

California Department of Education Statewide Assessment Division



California Alternate Performance Assessment Technical Report Spring 2009 Administration

**Submitted February 8, 2010
Educational Testing Service
Contract No. 5417**

Table of Contents

Acronyms and Initialisms Used in the <i>CAPA Technical Report</i>	vi
Chapter 1: Introduction	1
Background	1
Test Purpose	1
Content	2
Intended Population	2
Intended Use and Purpose of Test Scores	2
Testing Window	3
Significant Developments in 2009	3
New Standard Performance Level	3
Limitations of the Assessment	3
Score Interpretation	3
Comparing CAPA Results.....	4
Verify CAPA Test Level	4
Groups and Organizations Groups Involved in Test Development	4
State Board of Education (SBE).....	4
California Department of Education (CDE)	4
Contractors	5
Overview of the Technical Report	5
References	7
Chapter 2: An Overview of CAPA Processes	8
Task Development	8
Task Formats	8
Task Development Specifications	8
Item Banking	8
Task Refresh Rate	8
Test Assembly	9
Test Length	9
Test Blueprint.....	9
Content Rules and Task Selection.....	9
Psychometric Criteria	9
Test Administration	10
Test Security and Confidentiality.....	10
Procedures to Maintain Standardization	11
Test Variations, Accommodations, and Modifications	11
Scores	11
Aggregation Procedures	11
Individual Scores.....	12
Group Scores.....	12
Equating	13
Calibration	13
Scaling	13
References	15
Chapter 3: Task (Item) Development	16
Rules for Task Development	16
Tasks Development Specifications	16
Expected Task Ratio	17
Selection of Task Writers	18
Criteria for Selecting Task Writers	18
Task Writer Training.....	18
Task Review Process	18
Content Expert Reviews	20
Assessment Review Panels.....	20
Field Testing	22
Stand-Alone Field Testing.....	22
Embedded Field Test Tasks	22
Data Review Meetings	23
Item Banking	24
References	25

Chapter 4: Test Assembly	26
Test Length	26
Rules for Task Selection	26
Test Blueprints	26
Content Rules and Task Selection	26
Psychometric Criteria	27
Rules for Task Sequence and Layout	28
Chapter 5: Test Administration	29
Test Security and Confidentiality	29
ETS’s Office of Testing Integrity (OTI)	29
Test Development	29
Task Review by ARPs	29
Item Bank for Tasks	30
Transfer of Forms and Tasks to the CDE	30
Firewall	30
Printing	31
Test Administration	31
Test Delivery	31
Processing and Scoring	32
Data Management	32
Transfer of Scores via Secure Data Exchange	33
Statistical Analysis	33
Reporting and Posting Results	33
Student Confidentiality	33
Test Results	33
Procedures to Maintain Standardization	34
Test Administrators	34
CAPA Examiner’s Manual	35
District and Test Site Coordinator Manual	35
STAR Management System Manuals	35
Accommodations for Students with Disabilities	36
Identification	36
Scoring	36
Demographic Data Corrections	37
Testing Irregularities	37
Test Administration Incidents	37
References	38
Chapter 6: Performance Standards	39
Background	39
Standard Setting Procedure	39
Development of Competencies Lists	40
Standard Setting Methodology	40
Performance Profile Method	40
Results	41
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 and Subscores	47
Raw Score	47
Scale Score	47
Performance Levels	47
Score Verification Procedures	47
Scoring Key Verification Process	47
Score Verification Process	48
Overview of Score Aggregation Procedures	48
Individual Scores	48
Score Distributions and Summary Statistics	48
Reports to be Produced and Scores for Each Report	52
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.....	55
Appendix 7.A—Scale Score Distribution Tables.....	56
Appendix 7.B—Demographic Summaries.....	63
Chapter 8: Analyses.....	72
Samples Used for the Analyses.....	72
Classical Analyses.....	73
Average Item Score (AIS).....	73
Polyserial Correlation of the Task Score with the Total Test Score.....	73
Reliability Analyses.....	75
Subgroup Reliabilities and SEMs.....	76
Conditional Standard Errors of Measurement.....	76
Decision Classification Analyses.....	77
Validity Evidence.....	78
Purpose of the CAPA.....	78
The Constructs to Be Measured.....	79
The Scores Generated and the Interpretations and Uses of these Scores.....	79
Intended Test Population(s).....	79
Validity Evidence Collected.....	79
Evidence Based on Internal Structure.....	83
Evidence Based on Consequences of Testing.....	83
IRT Analyses.....	84
IRT Model-Data Fit Analyses.....	84
Model Fit Assessment Results.....	85
Summaries of Scaled IRT b-values.....	85
Scaling Results.....	85
DIF Analyses.....	86
References.....	89
Appendix 8.A—Classical Analyses: Task Statistics.....	91
Appendix 8.B—Reliabilities.....	108
Appendix 8.C—Intercorrelations Between Content Areas.....	122
Appendix 8.D—IRT Analyses.....	144
Chapter 9: Quality Control Procedures.....	164
Quality Control of Task Development.....	164
Task Specifications.....	164
Task Writers.....	164
Internal Contractor Reviews.....	164
Assessment Review Panel (ARP) Review.....	165
Statewide Pupil Assessment Review (SPAR) Panel Review.....	165
Data Review of Field Tested Tasks.....	165
Quality Control of the Item Bank.....	166
Quality Control of Test Materials.....	166
Collecting Test Materials.....	166
Processing Test Materials.....	167
Quality Control of Scanning.....	167
Post-scanning Edits.....	167
Quality Control of Image Editing.....	167
Quality Control of Answer Document Processing and Scoring.....	168
Accountability of Answer Documents.....	168
Processing of Answer Documents.....	168
Scoring and Reporting Specifications.....	168
Matching Information on CAPA Answer Documents.....	169
Storing Answer Documents.....	169
Quality Control of Psychometric Processes.....	169
Score Key Verification Procedures.....	169
Quality Control of Task Analyses, DIF, and the Scoring Process.....	170
Score Verification Process.....	170
Offloads to Test Development.....	171
Quality Control of Reporting.....	171
Excluding Student Scores from Summary Reports.....	172

Tables

Table 1.1	Description of the CAPA Assessment Levels.....	2
Table 1.2	CAPA Levels.....	4
Table 2.1	CAPA Item and Estimated Time Chart.....	9
Table 2.2	Subgroup Definitions.....	12
Table 3.1	CAPA ARP Member Qualifications, by Subject and Total.....	21
Table 3.2	Summary of Tasks and Forms Presented in the 2009 CAPA Administration.....	23
Table 4.1	Target Statistical Specifications for the CAPA.....	28
Table 6.1	Scale Scores Ranges for Performance Levels.....	42
Table 6.2	Percentage of Examinees in Each Performance Level.....	42
Table 7.1	Rubrics for CAPA Scoring.....	45
Table 7.2	Summary Statistics Describing Student Scores: ELA.....	49
Table 7.3	Summary Statistics Describing Student Scores: Mathematics.....	50
Table 7.4	Summary Statistics Describing Student Scores: Science.....	51
Table 7.5	Subgroup Definitions.....	52
Table 7.6	Types of CAPA Reports.....	53
Table 7.A.1	Scale Score Frequency Distributions: Level I, ELA.....	56
Table 7.A.2	Scale Score Frequency Distributions: Level I, Mathematics.....	56
Table 7.A.3	Scale Score Frequency Distributions: Level I, Science.....	57
Table 7.A.4	Scale Score Frequency Distributions: Level II, ELA.....	57
Table 7.A.5	Scale Score Frequency Distributions: Level II, Mathematics.....	58
Table 7.A.6	Scale Score Frequency Distributions: Level III, ELA.....	58
Table 7.A.7	Scale Score Frequency Distributions: Level III, Mathematics.....	59
Table 7.A.8	Scale Score Frequency Distributions: Level III, Science.....	59
Table 7.A.9	Scale Score Frequency Distributions: Level IV, ELA.....	60
Table 7.A.10	Scale Score Frequency Distributions: Level IV, Mathematics.....	60
Table 7.A.11	Scale Score Frequency Distributions: Level IV, Science.....	61
Table 7.A.12	Scale Score Frequency Distributions: Level V, ELA.....	61
Table 7.A.13	Scale Score Frequency Distributions: Level V, Mathematics.....	62
Table 7.A.14	Scale Score Frequency Distributions: Level V, Science.....	62
Table 7.B.1	Demographic Summary for ELA, All Examinees.....	63
Table 7.B.2	Demographic Summary for Mathematics, All Examinees.....	65
Table 7.B.3	Demographic Summary for Science, All Examinees.....	67
Table 7.C.1	Score Reports Reflecting CAPA Results.....	69
Table 8.1	CAPA 2009 Raw Score Means and Standard Deviations: Total P2 Population and Calibration Sample.....	73
Table 8.2	Average Item Score and Polyserial Correlation.....	74
Table 8.3	Reliabilities and Standard Errors of Measurement for the CAPA.....	76
Table 8.4	CAPA Content Area Correlations for CAPA Levels.....	82
Table 8.5	DIF Flags Based on the ETS DIF Classification Scheme.....	87
Table 8.6	Subgroup Classification for DIF Analyses.....	87
Table 8.A.1	AIS and Polyserial Correlation: Level I, ELA.....	91
Table 8.A.2	AIS and Polyserial Correlation: Level II, ELA.....	92
Table 8.A.3	AIS and Polyserial Correlation: Level III, ELA.....	93
Table 8.A.4	AIS and Polyserial Correlation: Level IV, ELA.....	94
Table 8.A.5	AIS and Polyserial Correlation: Level V, ELA.....	95
Table 8.A.6	AIS and Polyserial Correlation: Level I, Mathematics.....	96
Table 8.A.7	AIS and Polyserial Correlation: Level II, Mathematics.....	97
Table 8.A.8	AIS and Polyserial Correlation: Level III, Mathematics.....	98
Table 8.A.9	AIS and Polyserial Correlation: Level IV, Mathematics.....	99
Table 8.A.10	AIS and Polyserial Correlation: Level V, Mathematics.....	100
Table 8.A.11	AIS and Polyserial Correlation: Level I, Science.....	101
Table 8.A.12	AIS and Polyserial Correlation: Level III, Science.....	102
Table 8.A.13	AIS and Polyserial Correlation: Level IV, Science.....	103
Table 8.A.14	AIS and Polyserial Correlation: Level V, Science.....	104
Table 8.A.15	Frequency of Operational Task Scores: ELA.....	105
Table 8.A.16	Frequency of Operational Task Scores: Mathematics.....	106
Table 8.A.17	Frequency of Operational Task Scores: Science.....	107
Table 8.B.1	Reliabilities and SEMs by GENDER.....	108
Table 8.B.2	Reliabilities and SEMs by PRIMARY ETHNICITY.....	109
Table 8.B.3	Reliabilities and SEMs by PRIMARY ETHNICITY for Economically Disadvantaged.....	110
Table 8.B.4	Reliabilities and SEMs by PRIMARY ETHNICITY for Not Economically Disadvantaged.....	111
Table 8.B.5	Reliabilities and SEMs by PRIMARY ETHNICITY for Unknown Economic Status.....	112
Table 8.B.6	Reliabilities and SEMs by Disability.....	113
Table 8.B.7	Decision Accuracy and Decision Consistency: Level I, ELA.....	114
Table 8.B.8	Decision Accuracy and Decision Consistency: Level I, Mathematics.....	115

Table 8.B.9 Decision Accuracy and Decision Consistency: Level I, Science	115
Table 8.B.10 Decision Accuracy and Decision Consistency: Level II, ELA	116
Table 8.B.11 Decision Accuracy and Decision Consistency: Level II, Mathematics	116
Table 8.B.12 Decision Accuracy and Decision Consistency: Level III, ELA	117
Table 8.B.13 Decision Accuracy and Decision Consistency: Level III, Mathematics	117
Table 8.B.14 Decision Accuracy and Decision Consistency: Level III, Science	118
Table 8.B.15 Decision Accuracy and Decision Consistency: Level IV, ELA	118
Table 8.B.16 Decision Accuracy and Decision Consistency: Level IV, Mathematics	119
Table 8.B.17 Decision Accuracy and Decision Consistency: Level IV, Science	119
Table 8.B.18 Decision Accuracy and Decision Consistency: Level V, ELA	120
Table 8.B.19 Decision Accuracy and Decision Consistency: Level V, Mathematics	120
Table 8.B.20 Decision Accuracy and Decision Consistency: Level V, Science	121
Table 8.C.1 Raw Score Correlations by Gender: Level I	122
Table 8.C.2 Raw Score Correlations by Gender: Level II	122
Table 8.C.3 Raw Score Correlations by Gender: Level III	122
Table 8.C.4 Raw Score Correlations by Gender: Level IV	122
Table 8.C.5 Raw Score Correlations by Gender: Level V	122
Table 8.C.6 Raw Score Correlations by Ethnicity: Level I	123
Table 8.C.7 Raw Score Correlations by Ethnicity: Level II	123
Table 8.C.8 Raw Score Correlations by Ethnicity: Level III	124
Table 8.C.9 Raw Score Correlations by Ethnicity: Level IV	124
Table 8.C.10 Raw Score Correlations by Ethnicity: Level V	125
Table 8.C.11 Raw Score Correlations by Ethnicity for Economically Disadvantaged: Level I	125
Table 8.C.12 Raw Score Correlations by Ethnicity for Economically Disadvantaged: Level II	126
Table 8.C.13 Raw Score Correlations by Ethnicity for Economically Disadvantaged: Level III	126
Table 8.C.14 Raw Score Correlations by Ethnicity for Economically Disadvantaged: Level IV	127
Table 8.C.15 Raw Score Correlations by Ethnicity for Economically Disadvantaged: Level V	127
Table 8.C.16 Raw Score Correlations by Ethnicity for Not Economically Disadvantaged: Level I	128
Table 8.C.17 Raw Score Correlations by Ethnicity Not Economically Disadvantaged: Level II	128
Table 8.C.18 Raw Score Correlations by Ethnicity Not Economically Disadvantaged: Level III	129
Table 8.C.19 Raw Score Correlations by Ethnicity Not Economically Disadvantaged : Level IV	129
Table 8.C.20 Raw Score Correlations by Ethnicity Not Economically Disadvantaged : Level V	130
Table 8.C.21 Raw Score Correlations by Ethnicity for Unknown Economic Status : Level I	130
Table 8.C.22 Raw Score Correlations by Ethnicity for Unknown Economic Status : Level II	131
Table 8.C.23 Raw Score Correlations by Ethnicity for Unknown Economic Status : Level III	131
Table 8.C.24 Raw Score Correlations by Ethnicity for Unknown Economic Status : Level IV	132
Table 8.C.25 Raw Score Correlations by Ethnicity for Unknown Economic Status : Level V	132
Table 8.C.26 Raw Score Correlations by Economic Status: Level I	133
Table 8.C.27 Raw Score Correlations by Economic Status: Level II	133
Table 8.C.28 Raw Score Correlations by Economic Status : Level III	133
Table 8.C.29 Raw Score Correlations by Economic Status : Level IV	133
Table 8.C.30 Raw Score Correlations by Economic Status: Level V	133
Table 8.C.31 Raw Score Correlations by Disability: Level I	134
Table 8.C.32 Raw Score Correlations by Disability: Level II	135
Table 8.C.33 Raw Score Correlations by Disability: Level III	136
Table 8.C.34 Raw Score Correlations by Disability: Level IV	137
Table 8.C.35 Raw Score Correlations by Disability: Level V	138
Table 8.C.36 Inter-Rater Reliabilities for Operational Tasks: Level I	139
Table 8.C.37 Inter-Rater Reliabilities for Operational Tasks: Level II	140
Table 8.C.38 Inter-Rater Reliabilities for Operational Tasks: Level III	141
Table 8.C.39 Inter-Rater Reliabilities for Operational Tasks: Level IV	142
Table 8.C.40 Inter-Rater Reliabilities for Operational Tasks: Level V	143
Table 8.D.1 Item Classifications for Model-Data Fit Across All CAPA Levels	144
Table 8.D.2 Fit Classifications: Level I Tasks	144
Table 8.D.3 Fit Classifications: Level II Tasks	144
Table 8.D.4 Fit Classifications: Level III Tasks	144
Table 8.D.5 Fit Classifications: Level IV Tasks	144
Table 8.D.6 Fit Classifications: Level V Tasks	144
Table 8.D.7 IRT <i>b</i> -values for ELA, by Level	145
Table 8.D.8 IRT <i>b</i> -values for Mathematics, by Level	145
Table 8.D.9 IRT <i>b</i> -values for Science, by Level	145
Table 8.D.10 Score Conversions: Level I, ELA	146
Table 8.D.11 Score Conversions: Level II, ELA	147
Table 8.D.12 Score Conversions: Level III, ELA	148
Table 8.D.13 Score Conversions: Level IV, ELA	149

Table 8.D.14	Score Conversions: Level V, ELA.....	150
Table 8.D.15	Score Conversions: Level I, Mathematics	151
Table 8.D.16	Score Conversions: Level II, Mathematics	152
Table 8.D.17	Score Conversions: Level III, Mathematics	153
Table 8.D.18	Score Conversions: Level IV, Mathematics\.....	154
Table 8.D.19	Score Conversions: Level V, Mathematics	155
Table 8.D.20	Score Conversions: Level I, Science	156
Table 8.D.21	Score Conversions: Level III, Science	157
Table 8.D.22	Score Conversions: Level IV, Science.....	158
Table 8.D.23	Score Conversions: Level V, Science.....	159
Table 8.E.1	Item Exhibiting Significant DIF by Ethnic Group	160
Table 8.E.2	Items Exhibiting Significant DIF by Disability Group	160
Table 8.E.3	CAPA Disability Distributions: Level I	161
Table 8.E.4	CAPA Disability Distributions: Level II	162
Table 8.E.5	CAPA Disability Distributions: Level III	162
Table 8.E.6	CAPA Disability Distributions: Level IV	163
Table 8.E.7	CAPA Disability Distributions: Level V	163

Figures

Figure 3.1	The ETS Item Development Process for STAR	16
Figure 8.1	Decision Accuracy for Achieving a Performance Level.....	77
Figure 8.2	Decision Consistency for Achieving a Performance Level	77

Acronyms and Initialisms Used in the *CAPA Technical Report*

1PPC	1-parameter partial credit	IEP	individualized education program
ADA	Americans with Disabilities Act	IRT	task (item) response theory
AIS	average task (item) score	LEA	local educational agency
API	Academic Performance Index	MH	Mantel-Haenszel
ARP	Assessment Review Panel	NPS	nonpublic, nonsectarian school
AYP	adequate yearly progress	NSLP	National School Lunch Program
CAPA	California Alternate Performance Assessment	RACF	Random Access Control Facility
CMA	California Modified Assessment	RS	raw score
CDE	California Department of Education	SBE	State Board of Education
CDS	County-District-School	SD	standard deviation
CI	confidence interval	SEM	standard error of measurement
CSEMs	conditional standard errors of measurement	SFTP	secure file transfer protocol
CSTs	California Standards Tests	SGID	School and Grade Identification sheet
DIF	Differential Task (Item) Functioning	SKM	score key management
DPLT	designated primary language test	SMD	standardized mean difference
ESEA	Elementary and Secondary Education Act	SPAR	Statewide Pupil Assessment Review
ETS	Educational Testing Service	SR&T	Scoring, Reporting and Technology
GENASYS	Generalized Analysis System	STAR	Standardized Testing and Reporting
HumRRo	Human Resource Research Organization	STAR TAC	STAR Technical Assistance Center
ICC	task (item) characteristic curve	STS	Standards-based Tests in Spanish

Chapter 1: Introduction

Background

In 1997 and 1998, the California State Board of Education (SBE) adopted rigorous content standards in four major content areas: English–language arts (ELA), mathematics, history–social science, 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.

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). Senate Bill 1448, approved by the Legislature and the Governor in August 2004, reauthorized the STAR Program through January 1, 2011, in grades three through eleven. STAR Program testing in grade two has also been extended to the 2011 school year (spring 2011 administration) after Senate Bill 80 was passed in September 2007.

During its 2009 administration, the STAR Program had four components:

- California Standards based 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 plan (IEP) who meet the SBE-adopted eligibility criteria (In 2009, the CMA was administered for ELA in grades three through eight, for mathematics in grades three through seven, and for science in grades five and eight.)
- California Alternate Performance Assessment (CAPA), produced for students with significant cognitive disabilities who have an IEP and are not able to take the CSTs or the CMA
- 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) (In 2009, the STS was administered for reading/language arts in grades two through eleven, for grade-level mathematics in grades two through seven, for end-of-course [EOC] Algebra I in grades seven through eleven, and for EOC Geometry in grades eight through eleven.)

Test Purpose

The CAPA is designed to show how well students with significant cognitive disabilities are performing relative to California's content standards for ELA and mathematics in grades two through eleven and relative to 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. IEP teams determine on a student-by-student basis whether a student takes the CST/CMA or the CAPA. CAPA results are used in the calculation of each school and district Academic Performance Index (API).

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.

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 with the most significant cognitive disabilities (They may be in grades two through eleven.)
- 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 below displays the tests administered in 2009 by grade and content area.

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	—	Science	Science	Science
	Grades 5, 8, and 10 only		Grade 5 only	Grade 8 only	Grade 10 only

Intended Population

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

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 California *Education Code* Section 60602, <http://www.leginfo.ca.gov/cgi-bin/displaycode?section=edc&group=60001-61000&file=60600-60603>):

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

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

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

“60602. (d) It is the intent of the Legislature, insofar as is practically feasible and following the completion of annual testing, that the content, test structure, and test items in the assessments that are part of the Standardized Testing and Reporting Program become open and transparent to teachers, parents, and pupils, to assist all the stakeholders in working together to demonstrate improvement in pupil academic achievement. A planned change in annual test content, format, or design, should be made available to educators and the public well before the beginning of the school year in which the change will be implemented.”

In addition, STAR program assessments are used to provide data for state and federal accountability purposes.

Testing Window

The CAPA tests are administered at different times, depending on the progression of the school year within each particular school district. Specifically, schools must administer the CSTs, CMA, CAPA, and STS tests within a 21-day window, which begins 10 days before and ends 10 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, Title 5, Education, Division 1, Chapter 2, Subchapter 3.75, Article 2, § 855*; <http://www.cde.ca.gov/ta/tg/sr/starregs0207cln.doc>).

Significant Developments in 2009

New Standard Performance Level

A standard setting was held in the fall of 2008 to establish new cut scores for the below basic, basic, proficient, and advanced performance categories for Levels I through V in ELA and mathematics and Levels I and III through V in science. Also a new scale for reporting CAPA test results was developed in the spring of 2009. Thus the test results for spring 2009 CAPA administration were reported using new scales and cut scores for the four performance categories.

Limitations of the Assessment

Score Interpretation

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, 2009). It is also important to note that

statistically, a student’s score in a content area probably contains a standard measurement error and could vary somewhat if the student was retested.

Comparing CAPA Results

When comparing results for the CAPA, the user is limited to score comparisons within the same subject and CAPA level within or across test years. For example, it is appropriate to compare scores obtained on the 2009 CAPA Level II (Mathematics) test with those obtained on the 2010 CAPA Level II (Mathematics) test. Similarly it is appropriate to compare scores obtained on the 2009 CAPA Level IV (ELA) test with those obtained on the CAPA Level IV (ELA) test administered in 2010. 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 2009 cannot meaningfully be compared to results obtained in previous years.

Verify CAPA Test Level

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.

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 K–12 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 should be made on the basis of both the student’s CAPA performance from the previous year and on classroom assessments.

CAPA levels are shown in Table 1.2.

Table 1.2 CAPA Levels

CAPA Level	Grade Range	Subjects	Age Ranges for Ungraded Programs
I	2–11	ELA, mathematics, science	7–16
II	2 & 3	ELA, mathematics	7 & 8
III	4 & 5	ELA, mathematics, science	9 & 10
IV	6–8	ELA, mathematics, science	11–13
V	9–11	ELA, mathematics, science	14–16

Groups and Organizations Groups Involved in Test Development

State Board of Education (SBE)

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 *Education Code*.

California Department of Education (CDE)

The CDE oversees the California public school system, which is responsible for the education of more than seven million children and young adults in more than 9,000 schools. The CDE’s mission is to provide leadership, assistance, oversight, and resources so that every child in California has access to an educational system that meets world-class

standards. As part of its mission to promote district and school accountability for improving student achievement as defined by the SBE, the CDE oversees the development and administration of the STAR Program.

Contractors

Educational Testing Service

The CDE and the SBE contracted with Educational Testing Service (ETS) to develop and administer the STAR Program. As the prime contractor, ETS has overall responsibility for working with the CDE to improve its overall assessment system and to coordinate the work of ETS and its subcontractor Pearson. Activities directly conducted by ETS include: overall management of the program activities; development of all test questions; 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: production of all scannable test materials; packaging, distribution, and collection of test materials to school districts; 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 2009. The technical report contains eight additional chapters as follows:

- Chapter 2 presents a conceptual overview of processes involved in a testing cycle for the 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.
- Chapter 3 describes the procedures followed during the development of valid CAPA tasks; the chapter explains the process of field-testing new items and the review of tasks by contractors and content experts.
- Chapter 4 details the content and psychometric criteria applicable to the construction of CAPA for 2009.
- Chapter 5 presents the processes involved in the actual administration of the 2009 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 conducted to establish new cut scores. In addition, descriptions of students' proficiency classifications are also provided.
- 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 item-level analyses performed during the spring 2009 administration of the tests. These include the classical item analyses, differential item functioning (DIF), item response theory (IRT) and model-fit analyses, as well as documentation of the equating along with CAPA conversion tables. Also

discussed in this chapter are the procedures designed to ensure the validity of CAPA score uses and interpretations. Finally, the chapter summarizes the results of reliability analyses that include assessments of test reliability and the consistency and accuracy of the CAPA proficiency-level classifications.

- Chapter 9 highlights the importance of controlling and maintaining the quality of the CAPA. 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 Department of Education. (2009). *Interpreting 2009 STAR program test results*. Sacramento, CA. <http://www.cde.ca.gov/ta/tg/sr/documents/star09intrprslt.pdf>.

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 carry out 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 Development

Task Formats

Each CAPA item 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 for each task.

Task Development 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 task development specifications 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 test undergo an extensive review process that is designed to provide the best standards-based tests possible. Details about the task development specifications, the task review process, the item utilization plan, and the rules for arranging tasks on the forms are presented in Chapter 3, starting on page 16.

Item Banking

The newly developed tasks are field tested to obtain information about task performance and to obtain statistics that can be used to assemble operational forms. Once tasks have been field tested, ETS prepares the tasks and the associated statistics for review by the content experts. The tasks are then placed in the item bank along with their corresponding review information. Tasks that are accepted by the content experts 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 STAR 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 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 24 in Chapter 3.

Task Refresh Rate

The Item Utilization Plan assumes that each year, 50 percent of tasks on an operational form are refreshed; these tasks remain in the item bank for future use. In addition, the plan notes that five percent of the operational items are likely to become unusable because of

normal attrition and there is a need to focus development on what are called “critical” standards, which are standards that are difficult to measure well.

Test Assembly

Test Length

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 required within the school district’s selected testing window.

Table 2.1 CAPA Item and Estimated Time Chart

ITEM 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 Blueprint

ETS selects all CAPA test tasks to conform with 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 CAPA were approved by the SBE in 2006 for implementation beginning in 2008. CAPA blueprints 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 content-area test for the CAPA, test developers follow a number of rules. First and foremost, they select tasks that comply with the blueprint for that the CAPA level and content area to be assessed. Using an electronic item bank, assessment specialists typically begin by identifying a number of linking tasks. These are tasks that appeared in the previous year’s operational administration and they are used to equate the test forms administered each year. After the linking tasks are approved, assessment specialists populate the rest of the test form. In 2009, the CAPA test forms used in 2008 were re-used and a new score scale was developed, so no linking back to the previously used score scales was required.

During test development, 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 26.

Psychometric Criteria

For the CAPA, the test developers and psychometricians strive to accomplish three goals while developing a test:

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 is developed that outlines the desired psychometric properties of each CAPA test. Such rules are referred to as statistical targets. An assembly targets are developed for each CAPA: the total test target . These targets are provided to test developers before a test construction cycle begins. The test developers and psychometricians work together in making efforts to design the tests to these test targets. The staff also assesses the projected test characteristics during the preliminary review of the assembled forms. These target values are presented in Chapter 4, in Table 4.1 on page 28.

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 also described in Chapter 4.

Test Administration

It is of utmost priority to ETS to administer the CAPA in an appropriate, consistent, 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 the tests that ETS develops or handles, ETS maintains an Office of Testing Integrity (OTI). A detailed description of the OTI and its mission is presented in Chapter 5 on page 29.

In 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 process are discussed in detail in Chapter 5, starting on page 29.

- Test development process
- Task and data review process
- Item banking
- Transfer of forms and tasks to the CDE
- Security of electronic files via 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 CAPA tests, as described in this section.

Test Administrators

The CAPA are administered in conjunction with the other tests that comprise the STAR Program. In that respect, 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 coordinators, and test examiners. The responsibilities for each of the staff members are included in the *STAR District and Test Site Coordinator Manual* (CDE, 2009), which is presented in more detail on page 35 in Chapter 5.

Test Directions

ETS maintains a series of instructions compiled in detailed manuals that are available to the test administrators. Such documents include, but are not limited to, the following:

CAPA Examiner’s Manual—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

District and Test Site Coordinator Manual—Test administration procedures for district STAR coordinators and test site coordinators (see page 35 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

Test Variations, Accommodations, and Modifications

All public school students participate in the STAR Program, including English learners and students with disabilities. Students with an IEP 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.

Scores

The CAPA total raw scores equal the sum of examinees’ scores on the 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. Those raw scores are transformed to two-digit scale scores using the scaling process described starting on page 41 in Chapter 6. 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 classifying the scale scores into the following categories: 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 on page 47 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, independent charter school, district,

county, and state levels. The aggregated scores are generated both for individual scores as well as group scores. The following section presents the types of aggregation performed on CAPA scores.

Individual Scores

Table 7.2 through Table 7.4 in Chapter 7 are summary statistics for individual scores that describe student performance on each CAPA. Included in the tables are the means and standard deviations of student scores expressed in terms of both raw scores and scale scores; the raw score means and standard deviations expressed as percentages of the total raw score points in each test; and statistical information about the CAPA tasks.

Group 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 in Appendix 7B. In these tables, students are grouped by demographic characteristics, including gender, ethnicity, English proficiency, primary disability, and economic status. The tables show the numbers of valid cases in each group, scale score means and standard deviations as well as percentage in performance level for each demographic group. Table 2.2 defines the demographic groups included in the tables. Students’ economic status is determined by considering the education level of their parents and whether or not they were eligible for the National School Lunch Program (NSLP).

Table 2.2 Subgroup Definitions

Subgroup	Definition
Gender	Male
	Female
Ethnicity	American Indian or Alaska Native
	Asian
	– Chinese
	– Japanese
	– Korean
	– Vietnamese
	– Asian Indian
	– Laotian
	– Cambodian
	– Other Asian
	Pacific Islander
	– Native Hawaiian
	– Guamanian
	– Samoan
	– Tahitian
– Other Pacific Islander	
English Language Fluency	Filipino
	Hispanic or Latino
	African American
	White (not Hispanic)
Economic Status	English-language fluency
	Initially fluent English proficient
	English learner
	Reclassified fluent English proficient
Economic Status	Not economically disadvantaged
	Economically disadvantaged

Subgroup	Definition
Primary Disability	Mental retardation
	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

Equating

In 2009 there was no equating of the CAPA tests to previously used scales because the 2008 CAPA tests were re-used in 2009. In addition, a new scale for the CAPA tests was developed using 2009 test data.

Calibration

The operational tasks in each CAPA were calibrated using a proprietary version of the PARSCALE program and Rasch partial credit model. The estimation process was constrained by setting a common discrimination value for all tasks equal to 1.0/1.7 (or 0.588). This approach 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

The item calibrations for all CAPA tests were used to generate raw score to theta scoring tables. The thetas in these tables then were linearly transformed to a two-digit score scale that ranged from 15 to 60. Since 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:

$$ScaleScore = \left(35 - \hat{\theta}_{pro} \times \left(\frac{35 - 30}{\hat{\theta}_{pro} - \hat{\theta}_{bas}} \right) \right) + \left(\frac{35 - 30}{\hat{\theta}_{pro} - \hat{\theta}_{bas}} \right) \times \hat{\theta} \quad (2.1)$$

where,

$\hat{\theta}$ represents student ability

$\hat{\theta}_{pro}$ represents theta cut score for proficient on spring 2009 base scale

$\hat{\theta}_{bas}$ represents theta cut score for basic on spring 2009 base scale

Complete raw-to-scale score conversion tables for the 2009 CAPA are presented in Table 8.D.10 through Table 8.D.23 in Appendix 8.D, starting on page 146. The raw scores and corresponding unrounded converted scale scores are listed in those tables. The scale scores defining the cut scores for all performance levels are presented in Table 6.1, which is on page 42 in Chapter 6.

References

California Department of Education. (2009). *2009 STAR district and test site coordinator manual*. Sacramento, CA. http://www.startest.org/pdfs/STAR.coord_man.2009.pdf.

