California Department of Education
Standards and Assessment Division

California Alternate Performance Assessment
Technical Report
Spring 2008 Administration

February 2009

Educational Testing Service

Contract No. 5417

Table of Contents

Tables

Acronyms and Initialisms Used in the *California Alternate Performance Assessment Technical Report*

1PPC	1-parameter partial credit
AIS	average task (item) score
API	Academic Performance Index
ARP	Assessment Review Panel
AYP	adequate yearly progress
CAPA	California Alternate Performance Assessment
CMA	California Modified Assessment
CAT/6 Survey	California Achievement Tests, Sixth Edition Survey
CDE	California Department of Education
CI	confidence interval
CSEMs	conditional standard errors of measurement
CSTs	California Standards Tests
DIF	Differential Task (Item) Functioning
DPLT	designated primary language test
ELA	English–language arts
EM	expectation maximization
ETS	Educational Testing Service
FTP	file transfer protocol
GENASYS	Generalized Analysis System
ICCs	task (item) characteristic curves
IEP	individualized education program
IRT	task (item) response theory
NCLB	No Child Left Behind Act of 2001
NSLP	National School Lunch Program
OTI	Office of Testing Integrity
RACF	Random Access Control Facility
RS	raw score
SBE	State Board of Education
SD	standard deviation
SEM	standard error of measurement
SMD	standardized mean difference
SPAR	Statewide Pupil Assessment Review
STAR	Standardized Testing and Reporting
STS	Standards-based Tests in Spanish
WRMSD	weighted root mean square difference

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.

The primary goal of the STAR Program is to help measure how well students are mastering these content standards. During its 2008 administration, the STAR Program had six components:

- California Standardized Tests (CSTs), produced for California public schools
- California Achievement Tests, Sixth Edition Survey (CAT/6 Survey), given in grades three and seven and published by CTB/McGraw-Hill
- California Modified Assessment (CMA), an assessment of students' achievement of California's content standards for English–language arts, mathematics, and science, developed for students with disabilities who meet the CMA eligibility criteria approved by the SBE (In 2008, the CMA was administered to students in grades three, four, and five.)
- California Alternate Performance Assessment (CAPA), produced for students with significant cognitive disabilities who are not able to take the CSTs, the CMA, or the CAT/6 Survey
- Standards-based Tests in Spanish (STS), an assessment of students' achievement of California's content standards, given to Spanish-speaking English learners and administered as the STAR Program's designated primary language test (DPLT) (In 2008, the STS was administered to students in grades two through seven.)
- Aprenda: La prueba de logros en español, Tercera edición (Aprenda 3), given in grades eight and eleven and published by Harcourt Assessment Inc. (The STS replaced the Aprenda 3 as the DPLT in grades two through seven.)

Education Code Section 60602: Legislative Intent

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.

California Alternate Performance Assessment

Target Population

Students with significant cognitive disabilities in grades two through eleven who are unable to take the STAR CSTs even with accommodations or modifications or the CMA with accommodations take the CAPA. Participation in the CAPA and eligibility are determined by a student's individualized education program (IEP) team. Only students whose parents/guardians have submitted written requests to exempt them from STAR Program testing do not take the tests.

The five levels of the CAPA are as follows

- Level I, for students who are the most profoundly cognitively impaired. 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

Students in all five levels are tested in ELA and mathematics. In addition, students in grades five, eight, and ten take a grade-level science test. The CAPA assessments are designed to show how well students with severe cognitive disabilities are doing with respect to California's content standards. These content standards, approved by the SBE, describe what students should know and be able to do at each grade level.

Table 1.1, below, displays the tests administered in 2008 by grade and content area.

Table 1.1 Summary of CAPA Assessment Levels

Test Level	I	II	III	IV	V
Grades	2–11	2 and 3	4 and 5	6–8	9–11
Content Area	ELA	ELA	ELA	ELA	ELA
	Mathematics	Mathematics	Mathematics	Mathematics	Mathematics
	Science Grades 5, 8, and 10 only	–	Science Grade 5 only	Science Grade 8 only	Science Grade 10 only

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. These four field-test tasks differ across versions and allow for the administration of 32 unique tests.

The CAPA tests are administered at different times of year, depending on the progression of the school year within each particular school district. Specifically, schools must administer the CAPA tests within a 21-day window, which begins ten days before and ends ten days after the day on which 85 percent of the instructional year is completed. The CAPA tests are untimed.

Results of the CAPA are reported using scale scores ranging from 15 to 60 for each test. In addition, each student is assigned to one of the following proficiency levels: far below basic, below basic, basic, proficient, and advanced. The state's target is for all students to be classified as proficient or advanced. For all CAPA tests, the minimum scale scores defining basic and proficient are 30 and 35, respectively. The minimum scale scores defining below basic and advanced vary over the CAPA tests. The scale score information can be found in Appendix 3.A.

The total number of students to whom the 2008 CAPA was administered was 44,887.

Significant Development in 2008: Science

The tasks for science were operational for the first time in spring 2008. Data for these tests are included in this report.

Overview of the Technical Report

This technical report contains seven additional chapters, as follows:

- Chapter 2 describes the procedures followed to develop the CAPA tasks and to build the CAPA test forms for 2008. Characteristics of these forms also are presented in Chapter 2.
- Chapter 3 describes the scaling and equating procedures that were used.
- Chapter 4 details the procedures designed to ensure the content validity of the CAPA.
- Chapter 5 describes the kinds of score reports that are produced after each administration of the CAPA. It also summarizes the test-level analyses performed on scores obtained during the spring 2008 administration of the tests.
- Chapter 6 discusses the descriptive statistics at the task level for the operational and field-test tasks. Summaries of classical item analysis statistics, Rasch difficulty estimates, and evaluations of the Rasch model-data fit are included in Chapter 6.
- Chapter 7 highlights the importance of maintaining fairness for various CAPA subgroups. Chapter 7 summarizes demographic differences in performance, analyzing differential item functioning. Chapter 7 also includes a section describing procedures that were followed by Educational Testing Service (ETS) to ensure test security.
- Chapter 8 summarizes the reliability analyses, including test reliability and accuracy.

Each chapter contains summary tables in the body of the text. In addition, extended appendixes that report technical data for the CAPA forms are listed at the end of the relevant chapters.

Chapter 2: CAPA Development Procedures

The CAPA is constructed to measure students' achievement of the California content standards while meeting psychometric criteria for test difficulty and reliability. The psychometric criteria are evaluated using statistics from previous operational administrations or from field testing.

Test Assembly Procedures

One of the first steps in the development of a standardized test is the creation of the test blueprint. As with the other components of the STAR Program, the CAPA test blueprints were proposed by ETS, reviewed and approved by the respective Assessment Review Panels (ARPs), reviewed and approved by the California Department of Education (CDE), and presented to the SBE for adoption.

The California content standards were used as the basis for choosing tasks of the tests. The number of tasks in each cluster area was also described and made available in a the blueprint, a public document. The blueprints for the CAPA can be found on the following CDE Web pages:

http://www.cde.ca.gov/ta/tg/sr/documents/capaelablueprint.doc
http://www.cde.ca.gov/ta/tg/sr/documents/capamathblueprint.doc
http://www.cde.ca.gov/ta/tg/sr/documents/capasciblueprint.doc

A summary of the number of tasks specified in the blueprints for each cluster within content area and level is presented in the tables in Appendix 2.A.

Additional technical targets (for example, equal task difficulty and discrimination across test forms) for test construction are established on the basis of past characteristics of the tests, with the goal of maintaining parallel forms to the greatest extent possible.

Test Specifications

Statistical Specifications

The primary statistical targets used for the CAPA test assembly in 2008 are the test information function based on the item response theory (IRT) item parameters and an average polyserial correlation. The polyserial correlation is a measure of how well the tasks discriminate among test takers who differ in skill level. It is used when an interval variable is correlated with an ordinal variable that is assumed to reflect an underlying continuous variable. The polyserial correlation also is related to the overall reliability of the test. When using the Rasch model, the target information function makes it possible to choose items to produce a test that has the desired precision of measurement at all ability levels. The target mean and standard deviation (SD) of task *b*-values consistent with the information curves are also provided to test development staff to help with the test construction process.

The target statistical specifications are presented in Table 2.1, on the next page. The minimum target value for a polyserial is set at 0.60 for each test. The maximum average task (item) score (AIS) value is set at 80 percent of the score points available for each test and level, and the minimum value is set at 30 percent. The target mean *b*-value varies by test and level.

Table 2.1 Target Statistical Specifications for the CAPA

Subject	CAPA Level	Target Mean *b*	Target SD *b*	Min AIS Value	Max AIS Value	Mean AIS	Mean Polyserial	Min Polyserial
English–Language Arts	**I**	0.33	0.50	1.50	4.0	2.65	0.82	0.60
	II	–0.34	0.50	1.20	3.2	2.20	0.82	0.60
	III	0.01	0.50	1.20	3.2	2.20	0.82	0.60
	IV	–0.10	0.50	1.20	3.2	2.20	0.82	0.60
	V	0.08	0.50	1.20	3.2	2.20	0.82	0.60

Subject	CAPA Level	Target Mean *b*	Target SD *b*	Min AIS Value	Max AIS Value	Mean AIS	Mean Polyserial	Min Polyserial
Mathematics	I	–0.04	0.50	1.50	4.0	2.65	0.82	0.60
	II	0.16	0.50	1.20	3.2	2.20	0.82	0.60
	III	–0.31	0.50	1.20	3.2	2.20	0.82	0.60
	IV	–0.32	0.50	1.20	3.2	2.20	0.82	0.60
	V	–0.01	0.50	1.20	3.2	2.20	0.82	0.60
Science	I	0.14	0.50	1.50	4.0	2.65	0.82	0.60
	III	0.86	0.50	1.20	3.2	2.20	0.82	0.60
	IV	–0.64	0.50	1.20	3.2	2.20	0.82	0.60
	V	0.42	0.50	1.20	3.2	2.20	0.82	0.60

Content Specifications

ETS develops all of the CAPA test tasks to conform to the SBE-approved California content standards and the test blueprints. (See page 4 for the Web addresses of the CAPA blueprints.)

Task Development

ETS followed the SBE-approved Item Utilization Plan to guide the development of the tasks for each subject area. Task specification documents include the constructs to be measured and the California content standards included in the test blueprints. Those specifications help ensure that the CAPA tests consistently match the content standards from year to year. The task specifications also provide specific and important guidance to task writers and ensure that tasks are consistent in approach and written to measure students' achievement of the standards. 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. In summary, 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

In addition, the ELA task specifications contain guidelines for stimulus cards used to assess reading comprehension. These guidelines 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

Task Review Process

The tasks selected for each CAPA test undergo an extensive task review process that is designed to provide all California students with the best standards-based tests possible. This section summarizes the various reviews performed, which help to establish the validity of the scores from the 2008 CAPA tasks and test forms.

