

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



**California Alternate Performance
Assessment
Technical Report
Spring 2012 Administration**

**Final Draft Submitted March 15, 2013
Educational Testing Service
Contract No. 5417**

Table of Contents

Acronyms and Initialisms Used in the <i>CAPA Technical Report</i>	vii
Chapter 1: Introduction	1
Background	1
Test Purpose	1
Test Content	1
Intended Population	2
Intended Use and Purpose of Test Scores	3
Testing Window	3
Significant STAR Developments in 2012	4
Release of Results Delayed to Investigate Psychometric Impact of Social Network Security Breaches.....	4
Extension of the STAR Testing Window	4
Limitations of the Assessment	4
Score Interpretation	4
Out-of-Level Testing.....	5
Score Comparison	5
Groups and Organizations Groups Involved in Test Development	5
State Board of Education	5
California Department of Education	5
Contractors	5
Overview of the Technical Report	6
References	8
Chapter 2: An Overview of CAPA Processes	9
Task (Item) Development	9
Task Formats	9
Task (Item) Specifications.....	9
Item Banking.....	9
Task Refresh Rate	10
Test Assembly	10
Test Length	10
Test Blueprints	10
Content Rules and Task Selection.....	10
Psychometric Criteria.....	11
Test Administration	11
Test Security and Confidentiality.....	11
Procedures to Maintain Standardization	12
Test Variations, Accommodations, and Modifications	12
Scores	13
Aggregation Procedures	13
Equating	14
Calibration.....	14
Scaling.....	14
Linear Transformation.....	15
References	16
Chapter 3: Task (Item) Development	17
Rules for Task Development	17
Task Specifications	17
Expected Task Ratio.....	18
Selection of Task Writers	19
Criteria for Selecting Task Writers	19
Task (Item) Review Process	19
Contractor Review	19
Content Expert Reviews.....	21
Statewide Pupil Assessment Review Panel.....	23
Field Testing	23
Stand-alone Field Testing	23
Embedded Field-test Tasks	23
CDE Data Review	25
Item Banking	25
References	26

Chapter 4: Test Assembly	27
Test Length	27
Rules for Task Selection	27
Test Blueprints	27
Content Rules and Task Selection	27
Psychometric Criteria	28
Projected Psychometric Properties of the Assembled Tests	29
Rules for Task Sequence and Layout	29
Chapter 5: Test Administration	30
Test Security and Confidentiality	30
ETS's Office of Testing Integrity	30
Test Development	30
Task and Data Review	30
Item Banking	31
Transfer of Forms and Tasks to the CDE	31
Security of Electronic Files Using a Firewall	31
Printing and Publishing	32
Test Administration	32
Test Delivery	32
Processing and Scoring	33
Data Management	33
Transfer of Scores via Secure Data Exchange	34
Statistical Analysis	34
Reporting and Posting Results	34
Student Confidentiality	34
Student Test Results	34
Procedures to Maintain Standardization	35
Test Administrators	35
CAPA Examiner's Manual	36
District and Test Site Coordinator Manual	36
STAR Management System Manuals	37
Accommodations for Students with Disabilities	37
Identification	37
Adaptations	37
Scoring	38
Demographic Data Corrections	38
Testing Irregularities	38
Test Administration Incidents	38
References	39
Chapter 6: Performance Standards	40
Background	40
Standard Setting Procedure	40
Development of Competencies Lists	41
Standard Setting Methodology	42
Performance Profile Method	42
Results	42
References	44
Chapter 7: Scoring and Reporting	45
Procedures for Maintaining and Retrieving Individual Scores	45
Scoring and Reporting Specifications	46
Scanning and Scoring	46
Types of Scores	47
Raw Score	47
Scale Score	47
Performance Levels	47
Score Verification Procedures	47
Monitoring and Quality Control of Scoring	47
Score Verification Process	48
Overview of Score Aggregation Procedures	48
Individual Scores	48
Reports Produced and Scores for Each Report	53
Types of Score Reports	53
Score Report Contents	53

Score Report Applications.....	53
Criteria for Interpreting Test Scores	54
Criteria for Interpreting Score Reports	54
References	56
Appendix 7.A—Scale Score Distribution Tables	57
Appendix 7.B—Demographic Summaries	59
Appendix 7.C—Type of Score Report	65
Chapter 8: Analyses	68
Samples Used for the Analyses	68
Classical Analyses	69
Average Item Score.....	69
Polyserial Correlation of the Task Score with the Total Test Score.....	69
Reliability Analyses	71
Subgroup Reliabilities and SEMs.....	72
Conditional Standard Errors of Measurement.....	72
Decision Classification Analyses	73
Validity Evidence	74
Purpose of the CAPA.....	75
The Constructs to Be Measured.....	75
Interpretations and Uses of the Scores Generated.....	75
Intended Test Population(s).....	76
Validity Evidence Collected.....	76
Evidence Based on Response Processes.....	79
Evidence of Interrater Agreement.....	79
Evidence Based on Internal Structure.....	80
Evidence Based on Consequences of Testing.....	80
IRT Analyses	80
IRT Model-Data Fit Analyses.....	81
Model-fit Assessment Results.....	82
Evaluation of Scaling.....	82
Summaries of Scaled IRT b-values.....	83
Post-scaling Results.....	83
Differential Item Functioning Analyses	83
References	86
Appendix 8.A—Classical Analyses: Task Statistics	88
Appendix 8.B—Reliability Analyses	105
Appendix 8.C—Validity Analyses	119
Appendix 8.D—IRT Analyses	136
Appendix 8.E—DIF Analyses	152
Chapter 9: Quality Control Procedures	157
Quality Control of Task Development	157
Task Specifications.....	157
Task Writers.....	157
Internal Contractor Reviews.....	157
Assessment Review Panel Review.....	158
Statewide Pupil Assessment Review Panel Review.....	158
Data Review of Field-tested Tasks.....	158
Quality Control of the Item Bank	159
Quality Control of Test Form Development	159
Quality Control of Test Materials	160
Collecting Test Materials.....	160
Processing Test Materials.....	160
Quality Control of Scanning	160
Post-scanning Edits.....	161
Quality Control of Image Editing	161
Quality Control of Answer Document Processing and Scoring	161
Accountability of Answer Documents.....	161
Processing of Answer Documents.....	162
Scoring and Reporting Specifications.....	162
Matching Information on CAPA Answer Documents.....	162
Storing Answer Documents.....	162
Quality Control of Psychometric Processes	163
Quality Control of Task (Item) Analyses, DIF, and the Scoring Process.....	163

Score Verification Process.....	164
Year-to-Year Comparison Analyses.....	164
Offloads to Test Development.....	164
Quality Control of Reporting	164
Excluding Student Scores from Summary Reports.....	165
Reference.....	166
Chapter 10: Historical Comparisons.....	167
Base Year Comparisons	167
Examinee Performance.....	167
Test Characteristics	168
Appendix 10.A—Historical Comparisons Tables.....	169
Appendix 10.B—Historical Comparisons Tables.....	173

Tables

Table 1.1 Description of the CAPA Assessment Levels.....	2
Table 2.1 CAPA Item and Estimated Time Chart.....	10
Table 2.2 Scale Scores Ranges for Performance Levels.....	15
Table 3.1 Field-test Percentages for the CAPA	18
Table 3.2 CAPA ARP Member Qualifications, by Content Area and Total.....	22
Table 3.3 Summary of Tasks and Forms Presented in the 2012 CAPA.....	24
Table 4.1 Statistical Targets for CAPA Test Assembly	28
Table 4.2 Summary of 2012 CAPA Projected Statistical Attributes.....	29
Table 7.1 Rubrics for CAPA Scoring.....	45
Table 7.2 Summary Statistics Describing Student Scores: ELA	48
Table 7.3 Summary Statistics Describing Student Scores: Mathematics.....	49
Table 7.4 Summary Statistics Describing Student Scores: Science	50
Table 7.5 Percentage of Examinees in Each Performance Level	51
Table 7.6 Subgroup Definitions.....	52
Table 7.7 Types of CAPA Reports.....	53
Table 7.A.1 Scale Score Frequency Distributions: ELA, Levels I–V	57
Table 7.A.2 Scale Score Frequency Distributions: Mathematics, Levels I–V.....	57
Table 7.A.3 Scale Score Frequency Distributions: Science, Levels I–V.....	58
Table 7.B.1 Demographic Summary for ELA, All Examinees.....	59
Table 7.B.2 Demographic Summary for Mathematics, All Examinees.....	61
Table 7.B.3 Demographic Summary for Science, All Examinees.....	63
Table 7.C.1 Score Reports Reflecting CAPA Results.....	65
Table 8.1 CAPA Raw Score Means and Standard Deviations: Total P1 Population and Equating Sample.....	69
Table 8.2 Average Item Score and Polyserial Correlation.....	70
Table 8.3 Reliabilities and SEMs for the CAPA.....	72
Table 8.4 CAPA Content-area Correlations for CAPA Levels.....	79
Table 8.5 Evaluation of Common Items Between New and Reference Test Forms.....	82
Table 8.6 DIF Flags Based on the ETS DIF Classification Scheme.....	85
Table 8.7 Subgroup Classification for DIF Analyses	85
Table 8.A.1 AIS and Polyserial Correlation: Level I, ELA.....	88
Table 8.A.2 AIS and Polyserial Correlation: Level II, ELA.....	89
Table 8.A.3 AIS and Polyserial Correlation: Level III, ELA.....	90
Table 8.A.4 AIS and Polyserial Correlation: Level IV, ELA	91
Table 8.A.5 AIS and Polyserial Correlation: Level V, ELA	92
Table 8.A.6 AIS and Polyserial Correlation: Level I, Mathematics	93
Table 8.A.7 AIS and Polyserial Correlation: Level II, Mathematics	94
Table 8.A.8 AIS and Polyserial Correlation: Level III, Mathematics	95
Table 8.A.9 AIS and Polyserial Correlation: Level IV, Mathematics.....	96
Table 8.A.10 AIS and Polyserial Correlation: Level V, Mathematics.....	97
Table 8.A.11 AIS and Polyserial Correlation: Level I, Science.....	98
Table 8.A.12 AIS and Polyserial Correlation: Level III, Science.....	99
Table 8.A.13 AIS and Polyserial Correlation: Level IV, Science	100
Table 8.A.14 AIS and Polyserial Correlation: Level V, Science	101
Table 8.A.15 Frequency of Operational Task Scores: ELA.....	102
Table 8.A.16 Frequency of Operational Task Scores: Mathematics	103
Table 8.A.17 Frequency of Operational Task Scores: Science.....	104
Table 8.B.1 Reliabilities and SEMs by GENDER.....	105
Table 8.B.2 Reliabilities and SEMs by PRIMARY ETHNICITY	106
Table 8.B.3 Reliabilities and SEMs by PRIMARY ETHNICITY for Economically Disadvantaged	107
Table 8.B.4 Reliabilities and SEMs by PRIMARY ETHNICITY for Not Economically Disadvantaged.....	108

Table 8.B.5 Reliabilities and SEMs by PRIMARY ETHNICITY for Unknown Economic Status	109
Table 8.B.6 Reliabilities and SEMs by Disability	110
Table 8.B.7 Decision Accuracy and Decision Consistency: Level I, ELA	111
Table 8.B.8 Decision Accuracy and Decision Consistency: Level I, Mathematics	112
Table 8.B.9 Decision Accuracy and Decision Consistency: Level I, Science	112
Table 8.B.10 Decision Accuracy and Decision Consistency: Level II, ELA	113
Table 8.B.11 Decision Accuracy and Decision Consistency: Level II, Mathematics	113
Table 8.B.12 Decision Accuracy and Decision Consistency: Level III, ELA	114
Table 8.B.13 Decision Accuracy and Decision Consistency: Level III, Mathematics	114
Table 8.B.14 Decision Accuracy and Decision Consistency: Level III, Science	115
Table 8.B.15 Decision Accuracy and Decision Consistency: Level IV, ELA	115
Table 8.B.16 Decision Accuracy and Decision Consistency: Level IV, Mathematics	116
Table 8.B.17 Decision Accuracy and Decision Consistency: Level IV, Science	116
Table 8.B.18 Decision Accuracy and Decision Consistency: Level V, ELA	117
Table 8.B.19 Decision Accuracy and Decision Consistency: Level V, Mathematics	117
Table 8.B.20 Decision Accuracy and Decision Consistency: Level V, Science	118
Table 8.C.1 CAPA Content Area Correlations by Gender: Level I	119
Table 8.C.2 CAPA Content Area Correlations by Gender: Level II	119
Table 8.C.3 CAPA Content Area Correlations by Gender: Level III	119
Table 8.C.4 CAPA Content Area Correlations by Gender: Level IV	119
Table 8.C.5 CAPA Content Area Correlations by Gender: Level V	119
Table 8.C.6 CAPA Content Area Correlations by Ethnicity: Level I	120
Table 8.C.7 CAPA Content Area Correlations by Ethnicity: Level II	120
Table 8.C.8 CAPA Content Area Correlations by Ethnicity: Level III	120
Table 8.C.9 CAPA Content Area Correlations by Ethnicity: Level IV	120
Table 8.C.10 CAPA Content Area Correlations by Ethnicity: Level V	121
Table 8.C.11 CAPA Content Area Correlations by Ethnicity for Economically Disadvantaged: Level I	121
Table 8.C.12 CAPA Content Area Correlations by Ethnicity for Economically Disadvantaged: Level II	121
Table 8.C.13 CAPA Content Area Correlations by Ethnicity for Economically Disadvantaged: Level III	121
Table 8.C.14 CAPA Content Area Correlations by Ethnicity for Economically Disadvantaged: Level IV	122
Table 8.C.15 CAPA Content Area Correlations by Ethnicity for Economically Disadvantaged: Level V	122
Table 8.C.16 CAPA Content Area Correlations by Ethnicity for Not Economically Disadvantaged: Level I	122
Table 8.C.17 CAPA Content Area Correlations by Ethnicity for Not Economically Disadvantaged: Level II	122
Table 8.C.18 CAPA Content Area Correlations by Ethnicity for Not Economically Disadvantaged: Level III	123
Table 8.C.19 CAPA Content Area Correlations by Ethnicity for Not Economically Disadvantaged: Level IV	123
Table 8.C.20 CAPA Content Area Correlations by Ethnicity for Not Economically Disadvantaged: Level V	123
Table 8.C.21 CAPA Content Area Correlations by Ethnicity for Unknown Economic Status: Level I	123
Table 8.C.22 CAPA Content Area Correlations by Ethnicity for Unknown Economic Status: Level II	124
Table 8.C.23 CAPA Content Area Correlations by Ethnicity for Unknown Economic Status: Level III	124
Table 8.C.24 CAPA Content Area Correlations by Ethnicity for Unknown Economic Status: Level IV	124
Table 8.C.25 CAPA Content Area Correlations by Ethnicity for Unknown Economic Status: Level V	124
Table 8.C.26 CAPA Content Area Correlations by Economic Status: Level I	125
Table 8.C.27 CAPA Content Area Correlations by Economic Status: Level II	125
Table 8.C.28 CAPA Content Area Correlations by Economic Status: Level III	125
Table 8.C.29 CAPA Content Area Correlations by Economic Status: Level IV	125
Table 8.C.30 CAPA Content Area Correlations by Economic Status: Level V	125
Table 8.C.31 CAPA Content Area Correlations by Disability: Level I	126
Table 8.C.32 CAPA Content Area Correlations by Disability: Level II	127
Table 8.C.33 CAPA Content Area Correlations by Disability: Level III	128
Table 8.C.34 CAPA Content Area Correlations by Disability: Level IV	129
Table 8.C.35 CAPA Content Area Correlations by Disability: Level V	130
Table 8.C.36 Interrater Agreement Analyses for Operational Tasks: Level I	131
Table 8.C.37 Interrater Agreement Analyses for Operational Tasks: Level II	132
Table 8.C.38 Interrater Agreement Analyses for Operational Tasks: Level III	133
Table 8.C.39 Interrater Agreement Analyses for Operational Tasks: Level IV	134
Table 8.C.40 Interrater Agreement Analyses for Operational Tasks: Level V	135
Table 8.D.1 Item Classifications for Model-Data Fit Across All CAPA Levels	136
Table 8.D.2 Fit Classifications: Level I Tasks	136
Table 8.D.3 Fit Classifications: Level II Tasks	136
Table 8.D.4 Fit Classifications: Level III Tasks	136
Table 8.D.5 Fit Classifications: Level IV Tasks	136
Table 8.D.6 Fit Classifications: Level V Tasks	137
Table 8.D.7 IRT <i>b</i> -values for ELA, by Level	137
Table 8.D.8 IRT <i>b</i> -values for Mathematics, by Level	137
Table 8.D.9 IRT <i>b</i> -values for Science, by Level	137

Table 8.D.10	Score Conversions: Level I, ELA.....	138
Table 8.D.11	Score Conversions: Level II, ELA.....	139
Table 8.D.12	Score Conversions: Level III, ELA.....	140
Table 8.D.13	Score Conversions: Level IV, ELA.....	141
Table 8.D.14	Score Conversions: Level V, ELA.....	142
Table 8.D.15	Score Conversions: Level I, Mathematics.....	143
Table 8.D.16	Score Conversions: Level II, Mathematics.....	144
Table 8.D.17	Score Conversions: Level III, Mathematics.....	145
Table 8.D.18	Score Conversions: Level IV, Mathematics.....	146
Table 8.D.19	Score Conversions: Level V, Mathematics.....	147
Table 8.D.20	Score Conversions: Level I, Science.....	148
Table 8.D.21	Score Conversions: Level III, Science.....	149
Table 8.D.22	Score Conversions: Level IV, Science.....	150
Table 8.D.23	Score Conversions: Level V, Science.....	151
Table 8.E.1	Item Exhibiting Significant DIF by Ethnic Group.....	152
Table 8.E.2	Items Exhibiting Significant DIF by Disability Group.....	153
Table 8.E.3	CAPA Disability Distributions: Level I.....	154
Table 8.E.4	CAPA Disability Distributions: Level II.....	154
Table 8.E.5	CAPA Disability Distributions: Level III.....	155
Table 8.E.6	CAPA Disability Distributions: Level IV.....	155
Table 8.E.7	CAPA Disability Distributions: Level V.....	156
Table 10.A.1	Number of Examinees Tested, Scale Score Means and Standard Deviations of CAPA Across Base Year (2009), 2010, 2011, and 2012.....	169
Table 10.A.2	Percentage of Proficient and Above and Percentage of Advanced Across Base Year (2009), 2010, 2011, and 2012.....	169
Table 10.A.3	Observed Score Distributions of CAPA Across Base Year (2009), 2010, 2011, and 2012 for ELA.....	170
Table 10.A.4	Observed Score Distributions of CAPA Across Base Year (2009), 2010, 2011, and 2012 for Mathematics.....	171
Table 10.A.5	Observed Score Distributions of CAPA Across Base Year (2009), 2010, 2011, and 2012 for Science.....	172
Table 10.B.1	Average Item Score of CAPA Operational Test Items Across Base Year (2009), 2010, 2011, and 2012.....	173
Table 10.B.2	Mean IRT <i>b</i> -values for Operational Test Items Across Base Year (2009), 2010, 2011, and 2012.....	173
Table 10.B.3	Mean Polyserial Correlation of CAPA Operational Test Items Across Base Year (2009), 2010, 2011, and 2012.....	174
Table 10.B.4	Score Reliabilities and SEM of CAPA Across Base Year (2009), 2010, 2011, and 2012.....	174

Figures

Figure 3.1	The ETS Item Development Process for the STAR Program.....	17
Figure 8.1	Decision Accuracy for Achieving a Performance Level.....	74
Figure 8.2	Decision Consistency for Achieving a Performance Level.....	74

Acronyms and Initialisms Used in the *CAPA Technical Report*

1PPC	1-parameter partial credit	IEP	individualized education program
ADA	Americans with Disabilities Act	IRF	item response functions
AERA	American Educational Research Association	IRT	task (item) response theory
AIS	average task (item) score	IT	Information Technology
API	Academic Performance Index	LEA	local educational agency
ARP	Assessment Review Panel	MH	Mantel-Haenszel
AYP	Adequate Yearly Progress	MR/ID	Mental retardation/intellectual disability
CAPA	California Alternate Performance Assessment	NCME	National Council on Measurement Education
CCR	<i>California Code of Regulations</i>	NPS	nonpublic, nonsectarian school
CDE	California Department of Education	NSLP	National School Lunch Program
CDS	County-District-School	PSAA	Public School Accountability Act
CELDT	California English Language Development Test	QC	quality control
CI	confidence interval	RACF	Random Access Control Facility
CMA	California Modified Assessment	SBE	State Board of Education
CSEMs	conditional standard errors of measurement	SD	standard deviation
CSTs	California Standards Tests	SEM	standard error of measurement
DIF	Differential Task (Item) Functioning	SFTP	secure file transfer protocol
DPLT	designated primary language test	SGID	School and Grade Identification sheet
DQS	Data Quality Services	SMD	standardized mean difference
EC	Education Code	SPAR	Statewide Pupil Assessment Review
EM	expectation maximization	STAR	Standardized Testing and Reporting
ESEA	Elementary and Secondary Education Act	STAR TAC	STAR Technical Assistance Center
ETS	Educational Testing Service	STS	Standards-based Tests in Spanish
GENASYS	Generalized Analysis System	TIF	test information function
HumRRo	Human Resource Research Organization	WRMSD	weighted root-mean-square differences
ICC	task (item) characteristic curve		

Chapter 1: Introduction

Background

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

In order to measure and evaluate student achievement of the content standards, the state instituted the Standardized Testing and Reporting (STAR) Program. This Program, administered annually, was authorized in 1997 by state law (Senate Bill 376).

During its 2012 administration, the STAR Program had four components:

- California Standards Tests (CSTs), produced for California public schools to assess the California content standards for ELA, mathematics, history–social science, and science in grades two through eleven
- California Modified Assessment (CMA), an assessment of students’ achievement of California’s content standards for ELA, mathematics, and science, developed for students with an individualized education program (IEP) who meet the CMA eligibility criteria approved by the SBE
- California Alternate Performance Assessment (CAPA), produced for students with an IEP and who have significant cognitive disabilities and are not able to take the CSTs with accommodations and/or modifications or the CMA with accommodations
- Standards-based Tests in Spanish (STS), an assessment of students’ achievement of California’s content standards for Spanish-speaking English learners that is administered as the STAR Program’s designated primary language test (DPLT)

Test Purpose

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

CAPA results are used in the school and district Academic Performance Index (API) calculations. In addition, CAPA results in grades two through eight and grade ten for ELA and mathematics are used in determining Adequate Yearly Progress (AYP), which applies toward meeting the requirement of the federal Elementary and Secondary Education Act (ESEA) that all students score at the proficient level or above by 2014.

Test Content

Students in grades two through eleven who take the CAPA are administered one of the five levels of the CAPA ELA and mathematics tests. In addition, students in grades five, eight, and ten take a grade-level science test.

The five levels of the CAPA are as follows:

- Level I, for students who are in grades two through eleven with the most significant cognitive disabilities
- Level II, for students who are in grades two and three
- Level III, for students who are in grades four and five
- Level IV, for students who are in grades six through eight
- Level V, for students who are in grades nine through eleven

Table 1.1 displays CAPA levels for tests administered in 2012 by grade, content area, and age ranges for ungraded programs.

Table 1.1 Description of the CAPA Assessment Levels

Test Level	I	II	III	IV	V
Grades	2–11	2 and 3	4 and 5	6–8	9–11
Content Area	ELA	ELA	ELA	ELA	ELA
	Mathematics	Mathematics	Mathematics	Mathematics	Mathematics
	Science Grades 5, 8, and 10 only	–	Science Grade 5 only	Science Grade 8 only	Science Grade 10 only
Age Ranges for Ungraded Programs	7–16	7 & 8	9 & 10	11–13	14–16

Intended Population

All students enrolled in grades two through eleven in California public schools on the day testing begins are required to take the CSTs, the CMA (available for students in grades three through eleven in ELA, grades three through seven in mathematics, end-of-course Algebra I and Geometry, and grades five, eight, and ten in science), or 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.

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

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

Parents may submit a written request to have their child exempted from taking any or all parts of the tests within the STAR Program. Only students whose parents/guardians 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 STAR Program 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> (outside source):

“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.”

In addition, STAR Program assessments are used to provide data for school, district, and state purposes and to meet federal accountability requirements.

Testing Window

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

The CAPA tests are untimed. This assessment is administered individually and the testing time varies from one student to another, based on factors such as the student’s response time and attention span. A student may be tested with the CAPA over as many days as required within the school district’s testing window (*California Code of Regulations [CCR], Title 5, Education, Division 1, Chapter 2, Subchapter 3.75, Article 2, § 855*; in the California Department of Education [CDE] Web document linked at <http://www.cde.ca.gov/ta/tg/sr/resources.asp>).

Significant STAR Developments in 2012

Release of Results Delayed to Investigate Psychometric Impact of Social Network Security Breaches

This breach involved CST items only. No CAPA, CMA, or STS items were found to be exposed. However, release of results was delayed for all STAR tests.

The spring 2012 security breach of multiple test questions through social networking Web sites raised serious concerns that required comprehensive investigations and additional statistical analyses. In recognizing the importance and the need to provide valid and reliable results to the state, districts, and schools, both the CDE and Educational Testing Service (ETS) took every precaution necessary, including extensive statistical analyses, to ensure that all test results maintained the highest levels of psychometric integrity.

A preliminary investigation of the scope and magnitude of the security breach was completed by the end of May that resulted in the list of exposed test questions (test items) that were found on social networking sites. This process was needed to determine whether the exposed test questions were operational items that contributed to students' test scores, and more importantly, were linking items used in statistical analyses to maintain comparability of scores obtained on different test forms administered from one year to the next.

After the list of exposed items was determined, ETS proceeded to conduct statistical analyses to determine whether any test questions affected by the security breach needed to be removed from scoring to maintain the validity of test results. An exhaustive review of statistical results did not provide evidence of performance changes that differed from that of nonexposed items. From a statistical perspective, results did not indicate the need to remove exposed items from scoring. However, to minimize the potential impact of exposed items, schools with confirmed item exposure were removed from statistical analyses.

Statistical analyses also included an assessment of the impact of removing exposed items from score linking analyses. Results showed that for most of the affected CSTs, scale scores and raw score performance cuts produced with exposed items retained or removed from analyses were very similar. In addition, for all tests, there was no change in the percent of students classified as proficient or above when exposed items were retained or removed from the analyses.

Extension of the STAR Testing Window

The STAR testing window was extended by 4 days, to 25 days: 12 instructional days before and 12 instructional days after the date on which 85 percent of each school's, track's, or program's instructional days had been completed (formerly 21 days, 10 instructional days before and 10 instructional days after 85 percent of instructional days had been completed) (5 CCR Section 855 [a] [1]).

Limitations of the Assessment

Score Interpretation

Teachers and administrators should not use STAR results in isolation to make inferences about instructional needs. In addition, it is important to remember that a single test can provide only limited information. Other relevant information should be considered as well. It is advisable for parents to evaluate their child's strengths and weaknesses in the relevant topics by reviewing classroom work and progress reports in addition to the child's CAPA

results (CDE, 2012). It is important to note that student scores in a content area contain measurement error and could vary if students were retested.

Out-of-Level Testing

With the exception of Level I, each CAPA is designed to measure the content corresponding to a specific grade or grade span and is appropriate for students in the specific grade or grade span. Testing below a student's grade is not allowed for the CAPA or any test in the STAR Program; all students are required to take the test for the grade in which they are enrolled. School districts are advised to review all IEPs to ensure that any provision for testing below a student's grade level has been removed.

Score Comparison

When comparing results for the CAPA, the reviewer is limited to comparing results only within the same content area and CAPA level. For example, it is appropriate to compare scores obtained by students and/or schools on the 2012 CAPA Level II (Mathematics) test. Similarly, it is appropriate to compare scores obtained on the 2011 CAPA Level IV (ELA) test with those obtained on the CAPA Level IV (ELA) test administered in 2012. It is not appropriate to compare scores obtained on Levels II and IV of the ELA or mathematics tests, nor is it appropriate to compare ELA scores with mathematics scores. Since new score scales and cut scores were used for the 2009 CAPA tests, results from tests administered after 2009 cannot meaningfully be compared to results obtained in previous years.

Groups and Organizations Groups Involved in Test Development

State Board of Education

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

The SBE is responsible for assuring the compliance with programs that meet the requirement of the federal ESEA and the state's Public School Accountability Act (PSAA) and for reporting results in terms of the AYP and API, which measure the academic performance and growth of schools on a variety of academic measures. In order to provide information on student progress in public schools, as essential for those programs, the SBE supervises the administration and progress of the STAR Program.

California Department of Education

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

Contractors

Educational Testing Service

The CDE and the SBE contract with ETS to develop and administer the STAR Program. 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, school districts, and independently testing charter schools;
- Implementation and maintenance of the STAR 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 STAR Program. Activities conducted by Pearson include the following:

- Production of all scannable test materials;
- Packaging, distribution, and collection of testing materials to school districts and independently testing charter schools;
- Scanning and scoring of all responses, including performance scoring of the writing responses; and
- Production of all score reports and data files of test results.

Overview of the Technical Report

This technical report addresses the characteristics of the CAPA administered in spring 2012. The technical report contains nine additional chapters as follows:

- Chapter 2 presents a conceptual overview of processes involved in a testing cycle for a CAPA. 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 also included in this chapter. Also included are the references to various chapters that detail the processes briefly discussed in this chapter.
- Chapter 3 describes the procedures followed during the development of valid CAPA tasks; the chapter explains the process of field-testing new tasks and the review of tasks by contractors and content experts.
- Chapter 4 details the content and psychometric criteria that guided the construction of the CAPA for 2012.
- Chapter 5 presents the processes involved in the actual administration of the 2012 CAPA with an emphasis on efforts made to ensure standardization of the tests. It also includes a detailed section that describes the procedures that were followed by ETS to ensure test security.
- Chapter 6 describes the standard-setting process previously conducted to establish new cut scores.
- Chapter 7 details the types of scores and score reports that are produced at the end of each administration of the CAPA.
- Chapter 8 summarizes the results of the task (item)-level analyses performed during the spring 2012 administration of the tests. These include the classical item analyses, the reliability analyses that include assessments of test reliability and the consistency and accuracy of the CAPA performance-level classifications, and the procedures designed

to endure the validity of CAPA score uses and interpretations. Also discussed in this chapter are the item response theory (IRT) and model-fit analyses, as well as documentation of the equating along with CAPA conversion tables. Finally, the chapter summarizes the results of analyses investigating the differential item functioning (DIF) for each CAPA.

- Chapter 9 highlights the importance of controlling and maintaining the quality of the CAPA.
- Chapter 10 presents historical comparisons of various task (item)- and test-level results for the past three years and for the 2009 base year.

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

References

California *Code of Regulations, Title 5*, Education, Division 1, Chapter 2, Subchapter 3.75, Article 2, § 855.

California Department of Education. (2012). *STAR Program information packet for school district and school staff*. Sacramento, CA. Downloaded from <http://www.cde.ca.gov/tg/sr/resources.asp>

California Department of Education, EdSource, & the Fiscal Crisis Management Assistance Team. (2012). *Fiscal, demographic, and performance data on California's K–12 schools*. Sacramento, CA: Ed-Data. Downloaded from http://www.ed-data.k12.ca.us/App_Resx/EdDataClassic/fsTwoPanel.aspx?#!bottom=/layouts/EdDataClassic/profile.asp?Tab=1&level=04&reportNumber=16

Chapter 2: An Overview of CAPA Processes

This chapter provides an overview of the processes involved in a typical test development and administration cycle for the CAPA. Also described are the specifications maintained by ETS to implement each of those processes. The chapter is organized to provide a brief description of each process followed by a summary of the associated specifications. More details about the specifications and the analyses associated with each process are described in other chapters that are referenced in the sections that follow.

Task (Item) Development

Task Formats

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

Task (Item) Specifications

The CAPA tasks are developed to measure California content standards and designed to conform to principles of task writing defined by ETS (ETS, 2002). ETS maintains and updates a task specifications document, otherwise known as “task writer guidelines,” for each CAPA and has developed an item utilization plan to guide the development of the tasks for each content area. Task writing emphasis is determined in consultation with the CDE.

The task specifications describe the characteristics of the tasks that should be written to measure each content standard. The task specifications help ensure that the tasks in the CAPA measure the content standards in the same way. To do this, the task specifications provide detailed information to task writers that are developing tasks for the CAPA.

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

Item Banking

Before newly developed tasks are placed in the item bank, ETS prepares the tasks for review by content experts and various external review organizations such as the Assessment Review Panels (ARPs), which are described in Chapter 3 starting on page 21; and the Statewide Pupil Assessment Review (SPAR) panel, described in Chapter 3 starting on page 23.

Once the ARP review is complete, the tasks are placed in the item bank along with the associated information obtained at the review sessions. Tasks that are accepted by the content experts are updated to a “field-test ready” status. ETS then delivers the tasks to the CDE by means of a delivery of the California electronic item bank. Tasks are subsequently field-tested to obtain information about task performance and task (item) statistics that can be used to assemble operational forms. The CDE then reviews the task data and makes decisions about which tasks could be used operationally (see page 25 for more information about the CDE's data review). Any additional updates to task content and statistics are based on data collected from the operational use of the tasks. However, only the latest content of the task is retained in the bank at any time, along with the administration data from every administration that has included the task.

Further details on item banking are presented on page 25 in Chapter 3.

Task Refresh Rate

The item utilization plan assumes that each year, 25 percent of tasks on an operational form are refreshed (replaced); these tasks remain in the item bank for future use.

Test Assembly

Test Length

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

Table 2.1 CAPA Item and Estimated Time Chart

ITEMS AND ESTIMATED TIME CHART		
CAPA Content Area	Grades 2–11	
	Items	Times
English–Language Arts	12	45 minutes
Mathematics	12	45 minutes
Science	12	45 minutes

Test Blueprints

ETS selects all CAPA tasks to conform to the SBE-approved California content standards and test blueprints. The CAPA has been revised to better link it to the grade-level California content standards. The revised blueprints for the CAPA were approved by the SBE in 2006 for implementation beginning in 2008. The test blueprints for the CAPA can be found on the CDE STAR CAPA Blueprints Web page at <http://www.cde.ca.gov/ta/tg/sr/capablueprints.asp>.

Content Rules and Task Selection

When developing a new test form for a given CAPA level and content area, test developers follow a number of rules. First and foremost, they select tasks that meet the blueprint for that level and content area. Using the electronic item bank, assessment specialists begin by identifying a number of linking tasks. These are tasks that appeared in previous operational test administrations and are then used to equate the subsequent (new) test forms. After the linking tasks are approved, assessment specialists populate the rest of the test form.

Linking tasks are selected to proportionally represent the full blueprint. Each CAPA form is a collection of test tasks designed to reflect a reliable, fair, and valid measure of student learning within well-defined course content.

Another consideration is the difficulty of each task. Test developers strive to ensure that there are some easy and some hard tasks and that there are a number of tasks in the middle range of difficulty. The detailed rules are presented in Chapter 4, which begins on page 27.

Psychometric Criteria

The staff assesses the projected test characteristics during the preliminary review of the assembled forms. The statistical targets used for the 2012 test development and the projected characteristics of the assembled forms are presented starting from page 28 in Chapter 4.

The tasks in test forms are organized and sequenced to meet the requirements of the content area. Further details on the arrangement of tasks during test assembly are described on page 29 in Chapter 4.

Test Administration

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

Test Security and Confidentiality

All tests within the STAR Program are secure documents. For the CAPA 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 30.

In its 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 30.

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

Procedures to Maintain Standardization

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

Test Administrators

The CAPA is administered in conjunction with the other tests that comprise the STAR Program. ETS employs personnel who facilitate various processes involved in the standardization of an administration cycle.

Staff at school districts who are central to the processes include district STAR coordinators, test site coordinators, test examiners, test proctors, and observers. The responsibilities for each of the staff members are included in the *STAR District and Test Site Coordinator Manual* (CDE, 2012a); see page 36 in Chapter 5 for more information.

Test Directions

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

CAPA Examiner's Manual—The manual used by test examiners to administer and score the CAPA to be followed exactly so that all students have an equal opportunity to demonstrate their academic achievement (See page 36 in Chapter 5 for more information.)

District and Test Site Coordinator Manual—Test administration procedures for district STAR coordinators and test site coordinators (See page 36 in Chapter 5 for more information.)

STAR Management System manuals—Instructions for the Web-based modules that allow district STAR 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 STAR Management System (See page 37 in Chapter 5 for more information.)

Training in the form of “CAPA Train-the-Trainer” workshops is available in January and is presented in live workshops and a Webcast, which is later archived. A school district representative who takes the training can then train test site staff to train CAPA examiners and observers. Video segments that model CAPA task administration are made available during the school year; sample materials that support the training are available all year on the [startest.org](http://www.startest.org) Web site, at <http://www.startest.org/capa.html>.

Test Variations, Accommodations, and Modifications

All public school students participate in the STAR Program, including students with disabilities and English learners. Students with an IEP and who have significant cognitive disabilities may take the CAPA when they are unable to take the CSTs with or without accommodations and/or modifications or the CMA with accommodations.

