	0.48	0.47	0.36	1.00	0.39	1.04
Grade 10							
1. Cell Biology and Genetics	22	1.00	–	–	–	0.58	2.21
2. Evolution and Ecology	22	0.58	1.00	–	–	0.70	2.13
3. Physiology	10	0.56	0.62	1.00	–	0.63	1.41
4. Investigation and Experimentation	6	0.46	0.52	0.47	1.00	0.43	1.14

Table 8.B.2 Reliabilities and SEMs for the CMA for Science by Gender (Male)

CMA	N	Rel	SEM
Grade 5 Science	15,312	0.82	3.16
Grade 8 Science	12,788	0.83	3.39
Grade 10 Life Science	6,386	0.86	3.57

Table 8.B.3 Reliabilities and SEMs for the CMA for Science by Gender (Female)

CMA	N	Rel	SEM
Grade 5 Science	7,924	0.80	3.20
Grade 8 Science	6,424	0.79	3.44
Grade 10 Life Science	3,215	0.81	3.60

Table 8.B.4 Reliabilities and SEMs for the CMA for Science by Economic Status (Economically Disadvantaged)

CMA	N	Rel	SEM
Grade 5 Science	18,166	0.80	3.20
Grade 8 Science	14,808	0.81	3.43
Grade 10 Life Science	7,306	0.84	3.60

Table 8.B.5 Reliabilities and SEMs for the CMA for Science by Economic Status (Not Economically Disadvantaged)

CMA	N	Rel	SEM
Grade 5 Science	5,070	0.84	3.08
Grade 8 Science	4,404	0.84	3.35
Grade 10 Life Science	2,295	0.86	3.52

Table 8.B.6 Reliabilities and SEMs for the CMA for Science by English-language Fluency (English Only)

CMA	N	Rel	SEM
Grade 5 Science	12,266	0.83	3.13
Grade 8 Science	9,979	0.83	3.39
Grade 10 Life Science	4,888	0.86	3.55

Table 8.B.7 Reliabilities and SEMs for the CMA for Science by English-language Fluency (Initially Fluent English Proficient)

CMA	N	Rel	SEM
Grade 5 Science	197	0.81	3.07
Grade 8 Science	218	0.85	3.33
Grade 10 Life Science	198	0.87	3.51

Table 8.B.8 Reliabilities and SEMs for the CMA for Science by English-language Fluency (English Learner)

CMA	N	Rel	SEM
Grade 5 Science	10,361	0.79	3.22
Grade 8 Science	7,527	0.77	3.47
Grade 10 Life Science	3,608	0.80	3.66

Table 8.B.9 Reliabilities and SEMs for the CMA for Science by English-language Fluency (Reclassified Fluent English Proficient)

CMA	N	Rel	SEM
Grade 5 Science	402	0.83	3.08
Grade 8 Science	1,481	0.83	3.28
Grade 10 Life Science	898	0.86	3.42

Table 8.B.10 Reliabilities and SEMs for the CMA Science by English-language Fluency (Unknown)

CMA	N	Rel	SEM
Grade 5 Science	9	–	–
Grade 8 Science	7	–	–
Grade 10 Life Science	7	–	–

Table 8.B.11 Reliabilities and SEMs for the CMA for Science by Primary Ethnicity (African American)

CMA	N	Reliability	SEM
Grade 5 Science	2,182	0.80	3.23
Grade 8 Science	2,025	0.79	3.47
Grade10 Life Science	882	0.83	3.64

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

CMA	N	Reliability	SEM
Grade 5 Science	181	0.79	3.14
Grade 8 Science	164	0.79	3.39
Grade10 Life Science	82	0.83	3.54

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

CMA	N	Reliability	SEM
Grade 5 Science	747	0.84	3.17
Grade 8 Science	527	0.86	3.35
Grade10 Life Science	220	0.84	3.58

Table 8.B.14 Reliabilities and SEMs for the CMA for Science by Primary Ethnicity (Filipino)

CMA	N	Reliability	SEM
Grade 5 Science	200	0.78	3.18
Grade 8 Science	216	0.79	3.42
Grade10 Life Science	92	0.84	3.53

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

CMA	N	Reliability	SEM
Grade 5 Science	15,330	0.80	3.20
Grade 8 Science	12,703	0.80	3.43
Grade10 Life Science	6,363	0.84	3.60

Table 8.B.16 Reliabilities and SEMs for the CMA for Science by Primary Ethnicity (Pacific Islander)

CMA	N	Reliability	SEM
Grade 5 Science	99	0.79	3.19
Grade 8 Science	79	0.82	3.41
Grade 10 Life Science	28	0.89	3.48

Table 8.B.17 Reliabilities and SEMs for the CMA for Science by Primary Ethnicity (White)

CMA	N	Reliability	SEM
Grade 5 Science	3,962	0.84	3.04
Grade 8 Science	3,109	0.85	3.31
Grade 10 Life Science	1,725	0.87	3.48

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

CMA	N	Reliability	SEM
Grade 5 Science	535	0.84	3.10
Grade 8 Science	389	0.82	3.39
Grade 10 Life Science	209	0.87	3.51

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

CMA	N	Reliability	SEM
Grade 5 Science	418	0.80	3.21
Grade 8 Science	450	0.83	3.44
Grade 10 Life Science	194	0.84	3.60

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

CMA	N	Reliability	SEM
Grade 5 Science	35	0.80	3.12
Grade 8 Science	41	0.82	3.37
Grade 10 Life Science	27	0.82	3.52

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

CMA	N	Reliability	SEM
Grade 5 Science	337	0.86	3.14
Grade 8 Science	208	0.85	3.33
Grade 10 Life Science	83	0.85	3.55

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

CMA	N	Reliability	SEM
Grade 5 Science	125	0.78	3.15
Grade 8 Science	131	0.79	3.43
Grade 10 Life Science	56	0.85	3.49

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

CMA	N	Reliability	SEM
Grade 5 Science	2,006	0.82	3.13
Grade 8 Science	1,847	0.82	3.39
Grade 10 Life Science	978	0.85	3.56

Table 8.B.24 Reliabilities and SEMs for the CMA for Science by Primary Ethnicity-for-Not-Economically-Disadvantaged (Pacific Islander)

CMA	N	Reliability	SEM
Grade 5 Science	24	0.79	3.16
Grade 8 Science	20	0.86	3.27
Grade10 Life Science	12	0.90	3.27

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

CMA	N	Reliability	SEM
Grade 5 Science	1,941	0.84	2.99
Grade 8 Science	1,574	0.85	3.27
Grade10 Life Science	858	0.87	3.44

Table 8.B.26 Reliabilities and SEMs for the CMA for Science by Primary Ethnicity-for-Not-Economically-Disadvantaged (Unknown)

CMA	N	Reliability	SEM
Grade 5 Science	184	0.86	184
Grade 8 Science	133	0.84	133
Grade10 Life Science	87	0.86	87

Table 8.B.27 Reliabilities and SEMs for the CMA for Science by Primary Ethnicity-for-Economically-Disadvantaged (African American)