Internal Reviews

After the tasks have been written, ETS employs a series of internal reviews. The reviews establish the criteria used to judge the content validity of a task, making sure that each task is measuring what it is intended to measure. The internal reviews also examine the overall quality of the test 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:

- Internal content review
- Internal editorial review
- 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, 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 (or standards, for history–social science)
- 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 review 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 CDE and ARP review. 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

Assessment Review Panels (ARPs)

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 4.1. 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.

ARP Meetings for Review of CAPA Tasks

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

- Overview of the purpose and scope of the CAPA
- 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 involved more global issues, including—for ELA—the appropriateness, difficulty, and readability of reading stimulus cards. The ARPs also were trained on how to make recommendations for revising tasks. Guidelines for reviewing tasks were provided by ETS and approved by the CDE. The set of guidelines for reviewing tasks is summarized below:

Does the task:

- Measure the content standard?
- Match the test task specifications?
- Align with the construct being measured?

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

Is the stimulus (if any) for the task:

- Required in order to 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 reviewed a set of tasks independently and recorded their individual comments. The next step in the review process was for the group to discuss each task. The content-area assessment specialists facilitated the discussion and recorded all recommendations. Those recommendations were recorded in a master task-review booklet. Task review binders and other task evaluation materials also served to identify potential bias and sensitivity factors that the ARP considered as a part of its task reviews.

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

Statewide Pupil Assessment Review (SPAR) Panel

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 item 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.

Task Writer Training

ETS has developed an Item Utilization Plan to continue the development of tasks for the CAPA over the next five years. This plan includes strategies for continued coverage of all appropriate standards for all tests in each content area and levels.

Task writer training to be used for future task development was conducted over two days in Long Beach, California, in July 2008. 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.

The development of new tasks during this cycle was limited to a level that would allow for the replacement of tasks no longer available for use on operational forms. Thus, the task writers who participated were particularly experienced in writing to the standards assessed on the CAPA. All task writers met the following minimum qualifications:

- Possession of a bachelor's degree in the relevant content area or in the field of education with a special focus on a particular content of interest (An advanced degree in the relevant content area was desirable.)

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

Appendix 2.A—Test Assembly Specifications

Table. 2.A.1 2008 CAPA Test Assembly Specifications: English–Language Arts

	Number of Tasks				
	Level I	**Level II**	**Level III**	**Level IV**	**Level V**
Standard					
Reading/Word Analysis	0	1	1	1	0
Sight Word Reading	2	3	3	3	3
Reading Comprehension	2	1	2	2	3
Writing/Writing Strategies	1	2	1	1	1
Listening	2	0	0	0	0
Speaking Applications	1	1	1	1	1
Pre-test Tasks	4	4	4	4	4

Table. 2.A.2 2008 CAPA Test Assembly Specifications: Mathematics

	Number of Tasks				
	Level I	**Level II**	**Level III**	**Level IV**	**Level V**
Standard					
Number Sense	4	3	2	3	2
Counting and Money	0	0	1	1	1
Algebra and Functions	2	2	1	1	1
Measurement and Geometry	2	2	3	2	2
Statistics, Data Analysis, and Probability	0	1	1	1	2
Pre-test Tasks	4	4	4	4	4

Table. 2.A.3 2008 CAPA Test Assembly Specifications: Science

	Number of Tasks				
	Level I	**Level II**	**Level III**	**Level IV**	**Level V**
Standard					
Investigation and Experimentation	1		2	1	1
Physical Science	3		2	0	0
Life Science	2		2	0	0
Earth Science	2		2	1	2
Motion	0		0	1	0
Forces	0		0	1	0
Structure of Matter	0		0	1	0
Reactions	0		0	1	0
Periodic Table	0		0	1	0
Density and Buoyancy	0		0	1	0
Physics	0		0	0	1
Biology	0		0	0	3
Chemistry	0		0	0	1
Pre-test Tasks	4		4	4	4

Chapter 3: CAPA Equating Procedures

When test forms are created, two primary criteria must be satisfied. The first is content-based; tasks must be distributed within a test form according to content specifications. The second is statistical; tasks must have a specified distribution of difficulty or specified average difficulty and a specified average discrimination (correlation between the task score and the test score). These criteria help ensure that all forms of a test are comparable (that is, very similar in reliability and the construct that they measure). However, despite the efforts taken when a test is constructed, forms of a test will still differ in difficulty to a small degree. The equating process is used to adjust for differences in difficulty so that test takers' scores can be compared regardless of the test form they took.

Test Construction and Review

The CAPA is assembled to content and statistical specifications or targets. Each form contains some tasks that are the same as tasks used in the previous year, referred to as 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.

Post-Administration Operational Equating

The CAPA tests for ELA and mathematics are equated to the reference year using a common-item nonequivalent groups design and methods based on item response theory. The "base" or "reference" scale for the CAPA is established by the item calibrations carried out in 2007. Doing so establishes a scale to which the subsequent task calibrations can be linked. The 2008 tasks are placed on this scale through a set of common tasks that also were used in 2007.

The equating procedure for the CAPA involves three steps: task calibration, task parameter scaling, and true score equating. All three steps were completed for ELA and mathematics. Only calibration was completed for science because 2008 was the first operational year for that assessment. ETS uses a computer system called the Generalized Analysis System (GENASYS) for the IRT task equating and calibration work. As a part of this system, a proprietary version of the PARSCALE computer program (Muraki and Bock 1995) is used to estimate task parameters based on the one-parameter logistic (Rasch) model. Research at ETS has suggested that PARSCALE calibrations done in this manner produce results that are virtually identical to results based on WINSTEPS (Way, Kubiak, Henderson, and Julian 2002). The calibration procedures described below were applied to all CAPA assessments.

Science score were not scaled or equated to a previous year because no previous year's data were available. (In addition, cut scores were not available because they have not yet been established for the operational science tests.)

Calibration

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

$$P_{jk}(\theta) = \frac{\exp\left[\sum_{v=1}^{k} 1.7a_j(\theta - b_j - d_{jv})\right]}{\sum_{c=1}^{m_j} \exp\left[\sum_{v=1}^{c} 1.7a_j(\theta - b_j - d_{jv})\right]} \quad , \tag{3.1}$$

where

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

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

b_j is the difficulty of task *j*, and

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

For the task calibrations, the PARSCALE program 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. This is in keeping with previous CAPA equating and scaling procedures carried out using the WINSTEPS program (Linacre 2000). For the purpose of score equating, only the operational tasks are included for each test.

The PARSCALE calibrations were run in two stages, following procedures used with other ETS testing programs. In the first stage, estimation imposed normal constraints on the updated prior ability distribution. The estimates resulting from this first stage were used as starting values for a second PARSCALE run, in which the subject prior distribution was updated after each expectation maximization (EM) cycle with no constraints. For both stages, the metric of the scale was controlled by the constant discrimination parameters. This approach was used to obtain unscaled 2008 task parameter estimates. Each task was evaluated using fit statistics in conjunction with plots of model-data fit that were generated by the GENASYS system. Tasks flagged for potential misfit were evaluated with respect to their impact on test specifications, psychometric quality, and coverage of academic content standards.

Scaling

Calibrations of the 2008 forms in ELA and mathematics were scaled to the previously obtained reference scale estimates using linking tasks and the Stocking and Lord (1983) procedure. In the case of one-parameter model calibrations, this procedure is equivalent to setting the mean of the new task parameter estimates for the common tasks equal to the mean of the previously scaled estimates. As is commonly done in this approach, the linking process is carried out iteratively by inspecting differences between the transformed new and old (reference 2003) estimates for the linking tasks and removing tasks for which the item difficulty estimates changed significantly. Tasks with large weighted root-mean-square differences (WRMSD) between item characteristic curves (ICCs) based on the old and new difficulty estimates were removed from the linking set. The differences were calculated using the following formula:

$$WRMSD_j = \sqrt{\sum_{t=1}^{61} w_t \left[\sum_{c=1}^{m_j} \left(P_{jkn}(\theta) - P_{jkr}(\theta)\right)^2\right]} \tag{3.2}$$

where,

w_t is a weight equal to the proportion of estimated abilities from the transformed new form in score interval *t*,

$P_{jkn}(\theta)$ is the probability that an examinee with ability θ will perform in *kth* score category of task *j* on the transformed new form,

$P_{jkr}(\theta)$ is the probability that an examinee with ability θ will perform in *kth* score category of task *j* on the reference form, and

θ score intervals range from –3.0 to 3.0 in increments of 0.1.

Simply put, transformed new and old parameter estimates were evaluated using weighted (based on the reference form abilities) root mean square difference statistics that summarize differences in ICCs.

Based on established procedures, any linking items for which the WRMSD was greater than 0.625 for Level I and 0.500 for Levels II through V were eliminated. This criterion has produced reasonable results over time in similar equating work done with other testing programs at ETS. For the 2008 CAPA tests, no linking tasks were eliminated.

Table 3.1 presents, for the CAPA content area and level in ELA and mathematics, the number of common task between the 2008 (new) form and the test form to which it was linked (reference 2003), the number of tasks removed from the common task set, the correlation between the final set of new and reference difficulty estimates for the linking tasks, and the average WRMSD statistic (see equation 3.2) across the final set of common tasks.

Table 3.1 Common Tasks Between New and Reference Test Forms

Subject	Level	No. of Common Tasks	No. of Tasks Removed	Common Task Correlation	Average WRMSD
English–Language Arts	I	5	0	0.99	0.04
	II	5	0	0.98	0.07
	III	5	0	0.98	0.04
	IV	5	0	0.96	0.13
	V	5	0	0.94	0.10
Mathematics	I	5	0	0.88	0.09
	II	5	0	0.98	0.08
	III	5	0	0.99	0.05
	IV	5	0	0.99	0.14
	V	5	0	0.88	0.14

True Score Equating

Once the new calibrations for each test are transformed to the reference scale, IRT true score equating procedures are used to transform the new form number-correct scores to their respective reference form scale scores. The true score equating procedure is based on the relationship between raw scores and ability. For tests consisting entirely of *n* multiple-choice items, this is the well-known relationship defined in Lord (1980; eq. 4–5):

$$\xi(\theta) = \sum_{j=1}^{n} P_j(\theta), \tag{3.3}$$

where,

$P_j(\theta)$ is the probability of a correct response to task *j* at ability level θ (defined by the Rasch model),

ξ(θ) is the corresponding true score,

For all CAPA tests, ξ(θ) is based on *n* polytomously scored performance (constructed response) tasks[1], and the relationship can be defined as:

[1] See Chapter 5 for the scoring rubric.

$$\xi(\theta) = \sum_{j=1}^{n} \sum_{c=1}^{m_j} s_{jk} P_{jk}(\theta), \quad (3.4)$$

where,

s_{jk} is the value of the score associated with score category *k* of task j.