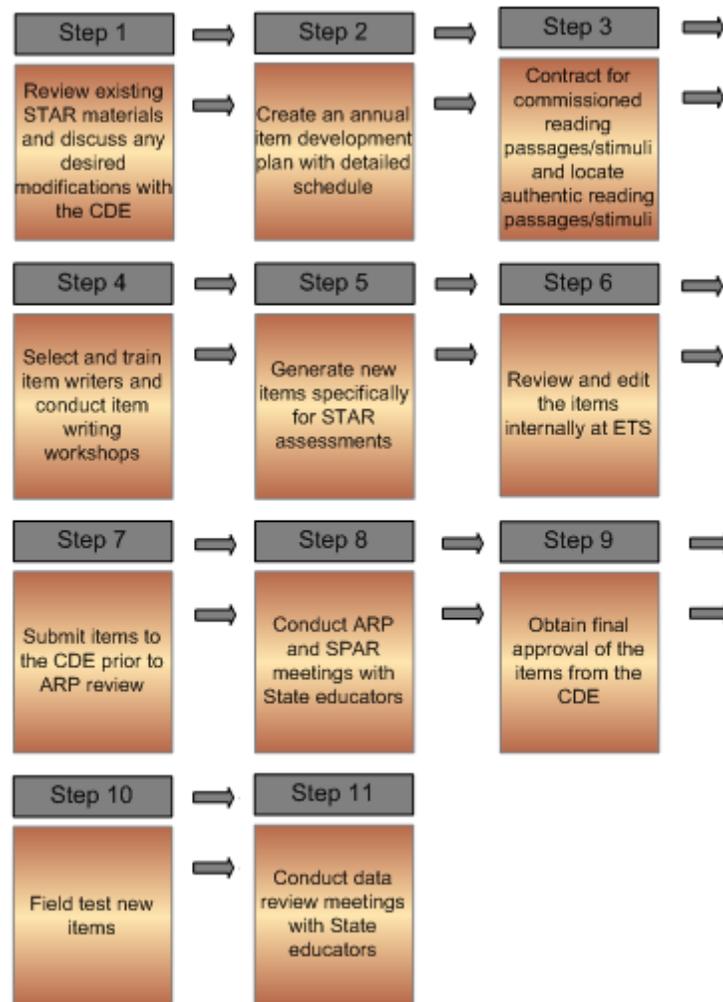
Educational Testing Service. (2002). *ETS standards for quality and fairness*. Office of Testing Integrity, Princeton, NJ: Educational Testing Service.

Linacre, J. M. (2000). *WINSTEPS: Rasch measurement* (Version 3.23). Chicago, IL: MESA Press.

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 STAR



Rules for Task Development

The development of CAPA tasks follow 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 item 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.

Tasks Development 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 subject 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 tests measure 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 stimuli (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

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 has developed the Item Utilization Plan to continue the development of CAPA tasks. The plan includes strategies for developing items 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 items to develop for each subject area.

The Item Utilization Plan assumes that each year, 50 percent of items on an operational form are refreshed; these items remain in the item bank for future use. In addition, the plan notes that five percent of the operational items are likely to become unusable because of normal attrition, and there is a need to focus development on what are called “critical” standards, which are standards that are difficult to measure well.

For the CAPA tests, ETS field test 150 percent for each operational form for each content area each year. Given that each operational form contains eight tasks, this means that twelve new tasks should be field-tested each year. This proportion would allow for a five percent attrition rate, while gradually increasing the overall size of the CAPA task bank.

Selection of Task Writers

Criteria for Selecting Task Writers

The tasks selected for each CAPA test are written by special panels of task writers that have a thorough understanding of the California content standards. Applicants for task writing are screened by senior ETS content staff. Only applicants 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 in 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 Writer Training

Task writer training was conducted over two days in Newport Beach, California, in 2009. An effort was made to evenly distribute the participants across the CAPA content areas. At this session, ETS test development specialists trained attendees in the basics of task writing. They also reviewed tasks that participants created during the training, offering feedback in both group and individual settings.

Task Review Process

After 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 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 and the ARP and, later, 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 from 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 criteria:

- 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
- 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 artwork, graphs, figures, or other illustrations

The assessment specialists also 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 accept the task and classification as written, suggest revisions, or recommend that the task be discarded. These steps occur prior to CDE 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 questions 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. In addition, 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

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 on matters related to task development. The composition of the ARPs is presented in Table 3.1, on the next page. The ARPs are responsible for reviewing all newly developed tasks for alignment to the California content standards. The ARPs reviewed the tasks for accuracy of content, clarity of phrasing, and quality. ETS provided the ARPs with the opportunity to review the tasks with the applicable field-test statistics and to make recommendations for the use of tasks in subsequent test forms. For example, the ARPs, in their examination of test tasks, could raise concerns related to age/level appropriateness and gender, racial/ethnic, 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 the following:

- Three or more years of general teaching experience in levels kindergarten through grade twelve and in the content areas (English–language arts, mathematics, or science)
- Possession of a bachelor’s degree or a higher degree in a grades or subject area related to English–language arts, mathematics, or science
- Knowledge and experience with the California content standards for English–language arts, mathematics, or science
- Special education credential
- Experience with more than one type of disability
- 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 different gender and ethnic groups as well as representation from different geographic regions 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 final approval. Upon approval, the applicant is notified that he or she has been selected to serve on the ARP committee. Table 3.1 shows the educational qualifications, present occupation, and credentials of the current CAPA ARP members.

Table 3.1 CAPA ARP Member Qualifications, by Subject and Total

CAPA	ELA	Math	Science		Total
Total	8	6	6		20
Occupation (Members may teach multiple levels.)					
Teacher or Program Specialist, Elementary/Middle School	3	2	0		5
Teacher or Program Specialist, High School	1	0	3		4
Teacher or Program Specialist, K–12	3	3	3		9
University Personnel	0	0	0		0
Other District Personnel (e.g., Director of Special Services, etc.)	1	1	0		2
Highest Degree Earned					
Bachelor's Degree	3	2	0		5
Master's Degree	5	4	6		15
Doctorate	0	0	0		0
Credential (Members may hold multiple credentials.)					
Elementary Teaching (multiple subjects)	4	3	0		7
Secondary Teaching (single subject)	0	1	4		5
Special Education	5	4	5		14
Reading Specialist	0	0	0		0
English Learner (CLAD, BCLAD)	1	0	1		2
Administrative	1	2	1		4
Other	0	0	0		0
None (teaching at the university level)	0	0	0		0

ARP Meetings for Review of CAPA Tasks

The 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's task specifications
- Overview of criteria for reviewing constructed-response writing tasks
- Review and evaluation of tasks for bias and sensitivity issues

Criteria also involve more global issues, 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 answer 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, panel 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. Those recommendations are recorded 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 consider as 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 (SPAR)

The SPAR panel is responsible for reviewing and approving the tests 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 *Education Code* Section 60614. If the SPAR panel rejects specific tasks, the tasks are replaced with other tasks that are acceptable to the SPAR panel that measure the same standard. For the SPAR panel meeting, the task development coordinator or an ETS content specialist, requested in advance by the CDE, attends the opening session and remains in a nearby location or near a telephone to be available to respond to any questions during the course of the meeting.

Field Testing

The primary purpose of field testing is 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 test, 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 situation, 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 scattered 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 items being field-tested within each operational test form. Different forms contain the same operational items and different field test tasks.

Allocation of Students to Forms

The operational test forms for a given CAPA version are spiraled among students in the state by assigning specific versions to school districts and independently testing charter schools so that a large representative sample of test takers responds to the tasks being field-tested that are embedded in these forms. The spiraling design ensures that a diverse sample of students take each task being field tested. 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 eight versions. Each version 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.

Table 3.2 provides information about the numbers of test forms, operational tasks, field-test tasks, and the approximate number of students in the P2¹ sample that took the operational and field-test tasks in Spring 2009. The sample sizes for the field tests are presented as ranges because the numbers of students who took a set of field-test items varied over the versions of CAPA.

Table 3.2 Summary of Tasks and Forms Presented in the 2009 CAPA Administration

Subject	Level	Operational			Field Test	
		# Items	Examinees Total (P2)	# Forms	# Items	Examinees Total (P2)
<i>English– Language Arts</i>	I	8	12,531	8	4	1,199–1,775
	II	8	6,587	8	4	636–793
	III	8	6,614	8	4	573–860
	IV	8	9,853	8	4	896–1,311
	V	8	10,517	8	4	1,031–1,352
<i>Mathematics</i>	I	8	12,484	8	4	1,188–1,767
	II	8	6,569	8	4	634–792
	III	8	6,602	8	4	572–859
	IV	8	9,831	8	4	894–1,312
	V	8	10,485	8	4	1,026–1,347
<i>Science</i>	I	8	3,296	8	4	304–489
	III	8	3,267	8	4	296–441
	IV	8	3,190	8	4	291–416
	V	8	3,396	8	4	315–432

Data Review Meetings

Once tasks have been field tested, ETS prepares the tasks and statistics for review by the ARPs. ETS assessment specialists facilitate the data review sessions with qualified psychometric staff on hand for technical assistance. Upon completion of the meeting, ETS provides the CDE with summaries of the recommendations based on the field-test analyses and committee reviews that are relevant to future form construction of the CAPA test. All final decisions on acceptance of tasks rest with the CDE in consultation with the SBE staff.

At data review meetings, the ARP members discuss tasks that have “poor” statistics and do not meet the psychometric criteria for task quality. The CDE defines the criteria for

¹ The P2 data file contains 100 percent of school district data that were received for ETS Statistical Analysis by approximately August 29, 2009

acceptable or unacceptable task statistics. These criteria ensure that the item (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 panel members also use the results of analyses for differential item functioning (DIF) to make judgments about the appropriateness of tasks for various subgroups.

Item Banking

Once the ARP review is complete, the tasks are placed in the item bank along with their corresponding review information. Tasks that are accepted by the ARP 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 STAR electronic item bank. Subsequent updates to tasks are based on field-test and operational use. However, only the latest version of the task is in the bank at any time, along with the administration data from every administration that has included the task.

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

Status and availability of an item are updated programmatically as items are presented for review, accepted or rejected, placed on a form for field testing, presented for statistical review, used operationally, or released. All rejection and release indications are monitored and controlled through ETS’s assessment development processes.

References

- Educational Testing Service (2002). *ETS standards for quality and fairness*. Office of Testing Integrity. Princeton, NJ: Educational Testing Service.
- 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 (Fourth edition)*. 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. The tests are also constructed to meet professional standards for validity and reliability. For the CAPA, the content standards and psychometric attributes are used as the basis for assembling the test forms.

Test Length

The number of tasks in each CAPA content area is decided 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 California content standards for each grade and content area. Also considered is the goal that the tests be short enough that most of the students complete the test in a reasonable amount of time.

All CAPA assessments consist of eight versions. Each version 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 students' scores. See Table 2.1 on page 9 for more details on the test length.

Rules for Task Selection

Test Blueprints

ETS develops all CAPA test tasks to conform to the SBE-approved California content standards and the CAPA 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 following 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 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 grade and content area. Using the 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 they are used to equate the test forms administered each year. 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.

For the task selection system in STAR test development, ETS continues to use the STAR Item Bank System. The item bank provides the test developer with a worksheet of tasks that are eligible to be selected for a new administration. Various statistical and classification information is available for these tasks, such as task number, status, standard code, answer key, and statistical information (for example, polyserial correlation coefficient, average task score, and distribution of score points, and IRT b -value).

The CAPA is assembled to content and statistical specifications or targets. Each form contains some tasks that are the same as those used in the previous year; these are called linking or equating tasks. The statistics used to select the linking tasks are obtained from the

previous year's operational administration. The nonlinking task statistics are generally based on the field tests.

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 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. In 2009, the CAPA test forms used in 2008 were re-used, so there is no linking task selection.

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 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 on page 16 contains further information about this process.

Psychometric Criteria

For CAPA, the test developers and psychometricians strive to accomplish three goals while developing a test:

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 is developed that outlines the desired psychometric properties of the CAPA. Such rules are referred to as statistical targets. Two types of assembly targets are developed for the CAPA: the total test target and the linking block target. These targets are provided to test developers before a test construction cycle begins.

The total test target or primary statistical targets used for assembling the CAPA for the 2008 administration were the test information function based on the item (task) response theory (IRT) item parameters and an average point biserial correlation. When using the IRT Rasch model, the target information function makes it possible to choose tasks to produce a test that has the desired precision of measurement at all ability levels. The target mean and standard deviation of item difficulty (*b*-values) consistent with the information curves were also provided to test development staff to help with the test construction process. The polyserial correlation is a measure of how well the items discriminate among test takers that differ in their ability, and it is related to the overall reliability of the test.

For the spring 2009 CAPA test, spring 2008 operational test forms were used again. Therefore the target information given in Table 4.1, on the next page, was that used to build the spring 2008 operational test forms. These target values were developed using data collected in a fall 2007 test.

Table 4.1 Target Statistical Specifications for the CAPA

Subject	CAPA Level	Target Mean <i>b</i>	Target SD <i>b</i>	Mean AIS	Mean Polyserial
<i>English– Language Arts</i>	I	0.33	0.50	2.65	0.82
	II	–0.34	0.50	2.20	0.82
	III	0.01	0.50	2.20	0.82
	IV	–0.10	0.50	2.20	0.82
	V	0.08	0.50	2.20	0.82
<i>Mathematics</i>	I	–0.04	0.50	2.65	0.82
	II	0.16	0.50	2.20	0.82
	III	–0.31	0.50	2.20	0.82
	IV	–0.32	0.50	2.20	0.82
	V	–0.01	0.50	2.20	0.82
<i>Science</i>	I	0.14	0.50	2.65	0.82
	III	0.86	0.50	2.20	0.82
	IV	–0.64	0.50	2.20	0.82
	V	0.42	0.50	2.20	0.82

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 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. This procedure was used to develop the 2008 CAPA forms, which were readministered in 2009.

Chapter 5: Test Administration

Test Security and Confidentiality

All tests within the STAR Program are secure documents. Every person having access to test materials is required to maintain 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 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 (OTI)

The OTI is a division of ETS that provides quality assurance services and resides in the ETS Legal Department. The Quality Assurance division publishes and maintains *ETS Standards for Quality and Fairness*, which supports 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
- Investigate any security breach
- Report on security activities

OTI helps prevent misconduct on the part of test takers and administrators, detect potential misconduct through empirically established indicators, and resolve situations in a fair and balanced way that reflects the laws and professional standards governing the integrity of testing.

Test Development

During the test development process, ETS staff members 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 hardcopy 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 Review by ARPs

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 strictly prohibited in the meeting rooms.

Item Bank for Tasks

When the ARP review is complete, the 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 STAR electronic item bank. Subsequent updates to tasks are based on data from field testing and the operational use of the items. Only the latest version of the task is in the bank at any time, along with the administration data from every administration that has included the task.

Security of the electronic task 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 offsite, to prevent loss from a system breakdown or a natural disaster.
- The off-site 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 authorize 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 administrator at the CDE and at ETS authorizes the users.

Transfer of Forms and Tasks to the CDE

ETS shares a secure file transfer protocol (SFTP) site with the CDE. SFTP is a standard 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, and 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.

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 Sacramento, California.

All electronic applications included in the STAR Management System (CDE, 2009a) 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

After tasks and test forms are approved, the files are sent for printing on a CD using a secure courier system. According to established procedures, the OTI pre-approves all printing vendors before they can work on secured confidential and proprietary testing material. The printing vendor must submit a completed ETS Printing Plan and 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 receipt of materials to packaging. The reputable carriers used by Pearson provide a specialized handling and delivery service that maintains test security and meets the CAPA 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 test 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 requesting access to the test 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 and 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 the highest priority to Pearson. Pearson requires the following standard safeguards for security at their 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 then stored in secure warehouses. After they are stored, they will not be handled again unless questions arise about a student's score. For example, a school district or a parent may request that a student's test responses be rescored. In such a case, the answer document is removed from storage, copied, and sent securely to the ETS facility in Sacramento, California, for hand scoring, after which the copy is destroyed. 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 and test booklets are destroyed after October 31 of 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 their facilities. Entrances are monitored by security personnel and a computerized badge-reading system is utilized. Upon entering the facilities, 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 Scoring, Reporting and Technology (SR&T) 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 then keeps the files on secure servers and adheres to the ETS Code of Ethics to prevent any unauthorized access.

Reporting and Posting Results

After statistical analysis has been completed on student data, the files flow in three different directions. Paper reports, some with individual student results and others with summary results, are produced. Encrypted files of summary results are also sent to the CDE by means of SFTP. Any summary results that fewer than ten students are not reported. The statistics from the results are also 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 qualified for the National School Lunch Program (NSLP), and so forth (CDE, 2009b). ETS takes precautions to prevent any of this information from becoming public or being used for anything 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.

Test Results

ETS also has security measures for files and reports that show students' scores and performance levels. ETS is committed to safeguarding this information from unauthorized access, disclosure, modification, or destruction. ETS has strict information security policies in place to protect the confidentiality of ETS and client data. Access by ETS staff access to production databases is limited to personnel with a business need to access that 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. Comprehensive disaster recovery facilities are available and tested regularly at the SunGard installation in Philadelphia, Pennsylvania. ETS routinely sends backup data cartridges and files for critical

software, applications, and documentation to a secure off-site storage facility for safekeeping.

Access to the ETS Computer Processing Center is controlled through the use of 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 in use at 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
- Internet postings
- 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 are administered in conjunction with other tests that comprise the STAR Program. In that respect, 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, 2009c). This manual is described in the next section.

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

District STAR Coordinator

Each local education agency¹ (LEA) designates a district STAR coordinator who is responsible for ensuring the proper and consistent administration of the STAR tests. They 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

¹ Local education agencies include public school districts, statewide benefit charter schools, state board-authorized charter schools, county of education programs, and charter schools testing independently from their home district.

test site coordinators, receiving scorable and nonscorable materials from schools after an administration, and returning the materials to the STAR contractor for processing.

Test Examiner

The CAPA is administered 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, 2009d) 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 him or her 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 ensure the comparability of scores, the test site coordinator and principal of the school should objectively and randomly select ten 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 complete 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 so that all students have an equal opportunity to demonstrate their academic achievement (CDE, 2009d).

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, 2009c). However, the manual is not intended as a substitute for the *California Code of Regulations, 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 the test administration for school districts, to verify contact information of those school districts, and to update the school district's shipping information. (CDE, 2009e)

- **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, 2009f)
- **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, 2009b)
- **Extended Data Corrections**—This module allows school districts to correct the data that were submitted during Pre-ID up to seven days prior to the end of the school district’s selected testing window. (CDE, 2009g)

Accommodations for Students with Disabilities

All 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

All students participate in the STAR program, including students with disabilities and English learners. Most students with disabilities and English learners take the California Standards Tests under standard conditions. Some students with disabilities and English learners, however, 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, 2009c). Because examiners may adapt the CAPA in light of a student’s instructional mode, accommodations and modifications do not apply to the CAPA.

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 Examiners’ Manual* (CDE, 2009d), on page 21 of the nonsecure manual.

Scoring

CAPA tasks are scored using a 5-point 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 ten 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 Demographics 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 levels; 3) Obtain a corrected data CD-ROM for school district records; or 4) Match unmatched records (CDE, 2009h).

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 CDE have determined that an irregularity has occurred, 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*.

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* and the *STAR District and Test Site Coordinator Manual* (CDE, 2009c). 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 incident.

References

- California Department of Education (2009a). *2009 STAR Management System*. <http://www.startest.org/sms.html>.
- California Department of Education (2009b). *2009 STAR Pre-ID instructions manual*. Sacramento, CA. http://www.startest.org/pdfs/STAR.pre-id_manual.2009.pdf.
- California Department of Education (2009c). *2009 STAR district and test site coordinator manual*. Sacramento, CA. http://www.startest.org/pdfs/STAR.coord_man.2009.pdf.
- California Department of Education (2009d). *2009 California Alternate Performance Assessment (CAPA) examiner's manual*. Sacramento, CA. http://www.startest.org/pdfs/CAPA.examiners_manual.nonsecure.2009.pdf.
- California Department of Education (2009e). *2009 STAR Test Administration Setup manual*. Sacramento, CA.
- California Department of Education (2009f). *2009 STAR Order Management manual*. Sacramento, CA. http://www.startest.org/pdfs/STAR.order_mgmt.2009.pdf.
- California Department of Education (2009g). *2009 STAR Extended Data Corrections manual*. Sacramento, CA. http://www.startest.org/pdfs/STAR.xdc_manual.2009.pdf.

Chapter 6: Performance Standards

Background

From September 16 to 18, 2008, ETS conducted a standard-setting workshop in Sacramento, California, to recommend cut scores that delineated performance standards for the CAPA for ELA and mathematics levels I through V, and the CAPA for science levels I and III through V.¹ The performance standards were defined by the SBE as (1) far below basic, (2) below basic, (3) basic, (4) proficient, and (5) advanced. Performance standards are developed from a general description of the performance level (policy level descriptors) and competencies lists, which operationally define each level. Cut scores numerically define the performance levels. This chapter describes the process of developing performance standards which were applied to the CAPA operational tests in the spring of 2009.

ETS employed carefully designed standard-setting procedures to facilitate the development of performance standards for each CAPA test. 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. See the technical report for the standard-setting (ETS, 2008) for more 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 standard. The process generally requires that a panel of subject matter experts and others with relevant perspectives (e.g., teachers, school administrators) be assembled. For the CAPA, panelists were recruited to include California educators with experience administering the CAPA, who have direct experience in the education of students who take the CAPA, and who are familiar with the California content standards. Panelists were recruited to be representative of the educators of the state’s CAPA-eligible students (ETS, 2008). Panelists were assigned to one of four panels (Level I, CAPA 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 the steps listed below.

- Prior to attending the workshop, all panelists received a pre-assignment. The task was to review, on their own, the content standards upon which the CAPA tests are based and take notes on their own expectations for students at each performance level. This allows the panelists to understand how their perceptions may relate to the complexity of content standards.

¹ The CAPA for Science is not assessed at Level II.

- At the start of the workshop, panelists received training which includes 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 include the percentage of students assessed in a previous test administration of the test that would fall into each performance level, given the panelists’ judgments of cut scores.
- Panelists next became familiar with the tasks by reviewing the actual test and the rubrics, and then assessing and discussing the demands of the tasks.
- Panelists then reviewed a description of each performance level (that is, the competencies list) as a group, noting the increasing demands of each subsequent level. In this step, they began to visualize the knowledge and skills of students in each performance level and the differences between levels.
- 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.
- After completing training in the method, confirmed through an evaluation questionnaire, panelists made individual judgments; and discussed feedback related to other panelists’ judgments and feedback based on student performance data (impact data²). Panelists revised their judgments during the process if they wished. The final recommended cut scores were based on an average of panelists’ judgments at the end of three rounds. For the CAPA, the cut scores recommended by the panelists and the CDE superintendent recommendation were presented for public comment at regional public hearings. Comments and recommendations are then presented to the SBE for approval.

Development of Competencies Lists

Prior to the CAPA standard-setting workshop, ETS facilitated a meeting in which a subset of the standard setting panelists were assembled to develop a list of competencies based on the California content standards and California 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 who have experience working with students who take the CAPA, and panelists were assigned to one of four panels (Level I, and CAPA 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 2009 is called the Performance

² No impact data were presented to the Level I panel due to the change in the Level I rubric.

Profile Method in this report. The procedure was a modification to the Performance Profile Method used for the CAPA standard setting in 2003 (Morgan, 2003).³ 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.⁴

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.⁵ 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, 2008).

Results

The recommended cut scores obtained from the standard-setting 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 ranged 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 standards usually vary by grade and subject area. They are derived using procedures based on item response theory (IRT). The raw cut

³ Modifications were made to materials used such as the structure of the profiles and feedback. Panelists were asked to think holistically in both the 2003 and 2008 workshops.

⁴ Cut scores for below basic and far below basic performance levels were set statistically.

⁵ In creating score distributions for selection of profiles and projection of impact data, data files were based on sampling and selection criteria supplied by the CDE.

scores for a given test are mapped to IRT *thetas* ($\hat{\theta}_S$) using the test characteristic function⁶ and then transformed to the scale score metric using the following equation:

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

where,

$\theta_{\text{proficient}}$ 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 standard are presented in Table 6.1. The cut score for each performance standard is the lower bound of each scale score range. The scale score ranges do not change from year to year.

Table 6.1 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 – 22	23 – 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 – 19	20 – 29	30 – 34	35 – 38	39 – 60
	II	15 – 17	18 – 29	30 – 34	35 – 40	41 – 60
	III	15 – 16	17 – 29	30 – 34	35 – 39	40 – 60
	IV	15 – 17	18 – 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

Table 6.2 presents the percentages of examinees meeting each performance standard in 2009.

Table 6.2 Percentage of Examinees in Each Performance Level

Content Area	CAPA Level	Far Below Basic	Below Basic	Basic	Proficient	Advanced
English–Language Arts	I	5%	8%	12%	24%	51%
	II	1%	4%	17%	37%	41%
	III	1%	3%	13%	41%	42%
	IV	2%	7%	15%	40%	37%
	V	1%	3%	15%	38%	42%
Mathematics	I	9%	10%	19%	32%	29%
	II	3%	14%	22%	29%	33%
	III	1%	9%	26%	34%	31%
	IV	3%	15%	21%	29%	31%
	V	2%	11%	20%	33%	34%

⁶ An IRT test characteristic curve is the sum of item characteristic curves (ICC), where an ICC represents the probability of correctly responding to an item conditioned on examinee ability.

Content Area	CAPA Level	Far Below Basic	Below Basic	Basic	Proficient	Advanced
Science	I	10%	13%	19%	26%	33%
	III	1%	5%	26%	50%	19%
	IV	1%	7%	34%	43%	15%
	V	2%	8%	29%	44%	17%

The numbers in the summary table may not match exactly the results reported on the CDE Web site because of slight differences in the samples used to compute the statistics. The P2 data file used for the analyses in this chapter.

References

- Educational Testing Service K–12 Statistical Analysis Group. (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: Educational Testing Service.
- Educational Testing Service. (2008). *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: Educational Testing Service.
- 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.
- Morgan, D.L. (2003). *CAPA standard setting technical report*. Unpublished manuscript, Princeton, NJ: Educational Testing Service.
- 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. Such 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, attempts are made to ensure sufficient data are collected for the major subgroups of students. Such data help ETS to conduct 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 or observer using a rubric that depends on the test level being assessed. 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. After the student has responded to a task, the test examiner marks the corresponding circle on the student answer document.

Scoring Rubric

The scoring rubric represents the guideline for scoring the task. The rubric varies according to the CAPA level. 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 that 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 that is able to complete a task successfully at one level of prompting may take weeks or months to increase his or her proficiency on that task to be able to complete the task successfully at a less intrusive level of prompting. The differences within prompting levels are the reason 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 tasks. For all other levels, the rubric is specific to the task. Both rubrics are presented in Table 7.1.

The CAPA tests are administered by a special education teacher or case carrier who regularly works with the student being tested. In addition, all test administrators must have completed the CAPA test examiner training (CDE, 2009a).

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 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
- **Reporting Cluster Names and Task Numbers**—Identifies the reporting clusters for each test and the number of tasks in each cluster

The scoring specifications are reviewed and revised by the CDE, ETS, and Pearson each year. After a version that all parties agree to 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 district must return scorable and nonscorable materials within five working days after the last day for each test administration period.

Types of Scores and Subscores

Raw Score

For the CAPA for ELA and mathematics, there are five test levels and eight tasks per level. For the CAPA for science, there are four test levels and eight tasks per level. Performance scoring for Level I is based on a rubric with a range of 0–5 with 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. 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 on each CAPA exam are transformed to two-digit scale scores using the calibration process described in Chapter 2 on page 13. Scale scores range from 15 to 60 on each CAPA content-area test. The scale scores of examinees that have been tested in different years in a given grade and content area can be compared. However, the raw scores of these examinees cannot be meaningfully compared, because these scores are affected by the difficulty of the test taken as well as the ability of the examinee. Test difficulty will change to some degree from year to year.

Performance Levels

Students taking each CAPA content-area test are classified into one of the following performance levels:

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

For all CAPA exams, 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

ETS and Pearson take various necessary measures to ascertain that the student scores are computed accurately.

Scoring Key Verification Process

Scoring keys, provided in the form planners, are produced by ETS and verified thoroughly by performing various quality control checks. The form planners contain the information about an assembled test form including test name, administration year, maximum possible score for each task, and the standards and statistics associated with each task. Various checks are performed before keys are finalized, as listed below:

- The form planners are checked for accuracy against the Form Planner Specification document and the Score Key and Score Conversion document before the keys are loaded into the score key management system (SKM) at ETS.
- The sequence of tasks in the form planners are matched with their sequence in the actual test booklets.

- The demarcations of various sections in the actual test book are checked against the list of demarcations provided by the ETS test development staff.
- Scoring is verified internally at Pearson. ETS independently generates scores and verifies Pearson's scoring of the data by comparing the two results. Any discrepancies are then resolved.
- The entire scoring system is tested using a test deck that includes typical and extremely atypical responses vectors.
- Classical Task analyses are run on an early sample of data to provide an additional check of tasks. Although rare, if a task is found to be problematic, a followup process is carried out for it to be excluded from further analyses.

Score Verification Process

ETS psychometricians employ special procedures that adjust for task difficulty of one test form to another. As a result of this process, scoring tables are produced. Such tables map the current year's raw score to an appropriate scale score. Pearson utilizes these tables to generate scale scores for each student.

ETS verifies Pearson's scale scores by adhering to procedures such as the following:

- 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 school districts is based on the availability of data for all schools included in those districts, known as "complete districts"
- Reviewing longitudinal data for reasonableness; the results of the analyses are used to look at the trends and trends for the complete districts
- Reviewing longitudinal data for reasonableness using 99 percent of the entire testing population; the results are used to evaluate the trends for the state as well as few large school districts

The results of the longitudinal analyses are provided to the CDE and jointly discussed. Any anomalies in the results are investigated further and jointly discussed. Scores are released after explanations that satisfy both the CDE and ETS are obtained.

Overview of Score Aggregation Procedures

In order to provide meaningful results to the stakeholders, CAPA scores for a given level and 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 as well as group scores. The following section presents 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 exam.

Score Distributions and Summary Statistics

Summary statistics are presented in Table 7.2 to Table 7.4 that describe student performance on each CAPA exam. Included in these tables are the numbers 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 to the summary statistics of operational tasks for each test. Scale score frequency distributions for ELA, mathematics, and science, based on the spring 2009 administration of

the CAPA, are presented in Appendix 7.A. The percentages of students in each performance category are presented in Table 6.2 in Chapter 6.

The numbers in the summary tables may not match exactly the results reported on the CDE's Web site, as there may be slight differences in the samples used to compute the statistics. The P2 data file was used for the analyses in this chapter.