Examiners may adapt the CAPA in light of a student’s instructional mode as specified in each student’s IEP or Section 504 plan in one of two ways: (1) suggested adaptations for particular tasks, as specified in the task preparation; and (2) core adaptations that are applicable for many of the tasks. Details of the adaptations are presented in the core adaptations of the *CAPA Examiner's Manual* (CDE, 2012b).

As noted on the CDE CAPA Participation Criteria Web page, “Since examiners may adapt the CAPA based on students’ instructional mode, accommodations and modifications do not apply to the CAPA.” (CDE, 2012c)

Scores

The CAPA total test raw scores equal the sum of examinees’ scores on the operational tasks. Raw scores for Level I range from 0 to 40; for the other CAPA levels, the raw-score range is from 0 to 32. Total test raw scores are transformed to two-digit scale scores using the scaling process described starting on page 14. CAPA results are reported through the use of these scale scores; the scores range from 15 to 60 for each test. Also reported are performance levels obtained by categorizing the scale scores into the following levels: far below basic, below basic, basic, proficient, and advanced. The state’s target is for all students to score at the proficient or advanced level.

Detailed descriptions of CAPA scores are described starting on page 45 in Chapter 7.

Aggregation Procedures

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

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

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

Individual Scores

Table 7.2 through Table 7.4 starting on page 48 in Chapter 7 provide summary statistics for individual scores aggregated at the state level, describing overall student performance on each CAPA. Included in the tables are the possible and actual ranges and the means and standard deviations of student scores, expressed in terms of both raw scores and scale scores. The tables also present statistical information about the CAPA tasks.

Demographic Subgroup Scores

Statistics summarizing CAPA student performance by content area and for selected groups of students are provided in Table 7.B.1 through Table 7.B.3 starting on page 59 in Appendix 7.B. In these tables, students are grouped by demographic characteristics, including gender, ethnicity, English-language fluency, primary disability, and economic status. The tables show the numbers of students with valid scores in each group, scale score means and standard deviations as well as percentage in performance level for each demographic group. Table 7.6 on page 52 provides definitions for the demographic groups included in the tables.

Equating

Each CAPA is equated to a reference form using a common-item nonequivalent groups data collection design and methods based on item response theory (IRT) (Hambleton & Swaminathan, 1985). The “base” or “reference” calibrations for the CAPA were established by calibrating samples of data from the 2009 administration. Doing so established a scale to which subsequent item calibrations could be linked. The 2012 task parameter estimates were placed on the reference 2009 scale using a set of linking items selected from the 2011 forms and readministered in 2012.

The procedure used for equating the CAPA involves three steps: calibration, scaling, and linear transformation. Each of those procedures, as described below, is applied to all CAPA tests.

Calibration

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

The PARSCALE calibrations are 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 are used as starting values for a second PARSCALE run, in which the subject prior distribution is updated after each expectation maximization (EM) cycle with no constraints. For both stages, the metric of the scale is controlled by the constant discrimination parameters.

Scaling

Calibrations of the 2012 tasks were linked to the previously obtained reference scale estimates using linking tasks and the Stocking and Lord (1983) procedure. In the case of the one-parameter model calibrations, this procedure is equivalent to setting the mean of the new task parameter estimates for the linking set equal to the mean of the previously scaled estimates. As noted earlier, the linking set is a collection of tasks in a current test form that also appeared in last year’s form and was scaled at that time.

The linking process is carried out iteratively by inspecting differences between the transformed new and old (reference) estimates for the linking tasks and removing tasks for which the difficulty estimates changed significantly. Tasks with large weighted root-mean-square differences (WRMSDs) between item characteristic curves (ICCs) based on the old and new difficulty estimates were removed from the linking set. The differences are 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 a given score for the transformed new form item at ability θ_j , and

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

Based on established procedures, any linking items for which the WRMSD was greater than 0.625 for Level I and 0.500 for Levels II through V were eliminated from the linking set. This criterion has produced reasonable results over time in similar equating work done with other testing programs at ETS.

Linear Transformation

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

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

where,

θ represents the student ability,

$\theta_{\text{proficient}}$ represents the theta cut score for proficient on the spring 2009 base scale, and

θ_{basic} represents the theta cut score for basic on the spring 2009 base scale.

Complete raw-to-scale score conversion tables for the 2012 CAPA are presented in Table 8.D.10 through Table 8.D.23 in Appendix 8.D, starting on page 138. The raw scores and corresponding transformed scale scores are listed in those tables.

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

Table 2.2 Scale Scores Ranges for Performance Levels

Content Area	CAPA Level	Far Below Basic	Below Basic	Basic	Proficient	Advanced
English–Language Arts	I	15	16 – 29	30 – 34	35 – 39	40 – 60
	II	15 – 18	19 – 29	30 – 34	35 – 39	40 – 60
	III	15 – 23	24 – 29	30 – 34	35 – 39	40 – 60
	IV	15 – 17	18 – 29	30 – 34	35 – 41	42 – 60
	V	15 – 22	23 – 29	30 – 34	35 – 39	40 – 60
Mathematics	I	15	16 – 29	30 – 34	35 – 38	39 – 60
	II	15 – 17	18 – 29	30 – 34	35 – 40	41 – 60
	III	15	16 – 29	30 – 34	35 – 39	40 – 60
	IV	15	16 – 29	30 – 34	35 – 40	41 – 60
	V	15 – 16	17 – 29	30 – 34	35 – 39	40 – 60
Science	I	15	16 – 29	30 – 34	35 – 38	39 – 60
	III	15 – 21	22 – 29	30 – 34	35 – 39	40 – 60
	IV	15 – 19	20 – 29	30 – 34	35 – 39	40 – 60
	V	15 – 20	21 – 29	30 – 34	35 – 38	39 – 60

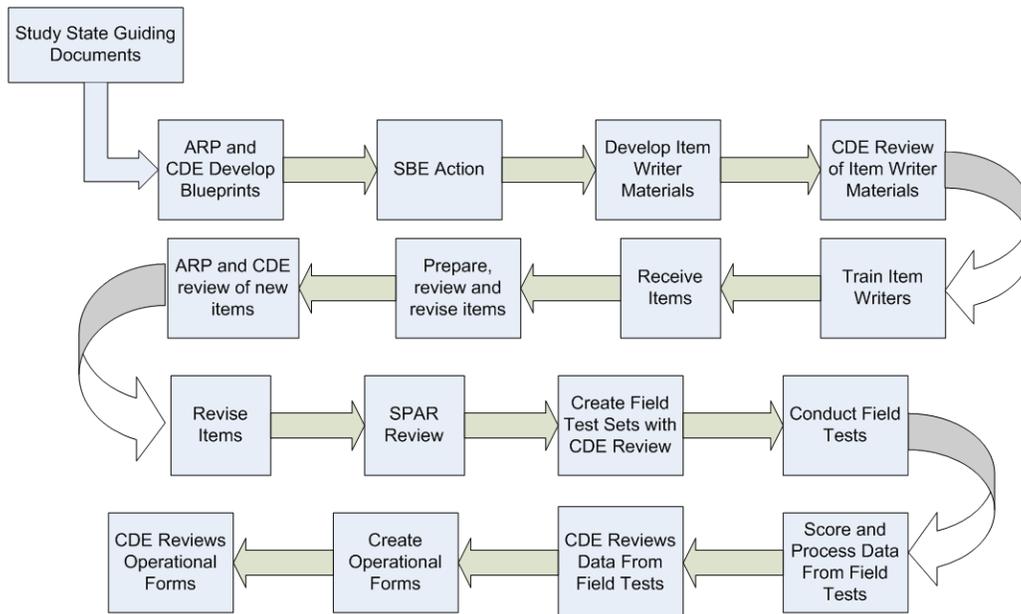
References

- California Department of Education. (2012a). *2012 STAR district and test site coordinator manual*. Sacramento, CA. Downloaded from http://www.startest.org/pdfs/STAR.coord_man.2012.pdf
- California Department of Education. (2012b). *2012 California Alternate Performance Assessment (CAPA) examiner's manual*. Sacramento, CA. Downloaded from http://www.startest.org/pdfs/CAPA.examiners_manual.nonsecure.2012.pdf
- California Department of Education. (2012c). *CAPA participation criteria*. Downloaded from <http://www.cde.ca.gov/TA/tg/sr/participcritria.asp>
- Educational Testing Service. (2002). *ETS standards for quality and fairness*. Princeton, NJ: Author.
- Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory: principles and applications*. Boston, MA: Kluwer-Nijhoff.
- Linacre, J. M. (2000). *WINSTEPS: Rasch measurement* (Version 3.23). Chicago, IL: MESA Press.
- Stocking, M. L., and Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7, pp. 201–10.

Chapter 3: Task (Item) Development

The CAPA tasks are developed to measure California’s content standards and designed to conform to principles of item writing defined by ETS (ETS, 2002). Each CAPA task goes through a comprehensive development cycle as is described in Figure 3.1, below.

Figure 3.1 The ETS Item Development Process for the STAR Program



Rules for Task Development

The development of CAPA tasks follows guidelines for task writing approved by the CDE. These guidelines direct a task writer to assess a task for the relevance of the information being assessed, its relevance to the California content standards, its match to the test and task specifications, and its appropriateness to the population being assessed. As described below, tasks are eliminated early in a rigorous task review process when they are only peripherally related to the test and task specifications, do not measure core outcomes reflected in the California content standards, or are not developmentally appropriate.

Task Specifications

ETS senior content staff leads the task writers in the task development and review process. In addition, experienced ETS content specialists and assessment editors review each task during the forms-construction process. The lead assessment specialists for each content area work directly with the other ETS assessment specialists to carefully review and edit each task for such technical characteristics as quality, match to content standards, and conformity with California-approved task-writing practices. ETS follows the SBE-approved item utilization plan to guide the development of the tasks for each content area. Task specification documents include a description of the constructs to be measured and the California content standards. Those specifications help to ensure that the CAPA measures the content standards in the same way each year. The task specifications also provide specific and important guidance to task writers.

The task specifications describe the general characteristics of the tasks for each content standard, indicate task types or content to be avoided, and define the content limits for the tasks. More specifically, the specifications include the following:

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

In addition, the ELA task specifications that contain guidelines for stimulus cards used to assess reading comprehension include the following:

- A list of topics to be avoided
- The acceptable ranges for the number of words on a stimulus card
- Expected use of artwork
- The target number of tasks attached to each reading stimulus card

Expected Task Ratio

ETS developed the item utilization plan to continue the development of CAPA tasks. The plan includes strategies for developing tasks that will permit coverage of all appropriate standards for all tests in each content area and at each grade level. ETS test development staff uses this plan to determine the number of tasks to develop for each content area.

The item utilization plan assumes that each year, 25 percent of items on an operational form would be refreshed (replaced); these items remain in the item bank for future use.

The item utilization plan also declares that an additional five percent of the operational items are likely to become unusable because of normal attrition, and notes that there is a need to focus development on “critical” standards, which are standards that are difficult to measure well or for which there are few usable items.

Each year, ETS field tests 16 tasks per CAPA level for both ELA and mathematics, and eight tasks per CAPA level for science. Given that each test contains eight operational tasks, the ratios of field-test to operational tasks are 200 percent for ELA and mathematics and 100 percent for science for each CAPA level. These task ratios would allow for a five percent attrition rate while gradually increasing the overall size of the CAPA item bank. The field-test percentages and task counts are presented in Table 3.1.

Table 3.1 Field-test Percentages for the CAPA

Content Area	Number of Operational Tasks per CAPA level	Field-test Percentage per CAPA level	Number of Tasks to Be Field-tested per CAPA level
English–Language Arts	8	200%	16
Mathematics	8	200%	16
Science	8	100%	8

Selection of Task Writers

Criteria for Selecting Task Writers

The tasks for each CAPA are written by individual task writers who have a thorough understanding of the California content standards. Applicants for task writing are screened by senior ETS content staff. Only those with strong content and teaching backgrounds are approved for inclusion in the training program for task writers. Because most of the participants are current or former California educators, they are particularly knowledgeable about the standards assessed by the CAPA. All task writers meet the following minimum qualifications:

- Possession of a bachelor's degree in the relevant content area or in the field of education with special focus on a particular content of interest; an advanced degree in the relevant content area is desirable
- Previous experience in writing tasks for standards-based assessments, including knowledge of the many considerations that are important when developing tasks to measure state-specific standards
- Previous experience in writing tasks in the content areas covered by CAPA levels
- Familiarity, understanding, and support of the California content standards
- Current or previous teaching experience in California, when possible
- Knowledge about the abilities of the students taking the tests

Task (Item) Review Process

The tasks selected for the CAPA undergo an extensive task review process that is designed to provide the best standards-based tests possible. This section summarizes the various reviews performed to ensure the quality of the CAPA tasks and test forms.

Contractor Review

Once the tasks have been written, ETS employs a series of internal reviews. The reviews establish the criteria used to judge the quality of the task content and are designed to ensure that each task is measuring what it is intended to measure. The internal reviews also examine the overall quality of the tasks before they are prepared for presentation to the CDE and the Assessment Review Panels (ARPs). Because of the complexities involved in producing defensible tasks for high-stakes programs such as the STAR Program, it is essential that many experienced individuals review each task before it is brought to the CDE, the ARPs, and Statewide Pupil Assessment Review (SPAR) panels.

The ETS review process for the CAPA includes the following:

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

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

1. Internal Content Review

Test tasks and materials undergo two reviews by the content-area assessment specialists. These assessment specialists make sure that the test tasks and related materials are in compliance with ETS's written guidelines for clarity, style, accuracy, and appropriateness for California students as well as in compliance with the approved task specifications.

Assessment specialists review each task on the basis of the following characteristics:

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

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

2. Internal Editorial Review

After the content-area assessment specialists review each task, a group of specially trained editors reviews each task in preparation for review by the CDE and the ARPs. The editors check tasks for clarity, correctness of language, appropriateness of language for the grade level assessed, adherence to the style guidelines, and conformity with accepted task-writing practices.

3. Internal Sensitivity Review

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

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

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

Content Expert Reviews

Assessment Review Panels

ETS is responsible for working with ARPs as tasks are developed for the CAPA. The ARPs are advisory panels to the CDE and ETS and provide guidance on matters related to task development for the CAPA. The ARPs are responsible for reviewing all newly developed tasks for alignment to the California content standards. The ARPs also review the tasks for accuracy of content, clarity of phrasing, and quality. In their examination of test tasks, the ARPs may raise concerns related to age/level appropriateness and gender, racial, ethnic, and/or socioeconomic bias.

Composition of ARPs

The ARPs are composed of current and former teachers, resource specialists, administrators, curricular experts, and other education professionals. Current school staff members must meet minimum qualifications to serve on the CAPA ARPs, including:

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

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

Current ARP members are recruited through an application process. Recommendations are solicited from school districts and county offices of education as well as from CDE and SBE staff. Applications are received and reviewed throughout the year. They are reviewed by the ETS assessment directors, who confirm that the applicant's qualifications meet the specified criteria. Applications that meet the criteria are forwarded to CDE and SBE staff for further review and agreement on ARP membership. Upon approval, the applicant is notified that he or she has been selected to serve on the ARP committee.

Table 3.2, on the next page, shows the educational qualifications, present occupation, and credentials of the current CAPA ARP members.

Table 3.2 CAPA ARP Member Qualifications, by Content Area and Total

CAPA	ELA	Math	Science	Total
Total	9	9	6	24
Occupation (Members may teach multiple levels.)				
Teacher or Program Specialist, Elementary/Middle School	5	3	2	10
Teacher or Program Specialist, High School	1	1	2	4
Teacher or Program Specialist, K–12	5	4	4	13
University Personnel	0	0	0	0
Other District Personnel (e.g., Director of Special Services, etc.)	2	1	0	3
Highest Degree Earned				
Bachelor’s Degree	4	4	2	10
Master’s Degree	5	5	4	14
Doctorate	0	0	0	0
K–12 Teaching Credentials and Experience (Members may hold multiple credentials.)				
Elementary Teaching (multiple subjects)	4	3	1	8
Secondary Teaching (single subject)	0	1	4	5
Special Education	5	7	5	17
Reading Specialist	0	0	0	0
English Learner (CLAD, BCLAD)	1	1	1	3
Administrative	1	1	2	4
Other	0	0	0	0
None (teaching at the university level)	0	0	0	0

ARP Meetings for Review of CAPA Tasks

ETS content-area assessment specialists facilitate the CAPA ARP meetings. Each meeting begins with a brief training session on how to review tasks. ETS provides this training, which consists of the following topics:

- Overview of the purpose and scope of the CAPA
- Overview of the CAPA’s test design specifications and blueprints
- Analysis of the CAPA task specifications
- Overview of criteria for reviewing constructed-response tasks
- Review and evaluation of tasks for bias and sensitivity issues

Criteria also involve more global factors, including—for ELA—the appropriateness, difficulty, and readability of reading stimulus cards. The ARPs also are trained on how to make recommendations for revising tasks.

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

Does the task:

- Measure the content standard?
- Match the test task specifications?
- Align with the construct being measured?
- Test worthwhile concepts or information?

- Reflect good and current teaching practices?
- Have wording that gives the student a full sense of what the task is asking?
- Avoid unnecessary wordiness?
- Reflect content that is free of bias against any person or group?

Is the stimulus, if any, for the task:

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

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

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

Statewide Pupil Assessment Review Panel

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

Field Testing

The primary purposes of field testing are to obtain information about task performance and to obtain statistics that can be used to assemble operational forms.

Stand-alone Field Testing

In 2002, for the new CAPA, a pool of tasks was initially constructed by administering the newly developed tasks in a stand-alone field test. In stand-alone field testing, examinees are recruited to take tests outside of the usual testing circumstances and the test results are typically not used for instructional or accountability purposes (Schmeiser & Welch, 2006).

Embedded Field-test Tasks

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

core set of operational tasks and different sets of field-test tasks. The numbers of embedded field-test tasks for the CAPA are shown in Table 3.3.

Allocation of Students to Forms

The test forms for a given CAPA are distributed by random assignment to school districts and independently testing charter schools so that a large representative sample of test takers responds to the field-test items embedded in these forms. The random assignment of specific forms ensures that a diverse sample of students take each field-test task. The students do not know which tasks are field-test tasks and which tasks are operational tasks; therefore, their motivation is not expected to vary over the two types of tasks (Patrick & Way, 2008).

Number of Forms and Sample Sizes

All CAPA assessments consist of four forms. Each form contains eight operational tasks that are the same and four unique tasks being field-tested. Scores on the field-test tasks are not counted toward student scores. See Table 2.1 on page 10 for more details on the test length.

Table 3.3 also shows the number of forms, operational tasks, field-test tasks, and the approximate number of students in the P1 data that took the operational and field-test tasks in spring 2012. The P1 data file contained test results for 100 percent of the entire test-taking population, and all the student records used in the August 30, 2012, reporting of STAR results.

The sample sizes for the field-test tasks are presented as ranges because the numbers of students who took a set of field-test tasks varied over the forms of CAPA.

Table 3.3 Summary of Tasks and Forms Presented in the 2012 CAPA

Content Area	Level	Operational		Field Test		
		No. Tasks	No. Examinees Total (P1)	No. Forms	No. Tasks	No. Examinees Total (P1)
English– Language Arts	I	8	14,098	4	16	2,879–3,654
	II	8	6,668	4	16†	1,245–1,579
	III	8	7,105	4	16	1,402–1,654
	IV	8	10,091	4	16	2,027–2,412
	V	8	10,424	4	16	2,137–2,636
Mathematics	I	8	14,065	4	16	2,872–3,642
	II	8	6,650	4	16	1,245–1,573
	III	8	7,094	4	16	1,398–1,655
	IV	8	10,068	4	16	2,022–2,406
	V	8	10,392	4	16	2,132–2,623
Science	I	8	3,564	4*	8	717–939
	III	8	3,556	4*	8	710–837
	IV	8	3,299	4*	8	628–829
	V	8	3,424	4*	8	709–852

* There are two unique forms and two repeated forms for science tests.

† Two field-test tasks are excluded from the analysis.

CDE Data Review

Once tasks have been field-tested, ETS prepares the tasks and the associated statistics for review by the CDE. ETS provides tasks with their statistical data, along with annotated comments sheets, for the CDE to use in its review. ETS conducts an introductory training to highlight any new issues and serve as a statistical refresher. CDE consultants then make decisions about which tasks should be included in the item bank. ETS psychometric and content staff are available to CDE consultants throughout this process.

Item Banking

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

After field-test or operational use, tasks that do not meet statistical specification may be rejected; such tasks are updated with a status of “rejected for statistical reasons” and remain unavailable in the bank. These statistics are obtained by the psychometrics group at ETS, which carefully evaluates each task for its level of difficulty and discrimination as well as conformance to the Rasch partial credit model. Psychometricians also determine if the task functions similarly for various subgroups of interest.

All unavailable items are marked with an availability indicator of “Unavailable,” a reason for rejection as described above, and cause alerts so they are not inadvertently included on subsequent test forms. Status and availability of a task are updated programmatically as tasks are presented for review, accepted or rejected, placed on a form for field-testing, presented for statistical review, and used operationally. All rejection indications are 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 STAR (CST, CMA, CAPA, and STS). CAHSEE and STAR 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 2005/SQL 2008, already deployed. ETS provides updated versions of the item bank to the CDE on an ongoing basis and works with the CDE to determine the optimum process if a change in databases is desired.

References

- Educational Testing Service (2002). *ETS standards for quality and fairness*. Princeton, NJ: Author.
- Patrick, R., & Way, D. (March, 2008). *Field testing and equating designs for state educational assessments*. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.
- Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R.L. Brennan (Ed.), *Educational measurement* (4th, ed.). Westport, CT: American Council on Education and Praeger Publishers.

Chapter 4: Test Assembly

The CAPA is constructed to measure students' performance relative to California's content standards approved by the SBE. They are also constructed to meet professional standards for validity and reliability. For each CAPA, the content standards and desired psychometric attributes are used as the basis for assembling the test forms.

Test Length

The number of tasks in each CAPA 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 level and content area. Also considered is the goal that the tests be short enough that most of the students complete it in a reasonable amount of time.

Each CAPA consists of 12 tasks, including eight operational tasks and four field-test tasks. For more details on the distribution of items at each level and content area, see Table 3.3 in Chapter 3 on page 24.

Rules for Task Selection

Test Blueprints

ETS develops all CAPA test tasks to conform to the SBE-approved California content standards and test blueprints. The CAPA blueprints were revised and approved by the SBE in 2006 for implementation beginning in 2008.

The California content standards were used as the basis for choosing tasks for the tests. The blueprints for the CAPA can be found on the CDE STAR CAPA Blueprints Web page at <http://www.cde.ca.gov/ta/tg/sr/capablueprints.asp>.

Content Rules and Task Selection

When developing a new test form for a given CAPA level and content area, test developers follow a number of rules. First and foremost, they select tasks that meet the blueprint for that level and content area. Using an electronic item bank, assessment specialists begin by identifying a number of linking tasks. These are tasks that appeared in the previous year's operational administration and are used to equate the test forms administered each year. Linking tasks are selected to proportionally represent the full blueprint. The selected linking tasks are also reviewed by psychometricians to ensure that the specific psychometric criteria are met.

After the linking tasks are approved, assessment specialists populate the rest of the test form. Their first consideration is the strength of the content and the match of each task to a specified content standard. In selecting tasks, team members also try to ensure that they include a variety of formats and content and that at least some of them include graphics for visual interest.

Another consideration is the difficulty of each task. Test developers strive to ensure that there are some easy and some hard tasks, and that there are a number of tasks in the middle range of difficulty. If tasks do not meet all content and psychometric criteria, staff reviews the other available tasks to determine if there are other selections that could improve the match of the test to all of the requirements. If such a match is not attainable, the content team works in conjunction with psychometricians and the CDE to determine which

combination of tasks will best serve the needs of the students taking the test. Chapter 3, starting on page 17, contains further information about this process.

Psychometric Criteria

The three goals of CAPA test development are 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 that outlines the desired psychometric properties of the CAPA has been developed. These rules are referred to as statistical targets. Total test assembly targets are developed for each CAPA. These targets are provided to test developers before a test construction cycle begins.

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

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

The polyserial correlation describes the relationship between student performance on a polytomously scored item and student performance on the test as a whole. It is used as a measure of how well an item discriminates among test takers that differ in their ability and is related to the overall reliability of the test.

Assembly Targets

The target values for the CAPA, presented in Table 4.1, were used to build the spring 2012 operational test forms. These specifications were developed from the analyses of test forms administered in 2009, the base year in which test results were reported using new scales and new cut scores for the five performance levels: far below basic, below basic, basic, proficient, and advanced.

Table 4.1 Statistical Targets for CAPA Test Assembly

Content Area	CAPA Level	Target Mean b	Target SD b	Mean AIS	Mean Polyserial
English–Language Arts	I	–0.39	0.50	2.75	0.80
	II	–0.56	0.50	2.20	0.80
	III	–0.49	0.50	2.20	0.80
	IV	–0.50	0.50	2.20	0.80
	V	–0.61	0.50	2.20	0.80

Content Area	CAPA Level	Target Mean <i>b</i>	Target SD <i>b</i>	Mean AIS	Mean Polyserial
Mathematics	I	-0.27	0.50	2.75	0.80
	II	-0.79	0.50	2.20	0.80
	III	-0.80	0.50	2.20	0.80
	IV	-0.73	0.50	2.20	0.80
	V	-0.79	0.50	2.20	0.80
Science	I	-0.27	0.50	2.75	0.80
	III	-0.76	0.50	2.20	0.80
	IV	-0.61	0.50	2.20	0.80
	V	-0.31	0.50	2.20	0.80

Projected Psychometric Properties of the Assembled Tests

Prior to the 2012 administration, ETS psychometricians performed a preliminary review of the technical characteristics of the assembled tests. Table 4.2 shows the projected statistical attributes of each CAPA based on banked item statistics from the most recent administration of the items comprising the newly assembled 2012 test forms. These values can be compared to the target values in Table 4.1.

Table 4.2 Summary of 2012 CAPA Projected Statistical Attributes

Content Area	CAPA Level	Mean <i>b</i>	SD <i>b</i>	Mean AIS	Min AIS	Max AIS	Mean Polyserial
English–Language Arts	I	-0.53	0.14	3.08	2.73	3.56	0.78
	II	-0.80	0.67	2.40	1.95	3.57	0.72
	III	-0.74	0.33	2.48	2.21	2.99	0.79
	IV	-0.79	0.35	2.33	1.64	2.60	0.78
	V	-0.94	0.52	2.56	1.97	3.12	0.76
Mathematics	I	-0.24	0.13	2.85	2.52	3.23	0.74
	II	-0.93	0.73	2.48	1.22	3.25	0.73
	III	-0.90	0.34	2.38	1.76	2.98	0.66
	IV	-0.82	0.63	2.48	1.48	3.00	0.70
	V	-1.07	0.34	2.64	2.01	3.35	0.75
Science	I	-0.31	0.15	2.87	2.39	3.17	0.80
	III	-1.13	0.30	2.58	2.22	2.91	0.69
	IV	-1.16	0.37	2.69	2.27	2.96	0.68
	V	-0.66	0.44	2.67	2.14	3.33	0.75

Rules for Task Sequence and Layout

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

Chapter 5: Test Administration

Test Security and Confidentiality

All tests within the STAR Program are secure documents. For the CAPA 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.

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

Task and Data Review

ETS enforces security measures at ARP meetings to protect the integrity of meeting materials using the following guidelines:

- Individuals who participate in the ARPs must sign a confidentiality agreement.
- Meeting materials are strictly managed before, during, and after the review meetings.
- Meeting participants are supervised at all times during the meetings.
- Use of electronic devices is prohibited in the meeting rooms.

Item Banking

When the ARP review is complete, the tasks are placed in the item bank. ETS then delivers the tasks to the CDE through the California electronic item bank. Subsequent updates to content and statistics associated with tasks are based on data collected from field testing and the operational use of the tasks. The latest version of the task is retained in the bank along with the data from every administration that has included the task.

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

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

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

Transfer of Forms and Tasks to the CDE

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

Security of Electronic Files Using a Firewall

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

All electronic applications included in the STAR Management System (CDE, 2012a) remain protected by the ETS firewall software at all times. Due to the sensitive nature of the student information processed by the STAR Management System, the firewall plays a significant role in maintaining an assurance of confidentiality in the users of this information. (It should

be noted that the STAR Management System neither stores nor processes tests or student test results.)

Printing and Publishing

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

Test Administration

Pearson receives testing materials from printers, packages them, and sends them to school districts. After testing, the school districts 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 STAR program schedule. The carriers provide inside delivery directly to the district STAR 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 district STAR 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 district STAR coordinator, who is responsible for returning them to the STAR Scoring and Processing Centers. The following measures are in place to ensure security of STAR testing materials:

- District STAR coordinators are required to sign and submit a "STAR Test (Including Field Tests) Security Agreement for District and Test Site Coordinators" form to the STAR Technical Assistance Center before ETS may ship any testing materials to the school district.
- Test site coordinators have to sign and submit a "STAR Test (Including Field Tests) Security Agreement for District and Test Site Coordinators" form to the district STAR coordinator before any testing materials may be delivered to the school/test site.
- Anyone having access to the testing materials must sign and submit a "STAR Test (Including Field Tests) Security Affidavit for Test Examiners, Proctors, Scribes, and Any Other Person Having Access to STAR Tests" form to the test site coordinator before receiving access to any testing materials.
- It is the responsibility of each person participating in the STAR 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 district STAR coordinator. The district STAR 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 unless questions arise about a student's score. School and district personnel are not allowed to look at a completed answer documents unless necessary for the purpose of 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 and follows 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 STAR 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
- 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.)
- Task-level statistics based on the results which are entered into the item bank

Student Confidentiality

To meet ESEA and state requirements, school districts 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, 2012b). 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 in Pre-ID files and reports.

Student Test Results

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

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

ETS has many facilities and procedures that protect computer files. Facilities, policies, software, and procedures such as firewalls, intrusion detection, and virus control are in place to provide for physical security, data security, and disaster recovery. ETS is certified in the BS 25999-2 standard for business continuity and conducts disaster recovery exercises

annually. ETS routinely backs up its data to either disk through deduplication or to tape, both of which are stored off site.

Access to the ETS Computer Processing Center is controlled by employee and visitor identification badges. The Center is secured by doors that can be unlocked only by the badges of personnel who have functional responsibilities within its secure perimeter. Authorized personnel accompany visitors to the 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 the test results of individual students in both electronic files and on paper reports during the following events:

- Scoring
- Transfer of scores by means of secure data exchange
- Reporting
- Posting of aggregate data
- Storage

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

Procedures to Maintain Standardization

The CAPA 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 CAPA tests, as described in this section.

Test Administrators

The CAPA is administered in conjunction with the other tests that comprise the STAR Program. ETS employs personnel who facilitate various processes involved in the standardization of an administration cycle.

The responsibilities for district and test site staff members are included in the *STAR District and Test Site Coordinator Manual* (CDE, 2012c). This manual is described in the next section.

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

District STAR Coordinator

Each local education agency (LEA) designates a district STAR coordinator who is responsible for ensuring the proper and consistent administration of the STAR 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.

District STAR coordinators are also responsible for securing testing materials upon receipt, distributing testing materials to schools, tracking the materials, training and answering

questions from district 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 STAR contractor for processing.

Test Site Coordinator

At each test site, the superintendent of the school district or the district STAR coordinator designates a STAR test site coordinator from among the employees of the school district. (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 school district after testing, and returning the materials to the school district. (CDE, 2012c)

Test Examiner

The CAPA is administered to students individually by test examiners who may be assisted by test proctors and scribes. A test examiner is an employee of a school district or an employee of a nonpublic, nonsectarian school (NPS) who has been trained to administer the tests and has signed a STAR Test Security Affidavit. For the CAPA, the test examiner must be a certificated or licensed school staff member (5 CCR Section 850 [q]). Test examiners must follow the directions in the *CAPA Examiner's Manual* (CDE, 2012d) exactly.

Test Proctor

A test proctor is an employee of the school district 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 STAR Program (5 CCR Section 850 [r]). Test proctors must sign STAR Test Security Affidavits (5 CCR Section 859 [c]).

Observer

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

CAPA Examiner's Manual

The *CAPA Examiner's Manual* describes the CAPA administrative procedures and scoring rubrics and contains the manipulative lists and all the tasks for all the CAPA content area tests at each level. Examiners must follow task preparation guidelines exactly (CDE, 2012d).

District 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 *STAR District and Test Site Coordinator Manual* contributes to this goal by providing information about the responsibilities of district and test site coordinators, as well as those of the other staff involved in the administration cycle (CDE, 2012c). 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.

STAR Management System Manuals

The STAR Management System is a series of secure, Web-based modules that allow district STAR 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 STAR Management System. The modules of the STAR Management System are as follows:

- **Test Administration Setup**—This module allows school districts to determine and calculate dates for scheduling test administrations for school districts, to verify contact information for those school districts, and to update the school district’s shipping information. (CDE, 2012e)
- **Order Management**—This module allows school districts to enter quantities of testing materials for schools. Its manual includes guidelines for determining which materials to order. (CDE, 2012f)
- **Pre-ID**—This module allows school districts to enter or upload student information including demographics and to identify the test(s) the student will take. This information is printed on student test booklets or answer documents or on labels that can be affixed to test booklets or answer documents. Its manual includes the CDE’s Pre-ID layout. (CDE, 2012b)
- **Extended Pre-ID Data Corrections**—This module allows school districts to correct the data that were submitted during Pre-ID prior to the last day of the school district’s selected testing window. (CDE, 2012b)

Accommodations for Students with Disabilities

All public school students participate in the STAR Program, including students with disabilities and English learners. ETS policy states that reasonable testing accommodations be provided to students with documented disabilities that are identified in the Americans with Disabilities Act (ADA). The ADA mandates that test accommodations be individualized, meaning that no single type of test accommodation may be adequate or appropriate for all individuals with any given type of disability. 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 CSTs under standard conditions. However, some students with disabilities and some English learners may need assistance when taking the tests. This assistance takes the form of test variations, accommodations, or modifications. The Matrices of Test Variations, Accommodations, and Modifications for administrations of California Statewide Assessments are provided in Appendix E of the *STAR District and Test Site Coordinator Manual* (CDE, 2012c). Because examiners may adapt the CAPA in light of a student’s instructional mode, accommodations and modifications do not apply to the CAPA.

Adaptations

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

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

The core adaptations may be appropriate for students across many of the CAPA tasks and are provided in the *CAPA Examiner's Manual* (CDE, 2012d), on page 23 of the nonsecure manual.

Scoring

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

Demographic Data Corrections

After reviewing student data, some school districts may discover demographic data or CAPA levels that are incorrect. The Demographic Data Corrections module of the STAR Management System gives school district the means to correct these data within a specified availability window. Districts may correct data to: (1) Have the school district's API/AYP recalculated; (2) Rescore uncoded or miscoded CAPA test levels; (3) Obtain a corrected data CD-ROM for school district records; or (4) Match unmatched records. (CDE, 2012g)

Testing Irregularities

Testing irregularities are circumstances that may compromise the reliability and validity of test results and, if more than five percent of the students tested are involved, could affect a school's API and AYP.

The district STAR coordinator is responsible for immediately notifying the CDE of any irregularities that occur before, during, or after testing. The test examiner is responsible for immediately notifying the district STAR coordinator of any security breaches or testing irregularities that occur in the administration of the test. Once the district STAR coordinator and the CDE have determined that an irregularity has occurred, the CDE instructs the district STAR coordinator on how and where to identify the irregularity on the answer document. The information and procedures to assist in identifying irregularities and notifying the CDE are provided in the *STAR District and Test Site Coordinator Manual* (CDE, 2012c).

Test Administration Incidents

A test administration incident is any event that occurs before, during, or after test administrations that does not conform to the instructions stated in the *CAPA Examiner's Manual* (CDE, 2012d) and the *STAR District and Test Site Coordinator Manual* (CDE, 2012c). These events include test administration errors and disruptions. Test administration incidents generally do not affect test results. These administration incidents are not reported to the CDE or the STAR Program testing contractor. The STAR test site coordinator should immediately notify the district STAR coordinator of any test administration incidents that occur. It is recommended by the CDE that districts and schools maintain records of these incidents.

References

- California Department of Education (2012a). *2012 STAR Management System*.
<http://www.startest.org/sms.html>
- California Department of Education (2012b). *2012 STAR Pre-ID and Extended Pre-ID Data Corrections instructions manual*. Sacramento, CA. Downloaded from
http://www.startest.org/pdfs/STAR.pre-id_xdc_manual.2012.pdf
- California Department of Education (2012c). *2012 STAR district and test site coordinator manual*. Sacramento, CA. Downloaded from http://www.startest.org/pdfs/STAR.coord_man.2012.pdf
- California Department of Education (2012d). *2012 California Alternate Performance Assessment (CAPA) examiner's manual*. Sacramento, CA. Downloaded from
http://www.startest.org/pdfs/CAPA.examiners_manual.nonsecure.2012.pdf
- California Department of Education (2012e). *2012 STAR Test Administration Setup manual*. Sacramento, CA. Downloaded from http://www.startest.org/pdfs/STAR.test_admin_setup.2012.pdf
- California Department of Education (2012f). *2012 STAR Order Management manual*. Sacramento, CA. Downloaded from http://www.startest.org/pdfs/STAR.order_mgmt.2012.pdf
- California Department of Education. (2012g). *2012 STAR Demographic Data Corrections manual*. Sacramento, CA. Downloaded from http://www.startest.org/pdfs/STAR.data_corrections_manual.2012.pdf

Chapter 6: Performance Standards

Background

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

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

Performance standards are developed from a general description of each performance level (policy-level descriptors) and the associated competencies lists, which operationally define each level. Cut scores numerically define the performance levels. This chapter describes the process of developing performance standards which were first applied to the CAPA operational tests in the spring of 2009.