For Level I, there are six possible scores per task: 0, 1, 2, 3, 4, and 5. For Levels II–V there are five possible scores: 0, 1, 2, 3, and 4. A score of zero is assigned only to students who fail to respond to the prompt.

For each integer score ξ_n on the new form, the true score equating procedure first solved for the corresponding ability. Next, the procedure used that ability level to find the corresponding score ξ_r on the reference form. Finally, each score ξ_r is transformed to the appropriate CAPA scale score scale using the reference form CAPA raw-score-to-scale-score conversion tables and linear interpolation.

Complete raw-to-scale score conversion tables for the 2008 CAPA ELA and mathematics are presented in Appendix 3.A. Scale scores were truncated at both ends of the scale so that the minimum reported scale score is 15 and the maximum reported scale score is 60. These tables also display the various proficiency category cut points.

Conditional Standard Errors of Measurement (CSEMs)

As a part of the IRT-based equating procedures, scale score conversion tables and 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 tasks are located, and larger at the extremes, where there are fewer tasks. An examinee's CSEM under the IRT framework is equal to the inverse of the square root of the test information function:

$$\mathrm{CSEM}(\hat{\theta}) = \frac{1}{\sqrt{\mathrm{I}(\theta)}} a, \quad (3.5)$$

where,

$\mathrm{CSEM}(\hat{\theta})$ is the conditional standard error of measurement of the scale score

$\mathrm{I}(\theta)$ is the test information function

a is the original scaling factor needed to transform theta to the scale score metric

a , the original scaling factor, was established following the standard setting. At this time, a linear relationship was established between the cut scores in the scale score metric at the basic and proficient levels and theta values in the ability metric. The multiplicative constant of that equation is the scaling factor or a

When a test has cut scores, it is important to provide CSEMs at the cut scores. The tables in Appendix 3.A present the scale score CSEMs at the lowest score that defines the below basic, basic, proficient, and advanced levels for each CAPA test. The CSEMs tended to be higher at the advanced cut points for both ELA and mathematics. The pattern of lower values of CSEMs at the basic and proficient levels are expected because (1) more tasks tend to be of middle difficulty; and (2) tasks at the extremes still provide information toward the middle of the scale. The result is more precise scores in the middle of the scale and less precise scores in the extremes of the scale.

Equating Samples

This section describes characteristics of the samples used to establish the 2003 reference forms for ELA and mathematics as well as the equating samples used to equate the CAPA in subsequent years. Beginning in 2003, equating samples have been composed of student records in a data file obtained near the end of May. To establish the 2003 reference forms for ELA and mathematics, ETS included in the equating samples those students with valid results on the CAPA. As anticipated, these data made up from 5 to 10 percent of the total CAPA testing population. Using these smaller student samples available in late May for equating was necessary to meet score reporting deadlines.

The 2008 equating samples were made up of valid student records obtained in early June. These data consisted of approximately 17 to 33 percent of the CAPA testing data that were available in the sample received in late August (referred to as the P2[2] data). The P2 data is the basis for the information presented in the technical report, with the exception of that related to equating. The number of students in the equating sample and the P2 data are presented in Table 3.2, below. Note that the sample sizes are included for science for reference, although science scores were not equated. Again, the use of student data available at the time of equating was necessitated by score reporting deadlines and was approved by the CDE.

Table 3.2 CAPA 2008 Raw Score Means and Standard Deviations: Total P2 Population and Equating Sample

Group	Level	P2			Equating Sample			
		N	Mean RS*	SD RS*	N	% of P2	Mean RS*	SD RS*
English–Language Arts	I	11,136	27.00	11.83	1,964	18%	26.77	12.08
	II	6,482	22.87	6.17	1,583	24%	22.83	6.01
	III	6,577	22.79	6.47	1,562	24%	23.12	6.53
	IV	10,372	19.74	7.25	2,340	23%	19.90	7.20
	V	10,320	21.07	7.29	3,468	34%	20.80	7.35
Mathematics	I	11,096	22.75	11.04	1,957	18%	22.71	11.20
	II	6,466	20.73	7.57	1,578	24%	20.75	7.64
	III	6,563	21.02	7.25	1,560	24%	21.49	7.36
	IV	10,361	18.68	7.66	2,333	23%	18.88	7.62
	V	10,283	21.22	7.89	3,452	34%	20.92	8.01
Science	I	2,946	22.66	11.86	510	17%	22.81	11.97
	III	3,123	21.06	6.79	715	23%	21.71	6.73
	IV	3,436	19.70	6.50	755	22%	19.82	6.19
	V	3,366	19.31	6.62	1,121	33%	19.29	6.67

* RS = raw score

[2] P2 contains data for the schools from which answer documents were received by ETS Statistical Analysis by approximately August 29, 2008.

References

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

Lord, F. 1980. *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Muraki, E. 1992. A Generalized Partial Credit Model: Application of an EM Algorithm. *Applied Psychological Measurement,* 16: 159–76.

Muraki, E. and Bock, R. D. 1995. *PARSCALE: Parameter scaling of rating data* (Version 2.2). Chicago, IL: Scientific Software, Inc.

Stocking, M.L., and F. M. Lord 1983. "Developing a Common Metric in Task Response Theory," *Applied Psychological Measurement, 7,* 201–10.

Way, W. D.; A. T. Kubiak; D. Henderson; and M. W. Julian 2002, April. Accuracy and Stability of Calibrations for Mixed-Task-Format Tests Using the 1-Parameter and Generalized Partial Credit Models. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Appendix 3.A—New Form Conversion Tables

Table 3.A.1 Score Conversions: Level I English–Language Arts

Raw Score	Scale Score	CSEM *	Performance Level
40	60	13.8	
39	60	13.2	
38	60	8.8	
37	59	7.0	
36	57	6.1	
35	55	5.4	
34	53	5.0	**Advanced**
33	51	4.7	
32	50	4.4	
31	49	4.2	
30	48	4.1	
29	47	4.0	
28	46	3.9	
27	45	3.8	
26	44	3.8	
25	43	3.7	
24	43	3.7	
23	42	3.7	
22	41	3.7	
21	40	3.7	**Proficient**
20	39	3.8	
19	38	3.8	
18	37	3.9	
17	36	4.0	
16	35	4.1	
15	34	4.3	
14	33	4.5	
13	32	4.7	**Basic**
12	30	4.9	
11	29	5.2	
10	27	5.6	**Below Basic**
9	26	6.0	
8	25	6.4	
7	24	6.8	
6	22	7.4	
5	21	8.0	
4	19	8.8	**Far Below Basic**
3	17	9.9	
2	15	11.9	
1	15	16.6	
0	15	16.7	

* Conditional standard error of measurement

Table 3.A.2 Score Conversions: Level I Mathematics

Raw Score	Scale Score	CSEM *	Performance Level
40	60	15.4	
39	58	10.2	
38	53	6.9	
37	50	5.5	**Advanced**
36	48	4.8	
35	46	4.3	
34	45	4.0	
33	43	3.8	
32	42	3.6	
31	41	3.5	
30	40	3.4	
29	39	3.3	
28	39	3.2	**Proficient**
27	38	3.2	
26	37	3.2	
25	36	3.2	
24	35	3.2	
23	35	3.2	
22	34	3.3	
21	33	3.3	
20	32	3.4	**Basic**
19	31	3.5	
18	30	3.6	
17	29	3.8	
16	28	4.0	
15	26	4.2	
14	25	4.4	
13	24	4.7	**Below Basic**
12	23	5.0	
11	22	5.2	
10	22	5.4	
9	21	5.6	
8	20	5.8	
7	20	6.0	
6	19	6.2	
5	18	6.6	
4	17	7.1	**Far Below Basic**
3	16	7.9	
2	15	9.4	
1	15	13.0	
0	15	13.1	

* Conditional standard error of measurement

Table 3.A.3 Score Conversions: Level II English–Language Arts

Raw Score	Scale Score	CSEM *	Performance Level
32	60	17.8	
31	48	5.0	
30	45	3.5	**Advanced**
29	43	2.8	
28	41	2.5	
27	40	2.2	
26	39	2.1	
25	38	2.0	
24	38	1.9	**Proficient**
23	37	1.9	
22	36	1.9	
21	36	1.9	
20	35	1.9	
19	34	2.0	
18	33	2.0	
17	33	2.1	**Basic**
16	32	2.1	
15	31	2.2	
14	30	2.2	
13	29	2.2	
12	28	2.2	
11	27	2.2	**Below Basic**
10	26	2.3	
9	25	2.3	
8	24	2.4	
7	23	2.6	
6	21	2.7	
5	20	2.9	
4	18	3.2	**Far Below Basic**
3	16	3.5	
2	15	4.1	
1	15	5.5	
0	15	5.5	

* Conditional standard error of measurement

Table 3.A.4 Score Conversions: Level II Mathematics

Raw Score	Scale Score	CSEM *	Performance Level
32	60	11.5	**Advanced**
31	53	5.7	
30	49	3.9	
29	47	3.1	
28	46	2.7	
27	44	2.5	
26	43	2.4	
25	43	2.3	
24	42	2.2	
23	41	2.1	
22	40	2.1	**Proficient**
21	40	2.1	
20	39	2.1	
19	38	2.1	
18	38	2.1	
17	37	2.1	
16	36	2.2	
15	35	2.3	
14	34	2.3	**Basic**
13	34	2.5	
12	33	2.6	
11	31	2.8	
10	30	3.1	
9	28	3.4	**Below Basic**
8	26	3.8	
7	24	4.0	**Far Below Basic**
6	21	4.1	
5	18	4.1	
4	16	4.2	
3	15	4.5	
2	15	5.1	
1	15	6.7	
0	15	6.8	

* Conditional standard error of measurement

Table 3.A.5 Score Conversions: Level III English–Language Arts

Raw Score	Scale Score	CSEM *	Performance Level
32	60	11.5	
31	59	8.2	
30	52	6.0	**Advanced**
29	47	5.0	
28	44	4.4	
27	42	4.0	
26	40	3.6	
25	38	3.3	
24	37	3.1	**Proficient**
23	36	3.0	
22	35	2.9	
21	34	2.8	
20	33	2.8	
19	32	2.8	**Basic**
18	31	2.8	
17	30	2.8	
16	29	2.9	
15	27	3.0	
14	26	3.0	**Below Basic**
13	25	3.1	
12	23	3.2	
11	23	3.4	
10	22	3.5	
9	22	3.6	
8	21	3.8	
7	21	4.0	
6	20	4.2	
5	19	4.5	**Far Below Basic**
4	18	4.8	
3	18	5.3	
2	16	6.2	
1	15	8.3	
0	15	8.4	