CMA	N	Reliability	SEM
Grade 5 Science	1,764	0.79	3.23
Grade 8 Science	1,575	0.78	3.48
Grade 10 Life Science	688	0.83	3.65

Table 8.B.28 Reliabilities and SEMs for the CMA for Science by Primary Ethnicity-for-Economically-Disadvantaged (American Indian)

CMA	N	Reliability	SEM
Grade 5 Science	146	0.79	3.15
Grade 8 Science	123	0.78	3.40
Grade 10 Life Science	55	0.84	3.55

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

CMA	N	Reliability	SEM
Grade 5 Science	410	0.82	3.19
Grade 8 Science	319	0.86	3.37
Grade 10 Life Science	137	0.83	3.60

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

CMA	N	Reliability	SEM
Grade 5 Science	75	0.76	3.24
Grade 8 Science	85	0.80	3.40
Grade10 Life Science	36	0.81	3.61

Table 8.B.31 Reliabilities and SEMs for the CMA for Science by Primary Ethnicity-for-Economically-Disadvantaged (Hispanic)

CMA	N	Reliability	SEM
Grade 5 Science	13,324	0.80	3.21
Grade 8 Science	10,856	0.80	3.44
Grade10 Life Science	5,385	0.83	3.61

Table 8.B.32 Reliabilities and SEMs for the CMA for Science by Primary Ethnicity-for-Economically-Disadvantaged (Pacific Islander)

CMA	N	Reliability	SEM
Grade 5 Science	75	0.78	3.20
Grade 8 Science	59	0.78	3.46
Grade 10 Life Science	16	0.82	3.64

Table 8.B.33 Reliabilities and SEMs for the CMA for Science by Primary Ethnicity-for-Economically-Disadvantaged (White)

CMA	N	Reliability	SEM
Grade 5 Science	2,021	0.83	3.09
Grade 8 Science	1,535	0.84	3.35
Grade 10 Life Science	867	0.87	3.52

Table 8.B.34 Reliabilities and SEMs for the CMA for Science by Primary Ethnicity-for-Economically-Disadvantaged (Unknown)

CMA	N	Reliability	SEM
Grade 5 Science	351	0.82	3.12
Grade 8 Science	256	0.80	3.42
Grade 10 Life Science	122	0.87	3.54

Table 8.B.35 Reliabilities and SEMs for the CMA for Science by Gender by Economic Status (Economically Disadvantaged)

CMA	Male N	Male Reliability	Male SEM	Female N	Female Reliability	Female SEM
Grade 5 Science	11,882	0.81	3.18	6,284	0.79	3.22
Grade 8 Science	9,812	0.82	3.41	4,996	0.78	3.46
Grade 10 Life Science	4,851	0.86	3.58	2,455	0.80	3.63

Table 8.B.36 Reliabilities and SEMs for the CMA for Science by Gender by Economic Status (Not Economically Disadvantaged)

CMA	Male N	Male Reliability	Male SEM	Female N	Female Reliability	Female SEM
Grade 5 Science	3,430	0.84	3.07	1,640	0.82	3.11
Grade 8 Science	2,976	0.85	3.33	1,428	0.82	3.38
Grade 10 Life Science	1,535	0.87	3.51	760	0.84	3.53

Table 8.B.37 Reliabilities and SEMs for the CMA for Science by Primary Disability (Autism)

CMA	N	Reliability	SEM
Grade 5 Science	1,762	0.84	3.16
Grade 8 Science	1,309	0.86	3.33
Grade 10 Life Science	599	0.88	3.48

Table 8.B.38 Reliabilities and SEMs for the CMA for Science by Primary Disability (Deaf-Blindness)

CMA	N	Reliability	SEM
Grade 5 Science	2	–	–
Grade 8 Science	4	–	–
Grade 10 Life Science	3	–	–

Table 8.B.39 Reliabilities and SEMs for the CMA for Science by Primary Disability (Emotional Disturbance)

CMA	N	Reliability	SEM
Grade 5 Science	356	0.84	3.15
Grade 8 Science	456	0.83	3.41
Grade 10 Life Science	300	0.87	3.53

Table 8.B.40 Reliabilities and SEMs for the CMA for Science by Primary Disability (Hearing Impairment)

CMA	N	Reliability	SEM
Grade 5 Science	244	0.81	3.22
Grade 8 Science	239	0.79	3.45
Grade 10 Life Science	175	0.81	3.63

Table 8.B.41 Reliabilities and SEMs for the CMA for Science by Primary Disability (Mental Retardation)

CMA	N	Reliability	SEM
Grade 5 Science	405	0.70	3.31
Grade 8 Science	409	0.69	3.53
Grade 10 Life Science	235	0.67	3.73

Table 8.B.42 Reliabilities and SEMs for the CMA for Science by Primary Disability (Multiple Disabilities)

CMA	N	Reliability	SEM
Grade 5 Science	26	0.82	3.23
Grade 8 Science	25	0.81	3.49
Grade 10 Life Science	20	0.83	3.64

Table 8.B.43 Reliabilities and SEMs for the CMA for Science by Primary Disability (Orthopedic Impairment)

CMA	N	Reliability	SEM
Grade 5 Science	129	0.82	3.21
Grade 8 Science	124	0.82	3.43
Grade 10 Life Science	73	0.82	3.61

Table 8.B.44 Reliabilities and SEMs for the CMA for Science by Primary Disability (Other Health Impairment)

CMA	N	Reliability	SEM
Grade 5 Science	2,629	0.83	3.12
Grade 8 Science	2,079	0.83	3.39
Grade 10 Life Science	951	0.87	3.54

Table 8.B.45 Reliabilities and SEMs for the CMA for Science by Primary Disability (Specific Learning Disability)

CMA	N	Reliability	SEM
Grade 5 Science	14,209	0.81	3.17
Grade 8 Science	12,668	0.81	3.41
Grade 10 Life Science	6,402	0.84	3.59

Table 8.B.46 Reliabilities and SEMs for the CMA for Science by Primary Disability (Speech or Language Impairment)

CMA	N	Reliability	SEM
Grade 5 Science	2,036	0.80	3.21
Grade 8 Science	965	0.80	3.43
Grade 10 Life Science	328	0.82	3.60

Table 8.B.47 Reliabilities and SEMs for the CMA for Science by Primary Disability (Traumatic Brain Injury)

CMA	N	Reliability	SEM
Grade 5 Science	39	0.80	3.17
Grade 8 Science	38	0.75	3.50
Grade 10 Life Science	20	0.88	3.56

Table 8.B.48 Reliabilities and SEMs for the CMA for Science by Primary Disability (Visual Impairment)

CMA	N	Reliability	SEM
Grade 5 Science	44	0.82	3.19
Grade 8 Science	47	0.80	3.47
Grade 10 Life Science	20	0.90	3.45

Table 8.B.49 Reliabilities and SEMs for the CMA for Science by Primary Disability (Unknown)

CMA	N	Reliability	SEM
Grade 5 Science	1,355	0.82	3.16
Grade 8 Science	849	0.81	3.44
Grade 10 Life Science	475	0.86	3.58