Table 7.2 Summary Statistics Describing Student Scores: ELA

Level	I	II	III	IV	V
Scale Score Information					
Number of examinees	12,531	6,587	6,614	9,853	10,517
Mean score	40.84	39.24	39.12	39.19	38.54
SD *	12.02	7.46	5.94	7.75	6.21
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	40	38	38	40	39
Reliability	0.91	0.84	0.86	0.88	0.89
SEM †	3.68	3.02	2.21	2.73	2.08
Raw Score Information					
Mean score	26.85	23.24	23.19	20.05	21.67
SD *	12.02	6.16	6.09	7.10	7.01
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	30	24	24	21	23
Reliability	0.91	0.84	0.86	0.88	0.89
SEM †	3.67	2.49	2.26	2.50	2.35
Task Information					
Number of tasks	8	8	8	8	8
Mean AIS ‡	3.37	2.91	2.91	2.51	2.73
SD AIS ‡	0.26	0.56	0.33	0.50	0.40
Min. AIS	2.81	2.36	2.30	1.65	1.95
Max. AIS	3.67	3.83	3.24	3.15	3.17
Possible range	0–5	0–4	0–4	0–4	0–4
Mean polyserial	0.81	0.75	0.75	0.78	0.79
SD polyserial	0.06	0.05	0.04	0.04	0.04
Min. polyserial	0.69	0.65	0.67	0.74	0.74
Max. polyserial	0.86	0.81	0.80	0.85	0.86
Mean Rasch difficulty	–0.74	–1.54	–1.52	–0.93	–1.19
SD Rasch difficulty	0.10	0.79	0.51	0.58	0.46
Min. Rasch difficulty	–0.82	–3.08	–2.15	–1.91	–1.73
Max. Rasch difficulty	–0.51	–0.87	–0.71	0.07	–0.33

* Standard Deviation | † Standard Error of Measurement | ‡ AIS = 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	12,484	6,569	6,602	9,831	10,485
Mean score	35.11	37.60	36.58	36.41	37.51
SD *	9.74	9.56	6.64	8.80	8.85
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	36	37	37	37	37
Reliability	0.87	0.88	0.87	0.88	0.87
SEM †	3.49	3.31	2.42	3.10	3.14
Raw Score Information					
Mean score	21.54	21.58	21.53	18.95	21.91
SD *	11.16	7.45	6.95	7.46	7.62
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	22	22	23	19	23
Reliability	0.87	0.88	0.87	0.88	0.87
SEM †	4.00	2.58	2.54	2.62	2.70
Task Information					
Number of tasks	8	8	8	8	8
Mean AIS ‡	2.70	2.70	2.70	2.37	2.76
SD AIS ‡	0.36	0.37	0.42	0.49	0.32
Min. AIS	2.12	2.05	2.18	1.78	2.23
Max. AIS	3.27	3.16	3.34	3.24	3.13
Possible range	0–5	0–4	0–4	0–4	0–4
Mean polyserial	0.79	0.78	0.76	0.79	0.78
SD polyserial	0.03	0.06	0.09	0.09	0.04
Min. polyserial	0.73	0.66	0.60	0.62	0.74
Max. polyserial	0.83	0.84	0.84	0.87	0.84
Mean Rasch difficulty	–0.29	–1.18	–1.29	–0.85	–1.21
SD Rasch difficulty	0.14	0.33	0.39	0.48	0.25
Min. Rasch difficulty	–0.52	–1.61	–1.87	–1.76	–1.45
Max. Rasch difficulty	–0.06	–0.64	–0.77	–0.24	–0.76

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

Table 7.4 Summary Statistics Describing Student Scores: Science

Level	I	III	IV	V
Scale Score Information				
Number of examinees	3,296	3,267	3,190	3,396
Mean score	35.59	36.24	35.56	35.35
SD *	11.25	5.45	5.53	5.34
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	0.91	0.85	0.85	0.87
SEM †	3.46	2.08	2.12	1.93
Raw Score Information				
Mean score	21.81	21.40	19.64	19.58
SD *	12.22	6.35	6.42	6.35
Possible range	0–40	0–32	0–32	0–32
Obtained range	0–40	0–32	0–32	0–32
Median	22	22	20	20
Reliability	0.91	0.85	0.85	0.87
SEM †	3.76	2.43	2.46	2.30
Task Information				
Number of tasks	8	8	8	8
Mean AIS ‡	2.75	2.71	2.47	2.47
SD AIS ‡	0.30	0.22	0.17	0.29
Min. AIS	2.39	2.35	2.25	2.06
Max. AIS	3.37	2.96	2.79	2.86
Possible range	0–5	0–4	0–4	0–4
Mean polyserial	0.82	0.75	0.75	0.78
SD polyserial	0.03	0.05	0.03	0.04
Min. polyserial	0.77	0.66	0.68	0.73
Max. polyserial	0.85	0.81	0.79	0.83
Mean Rasch difficulty	–0.23	–1.29	–0.95	–0.54
SD Rasch difficulty	0.16	0.29	0.17	0.30
Min. Rasch difficulty	–0.53	–1.71	–1.22	–0.94
Max. Rasch difficulty	–0.00	–0.96	–0.71	–0.07

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

Table 7.A.1 thru Table 7.A.14 in Appendix 7.A, starting on page 56, show the distributions of scale scores for each CAPA exam. The results are reported in terms of 16 score intervals.

Group Scores

Statistics summarizing student performance by content area and levels 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 proficiency, economic status, and primary disability. The tables show the numbers of valid cases in each group as well as scale score means and standard deviations for each demographic group. Table 7.5 defines 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).

Note that the statistics in these tables slightly differ from the statewide statistics reported on the CDE Web site because students with invalid scores were excluded from the tables.

Table 7.5 Subgroup Definitions

Subgroup	Definition
Gender	Male Female
Ethnicity	American Indian or Alaska Native Asian <ul style="list-style-type: none"> – Chinese – Japanese – Korean – Vietnamese – Asian Indian – Laotian – Cambodian – Other Asian Pacific Islander <ul style="list-style-type: none"> – Native Hawaiian – Guamanian – Samoan – Tahitian – Other Pacific Islander Filipino Hispanic or Latino African American White (not Hispanic)
English Language Fluency	English-language fluency Initially fluent English proficient English learner Reclassified fluent English proficient
Economic Status	Not economically disadvantaged Economically disadvantaged
Primary Disability	Mental Retardation 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 to be 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.

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 below.

Table 7.6 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 reports to the child's parents or guardians 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 results for each CAPA exam taken by the student. Scale scores are reported on a scale ranging from 15 to 60. Results for the CAPA also are reported by performance levels, which 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. Beginning in 2008, an additional score report, Ethnicity for Economic Status, is produced for the CAPA. This Subgroup Summary report aggregates and reports results by economic status within selected ethnic groups.

Score Report Applications

CAPA results provide parents and guardians with information about their children'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 to talk with teachers about ways to improve their children's achievement of the California content standards.

Schools may use the CAPA results to help make decisions about how to best 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 grade 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 progress toward achieving established goals. As mentioned previously, CAPA results are used to calculate each school'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 (e.g., 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 Elementary and Secondary Education Act (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 Adequate Yearly Progress (AYP) report. Each year, California schools 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, 2009b). 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 was retested.

Criteria for Interpreting Score Reports

The information presented on 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 levels. This is because the score scales are different for each content area and level. Comparisons between raw scores should be limited to comparisons within not only content area and level but also test year. Comparing scores obtained in different levels or content areas should be avoided, because the results are not on the same scale. The user may compare scores for the same content area and levels, within a school, between schools, or between a school and its district, its county, or the state. Since new score scales and cut scores were applied to the 2009 CAPA test results, results from 2009 cannot meaningfully be compared to results obtained in previous years. For more details on the criteria for interpreting information provided on the score reports, see the *2009 STAR Post-Test Guide* (CDE, 2009c).

References

- California Department of Education. (2009a). *2009 CAPA examiner's manual*. http://www.startest.org/pdfs/CAPA.examiners_manual.nonsecure.2009.pdf.
- California Department of Education. (2009b). *2009 STAR CST/CMA, CAPA, and STS printed reports*. <http://www.startest.org/pdfs/STAR.reports.2009.pdf>.
- California Department of Education. (2009c). *2009 STAR post-test guide*. http://www.startest.org/pdfs/STAR.post-test_guide.2009.pdf.
- Educational Testing Service. (2002). *ETS standards for quality and fairness*. Office of Testing Integrity. Princeton, NJ: Educational Testing Service, Princeton.
- 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.

Appendix 7.A—Scale Score Distribution Tables

Table 7.A.1 Scale Score Frequency Distributions: Level I, ELA

Scale Score	English—Language Arts			
	Frequency	Percent	Cumulative Frequency	Percent Below
60	2,230	17.80	2,230	82.20
57–59	–	–	–	–
54–56	–	–	–	–
51–53	624	4.98	2,854	77.22
48–50	388	3.10	3,242	74.13
45–47	299	2.39	3,541	71.74
42–44	1,708	13.63	5,249	58.11
39–41	1,784	14.24	7,033	43.88
36–38	1,567	12.50	8,600	31.37
33–35	1,559	12.44	10,159	18.93
30–32	694	5.54	10,853	13.39
27–29	545	4.35	11,398	9.04
24–26	140	1.12	11,538	7.92
21–23	128	1.02	11,666	6.90
18–20	128	1.02	11,794	5.88
15–17	737	5.88	12,531	0.00

Note: Scores not obtainable in 2009 are show as dashes (–).

Table 7.A.2 Scale Score Frequency Distributions: Level I, Mathematics

Scale Score	Mathematics			
	Frequency	Percent	Cumulative Frequency	Percent Below
60	603	4.83	603	95.17
57–59	–	–	–	–
54–56	–	–	–	–
51–53	–	–	–	–
48–50	237	1.90	840	93.27
45–47	382	3.06	1,222	90.21
42–44	934	7.48	2,156	82.73
39–41	1,465	11.74	3,621	70.99
36–38	2,775	22.23	6,396	48.77
33–35	2,628	21.05	9,024	27.72
30–32	1,053	8.43	10,077	19.28
27–29	407	3.26	10,484	16.02
24–26	492	3.94	10,976	12.08
21–23	174	1.39	11,150	10.69
18–20	177	1.42	11,327	9.27
15–17	1,157	9.27	12,484	0.00

Note: Scores not obtainable in 2009 are show as dashes (–).

Table 7.A.3 Scale Score Frequency Distributions: Level I, Science

Scale Score	Science			
	Frequency	Percent	Cumulative Frequency	Percent Below
60	280	8.50	280	91.50
57–59	–	–	–	–
54–56	–	–	–	–
51–53	–	–	–	–
48–50	81	2.46	361	89.05
45–47	69	2.09	430	86.95
42–44	267	8.10	697	78.85
39–41	394	11.95	1,091	66.90
36–38	588	17.84	1,679	49.06
33–35	611	18.54	2,290	30.52
30–32	271	8.22	2,561	22.30
27–29	108	3.28	2,669	19.02
24–26	207	6.28	2,876	12.74
21–23	–	–	–	–
18–20	49	1.49	2,925	11.26
15–17	371	11.26	3,296	0.00

Note: Scores not obtainable in 2009 are show as dashes (–).

Table 7.A.4 Scale Score Frequency Distributions: Level II, ELA

Scale Score	English–Language Arts			
	Frequency	Percent	Cumulative Frequency	Percent Below
60	405	6.15	405	93.85
57–59	–	–	–	–
54–56	–	–	–	–
51–53	–	–	–	–
48–50	375	5.69	780	88.16
45–47	375	5.69	1,155	82.47
42–44	795	12.07	1,950	70.40
39–41	1,090	16.55	3,040	53.85
36–38	1,776	26.96	4,816	26.89
33–35	1,081	16.41	5,897	10.48
30–32	362	5.50	6,259	4.98
27–29	154	2.34	6,413	2.64
24–26	89	1.35	6,502	1.29
21–23	28	0.43	6,530	0.87
18–20	12	0.18	6,542	0.68
15–17	45	0.68	6,587	0.00

Note: Scores not obtainable in 2009 are show as dashes (–).

Table 7.A.5 Scale Score Frequency Distributions: Level II, Mathematics

Scale Score	Mathematics			
	Frequency	Percent	Cumulative Frequency	Percent Below
60	417	6.35	417	93.65
57–59	–	–	–	–
54–56	–	–	–	–
51–53	386	5.88	803	87.78
48–50	–	–	–	–
45–47	338	5.15	1,141	82.63
42–44	682	10.38	1,823	72.25
39–41	886	13.49	2,709	58.76
36–38	1,049	15.97	3,758	42.79
33–35	1,053	16.03	4,811	26.76
30–32	658	10.02	5,469	16.75
27–29	547	8.33	6,016	8.42
24–26	137	2.09	6,153	6.33
21–23	209	3.18	6,362	3.15
18–20	34	0.52	6,396	2.63
15–17	173	2.63	6,569	0.00

Note: Scores not obtainable in 2009 are show as dashes (–).

Table 7.A.6 Scale Score Frequency Distributions: Level III, ELA

Scale Score	English–Language Arts			
	Frequency	Percent	Cumulative Frequency	Percent Below
60	199	3.01	199	96.99
57–59	–	–	–	–
54–56	–	–	–	–
51–53	–	–	–	–
48–50	304	4.60	503	92.39
45–47	426	6.44	929	85.95
42–44	934	14.12	1,863	71.83
39–41	1,341	20.28	3,204	51.56
36–38	2,044	30.90	5,248	20.65
33–35	891	13.47	6,139	7.18
30–32	258	3.90	6,397	3.28
27–29	111	1.68	6,508	1.60
24–26	45	0.68	6,553	0.92
21–23	34	0.51	6,587	0.41
18–20	5	0.08	6,592	0.33
15–17	22	0.33	6,614	0.00

Note: Scores not obtainable in 2009 are show as dashes (–).

Table 7.A.7 Scale Score Frequency Distributions: Level III, Mathematics

Scale Score	Mathematics			
	Frequency	Percent	Cumulative Frequency	Percent Below
60	134	2.03	134	97.97
57–59	–	–	–	–
54–56	–	–	–	–
51–53	–	–	–	–
48–50	230	3.48	364	94.49
45–47	–	–	–	–
42–44	762	11.54	1,126	82.94
39–41	1,274	19.30	2,400	63.65
36–38	1,579	23.92	3,979	39.73
33–35	1,105	16.74	5,084	22.99
30–32	837	12.68	5,921	10.32
27–29	320	4.85	6,241	5.47
24–26	200	3.03	6,441	2.44
21–23	39	0.59	6,480	1.85
18–20	33	0.50	6,513	1.35
15–17	89	1.35	6,602	0.00

Note: Scores not obtainable in 2009 are show as dashes (–).

Table 7.A.8 Scale Score Frequency Distributions: Level III, Science

Scale Score	Science			
	Frequency	Percent	Cumulative Frequency	Percent Below
	69	2.11	69	97.89
	–	–	–	–
	–	–	–	–
	–	–	–	–
	–	–	–	–
	105	3.21	174	94.67
	122	3.73	296	90.94
	493	15.09	789	75.85
	934	28.59	1,723	47.26
	1,093	33.46	2,816	13.80
	268	8.20	3,084	5.60
	104	3.18	3,188	2.42
	29	0.89	3,217	1.53
	20	0.61	3,237	0.92
	10	0.31	3,247	0.61
	20	0.61	3,267	0.00

Note: Scores not obtainable in 2009 are show as dashes (–).

Table 7.A.9 Scale Score Frequency Distributions: Level IV, ELA

Scale Score	English–Language Arts			
	Frequency	Percent	Cumulative Frequency	Percent Below
60	219	2.22	219	97.78
57–59	–	–	–	–
54–56	239	2.43	458	95.35
51–53	–	–	–	–
48–50	653	6.63	1,111	88.72
45–47	967	9.81	2,078	78.91
42–44	1,534	15.57	3,612	63.34
39–41	1,911	19.40	5,523	43.95
36–38	1,669	16.94	7,192	27.01
33–35	1,008	10.23	8,200	16.78
30–32	822	8.34	9,022	8.43
27–29	398	4.04	9,420	4.39
24–26	83	0.84	9,503	3.55
21–23	70	0.71	9,573	2.84
18–20	125	1.27	9,698	1.57
15–17	155	1.57	9,853	0.00

Note: Scores not obtainable in 2009 are show as dashes (–).

Table 7.A.10 Scale Score Frequency Distributions: Level IV, Mathematics

Scale Score	Mathematics			
	Frequency	Percent	Cumulative Frequency	Percent Below
60	269	2.74	269	97.26
57–59	–	–	–	–
54–56	–	–	–	–
51–53	391	3.98	660	93.29
48–50	295	3.00	955	90.29
45–47	687	6.99	1,642	83.30
42–44	689	7.01	2,331	76.29
39–41	1,436	14.61	3,767	61.68
36–38	1,687	17.16	5,454	44.52
33–35	1,229	12.50	6,683	32.02
30–32	1,319	13.42	8,002	18.60
27–29	888	9.03	8,890	9.57
24–26	286	2.91	9,176	6.66
21–23	257	2.61	9,433	4.05
18–20	75	0.76	9,508	3.29
15–17	323	3.29	9,831	0.00

Note: Scores not obtainable in 2009 are show as dashes (–).

Table 7.A.11 Scale Score Frequency Distributions: Level IV, Science

Scale Score	Science			
	Frequency	Percent	Cumulative Frequency	Percent Below
60	46	1.44	46	98.56
57–59	–	–	–	–
54–56	–	–	–	–
51–53	–	–	–	–
48–50	–	–	–	–
45–47	44	1.38	90	97.18
42–44	157	4.92	247	92.26
39–41	393	12.32	640	79.94
36–38	1,010	31.66	1,650	48.28
33–35	864	27.08	2,514	21.19
30–32	420	13.17	2,934	8.03
27–29	155	4.86	3,089	3.17
24–26	36	1.13	3,125	2.04
21–23	10	0.31	3,135	1.72
18–20	19	0.60	3,154	1.13
15–17	36	1.13	3,190	0.00

Note: Scores not obtainable in 2009 are show as dashes (–).

Table 7.A.12 Scale Score Frequency Distributions: Level V, ELA

Scale Score	English–Language Arts			
	Frequency	Percent	Cumulative Frequency	Percent Below
60	274	2.61	274	97.39
57–59	–	–	–	–
54–56	–	–	–	–
51–53	–	–	–	–
48–50	400	3.80	674	93.59
45–47	517	4.92	1,191	88.68
42–44	1,277	12.14	2,468	76.53
39–41	3,097	29.45	5,565	47.09
36–38	2,179	20.72	7,744	26.37
33–35	1,698	16.15	9,442	10.22
30–32	572	5.44	10,014	4.78
27–29	211	2.01	10,225	2.78
24–26	113	1.07	10,338	1.70
21–23	59	0.56	10,397	1.14
18–20	33	0.31	10,430	0.83
15–17	87	0.83	10,517	0.00

Note: Scores not obtainable in 2009 are show as dashes (–).

Table 7.A.13 Scale Score Frequency Distributions: Level V, Mathematics

Scale Score	Mathematics			
	Frequency	Percent	Cumulative Frequency	Percent Below
60	767	7.32	767	92.68
57–59	–	–	–	–
54–56	–	–	–	–
51–53	–	–	–	–
48–50	–	–	–	–
45–47	499	4.76	1,266	87.93
42–44	1,104	10.53	2,370	77.40
39–41	1,804	17.21	4,174	60.19
36–38	2,475	23.61	6,649	36.59
33–35	1,524	14.54	8,173	22.05
30–32	918	8.76	9,091	13.30
27–29	473	4.51	9,564	8.78
24–26	278	2.65	9,842	6.13
21–23	321	3.06	10,163	3.07
18–20	61	0.58	10,224	2.49
15–17	261	2.49	10,485	0.00

Note: Scores not obtainable in 2009 are show as dashes (–).

Table 7.A.14 Scale Score Frequency Distributions: Level V, Science

Scale Score	Science			
	Frequency	Percent	Cumulative Frequency	Percent Below
60	33	0.97	33	99.03
57–59	–	–	–	–
54–56	–	–	–	–
51–53	–	–	–	–
48–50	–	–	–	–
45–47	46	1.35	79	97.67
42–44	129	3.80	208	93.88
39–41	373	10.98	581	82.89
36–38	1,288	37.93	1,869	44.96
33–35	874	25.74	2,743	19.23
30–32	332	9.78	3,075	9.45
27–29	196	5.77	3,271	3.68
24–26	36	1.06	3,307	2.62
21–23	25	0.74	3,332	1.88
18–20	14	0.41	3,346	1.47
15–17	50	1.47	3,396	0.00

Note: Scores not obtainable in 2009 are show as dashes (–).

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	46,102	2%	5%	14%	35%	43%
Male	29,672	2%	5%	14%	35%	44%
Female	16,264	2%	5%	14%	36%	42%
Gender unknown	166	0%	6%	13%	35%	46%
American Indian	395	2%	5%	12%	32%	49%
Asian American	3,286	3%	8%	17%	36%	37%
Pacific Islander	245	2%	4%	17%	37%	40%
Filipino	1,419	2%	7%	17%	38%	36%
Hispanic	22,769	2%	5%	14%	35%	43%
African American	4,987	3%	5%	12%	35%	45%
White	12,331	2%	5%	14%	35%	44%
Ethnicity unknown	670	2%	6%	16%	31%	45%
English Only	28,420	2%	6%	14%	35%	43%
Initially—Fluent English Proficient	968	4%	6%	19%	35%	36%
English Learner	14,888	2%	5%	14%	35%	44%
Reclassified—Fluent English Proficient	1,045	2%	4%	15%	35%	44%
English Proficient unknown	781	1%	4%	10%	35%	50%
Mental Retardation	19,395	1%	5%	16%	38%	40%
Hard of Hearing	318	1%	6%	13%	38%	43%
Deafness	434	0%	2%	10%	46%	41%
Speech/Language Impairment	1,695	0%	1%	4%	28%	67%
Visual Impairment	536	6%	8%	17%	33%	36%
Emotional Disturbance	380	1%	2%	3%	24%	70%
Orthopedic Impairment	4,335	6%	10%	16%	34%	34%
Other Health Impairment	1,899	1%	3%	8%	34%	54%
Specific Learning Impairment	2,960	0%	0%	2%	23%	74%
Deaf Blindness	43	21%	16%	14%	16%	33%
Multiple Group	2,151	8%	10%	19%	32%	31%
Autism	10,732	2%	7%	16%	35%	40%
Traumatic Brain Injury	311	4%	6%	10%	32%	48%
Unknown	913	2%	5%	11%	35%	47%
Not Econ. Disadvantaged	16,460	3%	7%	15%	35%	40%
Economically Disadvantaged	28,293	2%	5%	14%	35%	45%
Unknown Economic Status	1,349	2%	4%	12%	33%	49%
Primary Ethnicity—Not Economically Disadvantaged						
American Indian	139	1%	9%	17%	30%	43%
Asian American	1,706	3%	9%	17%	36%	36%
Pacific Islander	100	2%	5%	19%	39%	35%
Filipino	925	2%	7%	18%	38%	35%
Hispanic	4,405	4%	7%	15%	33%	40%
African American	1,469	4%	7%	14%	35%	41%
White	7,483	2%	6%	15%	36%	41%
Ethnicity unknown	233	3%	9%	19%	31%	38%

	Number Tested	Percentage in Performance Level				
		Far Below Basic	Below Basic	Basic	Proficient	Advanced
Primary Ethnicity—Economically Disadvantaged						
American Indian	245	2%	2%	11%	33%	52%
Asian American	1,499	3%	7%	17%	36%	38%
Pacific Islander	138	2%	4%	15%	36%	43%
Filipino	457	3%	7%	15%	37%	38%
Hispanic	17,919	2%	5%	14%	35%	44%
African American	3,367	2%	4%	11%	36%	47%
White	4,454	1%	4%	12%	35%	48%
Ethnicity unknown	214	2%	6%	15%	30%	47%
Primary Ethnicity—Unknown Economic Status						
American Indian	11	0%	0%	0%	36%	64%
Asian American	81	1%	5%	16%	44%	33%
Pacific Islander	-	-	-	-	-	-
Filipino	37	0%	8%	27%	35%	30%
Hispanic	445	2%	3%	11%	35%	49%
African American	151	2%	3%	9%	32%	54%
White	394	2%	5%	11%	31%	52%
Ethnicity unknown	223	2%	4%	13%	30%	51%

* Results for groups with fewer than 11 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	45,971	4%	12%	21%	32%	31%
Male	29,595	4%	11%	20%	31%	33%
Female	16,209	5%	13%	22%	33%	28%
Gender unknown	167	1%	15%	21%	31%	32%
American Indian	395	2%	11%	20%	32%	35%
Asian American	3,272	5%	13%	23%	32%	27%
Pacific Islander	243	2%	13%	22%	38%	24%
Filipino	1,415	5%	13%	23%	32%	27%
Hispanic	22,713	4%	12%	21%	31%	32%
African American	4,971	5%	11%	21%	31%	32%
White	12,298	4%	12%	22%	32%	30%
Ethnicity unknown	664	4%	13%	20%	30%	33%
English Only	28,328	4%	13%	21%	31%	30%
Initially—Fluent English Proficient	961	6%	13%	24%	31%	26%
English Learner	14,861	4%	11%	21%	32%	33%
Reclassified—Fluent English Proficient	1,043	3%	11%	21%	32%	34%
English Proficient unknown	778	3%	9%	20%	32%	37%
Mental Retardation	19,357	3%	14%	24%	32%	27%
Hard of Hearing	317	3%	9%	18%	29%	41%
Deafness	433	0%	3%	10%	35%	51%
Speech/Language Impairment	1,695	0%	3%	9%	32%	57%
Visual Impairment	532	10%	15%	27%	27%	21%
Emotional Disturbance	378	3%	2%	9%	26%	60%
Orthopedic Impairment	4,311	11%	16%	24%	29%	20%
Other Health Impairment	1,891	2%	8%	18%	32%	40%
Specific Learning Impairment	2,955	0%	1%	5%	27%	67%
Deaf Blindness	43	26%	21%	26%	16%	12%
Multiple Group	2,144	13%	16%	25%	29%	17%
Autism	10,696	4%	12%	21%	33%	29%
Traumatic Brain Injury	310	8%	9%	16%	33%	34%
Unknown	909	4%	10%	18%	31%	37%
Not Econ. Disadvantaged	16,411	5%	13%	22%	32%	27%
Economically Disadvantaged	28,216	4%	11%	20%	32%	33%
Unknown Economic Status	1,344	3%	9%	19%	31%	38%
Primary Ethnicity—Not Economically Disadvantaged						
American Indian	139	1%	14%	24%	32%	29%
Asian American	1,702	6%	13%	23%	32%	26%
Pacific Islander	100	3%	20%	20%	39%	18%
Filipino	922	5%	13%	25%	31%	26%
Hispanic	4,390	7%	13%	21%	31%	28%
African American	1,460	7%	13%	21%	30%	28%
White	7,466	4%	13%	23%	32%	27%
Ethnicity unknown	232	5%	18%	23%	29%	25%

	Number Tested	Percentage in Performance Level				
		Far Below Basic	Below Basic	Basic	Proficient	Advanced
Primary Ethnicity—Economically Disadvantaged						
American Indian	245	3%	9%	18%	33%	37%
Asian American	1,490	5%	12%	24%	32%	28%
Pacific Islander	137	2%	8%	25%	36%	29%
Filipino	456	6%	12%	21%	34%	27%
Hispanic	17,879	4%	12%	21%	31%	33%
African American	3,360	4%	10%	20%	32%	34%
White	4,438	3%	10%	19%	33%	35%
Ethnicity unknown	211	3%	12%	18%	33%	33%
Primary Ethnicity—Unknown Economic Status						
American Indian	11	0%	9%	9%	18%	64%
Asian American	80	1%	16%	28%	25%	30%
Pacific Islander	-	-	-	-	-	-
Filipino	37	3%	19%	16%	32%	30%
Hispanic	444	3%	8%	17%	34%	38%
African American	151	3%	7%	25%	28%	38%
White	394	5%	9%	18%	31%	37%
Ethnicity unknown	221	3%	9%	17%	29%	42%

* Results for groups with fewer than 11 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	13,149	3%	8%	27%	41%	21%
Male	8,441	3%	8%	26%	40%	23%
Female	4,672	4%	8%	29%	42%	18%
Gender Unknown	36	3%	17%	28%	28%	25%
American Indian	133	0%	7%	23%	44%	26%
Asian American	943	5%	11%	30%	39%	15%
Pacific Islander	67	4%	7%	30%	36%	22%
Filipino	388	4%	11%	30%	38%	17%
Hispanic	6,280	3%	7%	27%	41%	22%
African American	1,482	3%	8%	27%	41%	21%
White	3,683	3%	8%	26%	41%	22%
Ethnicity unknown	173	4%	9%	29%	40%	18%
English Only	8,218	4%	8%	27%	40%	21%
Initially—Fluent English Proficient	283	4%	13%	28%	37%	18%
English Learner	4,098	3%	7%	27%	41%	21%
Reclassified—Fluent English Proficient	369	2%	5%	29%	43%	21%
English Proficient unknown	181	3%	8%	25%	44%	20%
Mental Retardation	5,934	2%	7%	31%	41%	19%
Hard of Hearing	97	2%	9%	22%	46%	21%
Deafness	137	1%	1%	16%	51%	31%
Speech/Language Impairment	324	0%	1%	12%	52%	35%
Visual Impairment	148	9%	14%	22%	34%	22%
Emotional Disturbance	133	2%	2%	14%	40%	43%
Orthopedic Impairment	1,275	10%	14%	26%	33%	17%
Other Health Impairment	484	2%	4%	23%	46%	26%
Specific Learning Impairment	847	0%	0%	11%	47%	42%
Deaf Blindness	-	-	-	-	-	-
Multiple Group	630	12%	13%	27%	36%	13%
Autism	2,844	3%	10%	29%	39%	19%
Traumatic Brain Injury	97	7%	6%	20%	44%	23%
Unknown	195	3%	10%	23%	42%	23%
Not Econ. Disadvantaged	4,820	5%	10%	28%	39%	19%
Economically Disadvantaged	8,011	3%	7%	27%	41%	22%
Unknown Economic Status	318	3%	7%	22%	48%	20%
Primary Ethnicity—Not Economically Disadvantaged						
American Indian	41	0%	5%	24%	41%	29%
Asian American	498	5%	13%	29%	40%	13%
Pacific Islander	21	5%	14%	33%	38%	10%
Filipino	252	5%	12%	32%	36%	16%
Hispanic	1,231	8%	10%	26%	37%	20%
African American	443	6%	10%	29%	37%	18%
White	2,274	3%	9%	28%	40%	20%
Ethnicity unknown	60	8%	10%	35%	35%	12%

Primary Ethnicity—Economically Disadvantaged						
American Indian	87	0%	8%	23%	44%	25%
Asian American	431	5%	10%	31%	37%	17%
Pacific Islander	46	4%	4%	28%	35%	28%
Filipino	131	2%	10%	24%	43%	21%
Hispanic	4,946	2%	7%	27%	41%	22%
African American	994	2%	7%	26%	43%	23%
White	1,318	3%	6%	24%	43%	24%
Ethnicity unknown	58	2%	10%	26%	41%	21%
Primary Ethnicity—Unknown Economic Status						
American Indian	-	-	-	-	-	-
Asian American	14	0%	7%	50%	43%	0%
Pacific Islander	-	-	-	-	-	-
Filipino	-	-	-	-	-	-
Hispanic	103	3%	4%	23%	52%	17%
African American	45	2%	7%	20%	56%	16%
White	91	4%	10%	15%	42%	29%
Ethnicity unknown	55	2%	5%	25%	44%	24%

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

Appendix 7.C – Type of Score Report

Table 7.C.1 Score Reports Reflecting CAPA Results

2009 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. Data presented include:</p> <ul style="list-style-type: none"> • Scale scores • Performance levels* 	<p>Because this report includes individual student results, it is not distributed beyond the student's school.</p> <p>Two colored copies of this report are provided for each student. One is for the student's current teacher, and one is to be distributed to parents/guardians by the school district.</p>
Student Record Label	
<p>These reports are printed on adhesive labels to be affixed to the student's permanent school records. Each pupil shall have an individual record of accomplishment that includes STAR testing results (see <i>California Education Code</i> Section 60607 [a]). Significant information includes:</p> <ul style="list-style-type: none"> • Scale scores and performance levels 	<p>Because this report includes individual student results, it is not distributed beyond the student's school.</p>
Student Master List	
<p>This report is an alphabetical roster of individual student results. It mainly includes:</p> <ul style="list-style-type: none"> • A scale score and a performance level for each content area tested 	<p>This report provides administrators and teachers with a quick reference to all students' results within each level or within each level and year-round schedule at a school.</p> <p>Because this report includes individual student results, it should not be distributed beyond the student's school.</p>
Student Master List Summary	
<p>This report summarizes student results at the school, district, county, and state levels for each content area and CAPA level. It does <i>not</i> include any individual student information. For each CAPA level, the following data are summarized:</p> <ul style="list-style-type: none"> • By content area tested: <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 percent correct 	<p>This report is a resource for evaluators, researchers, teachers, parents/guardians, community members, and administrators.</p> <p>One copy is sent to the school and one to the district.</p> <p>This report is also produced for 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.</p>

2009 STAR CAPA Student Reports	
Description	Distribution
<p>For each content area tested for CAPA:</p> <ul style="list-style-type: none"> • Mean scale score • Standard deviation of scale scores • Number and percent of students scoring at each performance level 	
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 is included:</p> <ul style="list-style-type: none"> • Total number tested in the subgroup • Percent tested in subgroup as a percent of all students tested • Number and percent of valid scores • Number tested who received scores • Mean scale scores • Standard deviation of scale scores • Number and percent of students scoring at each CAPA performance level 	<p>This report is a resource for evaluators, researchers, teachers, parents/guardians, community members, and administrators.</p> <p>One copy is sent to the school and one to the district. This report is also produced for 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.</p>

2009 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 both parents have not received a high school diploma OR the student participates in the free or reduced-price lunch program also known as the National School Lunch Program (NSLP). As with the standard Subgroup Summary, this disaggregation contains no individual student-identifying information and is aggregated at the school, 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 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 scores • 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 sent to the school and one copy to the district. This report is also produced for 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.</p>

*Performance levels are advanced, proficient, basic, below basic, and far below basic.

Chapter 8: Analyses

This chapter summarizes the item- and test-level statistics obtained for the California standards-testing program administered during the spring of 2009.

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 presents the AIS and polyserial correlation coefficient for each task in each operational test and each field-test task. Also presented in this appendix is information about the distribution of scores for the operational tasks and the field-test tasks. In addition, the mean, minimum and maximum of average item scores (AIS) and polyserial correlations for the polytomously scored operational tasks are presented in Table 8.2 on page 74.
2. Appendix 8.B presents results of the reliability analyses of total test scores and subscores for the population as a whole and for selected subgroups. Also presented are results of the analyses of the inter-rater reliability results for all tests. In this appendix, the results of the analyses of the consistency and accuracy of the performance classifications are presented.
3. Appendix 8.C presents the scoring tables that show the correlations between scores obtained in the different content areas measured by the CAPA. They are provided as an 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 results for various subgroups are summarized in Appendix 8.C.
4. Appendix 8.D presents the distribution of items based on their fit to the Rasch model. This appendix also includes summaries of Rasch item difficulty statistics (b-values) for the operational and field-test items. In this appendix, raw-to-scale score conversion tables also are listed.
5. Appendix 8.E presents the results of the DIF analyses applied to all operational and field-test items for which sufficient student samples were available. In this appendix, items 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 in the testing process and involved varying proportions of the full CAPA data. All CAPA analyses for this technical report were conducted using the P2 data file except for the item calibration. For the calibration and scaling for reporting, all valid cases available by early June were used.

