

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



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

**Final Submitted March 6, 2015
Educational Testing Service
Contract No. 5417**

Table of Contents

Acronyms and Initialisms Used in the <i>CMA Technical Report</i>	vi
Chapter 1: Introduction	1
Background	1
Test Purpose	1
Test Content	1
Intended Population	2
Intended Use and Purpose of Test Scores	2
Testing Window	3
Significant CAASPP Developments in 2014	3
Renamed the Program.....	3
Pre-equated All Results	3
Reduced the Number of Tests	3
Reduced the Number of Test Versions	3
Updated Universal Tools, Designated Supports, and Accommodations (formerly Accommodations and Variations).....	3
Suspended Reporting of Adequate Yearly Progress and the Academic Performance Index.....	3
Limitations of the Assessment	3
Score Interpretation	3
Out-of-Level Testing	4
Score Comparison	4
Groups and Organizations Involved with the CAASPP Assessment System	4
State Board of Education	4
California Department of Education	4
Contractors	5
Overview of the Technical Report	5
References	7
Chapter 2: An Overview of CMA 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	11
Non-embedded Supports.....	12
Individualized Aids (Previously Called Modifications)	12
Special Services Summaries	12
Scores	13
Aggregation Procedures	13
Equating	14
Post-Equating	14
Pre-Equating.....	14
References	17
Appendix 2.A—CMA Items and Estimated Time Chart	18
Appendix 2.B—Reporting Clusters	19
Science	19
Appendix 2.C—Universal Tools, Designated Supports, and Accommodations for the California Assessment of Student Performance and Progress	20
Appendix 2.D—Special Service Summary Tables	22
Chapter 3: Item Development	25
Rules for Item Development	25
Item Specifications.....	25
Expected Item Ratio.....	26
Selection of Item Writers	26

Criteria for Selecting Item Writers	26
Item Review Process	27
Contractor Review	27
Content Expert Reviews	28
Statewide Pupil Assessment Review Panel.....	31
Field Testing	31
Stand-alone Field Testing	31
Embedded Field-test Items	31
CDE Data Review	32
Item Banking	32
References	34
Chapter 4: Test Assembly	35
Test Length	35
Rules for Item Selection	35
Test Blueprint.....	35
Content Rules and Item Selection.....	35
Psychometric Criteria	36
Projected Psychometric Properties of the Assembled Tests.....	37
Rules for Item Sequence and Layout	38
Reference	39
Appendix 4.A—Technical Characteristics	40
Appendix 4.B—Cluster Targets	41
Chapter 5: Test Administration	44
Test Security and Confidentiality	44
ETS’s Office of Testing Integrity.....	44
Test Development.....	44
Item and Data Review.....	44
Item Banking.....	45
Transfer of Forms and Items to the CDE	45
Security of Electronic Files Using a Firewall	45
Printing and Publishing	46
Test Administration	46
Test Delivery.....	46
Processing and Scoring	47
Data Management	47
Transfer of Scores via Secure Data Exchange	48
Statistical Analysis	48
Reporting and Posting Results.....	48
Student Confidentiality	48
Student Test Results.....	48
Procedures to Maintain Standardization	49
Test Administrators	49
Directions for Administration	50
LEA and Test Site Coordinator Manual.....	50
Test Management System Manuals.....	51
Test Booklets	51
Universal Tools, Designated Supports, and Accommodations	51
Identification.....	51
Scoring.....	52
Testing Incidents	52
Social Media Security Breaches	52
Testing Improprieties	52
References	53
Chapter 6: Performance Standards	54
Background	54
Standard-Setting Procedure	54
Development of Competencies Lists.....	56
Standard-Setting Methodology	56
Bookmark Method.....	56
Results	57
References	58
Chapter 7: Scoring and Reporting	59
Procedures for Maintaining and Retrieving Individual Scores	59

Scoring and Reporting Specifications	59
Scanning and Scoring	60
Types of Scores and Subscores	60
Raw Score	60
Subscore	60
Scale Score	60
Performance Levels	60
Score Verification Procedures	61
Scoring Key Verification Process	61
Overview of Score Aggregation Procedures	61
Individual Scores	61
Reports Produced and Scores for Each Report	64
Types of Score Reports	64
Score Report Contents	64
Score Report Applications	65
Criteria for Interpreting Test Scores	65
Criteria for Interpreting Score Reports	65
Reference	67
Appendix 7.A—Scale Score Distribution Tables	68
Appendix 7.B—Demographic Summaries	69
Appendix 7.C—Types of Score Reports	75
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	84
The Constructs to Be Measured	84
Interpretations and Uses of the Scores Generated	85
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	90
References	93
Appendix 8.A—Classical Analyses	95
Appendix 8.B—Reliability Analyses	96
Appendix 8.C—IRT Analyses	105
Chapter 9: Quality Control Procedures	108
Quality Control of Item Development	108
Item Specifications	108
Item Writers	108
Internal Contractor Reviews	108
Assessment Review Panel Review	109
Statewide Pupil Assessment Review Panel Review	109
Data Review of Field-tested Items	109
Quality Control of the Item Bank	110
Quality Control of Test Form Development	110
Quality Control of Test Materials	111
Collecting Test Materials	111

Processing Test Materials.....	111
Quality Control of Scanning	111
Quality Control of Image Editing	112
Quality Control of Answer Document Processing and Scoring	112
Accountability of Answer Documents.....	112
Processing of Answer Documents.....	112
Scoring and Reporting Specifications.....	113
Storing Answer Documents.....	113
Quality Control of Psychometric Processes	113
Score Key Verification Procedures.....	113
Quality Control of Item Analyses and the Equating Process.....	113
Score Verification Process.....	115
Year-to-Year Comparison Analyses.....	115
Offloads to Test Development.....	115
Quality Control of Reporting	115
Electronic Reporting.....	116
Excluding Student Scores from Summary Reports.....	116
Reference	117
Chapter 10: Historical Comparisons	118
Base Year Comparisons	118
Examinee Performance	118
Test Characteristics	119
Appendix 10.A—Historical Comparisons Tables, Examinee Performance	120
Appendix 10.B—Historical Comparisons Tables, Test Characteristics	121

Tables

Table 2.1 Scale-Score Ranges for Performance Levels.....	16
Table 2.C.1 Matrix One Part 2: Non-Embedded Supports for the CMA.....	20
Table 2.D.1 Special Service Summary for Science, Grades Five, Eight, and Ten (Life Science).....	22
Table 3.1 Stand-alone Field-testing Timeline for the CMA.....	31
Table 4.1 Statistical Targets for CMA Test Assembly.....	37
Table 4.A.1 Summary of 2014 CMA Projected Raw Score Statistics.....	40
Table 4.A.2 Summary of 2014 CMA Projected Item Statistics.....	40
Table 7.1 Mean and Standard Deviation of Raw and Scale Scores for the CMA.....	62
Table 7.2 Percentages of Examinees in Each Performance Level.....	62
Table 7.3 Subgroup Definitions.....	63
Table 7.4 Types of CMA Reports.....	64
Table 7.A.1 Distribution of CMA Scale Scores for Science.....	68
Table 7.B.1 Demographic Summary for Science, Grade Five.....	69
Table 7.B.2 Demographic Summary for Science, Grade Eight.....	71
Table 7.B.3 Demographic Summary for Life Science (Grade 10).....	73
Table 7.C.1 Score Reports Reflecting CMA Results.....	75
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.....	81
Table 8.3 Scale Score CSEM at Performance-level Cut Points.....	82
Table 8.4 Original Year of Administration for the CMA.....	87
Table 8.A.1 Item-by-item p -value and Point Biserial for Science, Grades Five, Eight, and Ten—Current Year (2014) and Original Year of Administration.....	95
Table 8.B.1 Subscore Reliabilities and Intercorrelations for Science.....	96
Table 8.B.2 Reliabilities and SEMs for the CMA by Gender.....	96
Table 8.B.3 Reliabilities and SEMs for the CMA by Economic Status.....	96
Table 8.B.4 Reliabilities and SEMs for the CMA by English-language Fluency.....	96
Table 8.B.5 Reliabilities and SEMs for the CMA by Primary Ethnicity.....	97
Table 8.B.6 Reliabilities and SEMs for the CMA by Primary Ethnicity-for-Not-Economically-Disadvantaged.....	97
Table 8.B.7 Reliabilities and SEMs for the CMA by Primary Ethnicity-for-Economically-Disadvantaged.....	97
Table 8.B.8 Reliabilities and SEMs for the CMA by Gender by Economic Status.....	98
Table 8.B.9 Reliabilities and SEMs for the CMA by Primary Disability.....	98
Table 8.B.10 Reliabilities and SEMs for the CMA by Primary Disability (continued).....	98
Table 8.B.11 Overall Subgroup Reliabilities.....	99
Table 8.B.12 Overall Subgroup Reliabilities—Primary Ethnicity.....	99
Table 8.B.13 Overall Subgroup Reliabilities by Primary Ethnicity—Not Economically Disadvantaged.....	99
Table 8.B.14 Overall Subgroup Reliabilities by Primary Ethnicity—Economically Disadvantaged.....	99
Table 8.B.15 Overall Subgroup Reliabilities by Gender/Economic Status.....	99

Table 8.B.16 Overall Subgroup Reliabilities by Primary Disability	100
Table 8.B.17 Overall Subgroup Reliabilities by Primary Disability (continued).....	100
Table 8.B.18 Subscore Reliabilities and SEM for Science by Gender/Economic Status	100
Table 8.B.19 Subscore Reliabilities and SEM for Science by English-language Fluency	100
Table 8.B.20 Subscore Reliabilities and SEM for Science by Primary Ethnicity	101
Table 8.B.21 Subscore Reliabilities and SEM for Science by Primary Ethnicity-for-Not Economically Disadvantaged	101
Table 8.B.22 Subscore Reliabilities and SEM for Science by Primary Ethnicity-for-Economically Disadvantaged	102
Table 8.B.23 Subscore Reliabilities and SEM for Science by Disability	102
Table 8.B.24 Subscore Reliabilities and SEM for Science by Disability (continued)	103
Table 8.B.25 Reliability of Classification for Science, Grade Five	103
Table 8.B.26 Reliability of Classification for Science, Grade Eight	103
Table 8.B.27 Reliability of Classification for Life Science (Grade 10)	104
Table 8.C.1 Conversions for Science, Grade Five	105
Table 8.C.2 Conversions for Science, Grade Eight.....	106
Table 8.C.3 Conversions for Life Science, Grade Ten.....	107
Table 10.1 Base Years for the CMA	118
Table 10.A.1 Number of Examinees Tested, Scale Score Means, and Standard Deviations of CMA Across Base Year, 2012, 2013, and 2014.....	120
Table 10.A.2 Percentage of Proficient and Above and Percentage of Advanced Across Base Year, 2012, 2013, and 2014	120
Table 10.A.3 Observed Score Distributions of CMA Across Base Year, 2012, 2013, and 2014	120
Table 10.B.1 Mean Proportion Correct for Operational Test Items Across Base Year, 2012, 2013, and 2014.....	121
Table 10.B.2 Mean IRT <i>b</i> -values for Operational Test Items Across Base Year, 2012, 2013, and 2014.....	121
Table 10.B.3 Mean Point-Biserial Correlation for Operational Test Items Across Base Year, 2012, 2013, and 2014.....	121
Table 10.B.4 Score Reliabilities (Cronbach's Alpha) and SEM Across Base Year, 2012, 2013, and 2014.....	121

Figures

Figure 3.1 The ETS Item Development Process for the CAASPP System	25
Figure 4.A.1 Plots of Target Information Function and Projected Information for Total Test for Science	40
Figure 4.B.1 Plots of Target Information Functions and Projected Information for Clusters for Science, Grade Five	41
Figure 4.B.2 Plots of Target Information Functions and Projected Information for Clusters for Science, Grade Eight.....	42
Figure 4.B.3 Plots of Target Information Functions and Projected Information for Clusters for Life Science, Grade Ten ...	43
Figure 6.1 Bookmark Standard-setting Process for the CMA.....	56
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 Technical Report*

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

Chapter 1: Introduction

Background

In 1997 and 1998, the California State Board of Education (SBE) adopted content standards in four major content areas: English–language arts (ELA), mathematics, history–social science, and science. These standards were designed to provide state-level input into instruction curricula.

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

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

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

Test Purpose

The purpose of the CMA are 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.

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 biology. For a list of the CMA for Science reporting clusters and the standards they assess, see Appendix 2.B—Reporting Clusters on page 19.

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 science assessment or, for eligible students, a CMA science assessment; or, for students in grades two through eleven who meet the eligibility requirements, the CAPA. This requirement includes 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 science CSTs, the science CMA, or the CAPA is made by their IEP team.

The CMA are designed for students with an IEP who meet eligibility criteria adopted by the SBE. The decision to administer the CMA 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 content area the student is assigned to take. In addition, to be eligible to take the CMA, the student must have scored at the below basic or far below basic performance level on a previously administered CST.

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

Intended Use and Purpose of Test Scores

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

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

“60602. (a) (4) Provide information to pupils, parents or guardians, teachers, schools, and school districts on a timely basis so that the information can be used to further the development of the pupil and to improve the educational program.”

“60602. (c) It is the intent of the Legislature that parents, classroom teachers, other educators, governing board members of school districts, and the public be involved, in an active and ongoing basis, in the design and implementation of the statewide pupil assessment program and the development of assessment instruments.”

“60602. (d) It is the intent of the Legislature, insofar as is practically feasible and following the completion of annual testing, that the content, test structure, and test items in the assessments that are part of the Standardized Testing and Reporting Program become open and transparent to teachers, parents, and pupils, to assist all the stakeholders in working together to demonstrate improvement in pupil academic achievement. A planned

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

Testing Window

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

Significant CAASPP Developments in 2014

Renamed the Program

The CMA for Science paper-pencil tests administered in 2014 are now a component of the California Assessment of Student Performance and Progress (CAASPP) System pursuant to *CCR 5 § 850*.

Pre-equated All Results

Because intact test forms were used, raw-score-to-scale-score conversion tables were developed before tests were administered and used on these tests. This process was used on all CMA for Science forms.

Reduced the Number of Tests

Because California is in transition to the new assessment system, the number of non-computer-administered tests is reduced to only include grade-level science for the CSTs and CMA; and reading/language arts for the STS. (In 2014, there was no reduction in content areas for the CAPA, although only one version was administered.)

Reduced the Number of Test Versions

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

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

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

Suspended Reporting of Adequate Yearly Progress and the Academic Performance Index

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

Limitations of the Assessment

Score Interpretation

Teachers and administrators should not use CAASPP results in isolation to make inferences about instructional needs. In addition, it is important to remember that a single test can provide only limited information. Other relevant information should be considered as well. It is advisable for parents to evaluate their child’s strengths and weaknesses in the relevant topics by reviewing local assessments, classroom tests, student grades, classroom work, and teacher recommendations in addition to the child’s CMA results (CDE, 2013). It is also

important to note that student scores in a content area contain measurement error and could theoretically vary if students were retested.

Out-of-Level Testing

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

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

Groups and Organizations Involved with the CAASPP Assessment System

State Board of Education

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

California Department of Education

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

The CDE oversees California's public school system, which is responsible for the education of more than 6,200,000 children and young adults in more than 9,800 schools. California aims to provide a world-class education for all students, from early childhood to adulthood. The Department of Education serves California by innovating and collaborating with educators, schools, parents, and community partners which together, as a team, prepares students to live, work, and thrive in a highly connected world.

Contractors

Educational Testing Service

The CDE and the SBE contract with ETS to develop, administer, and report the CAASPP assessments. As the prime contractor, ETS has overall responsibility for working with the CDE to implement and maintain an effective assessment system and to coordinate the work of ETS and its subcontractor Pearson. Activities directly conducted by ETS include the following:

- Overall management of the program activities;
- Development of all test items;
- Construction and production of test booklets and related test materials;
- Support and training provided to counties, LEAs, and independently testing charter schools;
- Implementation and maintenance of the Test Management System for orders of materials and pre-identification services; and
- Completion of all psychometric activities.

Pearson

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

- Production of all scannable test materials;
- Packaging, distribution, and collection of testing materials to LEAs and independently testing charter schools;
- Scanning and scoring of all responses; and
- Production of all score reports and data files of test results.

Overview of the Technical Report

This technical report addresses the characteristics of the CMA administered in spring 2014. The technical report contains nine additional chapters as follows:

- Chapter 2 presents a conceptual overview of processes involved in a testing cycle for a CMA 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 items before the 2014 administration—in 2014, 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 forms reused in 2014.
- Chapter 5 presents the processes involved in the actual administration of the 2014 CMA 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.
- Chapter 7 details the types of scores and score reports that are produced at the end of each administration of the CMA and includes a discussion of quick-turnaround reporting.
- Chapter 8 summarizes the results of the test- and item-level analyses performed during the spring 2014 administration of the tests. These include the classical item analyses, the reliability analyses that include assessments of test reliability and the consistency and accuracy of the CMA performance-level classifications, and the procedures designed to ensure the validity of CMA score uses and interpretations. Also discussed in this chapter are item response theory (IRT), CMA 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.
- 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.* Downloaded 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&reportNumber=16

Chapter 2: An Overview of CMA Processes

This chapter provides an overview of the processes involved in a typical test development and administration cycle for the CMA. Also described are the specifications maintained by ETS to implement each of those processes. In 2014, three CMA in science were administered; intact forms—i.e., test forms from previous administrations—from different years were used. 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 three science tests of the CMA contain three-option multiple-choice items.

Item Specifications

There were no new items developed in 2014. Prior to the 2013 administration, the CMA items were developed to measure California content standards adopted by the state in 1997 and 1998 and designed to conform to principles of item writing defined by ETS (ETS, 2002). ETS maintained and updated an item specifications document, otherwise known as “item writer guidelines,” for each 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 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.

The items selected for the CMA 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 25.

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 25; and the Statewide Pupil Assessment Review (SPAR) panel, described in Chapter 3, starting on page 31.

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 32 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 32 in Chapter 3.

Item Refresh Rate

Prior to use intact forms in the 2014 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.

Test Assembly

Test Length

The number of operational items in each CMA 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 35 in Chapter 4.

Each CMA 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 administration, the field-test items were included as part of the intact test forms. The total number of items, including field-test items, in each CMA and the estimated time to complete a test form are presented in Appendix 2.A on page 18.

Test Blueprints

ETS selected all CMA items to conform to the SBE-approved California content standards and test blueprints. The test blueprints for the CMA for Science can be found on the CDE CAASPP Science Assessments Web page at <http://www.cde.ca.gov/ta/tg/ca/caasppscience.asp>.

Although the test blueprints specify the number of items at the individual standard level, scores for the CMA items are grouped into subcontent areas referred to as "reporting clusters." For each CMA reporting cluster, the percentage of questions correctly answered is reported on a student's score report. A description of the CMA reporting clusters and the standards that comprise each cluster are provided in Appendix 2.B, which starts on page 19.

Content Rules and Item Selection

Intact test forms from different years were used during the 2014 administration. (See Table 8.4 on page 87 for administration years.) Prior to the 2013 administration, test developers followed a number of rules when developing a new test form for a given grade and content area. First and foremost, they selected items that met the blueprint for that grade and content area. Using an electronic item bank, assessment specialists began by identifying a number of linking items. These were items that had appeared in previous operational test administrations and were then used to equate subsequent (new) test forms. After the linking items were approved, assessment specialists populated the rest of the test form.

Linking items were selected to proportionally represent the full blueprint. Each CMA 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 35.

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 administration and the projected characteristics of the forms are presented starting from page 36 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 38 in Chapter 4.

All the forms in the 2014 CMA 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 was originally administered.

Test Administration

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

Test Security and Confidentiality

All tests within the CAASPP System are secure documents. For the CMA 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 44.