Table 8.B.50 Overall Subgroup Reliabilities

CMA	Gender		Econ. Dis.		Language Fluency			
	Male	Female	No	Yes	EO	I-FEP	EL	R-FEP
Grade 5 Science	0.82	0.80	0.84	0.80	0.83	0.81	0.79	0.83
Grade 8 Science	0.83	0.79	0.84	0.81	0.83	0.85	0.77	0.83
Grade 10 Life Science	0.86	0.81	0.86	0.84	0.86	0.87	0.80	0.86

Table 8.B.51 Overall Subgroup Reliabilities—Primary Ethnicity

CMA	African American	American Indian	Asian	Filipino	Hispanic	Pacific Islander	White
Grade 5 Science	0.80	0.79	0.84	0.78	0.80	0.79	0.84
Grade 8 Science	0.79	0.79	0.86	0.79	0.80	0.82	0.85
Grade 10 Life Science	0.83	0.83	0.84	0.84	0.84	0.89	0.87

Table 8.B.52 Overall Subgroup Reliabilities by Primary Ethnicity—Not Economically Disadvantaged

CMA	African American	American Indian	Asian	Filipino	Hispanic	Pacific Islander	White
Grade 5 Science	0.80	0.80	0.86	0.78	0.82	0.79	0.84
Grade 8 Science	0.83	0.82	0.85	0.79	0.82	0.86	0.85
Grade 10 Life Science	0.84	0.82	0.85	0.85	0.85	0.90	0.87

Table 8.B.53 Overall Subgroup Reliabilities by Primary Ethnicity—Economically Disadvantaged

CMA	African American	American Indian	Asian	Filipino	Hispanic	Pacific Islander	White
Grade 5 Science	0.79	0.79	0.82	0.76	0.80	0.78	0.83
Grade 8 Science	0.78	0.78	0.86	0.80	0.80	0.78	0.84
Grade 10 Life Science	0.83	0.84	0.83	0.81	0.83	0.82	0.87

Table 8.B.54 Overall Subgroup Reliabilities by Gender/Economic Status

CMA	Economically Disadvantaged		Not Economically Disadvantaged	
	Male	Female	Male	Female
Grade 5 Science	0.81	0.79	0.84	0.82
Grade 8 Science	0.82	0.78	0.85	0.82
Grade 10 Life Science	0.86	0.80	0.87	0.84

Table 8.B.55 Overall Subgroup Reliabilities by Primary Disability

CMA	Autism	Deaf– Blindness	Emotional Dist.	Hearing	Mental Retard.	Mult. Disab.
Grade 8 Science	0.86	–	0.83	0.79	0.69	0.81
Grade 10 Life Science	0.88	–	0.87	0.81	0.67	0.83

Table 8.B.56 Overall Subgroup Reliabilities by Primary Disability (continued)

CMA	Orthoped. Impair.	Other Health Impair.	Specific Lrn Disab.	Speech or Lang Impair.	Traumatic Brain Injury	Visual Impair.
Grade 5 Science	0.82	0.83	0.81	0.80	0.80	0.82
Grade 8 Science	0.82	0.83	0.81	0.80	0.75	0.80
Grade 10 Life Science	0.82	0.87	0.84	0.82	0.88	0.90

Table 8.B.57 Subscore Reliabilities and SEM for Science by Gender/Economic Status

Subscore Area	N of Items	Male		Female		Not Econ. Dis.		Econ. Dis.		
		Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	
Grade 5 Science										
1. Physical Sciences	16	0.52	1.85	0.46	1.88	0.49	1.87	0.55	1.82	
2. Life Sciences	16	0.70	1.77	0.67	1.79	0.68	1.80	0.72	1.71	
3. Earth Sciences	16	0.63	1.82	0.62	1.83	0.61	1.84	0.65	1.78	
Grade 8 Science										
1. Motion	19	0.61	1.98	0.56	2.03	0.58	2.01	0.63	1.96	
2. Matter	23	0.66	2.27	0.60	2.28	0.63	2.28	0.68	2.24	
3. Earth Science	7	0.58	1.12	0.46	1.16	0.53	1.15	0.57	1.09	
4. Investigation and Experimentation	5	0.40	1.04	0.36	1.05	0.37	1.05	0.42	1.02	
Grade 10 Life Science										
1. Cell Biology and Genetics	22	0.61	2.20	0.53	2.21	0.57	2.22	0.61	2.18	
2. Evolution and Ecology	22	0.72	2.13	0.65	2.14	0.69	2.15	0.73	2.09	
3. Physiology	10	0.65	1.40	0.57	1.44	0.62	1.42	0.64	1.37	
4. Investigation and Experimentation	6	0.45	1.13	0.39	1.15	0.42	1.15	0.45	1.12	

Table 8.B.58 Subscore Reliabilities and SEM for Science by English-language Fluency

Subscore Area	N of Items	English Learner		English Only		I-FEP		R-FEP	
		Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM
Grade 5 Science									
1. Physical Sciences	16	0.46	1.88	0.53	1.84	0.47	1.83	0.54	1.82
2. Life Sciences	16	0.66	1.82	0.70	1.75	0.66	1.70	0.70	1.71
3. Earth Sciences	16	0.59	1.85	0.64	1.81	0.60	1.78	0.65	1.77
Grade 8 Science									
1. Motion	19	0.55	2.03	0.61	1.98	0.68	1.93	0.58	1.92
2. Matter	23	0.56	2.30	0.67	2.26	0.68	2.24	0.69	2.21
3. Earth Science	7	0.51	1.17	0.56	1.12	0.60	1.09	0.54	1.04
4. Investigation and Experimentation	5	0.33	1.06	0.40	1.03	0.42	1.03	0.43	0.99
Grade 10 Life Science									
1. Cell Biology and Genetics	22	0.50	2.24	0.60	2.20	0.58	2.17	0.61	2.14
2. Evolution and Ecology	22	0.63	2.18	0.72	2.12	0.77	2.08	0.70	2.05
3. Physiology	10	0.55	1.47	0.64	1.39	0.69	1.35	0.64	1.31
4. Investigation and Experimentation	6	0.32	1.17	0.45	1.13	0.41	1.12	0.50	1.07