Table 8.1 CAPA 2009 Raw Score Means and Standard Deviations: Total P2 Population and Calibration Sample

Group	Level	P2			Calibration Sample			
		N	Mean	SD	N	% of P2	Mean	SD
English– Language Arts	I	12,531	26.85	12.02	4,681	37%	26.67	12.03
	II	6,587	23.24	6.16	2,425	37%	23.58	6.11
	III	6,614	23.19	6.09	2,436	37%	23.50	6.12
	IV	9,853	20.05	7.10	3,547	36%	20.27	7.11
	V	10,517	21.67	7.01	4,113	39%	21.59	7.06
Mathematics	I	12,484	21.54	11.16	4,661	37%	21.29	11.07
	II	6,569	21.58	7.45	2,420	37%	21.96	7.52
	III	6,602	21.53	6.95	2,431	37%	21.87	6.98
	IV	9,831	18.95	7.46	3,542	36%	19.20	7.45
	V	10,485	21.91	7.62	4,098	39%	21.81	7.65
Science	I	3,296	21.81	12.22	1,231	37%	21.15	12.19
	III	3,267	21.40	6.35	1,180	36%	21.89	6.49
	IV	3,190	19.64	6.42	1,146	36%	19.77	6.28
	V	3,396	19.58	6.35	1,310	39%	19.64	6.43

Classical Analyses

Average Item Score (AIS)

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

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

where,

σ_{tot} is the standard deviation of the score; and

β is the item parameter to be estimated from the data 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. An item 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.

Table 8.A.1 through Table 8.A.14 present, for each item in the 2009 administration, the AIS and polyserial correlation. Some items were flagged for unusual statistics; these flags are shown in the tables. There are three types of flags. Although the flag definition appears in the heading at each table, the flags are displayed in the body of the tables only where applicable for the specific CAPA test 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

Subject	Level	No. of items	No. of Examinees	Mean		Minimum		Maximum	
				AIS	Polyserial	AIS	Polyserial	AIS	Polyserial
<i>ELA</i>	I	8	12,531	3.37	0.81	2.81	0.69	3.67	0.86
	II	8	6,587	2.91	0.75	2.36	0.65	3.83	0.81
	III	8	6,614	2.91	0.75	2.30	0.67	3.24	0.8
	IV	8	9,853	2.51	0.78	1.65	0.74	3.15	0.85
	V	8	10,517	2.73	0.79	1.95	0.74	3.17	0.86
<i>Math</i>	I	8	12,484	2.70	0.79	2.12	0.73	3.27	0.83
	II	8	6,569	2.70	0.78	2.05	0.66	3.16	0.84
	III	8	6,602	2.70	0.76	2.18	0.60	3.34	0.84
	IV	8	9,831	2.37	0.79	1.78	0.62	3.24	0.87
	V	8	10,485	2.76	0.78	2.23	0.74	3.13	0.84
<i>Science</i>	I	8	3,296	2.75	0.82	2.39	0.77	3.37	0.85
	III	8	3,267	2.71	0.75	2.35	0.66	2.96	0.81
	IV	8	3,190	2.47	0.75	2.25	0.68	2.79	0.79
	V	8	3,396	2.47	0.78	2.06	0.73	2.86	0.83

The distribution of students across task scores for each task in each CAPA level are presented 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 distributions 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 reported below:

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

where,

n is the number of items,

σ_i^2 is the variance of scores on the i -th item, and

σ_t^2 is the variance of the total score (either the total raw score or scale score).

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

$$\sigma_e = \sigma_t \sqrt{1 - \alpha} \quad (8.3)$$

where,

α is the reliability estimated using (8.2) above, and

σ_t is the standard deviation of the total raw scores.

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 for CAPA tests along with the number of items and examinees upon which those analyses were performed.

Table 8.3 Reliabilities and Standard Errors of Measurement for the CAPA

Subject Area	Level	No. of Items	No. of Examinees	Reliab.	Scale Score			Raw Score		
					Mean	S.D.	SEM	Mean	S.D.	SEM
<i>English– Language Arts</i>	I	8	12,531	0.91	40.84	12.02	3.68	26.85	12.02	3.67
	II	8	6,587	0.84	39.24	7.46	3.02	23.24	6.16	2.49
	III	8	6,614	0.86	39.12	5.94	2.21	23.19	6.09	2.26
	IV	8	9,853	0.88	39.19	7.75	2.73	20.05	7.10	2.50
	V	8	10,517	0.89	38.54	6.21	2.08	21.67	7.01	2.35
<i>Mathematics</i>	I	8	12,484	0.87	35.11	9.74	3.49	21.54	11.16	4.00
	II	8	6,569	0.88	37.60	9.56	3.31	21.58	7.45	2.58
	III	8	6,602	0.87	36.58	6.64	2.42	21.53	6.95	2.54
	IV	8	9,831	0.88	36.41	8.80	3.10	18.95	7.46	2.62
	V	8	10,485	0.87	37.51	8.85	3.14	21.91	7.62	2.70
<i>Science *</i>	I	8	3,296	0.91	35.59	11.25	3.46	21.81	12.22	3.76
	III	8	3,267	0.85	36.24	5.45	2.08	21.40	6.35	2.43
	IV	8	3,190	0.85	35.56	5.53	2.12	19.64	6.42	2.46
	V	8	3,396	0.87	35.35	5.34	1.93	19.58	6.35	2.30

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 language proficiency. As of 2009, reliability analyses are also presented for the subgroups categorized by whether or not examinees belonging to a certain ethnic subgroup were economically disadvantaged.

For each subgroup analysis, reliability and SEM information is reported for the total test scores. Table 8.B.1 through Table 8.B.6 present the overall reliabilities for the various subgroups. Note that the reliabilities are reported only for samples that are comprised of 11 or more examinees.

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 item response theory 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:

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

where,

CSEM($\hat{\theta}$) is the standard error of measurement and
I(θ) is the test information function.

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

Standard errors of measurement 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 in 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). RELCLASS-COMP estimates decision consistency using an estimated multivariate distribution of reported classifications on the current form of the exam and classifications on a hypothetical alternate form using the reliability of the test and strong true-score theory. RELCLASS-COMP also 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).

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 display the tables used in the analyses.

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	Mis-classification
	Achieves a performance level	Mis-classification	Correct classification

Figure 8.2 Decision Consistency for Achieving a Performance Level

		Decision made on the second form taken	
		Does not achieve a performance level	Achieves a performance level
Decision made on the first form taken	Does not achieve a performance level	Correct classification	Mis-classification
	Achieves a performance level	Mis-classification	Correct classification

The results of these analyses are presented in Table 8.B.7 through Table 8.B.20 in Appendix 8.B.

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

Also given are the results for the accuracy and consistency after the classifications have been collapsed into the two categories of below proficient versus proficient and above, which are the critical categories for AYP analyses.

Validity Evidence

Validity refers to the degree to which each interpretation or use of a test score is supported by evidence that is gathered (APA, AERA, & 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 does not involve a single study or gathering one particular kind of evidence. Validation involves multiple investigations and various kinds of evidence (APA, AERA, & 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 APA, AERA, and NCME's *The Standards for Educational and Psychological Testing* (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, 2009; see Appendix A).

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

Purpose of the CAPA

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

The Constructs to Be Measured

The CAPA is designed to show how well students with significant cognitive disabilities are performing relative to California content standards. The content standards were approved by the SBE and 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 subject area to be assessed the tasks to be presented, the administration instructions to be given, and the rules used to score examinee responses. They control as many aspects of the measurement procedure as possible, so that the testing conditions will remain the same over test administrations (Cronbach, 1971; Cronbach, Gleser, Nanda, & Rajaratnam, 1972) to minimize construct irrelevant score variance (Messick, 1989). The content blueprints for the CAPA can be found on the CDE 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.

The Scores Generated and the Interpretations and Uses of these Scores

Total scores expressed as scale scores, and student performance levels are generated for each subject area test. On the basis of a student's total score, an inference is drawn about how much knowledge and skill in the subject area the student has. The total score also is used to classify students in terms of their level of knowledge and skill in the subject area. The classifications are called performance levels and are labeled as follows: advanced, proficient, basic, below basic, and far below basic.

The tests that make up the CAPA testing program 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 (APA, AERA, & NCME, 1999, Standard 1.3). The user is advised to gather evidence to support these additional interpretations or uses (APA, AERA, & NCME, 1999, Standard, 1.4).

Intended Test Population(s)

Students with 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 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 APA, AERA, and NCME's *The Standards for Educational and Psychological Testing* (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 APA, AERA, 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, this does not provide compelling evidence to support a particular score interpretation or use, such evidence may prove useful in refuting rival explanations of test results. Evidence based on content includes the following:

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

Specifications and blueprints—ETS maintains item development 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 16.

Once the tasks are developed, ETS selects all CAPA test tasks to conform to the SBE-approved California content standards and test blueprints. Test blueprints for the components of the CAPA were proposed by ETS and reviewed and approved by the Assessment Review Panel (ARP). The ARP is an advisory panel to the CDE and ETS on areas related to item development for the CAPA. They 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 content 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 content and psychometric criteria applicable for the 2009 CAPA is presented in Chapter 4, starting on page 26.

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 go through multiple review cycles and involve multiple groups of reviewers. One of the reviews is carried out by an external reviewer, that is, the ARP. The ARP is 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 26.

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 22.

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

Human Resource Research Organization (HumRRO) performed an alignment study for the CAPA in April 2007. This reported was titled *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 CAPAs 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. Thus, the alignment is sufficient overall. A copy of the study is available 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 score on a test and measures of other variables external to the test can also provide evidence of validity when these relationships are found to be consistent with the definition of the construct that the test is intended to measure. As indicated in the APA, AERA, and NCME *Test Standards* (1999), the variables investigated can include other tests that measure the same construct and different constructs, criterion measures that score 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 items and all field-test items for which sufficient student samples were available.

The results of the DIF analyses are presented in Appendix 8.E. The vast majority of the items 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.

Intercorrelations Between Content Areas

To the degree that students' content area scores correlate as expected, evidence of the validity of regarding those scores as measures of the intended constructs is provided. Table 8.4 presents the correlations between scores on the CAPA tests, mean and standard deviation for raw score, and the numbers of students on which these correlations were based. At all test levels, the correlations between students' ELA and mathematics scores were in high. These tests' correlations with students Level I science scores also were high. For Levels III and above, the correlations of the ELA and mathematics scores with the science scores tended to be more moderate.

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

Table 8.4 CAPA Content Area Correlations for CAPA Levels

	N	Mean	S.D	ELA	Mathematics	Science
Level I						
ELA	12531	26.85	12.02	–	0.83	0.86
Mathematics	12484	21.54	11.16	–	–	0.86
Science	3296	21.81	12.22	–	–	–
Level II						
ELA	6587	23.24	6.16	–	0.80	NA
Mathematics	6569	21.58	7.45	–	–	NA
Science	NA	NA	NA	–	–	NA
Level III						
ELA	6614	23.19	6.09	–	0.83	0.81
Mathematics	6602	21.53	6.95	–	–	0.78
Science	3267	21.4	6.35	–	–	–
Level IV						
ELA	9853	20.05	7.1	–	0.81	0.77
Mathematics	9831	18.95	7.46	–	–	0.79
Science	3190	19.64	6.42	–	–	–
Level V						
ELA	10517	21.67	7.01	–	0.82	0.74
Mathematics	10485	21.91	7.62	–	–	0.73
Science	3396	19.58	6.35	–	–	–

Evidence Based on Response Processes

As noted in the APA, AERA, and NCME's *The Standards for Educational and Psychological Testing* (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 evidence provided by observers or judges involved in the scoring of examinee responses.

Evidence of Rater Reliability, Inter-rater Agreement

Rater consistency is critical to the 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' essay responses. In order to evaluate the reliability of the student scores, 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 inter-rater agreement.

Inter-rater Agreement

As noted previously, approximately ten percent of the test population's responses to the tasks were scored by two raters. The range of percentage of students for whom the raters were in exact agreement was 91 to 99 percent in ELA, 92 to 98 percent in mathematics, and 91 to 98 percent in Science for Level V.

The findings for each operational task for each test are presented in Table 8.C.36 through Table 8.C.40 in Appendix 8.C, which start on page 139.

Evidence Based on Internal Structure

As suggested by the *Standards*, 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 score 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 item 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 0.84 to 0.91.

Subgroup reliabilities—The reliabilities for various subgroups of the examinee population that differed in their demographic characteristics. The characteristics considered were gender, ethnicity, economic status, disability group, language proficiency, 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” (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 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, 1995), 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 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 items for the CAPA into discrete categories based on an evaluation of how well each item was fit by the Rasch 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.

The distributions of the operational and field-test tasks across the IRT model data fit classifications are presented in Table 8.D.1 through Table 8.D.6.

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 items 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 items and the field test items.

Scaling Results

Complete raw-to-scale score conversion tables for the 2009 CAPA are presented in Table 8.D.10 through Table 8.D.23 in Appendix 8.D, starting on page 146. The raw scores and corresponding unrounded converted scale scores are listed in those tables. For all of the

2009 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 categories are presented in Table 6.1, which is on page 42 in Chapter 6.

DIF Analyses

Analyses of differential item functioning (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 tasks being field-tested for which sufficient student samples were available. The sample size requirements for the field-test 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 unity 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 unity indicate DIF in favor of the focal group; a value of unity 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 1 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:

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

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.

The ETS classification system assigns tasks to one of 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:

- *A tasks or negligible DIF*: The Mantel Chi-square statistic is not statistically significant (at the 0.05 level) or $|SMD/SD| < 0.17$
- *B tasks or intermediate DIF*: The Mantel Chi-square statistic is statistically significant (at the 0.05 level) and $0.17 \leq |SMD/SD| < 0.25$
- *C tasks or large DIF*: The Mantel Chi-square statistic is statistically significant (at the 0.05 level) and $|SMD/SD| > 0.25$

In addition, the classifications identify which group is being advantaged. These classifications are displayed in Table 8.5. The categories have been used by all ETS testing programs for more than 13 years.

Table 8.5 DIF Flags Based on the ETS DIF Classification Scheme

Flag	Descriptor
A–	Low DIF favoring members of the reference group
B–	Moderate DIF favoring members of the reference group
C–	High DIF favoring members of the reference group
A+	Low DIF favoring members of the focal group
B+	Moderate DIF favoring members of the focal group
C+	High DIF favoring members of the focal group

Category C contains tasks with moderate to large values of DIF. As shown in Table 8.5, above, 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. 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 disadvantage a focal group (C-DIF) for future operational test forms unless their inclusion is deemed essential to meeting test-content specifications. Table 8.6 lists specific subgroups that were used for DIF analyses for the CAPA.

Table 8.6 Subgroup Classification for DIF Analyses

DIF Type	Reference Group	Focal Group
Gender	Male	Female
Race/Ethnicity	White	African American
		American Indian
		Asian
		Combined Asian Group (Asian/Pacific Islander/Filipino)
		Filipino
		Hispanic/Latin American
		Pacific Islander

DIF Type	Reference Group	Focal Group
Disability	Mental Retardation	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 subject area.

References

- AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, D.C: American Educational Research Association.
- Crocker, L. and Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, Vol. 16, pp. 292–334.
- Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). 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: Wiley.
- Dorans, N. J. and Holland, P. W. (1993). *DIF detection and description: Mantel-Haenszel and standardization*. In P. W. Holland and H. Wainer (Eds.), *Differential item functioning*. Hillsdale, NJ: Erlbaum, pp. 35–66.
- Dorans, N. J. and Kulick, E. (2006). Differential item functioning on the mini-mental state examination: an application of the Mantel-Haenszel and standardization procedures. *Medical Care*, pp. 44,107–14.
- Dorans, N. J. and Schmitt, A. P. (1993). Constructed response and differential item functioning: A pragmatic approach. In R.E. Bennett and W.C. Ward (Eds.), *Construction versus choice in cognitive measurement*. Hillsdale, NH: Erlbaum, pp. 135–165.
- Drasgow F. (1988). Polychoric and polyserial correlations. In L. Kotz and N. L. Johnson (Eds.), *Encyclopedia of statistical sciences*. New York: Wiley, pp. 7, 69-74.
- Educational Testing Service. (2002). *ETS standards for quality and fairness*. Princeton, NJ: Educational Testing Service.
- Feldt, L. S. and Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement*. New York: Macmillan.
- Holland, P. W. and Thayer, D. T. (1985). *An alternative definition of the ETS delta scale of item difficulty*. RR-85–43.
- Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed.). Washington, DC: American Council on Education and National Council on Measurement in Education.
- Livingston, S. A., and Lewis, C. (1995). Estimating the consistency and accuracy of classification based on test scores. *Journal on Educational Measurement*, Vol. 32, pp. 179–97.
- Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analyses of data from retrospective studies of disease. *Journal of the National Cancer Institute*, Vol. 22, pp. 719–48.

- Messick, S. Validity. (1989). In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13–103). Edited by R. L. Linn. New York: Macmillan.
- Muraki, E. and Bock, R. D. (1995). *PARSCALE: Parameter scaling of rating data* (Version 2.2). Chicago, IL: Scientific Software, Inc.

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	3.47	0.82	
1/7 *	2	2.46	0.66	
Operational	3	2.81	0.78	
Operational	4	3.33	0.83	
1/7 *	5	2.42	0.72	
Operational	6	3.67	0.69	
Operational	7	3.40	0.84	
1/7 *	8	2.46	0.71	
Operational	9	3.49	0.86	
Operational	10	3.50	0.86	
1/7 *	11	3.35	0.76	
Operational	12	3.27	0.78	
2/8 *	2	3.94	0.70	
2/8 *	5	2.88	0.70	
2/8 *	8	2.78	0.70	
2/8 *	11	2.97	0.77	
3	2	2.99	0.68	
3	5	2.94	0.73	
3	8	3.95	0.63	
3	11	3.07	0.73	
4	2	3.07	0.75	
4	5	3.88	0.64	
4	8	2.98	0.62	
4	11	3.64	0.68	
5	2	3.55	0.58	R
5	5	2.95	0.78	
5	8	2.37	0.78	
5	11	2.94	0.82	
6	2	3.40	0.73	
6	5	2.85	0.70	
6	8	4.18	0.71	H
6	11	3.44	0.63	

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

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	3.83	0.65	H
1/ 5 *	2	1.85	0.48	R
Operational	3	2.40	0.77	
Operational	4	2.58	0.81	
1/ 5 *	5	3.11	0.60	
Operational	6	2.78	0.80	
Operational	7	3.62	0.74	H
1/ 5 *	8	2.83	0.65	
Operational	9	2.61	0.77	
Operational	10	3.11	0.75	
1/ 5 *	11	3.42	0.67	H
Operational	12	2.36	0.71	
2/ 6 *	2	2.96	0.45	R
2/ 6 *	5	2.93	0.69	
2/ 6 *	8	2.19	0.72	
2/ 6 *	11	3.44	0.70	H
3/ 7 *	2	2.47	0.66	
3/ 7 *	5	2.10	0.44	R
3/ 7 *	8	3.34	0.66	H
3/ 7 *	11	2.49	0.63	
4/ 8 *	2	2.67	0.71	
4/ 8 *	5	3.29	0.72	H
4/ 8 *	8	2.56	0.69	
4/ 8 *	11	2.70	0.66	

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

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	3.22	0.67	H
1/ 5 *	2	3.22	0.71	H
Operational	3	2.55	0.80	
Operational	4	2.91	0.76	
1/ 5 *	5	2.27	0.75	
Operational	6	2.30	0.72	
Operational	7	3.24	0.75	H
1/ 5 *	8	2.55	0.78	
Operational	9	2.94	0.78	
Operational	10	3.03	0.77	
1/ 5 *	11	3.25	0.76	H
Operational	12	3.12	0.73	
2/ 6 *	2	2.64	0.67	
2/ 6 *	5	2.17	0.72	
2/ 6 *	8	2.88	0.57	R
2/ 6 *	11	2.25	0.65	
3/ 7 *	2	2.74	0.65	
3/ 7 *	5	3.14	0.69	
3/ 7 *	8	2.89	0.65	
3/ 7 *	11	3.26	0.73	H
4/ 8 *	2	2.57	0.58	R
4/ 8 *	5	3.33	0.68	H
4/ 8 *	8	2.80	0.66	
4/ 8 *	11	2.33	0.76	

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

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	3.15	0.75	
1/ 8 *	2	3.37	0.64	H
Operational	3	1.65	0.74	
Operational	4	3.07	0.78	
1/ 8 *	5	2.59	0.73	
Operational	6	2.58	0.82	
Operational	7	2.32	0.81	
1/ 8 *	8	2.80	0.66	
Operational	9	2.05	0.85	
Operational	10	2.59	0.74	
1/ 8 *	11	3.25	0.69	H
Operational	12	2.68	0.78	C-DIF
2	2	2.51	0.69	
2	5	3.30	0.46	R H
2	8	3.21	0.54	R H
2	11	3.06	0.70	
3	2	3.56	0.58	R H
3	5	3.02	0.70	
3	8	2.99	0.67	
3	11	3.39	0.74	H
4	2	2.96	0.50	R
4	5	3.67	0.64	H
4	8	2.70	0.64	
4	11	2.26	0.71	
5	2	2.39	0.77	
5	5	2.94	0.64	
5	8	2.04	0.81	
5	11	2.17	0.75	
6	2	2.11	0.80	
6	5	2.35	0.80	
6	8	2.66	0.64	
6	11	3.32	0.60	R H
7	2	2.56	0.76	
7	5	2.95	0.64	
7	8	2.28	0.76	
7	11	2.05	0.65	

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

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	3.17	0.79	
1/7 *	2	2.61	0.52	R
Operational	3	2.72	0.83	
Operational	4	2.67	0.74	
1/7 *	5	3.13	0.70	
Operational	6	2.98	0.86	
Operational	7	2.97	0.74	
1/7 *	8	2.04	0.84	
Operational	9	1.95	0.82	
Operational	10	2.96	0.78	C-DIF
1/7 *	11	2.58	0.80	
Operational	12	2.38	0.79	C-DIF
2/8 *	2	2.84	0.63	
2/8 *	5	2.89	0.77	C-DIF
2/8 *	8	3.28	0.65	H C-DIF
2/8 *	11	2.95	0.72	
3	2	3.5	0.73	H
3	5	2.57	0.73	
3	8	3.09	0.71	
3	11	2.93	0.69	
4	2	3.5	0.74	H
4	5	2.46	0.77	
4	8	2.91	0.61	
4	11	2.79	0.83	
5	2	2.98	0.79	
5	5	2.44	0.81	
5	8	2.84	0.70	
5	11	2.36	0.80	
6	2	2.95	0.64	
6	5	2.64	0.76	
6	8	3.09	0.63	
6	11	2.82	0.69	

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

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	3.27	0.83	
1/7 *	14	2.79	0.78	
Operational	15	2.92	0.79	
Operational	16	2.58	0.73	
1/7 *	17	2.58	0.73	
Operational	18	2.68	0.76	
Operational	19	2.12	0.80	
1/7 *	20	2.74	0.77	
Operational	21	2.71	0.81	
Operational	22	2.39	0.80	
1/7 *	23	2.07	0.67	
Operational	24	2.96	0.81	
2/8 *	14	3.02	0.77	
2/8 *	17	2.99	0.72	
2/8 *	20	3.01	0.74	
2/8 *	23	2.50	0.72	
3	14	2.47	0.63	
3	17	2.99	0.73	
3	20	3.02	0.68	
3	23	2.80	0.60	R
4	14	3.20	0.76	
4	17	2.37	0.69	
4	20	2.81	0.74	
4	23	2.49	0.69	
5	14	2.18	0.70	
5	17	2.66	0.76	
5	20	3.01	0.75	
5	23	2.19	0.71	
6	14	3.24	0.72	
6	17	2.67	0.72	
6	20	2.79	0.77	
6	23	2.74	0.73	

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

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	3.16	0.79	
1/ 5 *	14	3.4	0.72	H
Operational	15	2.73	0.84	
Operational	16	2.69	0.80	
1/ 5 *	17	3.01	0.63	
Operational	18	2.05	0.79	
Operational	19	2.58	0.66	
1/ 5 *	20	2.47	0.71	
Operational	21	2.36	0.83	
Operational	22	2.99	0.78	
1/ 5 *	23	2.67	0.77	
Operational	24	3.02	0.76	
2/ 6 *	14	2.99	0.73	
2/ 6 *	17	2.71	0.61	
2/ 6 *	20	2.7	0.72	
2/ 6 *	23	1.44	0.50	R
3/ 7 *	14	3.19	0.71	
3/ 7 *	17	3.28	0.66	H
3/ 7 *	20	2.3	0.35	R
3/ 7 *	23	2.68	0.54	R
4/ 8 *	14	3.39	0.63	H
4/ 8 *	17	3.38	0.66	H
4/ 8 *	20	2.89	0.62	
4/ 8 *	23	2.74	0.68	

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

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	3.02	0.83	
1/ 5 *	14	2.5	0.56	R
Operational	15	2.18	0.63	
Operational	16	3.08	0.77	
1/ 5 *	17	3.04	0.83	
Operational	18	2.82	0.84	
Operational	19	2.43	0.60	R
1/ 5 *	20	2	0.57	R
Operational	21	2.48	0.82	C-DIF
Operational	22	3.34	0.81	H
1/ 5 *	23	2.86	0.66	
Operational	24	2.28	0.77	
2/ 6 *	14	2.26	0.62	
2/ 6 *	17	2.01	0.54	R
2/ 6 *	20	3.29	0.72	H
2/ 6 *	23	3.24	0.57	R H
3/ 7 *	14	2.3	0.76	
3/ 7 *	17	3.08	0.70	
3/ 7 *	20	2.45	0.42	R
3/ 7 *	23	3.28	0.46	R H
4/ 8 *	14	1.91	0.58	R
4/ 8 *	17	2.13	0.74	
4/ 8 *	20	3.26	0.77	H
4/ 8 *	23	2.18	0.52	R

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

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	2.3	0.87	
1/ 8 *	14	1.22	0.44	R
Operational	15	1.78	0.80	
Operational	16	2.04	0.71	
1/ 8 *	17	2.91	0.53	R
Operational	18	2.95	0.86	
Operational	19	2.03	0.86	
1/ 8 *	20	3.23	0.79	H
Operational	21	3.24	0.62	H
Operational	22	2.41	0.83	
1/ 8 *	23	2.03	0.55	R
Operational	24	2.23	0.73	
2	14	3.11	0.53	R
2	17	2.88	0.72	
2	20	2.94	0.60	
2	23	2.48	0.54	R
3	14	3.06	0.73	
3	17	2.68	0.65	
3	20	2.64	0.84	
3	23	3.08	0.64	
4	14	2.57	0.51	R
4	17	2.97	0.69	
4	20	2.24	0.54	R
4	23	2.6	0.66	
5	14	2.78	0.57	R
5	17	1.86	0.86	
5	20	1.81	0.76	
5	23	2.04	0.56	R
6	14	2.93	0.46	R
6	17	2.67	0.84	
6	20	1.89	0.79	
6	23	2.4	0.86	
7	14	2.94	0.38	R
7	17	2.46	0.53	R
7	20	1.77	0.71	
7	23	2.92	0.79	

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

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	2.97	0.74	
1/7 *	14	2.73	0.74	
Operational	15	3.1	0.81	
Operational	16	2.82	0.74	
1/7 *	17	1.91	0.61	
Operational	18	2.77	0.83	
Operational	19	2.4	0.78	
1/7 *	20	2.43	0.58	R
Operational	21	2.68	0.75	
Operational	22	2.23	0.75	
1/7 *	23	3.03	0.83	
Operational	24	3.13	0.84	
2/8 *	14	2.65	0.75	
2/8 *	17	2.41	0.71	
2/8 *	20	3.29	0.77	H
2/8 *	23	2.44	0.69	
3	14	2.38	0.55	R
3	17	1.67	0.61	
3	20	3.15	0.76	
3	23	2.73	0.58	R
4	14	3.34	0.81	H
4	17	1.72	0.65	
4	20	2.44	0.53	R
4	23	2.84	0.76	
5	14	2.19	0.70	
5	17	3.02	0.84	
5	20	3.03	0.64	
5	23	2.98	0.84	
6	14	3.32	0.77	H
6	17	3.16	0.77	
6	20	3.15	0.62	
6	23	2.34	0.72	

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

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	2.70	0.85	
1/3/5/7 *	26	4.05	0.69	H
Operational	27	2.59	0.82	
Operational	28	2.83	0.82	
1/3/5/7 *	29	3.89	0.74	
Operational	30	2.83	0.85	
Operational	31	2.79	0.77	
1/3/5/7 *	32	2.76	0.78	
Operational	33	2.39	0.82	
Operational	34	2.52	0.83	
1/3/5/7 *	35	2.78	0.81	
Operational	36	3.37	0.83	
2/4/6/8 *	26	3.00	0.74	
2/4/6/8 *	29	3.24	0.68	
2/4/6/8 *	32	3.58	0.71	
2/4/6/8 *	35	2.96	0.77	

* 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	2.89	0.66	
1/3/5/7 *	26	3.22	0.69	H
Operational	27	2.96	0.73	
Operational	28	2.9	0.72	
1/3/5/7 *	29	2.13	0.62	
Operational	30	2.6	0.80	
Operational	31	2.44	0.81	
1/3/5/7 *	32	2.24	0.64	
Operational	33	2.7	0.78	
Operational	34	2.35	0.69	
1/3/5/7 *	35	3.06	0.63	
Operational	36	2.8	0.77	
2/4/6/8 *	26	3.39	0.66	H
2/4/6/8 *	29	2.51	0.56	R
2/4/6/8 *	32	3.02	0.73	
2/4/6/8 *	35	2.77	0.61	

* 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	2.44	0.76	
1/3/5/7 *	26	2.8	0.72	
Operational	27	2.46	0.74	
Operational	28	2.25	0.74	
1/3/5/7 *	29	2.39	0.71	
Operational	30	2.31	0.74	
Operational	31	2.79	0.68	
1/3/5/7 *	32	2.97	0.68	
Operational	33	2.61	0.75	
Operational	34	2.5	0.77	
1/3/5/7 *	35	2.27	0.71	
Operational	36	2.37	0.79	
2/4/6/8 *	26	2.08	0.68	
2/4/6/8 *	29	2.67	0.66	
2/4/6/8 *	32	2.61	0.72	
2/4/6/8 *	35	3.14	0.65	

* 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	2.32	0.74	
1/3/5/7 *	26	2.14	0.72	
Operational	27	2.25	0.74	
Operational	28	2.64	0.79	
1/3/5/7 *	29	2.37	0.68	
Operational	30	2.06	0.76	
Operational	31	2.78	0.83	
1/3/5/7 *	32	2.43	0.69	
Operational	33	2.64	0.82	
Operational	34	2.22	0.73	
1/3/5/7 *	35	1.89	0.68	
Operational	36	2.86	0.81	
2/4/6/8 *	26	2.71	0.51	R
2/4/6/8 *	29	2.07	0.67	
2/4/6/8 *	32	2.17	0.61	
2/4/6/8 *	35	2.52	0.60	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	1,100	8.79	1,263	10.11	1,198	9.59	1,305	10.47	1,148	9.19	1,174	9.41	1,173	9.40	1,063	8.53
	1	2,445	19.54	3,883	31.08	2,708	21.68	1,634	13.11	2,599	20.80	2,379	19.07	2,332	18.69	3,095	24.85
	2	533	4.26	1,120	8.97	590	4.72	404	3.24	576	4.61	556	4.46	546	4.38	587	4.71
	3	511	4.08	623	4.99	604	4.84	498	3.99	558	4.47	488	3.91	485	3.89	539	4.33
	4	1,114	8.90	819	6.56	1,081	8.65	1,351	10.84	945	7.56	817	6.55	901	7.22	1,017	8.17
	5	6,807	54.41	4,785	38.30	6,310	50.52	7,275	58.35	6,667	53.37	7,064	56.61	7,041	56.43	6,154	49.41
II	0	44	0.65	201	2.95	205	3.01	182	2.67	70	1.03	140	2.06	160	2.35	161	2.38
	1	95	1.39	1,955	28.73	1,927	28.33	1,518	22.31	207	3.04	1,468	21.57	585	8.60	1,596	23.59
	2	177	2.60	1,293	19.00	1,206	17.73	1,303	19.15	391	5.74	1,668	24.51	1,281	18.83	1,989	29.40
	3	379	5.56	1,605	23.59	660	9.70	445	6.54	944	13.85	1,113	16.36	1,208	17.76	1,605	23.73
	4	6,120	89.80	1,751	25.73	2,803	41.21	3,357	49.33	5,202	76.34	2,416	35.50	3,569	52.46	1,414	20.90
III	0	80	1.17	297	4.36	159	2.33	160	2.35	59	0.87	97	1.42	108	1.59	84	1.25
	1	404	5.92	998	14.64	621	9.12	961	14.10	329	4.82	1,394	20.47	928	13.62	801	11.88
	2	852	12.49	1,810	26.56	1,294	19.00	3,218	47.21	685	10.05	1,121	16.46	911	13.37	1,112	16.49
	3	2,144	31.43	2,124	31.17	2,343	34.40	1,747	25.63	2,648	38.83	412	6.05	1,588	23.31	1,053	15.62
	4	3,341	48.98	1,586	23.27	2,395	35.16	731	10.72	3,098	45.43	3,786	55.59	3,278	48.11	3,693	54.77
IV	0	102	1.02	742	7.44	432	4.33	417	4.18	485	4.86	275	2.76	206	2.07	373	3.77
	1	1,188	11.88	5,509	55.27	1,259	12.62	2,111	21.14	2,583	25.86	4,517	45.35	1,810	18.16	2,428	24.55
	2	1,384	13.84	1,422	14.27	988	9.90	1,713	17.15	2,395	23.98	1,535	15.41	2,584	25.92	1,478	14.94
	3	1,694	16.94	1,076	10.79	1,738	17.42	2,753	27.57	2,271	22.74	1,613	16.19	2,614	26.22	1,279	12.93
	4	5,630	56.31	1,219	12.23	5,559	55.72	2,992	29.96	2,255	22.57	2,020	20.28	2,755	27.64	4,332	43.80
V	0	172	1.63	259	2.46	286	2.72	200	1.90	161	1.53	324	3.08	430	4.09	282	2.69
	1	1,681	15.92	1,758	16.67	1,609	15.27	1,159	11.00	1,239	11.77	5,105	48.60	1,485	14.11	3,525	33.66
	2	1,280	12.12	1,772	16.80	2,225	21.12	1,853	17.59	1,861	17.68	1,503	14.31	1,288	12.24	1,372	13.10
	3	537	5.09	3,672	34.81	3,599	34.17	2,756	26.16	2,808	26.67	2,141	20.38	2,241	21.30	2,594	24.77
	4	6,890	65.25	3,088	29.27	2,815	26.72	4,569	43.36	4,459	42.35	1,432	13.63	5,079	48.27	2,698	25.77