* Conditional standard error of measurement

Table 3.A.6 Score Conversions: Level III Mathematics

Raw Score	Scale Score	CSEM *	Performance Level
32	60	12	
31	60	8.2	
30	55	5.8	
29	51	4.7	
28	48	4.0	**Advanced**
27	47	3.6	
26	45	3.3	
25	44	3.1	
24	43	2.9	
23	42	2.8	
22	41	2.7	
21	40	2.7	
20	39	2.7	**Proficient**
19	38	2.7	
18	37	2.7	
17	36	2.7	
16	35	2.8	
15	34	2.8	
14	33	2.9	**Basic**
13	32	3.1	
12	31	3.3	
11	29	3.6	
10	28	3.9	**Below Basic**
9	26	4.4	
8	25	4.9	
7	23	5.2	
6	22	5.3	
5	21	5.3	
4	19	5.4	**Far Below Basic**
3	18	5.7	
2	16	6.5	
1	15	8.5	
0	15	8.5	

* Conditional standard error of measurement

Table 3.A.7 Score Conversions: Level IV English–Language Arts

Raw Score	Scale Score	CSEM *	Performance Level
32	60	13.1	
31	56	6.9	
30	50	4.8	
29	48	4.0	**Advanced**
28	46	3.5	
27	44	3.3	
26	43	3.1	
25	41	2.9	
24	40	2.9	
23	39	2.8	
22	38	2.7	**Proficient**
21	37	2.7	
20	36	2.7	
19	35	2.7	
18	34	2.7	
17	33	2.7	
16	32	2.7	**Basic**
15	31	2.8	
14	30	2.9	
13	29	3.0	
12	28	3.1	**Below Basic**
11	26	3.3	
10	25	3.6	
9	23	3.9	
8	21	4.2	
7	20	4.5	
6	20	4.7	
5	19	4.9	**Far Below Basic**
4	18	5.2	
3	17	5.6	
2	16	6.4	
1	15	8.2	
0	15	7.9	

* Conditional standard error of measurement

Table 3.A.8 Score Conversions: Level IV Mathematics

Raw Score	Scale Score	CSEM *	Performance Level
32	60	16.3	
31	56	8.0	
30	50	5.3	
29	47	4.3	**Advanced**
28	45	3.7	
27	44	3.4	
26	42	3.2	
25	41	3.0	
24	40	2.9	
23	39	2.9	
22	38	2.8	**Proficient**
21	37	2.8	
20	36	2.8	
19	35	2.9	
18	34	2.9	
17	33	2.9	
16	32	3.0	**Basic**
15	31	3.1	
14	30	3.2	
13	29	3.3	
12	28	3.6	**Below Basic**
11	26	3.9	
10	24	4.4	
9	21	5.1	
8	20	5.9	
7	19	6.3	
6	18	6.1	
5	17	6.0	**Far Below Basic**
4	17	6.0	
3	16	6.2	
2	15	7.0	
1	15	9.1	
0	15	9.2	

* Conditional standard error of measurement

Table 3.A.9 Score Conversions: Level V English–Language Arts

Raw Score	Scale Score	CSEM *	Performance Level
32	60	9.9	
31	58	7.2	
30	51	5.1	
29	48	4.2	**Advanced**
28	45	3.7	
27	44	3.4	
26	42	3.1	
25	41	3.0	
24	39	2.8	
23	38	2.7	**Proficient**
22	37	2.7	
21	36	2.6	
20	35	2.5	
19	34	2.5	
18	33	2.5	
17	32	2.5	**Basic**
16	31	2.6	
15	30	2.6	
14	29	2.7	
13	28	2.8	**Below Basic**
12	27	3.0	
11	26	3.2	
10	24	3.5	
9	23	3.7	
8	23	4.0	
7	22	4.2	
6	21	4.4	
5	20	4.5	**Far Below Basic**
4	19	4.7	
3	18	5.0	
2	17	5.7	
1	15	7.5	
0	15	7.5	

* Conditional standard error of measurement

Table 3.A.10 Score Conversions: Level V Mathematics

Raw Score	Scale Score	CSEM *	Performance Level
32	60	10.3	**Advanced**
31	47	6.5	
30	43	4.4	
29	41	3.6	
28	40	3.1	**Proficient**
27	38	2.8	
26	37	2.6	
25	36	2.5	
24	36	2.4	
23	35	2.3	
22	34	2.3	**Basic**
21	33	2.2	
20	33	2.2	
19	32	2.2	
18	32	2.2	
17	31	2.3	
16	30	2.3	
15	29	2.4	**Below Basic**
14	29	2.5	
13	28	2.6	
12	27	2.8	
11	26	3.1	**Far Below Basic**
10	25	3.6	
9	24	4.3	
8	23	5.0	
7	22	5.4	
6	21	5.3	
5	19	5.1	
4	18	5.1	
3	17	5.4	
2	16	6.0	
1	15	7.9	
0	15	8	

* Conditional standard error of measurement

Chapter 4: Content Validity

This chapter summarizes evidence supporting the content validity of the CAPA. It is based on the spring 2008 test assembly process.

Validity Evidence Based on Test Content

CAPA tasks are developed to align with the content standards that are representative of the broader content domains: English–language arts, mathematics, and science. Thus, the content-related evidence of validity concerns the extent to which the test tasks represent these specified content standards.

A variety of steps are taken in the course of item development and adoption to maximize the content validity of the CAPA assessment. Items are developed by writers who have subject-area expertise and receive additional training from ETS. After development, these items are reviewed by ETS internal content-area experts. Using their expert knowledge, ETS staff review each item to evaluate the correspondence between the item's content and the standard that the item is written to measure. Item edits are made when necessary to improve this correspondence. Members of the ARP who have expertise in the subject area conduct a parallel review.

Also, for these reviews, ETS senior content staff worked directly with CDE content consultants. The CDE content consultants have extensive experience in K–12 assessments, particularly in their subject of expertise, and many are former teachers. At a minimum, each CDE content consultant holds a bachelor's degree; most have an advanced degree in their area of expertise. All ETS content and test development staff have extensive experience with K–12 assessments, experience in teaching students with a broad range of abilities, and an understanding of the California content standards. At a minimum, each holds a bachelor's degree; most ARP members have an advanced degree in their area of expertise.

Detailed information on the task and content evaluation process can also be found in Chapter 2 on page 4.

CAPA Assessment Review Panel

In addition to the thorough content reviews completed by ETS content-area experts and the CDE consultants, all CAPA tasks are reviewed by a content-area ARP. The ARPs are advisory panels to ETS on matters related to task development for the CAPA. Their credentials are presented in Table 4.1, on the next page.

Purpose

As described in Chapter 2, ETS is responsible for working with ARPs as tasks are developed for the CAPA tests. 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. ETS provides 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 the subsequent test forms. The ARPs may raise concerns in their examination of test tasks related to age- and level-appropriateness and to gender, racial/ethnic, and socioeconomic bias.

Because the ARPs are responsible for reviewing the newly developed tasks for alignment to the California content standards, they determine whether the tasks are:

- Measuring the California standards as appropriate for the CAPA testing population
- Free from bias
- Interesting and appropriate to students tested at any particular level

Composition

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 or higher degree in a level 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

School administrators, district/county content/program specialists, or university educators serving on the CAPA ARPs must meet similar qualifications:

- Three or more years of experience as a school administrator, district/county content/ program specialist, or university instructor in a level-specific area or area related to English–language arts, mathematics, or science
- Possession of a bachelor's or higher degree in a level-specific or subject area related to English–language arts, mathematics, or science
- Knowledge of and experience with the California content standards for English–language arts, mathematics, or science

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

Current ARP members are recruited through an application process. Recommendations are solicited from school districts and county offices of education as well as from to 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 4.1 shows the educational qualifications, present occupation, and credentials of the current CAPA ARP members.

Table 4.1 CAPA ARP Member Qualifications, by Subject and Total

	ELA	Math	Science		Grand Total
Total	8	6	7		21
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	4		10
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	7		16
Doctorate	0	0	0		0

	ELA	Math	Science		Grand Total
Credential (Members may hold multiple credentials.)					
Elementary Teaching (Multiple Subjects)	4	3	1		8
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	2		5
Other	0	0	0		0
None (teaching at the university level)	0	0	0		0

Currently, there are no term limits for ARP members. While most members serve on only one panel, some members serve on more than one to encourage consistency in the decisions made among the STAR testing programs. ETS and the CDE review the ARP membership annually for active participation. Members who have not attended a meeting within the past two years are notified that their invitation to participate may be withdrawn because of their lack of attendance. In addition, ETS and the CDE regularly review concerns about members whose conduct may be unprofessional and not conducive to the purpose of the ARP. If the concerns are determined to be valid, membership is revoked immediately.

CAPA Task Writers

The tasks selected for each CAPA test are written by special panels of task writers with expertise in 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. Thus, participants are particularly experienced in writing to the standards assessed on the CAPA. All task writers must meet the following minimum qualifications:

- Possession of a bachelor's or master's degree in the specified content area being tested
- Three or more years of general education teaching experience in the content areas (English–language arts, mathematics, or science)
- Knowledge about the abilities of the students taking the tests
- Knowledge and experience with California content standards in English–Language Arts, mathematics, or science.

Participants attend a general CAPA task-development training session, and then are given specific subject-area training. After viewing multiple examples of previously written CAPA tasks, participants are given task writing assignments. ETS facilitators provide feedback, and peer review methods are used to ensure the quality of the tasks.

Additional information about the task writing process is described in Chapter 2.

Chapter 5: Score Reports

This chapter describes analyses of the spring 2008 CAPA tasks and score reporting procedures. The sample used for analyses in this chapter contains the P2 data, which were available in late August.

Descriptions of Scores

Raw Score

For each CAPA test, the raw score is the total number of points a student obtains on the eight operational tasks in the test. At Level I, the highest obtainable raw score is 40; at Levels II through V, it is 32.

Scoring Rubric

For Level I ELA, mathematics, and science, all tasks are scored using a 5-point rubric. For all other levels, tasks are scored using a 4-point rubric. Both rubrics are presented in Table 5.1.

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 examiners must have completed the CAPA test examiner training. A detailed description of the test examiner requirements is available in the *CAPA Examiner's Manual*, linked on the ETS/STAR Web page at http://www.startest.org/archive.html (Outside Source).

Table 5.1 Rubrics for CAPA Scoring

Level I		Levels II–V	
Score Points	**Description**	**Score Points**	**Description**
5	Completes task successfully after initial cue and wait time.		
4	Completes task successfully after initial cues, wait time, **verbal/ gestural prompt,** and repeated cue.	4	Completes task with 100 percent accuracy.
3	Completes task successfully after initial cue, wait time, with **modeled/ physical prompt,** and repeated cue.	3	Partially completes task (scoring criteria specific to the task).
2	Attempts task after initial cue, wait time, **modeled/physical prompt**, and repeated cue.	2	Minimally completes task (scoring criteria specific to the task).
1	Orients toward task.	1	Attempts task.
0	Does not respond.	0	Does not respond.