Table 8.B.59 Subscore Reliabilities and SEM for Science by Primary Ethnicity

Subscore Area	N of Items	African American		American Indian		Asian		Filipino		Hispanic		Pacific Islander		White	
		Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM
Science Grade 5															
1. Physical Sciences	16	0.45	1.88	0.46	1.83	0.55	1.86	0.45	1.86	0.48	1.87	0.45	1.89	0.56	1.80
2. Life Sciences	16	0.68	1.81	0.65	1.77	0.73	1.78	0.63	1.81	0.67	1.80	0.64	1.80	0.71	1.68
3. Earth Sciences	16	0.60	1.87	0.60	1.82	0.65	1.82	0.63	1.80	0.60	1.84	0.66	1.80	0.66	1.76
Science Grade 8															
1. Motion	19	0.58	2.04	0.56	1.98	0.63	1.98	0.58	2.01	0.58	2.01	0.63	2.00	0.64	1.93
2. Matter	23	0.61	2.29	0.62	2.27	0.73	2.21	0.59	2.28	0.62	2.28	0.70	2.24	0.70	2.22
3. Earth Science	7	0.54	1.17	0.48	1.12	0.59	1.12	0.50	1.14	0.53	1.14	0.43	1.17	0.58	1.07
4. Investigation and Experimentation	5	0.30	1.06	0.27	1.05	0.50	1.00	0.33	1.05	0.37	1.05	0.29	1.06	0.44	1.01
Grade 10 Life Science															
1. Cell Biology and Genetics	22	0.56	2.23	0.47	2.21	0.62	2.21	0.53	2.22	0.56	2.22	0.68	2.14	0.62	2.17
2. Evolution and Ecology	22	0.68	2.18	0.69	2.10	0.65	2.15	0.63	2.09	0.68	2.15	0.75	2.09	0.75	2.07
3. Physiology	10	0.62	1.44	0.69	1.35	0.60	1.41	0.58	1.41	0.61	1.43	0.75	1.36	0.66	1.34
4. Investigation and Experimentation	6	0.40	1.16	0.43	1.11	0.45	1.11	0.46	1.11	0.41	1.15	0.37	1.11	0.47	1.11

Table 8.B.60 Subscore Reliabilities and SEM for Science by Primary Ethnicity-for-Not Economically Disadvantaged

Subscore Area	N of Items	African American		American Indian		Asian		Filipino		Hispanic		Pacific Islander		White	
		Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM
Grade 5 Science															
1. Physical Sciences	16	0.39	1.89	0.50	1.84	0.55	1.86	0.52	1.83	0.52	1.85	0.28	1.89	0.58	1.77
2. Life Sciences	16	0.70	1.79	0.64	1.78	0.77	1.74	0.65	1.78	0.69	1.75	0.68	1.77	0.70	1.65
3. Earth Sciences	16	0.58	1.86	0.50	1.81	0.67	1.80	0.56	1.80	0.60	1.81	0.69	1.77	0.67	1.73
Grade 8 Science															
1. Motion	19	0.63	2.02	0.57	1.97	0.61	1.97	0.57	2.02	0.59	1.98	0.63	1.90	0.65	1.90
2. Matter	23	0.65	2.28	0.65	2.26	0.72	2.20	0.57	2.30	0.64	2.27	0.77	2.17	0.70	2.21
3. Earth Science	7	0.62	1.14	0.48	1.13	0.60	1.10	0.46	1.15	0.53	1.12	0.45	1.11	0.58	1.04
4. Investigation and Experimentation	5	0.36	1.05	0.43	1.01	0.43	1.02	0.38	1.03	0.39	1.04	0.56	0.97	0.45	0.99
Grade 10 Life Science															
1. Cell Biology and Genetics	22	0.54	2.23	0.53	2.13	0.63	2.20	0.55	2.20	0.59	2.20	0.71	2.00	0.63	2.15
2. Evolution and Ecology	22	0.72	2.14	0.69	2.13	0.69	2.12	0.66	2.04	0.70	2.13	0.80	1.94	0.74	2.05
3. Physiology	10	0.64	1.41	0.74	1.29	0.49	1.43	0.61	1.41	0.62	1.40	0.73	1.34	0.66	1.33
4. Investigation and Experimentation	6	0.43	1.15	-0.17	1.21	0.34	1.14	0.49	1.09	0.44	1.13	0.41	1.01	0.48	1.09

Note: The small case count and the small number of items in the reporting cluster both contribute to the negative value found for the subgroup “American Indian” and the subscore area “Investigation and Experimentation.”

Table 8.B.61 Subscore Reliabilities and SEM for Science by Primary Ethnicity-for-Economically Disadvantaged

Subscore Area	N of Items	African American		American Indian		Asian		Filipino		Hispanic		Pacific Islander		White	
		Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM
Grade 5 Science															
1. Physical Sciences	16	0.47	1.88	0.44	1.84	0.55	1.86	0.32	1.90	0.47	1.88	0.48	1.88	0.53	1.82
2. Life Sciences	16	0.68	1.82	0.65	1.77	0.68	1.81	0.59	1.87	0.67	1.81	0.62	1.81	0.71	1.71
3. Earth Sciences	16	0.60	1.87	0.62	1.82	0.63	1.83	0.68	1.81	0.60	1.84	0.65	1.80	0.64	1.79
Grade 8 Science															
1. Motion	19	0.56	2.04	0.56	1.99	0.63	1.99	0.60	2.00	0.57	2.01	0.61	2.04	0.61	1.96
2. Matter	23	0.59	2.29	0.61	2.27	0.72	2.22	0.61	2.25	0.61	2.28	0.63	2.27	0.69	2.24
3. Earth Science	7	0.52	1.18	0.48	1.11	0.59	1.13	0.56	1.11	0.53	1.15	0.41	1.18	0.57	1.10
4. Investigation and Experimentation	5	0.29	1.07	0.20	1.07	0.54	0.99	0.24	1.08	0.37	1.05	0.14	1.09	0.43	1.02
Grade 10 Life Science															
1. Cell Biology and Genetics	22	0.56	2.23	0.45	2.24	0.61	2.21	0.48	2.23	0.55	2.22	0.48	2.23	0.60	2.19
2. Evolution and Ecology	22	0.67	2.19	0.70	2.10	0.63	2.16	0.57	2.17	0.68	2.15	0.61	2.21	0.75	2.09
3. Physiology	10	0.61	1.44	0.65	1.38	0.64	1.41	0.54	1.42	0.61	1.43	0.76	1.34	0.66	1.36
4. Investigation and Experimentation	6	0.39	1.16	0.57	1.06	0.51	1.10	0.40	1.15	0.40	1.15	0.15	1.21	0.46	1.12

Table 8.B.62 Subscore Reliabilities and SEM for Science by Disability

Subscore Area	N of Items	Autism		Emotional Disturbance		Hearing impairment		Mental Retardation		Multiple Disability	
		Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM
Grade 5											
1. Physical Science	16	0.54	1.86	0.55	1.85	0.50	1.89	0.40	1.91	0.50	1.89
2. Life Science	16	0.73	1.78	0.74	1.74	0.65	1.84	0.55	1.90	0.59	1.86
3. Earth Science	16	0.68	1.81	0.66	1.83	0.65	1.84	0.45	1.91	0.64	1.86
Grade 8											
1. Motion	19	0.67	1.96	0.64	1.99	0.57	2.04	0.50	2.10	0.35	2.12
2. Matter	23	0.72	2.22	0.65	2.27	0.60	2.27	0.41	2.30	0.66	2.26
3. Earth Science	7	0.56	1.09	0.59	1.13	0.43	1.18	0.29	1.27	0.62	1.18
4. Investigation and Experimentation	5	0.47	1.01	0.32	1.06	0.46	1.02	0.13	1.06	0.40	1.04
Life Science											
1. Cell Biology	22	0.66	2.18	0.61	2.19	0.50	2.23	0.40	2.26	0.48	2.25
2. Evolution	22	0.74	2.06	0.74	2.12	0.67	2.15	0.45	2.24	0.66	2.14
3. Physiology	10	0.66	1.35	0.69	1.36	0.54	1.46	0.48	1.50	0.60	1.49
4. Investigation and Experimentation	6	0.53	1.09	0.51	1.11	0.36	1.16	0.15	1.18	0.33	1.16