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	1,231	9.87	1,253	10.06	1,190	9.57	1,258	10.12	1,290	10.36	1,384	11.14	1,433	11.53	1,377	11.15
	1	2,960	23.74	3,852	30.92	4,925	39.60	4,494	36.15	6,119	49.15	4,231	34.04	5,068	40.78	3,576	28.95
	2	549	4.40	615	4.94	649	5.22	721	5.80	895	7.19	724	5.83	866	6.97	594	4.81
	3	505	4.05	707	5.67	738	5.93	724	5.82	677	5.44	704	5.66	740	5.95	618	5.00
	4	927	7.44	1,102	8.85	1,005	8.08	1,017	8.18	814	6.54	964	7.76	943	7.59	912	7.38
II	0	6,294	50.49	4,930	39.57	3,929	31.59	4,216	33.92	2,654	21.32	4,422	35.58	3,377	27.17	5,275	42.71
	1	198	2.91	220	3.24	175	2.58	245	3.61	172	2.54	218	3.21	167	2.46	109	1.62
	2	777	11.43	1,870	27.52	1,438	21.17	3,477	51.24	1,123	16.57	2,548	37.56	1,090	16.08	1,063	15.84
	3	700	10.30	684	10.07	1,322	19.46	594	8.75	1,815	26.77	897	13.22	873	12.88	1,020	15.20
	4	1,129	16.61	692	10.18	1,191	17.53	523	7.71	2,013	29.69	777	11.46	1,033	15.24	879	13.10
III	0	3,994	58.75	3,329	48.99	2,668	39.27	1,947	28.69	1,656	24.43	2,343	34.54	3,616	53.34	3,641	54.25
	1	76	1.12	116	1.70	116	1.70	122	1.79	104	1.53	156	2.29	97	1.42	140	2.06
	2	1,464	21.49	1,556	22.87	930	13.66	1,977	29.05	1,648	24.19	2,457	36.12	783	11.50	2,604	38.34
	3	668	9.80	2,927	43.01	824	12.11	518	7.61	1,801	26.43	767	11.28	541	7.95	1,284	18.90
	4	776	11.39	1,530	22.48	1,433	21.05	625	9.18	1,903	27.93	789	11.60	650	9.55	805	11.85
IV	0	3,830	56.21	676	9.93	3,503	51.47	3,564	52.37	1,357	19.92	2,633	38.71	4,738	69.58	1,959	28.84
	1	324	3.25	542	5.44	202	2.03	306	3.08	377	3.78	168	1.69	233	2.34	332	3.35
	2	3,816	38.27	5,919	59.40	4,547	45.66	2,461	24.78	5,042	50.60	916	9.22	3,740	37.59	3,171	31.96
	3	933	9.36	810	8.13	2,143	21.52	412	4.15	993	9.96	1,126	11.33	1,282	12.89	1,832	18.46
	4	2,206	22.12	628	6.30	783	7.86	816	8.22	945	9.48	1,969	19.81	1,063	10.68	3,060	30.84
V	0	2,692	27.00	2,065	20.72	2,283	22.93	5,938	59.78	2,608	26.17	5,759	57.95	3,631	36.50	1,527	15.39
	1	190	1.81	239	2.28	227	2.17	205	1.95	217	2.07	297	2.84	256	2.44	261	2.52
	2	2,465	23.42	1,877	17.90	2,568	24.57	1,498	14.28	4,189	39.89	3,550	33.92	3,901	37.21	2,130	20.55
	3	828	7.87	518	4.94	1,219	11.66	3,030	28.88	1,274	12.13	598	5.71	2,178	20.77	448	4.32
	4	1,023	9.72	1,801	17.17	1,368	13.09	1,583	15.09	1,001	9.53	927	8.86	1,637	15.61	759	7.32
	6,018	57.18	6,052	57.71	5,070	48.51	4,176	39.80	3,821	36.38	5,093	48.67	2,512	23.96	6,769	65.29	

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

Science Level	Score on Task	1		2		3		4		5		6		7		8	
		Count	Percent														
I	0	432	11.40	471	12.50	422	11.21	425	11.28	413	10.95	575	15.34	485	12.91	380	10.44
	1	1,295	34.17	1,330	36.29	1,191	31.63	1,219	32.34	1,215	32.23	1,396	37.24	1,348	35.89	741	20.35
	2	220	5.80	206	5.47	191	5.07	175	4.64	202	5.36	201	5.36	243	6.47	169	4.64
	3	209	5.51	222	5.89	220	5.84	184	4.88	232	6.15	220	5.87	204	5.43	141	3.87
	4	228	6.02	253	6.71	266	7.07	235	6.24	307	8.14	286	7.63	287	7.64	228	6.26
III	0	1,406	37.10	1,287	34.15	1,475	39.18	1,531	40.62	1,401	37.16	1,071	28.57	1,189	31.66	1,982	54.44
	1	56	1.61	61	1.76	30	0.87	127	3.65	84	2.41	75	2.16	68	1.96	73	2.13
	2	449	12.88	439	12.64	266	7.68	733	21.06	1,232	35.38	830	23.90	864	24.90	430	12.53
	3	366	10.50	703	20.24	837	24.16	661	18.99	625	17.95	420	12.09	1,045	30.12	651	18.96
	4	1,701	48.80	775	22.32	1,271	36.68	957	27.49	285	8.18	1,035	29.80	857	24.70	1,302	37.93
IV	0	914	26.22	1,495	43.05	1,061	30.62	1,003	28.81	1,256	36.07	1,113	32.05	636	18.33	977	28.46
	1	67	1.93	70	2.03	95	2.75	85	2.47	97	2.81	74	2.15	95	2.75	96	2.80
	2	1,017	29.35	701	20.31	1,148	33.23	915	26.55	570	16.53	635	18.42	666	19.29	915	26.71
	3	551	15.90	1,018	29.49	686	19.86	1,029	29.86	620	17.98	827	23.98	740	21.43	800	23.35
	4	962	27.76	969	28.07	837	24.23	755	21.91	851	24.68	1,013	29.38	1,371	39.70	857	25.01
V	0	868	25.05	694	20.10	689	19.94	662	19.21	1,310	37.99	899	26.07	581	16.83	758	22.12
	1	83	2.17	138	3.65	94	2.48	149	3.93	109	2.86	127	3.35	99	2.61	103	2.76
	2	949	24.86	979	25.86	606	15.97	1,267	33.40	427	11.22	482	12.73	1,264	33.35	647	17.36
	3	948	24.83	1,110	29.32	740	19.50	1,109	29.24	761	19.99	639	16.88	903	23.83	608	16.31
	4	1,287	33.71	932	24.62	1,492	39.33	740	19.51	1,409	37.01	1,860	49.13	766	20.21	730	19.59
	0	551	14.43	627	16.56	862	22.72	528	13.92	1,101	28.92	678	17.91	758	20.00	1,639	43.98

Appendix 8.B—Reliabilities

Table 8.B.1 Reliabilities and SEMs by GENDER

Subject Area	Level	No. of Items	Male		Female		Unknown Gender	
			Reliab.	SEM	Reliab.	SEM	Reliab.	SEM
<i>English– Language Arts</i>	I	8	0.90	3.69	0.91	3.64	0.86	4.14
	II	8	0.85	2.48	0.81	2.52	0.82	2.32
	III	8	0.87	2.25	0.85	2.28	0.88	2.36
	IV	8	0.87	2.52	0.88	2.45	0.89	2.50
	V	8	0.89	2.36	0.89	2.32	0.89	2.18
<i>Mathematics</i>	I	8	0.87	4.02	0.88	3.98	0.83	4.15
	II	8	0.88	2.57	0.87	2.58	0.88	2.49
	III	8	0.87	2.53	0.86	2.55	0.88	2.67
	IV	8	0.88	2.64	0.87	2.59	0.87	2.70
	V	8	0.87	2.70	0.87	2.71	0.88	2.49
<i>Science *</i>	I	8	0.90	3.76	0.91	3.76	0.92	3.27
	III	8	0.87	2.40	0.82	2.47	0.94	2.03
	IV	8	0.86	2.46	0.84	2.45	0.96	1.83
	V	8	0.87	2.29	0.86	2.31	0.75	2.47

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

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

Subject Area	Level	No. of Items	American Indian		Asian		Pacific Islander		Filipino	
			Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM
<i>English–Language Arts</i>	I	8	0.92	3.41	0.88	3.92	0.86	4.16	0.87	3.90
	II	8	0.83	2.45	0.85	2.56	0.78	2.45	0.84	2.44
	III	8	0.87	2.3	0.86	2.43	0.81	2.45	0.88	2.39
	IV	8	0.88	2.54	0.88	2.48	0.87	2.52	0.87	2.54
	V	8	0.90	2.15	0.90	2.43	0.90	2.31	0.88	2.48
<i>Mathematics</i>	I	8	0.83	4.27	0.85	4.06	0.83	4.18	0.83	4.15
	II	8	0.90	2.41	0.89	2.62	0.83	2.73	0.88	2.65
	III	8	0.87	2.59	0.85	2.65	0.87	2.60	0.87	2.57
	IV	8	0.86	2.69	0.88	2.66	0.84	2.69	0.88	2.63
	V	8	0.85	2.78	0.89	2.75	0.83	2.68	0.89	2.64
<i>Science *</i>	I	8	0.86	3.62	0.89	3.89	0.91	3.94	0.88	3.89
	III	8	0.85	2.44	0.85	2.53	0.84	2.33	0.87	2.4
	IV	8	0.84	2.36	0.85	2.45	0.77	2.50	0.88	2.48
	V	8	0.85	2.32	0.89	2.22	0.90	2.38	0.88	2.24
Subject Area	Level	No. of Items	Hispanic		African American		White		Unknown Ethnicity	
			Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM
<i>English–Language Arts</i>	I	8	0.92	3.57	0.92	3.6	0.89	3.79	0.90	3.72
	II	8	0.84	2.48	0.83	2.5	0.83	2.48	0.86	2.55
	III	8	0.85	2.24	0.84	2.20	0.87	2.26	0.92	2.22
	IV	8	0.88	2.45	0.88	2.48	0.87	2.54	0.89	2.46
	V	8	0.88	2.32	0.88	2.32	0.89	2.35	0.89	2.39
<i>Mathematics</i>	I	8	0.88	3.97	0.89	3.85	0.86	4.09	0.86	4.04
	II	8	0.88	2.58	0.88	2.59	0.88	2.54	0.91	2.59
	III	8	0.87	2.50	0.86	2.53	0.87	2.56	0.88	2.62
	IV	8	0.88	2.60	0.88	2.63	0.87	2.64	0.88	2.68
	V	8	0.87	2.68	0.87	2.73	0.88	2.71	0.86	2.74
<i>Science *</i>	I	8	0.91	3.75	0.92	3.49	0.9	3.82	0.91	3.61
	III	8	0.84	2.43	0.84	2.41	0.87	2.40	0.92	2.28
	IV	8	0.85	2.45	0.86	2.45	0.85	2.46	0.86	2.50
	V	8	0.86	2.30	0.85	2.31	0.87	2.31	0.85	2.43

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

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

Subject Area	Level	No. of Items	American Indian		Asian		Pacific Islander		Filipino	
			Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM
<i>English–Language Arts</i>	I	8	0.93	3.33	0.89	3.82	0.88	3.93	0.89	3.78
	II	8	0.79	2.37	0.86	2.58	0.81	2.33	0.86	2.43
	III	8	0.88	2.24	0.86	2.39	0.76	2.24	0.86	2.29
	IV	8	0.85	2.59	0.88	2.48	0.88	2.47	0.88	2.58
	V	8	0.88	2.18	0.91	2.40	0.92	2.15	0.89	2.51
<i>Mathematics</i>	I	8	0.80	4.67	0.85	4.07	0.87	4.05	0.84	4.14
	II	8	0.91	2.35	0.90	2.59	0.81	2.80	0.91	2.49
	III	8	0.88	2.57	0.84	2.65	0.86	2.54	0.87	2.53
	IV	8	0.86	2.70	0.87	2.67	0.84	2.65	0.86	2.63
	V	8	0.81	2.83	0.89	2.70	0.81	2.67	0.91	2.62
<i>Science *</i>	I	8	0.83	4.06	0.89	3.92	0.93	3.66	0.85	4.23
	III	8	0.89	2.29	0.86	2.59	0.80	2.28	0.86	2.33
	IV	8	0.85	2.39	0.85	2.41	0.80	2.38	0.83	2.54
	V	8	0.83	2.38	0.91	2.12	–	–	0.91	2.19
Subject Area	Level	No. of Items	Hispanic		African American		White		Unknown Ethnicity	
			Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM
<i>English–Language Arts</i>	I	8	0.91	3.55	0.92	3.55	0.90	3.74	0.91	3.61
	II	8	0.83	2.47	0.83	2.46	0.81	2.46	0.78	2.44
	III	8	0.85	2.23	0.86	2.15	0.86	2.19	0.93	2.10
	IV	8	0.88	2.45	0.86	2.47	0.86	2.53	0.92	2.26
	V	8	0.88	2.30	0.88	2.31	0.88	2.30	0.89	2.41
<i>Mathematics</i>	I	8	0.88	3.98	0.89	3.89	0.86	4.11	0.86	3.90
	II	8	0.88	2.56	0.87	2.56	0.88	2.48	0.85	2.70
	III	8	0.86	2.50	0.86	2.51	0.86	2.53	0.92	2.28
	IV	8	0.88	2.59	0.87	2.65	0.87	2.62	0.92	2.50
	V	8	0.87	2.67	0.86	2.72	0.87	2.67	0.77	3.02
<i>Science *</i>	I	8	0.91	3.74	0.92	3.50	0.90	3.79	0.90	3.85
	III	8	0.83	2.42	0.84	2.36	0.84	2.44	0.91	2.21
	IV	8	0.85	2.44	0.84	2.43	0.83	2.43	0.89	2.45
	V	8	0.86	2.30	0.83	2.30	0.89	2.23	0.68	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

Subject Area	Level	No. of Items	American Indian		Asian		Pacific Islander		Filipino	
			Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM
<i>English–Language Arts</i>	I	8	0.91	3.52	0.88	3.98	0.83	4.36	0.86	3.94
	II	8	0.84	2.58	0.83	2.55	0.79	2.43	0.84	2.45
	III	8	0.83	2.46	0.86	2.44	–	–	0.88	2.43
	IV	8	0.92	2.41	0.88	2.48	0.87	2.60	0.88	2.50
	V	8	0.94	2.10	0.90	2.43	0.89	2.46	0.87	2.48
<i>Mathematics</i>	I	8	0.85	3.98	0.85	4.05	0.76	4.36	0.83	4.16
	II	8	0.88	2.53	0.88	2.66	0.89	2.58	0.87	2.71
	III	8	0.84	2.62	0.85	2.64	–	–	0.88	2.58
	IV	8	0.86	2.61	0.88	2.66	0.85	2.55	0.89	2.63
	V	8	0.90	2.73	0.88	2.81	0.88	2.66	0.87	2.68
<i>Science *</i>	I	8	–	–	0.90	3.80	–	–	0.89	3.73
	III	8	–	–	0.85	2.47	–	–	0.87	2.44
	IV	8	0.85	2.20	0.85	2.48	–	–	0.89	2.46
	V	8	–	–	0.86	2.34	–	–	0.86	2.28
Subject Area	Level	No. of Items	Hispanic		African American		White		Unknown Ethnicity	
			Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM
<i>English–Language Arts</i>	I	8	0.92	3.63	0.92	3.71	0.89	3.82	0.86	4.19
	II	8	0.85	2.53	0.84	2.56	0.83	2.50	0.88	2.59
	III	8	0.88	2.27	0.81	2.36	0.88	2.30	0.91	2.35
	IV	8	0.89	2.44	0.90	2.47	0.87	2.55	0.88	2.57
	V	8	0.90	2.38	0.89	2.36	0.89	2.39	0.92	2.41
<i>Mathematics</i>	I	8	0.89	3.92	0.90	3.78	0.85	4.09	0.82	4.26
	II	8	0.88	2.62	0.89	2.62	0.87	2.59	0.93	2.54
	III	8	0.89	2.49	0.85	2.57	0.87	2.58	0.84	2.82
	IV	8	0.88	2.64	0.90	2.55	0.86	2.65	0.88	2.64
	V	8	0.88	2.71	0.87	2.74	0.88	2.74	0.91	2.57
<i>Science *</i>	I	8	0.91	3.75	0.93	3.46	0.90	3.83	0.89	3.92
	III	8	0.86	2.43	0.85	2.54	0.89	2.37	0.85	2.48
	IV	8	0.86	2.52	0.89	2.44	0.85	2.47	0.91	2.29
	V	8	0.89	2.28	0.88	2.28	0.86	2.35	0.90	2.45

* 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

Subject Area	Level	No. of Items	American Indian		Asian		Pacific Islander		Filipino	
			Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM
<i>English–Language Arts</i>	I	8	–	–	0.66	4.68	–	–	–	–
	II	8	–	–	0.87	2.46	–	–	–	–
	III	8	–	–	0.85	2.84	–	–	–	–
	IV	8	–	–	0.91	2.32	–	–	–	–
	V	8	–	–	0.88	2.38	–	–	0.93	2.27
<i>Mathematics</i>	I	8	–	–	0.87	3.85	–	–	–	–
	II	8	–	–	0.93	2.40	–	–	–	–
	III	8	–	–	0.77	3.00	–	–	–	–
	IV	8	–	–	0.88	2.51	–	–	–	–
	V	8	–	–	0.79	2.62	–	–	0.87	2.50
<i>Science *</i>	I	8	–	–	–	–	–	–	–	–
	III	8	–	–	–	–	–	–	–	–
	IV	8	–	–	–	–	–	–	–	–
	V	8	–	–	–	–	–	–	–	–
Subject Area	Level	No. of Items	Hispanic		African American		White		Unknown Ethnicity	
			Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM
<i>English–Language Arts</i>	I	8	0.92	3.59	0.92	3.49	0.90	3.46	0.94	3.03
	II	8	0.82	2.47	0.76	2.62	0.83	2.46	0.87	2.63
	III	8	0.79	2.31	0.76	2.00	0.85	2.28	0.91	2.15
	IV	8	0.84	2.47	0.86	2.52	0.92	2.42	0.86	2.42
	V	8	0.81	2.29	0.86	2.25	0.89	2.15	0.82	2.30
<i>Mathematics</i>	I	8	0.89	4.00	0.89	3.54	0.90	3.49	0.89	3.89
	II	8	0.86	2.70	0.85	2.66	0.88	2.47	0.92	2.45
	III	8	0.86	2.29	0.79	2.37	0.81	2.70	0.83	2.75
	IV	8	0.86	2.70	0.86	2.64	0.91	2.58	0.85	2.79
	V	8	0.74	2.81	0.81	2.73	0.88	2.60	0.82	2.72
<i>Science *</i>	I	8	0.91	4.06	–	–	0.86	3.97	–	–
	III	8	0.66	2.62	–	–	0.84	2.29	–	–
	IV	8	0.84	2.68	–	–	0.93	2.34	0.74	2.60
	V	8	0.78	2.35	0.79	2.44	0.92	2.18	0.72	2.47

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

Table 8.B.6 Reliabilities and SEMs by Disability

Subject Area	Level	No. of Items	Mental Retardation		Hard of Hearing		Deafness		Speech Impairment	
			Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM
<i>English–Language Arts</i>	I	8	0.91	3.38	0.90	3.65	0.90	3.17	0.91	2.47
	II	8	0.80	2.50	0.71	2.56	0.78	2.60	0.75	2.27
	III	8	0.83	2.28	0.85	2.30	0.86	2.19	0.73	2.05
	IV	8	0.88	2.42	0.89	2.45	0.84	2.26	0.80	2.35
	V	8	0.88	2.32	0.88	2.28	0.81	2.16	0.75	2.23
<i>Mathematics</i>	I	8	0.86	4.02	0.86	4.07	0.83	3.92	0.82	3.52
	II	8	0.86	2.57	0.87	2.52	0.85	2.65	0.81	2.40
	III	8	0.86	2.57	0.88	2.19	0.89	2.31	0.73	2.28
	IV	8	0.86	2.58	0.90	2.46	0.79	2.61	0.82	2.61
	V	8	0.86	2.73	0.88	2.50	0.70	2.53	0.67	2.73
<i>Science *</i>	I	8	0.91	3.67	0.87	3.81	0.85	3.77	0.86	3.65
	III	8	0.81	2.46	0.86	2.31	0.88	2.34	0.71	2.43
	IV	8	0.83	2.47	0.92	2.34	0.70	2.47	0.80	2.39
	V	8	0.85	2.31	0.86	2.27	0.78	2.14	0.75	2.27
Subject Area	Level	No. of Items	Visual Impairment		Emotional Disturbance		Orthopedic Impairment		Other Health Impairment	
			Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM
<i>English–Language Arts</i>	I	8	0.90	3.90	0.87	3.29	0.90	3.90	0.92	3.51
	II	8	0.88	2.53	0.92	2.26	0.83	2.47	0.81	2.35
	III	8	0.85	2.37	0.86	1.88	0.85	2.25	0.83	2.24
	IV	8	0.85	2.60	0.83	2.51	0.88	2.48	0.86	2.44
	V	8	0.89	2.33	0.86	2.04	0.90	2.36	0.87	2.19
<i>Mathematics</i>	I	8	0.86	4.00	0.90	3.65	0.88	3.84	0.89	3.88
	II	8	0.85	2.84	0.93	2.11	0.88	2.51	0.88	2.49
	III	8	0.84	2.82	0.80	2.41	0.86	2.56	0.85	2.46
	IV	8	0.87	2.62	0.88	2.56	0.87	2.57	0.86	2.65
	V	8	0.89	2.74	0.85	2.42	0.90	2.68	0.86	2.61
<i>Science *</i>	I	8	0.90	3.86	–	–	0.90	3.70	0.93	3.60
	III	8	0.88	2.72	0.84	2.58	0.84	2.41	0.80	2.44
	IV	8	0.84	2.63	0.79	2.31	0.87	2.38	0.81	2.40
	V	8	0.85	2.24	0.90	2.08	0.88	2.37	0.83	2.31
Subject Area	Level	No. of Items	Specific Learning Disability		Deaf-Blindness		Multiple Disabilities		Autism	
			Reliab.	SEM	Reliab.	SEM	Reliab.	SEM	Reliab.	SEM
<i>English–Language Arts</i>	I	8	0.92	2.63	0.95	3.28	0.91	3.81	0.85	3.91
	II	8	0.70	2.22	–	–	0.81	2.67	0.86	2.54
	III	8	0.67	1.84	–	–	0.84	2.38	0.89	2.34
	IV	8	0.74	2.36	–	–	0.88	2.47	0.89	2.45
	V	8	0.67	2.17	–	–	0.91	2.36	0.91	2.39
<i>Mathematics</i>	I	8	0.91	3.08	0.87	3.65	0.88	3.81	0.80	4.24
	II	8	0.74	2.26	–	–	0.88	2.58	0.88	2.67
	III	8	0.62	2.10	–	–	0.85	2.65	0.87	2.57
	IV	8	0.79	2.54	–	–	0.88	2.48	0.86	2.68
	V	8	0.62	2.55	–	–	0.89	2.72	0.89	2.73
<i>Science *</i>	I	8	0.89	2.86	–	–	0.90	3.66	0.85	4.01
	III	8	0.69	2.10	–	–	0.79	2.60	0.88	2.43
	IV	8	0.75	2.48	–	–	0.84	2.51	0.88	2.44
	V	8	0.74	2.34	–	–	0.88	2.29	0.90	2.27

Subject Area	Level	No. of Items	Traumatic Brain Injury		Unknown Disability	
			Reliab.	SEM	Reliab.	SEM
<i>English–Language Arts</i>	I	8	0.95	3.13	0.92	3.44
	II	8	0.78	2.49	0.83	2.64
	III	8	0.86	2.19	0.89	2.25
	IV	8	0.86	2.71	0.85	2.48
	V	8	0.85	2.25	0.86	2.27
<i>Mathematics</i>	I	8	0.94	3.33	0.90	3.70
	II	8	0.77	2.89	0.89	2.58
	III	8	0.86	2.39	0.89	2.47
	IV	8	0.88	2.77	0.85	2.74
	V	8	0.85	2.61	0.83	2.76
<i>Science *</i>	I	8	0.95	3.36	0.92	3.61
	III	8	0.80	2.39	0.93	2.20
	IV	8	0.83	2.47	0.83	2.52
	V	8	0.84	2.32	0.86	2.25

* 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	Advanced	Proficient	Basic	Below Basic	Far Below Basic	Category Total †
Decision Accuracy	30–40	0.46	0.05	0.00	0.00	0.00	0.51
	19–29	0.05	0.16	0.03	0.00	0.00	0.24
	11–18	0.00	0.03	0.07	0.02	0.00	0.12
All-forms Average *	4–10	0.00	0.00	0.03	0.05	0.01	0.08
	0–3	0.00	0.00	0.00	0.02	0.03	0.05
Estimated Proportion Correctly Classified: Total = 0.76, Proficient & Above = 0.93							
Decision Consistency	30–40	0.45	0.06	0.00	0.00	0.00	0.51
	19–29	0.06	0.12	0.04	0.01	0.00	0.24
	11–18	0.00	0.04	0.05	0.03	0.00	0.12
Alternate Form *	4–10	0.00	0.00	0.02	0.04	0.02	0.08
	0–3	0.00	0.00	0.00	0.02	0.03	0.05
Estimated Proportion Correctly Classified: Total = 0.69, 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	Advanced	Proficient	Basic	Below Basic	Far Below Basic	Category Total †
Decision Accuracy	29–40	0.25	0.04	0.00	0.00	0.00	0.29
	19–28	0.06	0.20	0.05	0.00	0.00	0.32
	11–18	0.00	0.05	0.12	0.03	0.00	0.19
All-forms Average *	6–10	0.00	0.00	0.04	0.05	0.01	0.10
	0–5	0.00	0.00	0.00	0.03	0.06	0.09
Estimated Proportion Correctly Classified: Total = 0.68, Proficient & Above = 0.89							
Decision Consistency	29–40	0.24	0.05	0.00	0.00	0.00	0.29
	19–28	0.08	0.16	0.07	0.01	0.00	0.32
	11–18	0.00	0.06	0.09	0.03	0.01	0.19
Alternate Form *	6–10	0.00	0.01	0.03	0.04	0.02	0.10
	0–5	0.00	0.00	0.01	0.02	0.06	0.09
Estimated Proportion Correctly Classified: Total = 0.58, 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	Advanced	Proficient	Basic	Below Basic	Far Below Basic	Category Total †
Decision Accuracy	29–40	0.30	0.03	0.00	0.00	0.00	0.33
	19–28	0.05	0.17	0.04	0.00	0.00	0.26
	11–18	0.00	0.05	0.10	0.03	0.00	0.19
All-forms Average *	5–10	0.00	0.00	0.04	0.07	0.02	0.13
	0–4	0.00	0.00	0.00	0.03	0.07	0.10
Estimated Proportion Correctly Classified: Total = 0.70, Proficient & Above = 0.90							
Decision Consistency	29–40	0.28	0.05	0.00	0.00	0.00	0.33
	19–28	0.07	0.13	0.05	0.01	0.00	0.26
	11–18	0.01	0.05	0.08	0.04	0.01	0.19
Alternate Form *	5–10	0.00	0.01	0.04	0.05	0.03	0.13
	0–4	0.00	0.00	0.00	0.03	0.07	0.10
Estimated Proportion Correctly Classified: Total = 0.62, Proficient & Above = 0.87							

* Values in table are proportions of the total sample.

† Inconsistencies with category cell entries are due to rounding.

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

	Placement Score	Advanced	Proficient	Basic	Below Basic	Far Below Basic	Category Total †
Decision Accuracy	26–32	0.35	0.06	0.00	0.00	0.00	0.41
	19–25	0.06	0.27	0.04	0.00	0.00	0.37
	13–18	0.00	0.05	0.10	0.01	0.00	0.17
All-forms Average *	4–12	0.00	0.00	0.01	0.03	0.00	0.04
	0–3	0.00	0.00	0.00	0.00	0.00	0.01
Estimated Proportion Correctly Classified: Total = 0.74 , Proficient & Above = 0.90							
Decision Consistency	26–32	0.33	0.08	0.00	0.00	0.00	0.41
	19–25	0.09	0.22	0.06	0.01	0.00	0.37
	13–18	0.00	0.06	0.08	0.03	0.00	0.17
Alternate Form *	4–12	0.00	0.00	0.01	0.03	0.00	0.04
	0–3	0.00	0.00	0.00	0.00	0.00	0.01
Estimated Proportion Correctly Classified: Total =0.65, Proficient & Above = 0.87							

* Values in table are proportions of the total sample.

† Inconsistencies with category cell entries are due to rounding.

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

	Placement Score	Advanced	Proficient	Basic	Below Basic	Far Below Basic	Category Total †
Decision Accuracy	27–32	0.27	0.06	0.00	0.00	0.00	0.33
	20–26	0.03	0.22	0.04	0.00	0.00	0.29
	14–19	0.00	0.05	0.13	0.03	0.00	0.22
All-forms Average *	7–13	0.00	0.00	0.04	0.10	0.00	0.14
	0–6	0.00	0.00	0.00	0.02	0.01	0.03
Estimated Proportion Correctly Classified: Total = 0.72, Proficient & Above = 0.90							
Decision Consistency	27–32	0.26	0.06	0.00	0.00	0.00	0.33
	20–26	0.06	0.18	0.05	0.00	0.00	0.29
	14–19	0.01	0.06	0.10	0.04	0.00	0.22
Alternate Form *	7–13	0.00	0.00	0.04	0.08	0.01	0.14
	0–6	0.00	0.00	0.00	0.02	0.01	0.03
Estimated Proportion Correctly Classified: Total =0.62, Proficient & Above =0.87							

* Values in table are proportions of the total sample.

† Inconsistencies with category cell entries are due to rounding.