Prompt Definitions

The following definitions are provided to clarify the vocabulary used in scoring the responses of students who require different types of prompting.

Prompt, verbal: Providing words of encouragement or phrases to help the student get started on the task (without telling the student how to complete the task or giving answers). An example of a verbal prompt is, "Pick up the crayon."

Prompt, gestural: Lightly touching the student on the shoulder to get his or her attention, gently moving the student's face to elicit eye contact with the examiner, nodding the head, or using gestures that signal messages. For example, the examiner makes a sweeping motion with his or her hand over the stimulus materials.

Prompt, modeled: To complete the task correctly for the student. For example, the test examiner picks up the correct manipulative or stimulus card, and then returns the card or manipulative to its initial position.

Prompt, physical (hand-over-hand): Modeling completion of the task, physically guiding the student to the task, or providing hand-over-hand guidance to complete the task. For example, the examiner demonstrates how to complete the task.

Scale Score

Raw scores on the CAPA for ELA and mathematics are converted to scale scores for comparison and reporting purposes. Scale scores on the CAPA range from 15 to 60.

The data in Table 5.2 and Table 5.3 present a summary of 2008 CAPA statistical information. Scale score frequency distributions for ELA and mathematics based on the spring 2008 administration of CAPA are presented in Appendix 5.A. Science scores were reported as raw scores and are presented in Appendix 5.B.

Table 5.2 Summary of 2008 CAPA Statistical Information: English–Language Arts and Mathematics

Level	I		II		III		IV		V	
Content	**ELA**	**Math**	**ELA**	**Math**	**ELA**	**Math**	**ELA**	**Math**	**ELA**	**Math**
Scale Score Information										
Number of examinees	11,136	11,096	6,482	6,466	6,577	6,563	10,372	10,361	10,320	10,283
Mean score	45.95	34.88	38.16	40.11	38.16	40.85	36.32	35.11	37.57	35.21
SD *	13.34	11.42	7.51	8.70	9.83	9.32	9.11	10.18	9.50	9.15
Possible range	15–60	15–60	15–60	15–60	15–60	15–60	15–60	15–60	15–60	15–60
Obtained range	15–60	15–60	15–60	15–60	15–60	15–60	15–60	15–60	15–60	15–60
Median	48	35	37	40	37	41	37	34	37	35
Reliability	0.93	0.91	0.84	0.88	0.87	0.88	0.88	0.88	0.89	0.88
SEM †	3.53	3.43	3.00	3.01	3.54	3.23	3.16	3.53	3.15	3.17
Raw Score Information										
Mean score	27.00	22.75	22.87	20.73	22.79	21.02	19.74	18.68	21.07	21.22
SD *	11.83	11.04	6.17	7.57	6.47	7.25	7.25	7.66	7.29	7.89
Possible range	0–40	0–40	0–32	0–32	0–32	0–32	0–32	0–32	0–32	0–32
Obtained range	0–40	0–40	0–32	0–32	0–32	0–32	0–32	0–32	0–32	0–32
Median	30	24	23	21	24	22	21	18	22	23
Reliability	0.93	0.91	0.84	0.88	0.87	0.88	0.88	0.88	0.89	0.88
SEM †	3.13	3.31	2.47	2.62	2.33	2.51	2.51	2.65	2.42	2.73
Task Information										
Number of tasks	8	8	8	8	8	8	8	8	8	8
Mean AIS ‡	3.36	2.84	2.86	2.59	2.85	2.62	2.47	2.34	2.64	2.66
SD AIS ‡	0.19	0.27	0.57	0.38	0.34	0.42	0.49	0.47	0.39	0.31
Min. AIS	2.92	2.44	2.32	1.93	2.22	2.09	1.63	1.74	1.89	2.14
Max. AIS	3.55	3.27	3.81	3.08	3.20	3.27	3.09	3.15	3.06	3.02
Possible range	0–5	0–5	0–4	0–4	0–4	0–4	0–4	0–4	0–4	0–4
Mean polyserial	0.83	0.84	0.75	0.79	0.76	0.78	0.79	0.79	0.80	0.79
SD polyserial	0.05	0.02	0.05	0.05	0.04	0.09	0.04	0.08	0.04	0.04
Min. polyserial	0.72	0.80	0.66	0.67	0.68	0.62	0.73	0.65	0.75	0.74
Max. polyserial	0.88	0.86	0.81	0.83	0.81	0.85	0.86	0.87	0.87	0.84
Mean Rasch difficulty	0.00	–0.23	–0.70	–0.04	–0.58	–0.14	–0.07	–0.13	–0.05	–0.42
SD Rasch difficulty	0.11	0.16	0.79	0.35	0.51	0.42	0.60	0.48	0.45	0.26
Min. Rasch difficulty	–0.07	–0.49	–2.21	–0.53	–1.22	–0.78	–1.09	–1.03	–0.59	–0.65
Max. Rasch difficulty	–0.26	–0.02	–0.04	–0.49	–0.22	–0.46	–0.96	–0.50	–0.75	–0.03

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

Table 5.3 Summary of 2008 CAPA Technical Characteristics: Science

Level	I	III	IV	V
Content	**Science**	**Science**	**Science**	**Science**
Raw Score Information				
Number of examinees	2,946	3,123	3,436	3,366
Mean score	22.66	21.06	19.70	19.31
SD *	11.86	6.79	6.50	6.62
Possible range	0–40	0–32	0–32	0–32
Obtained range	0–40	0–32	0–32	0–32
Median	24	22	20	20
Reliability	0.93	0.87	0.85	0.88
SEM †	3.14	2.45	2.52	2.29
Task Information				
Number of tasks	8	8	8	8
Mean AIS ‡	2.84	2.62	2.48	2.43
SD AIS ‡	0.27	0.22	0.16	0.30
Min. AIS	2.49	2.29	2.30	2.02
Max. AIS	3.38	2.88	2.78	2.80
Possible range	0–5	0–4	0–4	0–4
Mean polyserial	0.86	0.77	0.75	0.78
SD polyserial	0.02	0.05	0.03	0.03
Min. polyserial	0.81	0.70	0.69	0.74
Max. polyserial	0.88	0.82	0.80	0.84
Mean Rasch difficulty	–0.37	–1.03	–0.90	–0.47
SD Rasch difficulty	0.20	0.27	0.16	0.34
Min. Rasch difficulty	–0.73	–1.44	–1.18	–0.90
Max. Rasch difficulty	–0.08	–0.76	–0.71	0.03

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

Proficiency Levels

A student's score on each CAPA test is used to assign the student to one of the following proficiency levels:

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

For all CAPA tests for ELA and mathematics, a scale score of 35 provides the cut score separating basic performance from proficient performance, and a scale score of 30 differentiates basic performance from below basic performance. The cut scores defining the proficient/advanced and the below basic/far below basic boundaries vary slightly from test to test.

The percentage of students in each proficiency category from the total P2 sample is presented in Table 5.4 on page 33. This table provides the percentage in each category. Science is not included because performance levels had not been established when the spring 2008 administration occurred.

Table 5.4 Summary by Test Level and Subject Percentage of Examinees in Performance Levels

Subject	Test Level	Proficient/ Advanced	Advanced	Proficient	Basic	Below Basic	Far Below Basic
English–Language Arts	I	80%	57%	23%	5%	4%	11%
	II	72%	27%	45%	21%	5%	2%
	III	63%	35%	28%	21%	11%	5%
	IV	60%	31%	29%	19%	12%	9%
	V	64%	34%	30%	17%	10%	10%
Mathematics	I	55%	21%	34%	15%	15%	15%
	II	76%	45%	31%	16%	4%	4%
	III	74%	44%	30%	13%	9%	3%
	IV	49%	27%	22%	21%	14%	15%
	V	52%	21%	31%	23%	10%	16%

Purposes of Score Reporting

The tests that make up the STAR Program provide results or score summaries that are reported 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
4. Providing data for state and federal school accountability programs

Use of Score Reports

STAR program results provide parents and guardians with information about their children's progress. The results are a tool for increasing communication and collaboration between parents, guardians, and teachers. Along with teacher report cards and information from school and classroom tests, the STAR Student Reports can be used by parents and guardians to talk with teachers about ways to improve their children's achievement of the California content standards. Any discrepancies between performance reported on report cards and the scores reported on the STAR Student Report should also be discussed.

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

STAR program results help school districts and schools identify strengths and weaknesses in their instructional programs. Each year, school districts and school staffs examine STAR Program test results at each grade level and in each subject tested. Their findings are used to help determine:

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

The results from the STAR program are used for state and federal accountability programs to monitor each school's progress toward achieving established goals. STAR Program 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 and is used to rank the academic performance of schools, compare schools that have similar characteristics (such as size and ethnic makeup), identify low-performing and high-priority schools, and set yearly targets for academic growth.

STAR program results also are used to comply with federal No Child Left Behind (NCLB) 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 students, districtwide and at each school, are performing at or above the proficient level on the CSTs for English–Language Arts and Mathematics, or the CAPA.

Contents of Score Reports

The individual STAR Student Reports provide scale scores and performance-levels results for each CAPA test taken by the student for ELA and mathematics. As mentioned earlier, the scale scores range from 15 to 60, and results for the CAPA ELA and mathematics tests are also reported by performance levels: advanced, proficient, basic, below basic, or far below basic. Each performance level describes a students' level of proficiency in the content area tested.

In addition to individual student reports, several other reports are provided to different groups of stakeholders. A description of those reports is provided in Appendix 5.C.

Appendix 5.A—Scale Score Distribution Tables

Table 5.A.1 Scale Score Frequency Distributions: Level I English–Language Arts and Mathematics

Scale Score	English–Language Arts				Mathematics			
	Frequency	Percent	Cumulative Frequency	Percent Below	Frequency	Percent	Cumulative Frequency	Percent Below
60	2,617	23.50	2,617	76.50	493	4.44	493	95.56
57–59	1,111	9.98	3,728	66.52	191	1.72	684	93.84
54–56	303	2.72	4,031	63.80	–	–	–	–
51–53	744	6.68	4,775	57.12	163	1.47	847	92.37
48–50	936	8.41	5,711	48.72	581	5.24	1,428	87.13
45–47	907	8.14	6,618	40.57	670	6.04	2,098	81.09
42–44	1,000	8.98	7,618	31.59	539	4.86	2,637	76.23
39–41	655	5.88	8,273	25.71	1,710	15.41	4,347	60.82
36–38	512	4.60	8,785	21.11	1,148	10.35	5,495	50.48
33–35	477	4.28	9,262	16.83	1,396	12.58	6,891	37.90
30–32	283	2.54	9,545	14.29	895	8.07	7,786	29.83
27–29	258	2.32	9,803	11.97	510	4.60	8,296	25.23
24–26	441	3.96	10,244	8.01	522	4.70	8,818	20.53
21–23	211	1.89	10,455	6.12	664	5.98	9,482	14.55
18–20	89	0.80	10,544	5.32	728	6.56	10,210	7.98
15–17	592	5.32	11,136	0.00	886	7.98	11,096	0.00

Note: Dashes reflect scale scores that were not obtainable in 2008.