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 44.

- 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 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, as described in this section.

Test Administrators

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

Staff at LEAs who are central to the processes include LEA CAASPP coordinators, test site coordinators, test examiners, proctors, and scribes. The responsibilities of each of the staff members are included in the *CAASPP LEA and Test Site Coordinator Manual* (CDE, 2014); see page 50 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 examiners to administer the CMA to students to be followed exactly so that all students have an equal opportunity to demonstrate their academic achievement (See page 50 in Chapter 5 for more information.)

CAASPP LEA and Test Site Coordinator Manual—Test administration procedures for LEA CAASPP coordinators and test site coordinators (See page 50 in Chapter 5 for more information.)

Test Management System manuals—Instructions for the Web-based modules that allow LEA CAASPP coordinators to set up test administrations, order materials, and submit and correct student Pre-ID data; every module has its own user manual with detailed instructions on how to use the Test Management System (See page 51 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 IEPs and most English learners take the CMA under standard conditions. However, some students with IEPs and some English learners may need assistance when taking the CMA. 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 20 presents an adaptation of Matrix One of the “Universal Tools, Designated Supports, and Accommodations for the California Assessment of Student Performance and Progress.” Part 2 of Matrix One, found in Table 2.C.1, includes the non-

embedded supports; Appendix 2.C shows only the supports that were allowed for the CMA for Science in 2014 and were mapped to CMA answer documents so had data that could be collected. Table 2.C.1 also shows the answer document options in section A3 that are reported in Appendix 2.D and were defined but did not map to a specific universal tool, designated support, or accommodation, as well as the reported answer document options in section A4 that are unmapped.

The purpose of universal tools, designated supports, and accommodations for the CMA is to enable the students to take the assessments, rather than give them an advantage over other students or to artificially inflate their scores.

Non-embedded Supports

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

Individualized Aids (Previously Called Modifications)

Individualized aids, previously called modifications, fundamentally change what is being tested and may interfere with the construct being measured. All individualized aids must be identified, approved, and listed in the student’s IEP or Section 504 plan. Individualized aids are not available for the CMA for Science, as these tests have built-in modifications.

Special Services Summaries

The percentage of students using various universal tools, designated supports, and accommodations during the 2014 administration of the CMA 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.

Scores

Total test raw scores for the CMA in science equal the sum of examinees' scores on the operational multiple-choice test items.

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

In addition to scale scores for the total content-area test, performance on the associated reporting clusters is reported. The subscore or reporting cluster score is obtained by summing an examinee's scores on the items in each reporting cluster. That information is reported in terms of a percent-correct score.

Detailed descriptions of CMA scores are found in Chapter 7, which starts on page 59.

Aggregation Procedures

In order to provide meaningful results to the stakeholders, CMA 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 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 62 in Chapter 7, provide summary statistics for individual scores aggregated at the state level, describing overall student performance on each CMA. 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 student performance by content area and grade are provided in Table 7.B.1 on page 69 in Appendix 7.B.

Demographic Subgroup Scores

In Table 7.B.1 through Table 7.B.3, starting on page 69, 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 63 provides definitions for the demographic groups included in the tables.

Equating

Post-Equating

Prior to the 2013 administration, the CMA 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 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 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.

Pre-Equating

During the 2014 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 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 (EM) cycle with no constraints. For both stages, the metric of the scale was controlled by the constant discrimination parameters.

Scaling

Prior to the 2014 administration, calibrations of the items were linked to the previously obtained reference scale estimates using linking items and the Stocking and Lord (1983)

procedure. In the case of the one-parameter model calibrations, this procedure was equivalent to setting the mean of the new item parameter estimates for the linking set equal to the mean of the previously scaled estimates. As noted earlier, the linking set was a collection of items in a current test form that also appeared in the previous year's form and was scaled at that time.

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

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

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, 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 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 are presented in Table 8.C.1 through Table 8.C.3 in Appendix 8.C, starting on page 105. 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, regardless of when the form was administered, scale scores were adjusted at both ends of the scale so that the minimum reported scale score was 150 and the maximum reported scale score was 600. Raw scores of zero and perfect raw scores were assigned scale scores of 150 and 600, respectively.

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

Table 2.1 Scale-Score Ranges for Performance Levels

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

* Numbers indicate grade-level tests.

References

- California Department of Education. (2014). *2014 CAASPP LEA and test site coordinator manual*. Sacramento, CA. Downloaded from http://californiatac.org/rsc/pdfs/CAASPP.coord_man.2014.pdf
- California Department of Education. (2014). Universal Tools, Designated Supports, and Accommodations for the California Assessment of Student Performance and Progress. Sacramento, CA. <http://www.cde.ca.gov/ta/tg/ai/caasppmatrix1.asp>
- Educational Testing Service. (2002). *ETS standards for quality and fairness*. Princeton, NJ: Author.
- Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory: principles and applications*. Boston, MA: Kluwer-Nijhoff.
- 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 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

Science

Science Modified Standards Assessment (Grade Five)

Physical Sciences	16 items
Life Sciences	16 items
Earth Sciences	16 items

Science Modified Standards Assessment (Grade Eight)

Motion	19 items
Matter	23 items
Earth Science	7 items
Investigation and Experimentation	5 items

Science Modified Standards Assessment (Grade Ten)

Cell Biology and Genetics	22 items
Evolution and Ecology	22 items
Physiology	10 items
Investigation and Experimentation	6 items

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

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

Option	(U) Universal Tool (D) Designated Support (A) Accommodation	
Answer Document Section A3—Accommodations and Modifications		
B	Pupil marks in paper-pencil test booklet (other than responses including highlighting)	U
C	Scribe (previously known as “Essay responses dictated orally, in Manually Coded English, or in American Sign Language to a scribe, audio recorder, or speech-to-text converter” or “Student marks responses in test booklet and responses are transferred to a scorable answer document by an employee of the school, district, or nonpublic school” or “Student dictates multiple-choice question responses orally, or in Manually Coded English to a scribe, audio recorder, or speech-to-text converter for selected-response items”)	A
F	Alternate Response Options includes adapted keyboards, large keyboards, StickyKeys, MouseKeys, FilterKeys, adapted mouse, touch screen, head wand, and switches. (previously known as “Assistive device that does not interfere with the independent work of the student on the multiple choice and/or essay responses [writing portion of the test]”)	N/A
G	Braille (paper-pencil tests)	A
H	Large-print versions of a paper-pencil test (as available)	A
J, K	Breaks (previously known as “Extended Time” or “Test over more than one day for a test or test part to be administered in a single sitting” or “supervised breaks within a section of the test”)	U
L	Administration of the test to the pupil at the most beneficial time of day	A
M	Separate Setting (previously known as “Test individual student separately, provided that a test examiner directly supervises the student” or “Test student in a small group setting” or “Test administered at home or in hospital by test examiner”)	A
O	American Sign Language	A
S	Math Tools (i.e., non-embedded ruler, non-embedded protractor)	N/A
X	Abacus	A
Y	Leave blank	Unmapped
Z	Read Aloud (previously known as “Test questions and answer options read aloud to pupil or used audio CD presentation – excluding passages”)	A
Mark Nothing	Color Overlay (previously known as “Colored overlay, mask, or other means to maintain visual attention”)	U
Mark Nothing	Magnification (previously known as “Visual magnifying equipment”)	D
Mark Nothing	Noise buffers (e.g., individual carrel or study enclosure or noise-cancelling headphones)	D
Mark Nothing	Scratch Paper	U
Mark Nothing	Simplified or clarified test administration directions (does not apply to test questions)	U
Mark Nothing	Special lighting or acoustics, assistive devices (specific devices may require CAASPP contractor certification), and/or special or adaptive furniture	D

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 (STAR) within a test part provided that the test section is completed within a testing day. A test section is identified by a “STOP” at the end of it.	Unmapped
C	English learners (ELs) may have the opportunity to be tested separately with other ELs provided that the student is directly supervised by an employee of the school who has signed the test security affidavit and the student has been provided such a flexible setting as part of his/her regular instruction or assessment.	Unmapped
D	Translations (Glossary) (previously known as “Access to translation glossaries/word lists (English-to-primary language). Glossaries/Word lists shall not include definitions or formulas.)	D

- Universal Tools (U) Are available for all pupils. Pupils may turn the support(s) on/off when embedded as part of the technology platform for the computer-administered CAASPP tests or may choose to use it/them when provided as part of a paper-pencil test.
- Designated Supports (D) Are features that are available for use by any pupil for whom the need has been indicated prior to the assessment, by an educator or group of educators.
- Accommodations (A) For the CAASPP 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” in Table 2.D.1 presents the number of students using various special services on the CMA for Science in grade five. The column with the heading of “Pct. of Total” in the same table represents the percent of students using a service out of the total number of test-takers.
2. The total number of test-takers is the total of students listed under “Any universal tool, desig. support, 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.

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

Special Service Summary for Science, Grades Five, Eight, and Ten (Life Science)						
All Tested	Grade 5	Pct. of Total	Grade 8	Pct. of Total	Grade 10 Life Sci.	Pct. of Total
B: Marked in test booklet	664	2.46%	169	0.75%	55	0.42%
C: Scribe	26	0.10%	20	0.09%	14	0.11%
F: Alternate response options	N/A	N/A	N/A	N/A %	N/A	N/A
G: Braille	15	0.06%	6	0.03%	4	0.03%
H: Large-print versions of a paper-pencil test	78	0.29%	52	0.23%	26	0.20%
J: Breaks (Tested over more than one day)	953	3.53%	705	3.13%	138	1.07%
K: Breaks (Had supervised breaks)	4,080	15.09%	2,077	9.23%	1,308	10.10%
L: Most beneficial time of day	2,368	8.76%	980	4.36%	234	1.81%
M: Separate setting	45	0.17%	53	0.24%	14	0.11%
O: American Sign Language	74	0.27%	38	0.17%	36	0.28%
S: Math Tools	N/A	N/A	N/A	N/A	N/A	N/A
X: Abacus	641	2.37%	531	2.36%	145	1.12%
Y: Leave blank	228	0.84%	313	1.39%	78	0.60%
Z: Read aloud	8,434	31.20%	2,712	12.06%	430	3.32%
Univ. tool, desig. sup., and acc. is in Section 504 plan	3	0.01%	0	0.00%	1	0.01%
Univ. tool, desig. sup., and acc. is in IEP	10,894	40.30%	4,948	22.00%	1,938	14.96%
English Learner Test Variation A	13	0.05%	4	0.02%	2	0.02%
English Learner Test Variation B	35	0.13%	9	0.04%	2	0.02%
English Learner Test Variation C	74	0.27%	34	0.15%	8	0.06%
English Learner Test Variation D	3	0.01%	2	0.01%	14	0.11%
Any Universal tool, desig. support, and acc or Additional univ. tool, design. sup. for EL	11,751	43.47%	5,482	24.37%	2,051	15.84%
No Universal tool, desig. support, and acc or Additional univ. tool, design. sup. for EL	15,281	56.53%	17,012	75.63%	10,900	84.16%
English-Only Students	Grade 5	Pct. of Total	Grade 8	Pct. of Total	Grade 10 Life Sci.	Pct. of Total
B: Marked in test booklet	425	2.98%	104	0.91%	34	0.51%
C: Scribe	23	0.16%	13	0.11%	8	0.12%
F: Alternate response options	N/A	N/A	N/A	N/A	N/A	N/A
G: Braille test	5	0.04%	5	0.04%	4	0.06%
H: Large-print versions of a paper-pencil test	49	0.34%	32	0.28%	17	0.25%
J: Breaks (Tested over more than one day)	479	3.36%	354	3.11%	71	1.06%
K: Breaks (Had supervised breaks)	2,107	14.76%	1,063	9.35%	640	9.54%
L: Most beneficial time of day	1,256	8.80%	518	4.55%	144	2.15%
M: Separate setting	35	0.25%	30	0.26%	9	0.13%

Special Service Summary for Science, Grades Five, Eight, and Ten (Life Science)						
O: American Sign Language	49	0.34%	19	0.17%	23	0.34%
S: Math Tools	N/A	N/A	N/A	N/A	N/A	N/A
X: Abacus	386	2.70%	281	2.47%	90	1.34%
Y: Leave blank	116	0.81%	130	1.14%	40	0.60%
Z: Read aloud	4,279	29.98%	1,274	11.20%	192	2.86%
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	5,700	39.94%	2,416	21.24%	985	14.69%
English Learner Test Variation A	2	0.01%	0	0.00%	0	0.00%
English Learner Test Variation B	4	0.03%	0	0.00%	1	0.01%
English Learner Test Variation C	9	0.06%	8	0.07%	1	0.01%
English Learner Test Variation D	0	0.00%	0	0.00%	1	0.01%
Any Universal tool, desig. support, and acc or additional univ. tool, design. sup. for EL	6,124	42.91%	2,676	23.53%	1,042	15.54%
No Universal tool, desig. support, and acc or additional univ. tool, design. sup. for EL	8,149	57.09%	8,699	76.47%	5,664	84.46%
Initially Fluent English Proficient (I-FEP) Students	Grade 5	Pct. of Total	Grade 8	Pct. of Total	Grade 10 Life Sci.	Pct. of Total
B: Marked in test booklet	2	0.75%	0	0.00%	4	1.32%
C: Scribe	0	0.00%	0	0.00%	0	0.00%
F: Alternate response options	N/A	N/A	N/A	N/A	N/A	N/A %
G: Braille	1	0.37%	0	0.00%	0	0.00%
H: Large-print versions of a paper-pencil test	1	0.37%	0	0.00%	2	0.66%
J: Breaks (Tested over more than one day)	7	2.62%	6	1.58%	2	0.66%
K: Breaks (Had supervised breaks)	35	13.11%	36	9.50%	32	10.56%
L: Most beneficial time of day	27	10.11%	14	3.69%	4	1.32%
M: Separate setting	1	0.37%	1	0.26%	1	0.33%
O: American Sign Language	3	1.12%	5	1.32%	1	0.33%
S: Math Tools	N/A	N/A	N/A	N/A	N/A	N/A
X: Abacus	4	1.50%	6	1.58%	2	0.66%
Y: Leave blank	0	0.00%	5	1.32%	2	0.66%
Z: Examiner read test questions aloud	74	27.72%	36	9.50%	12	3.96%
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	101	37.83%	70	18.47%	48	15.84%
English Learner Test Variation A	0	0.00%	0	0.00%	0	0.00%
English Learner Test Variation B	0	0.00%	0	0.00%	0	0.00%
English Learner Test Variation C	0	0.00%	0	0.00%	0	0.00%
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	105	39.33%	78	20.58%	51	16.83%
No Universal tool, desig. support, and acc. or additional univ. tool, design. sup. for EL	162	60.67%	301	79.42%	252	83.17%
English Learner (EL) Students	Grade 5	Pct. of Total	Grade 8	Pct. of Total	Grade 10 Life Sci.	Pct. of Total
B: Marked in test booklet	222	1.86%	54	0.64%	16	0.36%
C: Dictated responses to a scribe	2	0.02%	6	0.07%	4	0.09%
F: Alternate response options	N/A	N/A	N/A	N/A	N/A	N/A
G: Braille	9	0.08%	1	0.01%	0	0.00%
H: Large-print versions of a paper-pencil test	26	0.22%	16	0.19%	5	0.11%
J: Breaks (Tested over more than one day)	457	3.83%	304	3.62%	36	0.80%
K: Breaks (Had supervised breaks)	1,853	15.54%	740	8.81%	463	10.34%
L: Most beneficial time of day	1,043	8.75%	338	4.03%	58	1.30%
M: Separate setting	8	0.07%	20	0.24%	4	0.09%

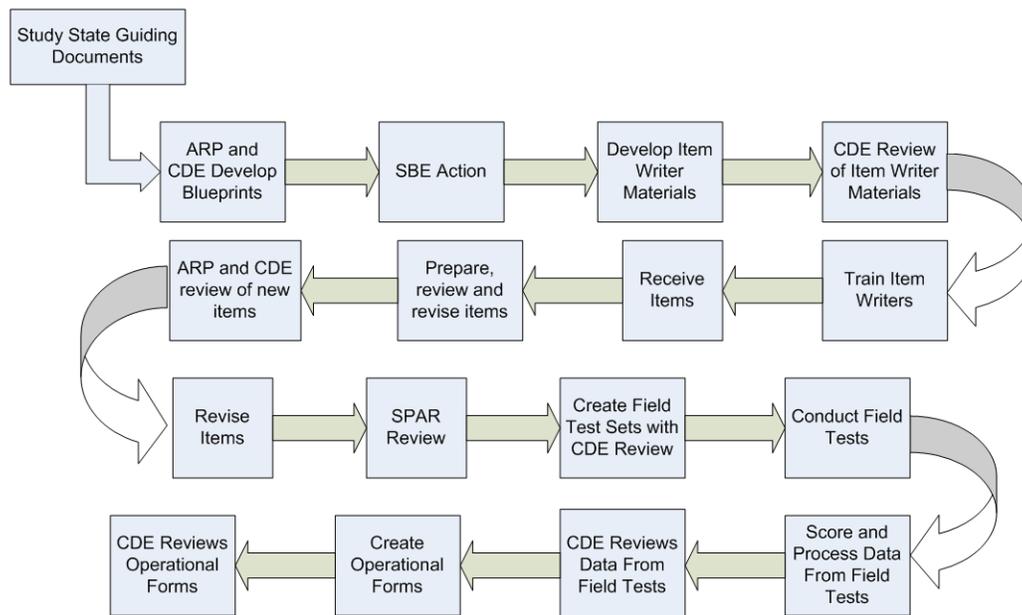
Special Service Summary for Science, Grades Five, Eight, and Ten (Life Science)						
O: American Sign Language	19	0.16%	13	0.15%	11	0.25%
S: Math Tools	N/A	N/A	N/A	N/A	N/A	N/A
X: Abacus	241	2.02%	194	2.31%	40	0.89%
Y: Leave blank	110	0.92%	138	1.64%	26	0.58%
Z: Examiner read test questions aloud	3,919	32.87%	1,108	13.20%	175	3.91%
Univ. tool, desig. sup., and acc. is in Section 504 plan	3	0.03%	0	0.00%	1	0.02%
Univ. tool, desig. sup., and acc. is in IEP	4,885	40.97%	1,926	22.94%	669	14.95%
English Learner Test Variation A	11	0.09%	4	0.05%	2	0.04%
English Learner Test Variation B	31	0.26%	8	0.10%	1	0.02%
English Learner Test Variation C	64	0.54%	24	0.29%	3	0.07%
English Learner Test Variation D	3	0.03%	1	0.01%	13	0.29%
Any Universal tool, desig. support, and acc. or additional univ. tool, design. sup. for EL	5,301	44.46%	2,132	25.40%	712	15.91%
No Universal tool, desig. support, and acc. or additional univ. tool, design. sup. for EL	6,621	55.54%	6,263	74.60%	3,764	84.09%
Reclassified Fluent English Proficient (R-FEP) Students	Grade 5	Pct. of Total	Grade 8	Pct. of Total	Grade 10 Life Sci.	Pct. of Total
B: Marked in test booklet	10	2.55%	6	0.40%	1	0.09%
C: Scribe	0	0.00%	1	0.07%	2	0.19%
F: Alternate response options	N/A	N/A	N/A	N/A	N/A	N/A
G: Braille	0	0.00%	0	0.00%	0	0.00%
H: Large-print versions of a paper-pencil test	0	0.00%	3	0.20%	1	0.09%
J: Breaks (Tested over more than one day)	6	1.53%	29	1.94%	12	1.13%
K: Breaks (Had supervised breaks)	63	16.07%	116	7.77%	105	9.91%
L: Most beneficial time of day	35	8.93%	69	4.62%	4	0.38%
M: Separate setting	1	0.26%	2	0.13%	0	0.00%
O: American Sign Language	2	0.51%	1	0.07%	1	0.09%
S: Math Tools	N/A	N/A	N/A	N/A	N/A	N/A
X: Abacus	5	1.28%	33	2.21%	10	0.94%
Y: Leave blank	1	0.26%	16	1.07%	6	0.57%
Read aloud	127	32.40%	155	10.39%	19	1.79%
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	159	40.56%	304	20.38%	145	13.68%
English Learner Test Variation A	0	0.00%	0	0.00%	0	0.00%
English Learner Test Variation B	0	0.00%	0	0.00%	0	0.00%
English Learner Test Variation C	1	0.26%	2	0.13%	3	0.28%
English Learner Test Variation D	0	0.00%	1	0.07%	0	0.00%
Any Universal tool, desig. support, and acc. or additional univ. tool, design. sup. for EL	165	42.09%	327	21.92%	151	14.25%
No Universal tool, desig. support, and acc. or additional univ. tool, design. sup. for EL	227	57.91%	1,165	78.08%	909	85.75%

Chapter 3: Item Development

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

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

Figure 3.1 The ETS Item Development Process for the CAASPP System



Rules for Item Development

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

Item Specifications

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

- A full statement of each academic content standard, as defined by the SBE (CDE, 2009)
- A description of each content strand
- The expected depth of knowledge (DOK) measured by items written for each standard (coded as 1, 2, 3, or 4; items assigned a DOK of 1 are the least cognitively complex, items assigned a DOK of 3 are the most cognitively complex, and the code of 4 would apply only to some writing tasks)
- The homogeneity of the construct measured by each standard

- A description of the kinds of item stems appropriate for multiple-choice items used to assess each standard
- A description of the kinds of distractors that are appropriate for multiple-choice items assessing each standard
- A description of appropriate data representations (such as charts, tables, graphs, or other illustrations) for mathematics 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, if applicable, for ELA items
- A description of specific kinds of items to be avoided, if any (for example, items with any negative expressions in the stem, e.g., “Which of the following is NOT. . .”)