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

	Placement Score	Advanced	Proficient	Basic	Below Basic	Far Below Basic	Category Total †
Decision Accuracy	26–32	0.35	0.07	0.00	0.00	0.00	0.42
	18–25	0.05	0.32	0.03	0.00	0.00	0.41
	10–17	0.00	0.03	0.10	0.00	0.00	0.13
All-forms Average *	4–9	0.00	0.00	0.01	0.01	0.00	0.03
	0–3	0.00	0.00	0.00	0.00	0.00	0.01
Estimated Proportion Correctly Classified: Total = 0.79, Proficient & Above = 0.93							
Decision Consistency	26–32	0.34	0.08	0.00	0.00	0.00	0.42
	18–25	0.08	0.27	0.06	0.00	0.00	0.41
	10–17	0.00	0.04	0.09	0.01	0.00	0.13
Alternate Form *	4–9	0.00	0.00	0.01	0.01	0.00	0.03
	0–3	0.00	0.00	0.00	0.00	0.00	0.01
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.13 Decision Accuracy and Decision Consistency: Level III, Mathematics

	Placement Score	Advanced	Proficient	Basic	Below Basic	Far Below Basic	Category Total †
Decision Accuracy	27–32	0.23	0.07	0.01	0.00	0.00	0.31
	20–26	0.04	0.26	0.04	0.00	0.00	0.34
	12–19	0.00	0.06	0.18	0.02	0.00	0.26
All-forms Average *	5–11	0.00	0.00	0.03	0.06	0.00	0.09
	0–4	0.00	0.00	0.00	0.01	0.00	0.01
Estimated Proportion Correctly Classified: Total = 0.73, Proficient & Above = 0.89							
Decision Consistency	27–32	0.22	0.08	0.01	0.00	0.00	0.31
	20–26	0.07	0.21	0.06	0.00	0.00	0.34
	12–19	0.00	0.07	0.15	0.03	0.00	0.26
Alternate Form *	5–11	0.00	0.00	0.03	0.05	0.01	0.09
	0–4	0.00	0.00	0.00	0.01	0.00	0.01
Estimated Proportion Correctly Classified: Total = 0.63, 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.14 Decision Accuracy and Decision Consistency: Level III, Science

	Placement Score	Advanced	Proficient	Basic	Below Basic	Far Below Basic	Category Total †
Decision Accuracy	28–32	0.13	0.06	0.00	0.00	0.00	0.19
	19–27	0.03	0.42	0.04	0.00	0.00	0.50
	11–18	0.00	0.06	0.18	0.01	0.00	0.26
All-forms Average *	4–10	0.00	0.00	0.02	0.03	0.00	0.05
	0–3	0.00	0.00	0.00	0.01	0.00	0.01
Estimated Proportion Correctly Classified: Total = 0.76, Proficient & Above = 0.90							
Decision Consistency	28–32	0.12	0.07	0.00	0.00	0.00	0.19
	19–27	0.06	0.37	0.07	0.00	0.00	0.50
	11–18	0.00	0.07	0.15	0.03	0.00	0.26
Alternate Form *	4–10	0.00	0.00	0.02	0.03	0.00	0.05
	0–3	0.00	0.00	0.00	0.00	0.00	0.01
Estimated Proportion Correctly Classified: Total = 0.67, 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	Advanced	Proficient	Basic	Below Basic	Far Below Basic	Category Total †
Decision Accuracy	24–32	0.31	0.06	0.00	0.00	0.00	0.37
	15–23	0.05	0.32	0.03	0.00	0.00	0.40
	10–14	0.00	0.05	0.08	0.02	0.00	0.15
All-forms Average *	4–9	0.00	0.00	0.02	0.05	0.00	0.07
	0–3	0.00	0.00	0.00	0.01	0.01	0.02
Estimated Proportion Correctly Classified: Total = 0.76, Proficient & Above = 0.92							
Decision Consistency	24–32	0.29	0.08	0.00	0.00	0.00	0.37
	15–23	0.07	0.27	0.05	0.01	0.00	0.40
	10–14	0.00	0.05	0.06	0.03	0.00	0.15
Alternate Form *	4–9	0.00	0.00	0.02	0.04	0.00	0.07
	0–3	0.00	0.00	0.00	0.01	0.01	0.02
Estimated Proportion Correctly Classified: Total = 0.67, 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.16 Decision Accuracy and Decision Consistency: Level IV, Mathematics

	Placement Score	Advanced	Proficient	Basic	Below Basic	Far Below Basic	Category Total †
Decision Accuracy	24–32	0.26	0.05	0.00	0.00	0.00	0.31
	17–23	0.04	0.20	0.05	0.00	0.00	0.29
	12–16	0.00	0.05	0.12	0.04	0.00	0.21
All-forms Average *	7–11	0.00	0.00	0.04	0.11	0.00	0.15
	0–6	0.00	0.00	0.00	0.03	0.00	0.03
Estimated Proportion Correctly Classified: Total = 0.69, Proficient & Above = 0.90							
Decision Consistency	24–32	0.25	0.06	0.00	0.00	0.00	0.31
	17–23	0.06	0.16	0.06	0.01	0.00	0.29
	12–16	0.01	0.06	0.09	0.05	0.01	0.21
Alternate Form *	7–11	0.00	0.01	0.05	0.08	0.02	0.15
	0–6	0.00	0.00	0.00	0.02	0.01	0.03
Estimated Proportion Correctly Classified: Total = 0.59, 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.17 Decision Accuracy and Decision Consistency: Level IV, Science

	Placement Score	Advanced	Proficient	Basic	Below Basic	Far Below Basic	Category Total †
Decision Accuracy	27–32	0.10	0.05	0.00	0.00	0.00	0.15
	19–26	0.02	0.35	0.05	0.00	0.00	0.43
	11–18	0.00	0.07	0.25	0.02	0.00	0.34
All-forms Average *	4–10	0.00	0.00	0.02	0.05	0.00	0.07
	0–3	0.00	0.00	0.00	0.01	0.00	0.01
Estimated Proportion Correctly Classified: Total = 0.74, Proficient & Above = 0.88							
Decision Consistency	27–32	0.10	0.06	0.00	0.00	0.00	0.15
	19–26	0.05	0.29	0.08	0.00	0.00	0.43
	11–18	0.01	0.09	0.21	0.04	0.00	0.34
Alternate Form *	4–10	0.00	0.00	0.02	0.04	0.00	0.07
	0–3	0.00	0.00	0.00	0.01	0.00	0.01
Estimated Proportion Correctly Classified: Total = 0.64, 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.18 Decision Accuracy and Decision Consistency: Level V, ELA

	Placement Score	Advanced	Proficient	Basic	Below Basic	Far Below Basic	Category Total †
Decision Accuracy	25–32	0.35	0.06	0.00	0.00	0.00	0.42
	16–24	0.04	0.31	0.03	0.00	0.00	0.38
	9–15	0.00	0.04	0.11	0.01	0.00	0.15
All-forms Average *	4–8	0.00	0.00	0.01	0.02	0.00	0.03
	0–3	0.00	0.00	0.00	0.01	0.00	0.01
Estimated Proportion Correctly Classified: Total = 0.79, Proficient & Above = 0.93							
Decision Consistency	25–32	0.34	0.07	0.00	0.00	0.00	0.42
	16–24	0.07	0.26	0.05	0.00	0.00	0.38
	9–15	0.00	0.05	0.09	0.02	0.00	0.15
Alternate Form *	4–8	0.00	0.00	0.01	0.02	0.00	0.03
	0–3	0.00	0.00	0.00	0.01	0.00	0.01
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	Advanced	Proficient	Basic	Below Basic	Far Below Basic	Category Total †
Decision Accuracy	27–32	0.28	0.06	0.00	0.00	0.00	0.34
	20–26	0.05	0.23	0.05	0.01	0.00	0.33
	12–19	0.00	0.04	0.15	0.01	0.00	0.20
All-forms Average *	6–11	0.00	0.01	0.04	0.06	0.01	0.11
	0–5	0.00	0.00	0.00	0.01	0.01	0.02
Estimated Proportion Correctly Classified: Total = 0.72, Proficient & Above = 0.90							
Decision Consistency	27–32	0.26	0.07	0.00	0.00	0.00	0.34
	20–26	0.08	0.18	0.07	0.01	0.00	0.33
	12–19	0.00	0.05	0.12	0.03	0.00	0.20
Alternate Form *	6–11	0.00	0.01	0.04	0.05	0.02	0.11
	0–5	0.00	0.00	0.00	0.01	0.01	0.02
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	Advanced	Proficient	Basic	Below Basic	Far Below Basic	Category Total †
Decision Accuracy	26–32	0.12	0.05	0.00	0.00	0.00	0.17
	19–25	0.06	0.33	0.05	0.00	0.00	0.44
	11–18	0.00	0.06	0.22	0.02	0.00	0.29
All-forms Average *	4–10	0.00	0.00	0.02	0.06	0.00	0.08
	0–3	0.00	0.00	0.00	0.01	0.01	0.02
Estimated Proportion Correctly Classified: Total = 0.73, Proficient & Above = 0.89							
Decision Consistency	26–32	0.11	0.07	0.00	0.00	0.00	0.17
	19–25	0.09	0.28	0.07	0.00	0.00	0.44
	11–18	0.00	0.07	0.19	0.03	0.00	0.29
Alternate Form *	4–10	0.00	0.00	0.02	0.05	0.00	0.08
	0–3	0.00	0.00	0.00	0.01	0.01	0.02
Estimated Proportion Correctly Classified: Total = 0.63, Proficient & Above = 0.85							

* Values in table are proportions of the total sample.

† Inconsistencies with category cell entries are due to rounding.

Appendix 8.C—Intercorrelations Between Content Areas

Table 8.C.1 Raw Score Correlations by Gender: Level I

	Male			Female			Unknown		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	–	0.83	0.85	–	0.84	0.86	–	0.77	0.79
Mathematics		–	0.85		–	0.87		–	0.75
Science			–			–			–
N	7783	7759	2041	4722	4699	1247	26	26	8
Raw Score Mean	27.07	21.78	22.03	26.50	21.17	21.52	26.38	17.62	13.38
Raw Score SD	11.86	11.04	12.13	12.27	11.36	12.36	11.18	10.15	11.46

Table 8.C.2 Raw Score Correlations by Gender: Level II

	Male			Female			Unknown		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	–	0.81	N/A	–	0.78	N/A	–	0.74	N/A
Mathematics		–	N/A		–	N/A		–	N/A
Science			–			–			–
N	4464	4451	N/A	2090	2085	N/A	33	33	N/A
Raw Score Mean	23.33	21.86	N/A	23.03	21.01	N/A	22.94	20.00	N/A
Raw Score SD	6.34	7.58	N/A	5.76	7.13	N/A	5.50	7.04	N/A

Table 8.C.3 Raw Score Correlations by Gender: Level III

	Male			Female			Unknown		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	–	0.83	0.82	–	0.81	0.78	–	0.88	0.96
Mathematics		–	0.80		–	0.74		–	0.87
Science			–			–			–
N	4397	4390	2185	2200	2195	1076	17	17	6
Raw Score Mean	23.26	21.82	21.57	23.05	20.97	21.07	22.41	19.65	19.83
Raw Score SD	6.20	6.99	6.62	5.85	6.83	5.75	6.93	7.68	7.99

Table 8.C.4 Raw Score Correlations by Gender: Level IV

	Male			Female			Unknown		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	–	0.81	0.77	–	0.80	0.76	–	0.76	0.87
Mathematics		–	0.79		–	0.79		–	0.95
Science			–			–			–
N	6337	6326	2043	3483	3472	1138	33	33	9
Raw Score Mean	20.10	19.37	19.83	19.94	18.18	19.29	21.70	21.09	18.89
Raw Score SD	7.09	7.59	6.60	7.11	7.13	6.04	7.38	7.37	8.99

Table 8.C.5 Raw Score Correlations by Gender: Level V

	Male			Female			Unknown		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	–	0.81	0.74	–	0.83	0.74	–	0.79	0.84
Mathematics		–	0.72		–	0.74		–	0.85
Science			–			–			–
N	6691	6669	2172	3769	3758	1211	57	58	13
Raw Score Mean	21.60	22.23	19.82	21.78	21.29	19.09	23.49	23.88	24.62
Raw Score SD	7.00	7.63	6.37	7.04	7.58	6.28	6.65	7.10	4.94

Table 8.C.6 Raw Score Correlations by Ethnicity: Level I

	American Indian			Asian			Pacific Islander			Filipino		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
Mathematics Science	—	0.76	0.80	—	0.81	0.85	—	0.74	0.81	—	0.81	0.82
N	75	75	23	994	992	272	74	72	19	432	431	118
Raw Score Mean	28.97	22.63	27.35	26.23	20.82	20.22	25.97	19.53	21.63	26.94	20.99	20.21
Raw Score SD	12.12	10.4	9.83	11.42	10.54	11.74	11.03	10.11	13.07	10.83	10.17	11.28
	Hispanic			African American			White			Unknown Ethnicity		
ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	
Mathematics Science	—	0.85	0.86	—	0.86	0.86	—	0.80	0.85	—	0.81	0.88
N	6478	6455	1668	1192	1185	310	3133	3122	852	153	152	34
Raw Score Mean	27.03	21.82	21.99	26.25	21.2	21.66	26.83	21.37	22.1	27.59	22.61	22.09
Raw Score SD	12.3	11.41	12.42	12.67	11.78	12.5	11.54	10.74	12	11.75	10.74	12.38

Table 8.C.7 Raw Score Correlations by Ethnicity: Level II

	American Indian			Asian			Pacific Islander			Filipino		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
Mathematics Science	—	0.79	N/A	—	0.8	N/A	—	0.78	N/A	—	0.82	N/A
N	58	58	N/A	480	479	N/A	37	37	N/A	203	203	N/A
Raw Score Mean	24.14	22.84	N/A	21.67	20.47	N/A	21.81	20.68	N/A	21.72	20.39	N/A
Raw Score SD	6.03	7.74	N/A	6.55	7.95	N/A	5.27	6.67	N/A	6.15	7.66	N/A
	Hispanic			African American			White			Unknown Ethnicity		
ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	
Mathematics Science	—	0.81	N/A	—	0.80	N/A	—	0.78	N/A	—	0.85	N/A
N	3349	3339	N/A	681	679	N/A	1667	1663	N/A	112	111	N/A
Raw Score Mean	23.53	21.78	N/A	23.22	21.68	N/A	23.4	21.74	N/A	21.63	19.36	N/A
Raw Score SD	6.12	7.39	N/A	6.08	7.41	N/A	6.02	7.27	N/A	6.86	8.53	N/A

Table 8.C.8 Raw Score Correlations by Ethnicity: Level III

	American Indian			Asian			Pacific Islander			Filipino		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	–	0.88	0.84	–	0.84	0.85	–	0.79	0.67	–	0.86	0.86
Mathematics	–	–	0.88	–	–	0.79	–	–	0.71	–	–	0.78
Science	–	–	–	–	–	–	–	–	–	–	–	–
N	57	57	40	468	468	224	29	29	15	226	225	114
Raw Score Mean	23.56	22.23	22	21.7	20.61	19.36	22.34	19.97	20.47	21.33	20.39	19.36
Raw Score SD	6.31	7.1	6.28	6.47	6.77	6.59	5.55	7.15	5.84	6.85	7.23	6.62
Unknown Ethnicity												
	Hispanic			African American			White			Unknown Ethnicity		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	–	0.82	0.79	–	0.8	0.76	–	0.84	0.82	–	0.83	0.86
Mathematics	–	–	0.77	–	–	0.76	–	–	0.8	–	–	0.86
Science	–	–	–	–	–	–	–	–	–	–	–	–
N	3307	3306	1607	728	725	386	1722	1715	848	77	77	33
Raw Score Mean	23.54	21.85	21.8	23.91	22.08	21.77	22.92	21.12	21.29	21.99	20.81	20.82
Raw Score SD	5.82	6.92	6.02	5.58	6.69	6.07	6.38	7.04	6.78	7.7	7.43	7.97

Table 8.C.9 Raw Score Correlations by Ethnicity: Level IV

	American Indian			Asian			Pacific Islander			Filipino		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	–	0.79	0.85	–	0.82	0.79	–	0.77	0.55	–	0.85	0.85
Mathematics	–	–	0.90	–	–	0.78	–	–	0.44	–	–	0.85
Science	–	–	–	–	–	–	–	–	–	–	–	–
N	88	88	33	687	683	241	49	49	18	271	269	82
Raw Score Mean	21.23	20.25	22.33	18.86	18.68	18.78	21.16	19.33	21.5	18.45	17.71	18.4
Raw Score SD	7.31	7.11	5.88	7.19	7.66	6.31	7.05	6.68	5.17	7.13	7.5	7.17
Unknown Ethnicity												
	Hispanic			African American			White			Unknown Ethnicity		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	–	0.81	0.78	–	0.8	0.75	–	0.8	0.76	–	0.86	0.79
Mathematics	–	–	0.81	–	–	0.79	–	–	0.76	–	–	0.78
Science	–	–	–	–	–	–	–	–	–	–	–	–
N	4802	4793	1476	1126	1127	386	2684	2678	908	146	144	46
Raw Score Mean	19.8	19.06	19.66	20.95	19.35	19.57	20.46	18.71	19.86	21.14	19.63	19.17
Raw Score SD	7.05	7.46	6.41	7.04	7.54	6.48	7.07	7.34	6.35	7.26	7.88	6.69

Table 8.C.10 Raw Score Correlations by Ethnicity: Level V

	American Indian			Asian			Pacific Islander			Filipino		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	—	0.78	0.73	—	0.83	0.72	—	0.76	0.92	—	0.81	0.80
Mathematics	—	—	0.6	—	—	0.69	—	—	0.75	—	—	0.76
Science	—	—	—	—	—	—	—	—	—	—	—	—
N	117	117	37	657	650	206	56	56	15	287	287	74
Raw Score Mean	22.5	22.16	19.59	19.59	20.45	17.24	21.46	23.05	19.27	20.15	21.32	18.68
Raw Score SD	6.91	7.29	6.09	7.77	8.17	6.64	7.46	6.6	7.54	7.24	7.81	6.6
	Hispanic			African American			White			Unknown Ethnicity		
ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	
ELA	—	0.81	0.73	—	0.82	0.74	—	0.82	0.74	—	0.81	0.73
Mathematics	—	—	0.72	—	—	0.72	—	—	0.74	—	—	0.75
Science	—	—	—	—	—	—	—	—	—	—	—	—
N	4833	4820	1529	1260	1255	400	3125	3120	1075	182	180	60
Raw Score Mean	21.46	21.9	19.68	22.47	22.14	19.91	22.21	22.12	19.81	21.95	22.37	19.75
Raw Score SD	6.83	7.56	6.21	6.75	7.47	6.01	7.07	7.67	6.51	7.2	7.22	6.22

Table 8.C.11 Raw Score 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	—	0.72	0.75	—	0.83	0.87	—	0.81	0.95	—	0.84	0.77
Mathematics	—	—	0.61	—	—	0.87	—	—	0.90	—	—	0.78
Science	—	—	—	—	—	—	—	—	—	—	—	—
N	35	35	12	450	447	123	36	35	11	149	149	39
Raw Score Mean	29.26	20.89	27.25	27.04	21.69	20.73	27.53	21.00	24.18	27.21	21.21	21.82
Raw Score SD	13.05	10.43	9.85	11.29	10.62	11.72	11.53	11.27	14.23	11.55	10.37	10.77
	Hispanic			African American			White			Unknown Ethnicity		
ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	
ELA	—	0.84	0.87	—	0.85	0.84	—	0.81	0.88	—	0.79	0.91
Mathematics	—	—	0.87	—	—	0.87	—	—	0.87	—	—	0.85
Science	—	—	—	—	—	—	—	—	—	—	—	—
N	4778	4765	1204	745	741	197	980	976	258	61	60	11
Raw Score Mean	27.87	22.53	22.97	27.42	22.36	22.97	27.32	22.00	22.40	28.11	22.82	22.91
Raw Score SD	11.98	11.25	12.18	12.31	11.61	12.24	11.66	10.85	12.22	11.76	10.59	12.41

Table 8.C.12 Raw Score 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	–	0.77	N/A	–	0.81	N/A	–	0.84	N/A	–	0.85	N/A
Mathematics Science	–	–	N/A	–	–	N/A	–	–	N/A	–	–	N/A
N	36	36	N/A	196	196	N/A	20	20	N/A	54	54	N/A
Raw Score Mean	25.67	23.67	N/A	21.77	20.41	N/A	22.55	20.50	N/A	22.28	20.70	N/A
Raw Score SD	5.20	7.92	N/A	6.91	8.03	N/A	5.34	6.42	N/A	6.49	8.45	N/A
	Hispanic			African American			White			Unknown Ethnicity		
ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	
ELA	–	0.81	N/A	–	0.79	N/A	–	0.81	N/A	–	0.78	N/A
Mathematics Science	–	–	N/A	–	–	N/A	–	–	N/A	–	–	N/A
N	2720	2712	N/A	479	479	N/A	665	665	N/A	38	37	N/A
Raw Score Mean	23.67	21.89	N/A	23.59	22.07	N/A	24.40	22.67	N/A	24.05	21.68	N/A
Raw Score SD	6.04	7.36	N/A	5.97	7.24	N/A	5.71	7.22	N/A	5.20	6.92	N/A

Table 8.C.13 Raw Score 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	–	0.91	0.85	–	0.85	0.85	–	0.79	0.66	–	0.86	0.87
Mathematics Science	–	–	0.93	–	–	0.81	–	–	0.70	–	–	0.76
N	42	42	30	204	203	93	19	19	12	77	77	44
Raw Score Mean	23.74	22.40	22.00	22.10	20.98	19.91	23.21	21.16	22.17	22.82	21.38	21.16
Raw Score SD	6.51	7.36	6.89	6.41	6.72	6.95	4.57	6.69	5.04	6.07	6.93	6.25
	Hispanic			African American			White			Unknown Ethnicity		
ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	
ELA	–	0.81	0.79	–	0.80	0.76	–	0.83	0.78	–	0.86	0.89
Mathematics Science	–	–	0.77	–	–	0.76	–	–	0.78	–	–	0.83
N	2720	2719	1323	531	530	280	652	650	331	26	26	14
Raw Score Mean	23.71	22.09	21.97	24.17	22.29	22.18	23.88	22.18	22.60	22.19	21.27	23.21
Raw Score SD	5.68	6.79	5.92	5.65	6.77	5.88	5.82	6.74	6.04	8.06	8.19	7.38

Table 8.C.14 Raw Score 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	-	0.77	0.81	-	0.79	0.76	-	0.75	0.48	-	0.83	0.79
Mathematics		-	0.85		-	0.76		-	0.24		-	0.76
Science			-			-			-			-
N	55	55	18	331	327	113	29	29	13	81	80	24
Raw Score Mean	22.20	21.04	23.17	18.85	18.90	19.33	22.17	20.41	22.15	18.77	17.96	19.42
Raw Score SD	6.59	7.11	6.12	7.06	7.49	6.19	7.13	6.71	5.38	7.43	7.01	6.25
	Hispanic			African American			White			Unknown Ethnicity		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	-	0.80	0.78	-	0.78	0.72	-	0.79	0.76	-	0.89	0.86
Mathematics		-	0.81		-	0.77		-	0.77		-	0.84
Science			-			-			-			-
N	3933	3927	1211	793	795	270	1009	1004	337	41	41	16
Raw Score Mean	19.83	19.18	19.79	21.29	19.72	20.17	21.33	19.74	21.04	20.41	19.02	18.69
Raw Score SD	6.97	7.44	6.31	6.70	7.36	6.11	6.85	7.29	5.94	8.06	8.80	7.39

Table 8.C.15 Raw Score 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	-	0.74	0.64	-	0.83	0.74	-	0.70	0.92	-	0.84	0.86
Mathematics		-	0.54		-	0.72		-	0.66		-	0.82
Science			-			-			-			-
N	77	77	27	318	317	102	34	34	10	96	96	24
Raw Score Mean	22.84	22.38	19.26	19.43	20.22	17.44	21.47	23.47	20.10	19.46	20.52	18.38
Raw Score SD	6.25	6.54	5.80	7.83	8.26	7.07	7.65	6.07	7.31	7.70	8.50	7.46
	Hispanic			African American			White			Unknown Ethnicity		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	-	0.81	0.72	-	0.82	0.74	-	0.81	0.73	-	0.79	0.52
Mathematics		-	0.72		-	0.71		-	0.75		-	0.54
Science			-			-			-			-
N	3768	3756	1208	819	815	247	1148	1143	392	48	47	17
Raw Score Mean	21.46	21.95	19.79	22.52	22.40	20.45	22.66	22.90	20.24	21.27	22.49	19.65
Raw Score SD	6.73	7.55	6.07	6.58	7.40	5.56	6.65	7.37	6.62	7.23	6.27	4.06

Table 8.C.16 Raw Score 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	–	0.83	0.90	–	0.79	0.83	–	0.65	0.52	–	0.79	0.85
Mathematics	–	–	0.76	–	–	0.84	–	–	0.55	–	–	0.87
Science	–	–	–	–	–	–	–	–	–	–	–	–
N	38	38	10	528	529	146	36	36	8	274	273	79
Raw Score Mean	28.63	23.97	27.90	25.43	20.03	19.82	24.61	17.94	18.13	27.08	21.12	19.42
Raw Score SD	11.49	10.28	10.72	11.56	10.41	11.90	10.63	8.85	11.21	10.39	10.07	11.50
	Hispanic			African American			White			Unknown Ethnicity		
ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	
ELA	–	0.87	0.83	–	0.87	0.87	–	0.80	0.85	–	0.81	0.84
Mathematics	–	–	0.86	–	–	0.88	–	–	0.82	–	–	0.86
Science	–	–	–	–	–	–	–	–	–	–	–	–
N	1622	1612	446	425	422	108	2078	2071	573	58	58	15
Raw Score Mean	24.61	19.71	19.26	24.12	19.14	19.76	26.56	21.04	22.03	26.07	21.03	20.73
Raw Score SD	12.85	11.60	12.65	13.04	11.87	12.73	11.49	10.68	11.96	11.11	10.03	11.93

Table 8.C.17 Raw Score Correlations by Ethnicity Not Economically Disadvantaged: Level II

	American Indian			Asian			Pacific Islander			Filipino		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	–	0.85	N/A	–	0.80	N/A	–	0.77	N/A	–	0.81	N/A
Mathematics	–	–	N/A	–	–	N/A	–	–	N/A	–	–	N/A
Science	–	–	–	–	–	–	–	–	–	–	–	–
N	20	20	N/A	264	263	N/A	15	15	N/A	142	142	N/A
Raw Score Mean	21.00	20.65	N/A	21.67	20.72	N/A	21.40	20.80	N/A	21.56	20.21	N/A
Raw Score SD	6.49	7.24	N/A	6.26	7.82	N/A	5.33	7.64	N/A	6.11	7.39	N/A
	Hispanic			African American			White			Unknown Ethnicity		
ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	
ELA	–	0.81	N/A	–	0.81	N/A	–	0.76	N/A	–	0.88	N/A
Mathematics	–	–	N/A	–	–	N/A	–	–	N/A	–	–	N/A
Science	–	–	–	–	–	–	–	–	–	–	–	–
N	537	535	N/A	175	173	N/A	932	928	N/A	44	44	N/A
Raw Score Mean	22.85	21.11	N/A	22.29	20.83	N/A	22.73	21.04	N/A	20.16	17.86	N/A
Raw Score SD	6.51	7.59	N/A	6.38	7.88	N/A	6.14	7.25	N/A	7.35	9.31	N/A

Table 8.C.18 Raw Score Correlations by Ethnicity Not Economically Disadvantaged: Level III

	American Indian			Asian			Pacific Islander			Filipino		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	–	0.78	0.87	–	0.82	0.86	–	0.79	1	–	0.86	0.85
Mathematics		–	0.62		–	0.79		–	0.77		–	0.81
Science			–			–			–			–
N	15	15	10	253	254	127	10	10	3	146	145	69
Raw Score Mean	23.07	21.73	22.00	21.36	20.25	18.99	20.70	17.70	13.67	20.55	19.85	18.30
Raw Score SD	5.91	6.56	4.19	6.48	6.82	6.31	7.04	7.80	3.51	7.15	7.38	6.65
Unknown Ethnicity												
Hispanic			African American			White			Unknown Ethnicity			
ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	Science
ELA	–	0.85	–	0.80	0.77	–	0.85	0.84	–	0.73	0.79	0.79
Mathematics		–		–	0.78		–	0.80		–	0.93	0.93
Science					–			–			–	–
N	532	532	254	179	98	1023	1018	499	29	29	13	13
Raw Score Mean	22.56	20.44	20.88	23.06	21.42	20.96	22.25	20.43	20.32	20.62	19.03	18.15
Raw Score SD	6.50	7.46	6.59	5.39	6.56	6.47	6.65	7.19	7.11	7.66	7.09	6.47

Table 8.C.19 Raw Score Correlations by Ethnicity Not Economically Disadvantaged : Level IV

	American Indian			Asian			Pacific Islander			Filipino		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	–	0.81	0.90	–	0.84	0.83	–	0.77	0.88	–	0.86	0.87
Mathematics		–	0.96		–	0.81		–	0.91		–	0.87
Science			–			–			–			–
N	31	31	13	337	337	123	18	18	5	185	184	57
Raw Score Mean	19.16	18.55	20.62	18.83	18.44	18.39	19.33	17.28	19.80	18.30	17.65	18.09
Raw Score SD	8.28	7.02	5.69	7.31	7.84	6.50	7.12	6.63	4.66	7.09	7.77	7.55
Unknown Ethnicity												
Hispanic			African American			White			Unknown Ethnicity			
ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	Science
ELA	–	0.83	0.78	–	0.84	0.80	–	0.80	0.74	–	0.87	0.84
Mathematics		–	0.81		–	0.83		–	0.74		–	0.85
Science			–			–			–			–
N	785	782	249	296	295	109	1597	1596	549	40	39	11
Raw Score Mean	19.36	18.33	19.15	19.77	18.14	18.11	19.87	17.99	19.15	19.40	17.56	16.45
Raw Score SD	7.44	7.58	6.85	7.79	7.93	7.22	7.08	7.21	6.37	7.26	7.60	7.47

Table 8.C.20 Raw Score Correlations by Ethnicity Not Economically Disadvantaged : Level V

	American Indian			Asian			Pacific Islander			Filipino		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	-	0.82	0.93	-	0.83	0.72	-	0.88	0.91	-	0.79	0.76
Mathematics		-	0.84		-	0.66		-	0.87		-	0.70
Science			-			-			-			-
N	35	35	8	324	319	102	21	21	5	178	178	47
Raw Score Mean	21.26	21.29	20.75	19.70	20.59	17.00	21.29	22.29	17.60	20.26	21.44	18.96
Raw Score SD	8.32	8.50	8.00	7.77	8.18	6.25	7.48	7.62	8.56	6.85	7.40	6.13
	Hispanic			African American			White			Unknown Ethnicity		
ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	
ELA	-	0.82	0.79	-	0.82	0.74	-	0.83	0.74	-	0.87	0.83
Mathematics		-	0.77		-	0.72		-	0.72		-	0.84
Science			-			-			-			-
N	929	929	282	392	391	128	1853	1853	653	62	62	21
Raw Score Mean	21.14	21.40	19.17	22.06	21.29	18.24	21.78	21.52	19.48	21.06	20.71	16.86
Raw Score SD	7.38	7.81	6.95	7.12	7.67	6.54	7.31	7.80	6.37	8.76	8.41	7.79

Table 8.C.21 Raw Score 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	-	1	N/A	-	0.84	0.90	-	N/A	N/A	-	0.77	N/A
Mathematics		-	N/A		-	0.72		-	N/A		-	N/A
Science			-			-			-			-
N	2	2	1	16	16	3	2	1	0	9	9	0
Raw Score Mean	30.50	27.50	23.00	29.94	22.88	19.00	22.50	25.00	N/A	18.11	13.33	N/A
Raw Score SD	13.44	13.44	N/A	8.00	10.65	3.61	9.19	N/A	N/A	8.96	7.37	N/A
	Hispanic			African American			White			Unknown Ethnicity		
ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	
ELA	-	0.89	0.84	-	0.81	0.91	-	0.76	0.67	-	0.83	0.94
Mathematics		-	0.90		-	0.67		-	0.82		-	1
Science			-			-			-			-
N	78	78	18	22	22	5	75	75	21	34	34	8
Raw Score Mean	26.03	21.73	24.22	27.59	21.68	11.20	28.00	22.52	20.33	29.26	24.91	23.50
Raw Score SD	12.74	11.90	13.47	12.61	10.90	8.53	11.00	10.95	10.71	12.82	11.99	14.50

Table 8.C.22 Raw Score 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	-	1	N/A	-	0.81	N/A	-	1	N/A	-	0.96	N/A
Mathematics Science		-	N/A		-	N/A		-	N/A		-	N/A
N	2	2	N/A	20	20	N/A	2	2	N/A	7	7	N/A
Raw Score Mean	28.00	30.00	N/A	20.75	17.85	N/A	17.50	21.50	N/A	20.86	21.71	N/A
Raw Score SD	4.24	1.41	N/A	6.95	8.85	N/A	3.54	0.71	N/A	4.41	7.65	N/A
Unknown Ethnicity												
	Hispanic			African American			White			Unknown Ethnicity		
ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	Science
ELA	-	0.77	N/A	-	0.86	N/A	-	0.83	N/A	-	0.84	N/A
Mathematics Science		-	N/A		-	N/A		-	N/A		-	N/A
N	92	92	N/A	27	27	N/A	70	70	N/A	30	30	N/A
Raw Score Mean	23.33	22.23	N/A	22.63	20.26	N/A	22.89	22.16	N/A	20.73	18.70	N/A
Raw Score SD	5.78	7.12	N/A	5.37	6.95	N/A	5.96	7.13	N/A	7.32	8.81	N/A

Table 8.C.23 Raw Score 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	-	N/A	N/A	-	0.87	0.95	-	N/A	N/A	-	0.90	N/A
Mathematics Science		-	N/A		-	0.94		-	N/A		-	N/A
N	0	0	0	11	11	4	0	0	0	3	3	1
Raw Score Mean	0	0	0	22.00	22.09	18.00	0	0	0	21.00	21.00	13.00
Raw Score SD	N/A	N/A	N/A	7.38	6.30	7.30	N/A	N/A	N/A	6.00	6.08	N/A
Unknown Ethnicity												
	Hispanic			African American			White			Unknown Ethnicity		
ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	Science
ELA	-	0.83	0.58	-	0.79	0.90	-	0.77	0.84	-	0.88	0.95
Mathematics Science		-	0.77		-	0.64		-	0.84		-	0.90
N	55	55	30	16	16	8	47	47	18	22	22	6
Raw Score Mean	24.38	23.67	22.10	25.13	22.56	17.63	24.32	21.47	24.11	23.55	22.59	21.00
Raw Score SD	5.04	6.07	4.47	4.10	5.18	5.42	5.94	6.18	5.67	7.33	6.68	11.40