Table 8.B.63 Subscore Reliabilities and SEM for Science by Disability (continued)

Subscore Area	N of Items	Orthopedic Impairment		Other Health Impairment		Specific Learning Disability		Speech or Language Impairment		Traumatic Brain Injury		Visual Impairment	
		Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM
Grade 5													
1. Physical Science	16	0.45	1.90	0.53	1.83	0.50	1.86	0.46	1.88	0.51	1.86	0.59	1.86
2. Life Science	16	0.68	1.81	0.71	1.74	0.68	1.78	0.68	1.81	0.69	1.77	0.66	1.80
3. Earth Science	16	0.68	1.83	0.63	1.80	0.61	1.83	0.60	1.84	0.59	1.82	0.67	1.83
Grade 8													
1. Motion	19	0.57	2.01	0.60	1.98	0.58	2.00	0.57	2.02	0.51	2.04	0.61	2.02
2. Matter	23	0.63	2.30	0.68	2.25	0.63	2.28	0.61	2.28	0.35	2.34	0.62	2.30
3. Earth Science	7	0.57	1.12	0.57	1.12	0.54	1.13	0.51	1.15	0.52	1.20	0.53	1.19
4. Investigation and Experimentation	5	0.49	1.01	0.39	1.04	0.38	1.04	0.36	1.05	0.39	1.05	0.45	1.03
Life Science													
1. Cell Biology	22	0.56	2.23	0.63	2.19	0.56	2.21	0.54	2.21	0.61	2.22	0.74	2.11
2. Evolution	22	0.66	2.15	0.74	2.12	0.69	2.14	0.67	2.14	0.76	2.07	0.84	2.04
3. Physiology	10	0.53	1.44	0.69	1.36	0.61	1.42	0.57	1.44	0.62	1.45	0.67	1.37
4. Investigation and Experimentation	6	0.23	1.18	0.47	1.13	0.41	1.14	0.37	1.15	0.29	1.17	0.34	1.12

Table 8.B.64 Reliability of Classification for Science, Grade Five

	Placement Score	Far Below Basic	Below Basic	Basic	Proficient	Advanced	Category Total
Decision Accuracy	0–14	0.00	0.02	0.00	0.00	0.00	0.03
	15–22	0.00	0.15	0.06	0.00	0.00	0.21
	23–29	0.00	0.04	0.21	0.06	0.00	0.31
	30–36	0.00	0.00	0.07	0.21	0.03	0.30
All-forms Average	37–48	0.00	0.00	0.00	0.05	0.10	0.14
	Estimated Proportion Correctly Classified: Total = 0.66, Proficient & Above = 0.87						
Decision Consistency	0–14	0.01	0.02	0.00	0.00	0.00	0.03
	15–22	0.02	0.12	0.07	0.01	0.00	0.21
	23–29	0.00	0.07	0.17	0.08	0.00	0.31
	30–36	0.00	0.01	0.08	0.16	0.05	0.30
Alternate Form	37–48	0.00	0.00	0.00	0.05	0.09	0.14
	Estimated Proportion Consistently Classified: Total = 0.55, Proficient & Above = 0.82						

Table 8.B.65 Reliability of Classification for Science, Grade Eight

	Placement Score	Far Below Basic	Below Basic	Basic	Proficient	Advanced	Category Total
Decision Accuracy	0–21	0.07	0.05	0.01	0.00	0.00	0.13
	22–26	0.03	0.10	0.06	0.00	0.00	0.19
	27–33	0.00	0.05	0.20	0.05	0.00	0.31
	34–40	0.00	0.00	0.06	0.15	0.02	0.24
All-forms Average	41–54	0.00	0.00	0.00	0.04	0.09	0.13
	Estimated Proportion Correctly Classified: Total = 0.62, Proficient & Above = 0.88						
Decision Consistency	0–21	0.07	0.05	0.02	0.00	0.00	0.13
	22–26	0.05	0.07	0.06	0.01	0.00	0.19
	27–33	0.02	0.06	0.16	0.07	0.00	0.31
	34–40	0.00	0.01	0.07	0.12	0.04	0.24
Alternate Form	41–54	0.00	0.00	0.01	0.04	0.08	0.13
	Estimated Proportion Consistently Classified: Total = 0.50, Proficient & Above = 0.83						

Table 8.B.66 Reliability of Classification for Life Science (Grade 10)

	Placement Score	Far Below Basic	Below Basic	Basic	Proficient	Advanced	Category Total
Decision Accuracy	0–23	0.10	0.05	0.00	0.00	0.00	0.15
	24–31	0.04	0.19	0.05	0.00	0.00	0.28
	32–39	0.00	0.06	0.19	0.04	0.00	0.28
	40–47	0.00	0.00	0.05	0.14	0.02	0.20
All-forms Average	48–60	0.00	0.00	0.00	0.03	0.05	0.07
	Estimated Proportion Correctly Classified: Total = 0.66, Proficient & Above = 0.91						
Decision Consistency	0–23	0.09	0.06	0.01	0.00	0.00	0.15
	24–31	0.06	0.15	0.07	0.00	0.00	0.28
	32–39	0.01	0.07	0.15	0.06	0.00	0.28
	40–47	0.00	0.01	0.06	0.11	0.03	0.20
Alternate Form	48–60	0.00	0.00	0.00	0.03	0.04	0.07
	Estimated Proportion Consistently Classified: Total = 0.55, Proficient & Above = 0.87						

Appendix 8.C—IRT Analyses

Table 8.C.1 Conversions for Science, Grade Five

Raw Scr.	Freq. Distrib.	Theta	Scale Score	Rprtd Score	Raw Scr.	Freq. Distrib.	Theta	Scale Score	Rprtd Score
0	0	N/A	150.0000	150	28	1,142	0.0712	337.5687	338
1	0	-4.3082	150.0000	150	29	1,108	0.1636	344.4507	344
2	0	-3.5850	150.0000	150	30	1,069	0.2577	351.4485	351
3	1	-3.1490	150.0000	150	31	1,068	0.3536	358.5914	359
4	0	-2.8304	150.0000	150	32	1,104	0.4520	365.9153	366
5	1	-2.5760	150.0000	150	33	1,036	0.5534	373.4561	373
6	0	-2.3620	156.4763	156	34	966	0.6582	381.2595	381
7	3	-2.1757	170.3378	170	35	953	0.7673	389.3793	389
8	10	-2.0097	182.6978	183	36	854	0.8815	397.8804	398
9	14	-1.8589	193.9220	194	37	772	1.0020	406.8435	407
10	35	-1.7200	204.2596	204	38	609	1.1300	416.3708	416
11	82	-1.5906	213.8913	214	39	548	1.2674	426.5976	427
12	92	-1.4688	222.9504	223	40	422	1.4166	437.7017	438
13	155	-1.3534	231.5396	232	41	358	1.5809	449.9342	450
14	230	-1.2433	239.7396	240	42	273	1.7653	463.6594	464
15	334	-1.1374	247.6155	248	43	165	1.9774	479.4452	479
16	416	-1.0353	255.2207	255	44	97	2.2298	498.2319	498
17	498	-0.9361	262.6009	263	45	77	2.5463	521.7867	522
18	629	-0.8395	269.7933	270	46	26	2.9801	554.0742	554
19	707	-0.7449	276.8329	277	47	10	3.7008	600.0000	600
20	696	-0.6519	283.7488	284	48	3	N/A	600.0000	600
21	819	-0.5603	290.5682	291					
22	824	-0.4697	297.3158	297					
23	993	-0.3796	304.0151	304					
24	916	-0.2900	310.6888	311					
25	981	-0.2004	317.3587	317					
26	1,028	-0.1105	324.0468	324					
27	1,112	-0.0201	330.7741	331					

Note: Performance-level cut scores are highlighted.