Expected Item Ratio

ETS prepared the item utilization plan for the development of CMA 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 CMA administration, field-test items were repeated as a part of the intact form.

Selection of Item Writers

Criteria for Selecting Item Writers

The items for each CMA were developed by individual item writers with a thorough understanding of the California content standards. Applicants for item writing were screened by senior ETS content staff. Only those with strong content and teaching backgrounds were approved for inclusion in the training program for item writers. Because most of the participants were current or former California educators, they were particularly knowledgeable about the standards assessed by the CMA. 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 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 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 CAASPP Program, 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 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. The ARPs are advisory panels to the CDE and ETS and provided guidance on matters related to item development for the CMA. The ARPs were responsible for reviewing all newly developed items for alignment to the California content standards. The ARPs also reviewed the items for accuracy of content, clarity of phrasing, and quality. In their examination of test items, the ARPs could raise concerns related to age/grade appropriateness and gender, racial, ethnic, and/or socioeconomic bias.

Composition of ARPs

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

- 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, 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.

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 Items

ETS content-area assessment specialists facilitated the CMA 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
- Overview of the CMA test design specifications and blueprints
- Analysis of the CMA item specifications
- Overview of criteria for evaluating multiple-choice test items and for reviewing constructed response writing tasks
- Review and evaluation of items for bias and sensitivity issues

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

- Overall technical quality
- Match to the California content standards
- Match to the construct being assessed by the standard
- Difficulty range
- Clarity
- Correctness of the answer

- Plausibility of the distractors
- Bias and sensitivity factors

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

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

Does the item:

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

Is the stimulus, if any, for the item:

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

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

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

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

Statewide Pupil Assessment Review Panel

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

Field Testing

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

Stand-alone Field Testing

For each new CMA 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 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 is presented in Table 3.1.

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

Content Area	CMA *	Field-test Year
	5	2007
Science	8	2008
	10 Life Science	2009

* Number indicates grade-level tests.

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 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 are not presented in this report, because

for the 2014 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 were spiraled among students in the state so that a large representative sample of test-takers responded to the field-test items embedded in these forms. The spiraling design ensured that a diverse sample of students took each field-test item. The students did not know which items were field-test items and which items were operational items; therefore, their motivation was not expected to vary over the two types of items (Patrick & Way, 2008).

CDE Data Review

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

Item Banking

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

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

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 (CSTs, CMA, CAPA, and STS). CAHSEE and CAASPP are currently consolidated in the California item banking system. ETS works with the CDE to obtain the data for assessments such as the CELDT, under contract with other vendors for inclusion into the item bank. ETS provides the item banking application using the 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. Downloaded 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 CMA were constructed to measure students' performance relative to California's content standards approved by the SBE. They were also constructed to meet professional standards for validity and reliability. For each CMA, 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 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 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 18.

Rules for Item Selection

Test Blueprint

All test items on CMA forms were selected to conform to the SBE-approved California content standards and test blueprints. The content blueprints for the CMA can be found on the 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 items are grouped into subcontent areas (reporting clusters). For each CMA reporting cluster, the percentage of questions correctly answered is reported on a student's score report. A list of the CMA reporting clusters by test and the number of items in the cluster that appear in each test are provided in Appendix 2.B—Reporting Clusters, which starts on page 19.

Content Rules and Item Selection

Intact test forms from previous testing administrations from different years were used during the 2014 administration. Prior to the 2014 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 25, contains further information about this process.

Psychometric Criteria

The three goals of CMA 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. Such rules are referred to as statistical targets.

Two types of assembly targets were developed for each CMA: 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 forms used in the 2014 administration were the test information function (TIF) and an average point-biserial correlation.

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

The graphs for the total test are presented in Figure 4.A.1, starting on page 40, for the science tests. These curves present the target TIF and the projected TIF for the total test.

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

Content Area	CMA *	Point Biserial		b-value		p-value	
		Mean	Minimum	Mean	St. Dev.	Minimum	Maximum
	5	0.37	0.14	-0.43	0.74	0.33	0.95
Science	8	0.33	0.14	-0.46	0.50	0.33	0.95
	10 Life Science	0.30	0.14	-0.22	0.50	0.33	0.95

* Numbers indicate grade-level tests.

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

Figure 4.B.1 through Figure 4.B.3, starting on page 41, present the target and projected information curves for the clusters in the administered tests.

Projected Psychometric Properties of the Assembled Tests

Prior to the 2014 administration, ETS psychometricians performed a preliminary review of the technical characteristics of the assembled tests. The expected or projected performance of examinees and the overall score reliability were estimated using the item-level statistics available in the California item bank for the selected items. The test reliability was based on Gulliksen's formula (Gulliksen, 1987) for estimating test reliability (r_{xx}) from item p -values and item point-biserial correlations:

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

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. Table 4.A.1 on page 40 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 CMA administration.

Table 4.A.2 on page 40 shows the mean observed statistics of the items on each CMA 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 2014 CMA Projected Raw Score Statistics

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

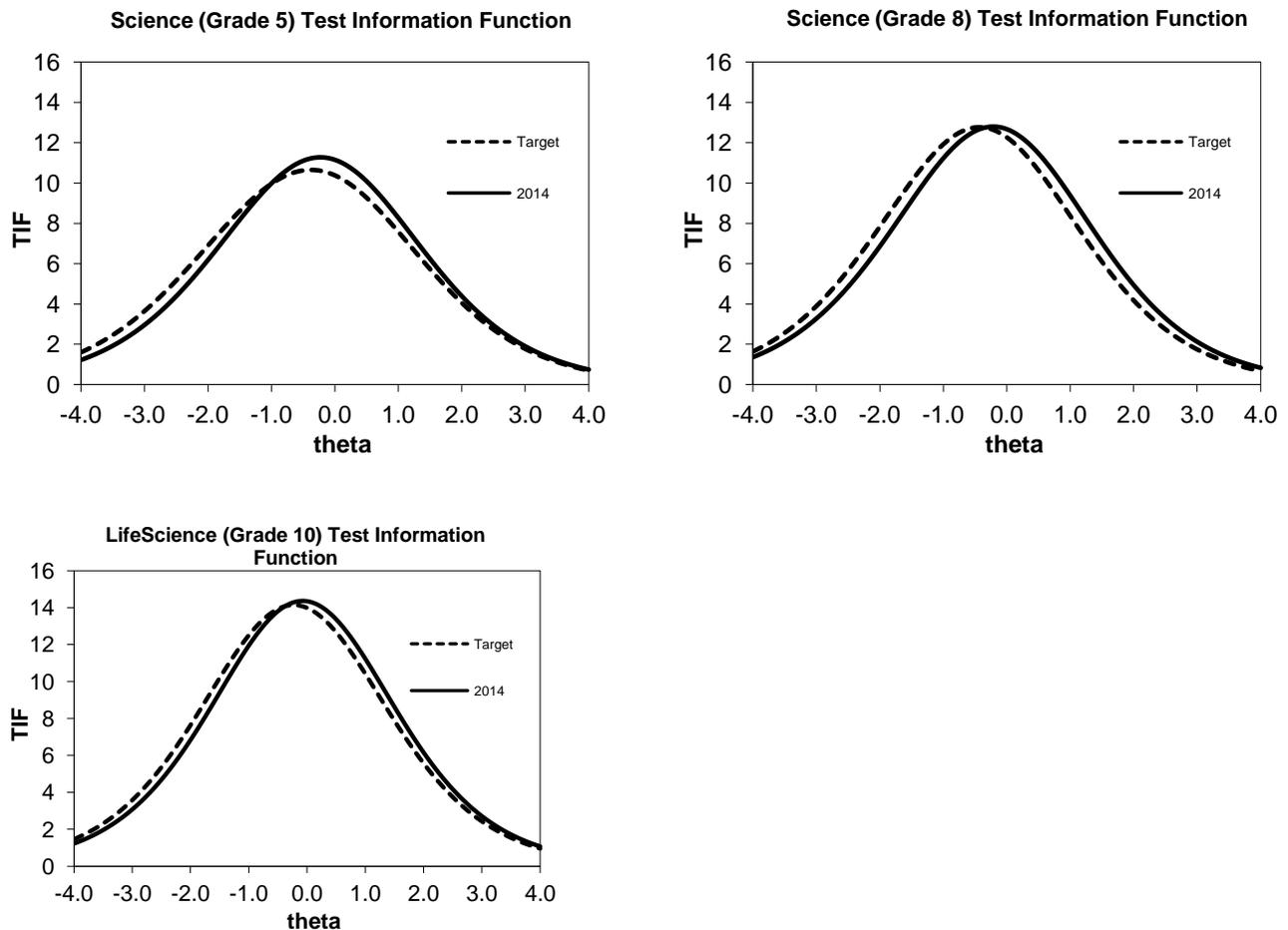
* Numbers indicate grade-level tests.

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

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

* Numbers indicate grade-level tests.

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



Appendix 4.B—Cluster Targets

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

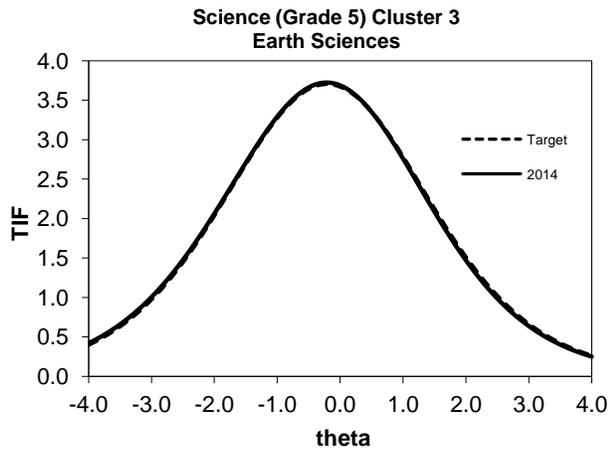
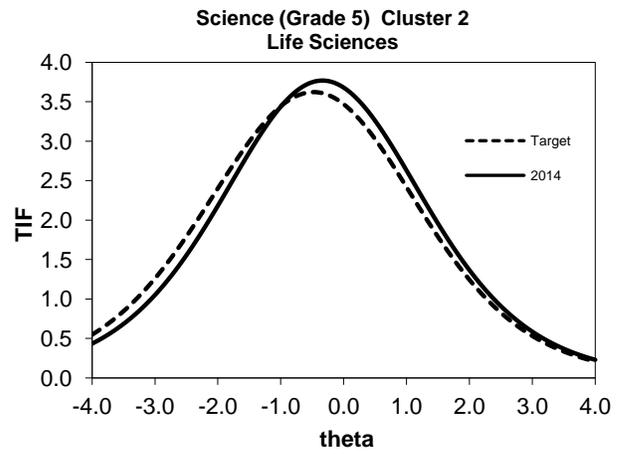
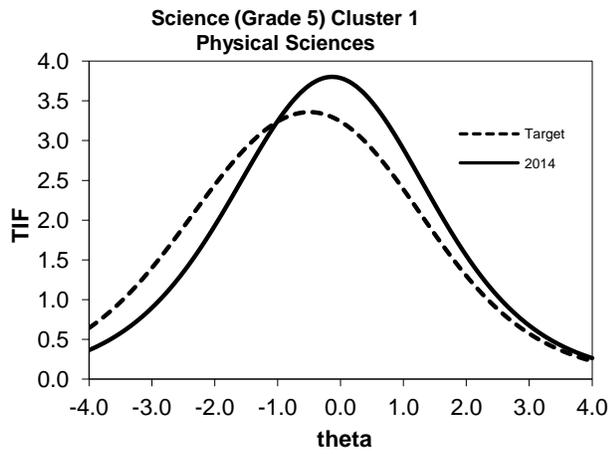


Figure 4.B.2 Plots of Target Information Functions and Projected Information for Clusters for Science, Grade Eight

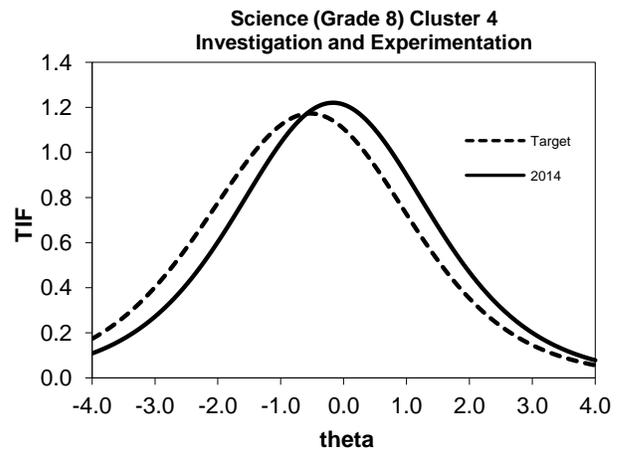
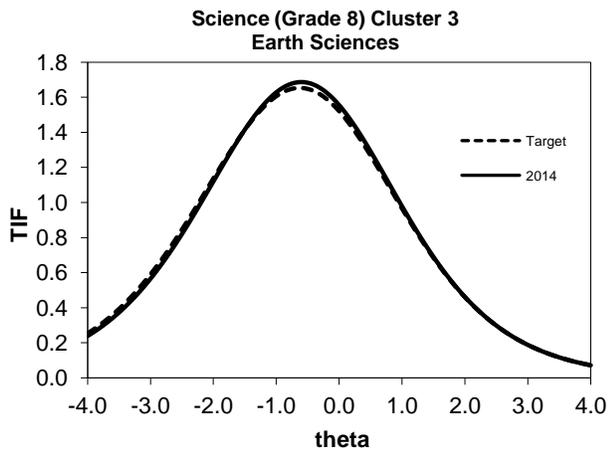
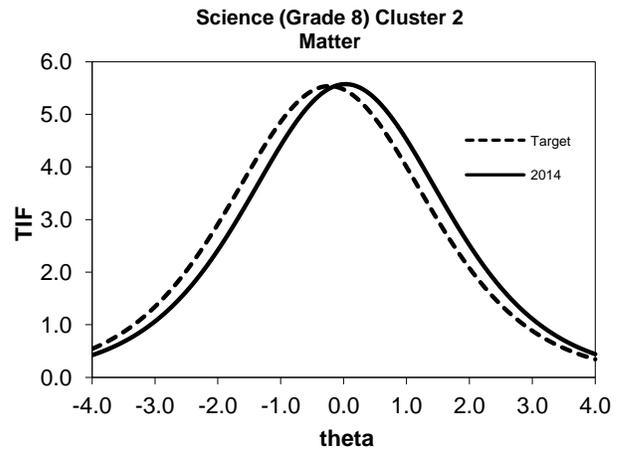
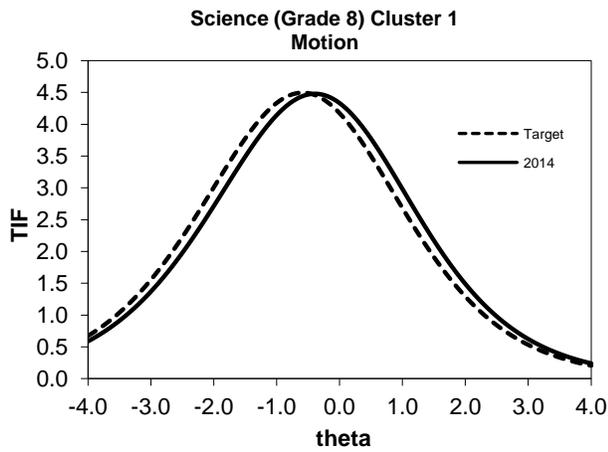
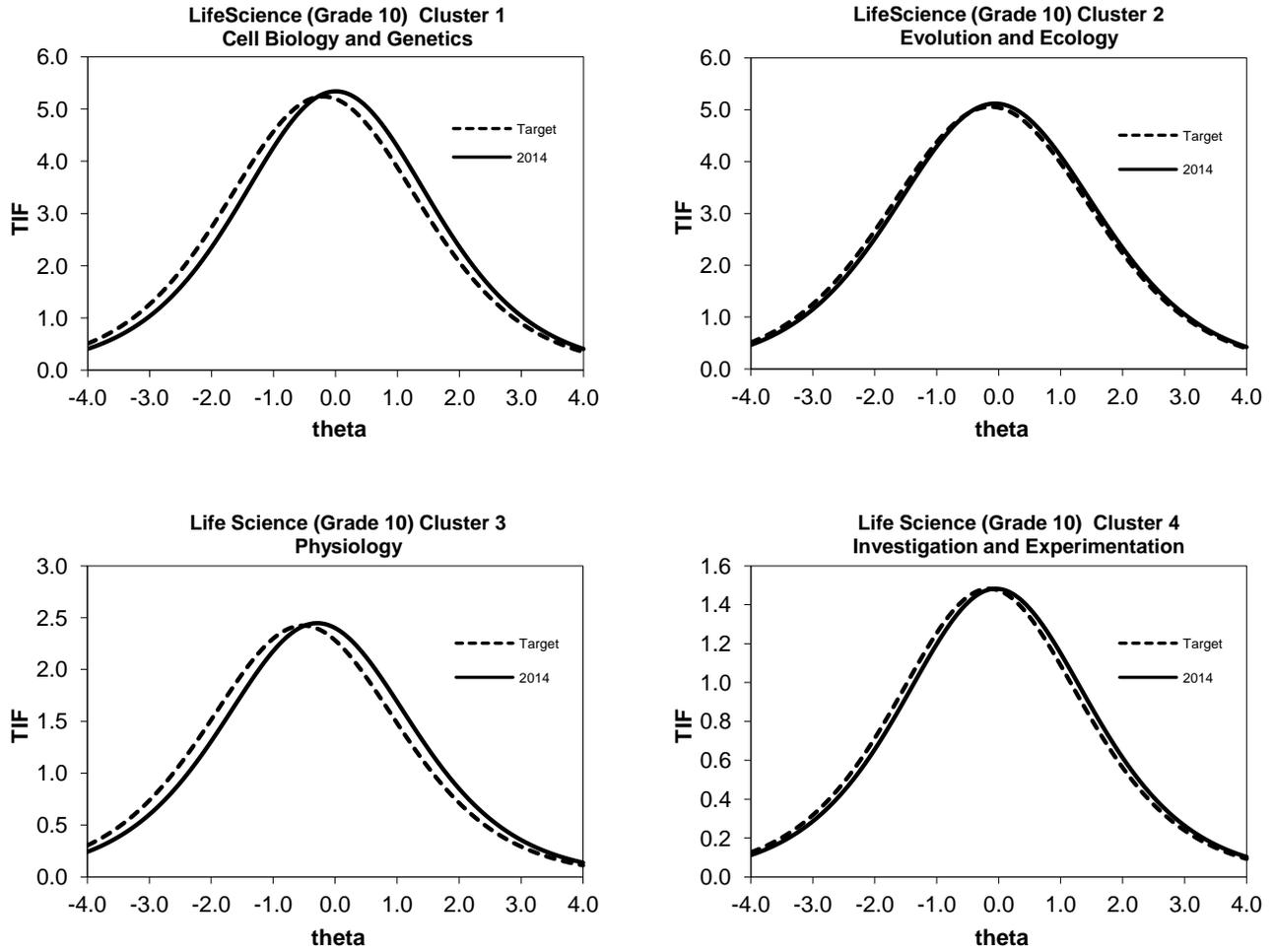


Figure 4.B.3 Plots of Target Information Functions and Projected Information for Clusters for Life Science, Grade Ten



Chapter 5: Test Administration

Test Security and Confidentiality

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

ETS's Office of Testing Integrity

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

During the test development process, ETS staff members consistently adhere to the following established security procedures:

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

Item and Data Review

As mentioned in Chapter 3, ARP meetings were not held in 2014 because there was no new item development for the 2014 CMA forms. However, before the 2014 administration,

ETS facilitated ARP meetings every year to review all newly developed CMA items and associated statistics. ETS enforced security measures at ARP meetings to protect the integrity of meeting materials using the following guidelines:

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

Item Banking

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

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

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

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

Transfer of Forms and Items to the CDE

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

Security of Electronic Files Using a Firewall

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

All electronic applications included in the Test Management System (CDE, 2014a) remain protected by the ETS firewall software at all times. Due to the sensitive nature of the student

information processed by the Test Management System, the firewall plays a significant role in maintaining an assurance of confidentiality in the users of this information.