Table 8.C.24 Raw Score 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	-	1	1	-	0.84	0.75	-	1	1	-	0.78	N/A
Mathematics		-	1		-	0.86		-	1		-	N/A
Science			-			-			-			-
N	2	2	2	19	19	5	2	2	2	5	5	1
Raw Score Mean	26.50	25.00	26.00	19.53	19.21	16.00	23.00	22.00	22.00	18.60	15.60	12.00
Raw Score SD	2.12	4.24	1.41	7.65	7.31	2.55	1.41	2.83	N/A	3.51	5.03	N/A
	Hispanic			African American			White			Unknown Ethnicity		
ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	
ELA	-	0.75	0.81	-	0.82	0.90	-	0.85	0.85	-	0.79	0.69
Mathematics		-	0.87		-	0.81		-	0.90		-	0.59
Science			-			-			-			-
N	84	84	16	37	37	7	78	78	22	65	64	19
Raw Score Mean	22.40	20.56	18.00	23.05	21.27	19.14	21.29	20.06	19.23	22.66	21.27	21.16
Raw Score SD	6.13	7.30	6.67	6.77	7.05	4.95	8.42	8.80	9.16	6.50	7.16	5.15

Table 8.C.25 Raw Score 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	-	0.98	1	-	0.64	1	-	N/A	N/A	-	0.81	0.99
Mathematics		-	1		-	1		-	N/A		-	0.99
Science			-			-			-			-
N	5	5	2	15	14	2	1	1	0	13	13	3
Raw Score Mean	25.80	25.00	19.50	20.73	22.79	19.50	25.00	25.00	0	23.69	25.62	16.67
Raw Score SD	5.02	9.80	0.71	7.01	5.65	3.54	N/A	N/A	N/A	8.36	6.98	8.74
	Hispanic			African American			White			Unknown Ethnicity		
ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	
ELA	-	0.67	0.51	-	0.73	0.54	-	0.80	0.86	-	0.73	0.48
Mathematics		-	0.52		-	0.70		-	0.85		-	0.51
Science			-			-			-			-
N	136	135	39	49	49	25	124	124	30	72	71	22
Raw Score Mean	23.67	24.08	20.00	25.00	24.51	23.16	24.58	23.79	21.23	23.17	23.75	22.59
Raw Score SD	5.21	5.56	4.97	5.91	6.21	5.28	6.45	7.51	7.70	5.40	6.45	4.64

Table 8.C.26 Raw Score Correlations by Economic Status: Level I

	Disadvantaged			Not Disadvantaged			Unknown Status		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	–	0.83	0.86	–	0.83	0.84	–	0.83	0.82
Mathematics		–	0.86		–	0.85		–	0.88
Science			–			–			–
N	7234	7208	1855	5059	5039	1385	238	237	56
Raw Score Mean	27.69	22.35	22.75	25.64	20.35	20.58	27.23	22.25	21.20
Raw Score SD	11.92	11.17	12.13	12.06	11.03	12.26	11.86	11.35	11.99

Table 8.C.27 Raw Score Correlations by Economic Status: Level II

	Disadvantaged			Not Disadvantaged			Unknown Status		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	–	0.81	N/A	–	0.79	N/A	–	0.81	N/A
Mathematics		–	N/A		–	N/A		–	N/A
Science			–			–			–
N	4208	4199	N/A	2129	2120	N/A	250	250	N/A
Raw Score Mean	23.68	21.96	N/A	22.43	20.88	N/A	22.53	21.26	N/A
Raw Score SD	6.04	7.38	N/A	6.31	7.53	N/A	6.09	7.56	N/A

Table 8.C.28 Raw Score Correlations by Economic Status : Level III

	Disadvantaged			Not Disadvantaged			Unknown Status		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	–	0.82	0.79	–	0.84	0.83	–	0.80	0.80
Mathematics		–	0.77		–	0.79		–	0.77
Science			–			–			–
N	4271	4266	2127	2189	2182	1073	154	154	67
Raw Score Mean	23.69	22.06	22.00	22.15	20.43	20.19	24.08	22.56	21.63
Raw Score SD	5.77	6.80	6.02	6.57	7.17	6.81	5.77	6.10	6.17

Table 8.C.29 Raw Score Correlations by Economic Status : Level IV

	Disadvantaged			Not Disadvantaged			Unknown Status		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	–	0.80	0.77	–	0.82	0.77	–	0.81	0.81
Mathematics		–	0.79		–	0.78		–	0.82
Science			–			–			–
N	6272	6258	2002	3289	3282	1116	292	291	72
Raw Score Mean	20.23	19.33	20.06	19.53	18.11	18.91	22.03	20.54	19.32
Raw Score SD	6.98	7.41	6.23	7.28	7.46	6.64	7.03	7.60	6.89

Table 8.C.30 Raw Score Correlations by Economic Status: Level V

	Disadvantaged			Not Disadvantaged			Unknown Status		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	–	0.81	0.73	–	0.82	0.76	–	0.73	0.69
Mathematics		–	0.72		–	0.73		–	0.69
Science			–			–			–
N	6308	6285	2027	3794	3788	1246	415	412	123
Raw Score Mean	21.70	22.09	19.81	21.38	21.37	19.02	23.93	24.00	21.31
Raw Score SD	6.82	7.54	6.20	7.39	7.82	6.58	5.92	6.49	5.88

Table 8.C.31 Raw Score Correlations by Disability: Level I

	Mental Retardation			Hard of Hearing			Deafness			Speech Impairment		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	–	0.81	0.85	–	0.86	0.89	–	0.78	0.69	–	0.81	0.91
Mathematics		–	0.84		–	0.82		–	0.84		–	0.83
Science			–			–			–			–
N	4703	4688	1313	75	75	17	65	65	18	104	103	13
Raw Score Mean	30.04	24.17	25.21	27.69	23.28	21.24	30.15	25.57	27.33	34.86	31.08	29.54
Raw Score SD	11.18	10.75	11.91	11.76	11.08	10.44	10.21	9.52	9.88	8.21	8.27	9.83
Visual Impairment												
Emotional Disturbance			Orthopedic Impairment			Other Health Impairment						
ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	
ELA	–	0.83	–	0.7	0.76	–	0.84	0.86	–	0.87	0.87	
Mathematics		–		–	0.95		–	0.87		–	0.86	
Science					–			–			–	
N	309	306	84	22	22	3	2537	2521	699	307	306	53
Raw Score Mean	22.6	17.4	18.06	32.32	28.14	26.67	21.88	17.15	17.05	27.53	22.62	19.62
Raw Score SD	12.41	10.84	12.48	9.24	11.75	12.22	12.16	11.01	11.7	12.29	11.53	13.78
Specific Learning Disability												
Deaf-Blindness			Multiple Disabilities			Autism						
ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	
ELA	–	0.84	0.74	–	0.88	0.54	–	0.86	0.87	–	0.75	0.78
Mathematics		–	0.72		–	0.95		–	0.87		–	0.78
Science			–			–			–			–
N	80	80	16	31	31	3	1200	1196	327	2804	2796	680
Raw Score Mean	32.48	30.79	31.5	18.29	12.87	9	20.94	16.2	15.68	28.58	23.05	23.46
Raw Score SD	9.57	10.13	8.49	14.32	10.02	3.61	12.8	10.99	11.61	10.13	9.59	10.5
Traumatic Brain Injury												
Unknown Disability												
ELA	Math	Science	ELA	Math	Science							
ELA	–	0.93	0.8	–	0.87	0.86						
Mathematics		–	0.89		–	0.91						
Science			–			–						
N	95	96	28	199	199	42						
Raw Score Mean	24.18	20.18	19.68	26.08	21.42	20.26						
Raw Score SD	14.3	13.68	15.46	12.53	11.94	12.55						

Table 8.C.32 Raw Score Correlations by Disability: Level II

	Mental Retardation			Hard of Hearing			Deafness			Speech Impairment		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	–	0.76	N/A	–	0.74	N/A	–	0.8	N/A	–	0.73	N/A
Mathematics	–	–	N/A	–	–	N/A	–	–	N/A	–	–	N/A
Science	–	–	–	–	–	–	–	–	–	–	–	–
N	2170	2168	N/A	42	42	N/A	40	40	N/A	694	694	N/A
Raw Score Mean	22.19	19.42	N/A	24.1	23.48	N/A	22.93	22.58	N/A	26.65	25.74	N/A
Raw Score SD	5.55	6.97	N/A	4.76	6.9	N/A	5.53	6.9	N/A	4.51	5.53	N/A
Visual Impairment												
Emotional Disturbance			Orthopedic Impairment			Other Health Impairment						
ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	
ELA	–	0.75	N/A	–	0.89	N/A	–	0.8	N/A	–	0.81	N/A
Mathematics	–	–	N/A	–	–	N/A	–	–	N/A	–	–	N/A
Science	–	–	–	–	–	–	–	–	–	–	–	–
N	41	41	N/A	39	39	N/A	309	309	N/A	324	323	N/A
Raw Score Mean	22.61	20.98	N/A	25.65	25.05	N/A	23.55	21.1	N/A	24.74	23.09	N/A
Raw Score SD	7.17	7.24	N/A	7.88	7.95	N/A	6.04	7.4	N/A	5.39	7.07	N/A
Specific Learning Disability												
Deaf-Blindness			Multiple Disabilities			Autism						
ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	
ELA	–	0.66	N/A	–	0.29	N/A	–	0.78	N/A	–	0.82	N/A
Mathematics	–	–	N/A	–	–	N/A	–	–	N/A	–	–	N/A
Science	–	–	–	–	–	–	–	–	–	–	–	–
N	558	556	N/A	4	4	N/A	154	153	N/A	2038	2028	N/A
Raw Score Mean	27.54	27.46	N/A	25.5	24.75	N/A	21.75	18.41	N/A	21.83	20.78	N/A
Raw Score SD	4.05	4.47	N/A	2.65	5.91	N/A	6.2	7.37	N/A	6.77	7.76	N/A
Traumatic Brain Injury												
Unknown Disability												
ELA	Math	Science	ELA	Math	Science							
ELA	–	0.75	N/A	–	0.82	N/A						
Mathematics	–	–	N/A	–	–	N/A						
Science	–	–	–	–	–	–						
N	28	28	N/A	145	144	N/A						
Raw Score Mean	23.54	22.07	N/A	22.52	21.84	N/A						
Raw Score SD	5.32	6.09	N/A	6.42	7.62	N/A						

Table 8.C.33 Raw Score Correlations by Disability: Level III

	Mental Retardation			Hard of Hearing			Deafness			Speech Impairment		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	—	0.80	0.75	—	0.72	0.79	—	0.78	0.81	—	0.7	0.67
Mathematics		—	0.73		—	0.78		—	0.87		—	0.58
Science			—			—			—			—
N	2711	2707	1391	45	45	27	64	64	34	379	380	158
Raw Score Mean	22.52	20.22	20.83	24.6	24.6	22	23.86	25.16	23.35	26.56	26.18	24.97
Raw Score SD	5.59	6.76	5.66	5.86	6.3	6.14	5.75	6.89	6.74	3.97	4.35	4.49
Other Health Impairment												
Visual Impairment			Emotional Disturbance			Orthopedic Impairment			Other Health Impairment			
ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	
ELA	—	0.73	0.49	—	0.83	0.72	—	0.80	0.81	—	0.79	0.79
Mathematics		—	0.87		—	0.83		—	0.8		—	0.71
Science			—			—			—			—
N	43	43	18	55	55	32	376	374	195	317	314	145
Raw Score Mean	22.19	19.98	18.44	26.75	26.02	25.31	23.36	20.37	21.88	24.91	23.2	23.29
Raw Score SD	6.09	7.15	7.84	5.04	5.42	6.42	5.73	6.83	6.04	5.38	6.31	5.4
Autism												
Specific Learning Disability			Deaf-Blindness			Multiple Disabilities			Autism			
ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	
ELA	—	0.64	0.56	—	1	N/A	—	0.80	0.79	—	0.86	0.84
Mathematics		—	0.5		—	N/A		—	0.74		—	0.82
Science			—			—			—			—
N	553	554	275	2	2	1	170	170	85	1759	1754	845
Raw Score Mean	28.03	27.38	26.94	22.5	23	9	21.69	19.4	20.02	21.61	20.51	19.41
Raw Score SD	3.22	3.38	3.79	13.44	12.73	N/A	5.93	6.87	5.73	6.91	7.17	7.05
Traumatic Brain Injury												
Unknown Disability			Unknown Disability			Unknown Disability			Unknown Disability			
ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	
ELA	—	0.84	0.79	—	0.86	0.83						
Mathematics		—	0.68		—	0.76						
Science			—			—						
N	38	38	20	102	102	41						
Raw Score Mean	24.03	22.61	23.2	23.26	21.67	20.34						
Raw Score SD	5.81	6.49	5.35	6.71	7.3	8.16						

Table 8.C.34 Raw Score Correlations by Disability: Level IV

	Mental Retardation			Hard of Hearing			Deafness			Speech Impairment		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	—	0.80	0.74	—	0.86	0.82	—	0.68	0.67	—	0.67	0.77
Mathematics	—	—	0.77	—	—	0.92	—	—	0.65	—	—	0.75
Science	—	—	—	—	—	—	—	—	—	—	—	—
N	4544	4539	1554	66	65	19	107	107	32	293	293	76
Raw Score Mean	19.09	17.59	18.97	18.73	19.29	18.16	20.08	23.45	22.88	24.16	23.93	23.61
Raw Score SD	6.86	6.89	5.98	7.47	7.76	8.2	5.59	5.64	4.51	5.31	6.22	5.37
	Visual Impairment			Emotional Disturbance			Orthopedic Impairment			Other Health Impairment		
ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	
ELA	—	0.79	0.81	—	0.72	0.63	—	0.83	0.84	—	0.79	0.72
Mathematics	—	—	0.91	—	—	0.59	—	—	0.84	—	—	0.74
Science	—	—	—	—	—	—	—	—	—	—	—	—
N	58	58	20	97	96	35	583	581	186	484	482	142
Raw Score Mean	20.26	18.03	19.8	24.35	23.92	24.66	19.87	17.43	19.05	22.27	20.57	21.63
Raw Score SD	6.71	7.27	6.58	6.08	7.49	5.06	7.19	7.11	6.68	6.61	6.99	5.5
	Specific Learning Disability			Deaf-Blindness			Multiple Disabilities			Autism		
ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	
ELA	—	0.59	0.55	—	N/A	N/A	—	0.82	0.82	—	0.82	0.81
Mathematics	—	—	0.69	—	—	N/A	—	—	0.80	—	—	0.82
Science	—	—	—	—	—	—	—	—	—	—	—	—
N	865	864	261	1	1	0	305	304	117	2171	2164	674
Raw Score Mean	25.2	25.07	23.86	18	13	N/A	18.23	16.68	18.8	18.83	18.45	18.45
Raw Score SD	4.63	5.59	4.99	N/A	N/A	N/A	7.01	7.22	6.29	7.5	7.94	7.11
	Traumatic Brain Injury			Unknown Disability								
ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	
ELA	—	0.77	0.54	—	0.80	0.73	—	—	—	—	—	—
Mathematics	—	—	0.62	—	—	0.67	—	—	—	—	—	—
Science	—	—	—	—	—	—	—	—	—	—	—	—
N	72	71	27	207	206	47	—	—	—	—	—	—
Raw Score Mean	22.4	20.25	21.48	21.96	20.58	20.83	—	—	—	—	—	—
Raw Score SD	7.27	7.88	5.94	6.44	7.06	6.14	—	—	—	—	—	—

Table 8.C.35 Raw Score Correlations by Disability: Level V

	Mental Retardation			Hard of Hearing			Deafness			Speech Impairment		
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
ELA	—	0.81	0.70	—	0.73	0.39	—	0.56	0.58	—	0.51	0.51
Mathematics		—	0.70		—	0.56		—	0.57		—	0.48
Science			—			—			—			—
N	5267	5255	1676	90	90	34	158	157	53	225	225	77
Raw Score Mean	20.9	20.85	18.86	21.99	23.86	20.03	22.93	26.18	22.74	25.74	26.35	23.45
Raw Score SD	6.69	7.42	5.91	6.6	7.07	6.1	5.01	4.63	4.57	4.43	4.78	4.57
Other Health Impairment												
Visual Impairment			Emotional Disturbance			Orthopedic Impairment			Other Health Impairment			
ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	
ELA	—	0.82	0.65	—	0.79	0.61	—	0.82	0.83	—	0.81	0.61
Mathematics		—	0.71		—	0.64		—	0.76		—	0.66
Science			—			—			—			—
N	85	84	26	166	166	63	530	526	195	467	466	144
Raw Score Mean	21.38	18.98	21.69	26.13	26.33	23.97	20.48	19.31	18.65	24.34	24.1	21.67
Raw Score SD	7.01	8.15	5.73	5.49	6.31	6.48	7.52	8.31	6.93	6.09	6.94	5.64
Autism												
Specific Learning Disability			Deaf-Blindness			Multiple Disabilities			Autism			
ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	
ELA	—	0.52	0.44	—	0.97	N/A	—	0.84	0.75	—	0.83	0.81
Mathematics		—	0.48		—	N/A		—	0.72		—	0.78
Science			—			—			—			—
N	904	901	295	5	5	0	322	321	101	1960	1954	645
Raw Score Mean	26.41	27.37	23.44	7.8	4	0	19.72	19.7	18.66	20.35	21.31	18.18
Raw Score SD	3.79	4.13	4.62	10.23	5.66	N/A	7.67	8.21	6.57	8.04	8.16	7.14
Traumatic Brain Injury												
Unknown Disability			Unknown Disability			Unknown Disability			Unknown Disability			
ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	
ELA	—	0.75	0.76	—	0.74	0.73	—	0.74	0.74	—	0.74	0.74
Mathematics		—	0.53		—	0.74		—	0.74		—	0.74
Science			—			—			—			—
N	78	77	22	260	258	65	—	—	—	—	—	—
Raw Score Mean	24	24	20.86	23.4	23.57	20.82						
Raw Score SD	5.89	6.66	5.85	6.02	6.61	6.12						

Table 8.C.36 Inter-Rater Reliabilities for Operational Tasks: Level I

Level I		First Rating			Second Rating			% Agreement			MAD *	Corr †
Subject	Task	N	Mean	SD	N	Mean	SD	Exact	Adjacent	Neither		
English– Language Arts	1	1,732	3.71	1.82	1,732	3.69	1.82	95.55	3.41	1.04	0.07	0.98
	3	1,732	2.99	1.96	1,732	2.97	1.95	94.63	3.46	1.91	0.10	0.97
	4	1,732	3.58	1.88	1,732	3.57	1.88	94.57	4.04	1.38	0.08	0.98
	6	1,732	3.94	1.73	1,732	3.92	1.74	94.23	3.29	2.49	0.12	0.94
	7	1,732	3.66	1.84	1,732	3.63	1.84	94.92	3.41	1.68	0.09	0.96
	9	1,732	3.75	1.83	1,732	3.73	1.85	95.61	2.89	1.50	0.08	0.97
	10	1,732	3.78	1.81	1,732	3.76	1.82	95.27	3.35	1.38	0.08	0.97
	12	1,732	3.54	1.87	1,732	3.52	1.87	95.79	2.77	1.44	0.08	0.97
Mathematics	1	1,711	3.55	1.89	1,711	3.53	1.89	95.50	3.33	1.17	0.08	0.97
	3	1,711	3.08	1.94	1,711	3.04	1.94	94.97	3.10	1.94	0.10	0.96
	4	1,711	2.80	1.93	1,711	2.79	1.93	94.97	3.16	1.87	0.10	0.96
	6	1,711	2.88	1.92	1,711	2.86	1.92	95.68	2.92	1.41	0.07	0.97
	7	1,711	2.21	1.78	1,711	2.19	1.77	95.32	3.39	1.29	0.07	0.97
	9	1,711	2.95	1.94	1,711	2.93	1.93	94.74	3.45	1.82	0.09	0.97
	10	1,711	2.57	1.89	1,711	2.53	1.89	94.51	3.86	1.64	0.09	0.97
	12	1,711	3.23	1.97	1,711	3.22	1.96	95.38	2.86	1.76	0.09	0.96
Science	1	446	2.94	1.95	446	2.89	1.97	95.96	3.14	0.90	0.07	0.97
	3	446	2.87	1.96	446	2.84	1.97	95.52	3.36	1.12	0.07	0.97
	4	446	2.97	1.97	446	2.94	1.98	93.72	4.93	1.34	0.10	0.97
	6	446	3.01	2.00	446	2.96	2.00	93.50	4.93	1.57	0.11	0.96
	7	446	2.97	1.95	446	2.95	1.96	95.74	2.47	1.79	0.08	0.97
	9	446	2.65	1.96	446	2.64	1.95	94.17	3.81	2.02	0.11	0.95
	10	446	2.65	1.93	446	2.61	1.95	94.62	3.59	1.78	0.09	0.97
	12	446	3.60	1.92	446	3.60	1.93	93.95	3.14	2.92	0.15	0.92

* Mean absolute difference between first and second ratings

† Pearson correlation between first and second ratings

Table 8.C.37 Inter-Rater Reliabilities for Operational Tasks: Level II

Level II		First Rating			Second Rating			% Agreement			MAD *	Corr †
Subject	Task	N	Mean	SD	N	Mean	SD	Exact	Adjacent	Neither		
English– Language Arts	1	1,250	3.86	0.50	1,250	3.86	0.51	99.20	0.64	0.16	0.01	0.97
	3	1,250	2.31	1.24	1,250	2.31	1.23	93.68	5.04	1.28	0.08	0.96
	4	1,250	2.45	1.37	1,250	2.45	1.37	94.40	4.08	1.52	0.08	0.97
	6	1,250	2.77	1.33	1,250	2.78	1.32	95.12	3.92	0.96	0.06	0.97
	7	1,250	3.68	0.73	1,250	3.68	0.73	97.60	2.08	0.32	0.03	0.96
	9	1,250	2.65	1.20	1,250	2.64	1.20	96.24	3.20	0.56	0.05	0.98
	10	1,250	3.14	1.10	1,250	3.13	1.09	95.20	4.48	0.32	0.05	0.98
	12	1,250	2.27	1.10	1,250	2.28	1.10	93.60	5.28	1.12	0.09	0.93
Mathematics	1	1,242	3.21	1.14	1,242	3.20	1.14	98.15	1.45	0.40	0.03	0.98
	3	1,242	2.83	1.38	1,242	2.84	1.36	97.50	1.85	0.64	0.04	0.98
	4	1,242	2.79	1.22	1,242	2.79	1.22	96.46	2.98	0.56	0.04	0.98
	6	1,242	2.02	1.38	1,242	2.01	1.38	96.78	2.17	1.04	0.05	0.97
	7	1,242	2.65	1.06	1,242	2.64	1.07	95.81	3.78	0.40	0.05	0.97
	9	1,242	2.36	1.35	1,242	2.36	1.36	97.42	1.85	0.72	0.04	0.98
	10	1,242	3.14	1.18	1,242	3.15	1.18	97.34	1.85	0.80	0.04	0.97
	12	1,242	3.03	1.22	1,242	3.07	1.18	95.65	2.74	1.61	0.08	0.92

* Mean absolute difference between first and second ratings

† Pearson correlation between first and second ratings

Table 8.C.38 Inter-Rater Reliabilities for Operational Tasks: Level III

Level III Subject	Task	First Rating			Second Rating			% Agreement			MAD *	Corr †
		N	Mean	SD	N	Mean	SD	Exact	Adjacent	Neither		
English– Language Arts	1	1,167	3.18	0.94	1,167	3.19	0.94	93.14	6.51	0.35	0.07	0.95
	3	1,167	2.61	1.04	1,167	2.61	1.04	94.17	5.31	0.52	0.07	0.96
	4	1,167	3.00	0.97	1,167	3.00	0.98	96.06	3.26	0.68	0.05	0.96
	6	1,167	2.31	0.80	1,167	2.33	0.79	92.80	6.00	1.20	0.08	0.91
	7	1,167	3.29	0.78	1,167	3.28	0.79	96.83	2.74	0.43	0.04	0.96
	9	1,167	2.95	1.28	1,167	2.96	1.27	96.40	2.91	0.68	0.05	0.97
	10	1,167	3.09	1.11	1,167	3.10	1.10	96.74	2.57	0.69	0.05	0.97
	12	1,167	3.14	1.11	1,167	3.17	1.08	95.63	2.83	1.55	0.08	0.91
Mathematics	1	1,166	3.10	1.23	1,166	3.10	1.24	98.11	1.63	0.26	0.02	0.99
	3	1,166	2.23	0.87	1,166	2.24	0.88	92.45	7.12	0.43	0.08	0.94
	4	1,166	3.08	1.11	1,166	3.09	1.11	97.34	2.23	0.43	0.03	0.98
	6	1,166	2.84	1.37	1,166	2.85	1.37	97.00	2.32	0.69	0.04	0.98
	7	1,166	2.48	1.04	1,166	2.49	1.05	95.63	3.69	0.69	0.05	0.96
	9	1,166	2.48	1.35	1,166	2.47	1.35	97.34	1.97	0.69	0.04	0.98
	10	1,166	3.47	1.03	1,166	3.48	1.02	98.63	1.11	0.26	0.02	0.98
	12	1,166	2.26	1.27	1,166	2.27	1.27	95.45	3.52	1.03	0.07	0.96
Science	1	551	2.80	0.94	551	2.82	0.93	94.56	4.54	0.91	0.07	0.95
	3	551	2.93	1.08	551	2.93	1.08	92.38	7.08	0.54	0.09	0.95
	4	551	2.97	0.89	551	2.98	0.88	94.56	4.17	1.27	0.07	0.92
	6	551	2.68	1.13	551	2.67	1.13	94.56	5.26	0.18	0.06	0.98
	7	551	2.42	1.35	551	2.43	1.35	96.55	2.72	0.72	0.05	0.98
	9	551	2.71	1.15	551	2.70	1.16	96.19	2.72	1.09	0.05	0.97
	10	551	2.35	1.09	551	2.34	1.08	96.37	3.27	0.36	0.04	0.98
	12	551	2.88	0.98	551	2.91	0.94	94.56	4.90	0.54	0.06	0.95

* Mean absolute difference between first and second ratings

† Pearson correlation between first and second ratings

Table 8.C.39 Inter-Rater Reliabilities for Operational Tasks: Level IV

Level IV		First Rating			Second Rating			% Agreement			MAD *	Corr †
Subject	Task	N	Mean	SD	N	Mean	SD	Exact	Adjacent	Neither		
English– Language Arts	1	1,369	3.24	1.05	1,369	3.23	1.06	97.74	1.97	0.30	0.03	0.98
	3	1,369	1.53	1.08	1,369	1.53	1.08	92.91	5.92	1.17	0.09	0.94
	4	1,369	3.17	1.18	1,369	3.16	1.18	93.50	5.62	0.87	0.08	0.96
	6	1,369	2.64	1.18	1,369	2.62	1.19	92.55	6.65	0.81	0.08	0.96
	7	1,369	2.34	1.14	1,369	2.32	1.14	90.87	8.11	1.02	0.10	0.95
	9	1,369	2.11	1.22	1,369	2.10	1.22	95.18	4.09	0.73	0.06	0.97
	10	1,369	2.65	1.09	1,369	2.66	1.09	96.20	3.21	0.58	0.05	0.97
	12	1,369	2.80	1.28	1,369	2.81	1.28	94.74	4.38	0.88	0.07	0.96
Mathematics	1	1,355	2.35	1.27	1,355	2.35	1.27	96.68	2.95	0.37	0.04	0.98
	3	1,355	1.76	1.24	1,355	1.77	1.24	97.34	2.29	0.37	0.03	0.98
	4	1,355	1.91	1.17	1,355	1.91	1.16	95.65	3.76	0.59	0.05	0.97
	6	1,355	3.04	1.33	1,355	3.04	1.32	97.42	2.07	0.52	0.03	0.98
	7	1,355	2.07	1.32	1,355	2.07	1.32	96.38	3.10	0.52	0.04	0.98
	9	1,355	3.31	1.00	1,355	3.30	1.01	97.20	2.07	0.74	0.04	0.95
	10	1,355	2.46	1.35	1,355	2.47	1.35	97.71	1.70	0.59	0.04	0.98
	12	1,355	2.25	1.10	1,355	2.24	1.09	95.20	4.06	0.74	0.06	0.96
Science	1	394	2.38	1.20	394	2.38	1.19	97.46	2.03	0.50	0.03	0.98
	3	394	2.39	1.05	394	2.41	1.05	97.72	2.03	0.25	0.03	0.99
	4	394	2.29	1.18	394	2.27	1.19	93.91	4.31	1.78	0.09	0.95
	6	394	2.30	1.12	394	2.32	1.11	94.92	4.06	1.02	0.07	0.96
	7	394	2.86	1.14	394	2.86	1.13	95.18	4.06	0.76	0.06	0.97
	9	394	2.48	1.10	394	2.52	1.09	93.91	4.57	1.52	0.08	0.94
	10	394	2.54	0.99	394	2.55	0.99	96.19	3.30	0.50	0.05	0.97
	12	394	2.34	1.13	394	2.35	1.14	92.13	6.35	1.52	0.10	0.95

* Mean absolute difference between first and second ratings

† Pearson correlation between first and second ratings

Table 8.C.40 Inter-Rater Reliabilities for Operational Tasks: Level V

Level V		First Rating			Second Rating			% Agreement			MAD *	Corr †
Subject	Task	N	Mean	SD	N	Mean	SD	Exact	Adjacent	Neither		
English– Language Arts	1	1,048	3.21	1.20	1,048	3.21	1.20	96.09	3.05	0.86	0.05	0.97
	3	1,048	2.83	1.07	1,048	2.82	1.09	92.56	6.97	0.48	0.08	0.95
	4	1,048	2.74	1.05	1,048	2.74	1.05	91.89	7.35	0.77	0.09	0.94
	6	1,048	3.08	1.02	1,048	3.08	1.03	93.32	6.11	0.57	0.08	0.95
	7	1,048	3.10	1.03	1,048	3.10	1.04	93.89	5.34	0.76	0.07	0.95
	9	1,048	1.91	1.16	1,048	1.92	1.15	91.89	6.58	1.53	0.10	0.94
	10	1,048	3.10	1.16	1,048	3.10	1.18	93.13	5.73	1.15	0.09	0.95
	12	1,048	2.44	1.24	1,048	2.43	1.25	94.94	4.20	0.87	0.07	0.96
Mathematics	1	1,036	3.06	1.27	1,036	3.06	1.27	97.97	1.35	0.67	0.03	0.98
	3	1,036	3.22	1.14	1,036	3.23	1.13	96.91	2.32	0.77	0.05	0.96
	4	1,036	2.89	1.31	1,036	2.89	1.30	94.50	4.15	1.35	0.08	0.96
	6	1,036	2.87	1.11	1,036	2.87	1.12	96.14	3.19	0.67	0.05	0.96
	7	1,036	2.43	1.37	1,036	2.43	1.36	93.92	4.15	1.94	0.09	0.95
	9	1,036	2.64	1.41	1,036	2.63	1.42	95.37	3.57	1.07	0.07	0.97
	10	1,036	2.35	1.21	1,036	2.35	1.22	95.95	2.90	1.16	0.06	0.96
	12	1,036	3.26	1.25	1,036	3.29	1.21	96.43	2.03	1.54	0.07	0.93
Science	1	349	2.31	1.02	349	2.34	1.00	93.41	5.44	1.15	0.08	0.94
	3	349	2.34	1.07	349	2.36	1.08	91.40	7.74	0.86	0.10	0.95
	4	349	2.70	1.03	349	2.67	1.06	93.70	5.16	1.15	0.08	0.95
	6	349	2.09	1.08	349	2.08	1.08	95.70	3.72	0.58	0.05	0.96
	7	349	2.90	0.98	349	2.87	0.99	96.56	2.87	0.58	0.05	0.96
	9	349	2.74	0.94	349	2.75	0.93	94.56	4.30	1.15	0.07	0.92
	10	349	2.17	1.14	349	2.16	1.13	94.27	4.87	0.87	0.07	0.95
	12	349	2.90	1.20	349	2.95	1.17	96.28	1.72	2.01	0.08	0.91

* 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	8	10	4
B	71	67	32
C	57	59	25
D	12	12	3
F	0	0	0

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

Fit	ELA Frequency	Mathematics Frequency	Science Frequency
A	0	2	0
B	9	21	8
C	17	8	7
D	6	1	1
F	0	0	0

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

Fit	ELA Frequency	Mathematics Frequency
A	0	2
B	16	14
C	7	7
D	1	1
F	0	0

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

Fit	ELA Frequency	Mathematics Frequency	Science Frequency
A	2	2	2
B	13	11	6
C	9	9	7
D	0	2	1
F	0	0	0

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

Fit	ELA Frequency	Mathematics Frequency	Science Frequency
A	2	4	2
B	17	13	8
C	14	12	6
D	3	7	0
F	0	0	0