Table 8.C.2 Conversions for Science, Grade Eight

Raw Scr.	Freq. Distrib.	Theta	Scale Score	Rprtd Score	Raw Scr.	Freq. Distrib.	Theta	Scale Score	Rprtd Score
0	1	N/A	150.0000	150	28	894	-0.1656	307.1429	307
1	0	-4.3764	150.0000	150	29	859	-0.0868	314.2075	314
2	0	-3.6553	150.0000	150	30	874	-0.0075	321.3040	321
3	0	-3.2217	150.0000	150	31	875	0.0723	328.4528	328
4	0	-2.9059	150.0000	150	32	831	0.1530	335.6734	336
5	0	-2.6546	150.0000	150	33	778	0.2347	342.9878	343
6	0	-2.4440	150.0000	150	34	863	0.3177	350.4194	350
7	1	-2.2614	150.0000	150	35	755	0.4022	357.9933	358
8	0	-2.0992	150.0000	150	36	679	0.4888	365.7421	366
9	7	-1.9526	150.0000	150	37	654	0.5776	373.6939	374
10	2	-1.8181	159.1781	159	38	565	0.6691	381.8879	382
11	14	-1.6934	170.3484	170	39	543	0.7638	390.3675	390
12	22	-1.5766	180.8078	181	40	535	0.8623	399.1846	399
13	42	-1.4663	190.6790	191	41	462	0.9652	408.4012	408
14	87	-1.3616	200.0581	200	42	428	1.0734	418.0936	418
15	121	-1.2615	209.0224	209	43	348	1.1881	428.3556	428
16	156	-1.1653	217.6352	218	44	316	1.3104	439.3137	439
17	238	-1.0724	225.9479	226	45	266	1.4423	451.1215	451
18	303	-0.9824	234.0058	234	46	197	1.5861	463.9953	464
19	412	-0.8949	241.8466	242	47	141	1.7451	478.2338	478
20	501	-0.8094	249.5037	250	48	106	1.9242	494.2709	494
21	607	-0.7256	257.0052	257	49	76	2.1312	512.8023	513
22	638	-0.6432	264.3783	264	50	59	2.3784	534.9410	535
23	700	-0.5621	271.6466	272	51	48	2.6897	562.8140	563
24	719	-0.4818	278.8320	279	52	14	3.1184	600.0000	600
25	821	-0.4023	285.9550	286	53	5	3.8341	600.0000	600
26	818	-0.3232	293.0352	293	54	3	N/A	600.0000	600
27	828	-0.2444	300.0917	300					

Note: Performance-level cut scores are highlighted.

Table 8.C.3 Conversions for Life Science, Grade Ten

Raw Scr.	Freq. Distrib.	Theta	Scale Score	Rprtd Score	Raw Scr.	Freq. Distrib.	Theta	Scale Score	Rprtd Score
0	0	N/A	150.0000	150	31	361	0.0060	294.1228	294
1	0	-4.2357	150.0000	150	32	371	0.0763	300.1578	300
2	0	-3.5228	150.0000	150	33	325	0.1470	306.2201	306
3	0	-3.0971	150.0000	150	34	373	0.2182	312.3231	312
4	0	-2.7888	150.0000	150	35	326	0.2900	318.4806	318
5	0	-2.5446	150.0000	150	36	354	0.3626	324.7072	325
6	0	-2.3408	150.0000	150	37	293	0.4362	331.0182	331
7	0	-2.1649	150.0000	150	38	350	0.5110	337.4301	337
8	0	-2.0092	150.0000	150	39	341	0.5872	343.9650	344
9	0	-1.8688	150.0000	150	40	315	0.6650	350.6385	351
10	0	-1.7405	150.0000	150	41	296	0.7448	357.4753	357
11	5	-1.6218	154.5435	155	42	255	0.8267	364.5012	365
12	7	-1.5110	164.0473	164	43	245	0.9112	371.7460	372
13	14	-1.4067	172.9923	173	44	244	0.9986	379.2440	379
14	19	-1.3078	181.4687	181	45	200	1.0895	387.0360	387
15	35	-1.2136	189.5483	190	46	232	1.1844	395.1692	395
16	64	-1.1233	197.2899	197	47	169	1.2839	403.7056	404
17	86	-1.0364	204.7412	205	48	161	1.3890	412.7138	413
18	118	-0.9524	211.9432	212	49	131	1.5006	422.2853	422
19	177	-0.8709	218.9300	219	50	110	1.6201	432.5357	433
20	184	-0.7916	225.7314	226	51	80	1.7494	443.6164	444
21	245	-0.7142	232.3731	232	52	66	1.8906	455.7308	456
22	244	-0.6383	238.8779	239	53	62	2.0474	469.1727	469
23	288	-0.5638	245.2666	245	54	25	2.2244	484.3522	484
24	319	-0.4904	251.5563	252	55	27	2.4293	501.9209	502
25	327	-0.4180	257.7648	258	56	17	2.6747	522.9617	523
26	353	-0.3464	263.9076	264	57	13	2.9844	549.5157	550
27	310	-0.2754	269.9992	270	58	5	3.4114	586.1369	586
28	318	-0.2048	276.0534	276	59	0	4.1258	600.0000	600
29	377	-0.1344	282.0834	282	60	1	N/A	600.0000	600
30	363	-0.0642	288.1023	288					

Note: Performance-level cut scores are highlighted.

Chapter 9: Quality Control Procedures

Rigorous quality control procedures were implemented throughout the test development, administration, scoring, and reporting processes. As part of this effort, Educational Testing Service (ETS) maintains an Office of Testing Integrity (OTI) that resides in the ETS legal department. The OTI provides quality assurance services for all testing programs administered by ETS. In addition, the Office of Professional Standards Compliance at ETS publishes and maintains the *ETS Standards for Quality and Fairness*, which supports the OTI's goals and activities. The purposes of the *ETS Standards for Quality and Fairness* (ETS, 2002) are to help ETS design, develop, and deliver technically sound, fair, and useful products and services; and to help the public and auditors evaluate those products and services.

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

Quality Control of Item Development

The item development process for the California Modified Assessment (CMA) for Science prior to the 2012–13 administration is described in detail in Chapter 3, starting on page 27; there was no new item development in 2014–15 because the forms were reused. The next sections highlight elements of the process devoted specifically to the quality control of the items that were previously developed and reused during the 2014–15 CMA for Science administration.