Printing and Publishing

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

Test Administration

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

Test Delivery

Test security requires accounting for all secure materials before, during, and after each test administration. The LEA CAASPP coordinators are, therefore, required to keep all testing materials in central locked storage except during actual test administration times. 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.
- 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 Pearson. Pearson requires the following standard safeguards for security at its sites:

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

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

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

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

Data Management

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

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

For mainframe processes, Pearson utilizes Random Access Control Facility (RACF) to limit and control access to all data files (test and production), source code, object code, databases, and tables. RACF controls who is authorized to alter, update, or even read the files. All attempts to access files on the mainframe by unauthorized users are logged and monitored. In addition, Pearson uses ChangeMan, a mainframe configuration management tool, to control versions of the software and data files. ChangeMan provides another level of security, combined with RACF, to place the correct tested version of code into production. Unapproved changes are not implemented without prior review and approval.

Transfer of Scores via Secure Data Exchange

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

Statistical Analysis

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

Reporting and Posting Results

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

- Paper reports, some with individual student results and others with summary results
- A file of individual student results—available for download through the electronic reporting function of the Test Management System’s QTR module—that shows students’ scale scores and performance levels
- Encrypted files of summary results (sent to the CDE by means of SFTP) (Any summary results that have fewer than 11 students are not reported.)
- Item-level statistics based on the results which are entered into the item bank

Student Confidentiality

To meet ESEA and state requirements, LEAs must collect demographic data about students. This includes information about students’ ethnicity, parent education, disabilities, whether the student qualifies for the National School Lunch Program (NSLP), and so forth (CDE, 2014b). ETS takes precautions to prevent any of this information from becoming public or being used for anything other than testing purposes. These procedures are applied to all documents in which these student demographic data may appear, including Pre-ID files and reports.

Student Test Results

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

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

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

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

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

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

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

Procedures to Maintain Standardization

The CMA 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, as described in this section.

Test Administrators

The CMA 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 LEA and Test Site Coordinator Manual* (CDE, 2014c). This manual is described in the next section.

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

LEA CAASPP Coordinator

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

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

LEA CAASPP coordinators are also responsible for securing testing materials upon receipt, distributing testing materials to schools, tracking the materials, training and answering questions from LEA staff and 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.

Test Site Coordinator

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

Test site coordinators are responsible for making sure that the school has the proper testing materials, distributing testing materials within a school, securing materials before, during, and after the administration period, answering questions from test examiners, preparing and packaging materials to be returned to the LEA after testing, and returning the materials to the LEA. (CDE, 2014c)

Test Examiner

The CMA are administered by test examiners who may be assisted by test proctors and scribes. A test examiner is an employee of an LEA or an employee of a nonpublic, nonsectarian school (NPS) who has been trained to administer the tests and has signed a CAASPP Test Security Affidavit. Test examiners must follow the directions in the *California Modified Assessment Directions for Administration (DFA)* (CDE, 2014d) exactly.

Test Proctor

A test proctor is an employee of an LEA or a person, assigned by an NPS to implement the IEP of a student, who has received training designed to prepare the proctor to assist the test examiner in the administration of tests within the CAASPP Program (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 DFAs are manuals used by test examiners to administer the CMA to students (CDE, 2014d). Test examiners must follow all directions and guidelines and read, word-for-word, the instructions to students in "SAY" boxes to ensure test standardization.

LEA and Test Site Coordinator Manual

Test administration procedures are to be followed exactly so that all students have an equal opportunity to demonstrate their academic achievement. The *CAASPP LEA and Test Site Coordinator Manual* contributes to this goal by providing information about the responsibilities of LEA and test site coordinators, as well as those of the other staff involved in the administration cycle (CDE, 2014c). However, the manual is not intended as a substitute for the CCR, Title 5, Education (5 CCR), or to detail all of the coordinator's responsibilities.

Test Management System Manuals

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

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

Test Booklets

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

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

Universal Tools, Designated Supports, and Accommodations

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 under standard conditions. However, some students with disabilities and some English learners may need assistance when taking the CMA. 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, test examiners are responsible for marking the universal tools, designated supports, and/or accommodations used on the students' answer documents. Because the CMA were developed with modifications built into the test, non-embedded accessibility supports are not allowed. Students who require additional modifications take the content-area CST with non-embedded accessibility supports.

Scoring

The purpose of universal tools, designated supports, and accommodations is to enable students to take the CMA, not to give them an advantage over other students or to inflate their scores artificially.

Testing Incidents

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

The LEA CAASPP coordinator is responsible for immediately notifying the CDE of any irregularities or breaches that occur before, during, or after testing. The test examiner is responsible for immediately notifying the LEA CAASPP coordinator of any security breaches or testing irregularities that occur in the administration of the test. Once the LEA CAASPP coordinator and the CDE have determined that an irregularity or breach has occurred, the CDE instructs the LEA CAASPP coordinator on how and where to identify the irregularity or breach on the student answer document. The information and procedures to assist in identifying incidents and notifying the CDE are provided in the *CAASPP LEA and Test Site Coordinator Manual* (CDE, 2014c).

Social Media Security Breaches

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

There were no social media security breaches associated with the CMA in 2014.

Testing Improprieties

A testing impropriety is any event that occurs before, during, or after test administrations that does not conform to the instructions stated in the *DFAs* (CDE, 2014d) and the *CAASPP LEA and Test Site Coordinator Manual* (CDE, 2014c). These events include test administration errors, disruptions, and student cheating. Testing improprieties generally do not affect test results and are not reported to the CDE or the CAASPP 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. (2014a). *2014 Test Management System*. Sacramento, CA. <http://caaspp.org/administration/tms/>
- California Department of Education. (2014b). *2014 CAASPP Pre-ID and Extended Pre-ID Data Corrections instructions manual*. Sacramento, CA. Downloaded from http://www.startest.org/pdfs/calif-tac.pre-id_xdc_manual.2014.pdf
- California Department of Education. (2014c). *2014 CAASPP LEA and test site coordinator manual*. Sacramento, CA. Downloaded from http://caaspp.org/rsc/pdfs/CAASPP.coord_man.2014.pdf
- California Department of Education. (2014d). *2014 California Modified Assessment directions for administration*. Sacramento, CA. Downloaded from http://caaspp.org/rsc/pdfs/CMA.grade-5_dfa.2014.pdf
- California Department of Education. (2014e). *2014 CAASPP Test Administration Setup manual*. Sacramento, CA. Downloaded from http://www.startest.org/pdfs/STAR.test_admin_setup.2014.pdf
- California Department of Education. (2014f). *2014 CAASPP Order Management manual*. Sacramento, CA. Downloaded from http://www.startest.org/pdfs/calif-tac.order_mgmt.2014.pdf

Chapter 6: Performance Standards

Background

The 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 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 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 EOC 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 EOC 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, 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.

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

California employed carefully designed standard-setting procedures to facilitate the development of performance standards for each CMA. The standard-setting method used for the CMA is the Bookmark method. 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 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 (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, 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.

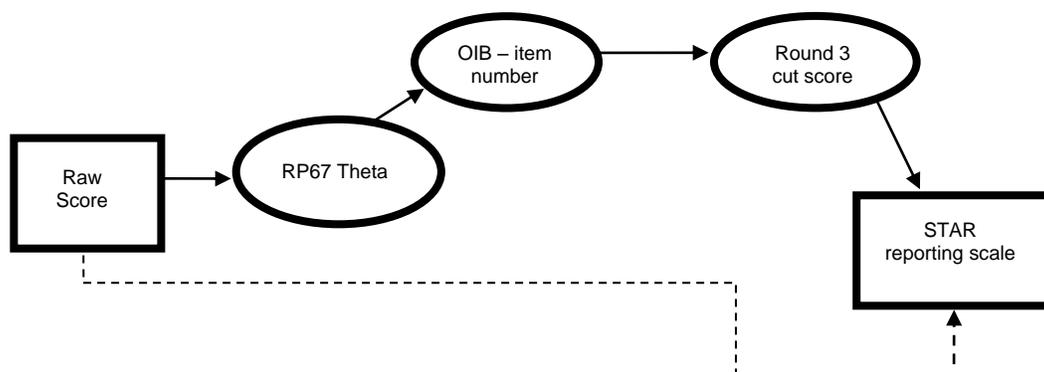
The Bookmark method is an item-mapping procedure in which panelists consider content covered by items in a specially constructed book where items are ordered from easiest to hardest based on operational student performance data from a previous test administration. The “item map,” which accompanies the 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 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 16. The cut score for each performance level is the lower bound of each scale-score range. The scale-score ranges do not change from year to year. Once established, they remain unchanged from administration to administration until such time that new performance standards are adopted.

Table 7.2 on page 62 in Chapter 7 presents the percentages of examinees meeting each performance level in 2014.

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

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

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

Scoring and Reporting Specifications

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

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

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

Scanning and Scoring

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

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

Types of Scores and Subscores

Raw Score

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

Subscore

The items in each CMA are aggregated into groups of related content standards to form reporting clusters. A subscore is a measure of an examinee's performance on the items in each reporting cluster. These results are reported both as raw scores and percent of items answered correctly. A description of the CMA reporting clusters is provided in Appendix 2.B of Chapter 2, starting on page 19.

Scale Score

Raw scores obtained on each CMA 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. 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 is categorized into one of the following performance levels:

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

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

Score Verification Procedures

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

Scoring Key Verification Process

Scoring keys, provided in the form planners, are produced by ETS and verified by performing multiple quality-control checks. The form planners contain the information about an assembled test form, including scoring keys, test name, administration year, subscore identification, and the standards and statistics associated with each item. The quality control checks that are performed before keys are finalized are listed below:

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

Overview of Score Aggregation Procedures

In order to provide meaningful results to the stakeholders, CMA 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 scores.

Individual Scores

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

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

Content Area	CMA *	No. of Items	No. of Examinees	Scale Score		Raw Score		Raw Score Percent Correct	
				Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
	5	48	26,744	345	56	28.55	7.36	59.49	15.34
Science	8	54	22,046	335	66	31.29	8.09	57.94	14.97
	10 Life Science	60	12,752	310	61	33.14	9.09	55.23	15.15

* Numbers indicate grade-level tests.

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

The numbers in the summary tables may not match exactly the results reported on the CDE’s Web site because of slight differences in the samples used to compute the statistics. The P1 data file was used for the analyses in this chapter. This file contained data collected from all LEAs but did not include corrections of demographic data through CALPADS. In addition, students with invalid scores were excluded from the tabled results.

Table 7.2 Percentages of Examinees in Each Performance Level

Content Area	CMA *	Far Below Basic	Below Basic	Basic	Proficient	Advanced	Proficient/Advanced †
	5	3%	20%	31%	31%	15%	47%
Science	8	12%	18%	31%	24%	15%	39%
	10 Life Science	16%	29%	29%	20%	6%	26%

* Numbers indicate grade-level tests.

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

Table 7.A.1 in Appendix 7.A starting on page 68 shows the distributions of scale scores for each CMA.

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.

Group Scores

Statistics summarizing student performance by each grade-level test for selected groups of students are provided starting on page 69 in Table 7.B.1 through Table 7.B.3 for the CMA.

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

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

Table 7.3 Subgroup Definitions

Subgroup	Definition
Gender	<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 – 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
Economic Status	<ul style="list-style-type: none"> • Not economically disadvantaged • Economically disadvantaged
Primary Disability	<ul style="list-style-type: none"> • Mental retardation/Intellectual disability (MR/ID) • Hard of hearing • Deafness • Speech or language impairment • Visual impairment • Emotional disturbance • Orthopedic impairment • Other health impairment • Specific learning impairment • Deaf blindness • Multiple disabilities • Autism • Traumatic brain injury

Reports Produced and Scores for Each Report

The tests that make up the CAASPP Program 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 reports. These categories and the specific reports in each category are given in Table 7.4.

Table 7.4 Types of CMA Reports

1. Summary Reports	<ul style="list-style-type: none"> ▪ CAASPP Student Master List Summary ▪ CAASPP Subgroup Summary (including Ethnicity for Economic Status)
2. Individual Reports	<ul style="list-style-type: none"> ▪ CAASPP Student Record Label ▪ CAASPP Student Master List ▪ CAASPP Student Report for CMA
3. Internet Reports	<ul style="list-style-type: none"> ▪ CMA Scores (state, county, LEA, school) ▪ CMA Summary Scores (state, county, LEA, school)

These reports are sent to the independently testing charter schools, counties, or school districts; the LEA forwards the appropriate reports to test sites or, in the case of the CAASPP Student Report, sends the report(s) to the child's parent or guardian and forwards a copy to the student's school or test site. Reports such as the CAASPP Student Report, Student Record Label, and Student Master List that include individual student results are not distributed beyond the student's school. Internet reports are described on the CDE Web site and are accessible to the public online at <http://caaspp.cde.ca.gov/>.

Because results were pre-equated, individual student scores were also available to LEAs prior to the release of summary reports, student record labels, and the master lists via electronic reporting, accessed using the QTR module to the Test Management System. This module permits LEAs to download a file containing student data that includes scale scores and performance levels for all tests taken.

Score Report Contents

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

Also given for each cluster is the average percent-correct for proficient students, which is presented as a range from the percent-correct score associated with the lowest proficient score on the total test to the percent-correct score associated with the lowest advanced score on the total test, less one percent. The average percent-correct estimates associated with the lowest proficient and advanced scores were obtained empirically for the tests that

have sample sizes of 25 or more examinees at both the minimum proficient and the minimum advanced score levels. In cases where the available sample sizes were less than 25, “data smoothing” was conducted before obtaining the averages (Lu & Smith, 2009).

Reports for students with disabilities and 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 Report and the other reports is provided in Appendix 7.C on page 75.

Score Report Applications

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

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

CMA results help LEAs and schools identify strengths and weaknesses in their instructional programs. Each year, LEAs and school staffs examine CMA 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 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 results (CDE, 2014a). It is also important to note that a student’s score in a content area contains measurement error and could vary somewhat if the student were retested.

Criteria for Interpreting Score Reports

The information presented in various reports must be interpreted with caution when making performance comparisons. When comparing scale score and performance-level results for the CMA, the user is limited to comparisons within the same content area and grade. This is because the score scales are different for each content area and grade. The user may compare scale scores for the same content area and grade, within a school, between schools, or between a school and its district, its county, or the state. The user can also

make comparisons within the same grade and content area across years. Comparing scores obtained in different grades or content areas should be avoided because the results are not on the same scale. Comparisons between raw scores or cluster scores should be limited to comparisons within not only content area and grade but also test year. For more details on the criteria for interpreting information provided on the score reports, see the *2014 CAASPP Post-Test Guide* (CDE, 2014b).