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

Standard Setting Procedure

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

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

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

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

Development of Competencies Lists

Prior to the CAPA standard-setting workshop, ETS facilitated a meeting in which a subset of the standard-setting panelists was assembled to develop lists of competencies based on the California content standards and policy-level descriptors. Four panels of educators were assembled to identify and discuss the competencies required of students in the CAPA levels and content areas for each performance level (below basic, basic, proficient, and advanced). Panels consisted of educators with experience working with students who take the CAPA. Panelists were assigned to one of four panels (Level I, ELA, mathematics, or science) based on experience working with students and administering the CAPA. At the conclusion of the meeting, the CDE reviewed the draft lists and delivered the final lists for

use in standard setting. The lists were used to facilitate the discussion and construction of the target student definitions during the standard-setting workshop.

Standard Setting Methodology

Performance Profile Method

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

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

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

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

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

Results

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

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

each test level and content area; this means that a student must earn a score of 35 or higher to achieve a proficient classification.

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

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

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

where,

θ represents the student ability

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

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

The scale-score ranges for each performance level are presented in Table 2.2 on page 15. The cut score for each performance level is the lower bound of each scale-score range. The scale-score ranges do not change from year to year. Once established, they remain unchanged from administration to administration until such time that new performance standards are adopted.

Table 7.5 on page 51 in Chapter 7 presents the percentages of examinees meeting each performance level in 2012.

References

- Educational Testing Service. (2003). *CAPA standard setting technical report* (California Department of Education Contract Number 5417). Princeton, NJ: Author.
- Educational Testing Service. (2008a). *Technical report on the standard setting workshop for the California Alternate Performance Assessment. December 29, 2008* (California Department of Education Contract Number 5417). Princeton, NJ: Author.
- Educational Testing Service K–12 Statistical Analysis Group. (2008b). *A study to examine the effects of changes to the CAPA Level I rubric involving the hand-over-hand prompt*, Unpublished memorandum. Princeton, NJ: Author.
- Jaeger, R.M. (1995a). Setting performance standards through two-stage judgmental policy capturing. *Applied Measurement in Education*, 8, pp. 15–40.
- Jaeger, R.M. (1995b). Setting standards for complex performances: An iterative, judgmental policy-capturing strategy. *Educational Measurement: Issues and Practice*, 14 (4), pp. 16–20.
- Plake, B. S., & Hamilton, R.K. (2001). The analytic judgment method for setting standards on complex performance assessments. In G.J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives*, (pp. 283–312). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Plake, B., Hamilton, R., & Jaeger, R.M. (1997). A new standard setting method for performance assessments: The dominant profile judgment method and some field-test results. *Educational and Psychological Measurement*, 57, pp. 400–11.
- Putnam, S.E., Pence, P. & Jaeger, R.M. (1995). A multi-stage dominant profile method for setting standards on complex performance assessments. *Applied Measurement in Education*, 8, pp. 57–83.

Chapter 7: Scoring and Reporting

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

The CAPA is composed entirely of performance tasks. Each content area includes eight performance tasks that are scored by a trained examiner using a rubric that depends on the test level being assessed. After the student has responded to a task, the examiner marks the score using the corresponding circle on the student's answer document.

Scoring Rubric

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

Beginning with the administration of the 2009 CAPA, the Level I rubric was changed to take into account issues related to scoring students who required a hand-over-hand prompt (ETS, 2008). ETS believed there was a significant difference between levels of prompting when dealing with this special population of students as evidenced by the amount of special education research that deals exclusively with prompting hierarchies. A child with significant cognitive disabilities who is able to complete a task successfully at one level of prompting may take weeks or months to increase his or her proficiency in that task in order to be able to complete the task successfully at a less intrusive level of prompting. The differences within prompting levels are the reason why ETS supported a rubric that differentiates between levels of prompting and scores the responses accordingly. For Level I ELA, mathematics, and science, all tasks are scored using the same rubric. For all other levels, the rubric is specific to the task. Both rubrics are presented in Table 7.1. Note that a score of zero in Level I indicates that the student did not orient toward a task after multiple prompts had been utilized. In Levels II–V, a score of zero implies that the student did not attempt the task. In both cases, the score is defined as "No Response" for the purpose of scoring the task.

Table 7.1 Rubrics for CAPA Scoring

Level I		Levels II–V	
Score Points	Description	Score Points	Description
5	Correct with no prompting		
4	Correct with verbal or gestural prompt	4	Completes task with 100 percent accuracy
3	Correct with modeled prompt	3	Partially completes task (as defined for each task)
2	Correct with hand-over-hand prompt (student completes task independently)	2	Minimally completes task (as defined for each task)

Level I		Levels II–V	
Score Points	Description	Score Points	Description
1	Orients to task or incorrect response after attempting the task independently	1	Attempts task
0	No response	0	Does not attempt task

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

Scoring and Reporting Specifications

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

- **General Reporting Specifications**—Provides the calculation rules for the information presented on STAR 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 Conversions**—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 school districts 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 including matching observer documents to examiner documents.

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 accord with the scoring specifications that have been approved by the CDE. Answer documents are designed to produce a single complete record for each student. This record includes demographic data and scanned responses for each student; once computed, the scored responses and the total test scores for a student are also merged into the same record. All scores 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 school district must return scorable and nonscorable materials within five working days after the selected last day of testing for each test administration period.

Types of Scores

Raw Score

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

Scale Score

Raw scores obtained on each CAPA test are converted to two-digit scale scores using the calibration process described in Chapter 2 on page 14. Scale scores range from 15 to 60 on each CAPA content-area test. The scale scores of examinees that have been tested in different years at a given CAPA test 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

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

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

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

Score Verification Procedures

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

Monitoring and Quality Control of Scoring

Scorer Selection

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

If the examiner or case carrier is not available to administer the test, it may be administered and scored by another CAPA-trained staff member such as a school psychologist; speech, physical, or occupational therapist; program specialist; or certified teacher, principal or assistant principal. This individual should have experience working with students with significant cognitive disabilities and must be trained to administer the CAPA (CDE, 2012a).

Quality Control

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

Score Verification Process

ETS psychometricians employ special procedures that adjust for differences in item difficulty of one test form to another. (See Chapter 2, Equating, on page 14 for details.) As a result of this process, scoring tables are produced. Such tables map the current year's raw score to an appropriate scale score. A series of quality control (QC) checks are carried out by ETS psychometricians to ensure the accuracy of each scoring table, as discussed in Chapter 9 on page 163.

Pearson utilizes the scoring tables to generate scale scores for each student. ETS verifies Pearson's scale scores by conducting QC and reasonableness checks, which are described in Chapter 9 on page 164.

Overview of Score Aggregation Procedures

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

Individual Scores

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

Score Distributions and Summary Statistics

Summary statistics that describes student performance on each CAPA are presented in Table 7.2 through Table 7.4. Included in these tables are the number of tasks in each test, the number of examinees taking each test, and the means and standard deviations of student scores expressed in terms of both raw scores and scale scores. In addition, summary statistics for the operational tasks on each test are provided.

Table 7.2 Summary Statistics Describing Student Scores: ELA

Level	I	II	III	IV	V
Scale Score Information					
Number of examinees	14,098	6,668	7,105	10,091	10,424
Mean score	40.76	38.82	39.56	39.02	38.72
SD *	11.04	6.91	6.46	8.45	6.04
Possible range	15–60	15–60	15–60	15–60	15–60
Obtained range	15–60	15–60	15–60	15–60	15–60
Median	41	39	40	40	39
Reliability	0.89	0.87	0.90	0.90	0.89
SEM †	3.70	2.45	2.01	2.71	2.02

Level	I	II	III	IV	V
Raw Score Information					
Mean score	24.92	19.00	20.13	18.71	20.45
SD *	11.60	6.55	6.96	7.26	6.80
Possible range	0–40	0–32	0–32	0–32	0–32
Obtained range	0–40	0–32	0–32	0–32	0–32
Median	27	19	21	19	21
Reliability	0.89	0.87	0.90	0.90	0.89
SEM †	3.89	2.32	2.17	2.33	2.27
Task Information					
Number of tasks	8	8	8	8	8
Mean AIS ‡	3.12	2.38	2.52	2.33	2.57
SD AIS ‡	0.29	0.54	0.28	0.33	0.40
Min. AIS	2.76	1.89	2.21	1.61	1.98
Max. AIS	3.61	3.57	3.01	2.66	3.07
Possible range	0–5	0–4	0–4	0–4	0–4
Mean polyserial	0.80	0.78	0.80	0.80	0.79
SD polyserial	0.02	0.07	0.08	0.04	0.06
Min. polyserial	0.77	0.68	0.68	0.72	0.67
Max. polyserial	0.83	0.85	0.88	0.84	0.86
Mean Rasch difficulty	–0.60	–0.76	–0.82	–0.87	–0.91
SD Rasch difficulty	0.14	0.79	0.39	0.43	0.50
Min. Rasch difficulty	–0.85	–2.40	–1.62	–1.37	–1.77
Max. Rasch difficulty	–0.46	0.03	–0.30	0.10	–0.27

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

Table 7.3 Summary Statistics Describing Student Scores: Mathematics

Level	I	II	III	IV	V
Scale Score Information					
Number of examinees	14,065	6,650	7,094	10,068	10,392
Mean score	36.15	37.28	36.34	37.14	37.49
SD *	9.00	8.50	5.54	7.50	8.08
Possible range	15–60	15–60	15–60	15–60	15–60
Obtained range	15–60	15–60	15–60	15–60	15–60
Median	37	38	36	38	37
Reliability	0.85	0.86	0.81	0.83	0.86
SEM †	3.46	3.18	2.41	3.07	3.03
Raw Score Information					
Mean score	22.85	19.62	19.04	19.92	21.03
SD *	10.89	6.93	6.14	6.50	7.36
Possible range	0–40	0–32	0–32	0–32	0–32
Obtained range	0–40	0–32	0–32	0–32	0–32
Median	24	20	19	21	22
Reliability	0.85	0.86	0.81	0.83	0.86
SEM †	4.19	2.60	2.67	2.66	2.76

Level	I	II	III	IV	V
Task Information					
Number of tasks	8	8	8	8	8
Mean AIS ‡	2.86	2.45	2.39	2.49	2.65
SD AIS ‡	0.28	0.62	0.44	0.55	0.42
Min. AIS	2.58	1.24	1.78	1.50	1.94
Max. AIS	3.25	3.20	2.96	2.97	3.35
Possible range	0–5	0–4	0–4	0–4	0–4
Mean polyserial	0.76	0.77	0.71	0.73	0.77
SD polyserial	0.03	0.09	0.11	0.11	0.05
Min. polyserial	0.72	0.62	0.55	0.61	0.71
Max. polyserial	0.81	0.86	0.83	0.85	0.84
Mean Rasch difficulty	–0.24	–0.96	–0.93	–0.81	–1.09
SD Rasch difficulty	0.14	0.74	0.36	0.58	0.36
Min. Rasch difficulty	–0.45	–1.92	–1.34	–1.41	–1.63
Max. Rasch difficulty	–0.05	0.50	–0.54	0.37	–0.41

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

Table 7.4 Summary Statistics Describing Student Scores: Science

Level	I	III	IV	V
Scale Score Information				
Number of examinees	3,564	3,556	3,299	3,424
Mean score	36.25	36.33	36.02	36.22
SD *	10.25	4.65	4.98	5.21
Possible range	15–60	15–60	15–60	15–60
Obtained range	15–60	15–60	15–60	15–60
Median	36	36	36	36
Reliability	0.88	0.84	0.81	0.85
SEM †	3.59	1.87	2.15	1.98
Raw Score Information				
Mean score	23.13	20.69	21.56	21.83
SD *	11.48	6.07	5.71	5.96
Possible range	0–40	0–32	0–32	0–32
Obtained range	0–40	0–32	0–32	0–32
Median	24	21	22	23
Reliability	0.88	0.84	0.81	0.85
SEM †	4.03	2.45	2.47	2.27
Task Information				
Number of tasks	8	8	8	8
Mean AIS ‡	2.91	2.60	2.69	2.74
SD AIS ‡	0.27	0.22	0.29	0.45
Min. AIS	2.37	2.24	2.17	2.15
Max. AIS	3.26	2.95	2.91	3.42
Possible range	0–5	0–4	0–4	0–4
Mean polyserial	0.79	0.72	0.70	0.76
SD polyserial	0.02	0.05	0.05	0.04

Level	I	III	IV	V
Min. polyserial	0.76	0.67	0.62	0.70
Max. polyserial	0.83	0.79	0.77	0.81
Mean Rasch difficulty	-0.31	-1.05	-1.11	-0.65
SD Rasch difficulty	0.16	0.30	0.34	0.53
Min. Rasch difficulty	-0.54	-1.49	-1.52	-1.41
Max. Rasch difficulty	0.00	-0.56	-0.45	0.02

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

The percentages of students in each performance level are presented in Table 7.5.

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

Table 7.5 Percentage of Examinees in Each Performance Level

Content Area	CAPA Level	Far Below Basic	Below Basic	Basic	Proficient	Advanced
English–Language Arts	I	6%	6%	8%	22%	59%
	II	2%	5%	13%	37%	43%
	III	2%	4%	12%	27%	54%
	IV	2%	9%	17%	32%	40%
	V	2%	3%	16%	36%	44%
Mathematics	I	6%	8%	19%	33%	34%
	II	3%	15%	17%	30%	35%
	III	1%	7%	22%	51%	20%
	IV	2%	10%	22%	39%	27%
	V	2%	9%	20%	36%	33%
Science	I	7%	10%	19%	30%	34%
	III	1%	3%	25%	53%	18%
	IV	1%	4%	28%	52%	14%
	V	1%	5%	24%	47%	23%

Table 7.A.1 through Table 7.A.3 in Appendix 7.A starting on page 57 show the distributions of scale scores for each CAPA. The results are reported in terms of three score intervals. A cell value of “N/A” indicates that there are no obtainable scale scores within that scale-score range for the particular CAPA.

Group Scores

Statistics summarizing student performance by content area for selected groups of students are provided in Table 7.B.1 through Table 7.B.3 for the CAPA.

In the tables, students are grouped by demographic characteristics including gender, ethnicity, English-language fluency, economic status, and primary disability. The tables show, for each demographic group, the numbers of valid cases and percentages of students in each performance level by demographic group.

Table 7.6 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 level are not reported and are presented as hyphens. Percentages in these tables may not sum up to 100 due to rounding.

Table 7.6 Subgroup Definitions

Subgroup	Definition
Gender	<ul style="list-style-type: none"> • Male • Female
Ethnicity	<ul style="list-style-type: none"> • African American • American Indian or Alaska Native • Asian <ul style="list-style-type: none"> – Asian Indian – Cambodian – Chinese – Hmong – Japanese – Korean – Laotian – Vietnamese – Other Asian • Hispanic or Latino • Pacific Islander <ul style="list-style-type: none"> – Guamanian – Native Hawaiian – Samoan – Tahitian – Other Pacific Islander • Filipino • 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 • Hard of hearing • Deafness • Speech/language impairment • Visual impairment • Emotional disturbance • Orthopedic impairment • Other health impairment • Specific learning impairment • Deaf blindness • Multiple group • Autism • Traumatic brain injury

Reports Produced and Scores for Each Report

The tests that make up the STAR Program provide results or score summaries that are reported for different purposes. The four major purposes include:

1. Communicating with parents and guardians;
2. Informing decisions needed to support student achievement;
3. Evaluating school programs; and
4. Providing data for state and federal accountability programs for schools and districts.

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

Types of Score Reports

There are three categories of CAPA reports. These categories and the specific reports in each category are given in the Table 7.7.

Table 7.7 Types of CAPA Reports

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

These reports are sent to the independently testing charter schools, counties, or school districts; the school district forwards the appropriate reports to test sites or, in the case of the STAR 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 STAR 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://star.cde.ca.gov/>.

Score Report Contents

The STAR Student Report provides scale scores and performance levels for each CAPA taken by the student. Scale scores are reported on a scale ranging from 15 to 60. The performance levels reported are: far below basic, below basic, basic, proficient, and advanced.

Further information about the STAR Student Report and the other reports is provided in Appendix 7.C on page 65.

Score Report Applications

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

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

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

The results from the CAPA are used for state and federal accountability programs to monitor each school's and district's progress toward achieving established goals. As mentioned previously, CAPA results are used to calculate each school's and district's Academic Performance Index (API). The API is a major component of California's Public School Accountability Act (PSAA) and is used to rank the academic performance of schools, compare schools with similar characteristics (for example, size and ethnic makeup), identify low-performing and high-priority schools, and set yearly targets for academic growth.

CAPA results also are used to comply with federal ESEA legislation that requires all schools to meet specific academic goals. The progress of each school toward achieving these goals is provided annually in an AYP report. Each year, California schools and districts must meet AYP goals by showing that a specified percentage of CAPA test-takers at the district and school level are performing at or above the proficient level on the CAPA for ELA and mathematics.

Criteria for Interpreting Test Scores

A school district may use CAPA 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 CAPA results (CDE, 2012b). It is also important to note that a student's score in a content area contains measurement error and could vary somewhat if the student were retested.

Criteria for Interpreting Score Reports

The information presented in various reports must be interpreted with caution when making performance comparisons. When comparing scale score and performance-level results for the CAPA, the user is limited to comparisons within the same content area and level. This is because the score scales are different for each content area and level. The user may compare scale scores for the same content area and level, within a school, between schools, or between a school and its district, its county, or the state. The user can also make comparisons within the same level and content area across years. Comparing scores obtained in different levels or content areas should be avoided because the results are not on the same scale. Comparisons between raw scores should be limited to comparisons within not only content area and level but also test year. Since new score scales and cut scores were applied beginning with the 2009 test results, results from this and subsequent

years cannot meaningfully be compared to results obtained in prior years. For more details on the criteria for interpreting information provided on the score reports, see the *2012 STAR Post-Test Guide* (CDE, 2012c).

References

- California Department of Education. (2012a). *2012 CAPA examiner's manual*. Sacramento, CA. Downloaded from http://www.startest.org/pdfs/CAPA.examiners_manual.nonsecure.2012.pdf
- California Department of Education. (2012b). *2012 STAR CST/CMA, CAPA, and STS printed reports*. Sacramento, CA. Downloaded from <http://www.startest.org/pdfs/STAR.reports.2012.pdf>
- California Department of Education. (2012c). *2012 STAR post-test guide*. Sacramento, CA. Downloaded from http://www.startest.org/pdfs/STAR.post-test_guide.2012.pdf
- Educational Testing Service. (2002). *ETS standards for quality and fairness*. Princeton, NJ: Author.
- Educational Testing Service. (2008) *A study to examine the effects of changes to the CAPA Level I rubric involving the hand-over-hand prompt*, Unpublished memorandum, Princeton, NJ: Author.

Appendix 7.A—Scale Score Distribution Tables

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

Table 7.A.1 Scale Score Frequency Distributions: ELA, Levels I–V

Scale Score	ELA I		ELA II		ELA III		ELA IV		ELA V	
	Freq	Pct.	Freq	Pct.	Freq	Pct.	Freq	Pct.	Freq	Pct.
60	1,554	11.02	53	0.79	71	1.00	131	1.30	173	1.66
57–59	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
54–56	473	3.36	83	1.24	N/A	N/A	208	2.06	N/A	N/A
51–53	352	2.50	142	2.13	149	2.10	235	2.33	N/A	N/A
48–50	276	1.96	192	2.88	178	2.51	692	6.86	247	2.37
45–47	1,551	11.00	771	11.56	973	13.69	1,334	13.22	777	7.45
42–44	1,876	13.31	960	14.40	1,305	18.37	1,415	14.02	1,556	14.93
39–41	2,843	20.17	1,294	19.41	1,788	25.17	1,669	16.54	2,902	27.84
36–38	1,940	13.76	1,483	22.24	1,068	15.03	1,220	12.09	2,241	21.50
33–35	975	6.92	926	13.89	874	12.30	1,178	11.67	1,364	13.09
30–32	570	4.04	292	4.38	255	3.59	887	8.79	672	6.45
27–29	405	2.87	182	2.73	212	2.98	358	3.55	197	1.89
24–26	154	1.09	99	1.48	84	1.18	310	3.07	130	1.25
21–23	126	0.89	81	1.21	38	0.53	81	0.80	43	0.41
18–20	N/A	N/A	31	0.46	22	0.31	137	1.36	28	0.27
15–17	1003	7.11	79	1.18	88	1.24	236	2.34	94	0.90

Table 7.A.2 Scale Score Frequency Distributions: Mathematics, Levels I–V

Scale Score	Math I		Math II		Math III		Math IV		Math V	
	Freq	Pct.	Freq	Pct.	Freq	Pct.	Freq	Pct.	Freq	Pct.
60	641	4.56	71	1.07	68	0.96	93	0.92	529	5.09
57–59	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
54–56	N/A	N/A	107	1.61	N/A	N/A	130	1.29	N/A	N/A
51–53	N/A	N/A	147	2.21	N/A	N/A	268	2.66	N/A	N/A
48–50	282	2.00	321	4.83	94	1.33	361	3.59	356	3.43
45–47	228	1.62	631	9.49	98	1.38	380	3.77	404	3.89
42–44	1,263	8.98	749	11.26	533	7.51	925	9.19	1,063	10.23
39–41	2,302	16.37	909	13.67	1,429	20.14	2,429	24.13	2,134	20.54
36–38	3,619	25.73	1,131	17.01	2,043	28.80	1,664	16.53	2,296	22.09
33–35	2,611	18.56	830	12.48	1,405	19.81	1,486	14.76	1,425	13.71
30–32	1,061	7.54	564	8.48	856	12.07	1,163	11.55	1,060	10.20
27–29	673	4.78	495	7.44	296	4.17	549	5.45	276	2.66
24–26	195	1.39	354	5.32	101	1.42	193	1.92	554	5.33
21–23	146	1.04	103	1.55	56	0.79	157	1.56	N/A	N/A
18–20	N/A	N/A	53	0.80	18	0.25	54	0.54	59	0.57
15–17	1,044	7.42	185	2.78	97	1.37	216	2.15	236	2.27

Table 7.A.3 Scale Score Frequency Distributions: Science, Levels I–V

Scale Score	Science I		Science III		Science IV		Science V	
	Freq	Pct.	Freq	Pct.	Freq	Pct.	Freq	Pct.
60	272	7.63	28	0.79	48	1.45	58	1.69
57–59	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
54–56	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
51–53	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
48–50	65	1.82	N/A	N/A	N/A	N/A	N/A	N/A
45–47	79	2.22	64	1.80	65	1.97	74	2.16
42–44	339	9.51	221	6.21	98	2.97	104	3.04
39–41	452	12.68	535	15.04	445	13.49	547	15.98
36–38	828	23.23	1,280	36.00	1,113	33.74	1,186	34.64
33–35	656	18.41	1,011	28.43	1,073	32.53	878	25.64
30–32	262	7.35	265	7.45	292	8.85	369	10.78
27–29	83	2.33	68	1.91	83	2.52	139	4.06
24–26	193	5.42	37	1.04	39	1.18	19	0.55
21–23	N/A	N/A	17	0.48	10	0.30	19	0.55
18–20	41	1.15	3	0.08	11	0.33	8	0.23
15–17	294	8.25	27	0.76	22	0.67	23	0.67

Appendix 7.B—Demographic Summaries

Table 7.B.1 Demographic Summary for ELA, All Examinees

	Number Tested	Percentage in Performance Level				
		Far Below Basic	Below Basic	Basic	Proficient	Advanced
All valid scores	48,386	3%	6%	13%	30%	49%
Male	31,522	3%	6%	13%	30%	49%
Female	16,583	3%	6%	13%	29%	49%
Gender unknown	281	5%	4%	15%	27%	49%
American Indian	379	2%	5%	9%	27%	56%
Asian American	3,294	4%	8%	16%	35%	38%
Pacific Islander	268	4%	9%	12%	31%	44%
Filipino	1,481	4%	8%	17%	31%	40%
Hispanic	25,211	3%	5%	13%	30%	50%
African American	4,739	3%	5%	11%	29%	52%
White	11,501	3%	6%	13%	29%	49%
Ethnicity unknown	1,513	3%	7%	13%	30%	47%
English Only	29,372	3%	6%	13%	30%	48%
Initially—Fluent English Proficient	805	4%	8%	16%	31%	42%
English Learner	1,688	2%	5%	12%	32%	49%
Reclassified—Fluent English Proficient	15,870	3%	5%	13%	30%	49%
English Proficient unknown	651	3%	4%	11%	29%	53%
Mental Retardation/Intellectual Disability	19,401	2%	5%	14%	32%	47%
Hard of Hearing	274	2%	4%	12%	34%	48%
Deafness	316	1%	4%	16%	44%	35%
Speech/Language Impairment	1,681	0%	1%	4%	25%	71%
Visual Impairment	478	11%	7%	12%	26%	45%
Emotional Disturbance	346	1%	1%	4%	22%	72%
Orthopedic Impairment	3,731	8%	8%	12%	29%	44%
Other Health Impairment	2,046	2%	3%	8%	28%	60%
Specific Learning Impairment	2,907	0%	0%	2%	17%	81%
Deaf Blindness	42	14%	17%	21%	31%	17%
Multiple Group	2,387	10%	9%	14%	30%	37%
Autism	13,925	3%	7%	15%	31%	43%
Traumatic Brain Injury	290	4%	3%	7%	29%	56%
Unknown	562	4%	6%	9%	30%	51%
Not Econ. Disadvantaged	31,347	3%	5%	12%	29%	51%
Economically Disadvantaged	16,002	4%	7%	14%	31%	44%
Unknown Economic Status	1,037	3%	4%	10%	28%	54%
Primary Ethnicity—Not Economically Disadvantaged						
American Indian	117	4%	4%	14%	28%	50%
Asian American	1,786	4%	7%	17%	35%	38%
Pacific Islander	98	4%	11%	9%	38%	38%
Filipino	900	5%	8%	17%	31%	38%
Hispanic	4,483	5%	7%	13%	29%	45%
African American	1,299	4%	6%	13%	31%	46%
White	6,692	4%	7%	14%	31%	45%
Ethnicity unknown	627	4%	8%	13%	30%	44%

	Number Tested	Percentage in Performance Level				
		Far Below Basic	Below Basic	Basic	Proficient	Advanced
Primary Ethnicity—Economically Disadvantaged						
American Indian	258	1%	5%	7%	26%	60%
Asian American	1,437	3%	8%	15%	35%	39%
Pacific Islander	164	5%	8%	13%	28%	46%
Filipino	545	3%	8%	17%	32%	42%
Hispanic	20,332	3%	5%	12%	30%	51%
African American	3,349	2%	4%	10%	29%	54%
White	4,563	2%	5%	11%	27%	56%
Ethnicity unknown	699	2%	6%	14%	30%	47%
Primary Ethnicity—Unknown Economic Status						
American Indian	4	—	—	—	—	—
Asian American	71	3%	7%	21%	28%	41%
Pacific Islander	6	—	—	—	—	—
Filipino	36	11%	8%	11%	25%	44%
Hispanic	396	3%	3%	10%	29%	55%
African American	91	1%	4%	5%	24%	65%
White	246	2%	4%	11%	28%	54%
Ethnicity unknown	187	5%	3%	8%	30%	54%

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

Table 7.B.2 Demographic Summary for Mathematics, All Examinees

	Number Tested	Percentage in Performance Level				
		Far Below Basic	Below Basic	Basic	Proficient	Advanced
All valid scores	48,269	3%	9%	20%	37%	30%
Male	31,446	3%	9%	19%	37%	32%
Female	16,547	4%	10%	21%	37%	28%
Gender unknown	276	3%	10%	20%	36%	32%
American Indian	379	3%	7%	20%	38%	33%
Asian American	3,281	4%	11%	20%	38%	26%
Pacific Islander	265	3%	11%	23%	32%	30%
Filipino	1,476	4%	11%	22%	36%	28%
Hispanic	25,163	3%	9%	20%	37%	31%
African American	4,722	3%	9%	19%	38%	31%
White	11,475	3%	10%	21%	37%	29%
Ethnicity unknown	1,508	3%	9%	22%	38%	28%
English Only	29,293	3%	10%	21%	37%	29%
Initially–Fluent English Proficient	805	4%	12%	23%	32%	28%
English Learner	1,687	2%	8%	20%	37%	33%
Reclassified–Fluent English Proficient	15,839	3%	9%	18%	37%	32%
English Proficient unknown	645	3%	8%	17%	38%	34%
Mental Retardation/Intellectual Disability	19,374	2%	11%	23%	37%	26%
Hard of Hearing	273	1%	6%	17%	41%	34%
Deafness	314	1%	5%	10%	40%	43%
Speech/Language Impairment	1,683	0%	3%	9%	39%	49%
Visual Impairment	476	11%	14%	22%	30%	24%
Emotional Disturbance	342	2%	2%	10%	32%	54%
Orthopedic Impairment	3,719	9%	12%	25%	33%	21%
Other Health Impairment	2,041	2%	7%	17%	39%	36%
Specific Learning Impairment	2,899	0%	1%	4%	37%	58%
Deaf Blindness	42	17%	31%	17%	21%	14%
Multiple Group	2,381	10%	15%	25%	31%	19%
Autism	13,879	3%	9%	20%	38%	30%
Traumatic Brain Injury	290	5%	6%	16%	36%	38%
Unknown	556	3%	10%	16%	40%	31%
Not Econ. Disadvantaged	31,291	3%	9%	19%	37%	32%
Economically Disadvantaged	15,952	4%	11%	22%	37%	26%
Unknown Economic Status	1,026	3%	7%	18%	38%	35%
Primary Ethnicity—Not Economically Disadvantaged						
American Indian	117	3%	8%	21%	36%	32%
Asian American	1,780	4%	12%	19%	40%	26%
Pacific Islander	97	4%	11%	26%	31%	28%
Filipino	896	5%	10%	22%	36%	26%
Hispanic	4,473	6%	11%	21%	35%	26%
African American	1,292	4%	10%	21%	37%	28%
White	6,671	4%	11%	22%	38%	26%
Ethnicity unknown	626	4%	10%	23%	38%	25%

	Number Tested	Percentage in Performance Level				
		Far Below Basic	Below Basic	Basic	Proficient	Advanced
Primary Ethnicity—Economically Disadvantaged						
American Indian	258	2%	6%	19%	38%	34%
Asian American	1,430	3%	11%	22%	37%	27%
Pacific Islander	162	3%	12%	22%	33%	31%
Filipino	544	2%	12%	23%	36%	28%
Hispanic	20,300	3%	9%	19%	37%	32%
African American	3,339	3%	8%	19%	38%	32%
White	4,560	3%	8%	19%	36%	34%
Ethnicity unknown	698	2%	9%	21%	38%	29%
Primary Ethnicity—Unknown Economic Status						
American Indian	4	—	—	—	—	—
Asian American	71	4%	7%	21%	38%	30%
Pacific Islander	6	—	—	—	—	—
Filipino	36	8%	3%	14%	22%	53%
Hispanic	390	3%	7%	16%	41%	33%
African American	91	2%	10%	18%	35%	35%
White	244	2%	5%	21%	37%	35%
Ethnicity unknown	184	3%	6%	16%	38%	37%

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

Table 7.B.3 Demographic Summary for Science, All Examinees

	Number Tested	Percentage in Performance Level				
		Far Below Basic	Below Basic	Basic	Proficient	Advanced
All valid scores	13,843	3%	6%	24%	45%	22%
Male	8,954	2%	6%	24%	45%	24%
Female	4,835	3%	6%	24%	47%	20%
Gender Unknown	54	2%	4%	39%	37%	19%
American Indian	126	1%	6%	14%	51%	28%
Asian American	892	3%	9%	28%	43%	17%
Pacific Islander	80	4%	10%	26%	43%	18%
Filipino	450	3%	10%	27%	42%	18%
Hispanic	7,107	3%	5%	23%	46%	23%
African American	1,419	2%	5%	23%	48%	22%
White	3,361	2%	5%	25%	43%	24%
Ethnicity unknown	408	3%	4%	27%	48%	18%
English Only	8,408	2%	6%	25%	45%	22%
Initially–Fluent English Proficient	235	3%	7%	33%	38%	19%
English Learner	530	2%	5%	22%	46%	25%
Reclassified–Fluent English Proficient	4,512	3%	5%	22%	46%	23%
English Proficient unknown	158	3%	3%	25%	44%	25%
Mental Retardation/Intellectual Disability	5,954	1%	5%	25%	47%	21%
Hard of Hearing	91	3%	2%	25%	48%	21%
Deafness	95	0%	1%	18%	60%	21%
Speech/Language Impairment	351	0%	1%	9%	53%	37%
Visual Impairment	141	9%	9%	26%	36%	21%
Emotional Disturbance	114	4%	1%	10%	51%	35%
Orthopedic Impairment	1,086	8%	10%	26%	39%	17%
Other Health Impairment	555	2%	3%	17%	53%	26%
Specific Learning Impairment	886	0%	0%	6%	51%	43%
Deaf Blindness	6	–	–	–	–	–
Multiple Group	673	9%	12%	27%	37%	15%
Autism	3,670	2%	7%	29%	42%	19%
Traumatic Brain Injury	90	1%	4%	14%	47%	33%
Unknown	131	2%	5%	27%	44%	22%
Not Econ. Disadvantaged	8,960	2%	5%	22%	47%	24%
Economically Disadvantaged	4,627	4%	8%	27%	42%	19%
Unknown Economic Status	256	2%	4%	23%	46%	25%
Primary Ethnicity—Not Economically Disadvantaged						
American Indian	47	2%	15%	19%	40%	23%
Asian American	495	4%	10%	30%	42%	14%
Pacific Islander	26	0%	15%	23%	50%	12%
Filipino	277	3%	12%	27%	39%	19%
Hispanic	1,257	6%	8%	25%	40%	20%
African American	403	3%	7%	26%	47%	17%
White	1,960	3%	6%	29%	41%	21%
Ethnicity unknown	162	3%	4%	31%	48%	14%

	Number Tested	Percentage in Performance Level				
		Far Below Basic	Below Basic	Basic	Proficient	Advanced
Primary Ethnicity—Economically Disadvantaged						
American Indian	78	0%	1%	10%	58%	31%
Asian American	381	3%	6%	26%	44%	20%
Pacific Islander	52	6%	8%	27%	38%	21%
Filipino	163	4%	7%	27%	45%	17%
Hispanic	5,755	2%	5%	22%	48%	24%
African American	992	2%	4%	22%	49%	24%
White	1,338	2%	4%	21%	45%	28%
Ethnicity unknown	201	2%	4%	24%	51%	18%
Primary Ethnicity—Unknown Economic Status						
American Indian	1	—	—	—	—	—
Asian American	16	0%	6%	38%	25%	31%
Pacific Islander	2	—	—	—	—	—
Filipino	10	—	—	—	—	—
Hispanic	95	0%	6%	23%	52%	19%
African American	24	4%	0%	17%	63%	17%
White	63	5%	3%	17%	41%	33%
Ethnicity unknown	45	4%	0%	27%	38%	31%

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

Appendix 7.C—Type of Score Report

Table 7.C.1 Score Reports Reflecting CAPA Results

2012 STAR CAPA Student Reports	
Description	Distribution
The CAPA 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) 	<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 school district 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 STAR 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. It includes the following data for each CAPA content area tested:</p> <ul style="list-style-type: none"> • Scale scores • Performance levels 	<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>

2012 STAR CAPA Student Reports	
Description	Distribution
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 CAPA grade and level, the following data are summarized by content area tested:</p> <ul style="list-style-type: none"> • Number of students enrolled • Number and percent of students tested • Number and percent of valid scores • Number tested with scores • Mean scale score • Scale score standard deviation • 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 school district.</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 is recommended that summary reports be retained for at least five years.</p>
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 (Disabilities among CAPA students include specific disabilities.) • 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. CAPA statistics are listed by CAPA level.</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 tested in the subgroup as a percent of all students tested • 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 school district.</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 for at least five years.</p>

2012 STAR CAPA Student Reports	
Description	Distribution
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 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. CAPA statistics are listed by CAPA level.</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 tested in the subgroup as a percent of all students tested • 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. One copy is packaged for the school and one for the school district.</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 for at least five years.</p>

Chapter 8: Analyses

This chapter summarizes the task (item)- and test-level statistics obtained for the CAPA administered during the spring of 2012.