Item Specifications

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

Item Writers

The items for each CMA for Science were written by item writers with a thorough understanding of the California content standards. The item writers were carefully screened and selected by senior ETS content staff and approved by the CDE. Only those with strong content and teaching backgrounds were invited to participate in an extensive training program for item writers.

Internal Contractor Reviews

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

Content Review

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

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

Editorial Review

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

Fairness Review

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

Assessment Director Review

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

Assessment Review Panel Review

The Assessment Review Panels (ARPs) were committees that advised the CDE and ETS on areas related to item development for the CMA for Science. The ARPs were responsible for reviewing all newly developed items for alignment to the California content standards. The ARPs also reviewed the items for accuracy of content, clarity of phrasing, and quality. See page 30 in Chapter 3 for additional information on the function of ARPs within the item-review process.

Statewide Pupil Assessment Review Panel Review

The Statewide Pupil Assessment Review Panel (SPAR) panel was responsible for reviewing and approving the achievement tests that were used statewide for the testing of students in California public schools in grades five, eight, and ten. The SPAR panel representatives ensured that the test items conformed to the requirements of *Education Code* Section 60602. If the SPAR panel rejected specific items, the items were replaced with other items. See page 33 in Chapter 3 for additional information on the function of the SPAR panel within the item-review process.

Data Review of Field-tested Items

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

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

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

The field-test items that appeared in the CMA for Science administered in 2014–15 were statistically reviewed in data review meetings the year they were originally administered. There was no data review of field-test items in 2014–15. See Table 8.4 on page 87 for the list of the original administrations of each test administered in 2014–15.

Quality Control of the Item Bank

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

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

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

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

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

Quality Control of Test Form Development

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

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

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

Quality Control of Test Materials

Collecting Test Materials

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

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

Processing Test Materials

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

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

Quality Control of Scanning

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

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

Quality Control of Image Editing

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

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

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

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

Quality Control of Answer Document Processing and Scoring

Accountability of Answer Documents

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

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

Processing of Answer Documents

Prior to processing operational answer documents and executing subsequent data processing programs, ETS conducts an end-to-end test. As part of this test, ETS prepares approximately 700 test cases covering all tests and many scenarios designed to exercise particular business rule logic. ETS marks answer documents for those 700 test cases. They are then scanned, scored, and aggregated. The results at various inspection points are checked by psychometricians and Data Quality Services staff. Additionally, a post-scan test file of approximately 50,000 records across the CAASPP System is scored and aggregated to test a broader range of scoring and aggregation scenarios. These procedures assure that students and LEAs receive the correct scores when the actual scoring process is carried out. In 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 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 61. The scoring specifications are reviewed and revised by the CDE and ETS each year. After a version that all parties endorse is finalized, the CDE issues a formal approval of the scoring and reporting specifications.

Storing Answer Documents

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

Quality Control of Psychometric Processes

Score Key Verification Procedures

ETS takes various necessary measures to ascertain that the scoring keys are applied to the student responses as expected and the student scores are computed accurately. Scoring keys, provided in the form planners, are produced by ETS and verified thoroughly by performing multiple quality control checks. The form planners contain the information about an assembled test form; other information in the form planner includes the test name, administration year, subscore identification, and standards and statistics associated with each item. The quality control checks that are performed before keys are finalized are listed on page 62 in Chapter 7.

Quality Control of Item Analyses and the Equating Process

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

review; their comments were reviewed by the psychometricians before items were approved to be included in the equating process.

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

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

Calibrations

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

Scaling

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

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

Scoring Tables

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

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

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

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

During the current administration, the data derived from prior item analyses are used to pre-equate the 2014–15 results. Key checks and classical item analyses as well as associated quality assurance checks are also conducted on the current data.

In addition, the scoring tables are reused and are checked against the scoring tables in the reuse-year technical report to ensure exact match. In addition, prior to reporting in 2014–15, every regular and special-version multiple-choice test was “certified” by ETS prior to being included in electronic reporting. To certify a test, psychometricians gathered a certain number of test cases and verified the accurate application of scoring keys and conversion tables.

Score Verification Process

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

Year-to-Year Comparison Analyses

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

- The first set of year-to-year comparison analyses looks at the tendencies and trends for the schools and LEAs for which ETS has received complete or near-complete results by mid-June.
- The second set of year-to-year comparison analyses uses over 90 percent of the entire testing populations 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 CDE and ETS are obtained.

Offloads to Test Development

During the original administration of the CMA for Science forms that are reused in 2014–15, the statistics based on classical item analyses were obtained. 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.

During the 2014–15 administration, only classical item statistics obtained on larger samples for all operational items are included in the statistical offloads.

Quality Control of Reporting

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

1. Comparing report formats to input sources from the CDE-approved samples
2. Validating and verifying the report data by querying the appropriate student data

3. Evaluating the production print execution performance by comparing the number of report copies, sequence of report order, and offset characteristics to the CDE's requirements
4. Proofreading reports by the CDE and ETS prior to any LEA mailings

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

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

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

Electronic Reporting

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

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

Excluding Student Scores from Summary Reports

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

Reference

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 Modified Assessment (CMA) results are routinely performed to identify the trends in examinee performance and test characteristics over time. Such comparisons were performed over the three most recent administrations—2013, 2014, and 2015—and the base year.

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

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

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

Table 10.1 Base Years for the CMA for Science

CMA	Base Year
Grade 5 Science	2009
Grade 8 Science	2010
Grade 10 Life Science	2011

The base years differ over CMA for Science. Reasons for these differences are as follows:

- In spring 2008, the CMA were first administered statewide for science in grade five. A standard setting was held in fall 2008 to establish cut scores for the below basic, basic, proficient, and advanced performance levels (the cut score for the far below basic performance level was set statistically). Spring 2009 was the first administration in which test results were reported using the new scales and cut scores for the five performance levels; thus, 2009 became the base year for these tests.
- In spring 2009, the CMA were first administered statewide for science in grade eight. A standard setting was held in fall 2009 to establish cut scores for the below basic, basic, proficient, and advanced performance levels (the cut score for the far below basic performance level was set statistically). Spring 2010 was the first administration in which test results were reported using the new scales and cut scores for the five performance levels; thus, 2010 became the base year for these tests.
- In spring 2010, the CMA were first administered statewide for grade ten Life Science. A standard setting was held in fall 2010 to establish cut scores for the below basic, basic, proficient, and advanced performance levels (the cut score for the far below basic performance level was set statistically). Spring 2011 was the first administration in which test results were reported using the new scales and cut scores for the five performance levels; thus, 2011 became the base year for these tests.

Examinee Performance

Table 10.A.1 on page 128 contains the number of examinees assessed and the means and standard deviations of examinees' scale scores in the base year and in 2013, 2014, and 2015 for each CMA for Science. As noted in previous chapters, the CMA for Science reporting scales range from 150 to 600 for all of the tests.