Reference

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

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

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

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

Appendix 7.A—Scale Score Distribution Tables

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

Scale Score	Grade 5	Grade 8	Grade 10
570 – 600	11	53	6
540 – 569	24	42	18
510 – 539	76	220	15
480 – 509	120	173	64
450 – 479	976	822	145
420 – 449	1,227	828	378
390 – 419	2,677	2,387	641
360 – 389	4,802	2,293	1,293
330 – 359	6,338	3,644	2,130
300 – 329	4,461	4,916	2,304
270 – 299	4,023	3,347	2,317
240 – 269	1,582	2,265	1,719
210 – 239	358	770	1,409
180 – 209	63	260	284
150 – 179	6	26	29

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 Performance Level					Mean Percent Correct in Content Area		
				Far Below Basic	Below Basic	Basic	Proficient	Advanced	Physical Sciences	Life Sciences	Earth Sciences
All valid scores	26,744	345	56	3%	20%	31%	31%	15%	57%	63%	59%
Male	17,601	347	58	3%	20%	29%	31%	17%	58%	63%	60%
Female	9,111	340	53	3%	21%	34%	31%	12%	55%	62%	58%
Gender unknown	32	322	55	6%	31%	28%	25%	9%	52%	53%	55%
American Indian	214	355	56	0%	16%	29%	40%	15%	59%	67%	61%
Asian American	849	345	58	4%	20%	27%	31%	18%	56%	62%	60%
Pacific Islander	112	337	48	3%	21%	35%	34%	8%	55%	61%	57%
Filipino	288	349	54	1%	18%	29%	36%	16%	57%	62%	63%
Hispanic	17,328	340	54	3%	22%	33%	30%	13%	56%	61%	58%
African American	2,585	334	54	4%	24%	32%	29%	10%	55%	60%	55%
White	4,745	366	60	1%	13%	25%	35%	27%	61%	69%	64%
Ethnicity unknown	623	352	57	2%	19%	28%	34%	18%	59%	65%	60%
English only	14,107	352	58	2%	17%	29%	33%	19%	58%	65%	61%
Initially fluent English prof.	262	368	63	2%	12%	23%	33%	30%	62%	69%	66%
English learner	11,808	335	53	3%	24%	34%	29%	11%	55%	59%	57%
Reclassified fluent English prof.	390	375	61	2%	11%	17%	37%	33%	63%	71%	68%
English prof. unknown	177	343	55	3%	19%	32%	33%	14%	56%	62%	59%
Autism	2,057	343	62	4%	23%	27%	29%	17%	56%	61%	60%
Deaf-blindness	2	–	–	–	–	–	–	–	–	–	–
Deafness	90	322	57	8%	27%	37%	18%	11%	51%	52%	56%
Emotional disturbance	518	343	57	3%	23%	27%	32%	15%	56%	63%	58%
Hard of hearing	209	335	54	3%	25%	33%	28%	11%	56%	56%	58%
MR/ID	390	294	45	12%	49%	27%	10%	2%	44%	46%	45%
Multiple disabilities	48	313	51	4%	44%	29%	19%	4%	47%	52%	52%
Orthopedic impairment	166	337	57	4%	21%	33%	31%	11%	53%	62%	56%
Other health impairment	2,884	354	57	2%	16%	28%	33%	20%	59%	66%	61%
Specific learning disability	16,992	346	56	2%	19%	31%	32%	16%	57%	63%	60%
Speech or language impairment	2,359	338	52	3%	21%	35%	29%	12%	55%	60%	58%
Traumatic brain injury	71	343	57	1%	21%	38%	25%	14%	55%	62%	59%
Visual impairment	66	353	66	6%	15%	18%	42%	18%	58%	66%	60%
Disability unknown	892	341	54	3%	19%	34%	32%	12%	56%	63%	58%

	No. Tested	Mean Scale Scores	Std. Dev. of Scale Scores	Percent in Performance Level					Mean Percent Correct in Content Area		
				Far Below Basic	Below Basic	Basic	Proficient	Advanced	Physical Sciences	Life Sciences	Earth Sciences
Not economically disadvantaged	5,765	361	60	2%	15%	26%	35%	24%	60%	67%	63%
Economically disadvantaged	20,636	341	55	3%	21%	32%	30%	13%	56%	61%	58%
Economic status unknown	343	334	55	3%	25%	31%	29%	12%	54%	60%	55%
Primary Ethnicity—Not Economically Disadvantaged											
American Indian	48	376	58	0%	10%	23%	44%	23%	65%	72%	66%
Asian American	351	351	59	2%	21%	26%	31%	21%	57%	64%	62%
Pacific Islander	35	340	44	0%	14%	51%	26%	9%	57%	62%	56%
Filipino	144	346	55	2%	18%	30%	38%	13%	55%	61%	63%
Hispanic	2,100	354	57	2%	16%	27%	35%	19%	58%	65%	62%
African American	457	340	58	3%	24%	30%	29%	14%	55%	63%	57%
White	2,395	373	60	1%	11%	22%	35%	31%	63%	71%	67%
Ethnicity unknown	235	361	55	1%	12%	29%	36%	22%	61%	68%	63%
Primary Ethnicity—Economically Disadvantaged											
American Indian	156	349	54	1%	18%	31%	37%	13%	57%	65%	59%
Asian American	489	341	58	5%	20%	29%	30%	15%	55%	61%	59%
Pacific Islander	76	336	50	4%	24%	28%	37%	8%	54%	60%	57%
Filipino	141	353	53	1%	18%	28%	34%	19%	58%	64%	63%
Hispanic	15,075	339	54	3%	22%	33%	29%	12%	55%	60%	58%
African American	2,094	333	53	4%	24%	32%	30%	9%	55%	60%	54%
White	2,280	360	58	1%	14%	26%	34%	23%	60%	68%	63%
Ethnicity unknown	325	349	58	2%	21%	27%	34%	17%	58%	64%	59%
Primary Ethnicity—Unknown Economic Status											
American Indian	10	—	—	—	—	—	—	—	—	—	—
Asian American	9	—	—	—	—	—	—	—	—	—	—
Pacific Islander	1	—	—	—	—	—	—	—	—	—	—
Filipino	3	—	—	—	—	—	—	—	—	—	—
Hispanic	153	328	55	5%	26%	34%	24%	11%	53%	58%	54%
African American	34	335	59	3%	24%	32%	24%	18%	53%	62%	55%
White	70	343	51	3%	17%	31%	36%	13%	57%	65%	56%
Ethnicity unknown	63	332	61	3%	30%	29%	27%	11%	55%	56%	56%

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

	No. Tested	Mean Scale Scores	Std. Dev. of Scale Scores	Percent in Performance Level						Mean Percent Correct in Content Area			
				Far Below Basic	Below Basic	Basic	Proficient	Advanced	Motion Matter	Earth Science	Investigation and Experimentation		
All valid scores	22,046	335	66	12%	18%	31%	24%	15%	63%	51%	69%	56%	
Male	14,608	337	69	12%	18%	30%	24%	16%	63%	52%	69%	56%	
Female	7,310	330	59	11%	20%	34%	25%	11%	61%	51%	68%	55%	
Gender unknown	128	327	60	14%	20%	30%	27%	9%	60%	51%	65%	54%	
American Indian	189	343	64	7%	16%	37%	23%	16%	66%	52%	71%	59%	
Asian American	598	348	74	8%	15%	32%	23%	21%	65%	55%	69%	60%	
Pacific Islander	88	328	58	13%	24%	23%	32%	9%	60%	51%	67%	56%	
Filipino	199	344	72	9%	19%	25%	30%	17%	62%	55%	70%	58%	
Hispanic	14,125	330	63	12%	19%	32%	24%	12%	62%	50%	68%	55%	
African American	2,238	319	60	16%	22%	33%	20%	9%	59%	48%	63%	51%	
White	3,804	357	72	8%	13%	25%	29%	25%	67%	57%	74%	61%	
Ethnicity unknown	805	337	64	11%	19%	31%	25%	15%	63%	52%	69%	56%	
English Only	11,115	340	68	11%	17%	29%	25%	17%	64%	53%	70%	57%	
Initially fluent English prof.	375	351	65	7%	14%	30%	28%	21%	67%	55%	73%	62%	
English learner	8,251	320	58	14%	23%	34%	21%	8%	60%	48%	66%	52%	
Reclassified fluent English prof.	1,470	368	66	4%	8%	26%	35%	27%	69%	59%	77%	66%	
English prof. unknown	835	334	66	13%	16%	31%	27%	13%	63%	51%	69%	55%	
Autism	1,291	346	77	12%	15%	27%	24%	21%	63%	55%	70%	59%	
Deaf-blindness	2	–	–	–	–	–	–	–	–	–	–	–	
Deafness	126	305	49	17%	29%	34%	17%	2%	54%	46%	61%	52%	
Emotional disturbance	653	329	72	17%	19%	28%	23%	13%	61%	50%	66%	54%	
Hard of hearing	178	333	67	12%	20%	28%	31%	10%	62%	51%	68%	57%	
MR/ID	430	284	45	33%	31%	27%	7%	1%	50%	42%	54%	41%	
Multiple disabilities	31	335	74	13%	19%	35%	6%	26%	60%	53%	68%	58%	
Orthopedic impairment	128	336	60	12%	11%	35%	30%	12%	62%	53%	68%	58%	
Other health impairment	2,295	342	69	11%	16%	29%	25%	19%	64%	53%	70%	57%	
Specific learning disability	14,927	335	64	11%	19%	32%	25%	14%	63%	51%	69%	56%	
Speech or language impairment	1,091	330	60	10%	19%	35%	26%	11%	61%	51%	68%	55%	
Traumatic brain injury	73	337	75	15%	16%	30%	23%	15%	61%	53%	68%	57%	
Visual impairment	58	334	72	16%	21%	28%	21%	16%	60%	54%	67%	53%	
Disability unknown	763	331	64	15%	17%	31%	24%	13%	62%	50%	67%	55%	
Not economically disadvantaged	4,617	352	70	9%	14%	27%	27%	22%	65%	56%	73%	60%	
Economically disadvantaged	16,129	329	63	13%	20%	32%	23%	12%	62%	50%	68%	54%	
Economic status unknown	1,300	337	70	13%	16%	30%	26%	15%	63%	52%	69%	56%	
Primary Ethnicity—Not Economically Disadvantaged													
American Indian	45	355	74	9%	11%	31%	29%	20%	65%	56%	75%	66%	
Asian American	206	350	73	7%	15%	30%	26%	22%	65%	56%	70%	61%	
Pacific Islander	23	344	55	0%	26%	17%	43%	13%	62%	55%	74%	63%	
Filipino	108	347	79	8%	21%	26%	27%	18%	62%	56%	69%	58%	
Hispanic	1,808	344	67	10%	15%	30%	26%	19%	64%	54%	72%	59%	

	No. Tested	Mean Scale Scores	Std. Dev. of Scale Scores	Percent in Performance Level					Mean Percent Correct in Content Area			
				Far Below Basic	Below Basic	Basic	Proficient	Advanced	Motion Matter	Earth Science	Investigation and Experimentation	
African American	396	328	63	13%	20%	32%	22%	13%	61%	50%	65%	55%
White	1,867	365	72	7%	12%	23%	30%	29%	68%	59%	76%	63%
Ethnicity unknown	164	348	70	11%	15%	23%	30%	21%	66%	55%	71%	57%
Primary Ethnicity—Economically Disadvantaged												
American Indian	133	337	59	8%	18%	38%	23%	14%	65%	51%	69%	57%
Asian American	340	346	74	9%	16%	34%	22%	20%	64%	54%	70%	59%
Pacific Islander	56	316	58	20%	25%	23%	25%	7%	57%	48%	65%	51%
Filipino	85	341	63	9%	18%	24%	34%	15%	63%	54%	69%	58%
Hispanic	11,760	328	62	13%	20%	33%	23%	11%	61%	49%	67%	54%
African American	1,699	317	59	16%	24%	33%	20%	8%	59%	48%	62%	50%
White	1,695	349	71	10%	14%	27%	28%	21%	65%	55%	72%	59%
Ethnicity unknown	361	333	63	10%	21%	34%	22%	13%	63%	51%	68%	55%
Primary Ethnicity—Unknown Economic Status												
American Indian	11	359	67	0%	18%	45%	9%	27%	70%	55%	81%	56%
Asian American	52	353	83	13%	8%	31%	23%	25%	67%	56%	66%	58%
Pacific Islander	9	—	—	—	—	—	—	—	—	—	—	—
Filipino	6	—	—	—	—	—	—	—	—	—	—	—
Hispanic	557	333	70	12%	19%	30%	25%	13%	62%	50%	69%	55%
African American	143	320	70	24%	13%	31%	24%	9%	61%	47%	62%	52%
White	242	354	76	10%	11%	26%	31%	22%	66%	56%	72%	59%
Ethnicity unknown	280	335	61	12%	17%	33%	27%	12%	62%	52%	69%	56%

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

	No. Tested	Mean Scale Scores	Std. Dev. of Scale Scores	Percent in Performance Level					Mean Percent Correct in Content Area			
				Far Below Basic	Below Basic	Basic	Proficient	Advanced	Cell Biology and Genetics	Evolution and Ecology	Physiology	Investigation and Experimentation
All valid scores	12,752	310	61	16%	29%	29%	20%	6%	54%	54%	59%	55%
Male	8,485	311	64	17%	29%	27%	20%	7%	54%	55%	60%	55%
Female	4,193	309	57	15%	29%	32%	19%	5%	54%	54%	59%	54%
Gender unknown	74	305	64	23%	28%	19%	22%	8%	54%	53%	56%	54%
American Indian	126	305	59	19%	28%	29%	19%	6%	53%	53%	60%	54%
Asian American	311	317	62	13%	28%	30%	22%	7%	56%	56%	61%	59%
Pacific Islander	64	308	63	16%	38%	19%	19%	9%	56%	52%	59%	56%
Filipino	149	328	62	8%	26%	27%	32%	8%	57%	59%	67%	60%
Hispanic	8,055	307	59	17%	30%	29%	18%	5%	53%	54%	58%	54%
African American	1,346	295	56	22%	33%	27%	16%	3%	51%	50%	56%	50%
White	2,354	328	68	12%	21%	28%	26%	12%	58%	59%	65%	59%
Ethnicity unknown	347	311	62	17%	26%	30%	21%	6%	55%	54%	59%	56%
English Only	6,585	315	64	15%	27%	28%	22%	8%	56%	55%	61%	56%
Initially fluent English prof.	299	325	63	11%	23%	32%	26%	8%	57%	59%	64%	59%
English learner	4,411	294	52	21%	34%	29%	13%	2%	50%	51%	55%	50%
Reclassified fluent English prof.	1,054	341	62	7%	19%	29%	32%	13%	60%	62%	68%	65%
English prof. unknown	403	305	58	18%	28%	29%	21%	3%	54%	53%	58%	55%
Autism	716	331	70	11%	24%	28%	24%	13%	58%	60%	65%	59%
Deaf-blindness	3	–	–	–	–	–	–	–	–	–	–	–
Deafness	113	296	47	13%	44%	31%	9%	3%	49%	53%	56%	50%
Emotional disturbance	452	307	69	22%	25%	25%	18%	9%	54%	53%	59%	54%
Hard of hearing	129	311	59	14%	30%	32%	18%	6%	54%	55%	61%	57%
MR/ID	262	264	43	39%	42%	15%	5%	0%	44%	43%	45%	42%
Multiple disabilities	29	290	50	21%	41%	21%	14%	3%	48%	50%	56%	49%
Orthopedic impairment	96	308	76	25%	26%	26%	13%	10%	53%	53%	59%	53%
Other health impairment	1,226	322	66	14%	24%	29%	23%	10%	57%	57%	64%	57%
Specific learning disability	8,852	309	59	16%	29%	29%	20%	6%	54%	54%	59%	54%
Speech or language impairment	468	310	55	12%	30%	35%	18%	5%	55%	54%	60%	56%
Traumatic brain injury	41	299	56	22%	32%	32%	7%	7%	53%	53%	55%	48%
Visual impairment	39	318	74	15%	38%	10%	21%	15%	57%	55%	59%	57%
Disability unknown	326	299	61	23%	27%	29%	17%	4%	52%	52%	55%	53%
Not economically disadvantaged	3,224	323	66	13%	24%	29%	25%	10%	57%	57%	64%	58%
Economically disadvantaged	9,010	305	59	18%	31%	29%	18%	5%	53%	53%	58%	53%
Economic status unknown	518	306	60	19%	27%	30%	20%	5%	53%	54%	59%	54%
Primary Ethnicity—Not Economically Disadvantaged												
American Indian	42	314	63	17%	21%	29%	29%	5%	55%	55%	64%	54%
Asian American	127	335	66	7%	25%	28%	28%	12%	59%	60%	67%	62%
Pacific Islander	24	308	61	13%	38%	21%	21%	8%	55%	51%	60%	57%
Filipino	79	336	60	5%	23%	29%	34%	9%	59%	62%	68%	61%
Hispanic	1,305	318	62	14%	26%	29%	23%	8%	56%	56%	62%	57%

	No. Tested	Mean Scale Scores	Std. Dev. of Scale Scores	Percent in Performance Level					Mean Percent Correct in Content Area			
				Far Below Basic	Below Basic	Basic	Proficient	Advanced	Cell Biology and Genetics	Evolution and Ecology	Physiology	Investigation and Experimentation
African American	332	303	58	17%	33%	27%	18%	5%	53%	52%	58%	51%
White	1,223	334	69	11%	19%	28%	27%	14%	60%	60%	66%	61%
Ethnicity unknown	92	312	61	17%	17%	41%	18%	5%	55%	55%	61%	56%
Primary Ethnicity—Economically Disadvantaged												
American Indian	83	299	56	20%	31%	29%	13%	6%	52%	51%	57%	53%
Asian American	177	306	55	17%	30%	32%	18%	3%	53%	53%	58%	57%
Pacific Islander	38	312	65	16%	39%	16%	18%	11%	56%	53%	61%	55%
Filipino	65	320	62	11%	29%	25%	28%	8%	56%	56%	65%	59%
Hispanic	6,471	304	58	17%	31%	29%	18%	5%	53%	53%	58%	53%
African American	955	292	55	23%	33%	27%	15%	2%	51%	49%	55%	49%
White	1,063	321	66	14%	24%	28%	24%	10%	57%	57%	63%	55%
Ethnicity unknown	158	315	64	13%	32%	27%	20%	8%	56%	55%	60%	57%
Primary Ethnicity—Unknown Economic Status												
American Indian	1	—	—	—	—	—	—	—	—	—	—	—
Asian American	7	—	—	—	—	—	—	—	—	—	—	—
Pacific Islander	2	—	—	—	—	—	—	—	—	—	—	—
Filipino	5	—	—	—	—	—	—	—	—	—	—	—
Hispanic	279	304	58	18%	30%	30%	17%	5%	52%	53%	60%	53%
African American	59	293	52	24%	31%	31%	14%	2%	50%	51%	54%	53%
White	68	329	67	10%	18%	35%	28%	9%	57%	60%	67%	56%
Ethnicity unknown	97	305	61	22%	26%	24%	25%	4%	54%	53%	55%	55%

Appendix 7.C—Types of Score Reports

Table 7.C.1 Score Reports Reflecting CMA Results

2014 CAASPP CMA PRINTED REPORTS	
DESCRIPTION	DISTRIBUTION
The CMA Student Report	
<p>This report provides parents/guardians and teachers with the student's results, presented in tables and graphs.</p> <p>Data presented include the following:</p> <ul style="list-style-type: none"> • Scale scores • Performance levels (advanced, proficient, basic, below basic, and far below basic) • Number and percent correct in each reporting cluster • Comparison of the student's scores on specific reporting clusters to the range of scores of students statewide who scored proficient on the total test 	<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 distributed by the LEA to parents/ guardians.</p>
Student Record Label	
<p>These reports are printed on adhesive labels to be affixed to the student's permanent school records. Each student shall have an individual record of accomplishment that includes CAASPP testing results (see California <i>EC</i> Section 60607[a]).</p> <p>Data presented include the following for each content area tested:</p> <ul style="list-style-type: none"> • Scale scores • Performance levels 	<p>This report includes individual student results and is not distributed beyond the student's school.</p>
Student Master List	
<p>This report is an alphabetical roster that presents individual student results. Data include the following:</p> <ul style="list-style-type: none"> • Percent correct for each reporting cluster within each content area tested • A scale score and a performance level for each content area tested 	<p>This report provides administrators and teachers with all students' results within each grade or within each grade and year-round schedule at a school.</p> <p>Because this report includes individual student results, it is not distributed beyond the student's school. It is recommended that Student Master List reports be retained until the grade level exits the school.</p>
Student Master List Summary	
<p>This report summarizes student results at the school, district, county, and state levels for each grade. It does not include any individual student information.</p> <p>For each CMA, the following data are summarized:</p> <ul style="list-style-type: none"> • By content area tests: <ul style="list-style-type: none"> – Number of students enrolled – Number and percent of students tested – Number and percent of valid scores 	<p>This report is a resource for evaluators, researchers, teachers, parents/guardians, community members, and administrators.</p> <p>One copy is packaged for the school and one for the LEA.</p> <p>This report is also produced for school districts, counties, and the state.</p> <p>Note: The data in this report may be shared with parents/guardians, community members, and the media only if the data are for 11 or more students. It</p>

2014 CAASPP CMA PRINTED REPORTS	
DESCRIPTION	DISTRIBUTION
<ul style="list-style-type: none"> – Number tested with scores – Mean percent correct • Mean scale score • Scale score standard deviation • Number and percent of students scoring at each performance level • The number of items for each reporting cluster and the mean percent correct • For grades four and seven, the percent of students achieving each writing response score 	is recommended that summary reports be retained until the grade level exits the school.
Subgroup Summary	
<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 <p>These reports contain no individual student-identifying information and are aggregated at the school, district, 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 • 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 	<p>This report is a resource for evaluators, researchers, teachers, parents/guardians, community members, and administrators.</p> <p>One copy is packaged for the school and one for the LEA.</p> <p>This report is also produced for school districts, counties, and the state.</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. It is recommended that summary reports be retained until the grade level exits the school.</p>
Subgroup Summary—Ethnicity for Economic Status	
<p>This report, a part of the Subgroup Summary, disaggregates and reports results by cross-referencing each ethnicity with economic status. The economic status for each student is “economically disadvantaged,” “not economically disadvantaged,” or “economic status unknown.” A student is defined as “economically disadvantaged” if the most educated parent of the student, as indicated in the answer document or Pre-ID, has not received a high school diploma or the student is eligible to participate in the free or reduced-price lunch program also known as the National School Lunch</p>	<p>This report is a resource for evaluators, researchers, teachers, parents/guardians, community members, and administrators.</p> <p>One copy is packaged for the school and one for the LEA.</p> <p>This report is also produced for school districts, counties, and the state.</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. It is recommended that summary reports be retained</p>

2014 CAASPP CMA PRINTED REPORTS	
DESCRIPTION	DISTRIBUTION
<p>Program (NSLP).</p> <p>As with the standard Subgroup Summary, this disaggregation contains no individual student-identifying information and is aggregated at the school, district, county, and state levels.</p> <p>For each subgroup within a report, and for the total number of students, the following data are included:</p> <ul style="list-style-type: none"> • Total number tested in the subgroup • Percent of enrollment tested in the subgroup • Number and percent of valid scores • Number tested who received scores • Mean scale scores • Standard deviation of scale scores • Number and percent of students scoring at each performance level 	<p>until the grade level exits the school.</p>

Chapter 8: Analyses

Background

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

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

1. Appendix 8.A on page 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 96 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 105 presents the scoring tables obtained as a result of the IRT equating process.