The statistics presented in this chapter are divided into five 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
5. Differential Item Functioning (DIF) Analyses

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

1. Appendix 8.A on page 88 presents the classical item analyses including average item score (AIS) and polyserial correlation coefficient, and associated flags, for the operational and field-test tasks of each test. Also presented in this appendix is information about the distribution of scores for the operational tasks. In addition, the mean, minimum, and maximum of AIS and polyserial correlation for each operational task are presented in Table 8.2 on page 70.
2. Appendix 8.B on page 105 presents results of the reliability analyses of total test scores for the population as a whole and for selected subgroups. Also presented are results of the analyses of the accuracy and consistency of the performance classifications.
3. Appendix 8.C on page 119 presents tables showing the correlations between scores obtained on the CAPA measured in the different content areas, which are provided as an example of the evidence of the validity of the interpretation and uses of CAPA scores. The results for the overall test population are presented in Table 8.4; the tables in Appendix 8.C summarize the results for various subgroups. Also included in Appendix 8.C are results of the rater agreement for each operational task.
4. Appendix 8.D on page 136 presents the results of IRT analyses including the distribution of tasks based on their fit to the Rasch model and the summaries of Rasch item difficulty statistics (*b*-values) for the operational and field-test tasks. In addition, the appendix presents the scoring tables obtained as a result of the IRT equating process. Information related to the evaluation of linking tasks is presented in Table 8.5 on page 82; these linking tasks were used in the equating process discussed later in this chapter.
5. Appendix 8.E on page 152 presents the results of the DIF analyses applied to all operational and field-test tasks for which sufficient student samples were available. In this appendix, tasks flagged for significant DIF are listed. Also given are the distributions of items across DIF categories.

Samples Used for the Analyses

CAPA analyses were conducted at different times after test administration and involved varying proportions of the full CAPA data.

IRT results for the operational items are based on the equating sample that includes all valid cases available in early June 2012. All other analyses for this technical report are based on

all valid cases in the P1 data, which contained test results for 100 percent of the entire test-taking population. Summary statistics describing the samples are presented in Table 8.1; the samples used to generate scoring tables are labeled as “Equating Samples.”

Table 8.1 CAPA Raw Score Means and Standard Deviations: Total P1 Population and Equating Sample

Content Area	Level	P1			Equating Sample			
		N	Mean	SD	N	% of P1	Mean	SD
English–Language Arts	I	14,098	24.92	11.60	8,305	59%	25.03	11.24
	II	6,668	19.00	6.55	4,239	64%	19.47	6.58
	III	7,105	20.13	6.96	4,391	62%	20.26	6.97
	IV	10,091	18.71	7.26	6,430	64%	18.93	7.16
	V	10,424	20.45	6.80	6,895	66%	20.60	6.67
Mathematics	I	14,065	22.85	10.89	8,289	59%	22.97	10.64
	II	6,650	19.62	6.93	4,235	64%	19.87	6.93
	III	7,094	19.04	6.14	4,384	62%	19.16	6.18
	IV	10,068	19.92	6.50	6,419	64%	20.12	6.43
	V	10,392	21.03	7.36	6,872	66%	21.25	7.24
Science	I	3,564	23.13	11.48	2,120	59%	23.39	11.25
	III	3,556	20.69	6.07	2,188	62%	20.84	6.13
	IV	3,299	21.56	5.71	2,124	64%	21.74	5.67
	V	3,424	21.83	5.96	2,279	67%	22.09	5.89

Classical Analyses

Average Item Score

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

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

Polyserial Correlation of the Task Score with the Total Test Score

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

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

$$r_{polyreg} = \frac{\hat{\beta}s_{tot}}{\sqrt{\hat{\beta}^2s_{tot}^2 + 1}} \quad (8.1)$$

where,

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

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

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

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

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

A descriptive summary of the classical item statistics for the overall test are presented in Table 8.2. The task-by-task values are presented in Table 8.A.1 through Table 8.A.14. Some tasks were flagged for unusual statistics; these flags are shown in the tables. Although the flag definition appears in the heading of each table, the flags are displayed in the body of the tables only where applicable for the specific CAPA presented. The flag classifications are as follows:

- Difficulty flags
 - A: Low average task score (below 1.5 at Level I; below 1.2 at Levels II–V)
 - H: High average task score (above 4.0 at Level I; above 3.2 at Levels II–V)
- Discrimination flag
 - R: Polyserial correlation less than .60
- Omit/nonresponse/flag
 - O: Omit/nonresponse rates greater than 5 percent

Table 8.2 Average Item Score and Polyserial Correlation

Content Area	Level	No. of items	No. of Examinees	Mean		Minimum		Maximum	
				AIS	Polyserial	AIS	Polyserial	AIS	Polyserial
English– Language Arts	I	8	14,098	3.12	0.80	2.76	0.77	3.61	0.83
	II	8	6,668	2.38	0.78	1.89	0.68	3.57	0.85
	III	8	7,105	2.52	0.80	2.21	0.68	3.01	0.88
	IV	8	10,091	2.33	0.80	1.61	0.72	2.66	0.84
	V	8	10,424	2.57	0.79	1.98	0.67	3.07	0.86
Mathematics	I	8	14,065	2.86	0.76	2.58	0.72	3.25	0.81
	II	8	6,650	2.45	0.77	1.24	0.62	3.20	0.86
	III	8	7,094	2.39	0.71	1.78	0.55	2.96	0.83
	IV	8	10,068	2.49	0.73	1.50	0.61	2.97	0.85
	V	8	10,392	2.65	0.77	1.94	0.71	3.35	0.84

Content Area	Level	No. of items	No. of Examinees	Mean		Minimum		Maximum	
				AIS	Polyserial	AIS	Polyserial	AIS	Polyserial
Science	I	8	3,564	2.91	0.79	2.37	0.76	3.26	0.83
	III	8	3,556	2.60	0.72	2.24	0.67	2.95	0.79
	IV	8	3,299	2.69	0.70	2.17	0.62	2.91	0.77
	V	8	3,424	2.74	0.76	2.15	0.70	3.42	0.81

As noted previously, the score distributions for individual operational tasks comprising each CAPA test are provided by content area and level in Table 8.A.15 through Table 8.A.17.

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 8.2:

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

where,

n is the number of tasks,

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

s_t^2 is the variance of the total score .

The standard error of measurement (SEM) provides a measure of score instability in the score metric. The SEM is defined by:

$$s_e = s_t \sqrt{1 - \alpha} \quad (8.3)$$

where,

α is the reliability estimated using equation 8.2, and

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

The SEM is particularly useful in determining the confidence interval (CI) that captures an examinee's true score. Assuming that measurement error is normally distributed, it can be said that upon infinite replications of the testing occasion, approximately 95 percent of the CIs of ± 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 8.3 gives the reliability and SEM for the CAPA, along with the number of items and examinees upon which those analyses were performed.

Table 8.3 Reliabilities and SEMs for the CAPA

Content Area	Level	No. of Items	No. of Examinees	Reliab.	Scale Score			Raw Score		
					Mean	S.D.	SEM	Mean	S.D.	SEM
English– Language Arts	I	8	14,098	0.89	40.76	11.04	3.70	24.92	11.6	3.89
	II	8	6,668	0.87	38.82	6.91	2.45	19.00	6.55	2.32
	III	8	7,105	0.90	39.56	6.46	2.01	20.13	6.96	2.17
	IV	8	10,091	0.90	39.02	8.45	2.71	18.71	7.26	2.33
	V	8	10,424	0.89	38.72	6.04	2.02	20.45	6.80	2.27
Mathematics	I	8	14,065	0.85	36.15	9.00	3.46	22.85	10.89	4.19
	II	8	6,650	0.86	37.28	8.50	3.18	19.62	6.93	2.60
	III	8	7,094	0.81	36.34	5.54	2.41	19.04	6.14	2.67
	IV	8	10,068	0.83	37.14	7.50	3.07	19.92	6.50	2.66
	V	8	10,392	0.86	37.49	8.08	3.03	21.03	7.36	2.76
Science	I	8	3,564	0.88	36.25	10.25	3.59	23.13	11.48	4.03
	III	8	3,556	0.84	36.33	4.65	1.87	20.69	6.07	2.45
	IV	8	3,299	0.81	36.02	4.98	2.15	21.56	5.71	2.47
	V	8	3,424	0.85	36.22	5.21	1.98	21.83	5.96	2.27

Subgroup Reliabilities and SEMs

The reliabilities of the CAPA were examined for various subgroups of the examinee population. The subgroups included in these analyses were defined by their gender, ethnicity, economic status, disability group, and English-language fluency. The reliability analyses are also presented by primary ethnicity within economic status.

Table 8.B.1 through Table 8.B.6 present the reliabilities and SEM information for the total test scores for each subgroup. Note that the reliabilities are reported only for samples that are comprised of 11 or more examinees. Also, in some cases, score reliabilities were not estimable and are presented in the tables as hyphens. Finally, results based on samples that contain 50 or fewer examinees should be interpreted with caution due to small sample sizes.

Conditional Standard Errors of Measurement

As part of the IRT-based equating procedures, scale-score conversion tables and conditional standard errors of measurement (CSEMs) are produced. CSEMs for CAPA scale scores are based on IRT and are calculated by the IRTEQUATE module in 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:

$$\text{CSEM}(\hat{\theta}) = \frac{1}{\sqrt{I(\hat{\theta})}} a \quad (8.4)$$

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 level and content area.

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

Table 8.D.10 through Table 8.D.23 in Appendix 8.D present the scale score CSEMs at the score required for a student to be classified in the below basic, basic, proficient, and advanced performance levels for the CAPA. The pattern of lower values of CSEMs at the basic and proficient levels are expected since (1) more items tend to be of middle difficulty; and (2) items at the extremes still provide information toward the middle of the scale. This results in more precise scores in the middle of the scale and less precise scores at the extremes of the scale.

Decision Classification Analyses

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

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

Decision consistency describes the extent to which examinees are classified in the same way as they would be on the basis of a single form of a test other than the one for which data are available. Decision consistency answers the following question: What is the agreement between the classifications based on two nonoverlapping, equally difficult, forms of the test? RELCLASS-COMP also estimates decision consistency using an estimated multivariate distribution of reported classifications on the current form of the 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.7 through Table 8.B.20 in Appendix 8.B starting on page 111.

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

The classifications are collapsed to below-proficient versus proficient and above, which are the critical categories for Adequate Yearly Progress (AYP) calculations, and are also presented in the tables.

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 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 CAPA 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 for the peer review of testing programs administered by states in response to the Elementary and Secondary Education Act (USDOE, 2001).

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

Purpose of the CAPA

As mentioned in Chapter 1, the CAPA is used in calculating school and district API. Additionally, the CAPA results for ELA and mathematics in grades two through eight and grade ten are used in determining AYP that applies toward meeting the requirement of the Elementary and Secondary Education Act (ESEA), which is to have all students score at proficient or above by 2014.

The Constructs to Be Measured

The CAPA is designed to show how well students with an IEP and who have significant cognitive disabilities perform relative to the California content standards. These content standards were approved by the SBE; they describe what students should know and be able to do at each level.

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

Interpretations and Uses of the Scores Generated

Total test scores expressed as scale scores, and student performance levels are generated for each grade-level test. The total test scale score is used to draw inferences about a

student's achievement in the content area and to classify the achievement into one of five performance levels: advanced, proficient, basic, below basic, and far below basic.

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

1. Communicating with parents and guardians;
2. Informing decisions needed to support student achievement;
3. Evaluating school programs; and
4. Providing data for state and federal accountability programs for schools.

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)

Students with an IEP and who have significant cognitive disabilities in grades two through eleven take the CAPA when they are unable to take the CSTs with or without accommodations or modifications or the CMA with accommodations. Participation in the CAPA and eligibility are determined by a student's IEP team. Only those students whose parents/guardians have submitted written requests to exempt them from STAR Program testing do not take the tests.

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 *Test 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 task specifications for the CAPA. The task specifications describe the characteristics of the tasks that should be written to measure each content standard. A thorough description of the specifications can be found in Chapter 3, starting on page 17. Once the tasks are developed and field-tested, ETS selects all CAPA test tasks to conform to the SBE-approved California content standards and test blueprints. Test blueprints for the CAPA were proposed by ETS and reviewed and approved by the Assessment Review Panels (ARPs), which are advisory panels to the CDE and ETS on areas related to task development for the CAPA. Test blueprints were also reviewed and approved by the CDE and presented to the SBE for adoption. There have been no recent changes in the blueprints for the CAPA; the blueprints were most recently revised and adopted by the SBE in 2006 for implementation beginning in 2008. The test blueprints for the CAPA can be found on the CDE STAR CAPA Blueprints Web page at <http://www.cde.ca.gov/ta/tg/sr/capablueprints.asp>.

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

Task review process—Chapter 3 explains in detail the extensive item review process applied to tasks written for use in the CAPA. In brief, tasks written for the CAPA undergo multiple review cycles and involve multiple groups of reviewers. One of the reviews is carried out by an external reviewer, that is, the ARPs. The ARPs are responsible for reviewing all newly developed tasks for alignment to the California content standards.

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

Guidelines for test construction were established with the goal of maintaining parallel forms to the greatest extent possible from year to year. Details can be found in Chapter 4, starting on page 27.

Additionally, an external review panel, the Statewide Pupil Assessment Review (SPAR), is 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 23.

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

Human Resource Research Organization (HumRRo) performed an alignment study for the CAPA in April 2007. The result of this study was a report entitled *Independent*

Evaluation of the Alignment of the California Standards Tests (CSTs) and the California Alternate Performance Assessment (CAPA).

HumRRO utilized the Webb alignment method to evaluate the alignment of the performance tasks field-tested in the 2007 CAPA to the California content standards. The Webb method requires a set of raters to evaluate each test item on two different dimensions: (1) the standard(s) targeted by items, and (2) the depth of knowledge required of students to respond to items. These ratings form the basis of the four separate Webb alignment analyses: categorical concurrence, depth-of-knowledge consistency, range-of-knowledge correspondence, and balance-of-knowledge representation. The results indicated that the performance tasks assess the majority of CAPA standards well across levels for both ELA and mathematics. A copy of the study is available as a CDE Web document at <http://www.cde.ca.gov/ta/tg/sr/documents/alignmentreport.pdf>.

Evidence Based on Relations to Other Variables

Empirical results concerning the relationships between the scores on a test and measures of other variables external to the test can also provide evidence of validity when these relationships are found to be consistent with the definition of the construct that the test is intended to measure. As indicated in the *Test 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 CAPA, DIF analyses were performed on all operational tasks and field-test tasks for which sufficient student samples were available.

The results of the DIF analyses are presented in Appendix 8.E, which starts on page 152. The vast majority of the tasks exhibited little or no significant DIF, suggesting that, in general, scores based on the CAPA tasks would have the same meaning for individuals who differed in their demographic characteristics.

Correlations Between Content-area Test Scores

To the degree that students' content-area test scores correlate as expected, evidence of the validity in regarding those scores as measures of the intended constructs is provided. Table 8.4 on the next page provides the correlations between scores on the 2012 CAPA content-area tests, and the numbers of students on which these correlations were based. Sample sizes for individual tests are shown in bold font on the diagonals of the correlation matrices, and the numbers of students on which the correlations were based are shown on the lower off-diagonals. The correlations are provided in the upper off-diagonals.

At Level I, the correlations between students' ELA, mathematics, and science scores were high. For Levels II and above, the correlations between content-area scores tended to be more moderate.

Table 8.C.1 through Table 8.C.35 in Appendix 8.C provide the content-area test score correlations by gender, ethnicity, English-language fluency, economic status, and disability. Similar patterns of correlations between students' ELA, mathematics, and science scores were found within the subgroups.

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

Table 8.4 CAPA Content-area Correlations for CAPA Levels

Level	Content	ELA	Mathematics	Science
I	ELA	14,098	0.80	0.81
	Mathematics	14,059	14,065	0.80
	Science	3,562	3,564	3,564
II	ELA	6,668	0.73	N/A
	Mathematics	6,648	6,650	N/A
	Science	N/A	N/A	N/A
III	ELA	7,105	0.77	0.74
	Mathematics	7,088	7,094	0.74
	Science	3,555	3,554	3,556
IV	ELA	10,091	0.75	0.68
	Mathematics	10,056	10,068	0.66
	Science	3,294	3,297	3,299
V	ELA	10,424	0.70	0.67
	Mathematics	10,377	10,392	0.66
	Science	3,420	3,422	3,424

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. Finally, evidence may also be derived from feedback provided by observers or judges involved in the scoring of examinee responses.

Evidence of Interrater Agreement

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

Interrater Agreement

As noted previously, approximately 10 percent of the test population's responses to the tasks were scored by two raters. Across all CAPA levels for ELA, mathematics and science, the percentage of students for whom the raters were in exact agreement ranged from 91 percent to 98 percent.

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 (task) scores and the relationship between these scores and scores on components of the test. To the extent that the score properties and relationships found are consistent with the definition of the construct measured by test, support is gained for interpreting these scores as measures of the construct.

For the CAPA, it is assumed that a single construct underlies the total scores obtained on each test. Evidence to support this assumption can be gathered from the results of task analyses, evaluations of internal consistency, and studies of model-data fit and reliability.

Reliability

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

Overall reliability—The reliability analyses are presented in Table 8.3. The results indicate that the reliabilities for all CAPA levels for ELA, mathematics, and science tended to be high, ranging from 0.81 to 0.90.

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

Evidence Based on Consequences of Testing

As observed in the *Standards*, tests are usually administered “with the expectation that some benefit will be realized from the intended use of the scores” (AERA, APA, & NCME, 1999, p. 18). When this is the case, evidence that the expected benefits accrue will provide support for intended use of the scores. The CDE and ETS are in the process of determining what kinds of information can be gathered to assess the consequences of the administration of the CAPA.

IRT Analyses

The IRT model used to calibrate the CAPA test tasks is the one-parameter partial credit (1PPC) model, a more restrictive version of the generalized partial-credit model (Muraki, 1992), in which all tasks are assumed to be equally discriminating. This model states that the probability that an examinee with ability θ will perform in the k th category of m_j ordered score categories of task j can be expressed as:

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

where,

m_j is the number of possible score categories ($c=1 \dots m_j$) for task j ,

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

b_j is the difficulty of task j , and

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

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

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

The parameters estimated for each task were evaluated for model-data fit, as described below.

IRT Model-Data Fit Analyses

ETS psychometricians classify operational and field-test tasks for the CAPA into discrete categories based on an evaluation of how well each task was fit by the Rasch partial credit model. The flagging procedure has categories of A, B, C, D, and F that are assigned based on an evaluation of graphical model-data fit information. Descriptors for each category are provided below.

Flag A

- Good fit of theoretical curve to empirical data along the entire ability range, may have some small divergence at the extremes
- Small Chi-square value relative to the other items in the calibration with similar sample sizes

Flag B

- Theoretical curve within error range across most of ability range, may have some small divergence at the extremes
- Acceptable Chi-square value relative to the other items in the calibration with similar sample sizes

Flag C

- Theoretical curve within error range at some regions and slightly outside of error range at remaining regions of ability range
- Moderate Chi-square value relative to the other items in the calibration with similar sample sizes
- This category often applies to items that appear to be functioning well, but that are not well fit by the Rasch model

Flag D

- Theoretical curve outside of error range at some regions across ability range
- Large Chi-square value relative to the other items in the calibration with similar sample sizes

Flag F

- Theoretical curve outside of error range at most regions across ability range
- Probability of answering item correctly may be higher at lower ability than higher ability (U-shaped empirical curve)
- Very large Chi-square value relative to the other items with similar sample sizes and classical item statistics tend also to be very poor

In general, items with flagging categories of A, B, or C are all considered acceptable. Ratings of D are considered questionable, and the ratings of F indicate a poor model fit.

Model-fit Assessment Results

The model-fit assessment is performed twice in the administration cycle. The assessment is first performed before scoring tables are produced and released. The assessment is performed again as part of the final item analyses when much larger samples are available. The flags produced as a result of this assessment are placed in the item bank. The test developers are asked to avoid the items flagged as D, if possible, and to carefully review them if they must be used. Test developers are instructed to avoid using items rated F for operational test assembly without a review by a psychometrician and by CDE content specialists.

The number of the operational and field-test tasks in each IRT model-data fit classification is presented in Table 8.D.1 through Table 8.D.6, which start on page 136.

Evaluation of Scaling

Calibrations of the 2012 forms were scaled to the previously obtained reference scale estimates in the item bank using the Stocking and Lord (1983) procedure. Details on the scaling procedures are provided on page 14 of Chapter 2.

The linking process is 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) on the basis of the old and new difficulty estimates are removed from the linking set. Based on established procedures, any linking items for which the WRMSD was greater than 0.625 for Level I and 0.500 for Levels II through V were eliminated. This criterion has produced reasonable results over time in similar equating work done with other testing programs at ETS. For the 2012 CAPA tests, no linking tasks were eliminated.

Table 8.5 presents, for each CAPA, the number of linking tasks between the 2012 (new) form and the test form to which it was linked (2011); the number of tasks removed from the linking task sets; the correlation between the final set of new and reference difficulty estimates for the linking tasks; and the average WRMSD statistic across the final set of linking tasks.

Table 8.5 Evaluation of Common Items Between New and Reference Test Forms

Content Area	Level	No. Linking Tasks	Linking Tasks Removed	Final Correlation	WRMSD*
English–Language Arts	I	5	0	0.98	0.09
	II	5	0	1.00	0.08
	III	5	0	1.00	0.07
	IV	5	0	1.00	0.09
	V	4	0	1.00	0.05

Content Area	Level	No. Linking Tasks	Linking Tasks Removed	Final Correlation	WRMSD*
Mathematics	I	5	0	0.95	0.05
	II	5	0	0.99	0.07
	III	5	0	1.00	0.05
	IV	5	0	1.00	0.03
	V	5	0	0.99	0.08
Science	I	5	0	0.45	0.10
	III	5	0	0.95	0.03
	IV	5	0	0.93	0.07
	V	5	0	1.00	0.05

* Average over retained tasks

Summaries of Scaled IRT b -values

Once the IRT b -values are placed on the item bank scale, analyses are performed to assess the overall test difficulty and the distribution of tasks in a particular range of item difficulty.

Table 8.D.7 through Table 8.D.9 present univariate statistics (mean, standard deviation, minimum, and maximum) for the scaled IRT b -values. The results for the overall test are presented separately for the operational tasks and the field-test tasks.

Post-scaling Results

As described on page 15 of Chapter 2, once the new item calibrations for each test are transformed to the base scale, transformed thetas are linearly converted using equation 2.2. to two-digit scale scores that ranged from 15 to 60. Complete raw-to-scale score conversion tables for the 2012 CAPA are presented in Table 8.D.10 through Table 8.D.23 in Appendix 8.D starting on page 138. The raw scores and corresponding rounded converted scale scores are listed in those tables.

For all of the 2012 CAPA, scale scores were truncated at both ends of the scale so that the minimum reported scale score was 15 and the maximum reported scale score was 60. The scale scores defining the cut scores for all performance levels are presented in Table 2.2, which is on page 15 in Chapter 2.

Differential Item Functioning Analyses

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

DIF analyses were performed on all operational tasks and all field-test tasks for which sufficient student samples were available. The sample size requirements for the DIF analyses were 100 in the focal group and 400 in the combined focal and reference groups. These sample sizes were based on standard operating procedures with respect to DIF analyses at ETS.

DIF analyses of the polytomously scored CAPA tasks are completed using two procedures. The first is the Mantel-Haenszel (MH) ordinal procedure, which is based on the Mantel procedure (Mantel, 1963; Mantel & Haenszel, 1959). The MH ordinal procedure compares the proportion of examinees in the reference and focal groups obtaining each task score after matching the examinees on their total test score. As with dichotomously scored tasks, the common odds ratio is estimated across the matched score groups. The resulting estimate is interpreted as the relative likelihood of obtaining a given task score for members of two groups that are matched on ability.

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

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

The standardized mean difference is computed as the following:

$$SMD = \frac{\sum_m w_m (E_{fm} - E_{rm})}{\sum_m w_m} \quad (8.6)$$

where,

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

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

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

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

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

In addition, the categories identify which group is being advantaged; categories are displayed in Table 8.6. The categories have been used by all ETS testing programs for more than 15 years.

Table 8.6 DIF Flags Based on the ETS DIF Classification Scheme

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

Category C contains tasks with large values of DIF. As shown in Table 8.6, tasks classified as C+ tend to be easier for members of the focal group than for members of the reference group with comparable total scores. Tasks classified as C– tend to be more difficult for members of the focal group than for members of the reference group whose total scores on the test are like those of the focal group.

The results of the DIF analyses are presented in Appendix 8.E, which starts on page 152. Table 8.E.1 and Table 8.E.2 list the tasks exhibiting significant DIF. Test developers are 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 is deemed essential to meeting test-content specifications.

Table 8.7 lists specific subgroups that were used for DIF analyses for the CAPA.

Table 8.7 Subgroup Classification for DIF Analyses

DIF Type	Reference Group	Focal Group
Gender	Male	Female
Race/Ethnicity	White	<ul style="list-style-type: none"> • African American • American Indian • Asian • Combined Asian Group (Asian/Pacific Islander/Filipino) • Filipino • Hispanic/Latin American • Pacific Islander
Disability	Mental Retardation/ Intellectual Disability (MR/ID)	<ul style="list-style-type: none"> • Autism • Deaf-Blindness • Deafness • Emotional Disturbance • Hard of Hearing • Multiple Disabilities • Orthopedic Impairment • Other Health Impairment • Specific Learning Disability • Speech or Language Impairment • Traumatic Brain Injury • Visual Impairment

Table 8.E.3 through Table 8.E.7 show the sample size for disability groups within test level and content area.

References

- AERA, APA, & NCME 1999. *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. New York, NY: Holt.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 292–334.
- Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, D. C.: American Council on Education.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York, NY: Wiley.
- Dorans, N. J. & Holland, P. W. (1993). *DIF detection and description: Mantel-Haenszel and standardization*. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp.35–66). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Dorans, N. J. & Kulick, E. (2006). Differential item functioning on the mini-mental state examination: An application of the Mantel-Haenszel and standardization procedures. *Medical Care*, *44*, 107–14.
- Dorans, N. J. & Schmitt, A. P. (1993). Constructed response and differential item functioning: A pragmatic approach. In R.E. Bennett & W.C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 135–65). Hillsdale, NH: Lawrence Erlbaum Associates, Inc.
- Dragow F. (1988). Polychoric and polyserial correlations. In L. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences* (Vol. 7, pp. 69–74). New York: Wiley.
- Educational Testing Service. (2002). *ETS standards for quality and fairness*. Princeton, NJ: Author.
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Washington, DC: American Council on Education and National Council on Measurement in Education.
- Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classification based on test scores. *Journal of Educational Measurement*, *32*, 179–97.
- Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure, *Journal of the American Statistical Association*, *58*, 690–700.
- Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analyses of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*, 719–48.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed. pp. 13–103). New York, NY: Macmillan.

- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159–76.
- Muraki, E. & Bock, R. D. (1995). *PARSCALE: Parameter scaling of rating data* (Computer software, Version 2.2). Chicago, IL: Scientific Software.
- Stocking, M. L., and Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, pp. 201–10.
- United States Department of Education. (2001). Elementary and Secondary Education Act (Public Law 107-11), Title VI, Chapter B, § 4, Section 6162. Retrieved from <http://www2.ed.gov/policy/elsec/leg/esea02/index.html>

Appendix 8.A—Classical Analyses: Task Statistics

Table 8.A.1 AIS and Polyserial Correlation: Level I, ELA

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

Version/ Field-Test Form	Task Position	AIS	Polyserial	Flag
Operational	1	3.18	.78	
1	2	3.44	.74	
Operational	3	2.76	.78	
Operational	4	3.61	.82	
1	5	3.57	.64	
Operational	6	3.47	.83	
Operational	7	2.87	.77	
1	8	2.83	.75	
Operational	9	2.90	.77	
Operational	10	3.09	.82	
1	11	3.35	.79	
Operational	12	3.09	.81	
2	2	3.47	.77	
2	5	2.76	.76	
2	8	2.96	.68	
2	11	3.14	.80	
3	2	3.16	.75	
3	5	2.98	.74	
3	8	4.09	.78	H
3	11	3.15	.74	
4	2	3.00	.68	
4	5	3.56	.53	R
4	8	3.27	.73	
4	11	3.47	.73	

Table 8.A.2 AIS and Polyserial Correlation: Level II, ELA**Flag values are as follows:****A** = low average task score**R** = low correlation with criterion**O** = high percent of omits/not responding**H** = high average task score

Version/ Field-Test Form	Task Position	AIS	Polyserial	Flag
Operational	1	1.96	.85	
1	2	3.02	.76	
Operational	3	3.57	.72	H
Operational	4	2.18	.84	
1	5	2.58	.66	
Operational	6	2.66	.77	
Operational	7	2.16	.85	
1	8	2.87	.56	R
Operational	9	2.41	.68	
Operational	10	2.17	.81	
1	11	3.00	.76	
Operational	12	1.89	.70	
2	2	2.65	.67	
2	8	3.08	.60	
2	11	2.90	.53	R
3	2	2.47	.45	R
3	5	2.48	.65	
3	8	2.76	.53	R
4	2	2.41	.47	R
4	5	3.18	.61	
4	8	2.61	.48	R
4	11	2.38	.78	

Table 8.A.3 AIS and Polyserial Correlation: Level III, ELA

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

Version/ Field-Test Form	Task Position	AIS	Polyserial	Flag
Operational	1	2.50	.85	
1	2	3.32	.64	H
Operational	3	2.64	.84	
Operational	4	3.01	.75	
1	5	3.09	.65	
Operational	6	2.27	.68	
Operational	7	2.74	.85	
1	8	3.47	.66	H
Operational	9	2.52	.86	
Operational	10	2.21	.88	
1	11	2.48	.46	R
Operational	12	2.26	.70	
2	2	3.74	.62	H
2	5	3.33	.68	H
2	8	2.67	.48	R
2	11	2.69	.54	R
3	2	2.74	.55	R
3	5	2.21	.66	
3	8	2.98	.67	
3	11	2.80	.67	
4	2	3.27	.69	H
4	5	3.04	.61	
4	8	3.60	.65	H
4	11	2.95	.59	R

Table 8.A.4 AIS and Polyserial Correlation: Level IV, ELA**Flag values are as follows:****A** = low average task score**R** = low correlation with criterion**O** = high percent of omits/not responding**H** = high average task score

Version/ Field-Test Form	Task Position	AIS	Polyserial	Flag	
Operational	1	2.49	.81		
1	2	3.59	.59	R	H
Operational	3	2.47	.81		
Operational	4	1.61	.75		
1	5	2.02	.78		
Operational	6	2.56	.84		
Operational	7	2.31	.82		
1	8	3.14	.63		
Operational	9	2.21	.82		
Operational	10	2.66	.72		
1	11	3.09	.58	R	
Operational	12	2.32	.82		
2	2	3.42	.63		H
2	5	2.65	.60		
2	8	2.90	.69		
2	11	2.00	.78		
3	2	3.20	.52	R	
3	5	2.16	.75		
3	8	2.44	.81		
3	11	2.16	.81		
4	2	2.92	.69		
4	5	2.75	.70		
4	8	2.30	.77		
4	11	2.40	.82		

Table 8.A.5 AIS and Polyserial Correlation: Level V, ELA

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

Version/ Field-Test Form	Task Position	AIS	Polyserial	Flag
Operational	1	3.05	.79	
1	2	2.66	.36	R
Operational	3	3.07	.67	
Operational	4	2.81	.76	
1	5	2.68	.79	
Operational	6	1.98	.79	
Operational	7	2.52	.83	
1	8	2.16	.79	
Operational	9	2.56	.86	
Operational	10	2.45	.86	
1	11	2.94	.76	
Operational	12	2.09	.76	
2	2	2.21	.36	R
2	5	2.55	.79	
2	8	2.92	.50	R
2	11	3.12	.67	
3	2	2.89	.76	
3	5	2.99	.62	
3	8	3.35	.69	H
3	11	3.24	.75	H
4	2	2.01	.59	R
4	5	2.32	.80	
4	8	3.43	.62	H
4	11	3.05	.72	

Table 8.A.6 AIS and Polyserial Correlation: Level I, Mathematics

Flag values are as follows:

A = low average task score

R = low correlation with criterion

O = high percent of omits/not responding

H = high average task score

Version/ Field-Test Form	Task Position	AIS	Polyserial	Flag
Operational	13	2.58	.78	
1	14	3.00	.63	
Operational	15	2.59	.72	
Operational	16	2.79	.73	
1	17	3.24	.71	
Operational	18	2.85	.73	
Operational	19	3.25	.81	
1	20	2.72	.70	
Operational	21	3.25	.76	
Operational	22	3.00	.78	
1	23	2.45	.65	
Operational	24	2.59	.79	
2	14	3.32	.74	
2	17	2.68	.62	
2	20	2.51	.60	R
2	23	2.65	.77	
3	14	3.08	.71	
3	17	3.03	.71	
3	20	2.93	.65	
3	23	2.85	.63	
4	14	3.11	.58	R
4	17	3.30	.69	
4	20	2.12	.64	
4	23	3.30	.76	

Table 8.A.7 AIS and Polyserial Correlation: Level II, Mathematics

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

Version/ Field-Test Form	Task Position	AIS	Polyserial	Flag
Operational	13	2.99	.84	
1	14	2.82	.60	R
Operational	15	3.20	.74	
Operational	16	2.08	.79	
1	17	3.07	.71	
Operational	18	2.27	.62	
Operational	19	2.65	.86	
1	20	2.74	.53	R
Operational	21	2.82	.78	
Operational	22	2.37	.85	
1	23	3.49	.74	H
Operational	24	1.24	.65	
2	14	2.30	.58	R
2	17	2.42	.71	
2	20	1.39	.52	R
2	23	2.93	.83	
3	14	1.32	.53	R
3	17	3.53	.70	H
3	20	3.09	.61	
3	23	3.65	.67	H
4	14	2.57	.70	
4	17	2.72	.64	
4	20	3.05	.75	
4	23	3.53	.68	H

Table 8.A.8 AIS and Polyserial Correlation: Level III, Mathematics**Flag values are as follows:****A** = low average task score**R** = low correlation with criterion**O** = high percent of omits/not responding**H** = high average task score

Version/ Field-Test Form	Task Position	AIS	Polyserial	Flag
Operational	13	2.47	.55	R
1	14	3.06	.73	
Operational	15	2.12	.80	
Operational	16	2.03	.67	
1	17	1.64	.52	R
Operational	18	2.96	.82	
Operational	19	2.86	.83	
1	20	2.61	.58	R
Operational	21	2.79	.64	
Operational	22	1.78	.61	
1	23	3.18	.69	
Operational	24	2.08	.78	
2	14	3.00	.72	
2	17	2.34	.72	
2	20	2.67	.62	
2	23	2.31	.69	
3	14	3.32	.51	R H
3	17	3.06	.74	
3	20	2.56	.72	
3	23	2.59	.65	
4	14	3.64	.72	H
4	17	2.54	.75	
4	20	2.25	.47	R
4	23	2.01	.57	R

Table 8.A.9 AIS and Polyserial Correlation: Level IV, Mathematics**Flag values are as follows:****A** = low average task score**R** = low correlation with criterion**O** = high percent of omits/not responding**H** = high average task score

Version/ Field-Test Form	Task Position	AIS	Polyserial	Flag
Operational	13	1.50	.65	
1	14	1.90	.74	
Operational	15	2.44	.85	
Operational	16	2.96	.85	
1	17	3.35	.54	R H
Operational	18	1.81	.77	
Operational	19	2.93	.84	
1	20	1.54	.56	R
Operational	21	2.97	.64	
Operational	22	2.64	.61	
1	23	3.05	.83	
Operational	24	2.66	.61	
2	14	3.26	.41	R H
2	17	2.46	.55	R
2	20	2.66	.80	
2	23	3.05	.83	
3	14	1.73	.67	
3	17	1.95	.66	
3	20	2.62	.82	
3	23	2.91	.55	R
4	14	1.90	.73	
4	17	3.05	.72	
4	20	1.70	.69	
4	23	2.42	.79	

Table 8.A.10 AIS and Polyserial Correlation: Level V, Mathematics

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

Version/ Field-Test Form	Task Position	AIS	Polyserial	Flag
Operational	13	2.57	.84	
1	14	2.47	.63	
Operational	15	3.35	.82	H
Operational	16	2.81	.72	
1	17	3.28	.78	H
Operational	18	2.30	.80	
Operational	19	2.59	.74	
1	20	2.20	.77	
Operational	21	2.69	.73	
Operational	22	1.94	.71	
1	23	2.52	.74	
Operational	24	2.94	.81	
2	14	2.13	.67	
2	17	1.76	.68	
2	20	2.45	.74	
2	23	3.43	.68	H
3	14	2.56	.69	
3	17	3.37	.76	H
3	20	2.46	.75	
3	23	2.73	.80	
4	14	3.04	.76	
4	17	3.31	.82	H
4	20	3.69	.68	H
4	23	3.27	.78	H

Table 8.A.11 AIS and Polyserial Correlation: Level I, Science**Flag values are as follows:****A** = low average task score**R** = low correlation with criterion**O** = high percent of omits/not responding**H** = high average task score

Version/Field-Test Form	Task Position	AIS	Polyserial	Flag
Operational	25	3.26	.76	
1/3*	26	2.76	.74	
Operational	27	3.03	.80	
Operational	28	2.95	.79	
1/3*	29	3.02	.66	
Operational	30	3.11	.83	
Operational	31	2.73	.82	
1/3*	32	2.48	.67	
Operational	33	2.84	.77	
Operational	34	2.98	.77	
1/3*	35	3.33	.56	R
Operational	36	2.37	.80	
2/4*	26	2.88	.74	
2/4*	29	2.58	.63	
2/4*	32	2.90	.71	
2/4*	35	2.95	.74	

* This task appeared on more than one field-test form.