Table 5.A.2 Scale Score Frequency Distributions: Level II English–Language Arts and Mathematics

Scale Score	English–Language Arts				Mathematics			
	Frequency	Percent	Cumulative Frequency	Percent Below	Frequency	Percent	Cumulative Frequency	Percent Below
60	347	5.35	347	94.65	351	5.43	351	94.57
57–59	–	–	–	–	–	–	–	–
54–56	–	–	–	–	–	–	–	–
51–53	–	–	–	–	317	4.90	668	89.67
48–50	328	5.06	675	89.59	273	4.22	941	85.45
45–47	359	5.54	1,034	84.05	607	9.39	1,548	76.06
42–44	356	5.49	1,390	78.56	1,112	17.20	2,660	58.86
39–41	1,067	16.46	2,457	62.10	1,012	15.65	3,672	43.21
36–38	1,840	28.39	4,297	33.71	1,043	16.13	4,715	27.08
33–35	1,244	19.19	5,541	14.52	914	14.14	5,629	12.94
30–32	478	7.37	6,019	7.14	347	5.37	5,976	7.58
27–29	228	3.52	6,247	3.63	128	1.98	6,104	5.60
24–26	118	1.82	6,365	1.80	159	2.46	6,263	3.14
21–23	46	0.71	6,411	1.10	29	0.45	6,292	2.69
18–20	34	0.52	6,445	0.57	46	0.71	6,338	1.98
15–17	37	0.57	6,482	0.00	128	1.98	6,466	0.00

Note: Dashes reflect scale scores that were not obtainable in 2008.

Table 5.A.3 Scale Score Frequency Distributions: Level III English–Language Arts and Mathematics

Scale Score	English–Language Arts				Mathematics			
	Frequency	Percent	Cumulative Frequency	Percent Below	Frequency	Percent	Cumulative Frequency	Percent Below
60	195	2.96	195	97.04	324	4.94	324	95.06
57–59	276	4.20	471	92.84	–	–	–	–
54–56	–	–	–	–	306	4.66	630	90.40
51–53	434	6.60	905	86.24	464	7.07	1,094	83.33
48–50	–	–	–	–	419	6.38	1,513	76.95
45–47	488	7.42	1,393	78.82	760	11.58	2,273	65.37
42–44	887	13.49	2,280	65.33	953	14.52	3,226	50.85
39–41	409	6.22	2,689	59.12	824	12.56	4,050	38.29
36–38	1,094	16.63	3,783	42.48	648	9.87	4,698	28.42
33–35	971	14.76	4,754	27.72	637	9.71	5,335	18.71
30–32	762	11.59	5,516	16.13	417	6.35	5,752	12.36
27–29	346	5.26	5,862	10.87	315	4.80	6,067	7.56
24–26	225	3.42	6,087	7.45	276	4.21	6,343	3.35
21–23	352	5.35	6,439	2.10	107	1.63	6,450	1.72
18–20	86	1.31	6,525	0.79	48	0.73	6,498	0.99
15–17	52	0.79	6,577	0.00	65	0.99	6,563	0.00

Note: Dashes reflect scale scores that were not obtainable in 2008.

Table 5.A.4 Scale Score Frequency Distributions: Level IV English–Language Arts and Mathematics

Scale Score	English–Language Arts				Mathematics			
	Frequency	Percent	Cumulative Frequency	Percent Below	Frequency	Percent	Cumulative Frequency	Percent Below
60	192	1.85	192	98.15	360	3.47	360	96.53
57–59	–	–	–	–	–	–	–	–
54–56	258	2.49	450	95.66	335	3.23	695	93.29
51–53	–	–	–	–	–	–	–	–
48–50	690	6.65	1,140	89.01	307	2.96	1,002	90.33
45–47	420	4.05	1,560	84.96	739	7.13	1,741	83.20
42–44	1,079	10.40	2,639	74.56	726	7.01	2,467	76.19
39–41	1,568	15.12	4,207	59.44	1066	10.29	3,533	65.90
36–38	1,486	14.33	5,693	45.11	1158	11.18	4,691	54.72
33–35	1,278	12.32	6,971	32.79	1256	12.12	5,947	42.60
30–32	1,137	10.96	8,108	21.83	1331	12.85	7,278	29.76
27–29	650	6.27	8,758	15.56	956	9.23	8,234	20.53
24–26	633	6.10	9,391	9.46	974	9.40	9,208	11.13
21–23	472	4.55	9,863	4.91	321	3.10	9,529	8.03
18–20	305	2.94	10,168	1.97	515	4.97	10,044	3.06
15–17	204	1.97	10,372	0.00	317	3.06	10,361	0.00

Note: Dashes reflect scale scores that were not obtainable in 2008.

Table 5.A.5 Scale Score Frequency Distributions: Level V English–Language Arts and Mathematics

Scale Score	English–Language Arts				Mathematics			
	Frequency	Percent	Cumulative Frequency	Percent Below	Frequency	Percent	Cumulative Frequency	Percent Below
60	220	2.13	220	97.87	635	6.18	635	93.82
57–59	354	3.43	574	94.44	–	–	–	–
54–56	–	–	–	–	–	–	–	–
51–53	469	4.54	1043	89.89	–	–	–	–
48–50	583	5.65	1626	84.24	–	–	–	–
45–47	580	5.62	2206	78.62	461	4.48	1,096	89.34
42–44	1,254	12.15	3460	66.47	407	3.96	1,503	85.38
39–41	1,140	11.05	4600	55.43	1162	11.30	2,665	74.08
36–38	1,470	14.24	6070	41.18	2176	21.16	4,841	52.92
33–35	1,207	11.70	7277	29.49	1691	16.44	6,532	36.48
30–32	990	9.59	8267	19.89	1134	11.03	7,666	25.45
27–29	787	7.63	9054	12.27	1010	9.82	8,676	15.63
24–26	439	4.25	9493	8.01	820	7.97	9,496	7.65
21–23	500	4.84	9993	3.17	486	4.73	9,982	2.93
18–20	167	1.62	10160	1.55	93	0.90	10,075	2.02
15–17	160	1.55	10320	0.00	208	2.02	10,283	0.00

Note: Dashes reflect scale scores that were not obtainable in 2008.

Appendix 5.B—Raw Score Distribution Tables

Table 5.B.1 Raw Score Frequency Distributions: Level I Science

Raw Score	Frequency	Percent	Cumulative Frequency	Percent Below	Raw Score	Frequency	Percent	Cumulative Frequency	Percent Below
40	216	7.33	216	92.67	19	95	3.22	1,898	35.57
39	75	2.55	291	90.12	18	60	2.04	1,958	33.54
38	65	2.21	356	87.92	17	79	2.68	2,037	30.86
37	103	3.50	459	84.42	16	80	2.72	2,117	28.14
36	66	2.24	525	82.18	15	66	2.24	2,183	25.90
35	55	1.87	580	80.31	14	50	1.70	2,233	24.20
34	99	3.36	679	76.95	13	55	1.87	2,288	22.34
33	64	2.17	743	74.78	12	35	1.19	2,323	21.15
32	59	2.00	802	72.78	11	41	1.39	2,364	19.76
31	108	3.67	910	69.11	10	40	1.36	2,404	18.40
30	83	2.82	993	66.29	9	49	1.66	2,453	16.73
29	54	1.83	1,047	64.46	8	88	2.99	2,541	13.75
28	115	3.90	1,162	60.56	7	66	2.24	2,607	11.51
27	65	2.21	1,227	58.35	6	28	0.95	2,635	10.56
26	75	2.55	1,302	55.80	5	36	1.22	2,671	9.33
25	125	4.24	1,427	51.56	4	33	1.12	2,704	8.21
24	85	2.89	1,512	48.68	3	36	1.22	2,740	6.99
23	56	1.90	1,568	46.78	2	39	1.32	2,779	5.67
22	95	3.22	1,663	43.55	1	37	1.26	2,816	4.41
21	73	2.48	1,736	41.07	0	130	4.41	2,946	0.00
20	67	2.27	1,803	38.80					

* Level I Science raw scores are based on eight tasks common across field-test forms.

Table 5.B.2 Raw Score Frequency Distributions: Level III Science

Raw Score	Frequency	Percent	Cumulative Frequency	Percent Below	Raw Score	Frequency	Percent	Cumulative Frequency	Percent Below
32	76	2.43	76	97.57	15	101	3.23	2,601	16.71
31	123	3.94	199	93.63	14	96	3.07	2,697	13.64
30	111	3.55	310	90.07	13	75	2.40	2,772	11.24
29	136	4.35	446	85.72	12	66	2.11	2,838	9.13
28	167	5.35	613	80.37	11	51	1.63	2,889	7.49
27	151	4.84	764	75.54	10	53	1.70	2,942	5.80
26	164	5.25	928	70.28	9	30	0.96	2,972	4.84
25	167	5.35	1,095	64.94	8	42	1.34	3,014	3.49
24	163	5.22	1,258	59.72	7	17	0.54	3,031	2.95
23	157	5.03	1,415	54.69	6	17	0.54	3,048	2.40
22	182	5.83	1,597	48.86	5	17	0.54	3,065	1.86
21	163	5.22	1,760	43.64	4	14	0.45	3,079	1.41
20	156	5.00	1,916	38.65	3	9	0.29	3,088	1.12
19	159	5.09	2,075	33.56	2	7	0.22	3,095	0.90
18	158	5.06	2,233	28.50	1	5	0.16	3,100	0.74
17	134	4.29	2,367	24.21	0	23	0.74	3,123	0.00
16	133	4.26	2,500	19.95					

* Level III Science raw scores are based on eight tasks common across field-test forms.