CMA for Science scale scores are used to classify student results into one of five performance levels: far below basic, below basic, basic, proficient, and advanced. The percentages of students qualifying for the proficient and advanced levels are presented in Table 10.A.3 on page 128; please note that this information may differ slightly from information found on the California Department of Education (CDE) California Assessment of Student Performance and Progress (CAASPP) reporting Web page at <http://caaspp.cde.ca.gov> due to differing dates on which data were accessed. The goal is for all students to achieve at or above the proficient level by 2014.

Table 10.A.5 through Table 10.A.7 show, for each CMA for Science, the distribution of scale scores observed in the base year, which differs according to test, and subsequent administrations in 2013, 2014, and 2015 as applicable. Frequency counts are provided for each scale score interval of 30. A frequency count of "N/A" indicates that there are no obtainable scale scores within that scale-score range. For all tests of the CMA for Science, a minimum score of 300 is required for a student to reach the basic level of performance, and a minimum score of 350 is required for a student to reach the proficient level of performance.

Test Characteristics

The item and test analysis results of the CMA for Science over the comparison years indicate that the CMA for Science meet the technical criteria established in professional standards for high-stakes tests.

Table 10.B.1 in Appendix 10.B, which starts on page 131, presents the average proportion correct values for the operational items in each CMA for Science. The mean proportion correct is affected by both the difficulty of the items and the abilities of the students administered the items.

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

The average point-biserial correlations for all of the CMA for Science are presented in Table 10.B.3. The reliabilities and SEM expressed in raw score units appear in Table 10.B.4 and Table 10.B.5. Like the average proportion correct, point-biserial correlations and reliabilities of the operational items 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 Across Base Year, 2013, 2014, and 2015

CMA	Base	2013	2014	2015
Grade 5 Science	18,657	27,709	26,744	23,236
Grade 8 Science	17,606	24,251	22,046	19,212
Grade 10 Life Science	10,786	15,629	12,752	9,601

Table 10.A.2 Scale Score Means and Standard Deviations of CMA for Science Across Base Year, 2013, 2014, and 2015

CMA	Base Mean	Base S.D.	2013 Mean	2013 S.D.	2014 Mean	2014 S.D.	2015 Mean	2015 S.D.
Grade 5 Science	335	58	345	56	345	56	343	57
Grade 8 Science	320	60	336	69	335	66	330	64
Grade 10 Life Science	294	58	303	62	310	61	312	63

Table 10.A.3 Percentage of Proficient and Above Across Base Year, 2013, 2014, and 2015

CMA	Base	2013	2014	2015
Grade 5 Science	42%	47%	47%	45%
Grade 8 Science	31%	42%	39%	37%
Grade 10 Life Science	18%	23%	26%	28%

Table 10.A.4 Percentage of Advanced Across Base Year, 2013, 2014, and 2015

CMA	Base	2013	2014	2015
Grade 5 Science	14%	14%	15%	14%
Grade 8 Science	10%	16%	15%	13%
Grade 10 Life Science	4%	6%	6%	7%

Table 10.A.5 Observed Score Distributions of CMA for Science Across Base Year, 2013, 2014, and 2015 for Science, Grade Five

Observed Score Distributions	Base	2013	2014	2015
570 – 600	5	14	11	13
540 – 569	15	41	24	26
510 – 539	46	87	76	77
480 – 509	82	167	120	97
450 – 479	387	603	976	796
420 – 449	789	2,007	1,227	970
390 – 419	2,013	3,211	2,677	2,235
360 – 389	2,602	3,929	4,802	4,059
330 – 359	3,725	6,964	6,338	5,499
300 – 329	4,028	4,467	4,461	3,918
270 – 299	2,396	4,096	4,023	3,675
240 – 269	1,694	1,653	1,582	1,478
210 – 239	731	391	358	329
180 – 209	123	67	63	59
150 – 179	21	12	6	5

Table 10.A.6 Observed Score Distributions of CMA for Science Across Base Year, 2013, 2014, and 2015 for Science, Grade Eight

Observed Score Distributions	Base	2013	2014	2015
570 – 600	18	36	53	22
540 – 569	17	73	42	48
510 – 539	27	219	220	135
480 – 509	110	212	173	106
450 – 479	279	1,048	822	604
420 – 449	625	1,041	828	664
390 – 419	1,011	2,882	2,387	1,968
360 – 389	2,027	2,788	2,293	1,898
330 – 359	2,780	3,970	3,644	3,227
300 – 329	4,069	3,844	4,916	4,330
270 – 299	3,291	4,304	3,347	3,058
240 – 269	2,333	2,462	2,265	2,158
210 – 239	759	986	770	697
180 – 209	221	316	260	272
150 – 179	39	70	26	25

Table 10.A.7 Observed Score Distributions of CMA for Science Across Base Year, 2013, 2014, and 2015 for Life Science (Grade Ten)

Observed Score Distributions	Base	2013	2014	2015
570 – 600	5	4	6	6
540 – 569	7	12	18	13
510 – 539	11	19	15	17
480 – 509	41	87	64	52
450 – 479	67	164	145	128
420 – 449	153	421	378	321
390 – 419	457	691	641	562
360 – 389	694	1,335	1,293	944
330 – 359	1,307	2,349	2,130	1,595
300 – 329	1,807	2,717	2,304	1,749
270 – 299	2,261	2,791	2,317	1,729
240 – 269	2,035	2,370	1,719	1,287
210 – 239	1,602	2,107	1,409	968
180 – 209	273	489	284	204
150 – 179	66	73	29	26

Appendix 10.B—Historical Comparisons Tables, Test Characteristics

Table 10.B.1 Mean Proportion Correct for Operational Test Items Across Base Year, 2013, 2014, and 2015

CMA	Base	2013	2014	2015
Grade 5 Science	0.61	0.61	0.59	0.59
Grade 8 Science	0.53	0.59	0.58	0.57
Grade 10 Life Science	0.50	0.53	0.55	0.56

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

CMA	Base	2013	2014	2015
Grade 5 Science	-0.54	-0.38	-0.29	-0.29
Grade 8 Science	-0.18	-0.26	-0.25	-0.25
Grade 10 Life Science	0.01	-0.06	-0.06	-0.06

Table 10.B.3 Mean Point-Biserial Correlation for Operational Test Items Across Base Year, 2013, 2014, and 2015

CMA	Base	2013	2014	2015
Grade 5 Science	0.33	0.32	0.32	0.32
Grade 8 Science	0.29	0.33	0.31	0.31
Grade 10 Life Science	0.30	0.32	0.31	0.32

Table 10.B.4 Score Reliabilities (Cronbach's Alpha) Across Base Year, 2013, 2014, and 2015

CMA	Base	2013	2014	2015
Grade 5 Science	0.82	0.81	0.82	0.82
Grade 8 Science	0.80	0.84	0.82	0.82
Grade 10 Life Science	0.83	0.85	0.84	0.85

Table 10.B.5 SEM Across Base Year, 2013, 2014, and 2015

CMA	Base	2013	2014	2015
Grade 5 Science	3.08	3.14	3.16	3.17
Grade 8 Science	3.44	3.36	3.39	3.41
Grade 10 Life Science	3.65	3.61	3.59	3.58