Table 8.A.12 AIS and Polyserial Correlation: Level III, Science

Flag values are as follows:

A = low average task score

R = low correlation with criterion

O = high percent of omits/not responding

H = high average task score

Version/Field-Test Form	Task Position	AIS	Polyserial	Flag
Operational	25	2.24	.68	
1/3*	26	2.51	.59	R
Operational	27	2.66	.67	
Operational	28	2.63	.79	
1/3*	29	2.96	.72	
Operational	30	2.50	.79	
Operational	31	2.76	.72	
1/3*	32	1.79	.70	
Operational	33	2.95	.73	
Operational	34	2.41	.71	
1/3*	35	2.57	.63	
Operational	36	2.63	.69	
2/4*	26	3.03	.59	R
2/4*	29	2.45	.72	
2/4*	32	2.42	.62	
2/4*	35	2.49	.70	

* This task appeared on more than one field-test form.

Table 8.A.13 AIS and Polyserial Correlation: Level IV, Science**Flag values are as follows:****A** = low average task score**R** = low correlation with criterion**O** = high percent of omits/not responding**H** = high average task score

Version/ Field-Test Form	Task Position	AIS	Polyserial	Flag
Operational	25	2.55	.62	
1/3*	26	2.29	.55	R
Operational	27	2.91	.74	
Operational	28	2.36	.68	
1/3*	29	2.48	.61	
Operational	30	2.90	.70	
Operational	31	2.91	.69	
1/3*	32	2.92	.68	
Operational	33	2.91	.73	
Operational	34	2.17	.77	
1/3*	35	2.40	.63	
Operational	36	2.82	.69	
2/4*	26	2.82	.65	
2/4*	29	3.03	.69	
2/4*	32	2.91	.66	
2/4*	35	2.69	.61	

* This task appeared on more than one field-test form.

Table 8.A.14 AIS and Polyserial Correlation: Level V, Science

Flag values are as follows:

A = low average task score

R = low correlation with criterion

O = high percent of omits/not responding

H = high average task score

Version/ Field-Test Form	Task Position	AIS	Polyserial	Flag
Operational	25	3.31	.77	H
1/3*	26	2.70	.46	R
Operational	27	2.37	.71	
Operational	28	2.75	.77	
1/3*	29	2.81	.63	
Operational	30	2.15	.76	
Operational	31	2.56	.70	
Operational	33	2.94	.79	
Operational	34	3.42	.81	H
1/3*	35	1.75	.53	R
Operational	36	2.44	.73	
2/4*	26	3.09	.66	
2/4*	29	2.77	.68	
2/4*	32	2.02	.54	R
2/4*	35	1.97	.59	R

* This task appeared on more than one field-test form.

Table 8.A.15 Frequency of Operational Task Scores: ELA

ELA Level	Score on Task	1		2		3		4		5		6		7		8	
		Count	Pct	Count	Percent												
I	0	1,603	11.09	1,590	11.00	1,395	9.65	1,519	10.51	1,638	11.34	1,698	11.75	1,601	11.08	1,759	12.17
	1	3,686	25.51	5,185	35.88	2,665	18.44	2,936	20.32	4,421	30.60	4,432	30.67	4,180	28.93	4,013	27.77
	2	652	4.51	632	4.37	502	3.47	610	4.22	871	6.03	1,048	7.25	610	4.22	590	4.08
	3	695	4.81	770	5.33	541	3.74	570	3.94	896	6.20	617	4.27	635	4.39	618	4.28
	4	1,328	9.19	1,158	8.01	1,072	7.42	1,063	7.36	1,562	10.81	926	6.41	1,041	7.20	1,106	7.65
5	6,485	44.88	5,114	35.39	8,274	57.26	7,751	53.64	5,061	35.03	5,728	39.64	6,382	44.17	6,363	44.04	
II	0	708	10.04	279	3.96	607	8.61	311	4.41	635	9.00	320	4.54	443	6.28	516	7.32
	1	2,280	32.33	245	3.47	1,848	26.21	1,432	20.31	2,115	29.99	1,526	21.64	2,277	32.29	2,578	36.56
	2	1,398	19.82	472	6.69	1,396	19.80	1,590	22.55	1,141	16.18	1,993	28.26	1,756	24.90	2,142	30.37
	3	2,177	30.87	972	13.78	2,375	33.68	1,147	16.26	2,103	29.82	1,766	25.04	1,126	15.97	1,119	15.87
	4	489	6.93	5,084	72.09	826	11.71	2,572	36.47	1,058	15.00	1,447	20.52	1,450	20.56	697	9.88
III	0	626	8.34	525	7.00	365	5.24	365	4.86	514	6.85	595	7.93	643	8.57	424	5.65
	1	1,401	18.67	887	11.82	547	7.29	1,222	16.29	1,255	16.73	1,889	25.18	2,249	29.97	1,067	14.22
	2	1,359	18.11	1,780	23.72	1,239	16.51	3,234	43.10	1,219	16.25	1,076	14.34	1,324	17.65	3,442	45.87
	3	2,417	32.21	2,440	32.52	2,521	33.60	1,855	24.72	1,785	23.79	1,471	19.61	1,928	25.70	1,894	25.24
	4	1,700	22.66	1,871	24.94	2,803	37.36	827	11.02	2,730	36.39	2,472	32.95	1,359	18.11	676	9.01
IV	0	575	5.46	750	7.12	1,067	10.12	716	6.79	577	5.47	598	5.67	550	5.22	838	7.95
	1	1,839	17.45	2,085	19.78	5,817	55.19	2,186	20.74	3,687	34.98	4,167	39.54	1,654	15.69	3,106	29.47
	2	2,313	21.94	2,761	26.20	1,493	14.17	1,884	17.87	1,515	14.37	1,129	10.71	2,519	23.90	1,800	17.08
	3	4,167	39.54	2,063	19.57	951	9.02	2,759	26.18	2,048	19.43	2,352	22.31	2,798	26.55	2,107	19.99
	4	1,646	15.62	2,881	27.33	1,212	11.50	2,995	28.42	2,713	25.74	2,294	21.76	3,019	28.64	2,689	25.51
V	0	808	7.40	630	5.77	739	6.77	783	7.17	854	7.82	875	8.02	925	8.48	799	7.32
	1	1,216	11.14	642	5.88	1,234	11.31	4,875	44.67	1,941	17.78	2,006	18.38	2,684	24.59	4,447	40.75
	2	1,144	10.48	1,323	12.12	1,987	18.21	1,516	13.89	2,122	19.44	1,909	17.49	1,753	16.06	1,760	16.13
	3	2,672	24.48	4,541	41.61	3,735	34.22	2,267	20.77	3,988	36.54	3,606	33.04	2,846	26.08	1,953	17.89
	4	5,074	46.49	3,778	34.62	3,219	29.49	1,473	13.50	2,009	18.41	2,518	23.07	2,706	24.79	1,955	17.91

Table 8.A.16 Frequency of Operational Task Scores: Mathematics

Math Level	Score on Task	1		2		3		4		5		6		7		8	
		Count	Percent														
I	0	2,127	14.72	1,749	12.10	1,519	10.51	1,667	11.54	1,544	10.69	1,400	9.69	1,735	12.01	2,108	14.59
	1	4,997	34.58	5,464	37.82	5,267	36.45	4,821	33.37	3,817	26.42	3,841	26.58	4,186	28.97	5,219	36.12
	2	874	6.05	721	4.99	617	4.27	661	4.57	513	3.55	559	3.87	681	4.71	725	5.02
	3	812	5.62	861	5.96	681	4.71	709	4.91	516	3.57	680	4.71	713	4.93	777	5.38
	4	1,070	7.41	1,464	10.13	1,194	8.26	1,211	8.38	915	6.33	1,176	8.14	1,245	8.62	991	6.86
	5	4,569	31.62	4,190	29.00	5,171	35.79	5,380	37.23	7,144	49.44	6,793	47.01	5,889	40.76	4,629	32.04
II	0	271	3.84	273	3.87	455	6.45	393	5.57	467	6.62	333	4.72	486	6.89	727	10.31
	1	1,277	18.11	477	6.76	3,556	50.43	1,861	26.39	2,137	30.30	1,295	18.36	2,538	35.99	5,036	71.41
	2	726	10.29	708	10.04	512	7.26	2,374	33.66	550	7.80	1,149	16.29	805	11.42	611	8.66
	3	1,297	18.39	2,369	33.59	485	6.88	677	9.60	603	8.55	1,339	18.99	767	10.88	380	5.39
	4	3,481	49.36	3,225	45.73	2,044	28.98	1,747	24.77	3,295	46.72	2,936	41.63	2,456	34.83	298	4.23
III	0	329	4.38	372	4.96	423	5.64	361	4.81	386	5.14	338	4.50	382	5.09	569	7.58
	1	1,401	18.67	3,019	40.24	2,939	39.17	1,588	21.16	1,667	22.22	653	8.70	3,998	53.29	3,266	43.53
	2	2,514	33.51	1,124	14.98	1,655	22.06	780	10.40	1,029	13.71	1,255	16.73	1,435	19.13	865	11.53
	3	1,524	20.31	1,678	22.36	1,497	19.95	792	10.56	690	9.20	3,962	52.81	676	9.01	1,229	16.38
	4	1,735	23.12	1,310	17.46	989	13.18	3,982	53.07	3,731	49.73	1,295	17.26	1,012	13.49	1,574	20.98
IV	0	761	7.22	636	6.03	634	6.02	865	8.21	617	5.85	558	5.29	563	5.34	645	6.12
	1	5,446	51.67	4,031	38.24	2,265	21.49	5,992	56.85	2,358	22.37	1,055	10.01	1,642	15.58	1,934	18.35
	2	3,489	33.10	685	6.50	615	5.83	778	7.38	805	7.64	2,019	19.16	2,489	23.61	2,151	20.41
	3	509	4.83	1,152	10.93	1,449	13.75	631	5.99	1,212	11.50	2,630	24.95	3,204	30.40	2,593	24.60
	4	335	3.18	4,036	38.29	5,577	52.91	2,274	21.57	5,548	52.64	4,278	40.59	2,642	25.07	3,217	30.52
V	0	738	6.76	714	6.54	658	6.03	738	6.76	755	6.92	789	7.23	869	7.96	922	8.45
	1	4,051	37.12	1,573	14.41	1,211	11.10	4,535	41.55	2,842	26.04	3,505	32.11	4,964	45.48	2,021	18.52
	2	618	5.66	296	2.71	3,784	34.67	1,138	10.43	1,921	17.60	621	5.69	1,979	18.13	1,302	11.93
	3	601	5.51	609	5.58	688	6.30	1,023	9.37	1,515	13.88	878	8.04	1,422	13.03	1,182	10.83
	4	4,906	44.95	7,722	70.75	4,573	41.90	3,480	31.89	3,881	35.56	5,121	46.92	1,680	15.39	5,487	50.27

Table 8.A.17 Frequency of Operational Task Scores: Science

Science Score on Level	Task	1		2		3		4		5		6		7		8	
		Count	Percent														
I	0	627	14.49	760	17.56	732	16.91	731	16.89	764	17.65	764	17.65	686	15.85	1,035	23.91
	1	1,077	24.88	1,127	26.04	1,258	29.07	1,130	26.11	1,391	32.14	1,286	29.71	1,270	29.34	1,528	35.3
	2	147	3.4	181	4.18	180	4.16	164	3.79	198	4.57	215	4.97	184	4.25	187	4.32
	3	166	3.84	216	4.99	196	4.53	150	3.47	212	4.9	208	4.81	188	4.34	228	5.27
	4	349	8.06	328	7.58	351	8.11	270	6.24	312	7.21	319	7.37	328	7.58	262	6.05
III	0	1,962	45.33	1,716	39.65	1,611	37.22	1,883	43.51	1,451	33.53	1,536	35.49	1,672	38.63	1,088	25.14
	1	291	7.48	233	5.99	303	7.79	262	6.73	276	7.09	264	6.78	296	7.61	273	7.01
	2	851	21.87	451	11.59	719	18.47	1,255	32.25	492	12.64	428	11	674	17.32	450	11.56
	3	1,247	32.04	1,317	33.84	701	18.01	609	15.65	910	23.38	681	17.5	1,104	28.37	1,063	27.31
	4	1,071	27.52	826	21.22	1,043	26.8	297	7.63	1,013	26.03	1,050	26.98	1,303	33.48	1,353	34.76
IV	0	432	11.1	1,065	27.36	1,126	28.93	1,469	37.74	1,201	30.86	1,469	37.74	515	13.23	753	19.35
	1	258	7.07	246	6.74	255	6.99	281	7.7	241	6.6	257	7.04	390	10.69	270	7.4
	2	523	14.33	432	11.84	842	23.07	466	12.77	148	4.06	398	10.91	1,127	30.89	533	14.61
	3	983	26.94	787	21.57	1,066	29.21	639	17.51	713	19.54	569	15.59	843	23.1	655	17.95
	4	1,219	33.41	711	19.48	772	21.16	884	24.23	1,760	48.23	1,233	33.79	470	12.88	929	25.46
V	0	666	18.25	1,473	40.37	714	19.57	1,379	37.79	787	21.57	1,192	32.67	819	22.44	1,262	34.58
	1	362	9.08	388	9.73	396	9.93	451	11.31	392	9.83	390	9.78	390	9.78	455	11.41
	2	235	5.89	828	20.76	493	12.36	1,143	28.66	638	16	311	7.8	228	5.72	1,150	28.84
	3	361	9.05	968	24.27	690	17.3	1,042	26.13	1,017	25.5	721	18.08	354	8.88	473	11.86
	4	963	24.15	1,331	33.38	1,515	37.99	744	18.66	1,079	27.06	1,321	33.12	577	14.47	926	23.22
	4	2,067	51.83	473	11.86	894	22.42	608	15.25	862	21.61	1,245	31.22	2,439	61.16	984	24.67

Appendix 8.B—Reliability Analyses

The reliabilities are reported only for samples that comprise 11 or more examinees. Also, in some cases in Appendix 8.B, score reliabilities were not estimable and are presented in the tables as hyphens. Finally, results based on samples that contain 50 or fewer examinees should be interpreted with caution due to small sample sizes.

Table 8.B.1 Reliabilities and SEMs by GENDER

Content Area	Level	No. of Tasks	Male		Female		Unknown Gender	
			Reliab.	SEM	Reliab.	SEM	Reliab.	SEM
English– Language Arts	I	8	0.88	3.92	0.90	3.83	0.90	3.92
	II	8	0.88	2.32	0.87	2.32	0.84	2.44
	III	8	0.90	2.17	0.90	2.17	0.89	2.21
	IV	8	0.90	2.34	0.90	2.31	0.92	2.21
	V	8	0.89	2.30	0.89	2.23	0.87	2.20
Mathematics	I	8	0.85	4.21	0.86	4.14	0.87	4.17
	II	8	0.86	2.61	0.85	2.57	0.85	2.58
	III	8	0.82	2.67	0.79	2.66	0.79	2.70
	IV	8	0.84	2.66	0.83	2.66	0.82	2.79
	V	8	0.86	2.76	0.86	2.74	0.84	2.75
Science	I	8	0.87	4.06	0.89	3.95	0.89	4.12
	III	8	0.85	2.43	0.80	2.49	0.84	2.55
	IV	8	0.82	2.49	0.81	2.43	0.83	2.24
	V	8	0.86	2.26	0.85	2.29	0.88	2.20

Table 8.B.2 Reliabilities and SEMs by PRIMARY ETHNICITY

Content Area	Level	No. of Tasks	American Indian		Asian		Pacific Islander		Filipino	
			Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM
English–Language Arts	I	8	0.87	4.08	0.86	4.10	0.88	3.99	0.87	4.02
	II	8	0.83	2.08	0.87	2.31	0.84	2.54	0.86	2.33
	III	8	0.85	2.24	0.90	2.21	0.92	2.15	0.91	2.14
	IV	8	0.91	2.18	0.88	2.33	0.90	2.30	0.89	2.34
	V	8	0.88	2.29	0.89	2.28	0.93	2.04	0.91	2.27
Mathematics	I	8	0.84	4.29	0.83	4.30	0.87	4.09	0.83	4.32
	II	8	0.79	2.60	0.86	2.58	0.87	2.54	0.85	2.65
	III	8	0.77	2.61	0.84	2.66	0.66	3.00	0.84	2.67
	IV	8	0.80	2.74	0.83	2.72	0.86	2.64	0.81	2.79
	V	8	0.87	2.77	0.87	2.69	0.90	2.50	0.87	2.77
Science	I	8	0.87	4.34	0.89	3.96	0.91	3.79	0.87	4.08
	III	8	0.79	2.28	0.82	2.52	0.77	2.61	0.87	2.39
	IV	8	0.75	2.56	0.80	2.44	0.88	2.30	0.83	2.41
	V	8	0.90	2.06	0.86	2.25	0.77	2.56	0.87	2.34
Content Area	Level	No. of Tasks	Hispanic		African American		White		Unknown Ethnicity	
			Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM
English–Language Arts	I	8	0.90	3.82	0.89	3.81	0.88	3.96	0.88	4.03
	II	8	0.87	2.32	0.88	2.28	0.88	2.33	0.88	2.37
	III	8	0.90	2.15	0.90	2.14	0.91	2.16	0.90	2.26
	IV	8	0.90	2.30	0.89	2.35	0.90	2.33	0.89	2.45
	V	8	0.89	2.24	0.88	2.27	0.89	2.30	0.88	2.33
Mathematics	I	8	0.86	4.11	0.86	4.14	0.83	4.30	0.83	4.32
	II	8	0.86	2.59	0.87	2.56	0.86	2.63	0.85	2.61
	III	8	0.81	2.65	0.80	2.67	0.81	2.68	0.80	2.70
	IV	8	0.82	2.66	0.84	2.61	0.85	2.65	0.84	2.67
	V	8	0.86	2.75	0.85	2.75	0.86	2.77	0.85	2.82
Science	I	8	0.88	4.00	0.87	4.07	0.87	4.05	0.84	4.26
	III	8	0.83	2.43	0.84	2.46	0.85	2.44	0.72	2.66
	IV	8	0.80	2.47	0.79	2.44	0.82	2.51	0.86	2.43
	V	8	0.85	2.27	0.85	2.23	0.86	2.27	0.85	2.29

Table 8.B.3 Reliabilities and SEMs by PRIMARY ETHNICITY for Economically Disadvantaged

Content Area	Level	No. of Tasks	American Indian		Asian		Pacific Islander		Filipino	
			Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM
English–Language Arts	I	8	0.84	4.33	0.87	4.00	0.90	3.80	0.88	3.95
	II	8	0.86	2.09	0.89	2.25	0.88	2.45	0.81	2.35
	III	8	0.84	2.20	0.91	2.16	0.91	2.23	0.91	2.18
	IV	8	0.91	2.20	0.88	2.34	0.88	2.28	0.85	2.42
	V	8	0.86	2.20	0.90	2.24	0.94	2.10	0.90	2.25
Mathematics	I	8	0.86	4.24	0.83	4.27	0.88	4.01	0.83	4.26
	II	8	0.81	2.64	0.87	2.57	0.90	2.54	0.79	2.66
	III	8	0.76	2.55	0.83	2.62	0.70	2.94	0.85	2.58
	IV	8	0.80	2.73	0.85	2.66	0.83	2.67	0.76	2.74
	V	8	0.86	2.74	0.88	2.61	0.91	2.43	0.87	2.71
Science *	I	8	–	–	0.88	4.00	0.92	3.88	0.86	4.16
	III	8	0.65	2.27	0.79	2.56	0.76	2.33	0.83	2.51
	IV	8	0.77	2.57	0.84	2.43	0.88	2.27	0.68	2.48
	V	8	0.83	2.14	0.78	2.27	0.83	2.53	0.89	2.38
Content Area	Level	No. of Tasks	Hispanic		African American		White		Unknown Ethnicity	
			Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM
English–Language Arts	I	8	0.89	3.80	0.89	3.76	0.89	3.86	0.89	3.84
	II	8	0.87	2.32	0.88	2.27	0.88	2.27	0.88	2.33
	III	8	0.90	2.15	0.90	2.13	0.91	2.13	0.88	2.36
	IV	8	0.90	2.30	0.89	2.36	0.89	2.34	0.88	2.43
	V	8	0.88	2.23	0.88	2.23	0.89	2.23	0.89	2.32
Mathematics	I	8	0.86	4.13	0.85	4.13	0.85	4.24	0.84	4.18
	II	8	0.86	2.59	0.87	2.55	0.86	2.57	0.85	2.58
	III	8	0.80	2.65	0.80	2.67	0.81	2.64	0.79	2.74
	IV	8	0.82	2.66	0.84	2.59	0.84	2.63	0.82	2.72
	V	8	0.86	2.74	0.85	2.74	0.86	2.72	0.84	2.83
Science	I	8	0.87	4.05	0.88	3.91	0.88	3.95	0.86	4.08
	III	8	0.82	2.43	0.84	2.43	0.85	2.41	0.76	2.54
	IV	8	0.79	2.48	0.80	2.44	0.80	2.50	0.82	2.64
	V	8	0.84	2.27	0.83	2.23	0.85	2.24	0.84	2.31

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

Table 8.B.4 Reliabilities and SEMs by PRIMARY ETHNICITY for Not Economically Disadvantaged

Content Area	Level	No. of Tasks	American Indian		Asian		Pacific Islander		Filipino	
			Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM
English– Language Arts *	I	8	0.90	3.71	0.85	4.18	0.84	4.29	0.85	4.11
	II	8	–	–	0.85	2.36	0.72	2.63	0.87	2.31
	III	8	0.86	2.22	0.90	2.26	0.93	2.05	0.92	2.10
	IV	8	0.91	2.15	0.88	2.32	0.92	2.09	0.90	2.29
	V	8	0.89	2.41	0.89	2.31	0.88	1.97	0.92	2.27
Mathematics *	I	8	0.81	4.36	0.83	4.34	0.84	4.25	0.82	4.38
	II	8	–	–	0.86	2.57	0.82	2.59	0.87	2.62
	III	8	0.80	2.69	0.84	2.69	0.62	3.17	0.83	2.75
	IV	8	0.80	2.78	0.81	2.76	0.89	2.57	0.83	2.76
	V	8	0.88	2.75	0.86	2.78	0.89	2.62	0.86	2.85
Science *	I	8	0.91	3.93	0.89	3.93	–	–	0.87	4.07
	III	8	–	–	0.83	2.49	–	–	0.90	2.33
	IV	8	–	–	0.75	2.44	–	–	0.87	2.38
	V	8	0.93	2.03	0.88	2.26	–	–	0.86	2.34
Content Area	Level	No. of Tasks	Hispanic		African American		White		Unknown Ethnicity	
			Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM
English– Language Arts	I	8	0.90	3.87	0.89	3.92	0.87	4.02	0.86	4.22
	II	8	0.88	2.32	0.88	2.29	0.88	2.37	0.88	2.43
	III	8	0.91	2.18	0.91	2.20	0.91	2.19	0.92	2.21
	IV	8	0.90	2.31	0.90	2.32	0.91	2.32	0.90	2.48
	V	8	0.90	2.28	0.88	2.36	0.89	2.33	0.88	2.40
Mathematics	I	8	0.87	4.09	0.86	4.16	0.82	4.33	0.81	4.45
	II	8	0.87	2.58	0.86	2.58	0.86	2.65	0.86	2.68
	III	8	0.83	2.64	0.82	2.67	0.81	2.70	0.81	2.73
	IV	8	0.83	2.69	0.83	2.61	0.85	2.66	0.83	2.68
	V	8	0.86	2.77	0.85	2.78	0.86	2.80	0.86	2.84
Science	I	8	0.90	3.82	0.83	4.33	0.86	4.13	0.81	4.44
	III	8	0.86	2.42	0.85	2.54	0.85	2.45	0.68	2.73
	IV	8	0.85	2.40	0.79	2.41	0.82	2.52	0.91	2.23
	V	8	0.87	2.26	0.88	2.20	0.86	2.31	0.81	2.30

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

Table 8.B.5 Reliabilities and SEMs by PRIMARY ETHNICITY for Unknown Economic Status

Content Area	Level	No. of Tasks	American Indian		Asian		Pacific Islander		Filipino	
			Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM
English– Language Arts *	I	8	–	–	0.92	3.78	–	–	–	–
	II	8	–	–	0.83	2.38	–	–	–	–
	III	8	–	–	–	–	–	–	–	–
	IV	8	–	–	0.71	2.29	–	–	–	–
	V	8	–	–	0.79	2.65	–	–	–	–
Mathematics *	I	8	–	–	0.89	3.84	–	–	–	–
	II	8	–	–	0.84	2.59	–	–	–	–
	III	8	–	–	–	–	–	–	–	–
	IV	8	–	–	0.82	2.78	–	–	–	–
	V	8	–	–	0.75	2.59	–	–	–	–
Science *	I	8	–	–	–	–	–	–	–	–
	III	8	–	–	–	–	–	–	–	–
	IV	8	–	–	–	–	–	–	–	–
	V	8	–	–	–	–	–	–	–	–
Content Area	Level	No. of Tasks	Hispanic		African American		White		Unknown Ethnicity	
			Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM
English– Language Arts	I	8	0.94	3.34	0.92	3.49	0.89	3.86	0.92	3.69
	II	8	0.84	2.43	0.81	2.11	0.84	2.30	0.85	2.33
	III	8	0.86	2.10	0.74	1.83	0.90	2.04	0.89	2.07
	IV	8	0.88	2.39	0.90	2.19	0.93	2.10	0.89	2.41
	V	8	0.85	2.26	0.69	2.32	0.89	2.10	0.87	2.18
Mathematics	I	8	0.92	3.72	0.88	4.05	0.84	4.25	0.84	4.34
	II	8	0.81	2.57	0.79	2.75	0.72	2.88	0.86	2.51
	III	8	0.66	2.81	0.60	2.71	0.76	2.99	0.81	2.52
	IV	8	0.80	2.61	0.75	3.13	0.83	2.75	0.87	2.53
	V	8	0.76	2.96	0.71	2.71	0.81	2.73	0.84	2.78
Science *	I	8	0.78	4.43	–	–	0.96	3.01	–	–
	III	8	0.85	2.45	–	–	0.92	2.43	0.65	2.94
	IV	8	0.53	2.56	–	–	0.83	2.40	0.85	2.08
	V	8	0.80	2.35	–	–	0.86	2.20	0.92	2.13

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

Table 8.B.6 Reliabilities and SEMs by Disability

Content Area	Level	No. of Tasks	MR/ID		Hard of Hearing		Deafness		Speech Impairment	
			Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM
English–Language Arts	I	8	0.89	3.75	0.84	3.82	0.83	4.24	0.86	3.21
	II	8	0.85	2.30	0.83	2.33	0.85	2.42	0.82	2.27
	III	8	0.88	2.15	0.92	2.09	0.89	2.12	0.83	2.07
	IV	8	0.89	2.27	0.90	2.31	0.80	2.40	0.83	2.27
	V	8	0.88	2.21	0.89	2.23	0.83	2.21	0.76	2.21
Mathematics	I	8	0.84	4.17	0.82	3.97	0.79	4.45	0.75	4.10
	II	8	0.84	2.53	0.71	2.84	0.88	2.48	0.80	2.52
	III	8	0.78	2.59	0.76	2.80	0.83	2.54	0.75	2.62
	IV	8	0.82	2.64	0.79	2.69	0.78	2.61	0.74	2.52
	V	8	0.85	2.76	0.81	2.81	0.79	2.40	0.70	2.71
Science *	I	8	0.87	3.98	0.90	3.97	0.82	4.44	–	–
	III	8	0.77	2.48	0.79	2.54	0.70	2.56	0.73	2.40
	IV	8	0.79	2.45	0.69	2.73	0.64	2.80	0.71	2.58
	V	8	0.84	2.28	0.67	2.54	0.59	2.25	0.73	2.29
Content Area	Level	No. of Tasks	Visual Impairment		Emotional Disturbance		Orthopedic Impairment		Other Health Impairment	
			Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM
English–Language Arts	I	8	0.90	3.81	0.93	2.99	0.89	3.93	0.92	3.55
	II	8	0.93	2.23	0.80	2.13	0.91	2.23	0.85	2.32
	III	8	0.92	2.07	0.76	2.10	0.90	2.15	0.86	2.15
	IV	8	0.92	2.25	0.85	2.37	0.89	2.37	0.88	2.34
	V	8	0.93	2.18	0.84	2.29	0.89	2.28	0.87	2.19
Mathematics	I	8	0.88	4.04	0.92	3.47	0.87	4.05	0.88	4.02
	II	8	0.90	2.64	0.77	2.47	0.87	2.57	0.85	2.51
	III	8	0.85	2.58	0.71	2.63	0.78	2.76	0.74	2.79
	IV	8	0.90	2.45	0.84	2.50	0.83	2.66	0.82	2.64
	V	8	0.92	2.66	0.84	2.65	0.87	2.75	0.84	2.72
Science *	I	8	0.89	4.06	–	–	0.88	3.91	0.91	3.76
	III	8	0.76	2.55	0.76	2.35	0.84	2.35	0.78	2.46
	IV	8	0.90	2.39	0.82	2.51	0.79	2.59	0.75	2.38
	V	8	0.91	2.37	0.89	2.20	0.84	2.29	0.84	2.13
Content Area	Level	No. of Tasks	Specific Learning Disability		Deaf-Blindness		Multiple Disabilities		Autism	
			Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM
English–Language Arts *	I	8	0.90	2.68	0.90	3.49	0.89	3.92	0.82	4.09
	II	8	0.78	2.19	–	–	0.88	2.26	0.88	2.34
	III	8	0.75	2.01	–	–	0.90	2.27	0.92	2.18
	IV	8	0.75	2.24	–	–	0.91	2.26	0.91	2.28
	V	8	0.77	2.06	–	–	0.91	2.21	0.90	2.31
Mathematics *	I	8	0.89	3.21	0.91	3.28	0.87	4.03	0.78	4.37
	II	8	0.76	2.39	–	–	0.86	2.47	0.86	2.66
	III	8	0.66	2.56	–	–	0.85	2.53	0.84	2.67
	IV	8	0.66	2.46	–	–	0.86	2.61	0.84	2.72
	V	8	0.67	2.58	–	–	0.88	2.74	0.86	2.80
Science *	I	8	0.80	2.68	–	–	0.87	3.97	0.83	4.22
	III	8	0.74	2.26	–	–	0.88	2.43	0.87	2.44
	IV	8	0.59	2.29	–	–	0.83	2.48	0.83	2.49
	V	8	0.71	2.14	–	–	0.87	2.15	0.87	2.31

Content Area	Level	No. of Tasks	Traumatic Brain Injury		Unknown Disability	
			Reliab.	SEM	Reliab.	SEM
English– Language Arts	I	8	0.91	3.70	0.90	3.92
	II	8	0.92	1.97	0.84	2.44
	III	8	0.79	2.09	0.89	2.21
	IV	8	0.85	2.45	0.92	2.21
	V	8	0.85	2.28	0.87	2.20
Mathematics	I	8	0.91	3.70	0.87	4.17
	II	8	0.85	2.65	0.85	2.58
	III	8	0.70	2.77	0.79	2.70
	IV	8	0.84	2.65	0.82	2.79
	V	8	0.86	2.59	0.84	2.75
Science	I	8	0.91	3.58	0.89	4.12
	III	8	0.86	2.22	0.84	2.55
	IV	8	0.79	2.08	0.83	2.24
	V	8	0.73	2.24	0.88	2.20

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

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

	Placement Score	Far Below Basic	Below Basic	Basic	Proficient	Advanced	Category Total †
Decision Accuracy	0 – 4	0.03	0.02	0.01	0.00	0.00	0.06
	5 – 8	0.01	0.02	0.02	0.01	0.00	0.06
	9 – 13	0.00	0.01	0.03	0.03	0.00	0.08
All-forms Average *	14 – 23	0.00	0.00	0.03	0.14	0.05	0.22
	24 – 40	0.00	0.00	0.00	0.05	0.54	0.59
Estimated Proportion Correctly Classified: Total = 0.77, Proficient & Above = 0.93							
Decision Consistency	0 – 4	0.03	0.02	0.01	0.00	0.00	0.06
	5 – 8	0.02	0.02	0.02	0.01	0.00	0.06
	9 – 13	0.01	0.01	0.02	0.03	0.00	0.08
Alternate Form *	14 – 23	0.00	0.01	0.03	0.11	0.06	0.22
	24 – 40	0.00	0.00	0.01	0.06	0.52	0.59
Estimated Proportion Correctly Classified: Total = 0.70, Proficient & Above = 0.90							

* Values in table are proportions of the total sample.

† Inconsistencies with category cell entries are due to rounding.

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

	Placement Score	Far Below Basic	Below Basic	Basic	Proficient	Advanced	Category Total †
Decision Accuracy	0 – 4	0.03	0.03	0.01	0.00	0.00	0.06
	5 – 9	0.01	0.04	0.04	0.00	0.00	0.08
	10 – 18	0.00	0.02	0.12	0.05	0.00	0.19
All-forms Average *	19 – 28	0.00	0.01	0.06	0.21	0.06	0.33
	29 – 40	0.00	0.00	0.00	0.05	0.28	0.34
Estimated Proportion Correctly Classified: Total = 0.68, Proficient & Above = 0.88							
Decision Consistency	0 – 4	0.03	0.02	0.01	0.00	0.00	0.06
	5 – 9	0.02	0.03	0.03	0.00	0.00	0.08
	10 – 18	0.01	0.03	0.10	0.06	0.00	0.19
Alternate Form *	19 – 28	0.00	0.01	0.07	0.16	0.08	0.33
	29 – 40	0.00	0.00	0.00	0.07	0.27	0.34
Estimated Proportion Correctly Classified: Total = 0.59, Proficient & Above = 0.85							

* Values in table are proportions of the total sample.

† Inconsistencies with category cell entries are due to rounding.

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

	Placement Score	Far Below Basic	Below Basic	Basic	Proficient	Advanced	Category Total †
Decision Accuracy	0 – 4	0.04	0.03	0.00	0.00	0.00	0.07
	5 – 10	0.01	0.06	0.04	0.00	0.00	0.10
	11 – 19	0.00	0.02	0.12	0.05	0.00	0.19
All-forms Average *	20 – 29	0.00	0.00	0.05	0.19	0.05	0.30
	30 – 40	0.00	0.00	0.00	0.05	0.29	0.34
Estimated Proportion Correctly Classified: Total = 0.70, Proficient & Above = 0.90							
Decision Consistency	0 – 4	0.04	0.02	0.01	0.00	0.00	0.07
	5 – 10	0.02	0.04	0.03	0.00	0.00	0.10
	11 – 19	0.01	0.03	0.10	0.06	0.00	0.19
Alternate Form *	20 – 29	0.00	0.01	0.06	0.15	0.07	0.30
	30 – 40	0.00	0.00	0.00	0.06	0.28	0.34
Estimated Proportion Correctly Classified: Total = 0.60, Proficient & Above = 0.86							

* Values in table are proportions of the total sample.

† Inconsistencies with category cell entries are due to rounding.

Table 8.B.10 Decision Accuracy and Decision Consistency: Level II, ELA

	Placement Score	Far Below Basic	Below Basic	Basic	Proficient	Advanced	Category Total †
Decision Accuracy	0 – 3	0.00	0.01	0.00	0.00	0.00	0.02
	4 – 9	0.00	0.04	0.01	0.00	0.00	0.05
	10 – 13	0.00	0.02	0.07	0.04	0.00	0.13
All-forms Average *	14 – 20	0.00	0.00	0.04	0.27	0.06	0.37
	21 – 32	0.00	0.00	0.00	0.05	0.38	0.43
	Estimated Proportion Correctly Classified: Total = 0.76, Proficient & Above = 0.91						
Decision Consistency	0 – 3	0.01	0.01	0.00	0.00	0.00	0.02
	4 – 9	0.00	0.04	0.01	0.00	0.00	0.05
	10 – 13	0.00	0.03	0.05	0.05	0.00	0.13
Alternate Form *	14 – 20	0.00	0.01	0.05	0.22	0.08	0.37
	21 – 32	0.00	0.00	0.00	0.07	0.36	0.43
	Estimated Proportion Correctly Classified: Total = 0.68, Proficient & Above = 0.89						

* Values in table are proportions of the total sample.

† Inconsistencies with category cell entries are due to rounding.

Table 8.B.11 Decision Accuracy and Decision Consistency: Level II, Mathematics

	Placement Score	Far Below Basic	Below Basic	Basic	Proficient	Advanced	Category Total †
Decision Accuracy	0 – 6	0.01	0.01	0.00	0.00	0.00	0.03
	7 – 12	0.01	0.09	0.04	0.01	0.00	0.15
	13 – 16	0.00	0.03	0.08	0.06	0.01	0.17
All-forms Average *	17 – 23	0.00	0.00	0.03	0.23	0.03	0.30
	24 – 32	0.00	0.00	0.01	0.07	0.28	0.35
	Estimated Proportion Correctly Classified: Total = 0.69, Proficient & Above = 0.89						
Decision Consistency	0 – 6	0.02	0.01	0.00	0.00	0.00	0.03
	7 – 12	0.02	0.08	0.04	0.02	0.00	0.15
	13 – 16	0.00	0.04	0.06	0.06	0.01	0.17
Alternate Form *	17 – 23	0.00	0.01	0.05	0.19	0.06	0.30
	24 – 32	0.00	0.00	0.01	0.08	0.26	0.35
	Estimated Proportion Correctly Classified: Total = 0.60, Proficient & Above = 0.85						

* Values in table are proportions of the total sample.