Samples Used for the Analyses

CMA 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 P1 data file, which contained the entire test-taking population and all the student records used in the July 29, 2014, reporting of CAASPP results.

During the 2014 administration, neither IRT calibrations nor scaling are implemented because of the intact form use and pre-equating. For the used intact forms without any replacement or edited items, the IRT results were derived based on the equating sample of the previous administration which can be found in Appendix D of the *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.

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

Content Area	CMA *	Admin. Year	No. of Items	No. of Examinees	Mean		Median	
					p -value	Pt-Bis	p -value	Pt-Bis
	5	2014	48	26,744	0.59	0.32	0.59	0.34
Science	8	2014	54	22,046	0.58	0.31	0.56	0.33
	10 Life Science	2014	60	12,752	0.55	0.31	0.57	0.30

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

Reliability Analyses

Reliability focuses on the extent to which differences in test scores reflect true differences in the knowledge, ability, or skill being tested, rather than fluctuations due to chance or random factors. The variance in the distribution of test scores—essentially, the differences among individuals—is partly due to real differences in the knowledge, skill, or ability being tested (true-score variance) and partly due to random unsystematic errors in the measurement process (error variance).

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

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

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

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

Content Area	CMA *	No. of Items	No. of Examinees	Reliability	Scale Score			Raw Score		
					Mean	Std. Dev.	SEM	Mean	Std. Dev.	SEM
	5	48	26,744	0.82	345	56	24.17	28.55	7.36	3.16
Science	8	54	22,046	0.82	335	66	27.63	31.29	8.09	3.39
	10 Life Science	60	12,752	0.84	310	61	24.29	33.14	9.09	3.59

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

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 96. 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.5 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.6 through Table 8.B.8, shows the same analyses for the subgroups based on primary ethnicity within economic status and gender within economic status. Table 8.B.9 and Table 8.B.10 present the reliabilities for subgroups based on primary disability.

Test-level reliabilities for the various subgroups are compiled in Table 8.B.11 through Table 8.B.17. The corresponding cluster-level reliabilities are provided in Table 8.B.18 through Table 8.B.24.

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 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.

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

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

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

Decision Classification Analyses

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

Decision accuracy describes the extent to which examinees are classified in the same way as they would be on the basis of the average of all possible forms of a test. Decision accuracy answers the following question: How does the actual classification of test-takers, based on their single-form scores, agree with the classification that would be made on the basis of their true scores, if their true scores were somehow known? RELCLASS-COMP estimates decision accuracy using an estimated multivariate distribution of reported classifications on the current form of the 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.25 through Table 8.B.27 in Appendix 8.B, starting on page 103.

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

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

Validity Evidence

Validity refers to the degree to which each interpretation or use of a test score is supported by evidence that is gathered (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999; ETS, 2002). It is a central concern underlying the development, administration, and scoring of a test and the uses and interpretations of test scores.

Validation is the process of accumulating evidence to support each proposed score interpretation or use. It involves more than a single study or gathering of one particular kind of evidence. Validation involves multiple investigations and various kinds of evidence (AERA, APA, & NCME, 1999; Cronbach, 1971; ETS, 2002; Kane, 2006). The process begins with test design and continues through the entire assessment process, including item development and field testing, analyses of item and test data, test scaling, scoring, and score reporting.

This section presents the evidence gathered to support the intended uses and interpretations of scores for the CMA testing program. The description is organized in the manner prescribed by *The Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999). These standards require a clear definition of the purpose of the test, which includes a description of the qualities—called constructs—that are to be assessed by a test, the population to be assessed, as well as how the scores are to be interpreted and used.

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

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

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

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

Purpose of the CMA

As mentioned in Chapter 1, the CMA for Science comprise the CAASPP System implementation of the remaining paper-pencil tests for students whose IEP or Section 504 plan require they take the CMA. The purpose of the CMA is to allow students with disabilities greater access to an assessment that helps measure their achievement with respect to California’s content standards.

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 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 can be found on the CDE STAR CMA Blueprints Web page at <http://www.cde.ca.gov/ta/tg/sr/cmablueprints.asp>. ETS developed all CMA test items to conform to the SBE-approved content standards and test blueprints.

Interpretations and Uses of the Scores Generated

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

Reporting cluster scores, also called subscores, are used to draw inferences about a student's achievement in each of several specific knowledge or skill areas covered by each test. Reporting cluster results compare an individual student's percent-correct score to the average percent-correct for the state as a whole. The range of scores for students who scored proficient on the total test is also provided for each cluster using a percent-correct metric. The reference points for this range are: (1) the average percent-correct for students who received the lowest score qualifying for the proficient performance level; and (2) the average percent-correct for students who received the lowest score qualifying for the advanced performance level, minus one percent. A detailed description of the uses and applications of CMA scores is presented in Chapter 7, which starts on page 59.

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

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

These are the only uses and interpretations of scores for which validity evidence has been gathered. If the user wishes to interpret or use the scores in other ways, the user is cautioned that the validity of doing so has not been established (AERA, APA, & NCME, 1999, Standard 1.3). The user is advised to gather evidence to support these additional interpretations or uses (AERA, APA, & NCME, 1999, Standard 1.4).

Intended Test Population(s)

California public school students in grades five, eight, and ten who meet certain eligibility criteria are the intended test population for the CMA in science. Only those students whose parents/ guardians have submitted written requests to exempt them from CAASPP System testing do not take a grade-level science test.

Validity Evidence Collected

Evidence Based on Content

According to *The Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999), analyses that demonstrate a strong relationship between a test's content and the construct that the test was designed to measure can provide important evidence of validity. In current K–12 testing, the construct of interest usually is operationally defined by state content standards and the test blueprints that specify the content, format, and scoring of items that are admissible measures of the knowledge and skills described in the content standards. Evidence that the items meet these specifications and represent the domain of knowledge and skills referenced by the standards supports the inference that students' scores on these items can appropriately be regarded as measures of the intended construct.

As noted in the AERA, APA, and NCME *Standards* (1999), evidence based on test content may involve logical analyses of test content in which experts judge the adequacy with which the test content conforms to the test specifications and represents the intended domain of content. Such reviews can also be used to determine whether the test content contains material that is not relevant to the construct of interest. Analyses of test content may also involve the use of empirical evidence of item quality.

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

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

Specifications and blueprints—ETS maintains item specifications for each CMA. 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 25. Once the items were developed and field-tested, ETS selected all CMA test items to conform to the SBE-approved California content standards and test blueprints. Test blueprints for the CMA 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. 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. The test blueprints for the CMA 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 is presented in Chapter 3, starting on page 25.

Item review process—Chapter 3 explains in detail the extensive item review process applied to items that were written for use in the CMA. In brief, items written for the CMA 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 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 35.

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 31.

Evidence Based on Relations to Other Variables

Empirical results concerning the relationships between the score on a test and measures of other variables external to the test can also provide evidence of validity when these relationships are found to be consistent with the definition of the construct that the test is intended to measure. As indicated in the *Standards* (AERA, APA, & NCME, 1999), the variables investigated can include other tests that measure the same construct and different constructs, criterion measures that scores on the test are expected to predict, as well as demographic characteristics of examinees that are expected to be related and unrelated to test performance.

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, 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's Technical Reports and Studies Web page at <http://www.cde.ca.gov/ta/tg/sr/technicalrpts.asp>. The year of original administration for each CMA for Science is shown in Table 8.4.

Table 8.4 Original Year of Administration for the CMA

Content Area	CMA *	Year
	5	2011
Science	8	2012
	10 Life Science	2012

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

Evidence Based on Response Processes

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

Evidence Based on Internal Structure

As suggested by the *Standards* (AERA, APA, & NCME, 1999), evidence of validity can also be obtained from studies of the properties of the item scores and the relationship between these scores and scores on components of the test. To the extent that the score properties and relationships found are consistent with the definition of the construct measured by the test, support is gained for interpreting these scores as measures of the construct.

For the CMA, 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 are presented in Table 8.A.1.

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

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 tests had average p -values that range from 0.55 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 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 0.84.

Reporting cluster reliabilities—For each CMA, 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 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.10.

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.25 through Table 8.B.27 in Appendix 8.B; these tables start on page 103. 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. 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 (ESEA).

Two factors corresponding to the ELA 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, 1999, p. 18). When this is the case, evidence that the expected benefits accrue will provide support for the intended use of the scores. The CDE and ETS are in the process of determining what kinds of information can be gathered to assess the consequences of administration of the CMA.

IRT Analyses

Post-Equating

Prior to 2013, the CMA 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 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 prior to 2013 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 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 item response theory (IRT), all the item parameter estimates were placed on the reference scale in their previous administrations

through the post-equating procedure described above. For all CMA using intact forms, 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's 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 2013 administration—once the IRT b -values were placed on the item bank scale, analyses were performed to assess the overall test difficulty, the difficulty level of reporting clusters, and the distribution of items in a particular range of item difficulty.

During the 2014 administration, 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. 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 item response theory (IRT) models that item parameters remain invariant across samples given a similar ability distribution. To produce results that are sufficiently accurate for high-stakes decisions, 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 administration, for all CMA 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 administered in 2014 are presented in Table 8.C.1 through Table 8.C.3 starting on page 105. The raw scores and corresponding transformed scale scores are listed in those tables. The scale scores were truncated at both ends of the scale so that the minimum reported scale score was 150 and the maximum reported scale score was 600. The scale scores defining the various performance-level cut points are presented in Table 2.1, which is in Chapter 2 on page 16.

Differential Item Functioning Analyses

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

Prior to the 2013 administration, DIF analyses were performed based on the FIA sample and were performed on all operational items and on all field-test items for which sufficient student samples were available. DIF analyses are not implemented during the 2014 administration because 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, can be found in Appendix E of the *CMA Technical Report* in the year the form was administered originally; see Table 8.4 on page 87 for administration years.

The statistical procedure of DIF analysis that was conducted prior to the 2013 administration is described in this section.

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

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

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

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

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

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

The factors considered in the DIF analyses included gender, ethnicity, level of English-language fluency, and primary disability. Note, however, that analyses of English learners on the CMA for ELA were presented for readers’ interest. Differential performance on an ELA item because of the language difficulties of nonnative speakers did not indicate that an item was unfair or biased.

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

Tables showed the distributions of field-test items across the DIF category classifications for the CMA. 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. 1999. *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, Grades Five, Eight, and Ten—Current Year (2014) and Original Year of Administration

Item-by-item p -value and Point Biserial for Science, Grades Five, Eight, and Ten												
CMA Years	Grade 5				Grade 8				Life Science (Grade 10)			
	2014		2011		2014		2012		2014		2012	
Items	p -value	Pt-Bis	p -value	Pt-Bis	p -value	Pt-Bis						
1	0.71	0.31	0.70	0.34	0.69	0.35	0.70	0.35	0.74	0.23	0.69	0.29
2	0.58	0.25	0.59	0.27	0.77	0.25	0.77	0.29	0.43	0.28	0.43	0.27
3	0.44	0.26	0.44	0.29	0.66	0.28	0.65	0.27	0.61	0.21	0.59	0.24
4	0.59	0.35	0.61	0.34	0.64	0.21	0.63	0.22	0.71	0.21	0.69	0.28
5	0.57	0.28	0.55	0.30	0.71	0.32	0.70	0.33	0.47	0.21	0.47	0.22
6	0.49	0.26	0.49	0.27	0.50	0.27	0.50	0.27	0.42	0.28	0.38	0.24
7	0.50	0.24	0.52	0.27	0.61	0.36	0.60	0.35	0.30	-0.05	0.31	-0.05
8	0.76	0.43	0.72	0.45	0.56	0.34	0.57	0.33	0.49	0.28	0.46	0.25
9	0.84	0.25	0.83	0.28	0.89	0.35	0.87	0.36	0.43	0.19	0.42	0.19
10	0.30	0.09	0.29	0.08	0.61	0.36	0.60	0.36	0.58	0.30	0.55	0.30
11	0.52	0.33	0.50	0.35	0.66	0.36	0.64	0.36	0.44	0.23	0.42	0.19
12	0.56	0.38	0.56	0.40	0.53	0.33	0.52	0.34	0.64	0.29	0.60	0.29
13	0.46	0.34	0.47	0.34	0.55	0.31	0.53	0.32	0.61	0.22	0.59	0.25
14	0.62	0.25	0.63	0.26	0.59	0.38	0.58	0.39	0.50	0.25	0.49	0.26
15	0.49	0.12	0.48	0.08	0.55	0.41	0.54	0.40	0.53	0.27	0.52	0.26
16	0.78	0.39	0.76	0.38	0.52	0.20	0.51	0.19	0.46	0.30	0.44	0.29
17	0.61	0.48	0.60	0.47	0.68	0.27	0.67	0.29	0.60	0.33	0.58	0.36
18	0.70	0.29	0.70	0.31	0.63	0.21	0.61	0.23	0.49	0.29	0.47	0.31
19	0.57	0.32	0.57	0.32	0.49	0.22	0.47	0.26	0.49	0.29	0.48	0.30
20	0.48	0.29	0.45	0.28	0.54	0.35	0.53	0.37	0.56	0.34	0.54	0.33
21	0.61	0.36	0.60	0.36	0.70	0.35	0.67	0.37	0.66	0.42	0.62	0.40
22	0.49	0.35	0.47	0.32	0.56	0.22	0.57	0.22	0.52	0.28	0.50	0.26
23	0.43	0.28	0.42	0.28	0.54	0.33	0.52	0.34	0.66	0.39	0.63	0.38
24	0.63	0.30	0.63	0.31	0.51	0.38	0.50	0.39	0.65	0.36	0.61	0.36
25	0.65	0.32	0.64	0.32	0.46	0.28	0.46	0.28	0.57	0.42	0.51	0.39
26	0.66	0.46	0.63	0.47	0.44	0.28	0.45	0.29	0.66	0.43	0.61	0.43
27	0.84	0.36	0.82	0.38	0.47	0.22	0.47	0.23	0.52	0.30	0.50	0.32
28	0.59	0.39	0.54	0.38	0.46	0.31	0.45	0.30	0.71	0.39	0.69	0.42
29	0.72	0.44	0.71	0.44	0.48	0.33	0.45	0.32	0.70	0.47	0.65	0.46
30	0.65	0.38	0.63	0.39	0.58	0.38	0.58	0.37	0.65	0.39	0.61	0.38
31	0.61	0.43	0.60	0.42	0.69	0.41	0.66	0.42	0.53	0.26	0.51	0.27
32	0.64	0.32	0.62	0.33	0.68	0.47	0.68	0.46	0.51	0.34	0.51	0.35
33	0.72	0.42	0.71	0.40	0.44	0.25	0.44	0.26	0.63	0.48	0.61	0.49
34	0.44	0.29	0.42	0.29	0.53	0.22	0.52	0.23	0.59	0.44	0.55	0.41
35	0.67	0.25	0.69	0.26	0.58	0.37	0.57	0.35	0.61	0.41	0.58	0.39
36	0.51	0.35	0.53	0.35	0.41	0.20	0.41	0.21	0.46	0.28	0.44	0.26
37	0.52	0.34	0.51	0.34	0.64	0.40	0.62	0.39	0.62	0.44	0.59	0.44
38	0.72	0.40	0.70	0.41	0.75	0.40	0.75	0.41	0.49	0.25	0.47	0.25
39	0.50	0.28	0.50	0.30	0.83	0.36	0.82	0.38	0.68	0.40	0.65	0.42
40	0.56	0.19	0.55	0.19	0.58	0.22	0.56	0.20	0.68	0.54	0.63	0.54
41	0.64	0.39	0.63	0.41	0.71	0.32	0.72	0.33	0.59	0.38	0.56	0.39
42	0.42	0.22	0.43	0.25	0.61	0.36	0.61	0.38	0.54	0.32	0.51	0.30
43	0.54	0.35	0.55	0.36	0.69	0.42	0.68	0.41	0.67	0.42	0.65	0.41
44	0.82	0.37	0.84	0.39	0.52	0.40	0.54	0.40	0.46	0.30	0.45	0.27
45	0.60	0.38	0.62	0.39	0.44	0.17	0.44	0.17	0.33	0.14	0.33	0.14
46	0.54	0.31	0.56	0.33	0.56	0.36	0.54	0.35	0.71	0.39	0.67	0.43
47	0.53	0.35	0.54	0.33	0.46	0.30	0.46	0.30	0.47	0.22	0.46	0.23
48	0.75	0.35	0.75	0.35	0.51	0.31	0.50	0.33	0.44	0.27	0.42	0.25
49	0.40	0.16	0.39	0.18	0.72	0.35	0.70	0.37
50	0.59	0.32	0.59	0.32	0.26	0.12	0.26	0.08
51	0.55	0.38	0.54	0.37	0.66	0.31	0.62	0.34
52	0.34	0.14	0.35	0.14	0.49	0.36	0.47	0.34
53	0.55	0.33	0.55	0.34	0.36	0.16	0.36	0.15
54	0.66	0.40	0.64	0.40	0.61	0.41	0.56	0.38
55	0.41	0.28	0.39	0.26
56	0.70	0.47	0.66	0.47
57	0.44	0.21	0.41	0.21
58	0.44	0.32	0.42	0.30
59	0.64	0.43	0.62	0.41
60	0.62	0.29	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 Reliabilities and Intercorrelations for Science							
Subscore Area	No. of items	Intercorrelations				Reliab.	SEM
Grade 5		1.	2.	3.	4.		
1. Physical Sciences	16	1.00	–	–		0.52	1.85
2. Life Sciences	16	0.54	1.00	–		0.69	1.77
3. Earth Sciences	16	0.49	0.59	1.00		0.62	1.82
Grade 8		1.	2.	3.	4.		
1. Motion	19	1.00	–	–	–	0.60	1.99
2. Matter	23	0.55	1.00	–	–	0.67	2.26
3. Earth Science	7	0.46	0.48	1.00	–	0.55	1.12
4. Investigation and Experimentation	5	0.50	0.48	0.36	1.00	0.40	1.04
Grade 10		1.	2.	3.	4.		
1. Cell Biology and Genetics	22	1.00	–	–	–	0.58	2.21
2. Evolution and Ecology	22	0.56	1.00	–	–	0.70	2.14
3. Physiology	10	0.54	0.61	1.00	–	0.62	1.42
4. Investigation and Experimentation	6	0.45	0.53	0.46	1.00	0.42	1.14