Table 5.B.3 Raw Score Frequency Distributions: Level IV Science

Raw Score	Frequency	Percent	Cumulative Frequency	Percent Below	Raw Score	Frequency	Percent	Cumulative Frequency	Percent Below
32	52	1.51	52	98.49	15	168	4.89	2,726	20.66
31	54	1.57	106	96.92	14	125	3.64	2,851	17.03
30	78	2.27	184	94.64	13	118	3.43	2,969	13.59
29	98	2.85	282	91.79	12	82	2.39	3,051	11.20
28	122	3.55	404	88.24	11	86	2.50	3,137	8.70
27	139	4.05	543	84.20	10	65	1.89	3,202	6.81
26	150	4.37	693	79.83	9	54	1.57	3,256	5.24
25	196	5.70	889	74.13	8	61	1.78	3,317	3.46
24	173	5.03	1,062	69.09	7	22	0.64	3,339	2.82
23	175	5.09	1,237	64.00	6	26	0.76	3,365	2.07
22	195	5.68	1,432	58.32	5	11	0.32	3,376	1.75
21	189	5.50	1,621	52.82	4	13	0.38	3,389	1.37
20	184	5.36	1,805	47.47	3	5	0.15	3,394	1.22
19	224	6.52	2,029	40.95	2	8	0.23	3,402	0.99
18	191	5.56	2,220	35.39	1	9	0.26	3,411	0.73
17	183	5.33	2,403	30.06	0	25	0.73	3,436	0.00
16	155	4.51	2,558	25.55					

* Level IV Science raw scores are based on eight tasks common across field-test forms.

Table 5.B.4 Raw Score Frequency Distributions: Level V Science

Raw Score	Frequency	Percent	Cumulative Frequency	Percent Below	Raw Score	Frequency	Percent	Cumulative Frequency	Percent Below
32	49	1.46	49	98.54	15	127	3.77	2661	20.94
31	32	0.95	81	97.59	14	109	3.24	2770	17.71
30	67	1.99	148	95.60	13	84	2.50	2854	15.21
29	66	1.96	214	93.64	12	84	2.50	2938	12.72
28	106	3.15	320	90.49	11	69	2.05	3007	10.67
27	126	3.74	446	86.75	10	60	1.78	3067	8.88
26	138	4.10	584	82.65	9	57	1.69	3124	7.19
25	157	4.66	741	77.99	8	91	2.70	3215	4.49
24	189	5.61	930	72.37	7	20	0.59	3235	3.89
23	194	5.76	1,124	66.61	6	17	0.51	3252	3.39
22	213	6.33	1,337	60.28	5	17	0.51	3269	2.88
21	219	6.51	1,556	53.77	4	10	0.30	3279	2.58
20	236	7.01	1,792	46.76	3	13	0.39	3292	2.20
19	198	5.88	1,990	40.88	2	14	0.42	3306	1.78
18	187	5.56	2,177	35.32	1	13	0.39	3319	1.40
17	216	6.42	2,393	28.91	0	47	1.40	3366	0.00
16	141	4.19	2,534	24.72					

* Level V Science raw scores are based on eight tasks common across field-test forms.

Appendix 5.C—Types of Score Reports Tables

Table 5.C.1 Score Reports Reflecting CAPA Results

2008 STAR CAPA Student Reports	
Description	**Distribution**
The CAPA Student Report	
This report provides parents/guardians and teachers with the student's results, presented in tables and graphs. Data presented include: • Scale scores for ELA and mathematics • Performance levels for ELA and mathematics • Percent correct for science • Descriptions of the performance levels for ELA and mathematics	Because this report includes individual student results, it is not distributed beyond the student's school. Two color 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 district.
Student Record Label	
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 California *Education Code* Section 60607 [a]). Significant information includes: • Scale scores and performance levels (ELA and Mathematics) • Percent correct (science)	Because this report includes individual student results, it is not distributed beyond the student's school.
Student Master List	
This report is an alphabetical roster of individual student results. It mainly includes: • A scale score and a performance level (ELA and Mathematics) • Percent correct (science)	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. Because this report includes individual student results, it is not distributed beyond the student's school.
Student Master List Summary	
This report summarizes student results at the school, district, county, and state levels for each levels. It does *not* include any individual student information. The following data are summarized by subject tested: • Number of students enrolled, number and percent of students tested, number and percent of valid scores, and number tested with scores • Mean percent correct, mean scale score, and scale score standard deviation for each subject area tested	This report is a resource for evaluators, researchers, teachers, parents/guardians, community members, and administrators. One copy is sent to the school and one to the district. This report is also produced for districts, counties, and the state. ***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.

2008 STAR CAPA Student Reports	
Description	**Distribution**
• Number and percent of students scoring at each performance level (ELA and mathematics) • Percent correct for science	
Subgroup Summary	
This set of reports disaggregates and reports results by the following subgroups: • All students • Disability status ***Note:*** Disabilities among CAPA students include specific disabilities. • Economic status • Gender • English proficiency • Primary ethnicity 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. For each subgroup within a report, and for the total number of students, the following is included: • 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 score (ELA and mathematics) • Standard deviation of scale score (ELA and mathematics) • Number and percent of students scoring at each CAPA performance level (ELA and mathematics) • Percent correct for science	This report is a resource for evaluators, researchers, teachers, parents/guardians, community members, and administrators. One copy is sent to the school and one to the district. This report is also produced for districts, counties, and the state. ***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.

2008 STAR CAPA Student Reports	
Description	**Distribution**
Subgroup Summary—Ethnicity for Economic Status	
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. For each subgroup within a report, and for the total number of students, the following are included: • 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 (ELA and mathematics only) • Standard deviation of scale score (ELA and mathematics only) • Number and percent of students scoring at each performance level (ELA and mathematics only)	This report is a resource for evaluators, researchers, teachers, parents/guardians, community members, and administrators. One copy is sent to the school and one copy to the district. This report is also produced for districts, counties, and the state. ***Note:*** The data on this report may be shared with parents/guardians, community members, and the media only if the data are for 11 or more students.

Chapter 6: Task Descriptive Statistics

This chapter provides statistics obtained for this assessment at the task level and information about the students who participated in the spring 2008 CAPA administration. The statistics presented include classical and IRT results.

The chapter is divided into three sections that cover the following:

1. Student participation
2. Classical task-level analyses, including average score on task (AIS) and polyserial correlations for each operational item
3. Summaries of Rasch model item difficulty statistics (*b*-values) for operational and field-test items and summaries of item classifications based on the fit of the data to the Rasch model

Participation

In 2008, a total of 44,887 students in grades two through eleven participated in the CAPA. Table 6.1 displays the number of students by level in the P2 data received in late August that were used for the analysis.

Table 6.1 Distribution of Students Across CAPA Test Levels

Test Level	Frequency	Percent	Cumulative Frequency	Cumulative Percent
I	11,136	24.8	11,136	24.8
II	6,482	14.4	17,618	39.2
III	6,577	14.7	24,195	53.9
IV	10,372	23.1	34,567	77.0
V	10,320	23.0	44,887	100.0

Table 6.2 summarizes information about the test forms and examinees included in the task analyses, including the numbers of test forms, operational tasks, field-test tasks, and the approximate number of students taking both operational and field-test tasks in the P2 sample. The sample sizes for the field tests are presented as a range because not all students were administered each field-test task. The values given are from the smallest number of students administered any one field-test task in a designated level and content area to the largest.

Table 6.2 Number of Items, Sample Size, and Forms Presented for the CAPA, 2008

Subject	Level	Operational		Field Test		
		# Items	Examinees Total (P2)	# Forms	# Items	Examinees Total (P2)
English–Language Arts	**I**	8	11,136	8	4	11,081–11,104
	II	8	6,482	8	4	6,395–6,471
	III	8	6,577	8	4	6,537–6,566
	IV	8	10,372	8	4	10,304–10,364
	V	8	10,320	8	4	10,309–10,316
Mathematics	**I**	8	11,096	8	4	11,047–11,059
	II	8	6,466	8	4	6,377–6,441
	III	8	6,563	8	4	6,501–6,534
	IV	8	10,361	8	4	10,287–10,335
	V	8	10,283	8	4	10,199–10,253
Science	**I**	8	2,946	8	4	2,877–2,942
	III	8	3,123	8	4	3,006–3,098
	IV	8	3,436	8	4	3,408–3,433
	V	8	3,366	8	4	3,276–3,312

Additional information about participation is presented in Appendix 6.A, which contains tables showing the number and percent of examinees who received each possible score point within each content area and test level.

Task Analyses

Statistics calculated for the tasks in the CAPA operational and field-test analyses are described as follows.

AIS

For polytomously scored tasks, this statistic indicates the average rating earned on the task. Desired values generally fall within the range of 30 percent to 80 percent of the maximum task score. Occasionally, tasks that fall outside this range can be justified for inclusion in an item bank or a test form on the basis of the quality and educational importance of the task content or to better measure 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 performance on the specific task and performance on the total test. The polyserial correlation is used when an interval variable is correlated with an ordinal variable that is assumed to reflect an underlying continuous variable.

Polyserial correlations are based on a polyserial regression model (Drasgow 1988). The model assumes that performance on a task and, thus, the item score Y, is determined by the examinees' position on an underlying latent variable η, which represents the examinee's ability to perform the task required by that item. The distribution of η for candidates with a given score x is assumed to be normal with mean = βx, where β is an item parameter to be estimated from the data. The model can be written as follows:

$$P(Y \le y_j \mid x) = P(\eta \le \alpha_j \mid x) = \Phi(\alpha_j - \beta x) \quad (6.1)$$

where:

y_j is the *j*th possible score on the item,

α_j is the value of η corresponding to y_j and

Φ is the unit normal cumulative distribution function.

The ETS proprietary software 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:

$$r_{polyreg} = \frac{\beta\sigma_{tot}}{\sqrt{\beta^2\sigma^2_{tot} + 1}} \quad (6.2)$$

where:

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

β is the item parameter to be estimated from the data using maximum likelihood.

As shown in the polyserial correlation formula, 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 this item.

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 .60. This value is higher than the minimum acceptable point biserial used with dichotomous items because the number of tasks is small and they are polytomous.

Appendix 6.B presents, for each item in the 2008 administration, the AIS and polyserial correlation. Some items were flagged for unusual statistics, and these flags are shown in the tables.

There are three types of flags. Although the flag definition appears in the headings 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

Differential item functioning (DIF) analyses are also performed on all operational items and all field-test items for which sufficient student samples are available. (See Chapter 7 for further discussion of DIF analysis.)

IRT Analyses

Summaries of IRT b-values

Table 6.3, Table 6.4, and Table 6.5 present the number of operational and field-test items and summary statistics for the IRT *b*-values after the scaling was completed.