† Inconsistencies with category cell entries are due to rounding.

Table 8.B.12 Decision Accuracy and Decision Consistency: Level III, ELA

	Placement Score	Far Below Basic	Below Basic	Basic	Proficient	Advanced	Category Total †
Decision Accuracy	0 – 3	0.00	0.01	0.00	0.00	0.00	0.02
	4 – 8	0.00	0.03	0.01	0.00	0.00	0.04
	9 – 13	0.00	0.01	0.08	0.03	0.00	0.12
All-forms Average *	14 – 20	0.00	0.00	0.02	0.21	0.04	0.27
	21 – 32	0.00	0.00	0.01	0.06	0.48	0.54
Estimated Proportion Correctly Classified: Total = 0.80, Proficient & Above = 0.93							
Decision Consistency	0 – 3	0.01	0.01	0.00	0.00	0.00	0.02
	4 – 8	0.00	0.03	0.01	0.00	0.00	0.04
	9 – 13	0.00	0.02	0.06	0.04	0.00	0.12
Alternate Form *	14 – 20	0.00	0.00	0.04	0.18	0.06	0.27
	21 – 32	0.00	0.00	0.01	0.07	0.46	0.54
Estimated Proportion Correctly Classified: Total = 0.73, Proficient & Above = 0.91							

* Values in table are proportions of the total sample.

† Inconsistencies with category cell entries are due to rounding.

Table 8.B.13 Decision Accuracy and Decision Consistency: Level III, Mathematics

	Placement Score	Far Below Basic	Below Basic	Basic	Proficient	Advanced	Category Total †
Decision Accuracy	0 – 3	0.00	0.01	0.00	0.00	0.00	0.01
	4 – 10	0.00	0.04	0.03	0.00	0.00	0.07
	11 – 15	0.00	0.02	0.12	0.07	0.00	0.22
All-forms Average *	16 – 24	0.00	0.00	0.05	0.42	0.04	0.51
	25 – 32	0.00	0.00	0.00	0.06	0.14	0.20
Estimated Proportion Correctly Classified: Total = 0.72, Proficient & Above = 0.88							
Decision Consistency	0 – 3	0.00	0.00	0.00	0.00	0.00	0.01
	4 – 10	0.00	0.04	0.03	0.00	0.00	0.07
	11 – 15	0.00	0.04	0.09	0.08	0.00	0.22
Alternate Form *	16 – 24	0.00	0.01	0.07	0.35	0.07	0.51
	25 – 32	0.00	0.00	0.00	0.07	0.13	0.20
Estimated Proportion Correctly Classified: Total = 0.62, Proficient & Above = 0.83							

* Values in table are proportions of the total sample.

† Inconsistencies with category cell entries are due to rounding.

Table 8.B.14 Decision Accuracy and Decision Consistency: Level III, Science

	Placement Score	Far Below Basic	Below Basic	Basic	Proficient	Advanced	Category Total †
Decision Accuracy	0 – 3	0.00	0.01	0.00	0.00	0.00	0.01
	4 – 9	0.00	0.02	0.01	0.00	0.00	0.03
	10 – 17	0.00	0.01	0.17	0.06	0.00	0.25
All-forms Average *	18 – 26	0.00	0.00	0.04	0.45	0.04	0.53
	27 – 32	0.00	0.00	0.00	0.07	0.11	0.18
Estimated Proportion Correctly Classified: Total = 0.75, Proficient & Above = 0.90							
Decision Consistency	0 – 3	0.00	0.00	0.00	0.00	0.00	0.01
	4 – 9	0.00	0.02	0.01	0.00	0.00	0.03
	10 – 17	0.00	0.02	0.15	0.07	0.00	0.25
Alternate Form *	18 – 26	0.00	0.00	0.07	0.39	0.08	0.53
	27 – 32	0.00	0.00	0.00	0.07	0.10	0.18
Estimated Proportion Correctly Classified: Total = 0.66, Proficient & Above = 0.86							

* Values in table are proportions of the total sample.

† Inconsistencies with category cell entries are due to rounding.

Table 8.B.15 Decision Accuracy and Decision Consistency: Level IV, ELA

	Placement Score	Far Below Basic	Below Basic	Basic	Proficient	Advanced	Category Total †
Decision Accuracy	0 – 4	0.01	0.01	0.00	0.00	0.00	0.02
	5 – 9	0.00	0.06	0.03	0.00	0.00	0.09
	10 – 13	0.00	0.03	0.08	0.05	0.01	0.17
All-forms Average *	14 – 21	0.00	0.00	0.03	0.26	0.03	0.32
	22 – 32	0.00	0.00	0.00	0.05	0.34	0.40
Estimated Proportion Correctly Classified: Total = 0.75, Proficient & Above = 0.91							
Decision Consistency	0 – 4	0.01	0.01	0.00	0.00	0.00	0.02
	5 – 9	0.01	0.05	0.03	0.00	0.00	0.09
	10 – 13	0.00	0.04	0.06	0.06	0.01	0.17
Alternate Form *	14 – 21	0.00	0.00	0.04	0.22	0.06	0.32
	22 – 32	0.00	0.00	0.00	0.07	0.32	0.40
Estimated Proportion Correctly Classified: Total = 0.67, Proficient & Above = 0.88							

* Values in table are proportions of the total sample.

† Inconsistencies with category cell entries are due to rounding.

Table 8.B.16 Decision Accuracy and Decision Consistency: Level IV, Mathematics

	Placement Score	Far Below Basic	Below Basic	Basic	Proficient	Advanced	Category Total †
Decision Accuracy	0 – 5	0.01	0.01	0.00	0.00	0.00	0.02
	6 – 11	0.00	0.06	0.04	0.00	0.00	0.10
	12 – 17	0.00	0.02	0.14	0.06	0.00	0.22
All-forms Average *	18 – 24	0.00	0.00	0.05	0.28	0.05	0.39
	25 – 32	0.00	0.00	0.00	0.06	0.21	0.27
Estimated Proportion Correctly Classified: Total = 0.69, Proficient & Above = 0.88							
Decision Consistency	0 – 5	0.01	0.01	0.00	0.00	0.00	0.02
	6 – 11	0.01	0.05	0.04	0.00	0.00	0.10
	12 – 17	0.00	0.04	0.11	0.07	0.00	0.22
Alternate Form *	18 – 24	0.00	0.01	0.07	0.23	0.08	0.39
	25 – 32	0.00	0.00	0.00	0.08	0.19	0.27
Estimated Proportion Correctly Classified: Total = 0.59, Proficient & Above = 0.84							

* Values in table are proportions of the total sample.

† Inconsistencies with category cell entries are due to rounding.

Table 8.B.17 Decision Accuracy and Decision Consistency: Level IV, Science

	Placement Score	Far Below Basic	Below Basic	Basic	Proficient	Advanced	Category Total †
Decision Accuracy	0 – 3	0.00	0.00	0.00	0.00	0.00	0.01
	4 – 11	0.00	0.03	0.01	0.00	0.00	0.04
	12 – 19	0.00	0.02	0.20	0.07	0.00	0.28
All-forms Average *	20 – 27	0.00	0.00	0.05	0.44	0.03	0.52
	28 – 32	0.00	0.00	0.00	0.08	0.07	0.14
Estimated Proportion Correctly Classified: Total = 0.73, Proficient & Above = 0.88							
Decision Consistency	0 – 3	0.00	0.00	0.00	0.00	0.00	0.01
	4 – 11	0.00	0.03	0.02	0.00	0.00	0.04
	12 – 19	0.00	0.03	0.17	0.08	0.00	0.28
Alternate Form *	20 – 27	0.00	0.00	0.08	0.37	0.07	0.52
	28 – 32	0.00	0.00	0.00	0.08	0.07	0.14
Estimated Proportion Correctly Classified: Total = 0.63, Proficient & Above = 0.84							

* Values in table are proportions of the total sample.

† Inconsistencies with category cell entries are due to rounding.

Table 8.B.18 Decision Accuracy and Decision Consistency: Level V, ELA

	Placement Score	Far Below Basic	Below Basic	Basic	Proficient	Advanced	Category Total †
Decision Accuracy	0 – 3	0.00	0.01	0.00	0.00	0.00	0.02
	4 – 8	0.00	0.02	0.01	0.00	0.00	0.03
	9 – 14	0.00	0.01	0.10	0.04	0.00	0.16
All-forms Average *	15 – 22	0.00	0.00	0.03	0.27	0.05	0.36
	23 – 32	0.00	0.00	0.00	0.05	0.39	0.44
Estimated Proportion Correctly Classified: Total = 0.79, Proficient & Above = 0.93							
Decision Consistency	0 – 3	0.01	0.01	0.00	0.00	0.00	0.02
	4 – 8	0.00	0.02	0.01	0.00	0.00	0.03
	9 – 14	0.00	0.02	0.08	0.05	0.00	0.16
Alternate Form *	15 – 22	0.00	0.00	0.05	0.23	0.07	0.36
	23 – 32	0.00	0.00	0.00	0.07	0.37	0.44
Estimated Proportion Correctly Classified: Total = 0.71, Proficient & Above = 0.90							

* Values in table are proportions of the total sample.

† Inconsistencies with category cell entries are due to rounding.

Table 8.B.19 Decision Accuracy and Decision Consistency: Level V, Mathematics

	Placement Score	Far Below Basic	Below Basic	Basic	Proficient	Advanced	Category Total †
Decision Accuracy	0 – 5	0.01	0.01	0.00	0.00	0.00	0.02
	6 – 10	0.01	0.04	0.04	0.00	0.00	0.09
	11 – 17	0.00	0.02	0.13	0.05	0.00	0.20
All-forms Average *	18 – 25	0.00	0.00	0.05	0.27	0.05	0.36
	26 – 32	0.00	0.00	0.00	0.06	0.26	0.33
Estimated Proportion Correctly Classified: Total = 0.72, Proficient & Above = 0.90							
Decision Consistency	0 – 5	0.01	0.01	0.00	0.00	0.00	0.02
	6 – 10	0.01	0.03	0.03	0.01	0.00	0.09
	11 – 17	0.00	0.03	0.11	0.06	0.00	0.20
Alternate Form *	18 – 25	0.00	0.00	0.06	0.22	0.08	0.36
	26 – 32	0.00	0.00	0.00	0.07	0.25	0.33
Estimated Proportion Correctly Classified: Total = 0.62, Proficient & Above = 0.86							

* Values in table are proportions of the total sample.

† Inconsistencies with category cell entries are due to rounding.

Table 8.B.20 Decision Accuracy and Decision Consistency: Level V, Science

	Placement Score	Far Below Basic	Below Basic	Basic	Proficient	Advanced	Category Total †
Decision Accuracy	0 – 2	0.00	0.00	0.00	0.00	0.00	0.01
	3 – 11	0.00	0.03	0.02	0.00	0.00	0.05
	12 – 19	0.00	0.00	0.19	0.05	0.00	0.24
All-forms Average *	20 – 26	0.00	0.01	0.06	0.35	0.05	0.47
	27 – 32	0.00	0.00	0.00	0.05	0.18	0.23
Estimated Proportion Correctly Classified: Total = 0.75, Proficient & Above = 0.88							
Decision Consistency	0 – 2	0.00	0.00	0.00	0.00	0.00	0.01
	3 – 11	0.00	0.03	0.02	0.00	0.00	0.05
	12 – 19	0.00	0.02	0.16	0.06	0.00	0.24
Alternate Form *	20 – 26	0.00	0.01	0.08	0.30	0.08	0.47
	27 – 32	0.00	0.00	0.00	0.06	0.17	0.23
Estimated Proportion Correctly Classified: Total = 0.66, Proficient & Above = 0.84							

* Values in table are proportions of the total sample.

† Inconsistencies with category cell entries are due to rounding.

Appendix 8.C—Validity Analyses

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

Table 8.C.1 CAPA Content Area Correlations by Gender: Level I

	Male			Female			Unknown		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	8,933	0.80	0.80	5,083	0.80	0.82	82	0.86	0.85
Mathematics	8,916	8,919	0.79	5,063	5,066	0.81	80	80	0.84
Science	2,245	2,246	2,246	1,304	1,305	1,305	13	13	13

Table 8.C.2 CAPA Content Area Correlations by Gender: Level II

	Male			Female			Unknown		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	4,601	0.74	N/A	2,023	0.73	N/A	44	0.78	N/A
Mathematics	4,588	4,589	N/A	2,018	2,019	N/A	42	42	N/A
Science	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 8.C.3 CAPA Content Area Correlations by Gender: Level III

	Male			Female			Unknown		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	4,770	0.77	0.74	2,288	0.77	0.74	47	0.82	0.58
Mathematics	4,759	4,762	0.75	2,283	2,286	0.73	46	46	0.68
Science	2,387	2,387	2,387	1,153	1,152	1,154	15	15	15

Table 8.C.4 CAPA Content Area Correlations by Gender: Level IV

	Male			Female			Unknown		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	6,592	0.76	0.68	3,452	0.75	0.68	47	0.78	–
Mathematics	6,566	6,575	0.66	3,443	3,446	0.66	47	47	–
Science	2,164	2,167	2,168	1,121	1,121	1,122	9	9	9

Table 8.C.5 CAPA Content Area Correlations by Gender: Level V

	Male			Female			Unknown		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	6,626	0.70	0.67	3,737	0.71	0.71	61	0.62	0.78
Mathematics	6,593	6,601	0.66	3,723	3,730	0.66	61	61	0.87
Science	2,151	2,151	2,153	1,252	1,254	1,254	17	17	17

Table 8.C.6 CAPA Content Area Correlations by Ethnicity: Level I

	American Indian			Asian			Pacific Islander			Filipino		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	91	0.76	0.87	1,181	0.79	0.82	84	0.83	0.88	451	0.82	0.81
Mathematics	90	90	0.88	1,177	1,177	0.78	83	83	0.80	450	450	0.83
Science	22	22	22	274	274	274	19	19	19	131	131	131
	Hispanic			African American			White			Unknown Ethnicity		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	7,547	0.81	0.80	1,256	0.80	0.78	3,180	0.79	0.82	308	0.80	0.79
Mathematics	7,532	7,535	0.80	1,251	1,252	0.79	3,172	3,174	0.80	304	304	0.85
Science	1,883	1,885	1,885	323	323	323	845	845	845	65	65	65

Table 8.C.7 CAPA Content Area Correlations by Ethnicity: Level II

	American Indian			Asian			Pacific Islander			Filipino		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	39	0.56	N/A	489	0.70	N/A	33	0.74	N/A	209	0.66	N/A
Mathematics	39	39	N/A	488	488	N/A	33	33	N/A	209	209	N/A
Science	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	Hispanic			African American			White			Unknown Ethnicity		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	3,640	0.74	N/A	581	0.75	N/A	1,475	0.74	N/A	202	0.78	N/A
Mathematics	3,629	3,630	N/A	580	581	N/A	1,468	1,468	N/A	202	202	N/A
Science	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 8.C.8 CAPA Content Area Correlations by Ethnicity: Level III

	American Indian			Asian			Pacific Islander			Filipino		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	65	0.74	0.77	515	0.79	0.73	44	0.78	0.59	188	0.78	0.82
Mathematics	65	65	0.61	514	514	0.70	44	44	0.54	187	187	0.87
Science	32	32	32	249	249	249	22	22	22	99	99	99
	Hispanic			African American			White			Unknown Ethnicity		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	3,775	0.76	0.71	722	0.78	0.79	1,596	0.77	0.76	200	0.78	0.65
Mathematics	3,766	3,770	0.74	721	722	0.78	1,591	1,592	0.74	200	200	0.74
Science	1,893	1,893	1,894	364	364	364	796	795	796	100	100	100

Table 8.C.9 CAPA Content Area Correlations by Ethnicity: Level IV

	American Indian			Asian			Pacific Islander			Filipino		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	86	0.71	0.70	669	0.75	0.72	58	0.80	0.81	338	0.75	0.80
Mathematics	86	86	0.58	667	667	0.66	57	57	0.82	336	338	0.73
Science	34	34	34	217	216	217	18	18	18	116	117	117
	Hispanic			African American			White			Unknown Ethnicity		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	5,193	0.75	0.67	1,034	0.77	0.61	2,434	0.77	0.68	279	0.75	0.70
Mathematics	5,176	5,185	0.66	1,028	1,028	0.62	2,428	2,429	0.67	278	278	0.69
Science	1,649	1,651	1,652	363	363	363	798	799	799	99	99	99

Table 8.C.10 CAPA Content Area Correlations by Ethnicity: Level V

	American Indian			Asian			Pacific Islander			Filipino		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	98	0.64	0.71	720	0.73	0.73	54	0.77	0.71	295	0.78	0.69
Mathematics	98	99	0.78	715	715	0.70	53	53	0.63	292	292	0.76
Science	37	38	38	230	230	230	21	21	21	103	103	103
	Hispanic			African American			White			Unknown Ethnicity		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	5,056	0.69	0.68	1,146	0.70	0.69	2,816	0.70	0.64	239	0.68	0.85
Mathematics	5,033	5,043	0.66	1,138	1,139	0.67	2,809	2,812	0.66	239	239	0.72
Science	1,674	1,675	1,676	368	368	369	921	921	921	66	66	66

Table 8.C.11 CAPA Content Area Correlations by Ethnicity for Economically Disadvantaged: Level I

	American Indian			Asian			Pacific Islander			Filipino		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	56	0.70	–	496	0.76	0.81	50	0.81	0.88	156	0.79	0.79
Mathematics	55	55	–	494	494	0.73	49	49	0.87	155	155	0.77
Science	8	8	8	124	124	124	13	13	13	45	45	45
	Hispanic			African American			White			Unknown Ethnicity		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	5,724	0.79	0.78	819	0.79	0.79	1,070	0.80	0.83	118	0.77	0.71
Mathematics	5,714	5,716	0.76	817	818	0.77	1,067	1,068	0.79	117	117	0.77
Science	1,394	1,396	1,396	213	213	213	293	293	293	23	23	23

Table 8.C.12 CAPA Content Area Correlations by Ethnicity for Economically Disadvantaged: Level II

	American Indian			Asian			Pacific Islander			Filipino		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	30	0.59	N/A	236	0.73	N/A	19	0.74	N/A	70	0.55	N/A
Mathematics	30	30	N/A	236	236	N/A	19	19	N/A	70	70	N/A
Science	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	Hispanic			African American			White			Unknown Ethnicity		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	3,003	0.74	N/A	444	0.77	N/A	644	0.74	N/A	93	0.80	N/A
Mathematics	2,994	2,994	N/A	443	444	N/A	643	643	N/A	93	93	N/A
Science	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 8.C.13 CAPA Content Area Correlations by Ethnicity for Economically Disadvantaged: Level III

	American Indian			Asian			Pacific Islander			Filipino		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	48	0.71	0.61	228	0.75	0.70	26	0.83	0.68	76	0.78	0.80
Mathematics	48	48	0.37	227	227	0.62	26	26	0.64	75	75	0.83
Science	24	24	24	108	108	108	12	12	12	37	37	37
	Hispanic			African American			White			Unknown Ethnicity		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	3,170	0.76	0.71	533	0.76	0.78	694	0.78	0.74	84	0.81	0.65
Mathematics	3,163	3,167	0.72	532	533	0.77	693	694	0.71	84	84	0.73
Science	1,608	1,609	1,609	263	263	263	344	344	344	45	45	45

Table 8.C.14 CAPA Content Area Correlations by Ethnicity for Economically Disadvantaged: Level IV

	American Indian			Asian			Pacific Islander			Filipino		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	60	0.75	0.71	309	0.76	0.78	38	0.80	0.79	127	0.64	0.63
Mathematics	60	60	0.56	308	308	0.70	38	38	0.77	127	129	0.43
Science	25	25	25	101	100	101	14	14	14	40	41	41
	Hispanic			African American			White			Unknown Ethnicity		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	4,352	0.74	0.66	750	0.77	0.62	1,041	0.76	0.64	122	0.71	0.60
Mathematics	4,338	4,347	0.65	744	744	0.64	1,041	1,042	0.65	122	122	0.65
Science	1,394	1,397	1,397	265	265	265	322	323	323	50	50	50

Table 8.C.15 CAPA Content Area Correlations by Ethnicity for Economically Disadvantaged: Level V

	American Indian			Asian			Pacific Islander			Filipino		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	64	0.60	0.77	361	0.74	0.64	35	0.81	0.67	116	0.71	0.64
Mathematics	64	65	0.79	358	358	0.66	34	34	0.77	115	115	0.74
Science	20	21	21	108	108	108	13	13	13	40	40	40
	Hispanic			African American			White			Unknown Ethnicity		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	4,083	0.68	0.67	803	0.69	0.66	1,114	0.68	0.65	85	0.58	0.86
Mathematics	4,067	4,076	0.66	799	800	0.61	1,111	1,113	0.64	85	85	0.65
Science	1,351	1,352	1,353	251	251	251	378	378	378	23	23	23

Table 8.C.16 CAPA Content Area Correlations by Ethnicity for Not Economically Disadvantaged: Level I

	American Indian			Asian			Pacific Islander			Filipino		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	35	0.88	0.92	666	0.81	0.83	31	0.86	–	286	0.82	0.82
Mathematics	35	35	0.88	664	664	0.82	31	31	–	286	286	0.86
Science	14	14	14	147	147	147	6	6	6	84	84	84
	Hispanic			African American			White			Unknown Ethnicity		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	1,734	0.84	0.84	415	0.82	0.76	2,045	0.79	0.83	164	0.80	0.78
Mathematics	1,729	1,729	0.86	412	412	0.82	2,041	2,042	0.80	164	164	0.88
Science	469	469	469	109	109	109	536	536	536	39	39	39

Table 8.C.17 CAPA Content Area Correlations by Ethnicity for Not Economically Disadvantaged: Level II

	American Indian			Asian			Pacific Islander			Filipino		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	9	–	N/A	234	0.67	N/A	14	0.72	N/A	132	0.70	N/A
Mathematics	9	9	N/A	233	233	N/A	14	14	N/A	132	132	N/A
Science	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	Hispanic			African American			White			Unknown Ethnicity		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	560	0.74	N/A	125	0.70	N/A	789	0.72	N/A	84	0.76	N/A
Mathematics	559	560	N/A	125	125	N/A	784	784	N/A	84	84	N/A
Science	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 8.C.18 CAPA Content Area Correlations by Ethnicity for Not Economically Disadvantaged: Level III

	American Indian			Asian			Pacific Islander			Filipino		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	15	0.85	–	282	0.82	0.74	17	0.73	–	107	0.79	0.83
Mathematics	15	15	–	282	282	0.74	17	17	–	107	107	0.89
Science	7	7	7	140	140	140	9	9	9	57	57	57
	Hispanic			African American			White			Unknown Ethnicity		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	546	0.76	0.72	177	0.81	0.81	876	0.75	0.77	90	0.78	0.75
Mathematics	546	546	0.79	177	177	0.82	872	872	0.76	90	90	0.84
Science	264	264	264	92	92	92	441	440	441	44	44	44

Table 8.C.19 CAPA Content Area Correlations by Ethnicity for Not Economically Disadvantaged: Level IV

	American Indian			Asian			Pacific Islander			Filipino		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	26	0.64	–	348	0.75	0.67	19	0.80	–	204	0.79	0.83
Mathematics	26	26	–	347	347	0.61	18	18	–	202	202	0.78
Science	9	9	9	111	111	111	4	4	4	74	74	74
	Hispanic			African American			White			Unknown Ethnicity		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	763	0.75	0.70	265	0.78	0.56	1,350	0.77	0.69	104	0.74	0.87
Mathematics	763	763	0.69	265	265	0.55	1,344	1,344	0.69	103	103	0.83
Science	231	231	231	93	93	93	457	457	457	33	33	33

Table 8.C.20 CAPA Content Area Correlations by Ethnicity for Not Economically Disadvantaged: Level V

	American Indian			Asian			Pacific Islander			Filipino		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	32	0.69	0.62	341	0.73	0.79	18	0.82	–	171	0.82	0.73
Mathematics	32	32	0.75	339	339	0.72	18	18	–	169	169	0.78
Science	17	17	17	115	115	115	7	7	7	62	62	62
	Hispanic			African American			White			Unknown Ethnicity		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	880	0.74	0.72	317	0.73	0.73	1,632	0.72	0.62	99	0.73	0.81
Mathematics	875	875	0.66	313	313	0.77	1,628	1,629	0.65	99	99	0.69
Science	293	293	293	109	109	109	526	526	526	28	28	28

Table 8.C.21 CAPA Content Area Correlations by Ethnicity for Unknown Economic Status: Level I

	American Indian			Asian			Pacific Islander			Filipino		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	0	–	–	19	0.96	–	3	–	–	9	–	–
Mathematics	0	0	–	19	19	–	3	3	–	9	9	–
Science	0	0	0	3	3	3	0	0	0	2	2	2
	Hispanic			African American			White			Unknown Ethnicity		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	89	0.84	0.83	22	0.84	–	65	0.82	0.68	26	0.93	–
Mathematics	89	90	0.88	22	22	–	64	64	0.83	23	23	–
Science	20	20	20	1	1	1	16	16	16	3	3	3

Table 8.C.22 CAPA Content Area Correlations by Ethnicity for Unknown Economic Status: Level II

	American Indian			Asian			Pacific Islander			Filipino		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	0	–	N/A	19	0.61	N/A	0	–	N/A	7	–	N/A
Mathematics	0	0	N/A	19	19	N/A	0	0	N/A	7	7	N/A
Science	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	Hispanic			African American			White			Unknown Ethnicity		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	77	0.62	N/A	12	0.78	N/A	42	0.68	N/A	25	0.76	N/A
Mathematics	76	76	N/A	12	12	N/A	41	41	N/A	25	25	N/A
Science	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 8.C.23 CAPA Content Area Correlations by Ethnicity for Unknown Economic Status: Level III

	American Indian			Asian			Pacific Islander			Filipino		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	2	–	–	5	–	–	1	–	–	5	–	–
Mathematics	2	2	–	5	5	–	1	1	–	5	5	–
Science	1	1	1	1	1	1	1	1	1	5	5	5
	Hispanic			African American			White			Unknown Ethnicity		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	59	0.57	0.71	12	0.82	–	26	0.86	0.95	26	0.77	0.49
Mathematics	57	57	0.64	12	12	–	26	26	0.86	26	26	0.66
Science	21	20	21	9	9	9	11	11	11	11	11	11

Table 8.C.24 CAPA Content Area Correlations by Ethnicity for Unknown Economic Status: Level IV

	American Indian			Asian			Pacific Islander			Filipino		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	0	–	–	12	0.73	–	1	–	–	7	–	–
Mathematics	0	0	–	12	12	–	1	1	–	7	7	–
Science	0	0	0	5	5	5	0	0	0	2	2	2
	Hispanic			African American			White			Unknown Ethnicity		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	78	0.70	0.39	19	0.51	–	43	0.82	0.68	53	0.87	0.66
Mathematics	75	75	0.55	19	19	–	43	43	0.52	53	53	0.57
Science	24	23	24	5	5	5	19	19	19	16	16	16

Table 8.C.25 CAPA Content Area Correlations by Ethnicity for Unknown Economic Status: Level V

	American Indian			Asian			Pacific Islander			Filipino		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	2	–	–	18	0.51	–	1	–	–	8	–	–
Mathematics	2	2	–	18	18	–	1	1	–	8	8	–
Science	0	0	0	7	7	7	1	1	1	1	1	1
	Hispanic			African American			White			Unknown Ethnicity		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	93	0.52	0.74	26	0.64	–	70	0.53	0.51	55	0.68	0.90
Mathematics	91	92	0.77	26	26	–	70	70	0.85	55	55	0.92
Science	30	30	30	8	8	9	17	17	17	15	15	15

Table 8.C.26 CAPA Content Area Correlations by Economic Status: Level I

	Disadvantaged			Not Disadvantaged			Unknown Status		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	8,489	0.79	0.79	5,376	0.82	0.83	233	0.86	0.78
Mathematics	8,468	8,472	0.77	5,362	5,363	0.83	229	230	0.87
Science	2,113	2,115	2,115	1,404	1,404	1,404	45	45	45

Table 8.C.27 CAPA Content Area Correlations by Economic Status: Level II

	Disadvantaged			Not Disadvantaged			Unknown Status		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	4,539	0.74	N/A	1,947	0.72	N/A	182	0.66	N/A
Mathematics	4,528	4,529	N/A	1,940	1,941	N/A	180	180	N/A
Science	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 8.C.28 CAPA Content Area Correlations by Economic Status: Level III

	Disadvantaged			Not Disadvantaged			Unknown Status		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	4,859	0.77	0.73	2,110	0.77	0.75	136	0.70	0.72
Mathematics	4,848	4,854	0.72	2,106	2,106	0.78	134	134	0.72
Science	2,441	2,442	2,442	1,054	1,053	1,054	60	59	60

Table 8.C.29 CAPA Content Area Correlations by Economic Status: Level IV

	Disadvantaged			Not Disadvantaged			Unknown Status		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	6,799	0.75	0.66	3,079	0.76	0.70	213	0.75	0.56
Mathematics	6,778	6,790	0.65	3,068	3,068	0.68	210	210	0.48
Science	2,211	2,215	2,216	1,012	1,012	1,012	71	70	71

Table 8.C.30 CAPA Content Area Correlations by Economic Status: Level V

	Disadvantaged			Not Disadvantaged			Unknown Status		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	6,661	0.69	0.67	3,490	0.73	0.68	273	0.56	0.66
Mathematics	6,633	6,646	0.65	3,473	3,474	0.67	271	272	0.78
Science	2,184	2,186	2,187	1,157	1,157	1,157	79	79	80

Table 8.C.31 CAPA Content Area Correlations by Disability: Level I

	MR/ID		Hard of Hearing		Deafness		Speech Impairment					
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science			
ELA	5,305	0.77	0.77	72	0.82	0.83	47	0.71	0.68	149	0.67	—
Mathematics	5,298	5,300	0.76	72	72	0.89	47	47	0.83	149	149	—
Science	1,412	1,413	1,413	18	18	18	13	13	13	8	8	8
	Visual Impairment											
			Emotional Disturbance		Orthopedic Impairment		Other Health Impairment					
ELA	277	0.85	0.87	23	0.82	—	2,193	0.82	0.85	400	0.83	0.89
Mathematics	276	276	0.74	23	23	—	2,181	2,184	0.84	399	399	0.87
Science	73	73	73	4	4	4	601	602	602	89	89	89
	Specific Learning Disability											
			Deaf-Blindness		Multiple Disabilities		Autism					
ELA	84	0.72	0.62	28	0.91	—	1,527	0.84	0.85	3,770	0.74	0.71
Mathematics	84	84	0.48	28	28	—	1,522	1,523	0.84	3,760	3,760	0.73
Science	13	13	13	5	5	5	381	381	381	893	893	893
	Traumatic Brain Injury											
			Unknown Disability									
ELA	94	0.85	0.87	129	0.86	0.86						
Mathematics	94	94	0.83	126	126	0.80						
Science	28	28	28	24	24	24						

Table 8.C.32 CAPA Content Area Correlations by Disability: Level II

	MR/ID			Hard of Hearing			Deafness			Speech Impairment		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	2,036	0.74	N/A	33	0.59	N/A	37	0.66	N/A	642	0.65	N/A
Mathematics	2,032	2,033	N/A	33	33	N/A	37	37	N/A	642	642	N/A
Science	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Visual Impairment												
Emotional Disturbance												
ELA	28	0.82	N/A	23	0.72	N/A	225	0.83	N/A	365	0.75	N/A
Mathematics	28	28	N/A	23	23	N/A	225	226	N/A	364	364	N/A
Science	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Orthopedic Impairment												
Other Health Impairment												
Specific Learning Disability												
Deaf-Blindness												
Multiple Disabilities												
ELA	537	0.61	N/A	5	-	N/A	113	0.75	N/A	2,493	0.74	N/A
Mathematics	535	535	N/A	5	5	N/A	113	113	N/A	2,482	2,482	N/A
Science	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Traumatic Brain Injury												
Unknown Disability												
ELA	35	0.81	N/A	96	0.67	N/A						
Mathematics	35	35	N/A	94	94	N/A						
Science	N/A	N/A	N/A	N/A	N/A	N/A						

Table 8.C.33 CAPA Content Area Correlations by Disability: Level III

	MR/ID			Hard of Hearing			Deafness			Speech Impairment		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	2,597	0.76	0.69	45	0.65	0.60	56	0.88	0.75	397	0.66	0.56
Mathematics	2,592	2,593	0.69	44	44	0.74	56	56	0.68	396	398	0.55
Science	1,345	1,344	1,345	26	26	26	29	29	29	192	192	192
	Visual Impairment											
	Emotional Disturbance			Orthopedic Impairment			Other Health Impairment					
ELA	34	0.89	0.79	46	0.60	0.60	303	0.74	0.78	356	0.75	0.69
Mathematics	33	33	0.73	46	46	0.73	302	302	0.74	356	356	0.62
Science	18	18	18	26	26	26	143	143	143	160	160	160
	Specific Learning Disability											
	Deaf-Blindness			Multiple Disabilities			Autism					
ELA	623	0.54	0.43	1	–	–	141	0.85	0.88	2,389	0.79	0.79
Mathematics	621	621	0.62	1	1	–	141	141	0.91	2,384	2,386	0.78
Science	320	319	320	0	0	0	82	82	82	1,162	1,163	1,163
	Traumatic Brain Injury											
	Unknown Disability											
ELA	42	0.45	0.57	75	0.73	0.70						
Mathematics	42	43	0.73	74	74	0.73						
Science	22	22	22	30	30	30						

Table 8.C.34 CAPA Content Area Correlations by Disability: Level IV

	MR/ID			Hard of Hearing			Deafness			Speech Impairment		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	4,428	0.74	0.67	53	0.74	0.81	82	0.64	0.54	307	0.65	0.60
Mathematics	4,414	4,419	0.66	53	53	0.69	81	81	0.77	307	307	0.43
Science	1,535	1,537	1,538	17	17	17	23	23	23	80	80	80
	Visual Impairment											
	Emotional Disturbance			Orthopedic Impairment			Other Health Impairment					
ELA	67	0.85	0.78	113	0.66	0.69	509	0.75	0.69	436	0.71	0.69
Mathematics	67	67	0.77	111	111	0.72	509	509	0.64	432	433	0.59
Science	22	22	22	43	43	43	171	171	171	126	127	127
	Specific Learning Disability											
	Deaf-Blindness			Multiple Disabilities			Autism					
ELA	773	0.59	0.45	5	-	-	292	0.75	0.68	2,836	0.77	0.67
Mathematics	770	773	0.37	5	5	-	291	291	0.77	2,828	2,830	0.67
Science	249	249	250	0	0	0	109	109	109	860	860	860
	Traumatic Brain Injury											
	Unknown Disability											
ELA	53	0.74	0.64	137	0.80	0.76						
Mathematics	52	52	0.77	136	137	0.69						
Science	19	19	19	40	40	40						

Table 8.C.35 CAPA Content Area Correlations by Disability: Level V

	MR/ID			Hard of Hearing			Deafness			Speech Impairment		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	5,035	0.71	0.68	71	0.68	0.60	94	0.42	0.46	186	0.45	0.51
Mathematics	5,018	5,029	0.69	71	71	0.35	93	93	0.65	186	187	0.44
Science	1,654	1,656	1,658	30	30	30	30	30	30	71	71	71
	Visual Impairment											
	Emotional Disturbance			Orthopedic Impairment			Other Health Impairment					
ELA	72	0.80	0.83	141	0.62	0.42	501	0.75	0.76	489	0.65	0.56
Mathematics	72	72	0.87	139	139	0.63	498	498	0.67	488	489	0.53
Science	28	28	28	41	41	41	170	170	170	179	179	179
	Specific Learning Disability											
	Deaf-Blindness			Multiple Disabilities			Autism					
ELA	890	0.48	0.44	3	–	–	314	0.79	0.77	2,437	0.71	0.70
Mathematics	886	886	0.40	3	3	–	313	313	0.79	2,420	2,421	0.70
Science	303	303	303	1	1	1	101	101	101	754	754	754
	Traumatic Brain Injury											
	Unknown Disability											
ELA	66	0.58	0.78	125	0.59	0.82						
Mathematics	66	66	0.58	124	125	0.59						
Science	21	21	21	37	37	37						