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

Fit	ELA Frequency	Mathematics Frequency	Science Frequency
A	4	0	0
B	16	8	10
C	10	23	5
D	2	1	1
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	-0.74	0.10	-0.82	-0.51
	Field-Test Items	24	-0.61	0.21	-1.09	-0.34
II	All Operational Items	8	-1.54	0.79	-3.08	-0.87
	Field-Test Items	16	-1.22	0.55	-2.03	-0.31
III	All Operational Items	8	-1.52	0.51	-2.15	-0.71
	Field-Test Items	16	-1.36	0.66	-2.82	-0.39
IV	All Operational Items	8	-0.93	0.58	-1.91	0.07
	Field-Test Items	28	-1.30	0.67	-2.61	-0.31
V	All Operational Items	8	-1.19	0.46	-1.73	-0.33
	Field-Test Items	24	-1.31	0.46	-2.34	-0.46

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	-0.29	0.14	-0.52	-0.06
	Field-Test Items	24	-0.25	0.12	-0.46	-0.03
II	All Operational Items	8	-1.18	0.33	-1.61	-0.64
	Field-Test Items	16	-1.35	0.52	-2.07	0.25
III	All Operational Items	8	-1.29	0.39	-1.87	-0.77
	Field-Test Items	16	-1.18	0.48	-1.90	-0.47
IV	All Operational Items	8	-0.85	0.48	-1.76	-0.24
	Field-Test Items	28	-1.02	0.58	-1.85	0.61
V	All Operational Items	8	-1.21	0.25	-1.45	-0.76
	Field-Test Items	24	-1.15	0.42	-1.70	-0.20

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	-0.23	0.16	-0.53	0.00
	Field-Test Items	8	-0.54	0.34	-1.13	-0.24
III	All Operational Items	8	-1.29	0.29	-1.71	-0.96
	Field-Test Items	8	-1.36	0.49	-1.99	-0.58
IV	All Operational Items	8	-0.95	0.17	-1.22	-0.71
	Field-Test Items	8	-1.01	0.38	-1.52	-0.47
V	All Operational Items	8	-0.54	0.30	-0.94	-0.07
	Field-Test Items	8	-0.39	0.30	-0.89	0.09

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

Raw Score	Theta	Scale Score	CSEM	Performance Level
40	20	60	17	
39	0.9978	52	9	
38	0.6258	48	6	
37	0.4359	46	5	
36	0.3075	44	4	
35	0.2092	43	3	Advanced
34	0.1287	42	3	
33	0.0598	42	3	
32	-0.001	41	3	
31	-0.056	40	3	
30	-0.1065	40	3	
29	-0.1537	39	2	
28	-0.1982	39	2	
27	-0.2407	38	2	
26	-0.2817	38	2	
25	-0.3215	37	2	
24	-0.3606	37	2	Proficient
23	-0.3991	36	2	
22	-0.4374	36	2	
21	-0.4758	35	2	
20	-0.5145	35	2	
19	-0.5539	35	2	
18	-0.5943	34	2	
17	-0.6361	34	2	
16	-0.6797	33	2	
15	-0.7257	33	2	
14	-0.7748	32	3	Basic
13	-0.828	31	3	
12	-0.8866	31	3	
11	-0.9525	30	3	
10	-1.0284	29	3	
9	-1.1187	28	4	
8	-1.2307	27	4	
7	-1.3765	25	5	Below Basic
6	-1.5752	23	5	
5	-1.8487	20	6	
4	-2.2053	16	7	
3	-2.643	15	7	
2	-3.1947	15	7	Far Below Basic
1	-4.0229	15	7	
0	-40	15	7	

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

Raw Score	Theta	Scale Score	CSEM	Performance Level
32	20	60	18	
31	1.8382	48	5	
30	1.2077	45	3	
29	0.8483	43	3	Advanced
28	0.5930	42	2	
27	0.3907	41	2	
26	0.2189	40	2	
25	0.0656	39	2	
24	-0.0763	38	2	
23	-0.2120	38	2	
22	-0.3450	37	2	Proficient
21	-0.4787	36	2	
20	-0.6156	36	2	
19	-0.7585	35	2	
18	-0.9093	34	2	
17	-1.0698	33	2	
16	-1.2399	33	2	
15	-1.4183	32	2	Basic
14	-1.6020	31	2	
13	-1.7881	30	2	
12	-1.9750	29	2	
11	-2.1632	28	2	
10	-2.3555	27	2	
9	-2.5563	26	2	
8	-2.7722	25	2	Below Basic
7	-3.0109	24	2	
6	-3.2810	23	3	
5	-3.5910	21	3	
4	-3.9505	19	3	
3	-4.3784	17	3	
2	-4.9236	15	4	
1	-5.7534	15	4	Far Below Basic
0	-40	15	4	

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

Raw Score	Theta	Scale Score	CSEM	Performance Level
32	20	60	17	
31	2.4573	48	4	
30	1.6585	45	3	
29	1.1544	43	2	Advanced
28	0.7818	42	2	
27	0.4888	41	2	
26	0.2491	40	2	
25	0.0460	39	2	
24	-0.1319	38	2	
23	-0.2925	38	2	
22	-0.4417	37	1	Proficient
21	-0.5837	37	1	
20	-0.7216	36	1	
19	-0.8581	36	1	
18	-0.9951	35	1	
17	-1.1346	34	1	
16	-1.2780	34	1	
15	-1.4268	33	2	
14	-1.5825	33	2	Basic
13	-1.7463	32	2	
12	-1.9199	31	2	
11	-2.1050	31	2	
10	-2.3041	30	2	
9	-2.5199	29	2	
8	-2.7560	28	2	
7	-3.0167	27	2	Below Basic
6	-3.3070	26	2	
5	-3.6328	25	2	
4	-4.0039	23	2	
3	-4.4398	22	3	
2	-4.9898	20	3	Far Below Basic
1	-5.8206	17	4	
0	-40	15	5	

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

Raw Score	Theta	Scale Score	CSEM	Performance Level
32	20	60	10	
31	2.4576	54	6	
30	1.8455	50	4	
29	1.4928	48	3	
28	1.2346	46	3	Advanced
27	1.0226	45	3	
26	0.8371	44	2	
25	0.6685	43	2	
24	0.5115	42	2	
23	0.3626	41	2	
22	0.2199	41	2	
21	0.0815	40	2	
20	-0.0538	39	2	
19	-0.1873	38	2	Proficient
18	-0.3202	37	2	
17	-0.4538	37	2	
16	-0.5893	36	2	
15	-0.7286	35	2	
14	-0.8735	34	2	
13	-1.0270	33	2	
12	-1.1929	32	2	Basic
11	-1.3764	31	3	
10	-1.5850	30	3	
9	-1.8279	29	3	
8	-2.1147	27	3	
7	-2.4501	25	3	Below Basic
6	-2.8306	23	4	
5	-3.2487	20	4	
4	-3.7048	18	4	
3	-4.2160	15	4	
2	-4.8329	15	4	Far Below Basic
1	-5.7238	15	4	
0	-40	15	4	

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

Raw Score	Theta	Scale Score	CSEM	Performance Level
32	20	60	18	
31	2.5773	48	4	
30	1.8453	45	3	
29	1.4126	43	2	
28	1.0996	42	2	Advanced
27	0.8488	41	2	
26	0.6355	41	2	
25	0.4468	40	2	
24	0.2757	39	2	
23	0.1176	39	2	
22	-0.0305	38	1	
21	-0.1710	37	1	
20	-0.3058	37	1	Proficient
19	-0.4369	36	1	
18	-0.5662	36	1	
17	-0.6958	35	1	
16	-0.8279	35	1	
15	-0.9652	34	1	
14	-1.1110	34	1	
13	-1.2693	33	2	
12	-1.4455	33	2	Basic
11	-1.6464	32	2	
10	-1.8807	31	2	
9	-2.1576	30	2	
8	-2.4831	29	2	
7	-2.8532	27	2	
6	-3.2545	26	2	Below Basic
5	-3.6752	24	2	
4	-4.1174	23	3	
3	-4.6021	21	3	
2	-5.1840	19	3	
1	-6.0358	15	4	Far Below Basic
0	-40	15	4	

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

Raw Score	Theta	Scale Score	CSEM	Performance Level
40	20	60	18	
39	1.6068	50	7	
38	1.2352	47	5	
37	1.0454	45	4	
36	0.9168	43	3	
35	0.8183	42	3	
34	0.7372	42	3	Advanced
33	0.6676	41	3	
32	0.6058	40	2	
31	0.5497	40	2	
30	0.4979	39	2	
29	0.4493	39	2	
28	0.4031	38	2	
27	0.3589	38	2	
26	0.316	37	2	
25	0.2741	37	2	
24	0.2329	37	2	
23	0.192	36	2	Proficient
22	0.1512	36	2	
21	0.11	35	2	
20	0.0682	35	2	
19	0.0254	35	2	
18	-0.0187	34	2	
17	-0.0648	34	2	
16	-0.1134	33	2	
15	-0.1652	33	2	
14	-0.2215	32	2	Basic
13	-0.2836	31	3	
12	-0.3538	31	3	
11	-0.4358	30	3	
10	-0.5352	29	3	
9	-0.663	28	4	
8	-0.8397	26	5	Below Basic
7	-1.1017	23	6	
6	-1.4751	20	6	
5	-1.9146	15	7	
4	-2.3733	15	7	
3	-2.8626	15	7	
2	-3.4406	15	7	Far Below Basic
1	-4.2833	15	7	
0	-40	15	7	

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

Raw Score	Theta	Scale Score	CSEM	Performance Level
32	20	60	15	Advanced
31	1.8176	51	7	
30	1.2554	46	4	
29	0.9453	44	4	
28	0.7249	43	3	
27	0.5484	41	3	
26	0.3968	40	3	Proficient
25	0.2608	39	3	
24	0.1353	38	3	
23	0.0170	37	2	
22	-0.0963	37	2	
21	-0.2063	36	2	
20	-0.3143	35	2	Basic
19	-0.4217	34	2	
18	-0.5297	33	2	
17	-0.6399	33	2	
16	-0.7537	32	2	
15	-0.8732	31	3	
14	-1.0009	30	3	Below Basic
13	-1.1403	29	3	
12	-1.2962	28	3	
11	-1.4756	27	3	
10	-1.6884	25	4	
9	-1.9469	23	4	
8	-2.2613	21	4	Far Below Basic
7	-2.6280	18	5	
6	-3.0271	15	5	
5	-3.4419	15	5	
4	-3.8741	15	5	
3	-4.3466	15	5	
2	-4.9156	15	5	Far Below Basic
1	-5.7545	15	5	
0	-40	15	5	

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

Raw Score	Theta	Scale Score	CSEM	Performance Level
32	20	60	17	
31	2.2423	48	5	
30	1.5373	44	3	Advanced
29	1.1305	42	3	
28	0.8473	41	3	
27	0.6293	40	2	
26	0.4500	39	2	
25	0.2952	38	2	Proficient
24	0.1567	37	2	
23	0.0294	37	2	
22	-0.0902	36	2	
21	-0.2047	36	2	
20	-0.3161	35	2	
19	-0.4261	34	2	
18	-0.5364	34	2	Basic
17	-0.6486	33	2	
16	-0.7644	33	2	
15	-0.8860	32	2	
14	-1.0162	31	2	
13	-1.1590	31	2	
12	-1.3205	30	2	
11	-1.5102	29	2	
10	-1.7447	28	3	Below Basic
9	-2.0496	26	3	
8	-2.4507	24	3	
7	-2.9304	22	4	
6	-3.4229	20	3	
5	-3.8952	17	3	
4	-4.3593	15	3	Far Below Basic
3	-4.8492	15	3	
2	-5.4277	15	3	
1	-6.2713	15	3	
0	-40	15	3	

Table 8.D.18 Score Conversions: Level IV, Mathematics¹

Raw Score	Theta	Scale Score	CSEM	Performance Level
32	20	60	12	
31	2.3443	52	6	
30	1.7547	48	4	
29	1.4380	46	3	
28	1.2174	45	3	Advanced
27	1.0429	43	3	
26	0.8940	42	3	
25	0.7604	41	2	
24	0.6364	41	2	
23	0.5181	40	2	
22	0.4032	39	2	
21	0.2896	38	2	
20	0.1760	37	2	Proficient
19	0.0611	37	2	
18	-0.0562	36	2	
17	-0.1770	35	2	
16	-0.3027	34	2	
15	-0.4349	33	3	
14	-0.5763	32	3	Basic
13	-0.7307	31	3	
12	-0.9046	30	3	
11	-1.1084	29	3	
10	-1.3583	27	4	
9	-1.6801	25	4	Below Basic
8	-2.1027	22	5	
7	-2.6161	18	5	
6	-3.1460	15	5	
5	-3.6489	15	5	
4	-4.1377	15	5	
3	-4.6490	15	5	Far Below Basic
2	-5.2470	15	5	
1	-6.1091	15	5	
0	-40	15	5	

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

Raw Score	Theta	Scale Score	CSEM	Performance Level
32	20	60	21	Advanced
31	1.6354	47	6	
30	1.1305	43	3	
29	0.8469	42	3	
28	0.6450	41	3	
27	0.4848	40	2	
26	0.3492	39	2	
25	0.2297	38	2	
24	0.1211	37	2	
23	0.0202	37	2	
22	-0.0753	36	2	
21	-0.1673	36	2	
20	-0.2571	35	2	Basic
19	-0.3460	34	2	
18	-0.4352	34	2	
17	-0.5260	33	2	
16	-0.6199	33	2	
15	-0.7188	32	2	
14	-0.8253	32	2	
13	-0.9431	31	2	
12	-1.0781	30	2	
11	-1.2401	29	3	
10	-1.4469	28	3	
9	-1.7321	26	4	
8	-2.1435	23	4	Far Below Basic
7	-2.6706	20	5	
6	-3.2058	17	4	
5	-3.7017	15	4	
4	-4.1782	15	4	
3	-4.6752	15	4	
2	-5.2581	15	4	
1	-6.1047	15	4	
0	-40	15	4	

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

Raw Score	Theta	Scale Score	CSEM	Performance Level
40	20	60	19	
39	1.5356	50	8	
38	1.1871	46	5	
37	1.0072	44	4	
36	0.8849	43	3	
35	0.7911	42	3	
34	0.7139	41	3	Advanced
33	0.6477	41	3	
32	0.5889	40	2	
31	0.5355	40	2	
30	0.486	39	2	
29	0.4396	39	2	
28	0.3955	38	2	
27	0.353	38	2	
26	0.3117	37	2	
25	0.2713	37	2	
24	0.2314	37	2	
23	0.1916	36	2	Proficient
22	0.1517	36	2	
21	0.1114	35	2	
20	0.0702	35	2	
19	0.028	35	2	
18	-0.0158	34	2	
17	-0.0616	34	2	
16	-0.11	33	2	
15	-0.1617	33	2	
14	-0.2179	32	2	Basic
13	-0.2798	31	3	
12	-0.3496	31	3	
11	-0.4305	30	3	
10	-0.5277	29	3	
9	-0.6505	28	4	
8	-0.8158	26	4	
7	-1.0529	24	5	Below Basic
6	-1.3891	20	6	
5	-1.8012	16	7	
4	-2.2476	15	7	
3	-2.7321	15	7	
2	-3.3084	15	7	Far Below Basic
1	-4.1506	15	7	
0	-40	15	7	

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

Raw Score	Theta	Scale Score	CSEM	Performance Level
32	20	60	30	
31	2.4292	45	4	
30	1.7039	42	2	Advanced
29	1.2755	41	2	
28	0.9700	40	2	
27	0.7313	39	2	
26	0.5329	38	1	
25	0.3603	37	1	
24	0.2047	37	1	
23	0.0602	36	1	Proficient
22	-0.0773	36	1	
21	-0.2108	35	1	
20	-0.3424	35	1	
19	-0.4742	35	1	
18	-0.6079	34	1	
17	-0.7451	34	1	
16	-0.8876	33	1	
15	-1.0375	33	1	Basic
14	-1.1972	32	1	
13	-1.3698	31	2	
12	-1.5594	31	2	
11	-1.7710	30	2	
10	-2.0111	29	2	
9	-2.2865	28	2	
8	-2.6016	27	2	
7	-2.9545	26	2	Below Basic
6	-3.3371	25	2	
5	-3.7426	23	2	
4	-4.1738	22	2	
3	-4.6510	20	2	
2	-5.2274	18	3	
1	-6.0744	15	4	Far Below Basic
0	-40	15	4	

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

Raw Score	Theta	Scale Score	CSEM	Performance Level
32	20	60	21	Advanced
31	2.8774	47	4	
30	2.1494	44	3	
29	1.7132	42	2	
28	1.3969	41	2	
27	1.1457	40	2	
26	0.9348	39	2	
25	0.7506	38	2	
24	0.5850	38	2	
23	0.4325	37	1	
22	0.2890	37	1	
21	0.1517	36	1	
20	0.0181	36	1	
19	-0.1140	35	1	
18	-0.2465	34	1	Basic
17	-0.3814	34	1	
16	-0.5210	33	1	
15	-0.6677	33	1	
14	-0.8247	32	2	
13	-0.9958	32	2	
12	-1.1862	31	2	
11	-1.4024	30	2	
10	-1.6527	29	2	
9	-1.9449	28	2	Below Basic
8	-2.2824	27	2	
7	-2.6573	25	2	
6	-3.0535	24	2	
5	-3.4610	22	2	
4	-3.8853	20	3	
3	-4.3501	19	3	
2	-4.9117	16	3	Far Below Basic
1	-5.7432	15	3	
0	-40	15	3	

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

Raw Score	Theta	Scale Score	CSEM	Performance Level
32	20	60	25	
31	3.5991	46	3	
30	2.8412	44	3	
29	2.3770	42	2	Advanced
28	2.0341	41	2	
27	1.7575	40	2	
26	1.5218	39	2	
25	1.3130	38	1	
24	1.1225	38	1	
23	0.9449	37	1	
22	0.7761	37	1	Proficient
21	0.6133	36	1	
20	0.4542	36	1	
19	0.2968	35	1	
18	0.1393	34	1	
17	-0.0202	34	1	
16	-0.1838	33	1	
15	-0.3541	33	1	Basic
14	-0.5340	32	1	
13	-0.7273	32	1	
12	-0.9389	31	2	
11	-1.1754	30	2	
10	-1.4451	29	2	
9	-1.7567	28	2	
8	-2.1150	27	2	
7	-2.5132	25	2	Below Basic
6	-2.9342	24	2	
5	-3.3662	23	2	
4	-3.8140	21	2	
3	-4.3018	19	3	
2	-4.8861	17	3	Far Below Basic
1	-5.7406	15	3	
0	-40	15	3	

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	VC208341	IV	12	Operational	0.346	White/Filipino	Filipino
	VC208675	V	10	Operational	–0.381	White/Filipino	White
	VC208660	V	12	Operational	0.407	White/ Filipino	Filipino
English– Language Arts Field-test Tasks	VC476360	V	5	2,8	–0.255	White/CoAsian	White
	VC476388	V	8	2,8	–0.275	White/CoAsian	White
Mathematics Operational Tasks	VC207333	III	21	Operational	0.437	White/Filipino	Filipino
Mathematics Field-test Tasks*	–	–	–	–	–	–	–
Science Operational Tasks*	–	–	–	–	–	–	–
Science Field-test Tasks*	–	–	–	–	–	–	–

* No science 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	VC208510	IV	1	Operational	0.431	MR/Autism	Autism
	VC208571	IV	4	Operational	–0.559	MR/Autism	MR
	VC208470	IV	6	Operational	–0.472	MR/Autism	MR
	VC208476	IV	7	Operational	–0.426	MR/Autism	MR
	VC208488	IV	9	Operational	0.314	MR/Autism	Autism
	VC208341	IV	12	Operational	0.683	MR/Autism	Autism
	VC208668	V	9	Operational	0.400	MR/Autism	Autism
	VC208675	V	10	Operational	–0.547	MR/Autism	MR
	VC208660	V	12	Operational	0.620	MR/Autism	Autism
English– Language Arts Field-test Tasks	VC458360	I	11	1, 7	–0.645	MR/OI	MR
	VC458377	I	11	4	–1.013	MR/OI	MR
	VC458397	I	8	5	–0.501	MR/OI	MR
	VC458360	I	11	1, 7	–0.514	MR/MD	MR
	VC458397	I	8	5	0.559	MR/Autism	Autism
	VC473405	II	8	2, 6	–0.293	MR/Autism	MR
	VC473409	II	2	3, 7	–0.246	MR/Autism	MR
	VC472221	III	11	2, 6	0.285	MR/Autism	Autism
	VC473012	III	11	4, 8	–0.294	MR/Autism	MR
	VC474233	IV	5	1, 8	–0.525	MR/Autism	MR
	VC472332	IV	5	3	–0.334	MR/Autism	MR
	VC473047	IV	11	3	–0.389	MR/Autism	MR
	VC472253	IV	2	5	0.742	MR/Autism	Autism
	VC471327	IV	8	5	–0.539	MR/Autism	MR
	VC476678	IV	11	5	–0.449	MR/Autism	MR
	VC471329	IV	5	7	–0.387	MR/Autism	MR
	VC471328	IV	8	7	–0.456	MR/Autism	MR
VC476682	IV	11	7	0.852	MR/Autism	Autism	
VC472965	V	2	2, 8	0.448	MR/Autism	Autism	

Content Area	Task No.	Level	Task#	Version	SMD	Comparison	In Favor Of
English– Language Arts Field-test Tasks (cont.)	VC476392	V	11	2, 8	–0.447	MR/Autism	MR
	VC476414	V	5	3	–0.447	MR/Autism	MR
	VC476415	V	5	4	–0.357	MR/Autism	MR
	VC476377	V	5	5	–0.379	MR/Autism	MR
	VC476416	V	11	5	–0.525	MR/Autism	MR
	VC476369	V	5	6	–0.392	MR/Autism	MR
Mathematics Operational Tasks	VC207429	III	15	OP	–0.305	MR/Autism	MR
	VC207333	III	21	OP	0.349	MR/Autism	Autism
	VC208066	V	21	OP	–0.433	MR/SL	MR
Mathematics Field-test Tasks	VC463974	II	17	1, 5	–0.348	MR/Autism	MR
	VC464012	II	17	2, 6	–0.312	MR/Autism	MR
	VC471040	II	20	2, 6	0.350	MR/Autism	Autism
	VC464053	II	14	4, 8	–0.325	MR/Autism	MR
	VC464015	II	20	4, 8	0.644	MR/Autism	Autism
	VC466254	III	17	3, 7	0.365	MR/Autism	Autism
	VC465932	III	17	4, 8	0.430	MR/Autism	Autism
	VC466244	III	20	4, 8	0.442	MR/Autism	Autism
	VC464468	IV	20	1, 8	0.362	MR/Autism	Autism
	VC469751	V	14	2, 8	0.294	MR/SL	Specific Learning Disability
VC473487	V	14	1, 7	0.420	MR/AU	Autism	
Science Operational Tasks*	VC206876	V	30	OP	0.277	MR/OI	Ortholmped
Science Field- test Tasks	VC431295	I	32	2,4,6,8	–0.937	MR/OI	MR

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

Disability	ELA		Mathematics		Science	
	Frequency	Percent	Frequency	Percent	Frequency	Percent
Mental Retardation	4,703	37.5	4,688	37.6	1,313	39.8
Hard of Hearing	75	0.6	75	0.6	17	0.5
Deafness	65	0.5	65	0.5	18	0.6
Speech or Language Impairment	104	0.8	103	0.8	13	0.4
Visual Impairment	309	2.5	306	2.5	84	2.6
Emotional Disturbance	22	0.2	22	0.2	3	0.1
Orthopedic Impairment	2,537	20.3	2,521	20.2	699	21.2
Other Health Impairment	307	2.5	306	2.5	53	1.6
Specific Learning Disability	80	0.6	80	0.6	16	0.5
Deaf-Blindness	31	0.3	31	0.3	3	0.1
Multiple Disabilities	1,200	9.6	1,196	9.6	327	9.9
Autism	2,804	22.4	2,796	22.4	680	20.6
Traumatic Brain Injury	95	0.8	96	0.8	28	0.9
Unknown	199	1.6	199	1.6	42	1.3
TOTAL	12,531	100.0	12,484	100.0	3,296	100.0

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

Disability	ELA		Mathematics	
	Frequency	Percent	Frequency	Percent
Mental Retardation	2,170	32.9	2,168	33.0
Hard of Hearing	42	0.6	42	0.6
Deafness	40	0.6	40	0.6
Speech or Language Impairment	694	10.5	694	10.6
Visual Impairment	41	0.6	41	0.6
Emotional Disturbance	40	0.6	39	0.6
Orthopedic Impairment	309	4.7	309	4.7
Other Health Impairment	324	4.9	323	4.9
Specific Learning Disability	558	8.5	556	8.5
Deaf-Blindness	4	0.1	4	0.1
Multiple Disabilities	154	2.3	153	2.3
Autism	2,038	30.9	2,028	30.9
Traumatic Brain Injury	28	0.4	28	0.4
Unknown	145	2.2	144	2.2
TOTAL	6,587	100.0	6,569	100.0

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

Disability	ELA		Mathematics		Science	
	Frequency	Percent	Frequency	Percent	Frequency	Percent
Mental Retardation	2,711	41.0	2,707	41.0	1,391	42.6
Hard of Hearing	45	0.7	45	0.7	27	0.8
Deafness	64	1.0	64	1.0	34	1.0
Speech or Language Impairment	379	5.7	380	5.8	158	4.8
Visual Impairment	43	0.7	43	0.7	18	0.6
Emotional Disturbance	55	0.8	55	0.8	32	1.0
Orthopedic Impairment	376	5.7	374	5.7	195	6.0
Other Health Impairment	317	4.8	314	4.8	145	4.4
Specific Learning Disability	553	8.4	554	8.4	275	8.4
Deaf-Blindness	2	0.0	2	0.0	1	0.0
Multiple Disabilities	170	2.6	170	2.6	85	2.6
Autism	1,759	26.6	1,754	26.6	845	25.9
Traumatic Brain Injury	38	0.6	38	0.6	20	0.6
Unknown	102	1.5	102	1.5	41	1.3
TOTAL	6,614	100.0	6,602	100.0	3,267	100.0

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

Disability	ELA		Mathematics		Science	
	Frequency	Percent	Frequency	Percent	Frequency	Percent
Mental Retardation	4,544	46.1	4,539	46.2	1,554	48.7
Hard of Hearing	66	0.7	65	0.7	19	0.6
Deafness	107	1.1	107	1.1	32	1.0
Speech or Language Impairment	293	3.0	293	3.0	76	2.4
Visual Impairment	58	0.6	58	0.6	20	0.6
Emotional Disturbance	97	1.0	96	1.0	35	1.1
Orthopedic Impairment	583	5.9	581	5.9	186	5.8
Other Health Impairment	484	4.9	482	4.9	142	4.5
Specific Learning Disability	865	8.8	864	8.8	261	8.2
Deaf-Blindness	1	0.0	1	0.0	0	0.0
Multiple Disabilities	305	3.1	304	3.1	117	3.7
Autism	2,171	22.0	2,164	22.0	674	21.1
Traumatic Brain Injury	72	0.7	71	0.7	27	0.9
Unknown	207	2.1	206	2.1	47	1.5
TOTAL	9,853	100.0	9,831	100.0	3,190	100.0

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

Disability	ELA		Mathematics		Science	
	Frequency	Percent	Frequency	Percent	Frequency	Percent
Mental Retardation	5,267	50.1	5,255	50.1	1,676	49.4
Hard of Hearing	90	0.9	90	0.9	34	1.0
Deafness	158	1.5	157	1.5	53	1.6
Speech or Language Impairment	225	2.1	225	2.2	77	2.3
Visual Impairment	85	0.8	84	0.8	26	0.8
Emotional Disturbance	166	1.6	166	1.6	63	1.9
Orthopedic Impairment	530	5.0	526	5.0	195	5.7
Other Health Impairment	467	4.4	466	4.4	144	4.2
Specific Learning Disability	904	8.6	901	8.6	295	8.7
Deaf-Blindness	5	0.1	5	0.1	0	0.0
Multiple Disabilities	322	3.1	321	3.1	101	3.0
Autism	1,960	18.6	1,954	18.6	645	19.0
Traumatic Brain Injury	78	0.7	77	0.7	22	0.7
Unknown	260	2.5	258	2.5	65	1.9
TOTAL	10,517	100.0	10,485	100.0	3,396	100.0

Chapter 9: Quality Control Procedures

ETS implements rigorous quality control procedures 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. OTI provides quality assurance services for all testing programs administered by ETS. In addition, the Quality Assurance division of ETS publishes and maintains the *ETS Standards for Quality and Fairness*, which supports 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 their 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 16. The following sections highlight elements of the process devoted specifically to the quality control of task development.

Task Specifications

ETS maintains task development specifications for 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 panels of task writers that 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 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 takes pride in the typographical integrity of the tasks presented to our clients and 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 will read each task before it is presented to CDE.

Assessment Review Panel (ARP) 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.

Statewide Pupil Assessment Review (SPAR) 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 *Education Code* Section 60614.

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 statistics are examined carefully at data review meetings, which is 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 panel members also use the results of analyses for differential task functioning (DIF) to make judgments about the appropriateness of tasks for various subgroups. The panelists respond to questions such as:

- Are there any instructional issues that have negatively affected the performance of the task?
- Is there a content problem within the task?

The panelists make recommendations about whether to accept or reject each task for inclusion in the STAR item bank.

Quality Control of the Item Bank

After the data review meetings, 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 STAR electronic item bank. The item bank database for CAPA tasks is maintained by a staff of application systems programmers, led by the Item Bank Manager, at ETS. All processes are logged; all change requests—including item bank updates for task availability status—are tracked; and all output and STAR item bank deliveries are quality controlled for accuracy.

Quality of the item bank and secure transfer of the STAR item bank to CDE is very important. The ETS internal task 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 to deliver the STAR electronic item bank to the CDE.

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 offsite, 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 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 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 system, 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 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 by hand.

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 coordinator 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.
5. Make changes and resubmit the post-edit program if errors are identified that require correction.
6. Print a listing of the post-edit file, the correction card file and the original scan file.

Pearson checks correction cards against the post file for corrections 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.
- All school districts and grades are compared to the CDE County-District-School (CDS) Master File.

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

Processing of Answer Documents

Prior to processing operational answer sheets 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 sheets 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 are scored and aggregated to test a broader range of scoring and aggregation scenarios. These procedures assure that students and school districts get 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
- Matching Criteria for ten percent of the CAPA tests that are scored more than once.

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

Quality Control of Psychometric Processes

Score Key Verification Procedures

ETS and Pearson take various necessary measures to ascertain that the scoring keys are applied to the student responses as expected, and the student scores are computed accurately. Scoring keys, provided in the form planners, are produced by ETS and verified thoroughly by performing various quality control checks. The form planners contain the information about an assembled test form including scoring keys, test name, administration year, subscore identification, and the standards and statistics associated with each task. Various checks are performed before keys are finalized, as listed below:

1. The form planners are checked for accuracy against the Form Planner Specification document and the Score Key and Score Conversion document before the keys are loaded into the score key management system (SKM) at ETS.
2. The sequence of tasks in the form planners are matched with their sequence in the actual test booklets.
3. The demarcations of various sections in the actual test book are checked against the list of demarcations provided by the test development staff.

4. Scoring is verified internally at Pearson. ETS independently generates scores and verifies Pearson's scoring of the data by comparing the two results. Any discrepancies are then resolved.
5. The entire scoring system is tested using a test deck that includes typical and extremely atypical responses vectors, as described earlier in section "Processing of Answer Documents" on page 168.
6. Classical item analyses are run on an early sample of data to provide an additional check of the keys. Although rare, if a task is found to be problematic, a followup process is performed that will exclude it from further analyses.

Quality Control of Task 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 CST. Quality assurance checks also include a review 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; the 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 above-mentioned 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 managers 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, 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 difficulties, standard errors for the Rasch task difficulties, 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.

Scoring Tables

Once equating activities are complete and raw-to-scale scoring tables are generated, the psychometricians carry out quality control checks on each scoring table. Scoring tables are checked to verify that all raw scores are included in the tables, scale scores increase as raw scores increase, and that the cut points for the performance levels are correctly identified. 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 compute scale scores for each student. ETS verifies Pearson's scale scores by following procedures, such as:

- 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 on all schools included in those districts, known as “complete districts”

- Reviewing longitudinal data for reasonableness; the results of the analyses are used to look at the trends and trends for the complete districts
- Reviewing longitudinal data for reasonableness using 99 percent of the entire testing population; the results are used to look at the trends for the state as well as few large districts

The results of the longitudinal analyses are provided to the CDE and jointly discussed. Any anomalies in the results are investigated further and jointly discussed. Scores are released after explanations that satisfy both 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 and 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:

- Comparing report formats to input sources from the CDE-approved samples
- Validating and verifying the report data by querying the appropriate student data
- 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
- Proofreading of reports at 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 CDE Master File is provided by the CDE 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 were 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, folded 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 test, Pearson proceeds with the first production batch test. The first production batch is selected to validate a subset of districts that contain examples of key reporting characteristics representative of the state

as a whole. The first production batch test incorporates client-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 tested but marked no answers and was absent, was not tested due to parent/guardian request, or did not complete the test due to illness.