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

Content Area	CMA *	Male			Female			Unknown		
		N	Rel	SEM	N	Rel	SEM	N	Rel	SEM
Science	5	17,601	0.82	3.14	9,111	0.80	3.18	32	0.82	3.27
	8	14,608	0.84	3.37	7,310	0.79	3.43	128	0.78	3.46
	10 Life Science	8,485	0.85	3.58	4,193	0.82	3.61	74	0.87	3.58

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

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

Content Area	CMA *	Econ Disadv.			Not Econ Disadv.			Unknown		
		N	Rel	SEM	N	Rel	SEM	N	Rel	SEM
Science	5	20,636	0.81	3.18	5,765	0.83	3.08	343	0.81	3.20
	8	16,129	0.81	3.42	4,617	0.84	3.31	1,300	0.84	3.38
	10 Life Science	9,010	0.83	3.61	3,224	0.86	3.53	518	0.84	3.61

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

Table 8.B.4 Reliabilities and SEMs for the CMA by English-language Fluency

Content Area	CMA *	English Only			I-FEP ¹			English Learner			R-FEP ²			Unknown		
		N	Rel	SEM	N	Rel	SEM	N	Rel	SEM	N	Rel	SEM	N	Rel	SEM
Science	5	14,107	0.82	3.13	262	0.85	3.01	11,808	0.79	3.21	390	0.84	2.98	177	0.81	3.17
	8	11,115	0.84	3.36	375	0.82	3.33	8,251	0.78	3.45	1,470	0.82	3.26	835	0.82	3.39
	10 Life Science	6,585	0.86	3.57	299	0.84	3.55	4,411	0.79	3.66	1,054	0.84	3.46	403	0.83	3.62

¹ Initially Fluent English Proficient

² Reclassified Fluent English Proficient

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

Table 8.B.5 Reliabilities and SEMs for the CMA by Primary Ethnicity

Content Area	CMA*	African American			American Indian			Asian			Filipino			Hispanic			Pacific Islander			White			Unknown		
		N	Rel	SEM	N	Rel	SEM	N	Rel	SEM	N	Rel	SEM	N	Rel	SEM	N	Rel	SEM	N	Rel	SEM	N	Rel	SEM
	5	2,585	0.80	3.21	214	0.80	3.12	849	0.83	3.15	288	0.80	3.14	17,328	0.80	3.18	112	0.75	3.24	4,745	0.83	3.04	623	0.82	3.12
Science	8	2,238	0.80	3.46	189	0.81	3.37	598	0.85	3.33	199	0.84	3.37	14,125	0.81	3.41	88	0.79	3.45	3,804	0.85	3.28	805	0.82	3.39
	10 Life Science	1,346	0.82	3.65	126	0.84	3.62	311	0.85	3.56	149	0.84	3.54	8,055	0.83	3.61	64	0.86	3.59	2,354	0.87	3.50	347	0.85	3.59

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

Table 8.B.6 Reliabilities and SEMs for the CMA by Primary Ethnicity-for-Not-Economically-Disadvantaged

Content Area	CMA *	African American			American Indian			Asian			Filipino			Hispanic			Pacific Islander			White			Unknown		
		N	Rel	SEM	N	Rel	SEM	N	Rel	SEM	N	Rel	SEM	N	Rel	SEM	N	Rel	SEM	N	Rel	SEM	N	Rel	SEM
	5	457	0.83	3.18	48	0.82	2.97	351	0.84	3.13	144	0.82	3.14	2,100	0.82	3.11	35	0.69	3.24	2,395	0.83	3.01	235	0.80	3.09
Science	8	396	0.81	3.43	45	0.85	3.29	206	0.85	3.32	108	0.86	3.35	1,808	0.83	3.35	23	0.76	3.42	1,867	0.85	3.24	164	0.85	3.32
	10 Life Science	332	0.83	3.64	42	0.86	3.58	127	0.86	3.47	79	0.82	3.51	1,305	0.85	3.57	24	0.85	3.60	1,223	0.87	3.46	92	0.85	3.59

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

Table 8.B.7 Reliabilities and SEMs for the CMA by Primary Ethnicity-for-Economically-Disadvantaged

Content Area	CMA *	African American			American Indian			Asian			Filipino			Hispanic			Pacific Islander			White			Unknown		
		N	Rel	SEM	N	Rel	SEM	N	Rel	SEM	N	Rel	SEM	N	Rel	SEM	N	Rel	SEM	N	Rel	SEM	N	Rel	SEM
	5	2,094	0.80	3.22	156	0.79	3.15	489	0.83	3.16	141	0.79	3.14	15,075	0.80	3.19	76	0.78	3.24	2,280	0.83	3.08	325	0.83	3.13
Science	8	1,699	0.79	3.47	133	0.79	3.39	340	0.85	3.35	85	0.81	3.39	11,760	0.80	3.42	56	0.79	3.48	1,695	0.85	3.32	361	0.81	3.41
	10 Life Science	955	0.82	3.66	83	0.82	3.65	177	0.82	3.62	65	0.84	3.57	6,471	0.83	3.62	38	0.86	3.59	1,063	0.86	3.53	158	0.85	3.59

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

Table 8.B.8 Reliabilities and SEMs for the CMA by Gender by Economic Status

Content Area	CMA *	Economically Disadvantaged									Not Economically Disadvantaged								
		Male			Female			Unknown			Male			Female			Unknown		
		N	Rel	SEM	N	Rel	SEM	N	Rel	SEM	N	Rel	SEM	N	Rel	SEM	N	Rel	SEM
Science	5	13,514	0.82	3.16	7,115	0.79	3.21	7	–	–	3,851	0.84	3.06	1,910	0.81	3.10	4	–	–
	8	10,657	0.83	3.39	5,424	0.77	3.45	48	0.81	3.44	3,148	0.85	3.29	1,456	0.82	3.35	13	0.75	3.53
	10 Life Science	5,980	0.85	3.60	3,010	0.81	3.64	20	0.85	3.60	2,186	0.87	3.53	1,026	0.84	3.54	12	0.90	3.54

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

Table 8.B.9 Reliabilities and SEMs for the CMA by Primary Disability

Content Area	CMA *	Autism			Deaf-Blindness			Deafness			Emotional Dist.			Hard of Hearing			MR/ID			Mult. Disab.		
		N	Rel	SEM	N	Rel	SEM	N	Rel	SEM	N	Rel	SEM	N	Rel	SEM	N	Rel	SEM	N	Rel	SEM
Science	5	2,057	0.84	3.14	2	–	–	90	0.83	3.20	518	0.82	3.16	209	0.80	3.22	390	0.72	3.29	48	0.78	3.30
	8	1,291	0.86	3.32	2	–	–	126	0.72	3.46	653	0.85	3.38	178	0.82	3.40	430	0.66	3.55	31	0.86	3.37
	10 Life Science	716	0.87	3.49	3	–	–	113	0.75	3.65	452	0.88	3.57	129	0.83	3.60	262	0.70	3.72	29	0.76	3.74

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

Table 8.B.10 Reliabilities and SEMs for the CMA by Primary Disability (continued)

Content Area	CMA *	Orthoped. Impair.			Other Health Impair.			Specific Lrn Disab.			Speech or Lang Impair.			Traumatic Brain Injury			Visual Impair.			Unknown		
		N	Rel	SEM	N	Rel	SEM	N	Rel	SEM	N	Rel	SEM	N	Rel	SEM	N	Rel	SEM	N	Rel	SEM
Science	5	166	0.82	3.19	2,884	0.82	3.12	16,992	0.81	3.16	2,359	0.79	3.20	71	0.81	3.18	66	0.85	3.11	892	0.80	3.18
	8	128	0.80	3.41	2,295	0.84	3.36	14,927	0.82	3.39	1,091	0.79	3.43	73	0.85	3.37	58	0.84	3.41	763	0.82	3.40
	10 Life Science	96	0.89	3.58	1,226	0.86	3.54	8,852	0.84	3.60	468	0.81	3.60	41	0.81	3.63	39	0.89	3.53	326	0.84	3.62

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

Table 8.B.11 Overall Subgroup Reliabilities

Content Area	CMA *	Gender		Econ. Dis.		Language Fluency			
		Male	Female	No	Yes	EO	I-FEP	EL	R-FEP
Science	5	0.82	0.80	0.83	0.81	0.82	0.85	0.79	0.84
	8	0.84	0.79	0.84	0.81	0.84	0.82	0.78	0.82
	10 Life Science	0.85	0.82	0.86	0.83	0.86	0.84	0.79	0.84

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

Table 8.B.12 Overall Subgroup Reliabilities—Primary Ethnicity

Content Area	CMA *	Primary Ethnicity						
		African American	American Indian	Asian	Filipino	Hispanic	Pacific Islander	White
Science	5	0.80	0.80	0.83	0.80	0.80	0.75	0.83
	8	0.80	0.81	0.85	0.84	0.81	0.79	0.85
	10 Life Science	0.82	0.84	0.85	0.84	0.83	0.86	0.87

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

Table 8.B.13 Overall Subgroup Reliabilities by Primary Ethnicity—Not Economically Disadvantaged

Content Area	CMA *	Ethnicity						
		African American	American Indian	Asian	Filipino	Hispanic	Pacific Islander	White
Science	5	0.83	0.82	0.84	0.82	0.82	0.69	0.83
	8	0.81	0.85	0.85	0.86	0.83	0.76	0.85
	10 Life Science	0.83	0.86	0.86	0.82	0.85	0.85	0.87

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

Table 8.B.14 Overall Subgroup Reliabilities by Primary Ethnicity—Economically Disadvantaged

Content Area	CMA*	Ethnicity						
		African American	American Indian	Asian	Filipino	Hispanic	Pacific Islander	White
Science	5	0.80	0.79	0.83	0.79	0.80	0.78	0.83
	8	0.79	0.79	0.85	0.81	0.80	0.79	0.85
	10 Life Science	0.82	0.82	0.82	0.84	0.83	0.86	0.86

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

Table 8.B.15 Overall Subgroup Reliabilities by Gender/Economic Status

Content Area	CMA *	Economically Disadvantaged		Not Economically Disadvantaged	
		Male	Female	Male	Female
Science	5	0.82	0.79	0.84	0.81
	8	0.83	0.77	0.85	0.82
	10 Life Science	0.85	0.81	0.87	0.84

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

Table 8.B.16 Overall Subgroup Reliabilities by Primary Disability

Content Area	CMA *	Autism	Ethnicity					MR/ID	Mult. Disab.
			Deaf–Blindness	Deafness	Emotional Dist.	Hard of Hearing			
Science	5	0.84	–	0.83	0.82	0.80	0.72	0.78	
	8	0.86	–	0.72	0.85	0.82	0.66	0.86	
	10 Life Science	0.87	–	0.75	0.88	0.83	0.70	0.76	

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

Table 8.B.17 Overall Subgroup Reliabilities by Primary Disability (continued)

Content Area	CMA *	Orthoped. Impair.	Other Health Impair.	Specific Lrn Disab.	Speech or Lang Impair.	Traumatic Brain Injury	Visual Impair.
Science	5	0.82	0.82	0.81	0.79	0.81	0.85
	8	0.80	0.84	0.82	0.79	0.85	0.84
	10 Life Science	0.89	0.86	0.84	0.81	0.81	0.89

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

Table 8.B.18 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	0.54	1.84	0.48	1.87	0.54	1.82	0.51	1.86	
2. Life Sciences	16	0.70	1.77	0.65	1.79	0.71	1.71	0.67	1.79	
3. Earth Sciences	16	0.63	1.81	0.61	1.83	0.64	1.78	0.61	1.83	
Grade 8 Science										
1. Motion	19	0.62	1.97	0.55	2.02	0.63	1.94	0.59	2.00	
2. Matter	23	0.69	2.25	0.61	2.28	0.70	2.23	0.64	2.27	
3. Earth Science	7	0.58	1.10	0.47	1.15	0.57	1.06	0.53	1.13	
4. Investigation and Experimentation	5	0.42	1.03	0.37	1.05	0.46	1.00	0.38	1.04	
Grade 10 Life Science										
1. Cell Biology and Genetics	22	0.60	2.21	0.55	2.22	0.61	2.19	0.57	2.22	
2. Evolution and Ecology	22	0.71	2.13	0.65	2.15	0.73	2.10	0.68	2.15	
3. Physiology	10	0.64	1.41	0.56	1.45	0.64	1.38	0.60	1.44	
4. Investigation and Experimentation	6	0.43	1.14	0.40	1.15	0.45	1.12	0.40	1.15	

Table 8.B.19 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	0.48	1.87	0.54	1.84	0.61	1.77	0.57	1.78	
2. Life Sciences	16	0.65	1.82	0.70	1.74	0.72	1.68	0.72	1.64	
3. Earth Sciences	16	0.59	1.84	0.64	1.81	0.66	1.74	0.67	1.71	

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 8 Science									
1. Motion	19	0.57	2.03	0.62	1.97	0.60	1.93	0.55	1.91
2. Matter	23	0.58	2.29	0.69	2.25	0.69	2.24	0.69	2.20
3. Earth Science	7	0.50	1.15	0.57	1.10	0.61	1.04	0.52	1.01
4. Investigation and Experimentation	5	0.36	1.05	0.41	1.03	0.43	1.01	0.41	0.99
Grade 10 Life Science									
1. Cell Biology and Genetics	22	0.49	2.24	0.61	2.20	0.56	2.20	0.57	2.16
2. Evolution and Ecology	22	0.62	2.18	0.72	2.12	0.70	2.11	0.71	2.05
3. Physiology	10	0.54	1.47	0.64	1.40	0.65	1.38	0.61	1.34
4. Investigation and Experimentation	6	0.33	1.17	0.44	1.14	0.44	1.12	0.43	1.09

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

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	0.49	1.87	0.56	1.82	0.51	1.86	0.44	1.86	0.51	1.86	0.43	1.90	0.55	1.80
2. Life Sciences	16	0.68	1.80	0.66	1.73	0.72	1.76	0.71	1.77	0.67	1.79	0.63	1.84	0.70	1.68
3. Earth Sciences	16	0.61	1.86	0.59	1.82	0.66	1.80	0.58	1.80	0.60	1.83	0.54	1.85	0.65	1.77
Science Grade 8															
1. Motion	19	0.55	2.03	0.56	1.95	0.64	1.96	0.60	2.00	0.58	2.00	0.56	2.03	0.65	1.91
2. Matter	23	0.63	2.29	0.67	2.26	0.72	2.22	0.69	2.24	0.64	2.27	0.60	2.30	0.71	2.21
3. Earth Science	7	0.53	1.17	0.51	1.10	0.57	1.10	0.52	1.10	0.53	1.13	0.55	1.13	0.58	1.05
4. Investigation and Experimentation	5	0.32	1.06	0.37	1.03	0.45	1.01	0.38	1.04	0.39	1.04	0.33	1.05	0.47	0.99
Grade 10 Life Science															
1. Cell Biology and Genetics	22	0.55	2.24	0.53	2.25	0.66	2.17	0.65	2.17	0.56	2.22	0.63	2.20	0.62	2.17
2. Evolution and Ecology	22	0.66	2.18	0.71	2.14	0.68	2.13	0.70	2.11	0.68	2.15	0.64	2.18	0.74	2.08
3. Physiology	10	0.57	1.46	0.62	1.42	0.54	1.43	0.53	1.39	0.60	1.43	0.70	1.37	0.67	1.35
4. Investigation and Experimentation	6	0.35	1.17	0.42	1.15	0.41	1.13	0.21	1.16	0.40	1.15	0.30	1.17	0.47	1.11

Table 8.B.21 Subscore Reliabilities and SEM for Science by Primary Ethnicity-for-Not Economically Disadvantaged

Subscore Reliabilities and SEM for Science by Primary Ethnicity-for-Not Economically Disadvantaged															
Subscore Area	N of 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	0.52	1.86	0.55	1.76	0.51	1.85	0.43	1.87	0.51	1.84	0.40	1.88	0.57	1.78
2. Life Sciences	16	0.71	1.77	0.67	1.63	0.72	1.75	0.73	1.76	0.69	1.74	0.56	1.85	0.71	1.65
3. Earth Sciences	16	0.64	1.84	0.65	1.73	0.69	1.77	0.60	1.79	0.63	1.79	0.27	1.88	0.64	1.75
Grade 8 Science															
1. Motion	19	0.57	2.01	0.65	1.94	0.66	1.95	0.63	1.99	0.59	1.97	0.43	2.04	0.65	1.89
2. Matter	23	0.67	2.27	0.75	2.19	0.72	2.21	0.71	2.23	0.68	2.24	0.68	2.26	0.71	2.20
3. Earth Science	7	0.54	1.15	0.53	1.03	0.51	1.11	0.46	1.13	0.55	1.08	0.32	1.10	0.58	1.01
4. Investigation and Experimentation	5	0.38	1.04	0.27	1.02	0.42	1.01	0.45	1.02	0.46	1.01	-0.01	1.08	0.48	0.98

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 10 Life Science															
1. Cell Biology and Genetics	22	0.54	2.24	0.54	2.25	0.70	2.12	0.62	2.17	0.57	2.21	0.67	2.18	0.63	2.16
2. Evolution and Ecology	22	0.65	2.18	0.73	2.13	0.71	2.07	0.62	2.10	0.71	2.12	0.55	2.20	0.76	2.06
3. Physiology	10	0.58	1.45	0.68	1.38	0.58	1.37	0.45	1.38	0.62	1.40	0.66	1.41	0.68	1.33
4. Investigation and Experimentation	6	0.40	1.16	0.35	1.17	0.43	1.10	-0.01	1.19	0.44	1.13	0.12	1.21	0.48	1.10

Table 8.B.22 Subscore Reliabilities and SEM for Science by Primary Ethnicity-for-Economically Disadvantaged

Subscore Reliabilities and SEM for Science by Primary Ethnicity-for-Economically Disadvantaged															
Subscore Area	N of Items	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	0.48	1.87	0.54	1.84	0.52	1.86	0.45	1.84	0.50	1.87	0.44	1.91	0.54	1.81
2. Life Sciences	16	0.67	1.81	0.66	1.75	0.72	1.77	0.67	1.77	0.66	1.80	0.66	1.82	0.69	1.71
3. Earth Sciences	16	0.60	1.86	0.58	1.83	0.64	1.82	0.54	1.81	0.60	1.84	0.61	1.84	0.66	1.78
Grade 8 Science															
1. Motion	19	0.54	2.04	0.53	1.95	0.63	1.97	0.55	2.00	0.58	2.01	0.58	2.05	0.65	1.93
2. Matter	23	0.61	2.29	0.64	2.28	0.70	2.23	0.68	2.25	0.63	2.28	0.52	2.32	0.70	2.23
3. Earth Science	7	0.51	1.18	0.48	1.14	0.59	1.09	0.60	1.07	0.53	1.13	0.57	1.14	0.57	1.08
4. Investigation and Experimentation	5	0.30	1.06	0.40	1.03	0.47	1.01	0.23	1.07	0.38	1.05	0.34	1.06	0.45	1.01
Grade 10 Life Science															
1. Cell Biology and Genetics	22	0.55	2.24	0.53	2.25	0.60	2.21	0.64	2.19	0.55	2.22	0.61	2.22	0.61	2.19
2. Evolution and Ecology	22	0.66	2.17	0.70	2.15	0.65	2.17	0.72	2.13	0.67	2.16	0.67	2.18	0.72	2.11
3. Physiology	10	0.57	1.47	0.59	1.45	0.48	1.47	0.62	1.39	0.59	1.44	0.70	1.37	0.67	1.37
4. Investigation and Experimentation	6	0.33	1.17	0.44	1.15	0.39	1.15	0.41	1.11	0.40	1.15	0.42	1.15	0.45	1.13