Table 8.C.36 Interrater Agreement Analyses for Operational Tasks: Level I

Level I		First Rating			Second Rating			% Agreement			MAD *	Corr †
Content Area	Task	N	Mean	SD	N	Mean	SD	Exact	Adjacent	Neither		
English– Language Arts	1	2,082	3.33	1.90	2,082	3.32	1.91	96.88	2.50	0.63	0.05	0.99
	3	2,082	2.92	1.93	2,082	2.93	1.93	96.11	2.64	1.26	0.07	0.97
	4	2,082	3.81	1.77	2,082	3.79	1.78	95.82	2.64	1.53	0.08	0.96
	6	2,082	3.72	1.82	2,082	3.71	1.82	96.21	2.64	1.15	0.07	0.97
	7	2,082	2.98	1.88	2,082	2.97	1.89	94.67	3.55	1.77	0.10	0.96
	9	2,082	3.13	1.94	2,082	3.12	1.94	95.77	2.55	1.68	0.08	0.97
	10	2,082	3.26	1.94	2,082	3.26	1.94	96.11	2.88	1.01	0.07	0.97
	12	2,082	3.34	1.94	2,082	3.33	1.94	96.64	2.35	1.01	0.06	0.97
Mathematics	1	2,068	2.80	1.94	2,068	2.79	1.93	96.28	2.90	0.83	0.06	0.98
	3	2,068	2.73	1.89	2,068	2.72	1.89	95.79	3.05	1.16	0.07	0.98
	4	2,068	2.88	1.93	2,068	2.86	1.94	96.91	1.93	1.15	0.06	0.98
	6	2,068	2.99	1.96	2,068	2.99	1.96	95.55	2.71	1.74	0.09	0.96
	7	2,068	3.50	1.90	2,068	3.49	1.90	97.10	2.03	0.87	0.05	0.98
	9	2,068	3.47	1.88	2,068	3.47	1.88	96.81	2.13	1.06	0.06	0.97
	10	2,068	3.22	1.93	2,068	3.20	1.93	96.03	2.56	1.40	0.08	0.97
	12	2,068	2.81	1.98	2,068	2.82	1.97	94.97	2.95	2.08	0.12	0.95
Science	1	511	3.38	1.92	511	3.36	1.93	97.26	1.76	0.98	0.05	0.98
	3	511	3.28	1.95	511	3.29	1.95	96.67	1.96	1.38	0.06	0.98
	4	511	3.18	1.96	511	3.17	1.96	97.85	1.57	0.59	0.04	0.99
	6	511	3.38	1.96	511	3.40	1.95	97.26	1.57	1.17	0.06	0.98
	7	511	2.95	1.96	511	2.97	1.96	95.69	2.54	1.77	0.09	0.97
	9	511	2.97	1.95	511	2.99	1.95	96.87	1.37	1.76	0.08	0.96
	10	511	3.24	1.95	511	3.20	1.95	96.09	2.54	1.38	0.08	0.97
	12	511	2.70	1.99	511	2.70	1.99	94.13	2.94	2.94	0.15	0.93

* Mean absolute difference between first and second ratings

† Pearson correlation between first and second ratings

Table 8.C.37 Interrater Agreement Analyses for Operational Tasks: Level II

Level II Content Area	Task	First Rating			Second Rating			% Agreement			MAD *	Corr †
		N	Mean	SD	N	Mean	SD	Exact	Adjacent	Neither		
English– Language Arts	1	1,490	1.95	1.06	1,490	1.95	1.06	96.31	3.42	0.27	0.04	0.98
	3	1,490	3.61	0.79	1,490	3.60	0.80	96.44	3.09	0.47	0.04	0.95
	4	1,490	2.16	1.09	1,490	2.17	1.09	94.03	5.10	0.87	0.07	0.96
	6	1,490	2.68	1.21	1,490	2.68	1.22	96.17	3.56	0.27	0.04	0.98
	7	1,490	2.17	1.19	1,490	2.17	1.19	96.11	3.62	0.27	0.04	0.98
	9	1,490	2.34	1.09	1,490	2.34	1.10	94.97	4.63	0.41	0.06	0.97
	10	1,490	2.16	1.18	1,490	2.17	1.18	95.91	3.49	0.60	0.05	0.97
	12	1,490	1.87	1.04	1,490	1.88	1.05	94.56	4.70	0.73	0.07	0.95
Mathematics	1	1,483	3.01	1.15	1,483	3.02	1.14	97.91	1.89	0.20	0.02	0.99
	3	1,483	3.23	0.92	1,483	3.24	0.90	96.49	2.56	0.94	0.05	0.94
	4	1,483	1.98	1.37	1,483	1.98	1.37	96.29	2.76	0.94	0.05	0.97
	6	1,483	2.25	1.17	1,483	2.25	1.17	96.16	3.30	0.54	0.05	0.97
	7	1,483	2.75	1.39	1,483	2.76	1.38	97.57	1.89	0.54	0.03	0.98
	9	1,483	2.80	1.20	1,483	2.80	1.21	97.10	2.49	0.40	0.04	0.98
	10	1,483	2.41	1.37	1,483	2.41	1.37	97.98	1.62	0.40	0.03	0.99
	12	1,483	1.20	0.78	1,483	1.19	0.79	95.89	3.44	0.67	0.05	0.94

* Mean absolute difference between first and second ratings

† Pearson correlation between first and second ratings

Table 8.C.38 Interrater Agreement Analyses for Operational Tasks: Level III

Level III		First Rating			Second Rating			% Agreement			MAD *	Corr †
Content Area	Task	N	Mean	SD	N	Mean	SD	Exact	Adjacent	Neither		
English– Language Arts	1	1,512	2.56	1.13	1,512	2.56	1.14	94.97	4.63	0.39	0.06	0.97
	3	1,512	2.72	1.05	1,512	2.71	1.06	94.58	5.09	0.33	0.06	0.97
	4	1,512	3.04	0.97	1,512	3.04	0.98	97.42	1.98	0.59	0.03	0.97
	6	1,512	2.27	0.90	1,512	2.27	0.90	95.30	4.37	0.33	0.05	0.96
	7	1,512	2.83	1.17	1,512	2.82	1.18	96.63	3.11	0.26	0.04	0.98
	9	1,512	2.57	1.29	1,512	2.55	1.29	95.44	4.23	0.34	0.05	0.98
	10	1,512	2.31	1.19	1,512	2.30	1.19	94.64	4.96	0.40	0.06	0.97
	12	1,512	2.28	0.84	1,512	2.28	0.84	93.78	5.56	0.67	0.07	0.93
Mathematics	1	1,516	2.53	1.05	1,516	2.52	1.05	96.44	3.23	0.33	0.04	0.98
	3	1,516	2.15	1.16	1,516	2.15	1.16	96.97	2.37	0.66	0.04	0.98
	4	1,516	1.93	1.07	1,516	1.95	1.08	94.13	4.62	1.25	0.07	0.95
	6	1,516	3.05	1.25	1,516	3.05	1.25	96.90	2.44	0.66	0.04	0.98
	7	1,516	2.92	1.29	1,516	2.92	1.30	97.49	1.98	0.52	0.03	0.98
	9	1,516	2.81	0.81	1,516	2.80	0.81	96.44	3.30	0.27	0.04	0.96
	10	1,516	1.72	1.06	1,516	1.72	1.06	97.49	1.85	0.66	0.04	0.97
	12	1,516	2.13	1.25	1,516	2.14	1.25	95.91	3.17	0.92	0.06	0.97
Science	1	748	2.24	0.96	748	2.23	0.96	94.52	5.21	0.26	0.06	0.96
	3	748	2.63	1.06	748	2.64	1.06	95.86	3.21	0.93	0.05	0.97
	4	748	2.66	1.12	748	2.67	1.10	94.79	4.81	0.40	0.06	0.97
	6	748	2.55	1.37	748	2.54	1.38	96.93	2.27	0.79	0.04	0.98
	7	748	2.80	1.04	748	2.80	1.03	96.39	2.94	0.66	0.04	0.97
	9	748	2.99	1.05	748	3.02	1.03	95.45	3.88	0.66	0.06	0.95
	10	748	2.46	0.97	748	2.46	0.98	96.52	3.34	0.13	0.04	0.98
	12	748	2.72	0.94	748	2.72	0.94	94.52	4.81	0.67	0.06	0.95

* Mean absolute difference between first and second ratings

† Pearson correlation between first and second ratings

Table 8.C.39 Interrater Agreement Analyses for Operational Tasks: Level IV

Level IV		First Rating			Second Rating			% Agreement			MAD *	Corr †
Content Area	Task	N	Mean	SD	N	Mean	SD	Exact	Adjacent	Neither		
English– Language Arts	1	1,596	2.48	0.98	1,596	2.48	0.99	95.43	4.20	0.38	0.05	0.97
	3	1,596	2.52	1.18	1,596	2.51	1.19	93.67	5.89	0.43	0.07	0.97
	4	1,596	1.51	1.04	1,596	1.52	1.05	93.11	5.51	1.38	0.09	0.94
	6	1,596	2.61	1.18	1,596	2.60	1.18	92.98	6.64	0.37	0.07	0.97
	7	1,596	2.38	1.24	1,596	2.39	1.24	96.99	2.63	0.38	0.04	0.98
	9	1,596	2.28	1.23	1,596	2.28	1.24	97.12	2.57	0.31	0.03	0.99
	10	1,596	2.71	1.10	1,596	2.71	1.11	96.12	3.26	0.63	0.05	0.97
	12	1,596	2.34	1.25	1,596	2.32	1.25	92.29	6.77	0.94	0.09	0.96
Mathematics	1	1,585	1.44	0.71	1,585	1.45	0.72	95.65	4.10	0.25	0.05	0.95
	3	1,585	2.48	1.39	1,585	2.47	1.40	96.59	2.84	0.57	0.04	0.98
	4	1,585	3.02	1.28	1,585	3.03	1.27	97.54	1.89	0.57	0.04	0.98
	6	1,585	1.80	1.28	1,585	1.81	1.29	96.28	2.71	1.01	0.05	0.97
	7	1,585	2.98	1.27	1,585	2.98	1.28	97.35	2.33	0.32	0.03	0.98
	9	1,585	3.00	1.07	1,585	2.99	1.08	96.59	2.59	0.82	0.05	0.96
	10	1,585	2.67	1.06	1,585	2.67	1.05	92.68	6.62	0.69	0.08	0.96
	12	1,585	2.68	1.14	1,585	2.68	1.14	95.52	3.47	1.01	0.06	0.95
Science	1	441	2.63	0.95	441	2.63	0.97	93.88	5.22	0.91	0.07	0.94
	3	441	3.00	1.11	441	2.99	1.12	94.56	4.54	0.90	0.07	0.95
	4	441	2.33	1.05	441	2.34	1.05	92.97	5.67	1.37	0.10	0.92
	6	441	3.07	1.05	441	3.05	1.08	94.56	4.76	0.69	0.07	0.95
	7	441	2.98	0.75	441	3.00	0.75	96.60	3.40	0.00	0.03	0.97
	9	441	2.97	0.97	441	2.98	0.98	95.46	4.31	0.23	0.05	0.97
	10	441	2.24	1.24	441	2.26	1.25	92.97	5.67	1.36	0.09	0.96
	12	441	2.82	1.13	441	2.85	1.12	95.46	3.40	1.14	0.06	0.96

* Mean absolute difference between first and second ratings

† Pearson correlation between first and second ratings

Table 8.C.40 Interrater Agreement Analyses for Operational Tasks: Level V

Level V		First Rating			Second Rating			% Agreement			MAD *	Corr †
Content Area	Task	N	Mean	SD	N	Mean	SD	Exact	Adjacent	Neither		
English– Language Arts	1	1,218	3.07	1.14	1,218	3.06	1.14	93.19	5.75	1.07	0.08	0.95
	3	1,218	3.11	0.86	1,218	3.11	0.85	96.22	3.12	0.65	0.05	0.95
	4	1,218	2.79	1.05	1,218	2.79	1.04	94.42	4.93	0.65	0.06	0.96
	6	1,218	1.89	1.13	1,218	1.90	1.15	93.19	5.34	1.48	0.09	0.95
	7	1,218	2.54	1.07	1,218	2.53	1.08	90.80	7.96	1.23	0.11	0.94
	9	1,218	2.60	1.11	1,218	2.61	1.11	92.78	6.32	0.90	0.08	0.96
	10	1,218	2.46	1.19	1,218	2.45	1.19	93.51	5.17	1.31	0.08	0.95
	12	1,218	2.03	1.20	1,218	2.01	1.20	92.45	5.58	1.97	0.10	0.94
Mathematics	1	1,212	2.54	1.46	1,212	2.55	1.45	96.95	2.23	0.82	0.04	0.98
	3	1,212	3.37	1.18	1,212	3.38	1.17	97.61	1.40	0.99	0.04	0.97
	4	1,212	2.74	1.13	1,212	2.74	1.13	94.97	3.88	1.16	0.07	0.96
	6	1,212	2.23	1.36	1,212	2.24	1.36	95.46	3.14	1.40	0.07	0.96
	7	1,212	2.57	1.29	1,212	2.55	1.30	93.81	4.87	1.32	0.08	0.95
	9	1,212	2.68	1.41	1,212	2.68	1.41	96.04	2.48	1.49	0.07	0.96
	10	1,212	1.88	1.14	1,212	1.89	1.14	94.14	4.29	1.57	0.08	0.94
	12	1,212	2.84	1.32	1,212	2.86	1.30	95.05	3.05	1.90	0.09	0.94
Science	1	381	3.40	0.83	381	3.38	0.85	98.16	1.31	0.52	0.03	0.97
	3	381	2.38	1.00	381	2.38	0.97	94.49	5.25	0.26	0.06	0.96
	4	381	2.76	0.98	381	2.76	1.00	96.06	3.94	0.00	0.04	0.98
	6	381	2.22	1.09	381	2.21	1.09	94.23	5.77	0.00	0.06	0.98
	7	381	2.60	1.08	381	2.62	1.09	91.60	6.30	2.10	0.11	0.92
	9	381	2.98	0.91	381	2.95	0.94	94.49	4.20	1.31	0.07	0.93
	10	381	3.43	0.91	381	3.40	0.93	95.54	3.41	1.04	0.07	0.92
	12	381	2.52	1.24	381	2.52	1.24	92.65	4.46	2.88	0.12	0.92

* Mean absolute difference between first and second ratings

† Pearson correlation between first and second ratings

Appendix 8.D—IRT Analyses

Table 8.D.1 Item Classifications for Model-Data Fit Across All CAPA Levels

Fit Classification	ELA No. of Items	Mathematics No. of Items	Science No. of Items
A	36	31	34
B	64	60	27
C	14	27	2
D	4	2	0
F	0	0	0

Table 8.D.2 Fit Classifications: Level I Tasks

Fit	ELA Frequency	Mathematics Frequency	Science Frequency
A	7	4	5
B	16	14	10
C	0	6	1
D	1	0	0
F	0	0	0

Table 8.D.3 Fit Classifications: Level II Tasks

Fit	ELA Frequency	Mathematics Frequency
A	6	4
B	12	14
C	3	6
D	1	0
F	0	0

Table 8.D.4 Fit Classifications: Level III Tasks

Fit	ELA Frequency	Mathematics Frequency	Science Frequency
A	11	3	15
B	10	15	1
C	3	5	0
D	0	1	0
F	0	0	0

Table 8.D.5 Fit Classifications: Level IV Tasks

Fit	ELA Frequency	Mathematics Frequency	Science Frequency
A	5	7	7
B	15	9	9
C	4	7	0
D	0	1	0
F	0	0	0

Table 8.D.6 Fit Classifications: Level V Tasks

Fit	ELA Frequency	Mathematics Frequency	Science Frequency
A	7	13	7
B	11	8	7
C	4	3	1
D	2	0	0
F	0	0	0

Table 8.D.7 IRT *b*-values for ELA, by Level

Level		Number of Items	Mean	Standard Deviation	Min	Max
I	All Operational Items	8	-0.60	0.14	-0.85	-0.46
	Field-Test Items	16	-0.63	0.15	-1.06	-0.46
II	All Operational Items	8	-0.76	0.79	-2.40	0.03
	Field-Test Items	14	-1.26	0.34	-1.73	-0.63
III	All Operational Items	8	-0.82	0.39	-1.62	-0.30
	Field-Test Items	16	-1.60	0.57	-2.77	-0.63
IV	All Operational Items	8	-0.87	0.43	-1.37	0.10
	Field-Test Items	16	-1.34	0.80	-2.70	-0.31
V	All Operational Items	8	-0.91	0.50	-1.77	-0.27
	Field-Test Items	16	-1.21	0.50	-2.23	-0.47

Table 8.D.8 IRT *b*-values for Mathematics, by Level

Level		Number of Items	Mean	Standard Deviation	Min	Max
I	All Operational Items	8	-0.24	0.14	-0.45	-0.05
	Field-test Items	16	-0.25	0.14	-0.44	0.11
II	All Operational Items	8	-0.96	0.74	-1.92	0.50
	Field-test Items	16	-1.29	0.82	-2.46	0.36
III	All Operational Items	8	-0.93	0.36	-1.34	-0.54
	Field-test Items	16	-1.20	0.57	-2.44	-0.39
IV	All Operational Items	8	-0.81	0.58	-1.41	0.37
	Field-test Items	16	-0.86	0.63	-1.93	0.12
V	All Operational Items	8	-1.09	0.36	-1.63	-0.41
	Field-test Items	16	-1.26	0.50	-2.27	-0.34

Table 8.D.9 IRT *b*-values for Science, by Level

Level		Number of Items	Mean	Standard Deviation	Min	Max
I	All Operational Items	8	-0.31	0.16	-0.54	0.00
	Field-test Items	8	-0.28	0.09	-0.40	-0.10
III	All Operational Items	8	-1.05	0.30	-1.49	-0.56
	Field-test Items	8	-0.95	0.48	-1.65	-0.04
IV	All Operational Items	8	-1.11	0.34	-1.52	-0.45
	Field-test Items	8	-1.16	0.27	-1.50	-0.78
V	All Operational Items	8	-0.65	0.53	-1.41	0.02
	Field-test Items	8	-0.39	0.60	-1.16	0.47

Table 8.D.10 Score Conversions: Level I, ELA

Raw Score	Freq. Distrib.	Theta	Scale Score	CSEM	Performance Level
40	1,554	N/A	60	0	
39	473	1.2198	55	6	
38	352	0.8570	51	6	
37	276	0.6725	49	5	
36	878	0.5481	47	4	
35	352	0.4529	46	3	
34	321	0.3749	45	3	
33	252	0.3080	44	3	
32	737	0.2487	44	3	Advanced
31	354	0.1948	43	3	
30	276	0.1451	43	3	
29	257	0.0985	42	2	
28	693	0.0541	41	2	
27	350	0.0116	41	2	
26	262	-0.0297	41	2	
25	255	-0.0701	40	2	
24	670	-0.1099	40	2	
23	345	-0.1495	39	2	
22	268	-0.1891	39	2	
21	239	-0.2291	38	2	
20	549	-0.2697	38	2	
19	294	-0.3114	37	2	
18	232	-0.3544	37	2	Proficient
17	219	-0.3993	36	2	
16	407	-0.4466	36	3	
15	250	-0.4971	35	3	
14	234	-0.5518	35	3	
13	175	-0.6120	34	3	
12	316	-0.6797	33	3	
11	218	-0.7581	32	3	Basic
10	154	-0.8522	31	4	
9	198	-0.9706	30	4	
8	405	-1.1296	28	5	
7	154	-1.3568	25	6	Below Basic
6	126	-1.6799	22	7	
5	138	-2.0810	17	5	
4	121	-2.5198	15	2	
3	101	-2.9985	15	0	
2	121	-3.5698	15	0	Far Below Basic
1	141	-4.4076	15	0	
0	381	N/A	15	0	

Table 8.D.11 Score Conversions: Level II, ELA

Raw Score	Freq. Distrib.	Theta	Scale Score	CSEM	Performance Level
32	53	N/A	60	0	Advanced
31	83	3.4951	56	4	
30	142	2.6734	52	4	
29	192	2.1618	49	3	
28	217	1.7845	47	3	
27	251	1.4833	46	3	
26	303	1.2305	45	2	
25	299	1.0109	44	2	
24	337	0.8146	43	2	
23	324	0.6348	42	2	
22	330	0.4668	41	2	
21	317	0.3067	40	2	Proficient
20	293	0.1513	39	2	
19	354	-0.0021	39	2	
18	356	-0.1562	38	2	
17	354	-0.3136	37	2	
16	417	-0.4771	36	2	
15	356	-0.6496	36	2	
14	350	-0.8339	35	2	Basic
13	329	-1.0329	34	2	
12	247	-1.2482	33	2	
11	159	-1.4810	31	2	
10	133	-1.7318	30	2	Below Basic
9	103	-2.0019	29	3	
8	79	-2.2937	28	3	
7	47	-2.6099	26	3	
6	52	-2.9525	24	3	
5	37	-3.3236	23	3	
4	44	-3.7305	21	3	Far Below Basic
3	31	-4.1932	18	4	
2	22	-4.7633	16	2	
1	19	-5.6101	15	0	
0	38	N/A	15	0	

Table 8.D.12 Score Conversions: Level III, ELA

Raw Score	Freq. Distrib.	Theta	Scale Score	CSEM	Performance Level
32	71	N/A	60	0	
31	149	3.5309	52	5	
30	178	2.7473	49	3	
29	281	2.2505	47	3	
28	325	1.8720	46	2	
27	367	1.5600	45	2	
26	400	1.2917	44	2	Advanced
25	459	1.0548	43	2	
24	446	0.8413	42	2	
23	412	0.6454	41	2	
22	385	0.4623	41	2	
21	365	0.2881	40	2	
20	326	0.1193	39	2	
19	300	-0.0471	39	2	
18	294	-0.2140	38	2	
17	248	-0.3842	37	2	Proficient
16	282	-0.5605	37	2	
15	244	-0.7457	36	2	
14	244	-0.9428	35	2	
13	220	-1.1544	34	2	
12	220	-1.3830	34	2	
11	190	-1.6304	33	2	Basic
10	122	-1.8972	32	2	
9	133	-2.1826	30	2	
8	102	-2.4855	29	2	
7	67	-2.8047	28	2	
6	43	-3.1410	27	2	Below Basic
5	41	-3.4989	25	2	
4	43	-3.8895	24	2	
3	38	-4.3357	22	3	
2	22	-4.8910	20	3	Far Below Basic
1	29	-5.7258	17	2	
0	59	N/A	15	0	

Table 8.D.13 Score Conversions: Level IV, ELA

Raw Score	Freq. Distrib.	Theta	Scale Score	CSEM	Performance Level
32	131	N/A	60	0	
31	208	2.9506	56	4	
30	235	2.2610	52	5	
29	314	1.8616	50	3	
28	378	1.5738	48	3	
27	405	1.3423	47	3	Advanced
26	488	1.1436	46	3	
25	441	0.9661	45	2	
24	478	0.8033	44	2	
23	472	0.6510	43	2	
22	465	0.5063	42	2	
21	454	0.3667	41	2	
20	393	0.2300	41	2	
19	421	0.0939	40	2	
18	401	-0.0437	39	2	Proficient
17	406	-0.1851	38	2	
16	412	-0.3328	37	2	
15	402	-0.4900	36	2	
14	364	-0.6604	35	2	
13	381	-0.8488	34	3	
12	433	-1.0616	33	3	Basic
11	453	-1.3072	32	3	
10	434	-1.5965	30	3	
9	358	-1.9403	28	4	
8	310	-2.3420	26	4	
7	81	-2.7857	23	4	Below Basic
6	78	-3.2448	20	4	
5	59	-3.7048	18	4	
4	57	-4.1730	15	2	
3	59	-4.6760	15	0	
2	32	-5.2716	15	0	Far Below Basic
1	32	-6.1339	15	0	
0	56	N/A	15	0	

Table 8.D.14 Score Conversions: Level V, ELA

Raw Score	Freq. Distrib.	Theta	Scale Score	CSEM	Performance Level
32	173	N/A	60	0	Advanced
31	247	3.1369	50	6	
30	322	2.3810	47	3	
29	455	1.9239	45	2	
28	445	1.5912	44	2	
27	508	1.3253	43	2	
26	603	1.0992	42	2	
25	596	0.8980	41	2	
24	597	0.7125	41	2	
23	594	0.5370	40	2	
22	571	0.3676	39	2	Proficient
21	544	0.2018	39	2	
20	524	0.0377	38	2	
19	469	-0.1259	38	2	
18	465	-0.2902	37	2	
17	415	-0.4564	36	2	
16	368	-0.6259	36	2	
15	383	-0.8003	35	2	
14	352	-0.9818	34	2	Basic
13	309	-1.1733	34	2	
12	320	-1.3782	33	2	
11	280	-1.6004	32	2	
10	206	-1.8444	31	2	
9	186	-2.1141	30	2	
8	138	-2.4123	29	2	Below Basic
7	59	-2.7391	28	2	
6	53	-3.0927	26	2	
5	39	-3.4726	25	2	
4	38	-3.8848	24	3	
3	43	-4.3495	22	3	Far Below Basic
2	28	-4.9189	20	3	
1	27	-5.7633	16	2	
0	67	N/A	15	0	

Table 8.D.15 Score Conversions: Level I, Mathematics

Raw Score	Freq. Distrib.	Theta	Scale Score	CSEM	Performance Level	
40	641	N/A	60	0		
39	282	1.5362	50	9		
38	228	1.1836	46	5		
37	235	1.0046	44	4		
36	685	0.8839	43	3		
35	343	0.7917	42	3	Advanced	
34	277	0.7162	41	3		
33	226	0.6514	41	2		
32	824	0.5942	40	2		
31	380	0.5422	40	2		
30	310	0.4943	39	2		
29	285	0.4494	39	2		
28	870	0.4068	38	2		
27	360	0.3659	38	2		
26	379	0.3264	38	2		
25	264	0.2877	37	2		
24	767	0.2495	37	2	Proficient	
23	369	0.2117	36	2		
22	365	0.1738	36	2		
21	245	0.1356	36	2		
20	658	0.0968	35	2		
19	377	0.0570	35	2		
18	299	0.0159	34	2		
17	221	-0.0269	34	2		
16	526	-0.0721	34	2		
15	286	-0.1204	33	2		
14	244	-0.1727	33	2	Basic	
13	200	-0.2304	32	2		
12	414	-0.2954	31	3		
11	246	-0.3708	31	3		
10	201	-0.4615	30	3		
9	185	-0.5763	29	4		
8	488	-0.7318	27	4		
7	195	-0.9586	25	5	Below Basic	
6	146	-1.2920	22	6		
5	131	-1.7127	17	5		
4	126	-2.1696	15	2		
3	120	-2.6629	15	0		
2	119	-3.2464	15	0	Far Below Basic	
1	125	-4.0950	15	0		
0	423	N/A	15	0		

Table 8.D.16 Score Conversions: Level II, Mathematics

Raw Score	Freq. Distrib.	Theta	Scale Score	CSEM	Performance Level
32	71	N/A	60	0	Advanced
31	107	2.4481	55	5	
30	147	1.9024	51	5	
29	321	1.5520	49	4	
28	267	1.2764	47	4	
27	364	1.0484	45	3	
26	376	0.8560	44	3	
25	373	0.6896	42	3	
24	305	0.5414	41	3	
23	285	0.4055	40	3	Proficient
22	319	0.2777	39	3	
21	287	0.1550	38	3	
20	288	0.0349	38	3	
19	294	-0.0846	37	3	
18	262	-0.2056	36	3	
17	260	-0.3300	35	3	Basic
16	272	-0.4602	34	3	
15	298	-0.5988	33	3	
14	277	-0.7493	32	3	
13	287	-0.9160	31	3	Below Basic
12	247	-1.1049	29	3	
11	248	-1.3232	28	4	
10	215	-1.5798	26	4	
9	139	-1.8828	24	4	
8	103	-2.2364	21	4	
7	53	-2.6345	18	5	Far Below Basic
6	43	-3.0607	15	2	
5	24	-3.5015	15	0	
4	28	-3.9595	15	0	
3	22	-4.4583	15	0	
2	21	-5.0549	15	0	
1	16	-5.9236	15	0	
0	31	N/A	15	0	

Table 8.D.17 Score Conversions: Level III, Mathematics

Raw Score	Freq. Distrib.	Theta	Scale Score	CSEM	Performance Level
32	68	N/A	60	0	Advanced
31	94	2.5491	49	7	
30	98	1.8862	46	4	
29	135	1.5141	44	3	
28	181	1.2533	43	2	
27	217	1.0483	42	2	
26	287	0.8755	41	2	
25	328	0.7228	40	2	
24	397	0.5834	39	2	
23	417	0.4529	39	2	Proficient
22	437	0.3284	38	2	
21	413	0.2078	38	2	
20	416	0.0895	37	2	
19	392	-0.0280	36	2	
18	385	-0.1462	36	2	
17	354	-0.2665	35	2	
16	373	-0.3910	35	2	
15	326	-0.5222	34	2	Basic
14	352	-0.6637	33	2	
13	290	-0.8208	32	2	
12	311	-1.0010	32	2	
11	255	-1.2153	31	2	
10	192	-1.4784	29	3	Below Basic
9	104	-1.8055	28	3	
8	101	-2.2016	26	3	
7	31	-2.6475	23	3	
6	25	-3.1070	21	3	
5	18	-3.5604	19	3	
4	22	-4.0156	17	3	
3	19	-4.5017	15	1	Far Below Basic
2	18	-5.0787	15	0	
1	10	-5.9221	15	0	
0	28	N/A	15	0	

Table 8.D.18 Score Conversions: Level IV, Mathematics

Raw Score	Freq. Distrib.	Theta	Scale Score	CSEM	Performance Level
32	93	N/A	60	0	Advanced
31	130	2.7999	55	5	
30	268	2.1354	51	5	
29	361	1.6953	48	4	
28	380	1.3732	46	4	
27	442	1.1271	44	3	
26	483	0.9277	43	3	
25	549	0.7570	41	3	
24	645	0.6048	40	3	Proficient
23	606	0.4653	39	3	
22	629	0.3349	39	2	
21	596	0.2112	38	2	
20	569	0.0922	37	2	
19	499	-0.0238	36	2	
18	385	-0.1386	35	2	
17	385	-0.2538	34	2	Basic
16	349	-0.3716	34	2	
15	367	-0.4945	33	2	
14	411	-0.6257	32	3	
13	380	-0.7697	31	3	
12	372	-0.9330	30	3	
11	305	-1.1249	28	3	Below Basic
10	244	-1.3593	27	4	
9	193	-1.6530	25	4	
8	157	-2.0170	22	4	
7	54	-2.4370	19	4	
6	41	-2.8779	16	3	
5	26	-3.3183	15	1	Far Below Basic
4	38	-3.7645	15	0	
3	22	-4.2441	15	0	
2	21	-4.8160	15	0	
1	22	-5.6552	15	0	
0	46	N/A	15	0	

Table 8.D.19 Score Conversions: Level V, Mathematics

Raw Score	Freq. Distrib.	Theta	Scale Score	CSEM	Performance Level
32	529	N/A	60	0	Advanced
31	356	1.8351	48	9	
30	404	1.3388	45	4	
29	564	1.0556	43	3	
28	499	0.8526	42	3	
27	510	0.6911	41	2	
26	549	0.5546	40	2	
25	512	0.4342	39	2	Proficient
24	563	0.3246	39	2	
23	498	0.2224	38	2	
22	485	0.1248	37	2	
21	466	0.0301	37	2	
20	438	-0.0636	36	2	
19	409	-0.1577	36	2	
18	378	-0.2535	35	2	Basic
17	366	-0.3526	34	2	
16	353	-0.4566	34	2	
15	328	-0.5673	33	2	
14	247	-0.6870	32	2	
13	268	-0.8193	32	2	
12	262	-0.9697	31	2	
11	283	-1.1482	30	3	Below Basic
10	276	-1.3740	28	3	
9	278	-1.6839	26	4	
8	276	-2.1219	24	4	
7	59	-2.6510	20	4	
6	41	-3.1691	17	4	Far Below Basic
5	30	-3.6501	15	2	
4	30	-4.1168	15	0	
3	27	-4.6074	15	0	
2	26	-5.1857	15	0	
1	20	-6.0291	15	0	
0	62	N/A	15	0	

Table 8.D.20 Score Conversions: Level I, Science

Raw Score	Freq. Distrib.	Theta	Scale Score	CSEM	Performance Level
40	272	N/A	60	0	
39	65	1.4879	49	10	
38	79	1.1416	46	5	
37	58	0.9639	44	4	
36	202	0.8432	43	3	
35	79	0.7506	42	3	Advanced
34	62	0.6746	41	3	
33	62	0.6092	40	2	
32	190	0.5513	40	2	
31	86	0.4988	39	2	
30	52	0.4503	39	2	
29	57	0.4048	38	2	
28	187	0.3617	38	2	
27	101	0.3203	37	2	
26	78	0.2803	37	2	
25	54	0.2411	37	2	Proficient
24	203	0.2026	36	2	
23	83	0.1644	36	2	
22	65	0.1261	36	2	
21	65	0.0876	35	2	
20	167	0.0484	35	2	
19	84	0.0082	34	2	
18	78	-0.0334	34	2	
17	61	-0.0768	34	2	
16	132	-0.1226	33	2	
15	69	-0.1717	33	2	Basic
14	58	-0.2251	32	2	
13	51	-0.2841	31	3	
12	91	-0.3510	31	3	
11	62	-0.4291	30	3	
10	38	-0.5239	29	3	
9	45	-0.6456	28	4	
8	133	-0.8136	26	5	Below Basic
7	60	-1.0633	24	6	
6	41	-1.4267	20	6	
5	40	-1.8660	16	4	
4	31	-2.3297	15	0	
3	31	-2.8253	15	0	
2	42	-3.4099	15	0	Far Below Basic
1	28	-4.2594	15	0	
0	122	N/A	15	0	

Table 8.D.21 Score Conversions: Level III, Science

Raw Score	Freq. Distrib.	Theta	Scale Score	CSEM	Performance Level
32	28	N/A	60	0	Advanced
31	64	2.8031	46	8	
30	91	2.0463	43	3	
29	130	1.5879	42	2	
28	149	1.2562	41	2	
27	168	0.9954	40	2	
26	218	0.7792	39	2	
25	225	0.5930	38	1	
24	213	0.4272	38	1	
23	206	0.2753	37	1	
22	210	0.1327	37	1	
21	216	-0.0045	36	1	
20	210	-0.1389	36	1	
19	187	-0.2733	35	1	
18	216	-0.4097	35	1	Basic
17	198	-0.5504	34	1	
16	169	-0.6975	34	1	
15	131	-0.8533	33	1	
14	110	-1.0206	33	1	
13	91	-1.2024	32	2	
12	70	-1.4025	31	2	
11	52	-1.6249	31	2	
10	52	-1.8739	30	2	
9	33	-2.1520	29	2	
8	35	-2.4591	28	2	
7	14	-2.7916	26	2	
6	10	-3.1448	25	2	
5	13	-3.5176	24	2	Far Below Basic
4	8	-3.9179	22	2	
3	9	-4.3675	21	3	
2	3	-4.9203	19	3	
1	5	-5.7471	16	2	
0	22	N/A	15	0	

Table 8.D.22 Score Conversions: Level IV, Science

Raw Score	Freq. Distrib.	Theta	Scale Score	CSEM	Performance Level
32	48	N/A	60	0	Advanced
31	65	2.5662	45	9	
30	98	1.8603	43	3	
29	121	1.4455	41	2	
28	146	1.1477	40	2	
27	178	0.9114	39	2	Proficient
26	214	0.7118	38	2	
25	236	0.5355	38	2	
24	227	0.3748	37	2	
23	222	0.2246	36	1	
22	214	0.0815	36	1	
21	229	-0.0571	35	1	
20	196	-0.1933	35	1	
19	205	-0.3291	34	1	Basic
18	176	-0.4661	34	1	
17	153	-0.6063	33	1	
16	114	-0.7519	33	2	
15	110	-0.9051	32	2	
14	83	-1.0689	31	2	
13	54	-1.2467	31	2	
12	45	-1.4425	30	2	
11	32	-1.6606	29	2	Below Basic
10	28	-1.9050	28	2	
9	23	-2.1781	27	2	
8	30	-2.4797	26	2	
7	9	-2.8062	25	2	
6	6	-3.1543	23	2	
5	4	-3.5237	22	2	
4	5	-3.9226	20	3	
3	6	-4.3726	18	3	Far Below Basic
2	2	-4.9269	16	2	
1	5	-5.7561	15	1	
0	15	N/A	15	0	

Table 8.D.23 Score Conversions: Level V, Science

Raw Score	Freq. Distrib.	Theta	Scale Score	CSEM	Performance Level
32	58	N/A	60	0	Advanced
31	74	3.3826	45	8	
30	104	2.6069	43	3	
29	151	2.1304	41	2	
28	188	1.7791	40	2	
27	208	1.4969	39	2	
26	216	1.2577	38	2	
25	245	1.0470	38	2	
24	236	0.8559	37	1	
23	250	0.6782	36	1	
22	239	0.5097	36	1	
21	217	0.3471	35	1	
20	210	0.1881	35	1	Basic
19	185	0.0309	34	1	
18	142	-0.1257	34	1	
17	124	-0.2829	33	1	
16	111	-0.4417	32	1	
15	86	-0.6035	32	1	
14	66	-0.7701	31	1	
13	59	-0.9440	31	1	
12	47	-1.1285	30	1	
11	44	-1.3275	29	2	
10	39	-1.5458	29	2	
9	26	-1.7883	28	2	
8	30	-2.0587	27	2	Far Below Basic
7	4	-2.3581	26	2	
6	10	-2.6852	25	2	
5	5	-3.0397	24	2	
4	7	-3.4276	22	2	
3	12	-3.8693	21	2	
2	8	-4.4168	19	3	
1	4	-5.2403	16	2	
0	19	N/A	15	0	

Appendix 8.E—DIF Analyses

Table 8.E.1 Item Exhibiting Significant DIF by Ethnic Group

Content Area	Task No.	Level	Task#	Version	SMD	Comparison	In Favor Of
English– Language Arts Operational Tasks	VC472221	3	6	Operational	0.272	White–Filipino	Filipino
	VC335049	4	7	Operational	0.325	White–Filipino	Filipino
	VC208668	5	6	Operational	0.310	White–Asian	Asian
	VC208654	5	12	Operational	0.386	White–Asian	Asian
	VC208654	5	12	Operational	0.406	White–Filipino	Filipino
	VC208654	5	12	Operational	0.371	White–CombAsian	CombAsian
English– Language Arts Field-test Tasks	VE630635	3	11	1	0.230	White–CombAsian	CombAsian
	VE630774	4	11	2	–0.351	White–CombAsian	White
	VE631901	5	5	1	–0.404	White–Asian	White
	VE631901	5	5	1	–0.363	White–CombAsian	White
	VE631926	5	11	1	–0.329	White–Asian	White
	VE631913	5	2	4	0.343	White–CombAsian	CombAsian
Mathematics Operational Tasks*	–	–	–	–	–	–	–
Mathematics Field-test Tasks	VE630729	4	23	3	–0.326	White–CombAsian	White
Science Operational Tasks*	–	–	–	–	–	–	–
Science Field- test Tasks*	–	–	–	–	–	–	–

* No items exhibited significant ethnic DIF.

Table 8.E.2 Items Exhibiting Significant DIF by Disability Group

Content Area	Task No.	Level	Task#	Version	SMD	Comparison	In Favor Of
English–Language Arts Operational Tasks	VC208470	4	6	Operational	–0.315	MR/ID–Autism	MR/ID
	VC335049	4	7	Operational	–0.390	MR/ID–Autism	MR/ID
	VE093557	4	9	Operational	0.439	MR/ID–Autism	Autism
	VC208350	4	10	Operational	0.295	MR/ID–Autism	Autism
	VC208668	5	6	Operational	0.493	MR/ID–Autism	Autism
	VC208654	5	12	Operational	0.526	MR/ID–Autism	Autism
English–Language Arts Field-test Tasks	VE630339	1	8	1	–0.652	MR/ID–Ortho Imped	MR/ID
	VE630339	1	8	1	–0.549	MR/ID–Multiple Disab	MR/ID
	VE630341	1	5	2	–0.748	MR/ID– Ortho Imped	MR/ID
	VE630355	1	5	4	0.455	MR/ID– Ortho Imped	MR/ID
	VE630355	1	5	4	–0.540	MR/ID–Autism	MR/ID
	VE630556	2	8	1	0.315	MR/ID–Autism	Autism
	VE630549	2	5	2	0.450	MR/ID–Autism	Autism
	VE630571	2	2	3	0.472	MR/ID–Autism	Autism
	VE630568	2	5	3	0.512	MR/ID–Autism	Autism
	VE630564	2	8	3	0.261	MR/ID–Autism	Autism
	VE630394	2	2	4	0.291	MR/ID–Autism	Autism
	VE630565	2	5	4	0.498	MR/ID–Autism	Autism
	VE630782	4	5	2	–0.398	MR/ID–Spfc Learn	MR/ID
	VE630774	4	11	2	–0.381	MR/ID–Autism	MR/ID
	VE631828	5	2	1	0.370	MR/ID–Spfc Learn	Spfc Learn
	VE631901	5	5	1	0.253	MR/ID–Spfc Learn	Spfc Learn
	VE631903	5	8	1	0.392	MR/ID–Autism	Autism
	VE631926	5	11	1	–0.434	MR/ID–Autism	MR/ID
	VE631818	5	2	2	0.367	MR/ID–Spfc Learn	Spfc Learn
	VE631864	5	5	2	0.218	MR/ID–Spfc Learn	Spfc Learn
	VE631907	5	8	2	0.571	MR/ID–Autism	Autism
	VE631874	5	2	3	–0.297	MR/ID–Autism	MR/ID
	VE631913	5	2	4	0.294	MR/ID–Spfc Learn	Spfc Learn
VE631913	5	2	4	0.787	MR/ID–Autism	Autism	
Mathematics Operational Tasks	VE436484	2	24	Operational	–0.225	MR/ID–Speech Imp	MR/ID
	VE436484	2	24	Operational	–0.369	MR/ID–Spfc Learn	MR/ID
	VE436484	2	24	Operational	–0.247	MR/ID–Autism	MR/ID
	VE098600	3	13	Operational	–0.362	MR/ID–Speech Imp	MR/ID
	VE098600	3	13	Operational	–0.304	MR/ID–Spfc Learn	MR/ID
	VC335678	3	22	Operational	–0.300	MR/ID–Spfc Learn	MR/ID
	VC465936	3	24	Operational	0.421	MR/ID–Speech Imp	Speech Imp
	VC465936	3	24	Operational	0.448	MR/ID–Autism	Autism
	VC208066	5	21	Operational	–0.548	MR/ID–Speech Imp	MR/ID
	VC208066	5	21	Operational	–0.376	MR/ID–Spfc Learn	MR/ID
Mathematics Field-test Tasks	VE627769	2	14	3	–0.219	MR/ID–Autism	MR/ID
	VE627907	3	14	1	0.385	MR/ID–Autism	Autism
	VE627822	3	23	1	0.375	MR/ID–Autism	Autism
	VE627828	3	20	2	0.572	MR/ID–Autism	Autism
	VE627849	3	17	3	0.361	MR/ID–Autism	Autism
	VE630702	4	17	2	0.372	MR/ID–Spfc Learn	Spfc Learn
	VE630673	4	17	4	0.360	MR/ID–Autism	Autism
	VE631703	5	20	1	0.353	MR/ID–Spfc Learn	Spfc Learn
VE631715	5	17	2	0.425	MR/ID–Spfc Learn	Spfc Learn	
Science Operational Tasks*	–	–	–	–	–	–	–
Science Field-test Tasks	VE629180	4	26	2/4	–0.311	MR/ID–Autism	MR/ID

* No items exhibited significant disability DIF.

Table 8.E.3 CAPA Disability Distributions: Level I

Disability	ELA		Mathematics		Science	
	Frequency	Percent	Frequency	Percent	Frequency	Percent
Mental retardation/Intellectual disability	5,305	37.6%	5,300	37.7%	1,413	39.6%
Hard of hearing	72	0.5%	72	0.5%	18	0.5%
Deafness	47	0.3%	47	0.3%	13	0.4%
Speech or language impairment*	149	1.1%	149	1.1%	—	—
Visual impairment	277	2.0%	276	2.0%	73	2.0%
Emotional disturbance*	23	0.2%	23	0.2%	—	—
Orthopedic impairment	2,193	15.6%	2,184	15.5%	602	16.9%
Other health impairment	400	2.8%	399	2.8%	89	2.5%
Specific learning disability	84	0.6%	84	0.6%	13	0.4%
Deaf-blindness	28	0.2%	28	0.2%	—	—
Multiple disabilities	1,527	10.8%	1,523	10.8%	381	10.7%
Autism	3,770	26.7%	3,760	26.7%	893	25.1%
Traumatic brain injury	94	0.7%	94	0.7%	28	0.8%
Unknown	129	0.9%	126	0.9%	24	0.7%
TOTAL	14,098	100.0%	14,065	100.0%	3,564	100.0%

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

Table 8.E.4 CAPA Disability Distributions: Level II

Disability	ELA		Mathematics	
	Frequency	Percent	Frequency	Percent
Mental retardation/Intellectual disability	2,036	30.5%	2,033	30.6%
Hard of hearing	33	0.5%	33	0.5%
Deafness	37	0.6%	37	0.6%
Speech or language impairment	642	9.6%	642	9.7%
Visual impairment	28	0.4%	28	0.4%
Emotional disturbance	23	0.3%	23	0.3%
Orthopedic impairment	225	3.4%	226	3.4%
Other health impairment	365	5.5%	364	5.5%
Specific learning disability	537	8.1%	535	8.0%
Deaf-blindness*	—	—	—	—
Multiple disabilities	113	1.7%	113	1.7%
Autism	2,493	37.4%	2,482	37.3%
Traumatic brain injury	35	0.5%	35	0.5%
Unknown	96	1.4%	94	1.4%
TOTAL	6,668	100.0%	6,650	100.0%

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

Table 8.E.5 CAPA Disability Distributions: Level III

Disability	ELA		Mathematics		Science	
	Frequency	Percent	Frequency	Percent	Frequency	Percent
Mental retardation/Intellectual disability	2,597	36.6%	2,593	36.6%	1,345	37.8%
Hard of hearing	45	0.6%	44	0.6%	26	0.7%
Deafness	56	0.8%	56	0.8%	29	0.8%
Speech or language impairment	397	5.6%	398	5.6%	192	5.4%
Visual impairment	34	0.5%	33	0.5%	18	0.5%
Emotional disturbance	46	0.6%	46	0.6%	26	0.7%
Orthopedic impairment	303	4.3%	302	4.3%	143	4.0%
Other health impairment	356	5.0%	356	5.0%	160	4.5%
Specific learning disability	623	8.8%	621	8.8%	320	9.0%
Deaf–blindness*	–	–	–	–	–	–
Multiple disabilities	141	2.0%	141	2.0%	82	2.3%
Autism	2,389	33.6%	2,386	33.6%	1,163	32.7%
Traumatic brain injury	42	0.6%	43	0.6%	22	0.6%
Unknown	75	1.1%	74	1.0%	30	0.8%
TOTAL	7,105	100.0%	7,094	100.0%	3,556	100.0%

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

Table 8.E.6 CAPA Disability Distributions: Level IV

Disability	ELA		Mathematics		Science	
	Frequency	Percent	Frequency	Percent	Frequency	Percent
Mental retardation/Intellectual disability	4,428	43.9%	4,419	43.9%	1,538	46.6%
Hard of hearing	53	0.5%	53	0.5%	17	0.5%
Deafness	82	0.8%	81	0.8%	23	0.7%
Speech or language impairment	307	3.0%	307	3.0%	80	2.4%
Visual impairment	67	0.7%	67	0.7%	22	0.7%
Emotional disturbance	113	1.1%	111	1.1%	43	1.3%
Orthopedic impairment	509	5.0%	509	5.1%	171	5.2%
Other health impairment	436	4.3%	433	4.3%	127	3.8%
Specific learning disability	773	7.7%	773	7.7%	250	7.6%
Deaf–blindness*	–	–	–	–	–	–
Multiple disabilities	292	2.9%	291	2.9%	109	3.3%
Autism	2,836	28.1%	2,830	28.1%	860	26.1%
Traumatic brain injury	53	0.5%	52	0.5%	19	0.6%
Unknown	137	1.4%	137	1.4%	40	1.2%
TOTAL	10,091	100.0%	10,068	100.0%	3,299	100.0%

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

Table 8.E.7 CAPA Disability Distributions: Level V

Disability	ELA		Mathematics		Science	
	Frequency	Percent	Frequency	Percent	Frequency	Percent
Mental retardation/Intellectual disability	5,035	48.3%	5,029	48.4%	1,658	48.4%
Hard of hearing	71	0.7%	71	0.7%	30	0.9%
Deafness	94	0.9%	93	0.9%	30	0.9%
Speech or language impairment	186	1.8%	187	1.8%	71	2.1%
Visual impairment	72	0.7%	72	0.7%	28	0.8%
Emotional disturbance	141	1.4%	139	1.3%	41	1.2%
Orthopedic impairment	501	4.8%	498	4.8%	170	5.0%
Other health impairment	489	4.7%	489	4.7%	179	5.2%
Specific learning disability	890	8.5%	886	8.5%	303	8.8%
Deaf-blindness*	—	—	—	—	—	—
Multiple disabilities	314	3.0%	313	3.0%	101	2.9%
Autism	2,437	23.4%	2,421	23.3%	754	22.0%
Traumatic brain injury	66	0.6%	66	0.6%	21	0.6%
Unknown	125	1.2%	125	1.2%	37	1.1%
TOTAL	10,424	100.0%	10,392	100.0%	3,424	100.0%

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

Chapter 9: Quality Control Procedures

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

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

Quality Control of Task Development

The task development process for the CAPA is described in detail in Chapter 3, starting on page 17. The next sections highlight elements of the process devoted specifically to the quality control of task development.

Task Specifications

ETS maintains task specifications for the CAPA and has developed an item utilization plan to guide the development of the tasks for each content area. Task writing emphasis is determined in consultation with the CDE. Adherence to the specifications ensures the maintenance of quality and consistency of the task development process.

Task Writers

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

Internal Contractor Reviews

Once tasks have been written, ETS assessment specialists make sure that each task goes through an intensive internal review process. Every step of this process is designed to produce tasks that exceed industry standards for quality. It includes 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 help to ensure that all tasks meet the specifications for the process.

Content Review

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

The artwork and graphics for the tasks are created during the internal content review period so assessment specialists can evaluate the correctness and appropriateness of the art early in the task development process. ETS selects visuals that are relevant to the task content

and that are easily understood so students do not struggle to determine the purpose or meaning of the questions.

Editorial Review

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

Fairness Review

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

Assessment Director Review

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

Assessment Review Panel Review

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

Statewide Pupil Assessment Review Panel Review

The SPAR panel is responsible for reviewing and approving the achievement tests that are to be used statewide for the testing of students in California public schools in grades two through eleven. The SPAR panel representatives ensure that the CAPA tasks conform to the requirements of *EC* Section 60602. See page 23 in Chapter 3 for additional information on the function of the SPAR panel within the task-review process.

Data Review of Field-tested Tasks

ETS field tests newly developed tasks to obtain statistical information about task performance. This information is used to evaluate tasks that are candidates for use in operational test forms. The tasks and task (item) statistics are examined carefully at data review meetings, where content experts discuss tasks that have poor statistics and do not meet the psychometric criteria for task quality. The CDE defines the criteria for acceptable or unacceptable task statistics. These criteria ensure that the task (1) has an appropriate level of difficulty for the target population; (2) discriminates well between examinees that differ in ability; and (3) conforms well to the statistical model underlying the measurement of the intended constructs. The results of analyses for differential item functioning (DIF) are used to make judgments about the appropriateness of items for various subgroups.

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

Quality Control of the Item Bank

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

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

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

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

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

Quality Control of Test Form Development

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

ETS also includes quality checks throughout preparation of the form planners. A form planner specifications document is developed by the test development team lead with input from ETS's item bank and statistics groups; this document is then reviewed by all team members who build forms at a training specific to form planners before the form-building process starts. After trained content team members sign off on a form planner, a representative from the internal QC group reviews each file for accuracy against the

specifications document. Assessment directors review and sign off on form planners prior to processing.

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

Quality Control of Test Materials

Collecting Test Materials

Once the tests are administered, school districts 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 districts contain color-coded labels identifying scorable and nonscorable materials and labels with bar-coded information identifying the school and district. The school districts 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 *CAPA Examiner's Manuals*, are returned to the Security Processing Department in Pearson's Cedar Rapids, Iowa facility. ETS and Pearson closely monitor the return of materials. The STAR Technical Assistance Center (TAC) at ETS monitors returns and notifies school districts that do not return their materials in a timely manner. STAR TAC contacts the district STAR coordinators and works with them to facilitate the return of the test materials.

Processing Test Materials

Upon receipt of the testing 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 their facilities. The materials are removed carefully from the shipping cartons and examined for a number of conditions, including physical damage, shipping errors, and omissions. A visual inspection to compare the number of students recorded on the School and Grade Identification (SGID) sheet with the number of answer documents in the stack is also conducted.

Pearson's image scanning process captures security information electronically and compares scorable material quantities reported on SGIDs to actual documents scanned. School districts 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 STAR 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 tasks on the diagnostic sheets, the scanner is recalibrated to work properly before being allowed to continue processing student documents.

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

Post-scanning Edits

After scanning, there are three opportunities for demographic data to be edited:

- After scanning, by Pearson online editors
- After Pearson's online editing, by district STAR coordinators (demographic edit)
- After paper reporting, by district STAR coordinators

Demographic edits completed by the Pearson editors and by the district STAR coordinators online are included in the data used for the paper reporting and for the technical reports.

Quality Control of Image Editing

Prior to submitting any STAR 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 task, and the appropriate problem and resolution text that instructs the editor on the actions that should be taken.

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

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

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, districts, and subgroups:

- Grade counts are compared to the District Master File Sheets.
- Document counts are compared to the School Master File Sheets.
- Document counts are compared to the SGIDs.

Any discrepancies identified in the steps outlined above are followed up by Pearson staff with the school districts 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 STAR Program is scored and aggregated to test a broader range of scoring and aggregation scenarios. These procedures assure that students and school districts receive the correct scores when the actual scoring process is carried out.

Scoring and Reporting Specifications

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

- General Reporting Specifications
- Form Planner Specifications
- Aggregation Rules
- "What If" . . . List
- Edit Specifications (which include matching information from observer documents to examiner documents for 10 percent of the CAPA that is administered)

Each of these documents is explained in detail in Chapter 7, starting on page 46. 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.

Matching Information on CAPA Answer Documents

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 must comply with the ETS scoring specifications.

All STAR answer documents contain uniquely numbered lithocodes that are both scannable and eye-readable. The lithocodes allow all pages of the document to be linked throughout processing, even after the documents have been slit into single sheets for scanning. For those students using more than one score, lithocodes link their demographics and responses within a document, while matching criteria are used to create a single record for all of the student's documents. The documents are matched within grade using the match criteria approved by the CDE.

Storing Answer Documents

After the answer documents have been scanned, edited, scored, and 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

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

The psychometric analyses conducted at ETS undergo comprehensive quality checks by a team of psychometricians and data analysts. Detailed checklists are consulted by members of the team for each of the statistical procedures performed on each CAPA. Quality assurance checks also include a comparison of the current year's statistics to statistics from previous years. The results of preliminary classical task analyses that provide a check on scoring keys are also reviewed by a senior psychometrician. The tasks that are flagged for questionable statistical attributes are sent to test development staff for their review; their comments are reviewed by the psychometricians before tasks are approved to be included in the equating process.

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

A few additional checks are performed for each process as described below.

Calibrations

During the calibration process, which is described in detail in Chapter 2 starting on page 14, checks are made to ascertain that the correct options for the analyses are selected. Checks are also made on the number of tasks, number of examinees with valid scores, IRT Rasch task difficulty estimates, standard errors for the Rasch task difficulty estimates, and the match of selected statistics to the results on the same statistics obtained during preliminary task analyses. Psychometricians also perform detailed reviews of plots and statistics to investigate if the data fit the model.

Scaling

During the scaling process, checks are made to ensure the following:

- The correct items are used for linking;
- The scaling evaluation process, including stability analysis and subsequent removal of items from the linking set (if any), is implemented according to specification (see details in the "Evaluation of Scaling" section in Chapter 8, on page 82); and
- The resulting scaling constants are correctly applied to transform the new item difficulty estimates on to the item bank scale.

Scoring Tables

Once the equating activities are complete and raw-to-scale score conversion tables are generated, the psychometricians carry out quality control checks on each scoring table.

Scoring tables are checked to verify the following:

- All raw scores are included in the tables;
- Scale scores increase as raw scores increase;
- The minimum reported scale score is 15 and maximum reported scale score is 60; and
- The cut points for the performance levels are correctly identified.

As a check on the reasonableness of the performance levels, psychometricians compare results from the current year with results from the past year at the cut points and the

percentage of all students in each performance level within the equating samples. After all quality control steps are completed and any differences are resolved, a senior psychometrician checks the scoring tables as the final step in quality control.

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 school districts and comparing these scores with those generated by Pearson. The selection of districts is based on the availability of data for all schools included in those districts, known as "pilot districts."

Year-to-Year Comparison Analyses

Year-to-year comparison analyses are conducted each year for quality control of the scoring procedure in general and as reasonableness checks for the CAPA results. Year-to-year comparison analyses use over 90 percent of the entire testing population to look at the tendencies and trends for the state as a whole as well as a few large districts.

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

Offloads to Test Development

The statistics based on classical task analyses and the IRT analyses are obtained at two different times in the testing cycle. The first time, the statistics are obtained on the equating samples to ensure the quality of equating, and then on larger sample sizes to ensure the stability of the statistics that are to be used for future test assembly. Statistics used to generate DIF flags are also obtained from the larger samples. The resulting classical, IRT, and DIF statistics for all items are provided to test development staff in specially designed Excel spreadsheets called "statistical offloads." The offloads are thoroughly checked by the psychometric staff before their release for test development review.

Quality Control of Reporting

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

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 school district mailings

All reports are required to include a single, accurate CDS code, a charter school number (if applicable), a school district 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.

For students for whom there is more than one answer document, the matching process, as described previously, provides for the creation of individual student records from which reports are created.

After the reports are validated against the CDE's requirements, a set of reports for pilot districts are 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 districts, Pearson proceeds with the first production batch test. The first production batch is selected to validate a subset of school districts that contain examples of key reporting characteristics representative of the state as a whole. The first production batch test incorporates CDE-selected school districts and provides the last check prior to generating all reports and mailing them to the districts.

Excluding Student Scores from Summary Reports

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

Reference

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

Chapter 10: Historical Comparisons

Base Year Comparisons

Historical comparisons of the CAPA results are routinely performed to identify the trends in examinee performance and test characteristics over time. Such comparisons were performed over a period of the three most recent years of administration—2010, 2011, and 2012—and the 2009 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 score reliability, and SEM, as well as the mean IRT *b*-value for each CAPA.

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

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

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

Examinee Performance

Table 10.A.1 on page 169 contains the number of examinees assessed and the means and standard deviations of examinees' scale scores in the base year (2009) and in 2010, 2011, and 2012 for each CAPA. As noted in previous chapters, the CAPA reporting scales range from 15 to 60 for all content areas and levels.

CAPA 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 169; please note that this information may differ slightly from information found on the CDE's STAR reporting Web page at <http://star.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. This goal for all students is consistent with school growth targets for state accountability and the federal requirements under the Elementary and Secondary Education Act.

Table 10.A.3 through Table 10.A.5 show for each CAPA the distribution of scale scores observed in the base year, in 2010, 2011, and 2012. Frequency counts are provided for each scale score interval of 3. A frequency count of "N/A" indicates that there are no obtainable scale scores within that scale-score range. For all CAPA, a minimum score of 30 is required for a student to reach the basic level of performance and a minimum score of 35 is required for a student to reach the proficient level of performance.

Test Characteristics

The item and test analysis results of the CAPA over the past several years indicate that the CAPA meets the technical criteria established in professional standards for high-stakes tests. In addition, every year, efforts are made to improve the technical quality of each CAPA.

Table 10.B.1 and Table 10.B.2 in Appendix 10.B, which starts on page 173, present, respectively, the average item scores and the mean equated IRT b -values for the tasks in each CAPA based on the equating samples. The average task scores are affected by both the difficulty of the items and the abilities of the students administered the tasks. The mean equated IRT b -values reflect only average item difficulty. Please note that comparisons of mean b -values should only be made within a given test; they should not be compared across test levels or content areas.

The average polyserial correlations for the CAPA are presented in Table 10.B.3. The reliabilities and standard errors of measurement (SEM) expressed in raw score units appear in Table 10.B.4. Like the average item score, polyserial correlations and reliabilities are affected by both item characteristics and student characteristics.

Appendix 10.A—Historical Comparisons Tables

Table 10.A.1 Number of Examinees Tested, Scale Score Means and Standard Deviations of CAPA Across Base Year (2009), 2010, 2011, and 2012

Content Area	CAPA	Number of Examinees (valid scores)				Scale Score Mean and Standard Deviation							
		Base	2010	2011	2012	Base		2010		2011		2012	
		Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
English–Language Arts	I	12,531	13,143	13,719	14,098	40.84	12.02	40.68	11.33	40.79	10.63	40.76	11.04
	II	6,587	6,682	6,643	6,668	39.24	7.46	38.54	6.25	38.72	5.70	38.82	6.91
	III	6,614	6,782	7,112	7,105	39.12	5.94	39.29	5.83	39.53	6.17	39.56	6.46
	IV	9,853	9,705	9,858	10,091	39.19	7.75	39.15	8.41	39.20	7.44	39.02	8.45
	V	10,517	10,443	10,217	10,424	38.54	6.21	38.73	6.59	38.88	6.44	38.72	6.04
Mathematics	I	12,484	13,111	13,689	14,065	35.11	9.74	35.87	9.30	36.08	8.70	36.15	9.00
	II	6,569	6,673	6,624	6,650	37.60	9.56	37.34	8.37	37.45	7.64	37.28	8.50
	III	6,602	6,770	7,098	7,094	36.58	6.64	36.50	5.80	36.33	4.99	36.34	5.54
	IV	9,831	9,676	9,845	10,068	36.41	8.80	37.15	8.91	37.26	8.11	37.14	7.50
	V	10,485	10,420	10,196	10,392	37.51	8.85	37.52	8.55	37.32	7.63	37.49	8.08
Science	I	3,296	3,490	3,512	3,564	35.59	11.25	36.49	11.13	36.20	10.67	36.25	10.25
	III	3,267	3,237	3,391	3,556	36.24	5.45	36.06	5.02	36.42	5.27	36.33	4.65
	IV	3,190	3,154	3,155	3,299	35.56	5.53	36.24	5.36	36.23	5.72	36.02	4.98
	V	3,396	3,325	3,245	3,424	35.35	5.34	35.69	4.79	35.82	5.08	36.22	5.21

Table 10.A.2 Percentage of Proficient and Above and Percentage of Advanced Across Base Year (2009), 2010, 2011, and 2012

Content Area	CAPA	% Proficient and Above				% Advanced			
		Base	2010	2011	2012	Base	2010	2011	2012
English–Language Arts	I	75%	78%	80%	81%	51%	55%	56%	59%
	II	78%	80%	85%	80%	41%	42%	42%	43%
	III	83%	86%	84%	81%	42%	46%	46%	54%
	IV	77%	72%	79%	72%	37%	43%	35%	40%
	V	80%	79%	79%	80%	42%	45%	48%	44%
Mathematics	I	61%	64%	69%	67%	29%	32%	33%	34%
	II	62%	66%	66%	65%	33%	35%	35%	35%
	III	65%	70%	72%	71%	31%	25%	18%	20%
	IV	60%	68%	67%	66%	31%	33%	32%	27%
	V	67%	68%	68%	69%	34%	37%	34%	33%
Science	I	59%	60%	58%	64%	33%	35%	35%	34%
	III	69%	70%	68%	71%	19%	16%	18%	18%
	IV	58%	71%	65%	66%	15%	18%	20%	14%
	V	61%	64%	68%	70%	17%	19%	20%	23%

Table 10.A.3 Observed Score Distributions of CAPA Across Base Year (2009), 2010, 2011, and 2012 for ELA

Observed Score Distributions	Level I			Level II			Level III			Level IV			Level V								
	Base	2010	2011	2012	Base	2010	2011	2012	Base	2010	2011	2012	Base	2010	2011	2012					
60	2,230	1,648	1,534	1,554	405	78	70	53	199	86	158	71	219	150	213	131	274	221	266	173	
57-59	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
54-56	N/A	456	488	473	N/A	N/A	N/A	83	N/A	N/A	N/A	N/A	239	189	235	208	N/A	N/A	N/A	N/A	N/A
51-53	624	N/A	N/A	352	N/A	100	91	142	N/A	118	N/A	149	N/A	236	N/A	235	N/A	N/A	N/A	N/A	N/A
48-50	388	615	580	276	375	133	156	192	304	249	300	178	653	650	664	692	400	266	346	247	
45-47	299	1,571	1,597	1,551	375	491	464	771	426	693	862	973	967	1,210	865	1,334	517	876	1,029	777	
42-44	1,708	1,555	1,768	1,876	795	1,011	978	960	934	1,191	1,004	1,305	1,534	1,764	1,475	1,415	1,277	1,634	1,320	1,556	
39-41	1,784	2,126	2,560	2,843	1,090	1,598	1,803	1,294	1,341	1,514	1,809	1,788	1,911	1,214	2,087	1,669	3,097	2,703	2,493	2,902	
36-38	1,567	1,885	1,883	1,940	1,776	1,613	1,487	1,483	2,044	1,502	1,360	1,068	1,669	1,636	1,795	1,220	2,179	2,251	2,190	2,241	
33-35	1,559	923	953	975	1,081	836	1,060	926	891	934	1,030	874	1,008	845	1,183	1,178	1,698	1,167	1,284	1,364	
30-32	694	599	853	570	362	507	230	292	258	255	326	255	822	787	611	887	572	824	782	672	
27-29	545	519	428	405	154	86	129	182	111	74	115	212	398	301	340	358	211	146	250	197	
24-26	140	167	151	154	89	88	64	99	45	58	79	84	83	333	109	310	113	42	91	130	
21-23	128	135	123	126	28	36	28	81	34	49	21	38	70	65	101	81	59	106	50	43	
18-20	128	113	156	N/A	12	25	24	31	5	14	16	22	125	57	37	137	33	57	32	28	
15-17	737	831	645	1,003	45	80	59	79	22	45	32	88	155	268	143	236	87	150	84	94	

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

Table 10.A.4 Observed Score Distributions of CAPA Across Base Year (2009), 2010, 2011, and 2012 for Mathematics

Observed Score Distributions	Level I			Level II			Level III			Level IV			Level V									
	Base	2010	2011	2012	Base	2010	2011	2012	Base	2010	2011	2012	Base	2010	2011	2012	Base	2010	2011	2012		
60	603	630	534	641	417	110	92	71	134	76	31	68	269	297	202	93	767	600	404	404	529	
57–59	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
54–56	N/A	N/A	N/A	N/A	N/A	138	90	107	N/A	N/A	N/A	N/A	N/A	N/A	N/A	130	N/A	N/A	N/A	N/A	N/A	N/A
51–53	N/A	N/A	N/A	N/A	386	N/A	N/A	147	N/A	N/A	N/A	N/A	391	364	N/A	268	N/A	N/A	N/A	N/A	N/A	N/A
48–50	237	297	245	282	N/A	208	143	321	230	95	47	94	295	383	296	361	N/A	N/A	381	269	356	
45–47	382	259	247	228	338	887	613	631	N/A	126	146	98	687	371	656	380	499	369	316	404	404	
42–44	934	1,111	1,235	1,263	682	659	1,071	749	762	625	497	533	689	1,344	1,246	925	1,104	997	1,158	1,063	1,063	
39–41	1,465	1,963	2,286	2,302	886	958	941	909	1,274	1,164	1,384	1,429	1,436	1,519	1,585	2,429	1,804	1,938	1,849	2,134	2,134	
36–38	2,775	3,252	3,636	3,619	1,049	1,172	1,052	1,131	1,579	1,979	2,550	2,043	1,687	1,611	1,957	1,664	2,475	2,356	2,439	2,296	2,296	
33–35	2,628	2,568	2,525	2,611	1,053	828	1,109	830	1,105	1,576	1,582	1,405	1,229	1,233	1,087	1,486	1,524	1,446	1,745	1,425	1,425	
30–32	1,053	946	973	1,061	658	729	803	564	837	645	413	856	1,319	644	1,102	1,163	918	917	1,022	1,060	1,060	
27–29	407	595	609	673	547	390	317	495	320	220	198	296	888	772	777	549	473	601	286	276	276	
24–26	492	216	174	195	137	259	97	354	200	133	81	101	286	562	219	193	278	245	434	554	554	
21–23	174	171	159	146	209	127	142	103	39	21	58	56	257	272	190	157	321	238	48	N/A	N/A	
18–20	177	162	161	N/A	34	32	29	53	33	14	18	18	75	64	54	54	61	90	37	59	59	
15–17	1,157	941	905	1,044	173	176	125	185	89	96	93	97	323	240	218	216	261	242	189	236	236	

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

Table 10.A.5 Observed Score Distributions of CAPA Across Base Year (2009), 2010, 2011, and 2012 for Science

Observed Score Distributions	Level II				Level III				Level IV				Level V			
	Base	2010	2011	2012	Base	2010	2011	2012	Base	2010	2011	2012	Base	2010	2011	2012
60	280	346	293	272	69	48	69	28	46	41	61	48	33	27	33	58
57-59	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
54-56	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
51-53	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
48-50	81	113	90	65	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
45-47	69	85	73	79	105	59	N/A	64	44	68	71	65	46	33	46	74
42-44	267	264	338	339	122	91	228	221	157	203	225	98	129	102	167	104
39-41	394	427	441	452	493	482	531	535	393	398	431	445	373	475	413	547
36-38	588	668	706	828	934	1,220	1,154	1,280	1,010	1,144	858	1,113	1,288	1,266	1,121	1,186
33-35	611	598	594	656	1,093	804	941	1,011	864	791	916	1,073	874	922	903	878
30-32	271	330	299	262	268	352	300	265	420	310	382	292	332	306	353	369
27-29	108	130	164	83	104	107	92	68	155	77	106	83	196	105	133	139
24-26	207	144	125	193	29	23	34	37	36	60	53	39	36	23	23	19
21-23	N/A	40	49	N/A	20	23	16	17	10	19	15	10	25	15	16	19
18-20	49	43	43	41	10	6	11	3	19	8	13	11	14	11	4	8
15-17	371	302	297	294	20	22	15	27	36	35	24	22	50	40	33	23

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

Appendix 10.B—Historical Comparisons Tables

Table 10.B.1 Average Item Score of CAPA Operational Test Items Across Base Year (2009), 2010, 2011, and 2012

Content Area	Level	Average Item Score			
		Base	2010	2011	2012
English–Language Arts	I	3.37	3.18	3.21	3.12
	II	2.91	2.43	2.51	2.38
	III	2.91	2.61	2.86	2.52
	IV	2.51	2.34	2.54	2.33
	V	2.73	2.54	2.74	2.57
Mathematics	I	2.70	2.80	2.85	2.86
	II	2.70	2.57	2.55	2.45
	III	2.70	2.48	2.32	2.39
	IV	2.37	2.57	2.47	2.49
	V	2.76	2.62	2.51	2.65
Science	I	2.75	2.98	2.91	2.91
	III	2.71	2.48	2.69	2.60
	IV	2.47	2.57	2.69	2.69
	V	2.47	2.44	2.58	2.74

Table 10.B.2 Mean IRT *b*-values for Operational Test Items Across Base Year (2009), 2010, 2011, and 2012

Content Area	Level	Mean IRT <i>b</i> -value			
		Base	2010	2011	2012
English–Language Arts	I	–0.74	–0.63	–0.61	–0.60
	II	–1.54	–0.91	–0.99	–0.76
	III	–1.52	–0.95	–1.32	–0.82
	IV	–0.93	–0.79	–1.05	–0.87
	V	–1.19	–0.89	–1.16	–0.91
Mathematics	I	–0.29	–0.23	–0.21	–0.24
	II	–1.18	–1.04	–1.02	–0.96
	III	–1.29	–0.95	–0.87	–0.93
	IV	–0.85	–1.02	–0.86	–0.81
	V	–1.21	–1.05	–0.98	–1.09
Science	I	–0.23	–0.37	–0.33	–0.31
	III	–1.29	–1.01	–1.29	–1.05
	IV	–0.95	–0.98	–1.19	–1.11
	V	–0.54	–0.39	–0.52	–0.65

Table 10.B.3 Mean Polyserial Correlation of CAPA Operational Test Items Across Base Year (2009), 2010, 2011, and 2012

Content Area	Level	Mean Polyserial Correlation			
		Base	2010	2011	2012
English–Language Arts	I	0.81	0.81	0.79	0.80
	II	0.75	0.75	0.72	0.78
	III	0.75	0.78	0.78	0.80
	IV	0.78	0.80	0.77	0.80
	V	0.79	0.81	0.80	0.79
Mathematics	I	0.79	0.77	0.75	0.76
	II	0.78	0.77	0.73	0.77
	III	0.76	0.73	0.68	0.71
	IV	0.79	0.79	0.75	0.73
	V	0.78	0.78	0.76	0.77
Science	I	0.82	0.81	0.80	0.79
	III	0.75	0.74	0.73	0.72
	IV	0.75	0.73	0.73	0.70
	V	0.78	0.74	0.76	0.76

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

Table 10.B.4 Score Reliabilities and SEM of CAPA Across Base Year (2009), 2010, 2011, and 2012

Content Area	Level	Reliability				SEM			
		Base	2010	2011	2012	Base	2010	2011	2012
English–Language Arts	I	0.91	0.90	0.88	0.89	3.67	3.80	3.86	3.89
	II	0.84	0.86	0.82	0.87	2.49	2.51	2.59	2.32
	III	0.86	0.88	0.88	0.90	2.26	2.26	2.19	2.17
	IV	0.88	0.90	0.86	0.90	2.50	2.32	2.48	2.33
	V	0.89	0.91	0.89	0.89	2.35	2.27	2.19	2.27
Mathematics	I	0.87	0.86	0.84	0.85	4.00	4.21	4.27	4.19
	II	0.88	0.86	0.82	0.86	2.58	2.64	2.64	2.60
	III	0.87	0.83	0.77	0.81	2.54	2.59	2.63	2.67
	IV	0.88	0.88	0.85	0.83	2.62	2.57	2.64	2.66
	V	0.87	0.88	0.85	0.86	2.70	2.62	2.73	2.76
Science	I	0.91	0.89	0.89	0.88	3.76	3.85	3.90	4.03
	III	0.85	0.85	0.84	0.84	2.43	2.51	2.49	2.45
	IV	0.85	0.84	0.84	0.81	2.46	2.47	2.32	2.47
	V	0.87	0.84	0.86	0.85	2.30	2.35	2.26	2.27