Table 8.B.23 Subscore Reliabilities and SEM for Science by Disability

Subscore Reliabilities and SEM for Science by Disability															
Subscore Area	N of Items	Autism		Deafness		Emotional Disturbance		Hard of Hearing		MR/ID		Multiple Disability			
		Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM		
Grade 5															
1. Physical Science	16	0.54	1.85	0.61	1.84	0.53	1.84	0.53	1.87	0.40	1.90	0.46	1.90		
2. Life Science	16	0.74	1.77	0.59	1.87	0.70	1.77	0.68	1.82	0.57	1.89	0.65	1.88		
3. Earth Science	16	0.66	1.80	0.69	1.81	0.65	1.83	0.54	1.86	0.49	1.90	0.52	1.91		
Grade 8															
1. Motion	19	0.68	1.96	0.46	2.05	0.65	1.98	0.58	2.01	0.45	2.11	0.75	1.96		
2. Matter	23	0.72	2.21	0.48	2.27	0.70	2.25	0.70	2.25	0.37	2.31	0.65	2.28		
3. Earth Science	7	0.60	1.08	0.36	1.20	0.60	1.12	0.42	1.16	0.45	1.24	0.52	1.10		
4. Investigation and Experimentation	5	0.47	1.01	0.41	1.03	0.43	1.03	0.38	1.04	-0.01	1.09	0.39	1.04		
Life Science															
1. Cell Biology	22	0.68	2.17	0.42	2.23	0.64	2.20	0.58	2.21	0.45	2.25	0.43	2.29		
2. Evolution	22	0.73	2.07	0.64	2.15	0.75	2.13	0.61	2.17	0.45	2.24	0.64	2.25		
3. Physiology	10	0.65	1.37	0.28	1.52	0.68	1.39	0.61	1.42	0.39	1.51	0.57	1.46		
4. Investigation and Experimentation	6	0.48	1.11	0.23	1.17	0.49	1.13	0.45	1.12	0.16	1.19	0.03	1.20		

Table 8.B.24 Subscore Reliabilities and SEM for Science by Disability (continued)

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	0.41	1.89	0.53	1.83	0.52	1.85	0.47	1.88	0.50	1.87	0.57	1.85
2. Life Science	16	0.72	1.76	0.70	1.74	0.68	1.77	0.65	1.81	0.67	1.78	0.71	1.73
3. Earth Science	16	0.63	1.84	0.64	1.80	0.62	1.82	0.59	1.83	0.69	1.80	0.67	1.82
Grade 8													
1. Motion	19	0.57	2.01	0.63	1.96	0.59	1.99	0.54	2.02	0.65	2.00	0.69	2.00
2. Matter	23	0.66	2.24	0.70	2.25	0.66	2.27	0.63	2.27	0.73	2.22	0.61	2.29
3. Earth Science	7	0.39	1.16	0.59	1.09	0.54	1.12	0.51	1.14	0.55	1.11	0.58	1.12
4. Investigation and Experimentation	5	0.39	1.04	0.43	1.02	0.40	1.04	0.36	1.05	0.41	1.03	0.53	1.01
Life Science													
1. Cell Biology	22	0.68	2.21	0.61	2.19	0.56	2.21	0.52	2.22	0.35	2.26	0.73	2.12
2. Evolution	22	0.78	2.12	0.72	2.12	0.68	2.14	0.64	2.14	0.68	2.14	0.74	2.15
3. Physiology	10	0.72	1.37	0.65	1.37	0.60	1.43	0.56	1.44	0.54	1.49	0.77	1.34
4. Investigation and Experimentation	6	0.39	1.16	0.46	1.13	0.41	1.15	0.37	1.15	0.36	1.15	0.60	1.09

Table 8.B.25 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	0.14	0.06	0.00	0.00	0.20
	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	0.31
	37–48	0.00	0.00	0.00	0.05	0.10	0.15
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	0.11	0.06	0.01	0.00	0.20
	23–29	0.00	0.06	0.16	0.08	0.00	0.31
	30–36	0.00	0.01	0.08	0.17	0.06	0.31
	37–48	0.00	0.00	0.00	0.05	0.10	0.15
Estimated Proportion Consistently Classified: Total = 0.55, Proficient & Above = 0.82							

Table 8.B.26 Reliability of Classification for Science, Grade Eight

	Placement Score	Far Below Basic	Below Basic	Basic	Proficient	Advanced	Category Total
Decision Accuracy	0–21	0.06	0.05	0.01	0.00	0.00	0.12
	22–26	0.02	0.10	0.06	0.00	0.00	0.18
	27–33	0.00	0.05	0.20	0.05	0.00	0.31
	34–40	0.00	0.00	0.06	0.16	0.02	0.24
	41–54	0.00	0.00	0.00	0.04	0.10	0.15
Estimated Proportion Correctly Classified: Total = 0.61, Proficient & Above = 0.88							
Decision Consistency	0–21	0.06	0.04	0.02	0.00	0.00	0.12
	22–26	0.04	0.07	0.06	0.01	0.00	0.18
	27–33	0.02	0.06	0.15	0.07	0.01	0.31
	34–40	0.00	0.01	0.07	0.12	0.04	0.24
	41–54	0.00	0.00	0.01	0.04	0.10	0.15
Estimated Proportion Consistently Classified: Total = 0.50, Proficient & Above = 0.83							

Table 8.B.27 Reliability of Classification for Life Science (Grade 10)

	Placement Score	Far Below	Below Basic	Basic	Proficient	Advanced	Category Total
Decision Accuracy	0–23	0.10	0.06	0.00	0.00	0.00	0.16
	24–31	0.04	0.20	0.05	0.00	0.00	0.29
All-forms Average	32–39	0.00	0.06	0.19	0.04	0.00	0.29
	40–47	0.00	0.00	0.05	0.13	0.01	0.20
	48–60	0.00	0.00	0.00	0.03	0.04	0.06
Estimated Proportion Correctly Classified: Total = 0.65, Proficient & Above = 0.91							
Decision Consistency	0–23	0.09	0.06	0.01	0.00	0.00	0.16
	24–31	0.06	0.15	0.07	0.00	0.00	0.29
Alternate Form	32–39	0.01	0.07	0.15	0.06	0.00	0.29
	40–47	0.00	0.01	0.06	0.11	0.03	0.20
	48–60	0.00	0.00	0.00	0.03	0.04	0.06
Estimated Proportion Consistently Classified: Total = 0.54, Proficient & Above = 0.87							

Appendix 8.C—IRT Analyses

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

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	1,290	0.0712	337.5687	338
1	0	-4.3082	150.0000	150	29	1,282	0.1636	344.4507	344
2	1	-3.5850	150.0000	150	30	1,276	0.2577	351.4485	351
3	1	-3.1490	150.0000	150	31	1,282	0.3536	358.5914	359
4	1	-2.8304	150.0000	150	32	1,221	0.4520	365.9153	366
5	0	-2.5760	150.0000	150	33	1,235	0.5534	373.4561	373
6	2	-2.3620	156.4763	156	34	1,253	0.6582	381.2595	381
7	1	-2.1757	170.3378	170	35	1,093	0.7673	389.3793	389
8	12	-2.0097	182.6978	183	36	981	0.8815	397.8804	398
9	14	-1.8589	193.9220	194	37	928	1.0020	406.8435	407
10	37	-1.7200	204.2596	204	38	768	1.1300	416.3708	416
11	61	-1.5906	213.8913	214	39	680	1.2674	426.5976	427
12	111	-1.4688	222.9504	223	40	547	1.4166	437.7017	438
13	186	-1.3534	231.5396	232	41	428	1.5809	449.9342	450
14	250	-1.2433	239.7396	240	42	326	1.7653	463.6594	464
15	339	-1.1374	247.6155	248	43	222	1.9774	479.4452	479
16	418	-1.0353	255.2207	255	44	120	2.2298	498.2319	498
17	575	-0.9361	262.6009	263	45	76	2.5463	521.7867	522
18	666	-0.8395	269.7933	270	46	24	2.9801	554.0742	554
19	766	-0.7449	276.8329	277	47	9	3.7008	600.0000	600
20	814	-0.6519	283.7488	284	48	2	N/A	600.0000	600
21	814	-0.5603	290.5682	291					
22	963	-0.4697	297.3158	297					
23	1,030	-0.3796	304.0151	304					
24	1,162	-0.2900	310.6888	311					
25	1,089	-0.2004	317.3587	317					
26	1,180	-0.1105	324.0468	324					
27	1,208	-0.0201	330.7741	331					

Note: Performance-level cut scores are highlighted.

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

Grade Eight									
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	1,035	-0.1656	307.1429	307
1	0	-4.3764	150.0000	150	29	1,016	-0.0868	314.2075	314
2	0	-3.6553	150.0000	150	30	969	-0.0075	321.3040	321
3	0	-3.2217	150.0000	150	31	1,000	0.0723	328.4528	328
4	0	-2.9059	150.0000	150	32	970	0.1530	335.6734	336
5	0	-2.6546	150.0000	150	33	928	0.2347	342.9878	343
6	0	-2.4440	150.0000	150	34	859	0.3177	350.4194	350
7	1	-2.2614	150.0000	150	35	887	0.4022	357.9933	358
8	0	-2.0992	150.0000	150	36	819	0.4888	365.7421	366
9	0	-1.9526	150.0000	150	37	768	0.5776	373.6939	374
10	10	-1.8181	159.1781	159	38	706	0.6691	381.8879	382
11	15	-1.6934	170.3484	170	39	695	0.7638	390.3675	390
12	26	-1.5766	180.8078	181	40	633	0.8623	399.1846	399
13	41	-1.4663	190.6790	191	41	559	0.9652	408.4012	408
14	67	-1.3616	200.0581	200	42	500	1.0734	418.0936	418
15	126	-1.2615	209.0224	209	43	441	1.1881	428.3556	428
16	178	-1.1653	217.6352	218	44	387	1.3104	439.3137	439
17	252	-1.0724	225.9479	226	45	343	1.4423	451.1215	451
18	340	-0.9824	234.0058	234	46	270	1.5861	463.9953	464
19	423	-0.8949	241.8466	242	47	209	1.7451	478.2338	478
20	511	-0.8094	249.5037	250	48	173	1.9242	494.2709	494
21	620	-0.7256	257.0052	257	49	129	2.1312	512.8023	513
22	711	-0.6432	264.3783	264	50	91	2.3784	534.9410	535
23	752	-0.5621	271.6466	272	51	42	2.6897	562.8140	563
24	816	-0.4818	278.8320	279	52	32	3.1184	600.0000	600
25	888	-0.4023	285.9550	286	53	17	3.8341	600.0000	600
26	891	-0.3232	293.0352	293	54	4	N/A	600.0000	600
27	896	-0.2444	300.0917	300					

Note: Performance-level cut scores are highlighted.

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

Grade Ten									
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	31	476	0.0060	294.1228	294
1	0	-4.2357	150.0000	150	32	444	0.0763	300.1578	300
2	0	-3.5228	150.0000	150	33	464	0.1470	306.2201	306
3	0	-3.0971	150.0000	150	34	467	0.2182	312.3231	312
4	0	-2.7888	150.0000	150	35	483	0.2900	318.4806	318
5	1	-2.5446	150.0000	150	36	446	0.3626	324.7072	325
6	0	-2.3408	150.0000	150	37	496	0.4362	331.0182	331
7	0	-2.1649	150.0000	150	38	440	0.5110	337.4301	337
8	0	-2.0092	150.0000	150	39	410	0.5872	343.9650	344
9	0	-1.8688	150.0000	150	40	389	0.6650	350.6385	351
10	2	-1.7405	150.0000	150	41	395	0.7448	357.4753	357
11	6	-1.6218	154.5435	155	42	357	0.8267	364.5012	365
12	3	-1.5110	164.0473	164	43	341	0.9112	371.7460	372
13	16	-1.4067	172.9923	173	44	304	0.9986	379.2440	379
14	19	-1.3078	181.4687	181	45	291	1.0895	387.0360	387
15	48	-1.2136	189.5483	190	46	242	1.1844	395.1692	395
16	85	-1.1233	197.2899	197	47	206	1.2839	403.7056	404
17	132	-1.0364	204.7412	205	48	193	1.3890	412.7138	413
18	193	-0.9524	211.9432	212	49	151	1.5006	422.2853	422
19	227	-0.8709	218.9300	219	50	138	1.6201	432.5357	433
20	291	-0.7916	225.7314	226	51	89	1.7494	443.6164	444
21	348	-0.7142	232.3731	232	52	83	1.8906	455.7308	456
22	350	-0.6383	238.8779	239	53	62	2.0474	469.1727	469
23	365	-0.5638	245.2666	245	54	39	2.2244	484.3522	484
24	438	-0.4904	251.5563	252	55	25	2.4293	501.9209	502
25	452	-0.4180	257.7648	258	56	15	2.6747	522.9617	523
26	464	-0.3464	263.9076	264	57	18	2.9844	549.5157	550
27	446	-0.2754	269.9992	270	58	2	3.4114	586.1369	586
28	468	-0.2048	276.0534	276	59	3	4.1258	600.0000	600
29	468	-0.1344	282.0834	282	60	1	N/A	600.0000	600
30	459	-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, 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, 2012) 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 CMA prior to the 2014 administration is described in detail in Chapter 3, starting on page 25. 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 CMA administration.

Item Specifications

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

Item Writers

The items for each CMA 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 ARPs were committees that advised the CDE and ETS on areas related to item development for the CMA. 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 28 in Chapter 3 for additional information on the function of ARPs within the item-review process.

Statewide Pupil Assessment Review Panel Review

The SPAR panel was responsible for reviewing and approving the achievement tests that were used statewide for the testing of students in California public schools in grades five, eight, and ten. The SPAR panel representatives ensured that the test items conformed to the requirements of *EC* Section 60602. If the SPAR panel rejected specific items, the items were replaced with other items. See page 25 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 administered in 2014 were statistically reviewed in data review meetings the year they were originally administered. There was no data review of field-test items in 2014. See Table 8.4 on page 87 for the list of the original administrations of each test administered in 2014.

Quality Control of the Item Bank

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

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

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

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

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

Quality Control of Test Form Development

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

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

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

Quality Control of Test Materials

Collecting Test Materials

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

All scorable materials are delivered to the Pearson scanning and scoring facilities in Iowa City, Iowa. The nonscorable materials, including test booklets, are returned to the Security Processing Department in Pearson's Cedar Rapids, Iowa, facility. ETS and Pearson closely monitor the return of materials. The California Technical Assistance Center (CalTAC) at ETS monitors returns and notifies LEAs that do not return their materials in a timely manner. CalTAC contacts the LEA CAASPP coordinators and works with them to facilitate the return of the test materials.

Processing Test Materials

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

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

Quality Control of Scanning

Before any CAASPP documents are scanned, Pearson conducts a complete check of the scanning system. ETS and Pearson create test decks for every test and form. Each test deck consists of approximately 25 answer documents marked to cover response ranges, demographic data, blanks, double marks, and other responses. Fictitious students are created to verify that each marking possibility is processed correctly by the scanning program. The output file generated as a result of this activity is thoroughly checked against

each answer document after 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, Pearson creates a mock set of documents to test all of the errors listed in the edit specifications. The set of test documents is used to verify that each image of the document is saved so that an editor 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.

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

Quality Control of Answer Document Processing and Scoring

Accountability of Answer Documents

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

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

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

Processing of Answer Documents

Prior to processing operational answer documents and executing subsequent data processing programs, ETS conducts an end-to-end test. As part of this test, ETS prepares

approximately 700 test cases covering all tests and many scenarios designed to exercise particular business rule logic. ETS marks answer documents for those 700 test cases. They are then scanned, scored, and aggregated. The results at various inspection points are checked by psychometricians and Data Quality Services staff. Additionally, a post-scan test file of approximately 50,000 records across the CAASPP System is scored and aggregated to test a broader range of scoring and aggregation scenarios. These procedures assure that students and LEAs receive the correct scores when the actual scoring process is carried out. In 2014, end-to-end testing also included the inspection of results in electronic reporting.

Scoring and Reporting Specifications

ETS develops standardized scoring procedures and specifications so testing materials are processed and scored accurately. These documents include:

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

Each of these documents is explained in detail in Chapter 7, starting on page 59. The scoring specifications are reviewed and revised by the CDE, ETS, and Pearson each year. After a version that all parties endorse is finalized, the CDE issues a formal approval of the scoring and reporting specifications.

Storing Answer Documents

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

Quality Control of Psychometric Processes

Score Key Verification Procedures

ETS and Pearson take various necessary measures to ascertain that the scoring keys are applied to the student responses as expected and the student scores are computed accurately. Scoring keys, provided in the form planners, are produced by ETS and verified thoroughly by performing multiple quality control checks. The form planners contain the information about an assembled test form; other information in the form planner includes the test name, administration year, subscore identification, and standards and statistics associated with each item. The quality control checks that are performed before keys are finalized are listed on page 61 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 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 14, checks were made to ascertain that the correct options for the analyses were selected. Checks were also made on the number of items, number of examinees with valid scores, IRT Rasch item difficulty estimates, standard errors for the Rasch item difficulty estimates, and the match of selected statistics to the results on the same statistics obtained during preliminary item analyses. Psychometricians also performed detailed reviews of plots and statistics to investigate if the model fit the data.

Scaling

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

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

Scoring Tables

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

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

As a check on the reasonableness of the performance levels, when the tests were originally administered, psychometricians compared results from the current year with results from the

past year at the cut points and the percentage of students in each performance level within the equating samples. After all quality control steps were completed and any differences were resolved, a senior psychometrician inspected the scoring tables as the final step in quality control before ETS delivered them to Pearson.

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

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, every regular and special-version multiple-choice test was “certified” by ETS prior to being included in electronic reporting. To certify a test, psychometricians gathered a certain number of test cases and verified the accurate application of scoring keys and conversion tables.

Score Verification Process

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

Year-to-Year Comparison Analyses

Year-to-year comparison analyses are conducted each year for quality control of the scoring procedure in general and as reasonableness checks for the CMA 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 forms that are reused in 2014, 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 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, ETS, and Pearson prior to any LEA mailings

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

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

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

Electronic Reporting

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

Before an LEA could download a student data file, ETS statisticians approved a QC file of test results data and ETS IT successfully processed the QC file. Once the data were deemed reliable and Pearson has processed a scorable answer document for every student who took a CMA 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 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—2012, 2013, and 2014—and the base year.

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

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

Table 10.1 Base Years for the CMA

Content Area	CMA	Base Year
	5	2009
Science	8	2010
	10 Life Science	2011

The base years differ over CMA grades and content areas. 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 120 contains the number of examinees assessed and the means and standard deviations of examinees' scale scores in the base year and in 2012, 2013, and

2014 for each CMA. As noted in previous chapters, the CMA reporting scales range from 150 to 600 for all of the tests.

CMA scale scores are used to classify student results into one of five performance levels: far below basic, below basic, basic, proficient, and advanced. The percentages of students qualifying for the proficient and advanced levels are presented in Table 10.A.2 on page 120; please note that this information may differ slightly from information found on the CDE's CAASPP reporting Web page at <http://caaspp.cde.ca.gov> due to differing dates on which data were accessed. The goal is for all students to achieve at or above the proficient level by 2014.

Table 10.A.3 shows for each CMA the distribution of scale scores observed in the base year, which differs according to test, and subsequent administrations in 2012, 2013, and 2014 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, 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 over the comparison years indicate that the CMA meet the technical criteria established in professional standards for high-stakes tests.

Table 10.B.1 in Appendix 10.B, which starts on page 121, presents the average proportion correct values for the operational items in each CMA. 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 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 standard error of measurement (SEM) expressed in raw score units appear in Table 10.B.4. Like the average proportion correct, point-biserial correlations and reliabilities of the operational items are affected by both item characteristics and student characteristics.