Table 6.3 IRT *b*-values for English–Language Arts by Level

Level		Number of Items	Mean	Standard Deviation	Min	Max
I	All Operational Items	8	0.00	0.11	–0.07	0.26
	Field-Test Items	24	0.09	0.27	–0.49	0.43
II	All Operational Items	8	–0.70	0.79	–2.21	–0.04
	Field-Test Items	16	–0.60	0.42	–1.32	–0.09
III	All Operational Items	8	–0.58	0.51	–1.22	0.22
	Field-Test Items	16	–0.70	0.49	–1.50	–0.01
IV	All Operational Items	8	–0.07	0.60	–1.09	0.96
	Field-Test Items	28	–0.41	0.54	–1.40	1.00
V	All Operational Items	8	–0.05	0.45	–0.59	0.75
	Field-Test Items	24	–0.14	0.44	–0.98	0.59

Table 6.4 IRT b-values for Mathematics by Level

Level		Number of Items	Mean	Standard Deviation	Min	Max
I	All Operational Items	8	–0.23	0.16	–0.49	–0.02
	Field-Test Items	24	–0.21	0.16	–0.55	0.09
II	All Operational Items	8	–0.04	0.35	–0.53	0.49
	Field-Test Items	16	–0.22	0.60	–0.93	1.57
III	All Operational Items	8	–0.14	0.42	–0.78	0.46
	Field-Test Items	16	–0.13	0.47	–0.95	0.63
IV	All Operational Items	8	–0.13	0.48	–1.03	0.50
	Field-Test Items	28	–0.24	0.46	–1.14	0.51
V	All Operational Items	8	–0.42	0.26	–0.65	0.03
	Field-Test Items	24	–0.47	0.33	–1.06	0.37

Table 6.5 IRT b-values for Science by Level

Level		Number of Items	Mean	Standard Deviation	Min	Max
I	All Operational Items	8	–0.37	0.20	–0.73	–0.08
	Field-Test Items	8	–0.57	0.27	–0.98	–0.26
III	All Operational Items	8	–1.03	0.27	–1.44	–0.76
	Field-Test Items	8	–1.22	0.55	–2.01	–0.35
IV	All Operational Items	8	–0.90	0.16	–1.18	–0.71
	Field-Test Items	8	–1.09	0.26	–1.49	–0.76
V	All Operational Items	8	–0.47	0.34	–0.90	0.03
	Field-Test Items	8	–0.67	0.47	–1.66	0.01

IRT Model-Data Fit Analyses

Because the Rasch model is used in scaling and equating the CAPA, an important part of IRT task analyses is the assessment of model-data fit. ETS statisticians classified operational and field-test tasks for the CAPA into discrete categories on the basis of an evaluation of how well each task was fit by the Rasch model. The flagging procedure has categories of A, B, C, D, and F, which are assigned on the basis of an evaluation of graphical model-data fit information. Descriptors for each category are as follows.

Flag A

- Good fit of theoretical curves to empirical data along the entire ability range for all categories; may have some small divergence at the extremes
- Small Chi-square value relative to the other tasks in the calibration with similar sample sizes

Flag B

- Theoretical curves within error range across most of ability range for most categories; may have some small divergence at the extremes
- Acceptable Chi-square value relative to the other tasks in the calibration with similar sample sizes

Flag C

- Theoretical curves within error range at some regions and slightly outside of error range at remaining regions of ability range for some categories
- Moderate Chi-square value relative to the other tasks in the calibration with similar sample sizes
- Often applies to tasks that appear to be functioning well but are not well fit by the Rasch model

Flag D

- Theoretical curves outside of error range at some regions across ability range for most categories
- Large Chi-square value relative to the other tasks in the calibration with similar sample sizes

Flag F

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

In general, tasks with flagging categories of A, B, or C are all considered acceptable. Ratings of D are considered questionable—test developers are asked to avoid these tasks if possible and to carefully review them if they must be used. Test developers are instructed to avoid using tasks rated F for operational test assembly without a review by a psychometrician. In some situations in which the available task pool is small, the use of an item having an IRT fit flag of F may not be avoidable.

A summary that includes all CAPA levels of the results of the IRT model-data fit classifications is presented in Table 6.6.

Table 6.6 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	7	3	3
B	85	62	24
C	54	75	34
D	2	6	3
F	0	2	0

The tables in Appendix 6.C also display the number of items in each fit classification by level for each content area.

Reference

Drasgow F. 1988. “Polychoric and Polyserial Correlations,” in *Encyclopedia of Statistical Sciences*. Edited by L. Kotz and N. L. Johnson. New York: Wiley, 7: 69–74.

Appendix 6.A—Frequency of Operational Task Scores Tables

Table 6.A.1 Frequency of Operational Task Scores: English–Language Arts

ELA Level	Score on Task	1		2		3		4		5		6		7		8	
		Count	Percent	Count	Percent	Count	Percent	Count	Percent	Count	Percent	Count	Percent	Count	Percent	Count	Percent
I	0	999	8.98	1,074	9.68	1,078	9.70	1,121	10.10	1,002	9.01	1,049	9.45	1,040	9.37	977	8.82
	1	1,220	10.97	1,439	12.97	1,337	12.03	1,134	10.22	1,152	10.36	1,211	10.91	1,200	10.81	1,205	10.88
	2	1,469	13.21	2,891	26.05	1,701	15.31	1,029	9.27	1,708	15.37	1,491	13.43	1,551	13.98	2,007	18.11
	3	748	6.72	1,066	9.61	786	7.07	723	6.51	813	7.31	726	6.54	711	6.41	761	6.87
	4	1,181	10.62	861	7.76	984	8.85	1,172	10.56	989	8.90	878	7.91	884	7.97	960	8.66
	5	5,507	49.51	3,767	33.94	5,228	47.04	5,919	53.33	5,452	49.05	5,743	51.75	5,710	51.46	5,170	46.66
II	0	38	0.56	208	3.08	197	2.92	192	2.84	66	0.98	149	2.21	169	2.50	144	2.14
	1	128	1.89	2,060	30.46	1,911	28.31	1,627	24.09	205	3.03	1,534	22.71	669	9.90	1,657	24.65
	2	194	2.87	1,231	18.20	1,251	18.53	1,354	20.05	403	5.96	1,827	27.05	1,229	18.19	2,029	30.19
	3	422	6.24	1,647	24.35	664	9.84	488	7.23	982	14.52	1,162	17.20	1,310	19.39	1,622	24.13
	4	5,980	88.44	1,617	23.91	2,728	40.41	3,092	45.79	5,108	75.52	2,083	30.84	3,379	50.01	1,269	18.88
III	0	102	1.50	346	5.09	194	2.86	191	2.82	77	1.13	129	1.90	143	2.11	119	1.77
	1	455	6.70	1,062	15.63	682	10.05	1,132	16.69	371	5.46	1,388	20.44	1,063	15.66	1,007	15.00
	2	833	12.27	1,809	26.63	1,302	19.19	3,069	45.25	706	10.39	1,116	16.43	877	12.92	1,052	15.67
	3	2,064	30.40	2,031	29.89	2,319	34.18	1,651	24.34	2,569	37.82	457	6.73	1,565	23.05	971	14.46
	4	3,336	49.13	1,546	22.76	2,287	33.71	740	10.91	3,070	45.19	3,701	54.50	3,142	46.27	3,565	53.10
IV	0	107	1.01	932	8.85	520	4.95	505	4.79	577	5.47	330	3.14	269	2.56	454	4.35
	1	1,402	13.29	5,655	53.70	1,322	12.58	2,336	22.18	2,820	26.76	4,835	46.02	1,972	18.74	2,668	25.56
	2	1,593	15.10	1,580	15.00	1,089	10.36	1,801	17.10	2,471	23.44	1,524	14.51	2,766	26.29	1,552	14.87
	3	1,812	17.17	1,138	10.81	1,919	18.26	2,913	27.65	2,352	22.31	1,640	15.61	2,684	25.51	1,313	12.58
	4	5,638	53.43	1,225	11.63	5,662	53.86	2,979	28.28	2,320	22.01	2,177	20.72	2,832	26.91	4,451	42.64
V	0	219	2.11	367	3.54	411	3.97	264	2.55	222	2.14	419	4.06	558	5.40	396	3.85
	1	1,889	18.21	1,839	17.75	1,678	16.22	1,216	11.75	1,422	13.73	4,981	48.22	1,561	15.10	3,625	35.26
	2	1,233	11.88	1,773	17.11	2,233	21.58	1,905	18.40	1,944	18.76	1,547	14.98	1,289	12.47	1,322	12.86
	3	544	5.24	3,568	34.44	3,532	34.14	2,654	25.64	2,706	26.12	2,063	19.97	2,076	20.08	2,496	24.28
	4	6,491	62.56	2,814	27.16	2,493	24.09	4,314	41.67	4,066	39.25	1,319	12.77	4,853	46.95	2,442	23.75

Table 6.A.2 Frequency of Operational Task Scores: Mathematics

Math Level	Score on Task	1		2		3		4		5		6		7		8	
		Count	Percent	Count	Percent	Count	Percent	Count	Percent	Count	Percent	Count	Percent	Count	Percent	Count	Percent
I	0	1,106	9.98	1,137	10.28	1,114	10.08	1,109	10.04	1,097	9.91	1,263	11.42	1,250	11.31	1,236	11.27
	1	1,296	11.70	1,390	12.57	1,437	13.00	1,460	13.21	1,636	14.78	1,782	16.12	1,983	17.94	1,380	12.58
	2	1,922	17.35	2,613	23.62	3,352	30.32	3,070	27.79	4,466	40.35	2,726	24.66	3,294	29.80	2,346	21.39
	3	770	6.95	942	8.52	938	8.49	900	8.15	992	8.96	932	8.43	1,075	9.72	814	7.42
	4	907	8.19	994	8.99	952	8.61	1,062	9.61	790	7.14	896	8.10	940	8.50	894	8.15
	5	5,076	45.82	3,985	36.03	3261	29.50	3,448	31.21	2,088	18.86	3,456	31.26	2,512	22.72	4,296	39.18
II	0	222	3.29	263	3.90	210	3.12	260	3.86	205	3.05	277	4.11	210	3.13	139	2.08
	1	868	12.86	2,107	31.24	1572	23.32	3,726	55.27	1,275	18.94	2,723	40.45	1,299	19.34	1,016	15.22
	2	791	11.72	702	10.41	1408	20.89	575	8.53	1,927	28.63	843	12.52	1,012	15.07	1,081	16.19
	3	1,115	16.51	674	9.99	1161	17.23	504	7.48	1,864	27.69	753	11.19	1,055	15.71	904	13.54
	4	3,756	55.63	2,999	44.46	2389	35.45	1,676	24.86	1,460	21.69	2,136	31.73	3,141	46.76	3,536	52.97
III	0	129	1.90	175	2.58	169	2.49	161	2.37	144	2.12	233	3.44	126	1.86	183	2.71
	1	1,452	21.38	1,628	24.03	963	14.21	2,089	30.80	1,701	25.07	2,565	37.87	863	12.74	2,626	38.82
	2	739	10.88	2,850	42.06	848	12.51	527	7.77	1,815	26.75	696	10.28	543	8.01	1,292	19.10
	3	724	10.66	1,519	22.42	1417	20.91	652	9.61	1,890	27.86	790	11.66	666	9.83	796	11.77
	4	3,746	55.17	604	8.91	3379	49.87	3,353	49.44	1,235	18.20	2,489	36.75	4,578	67.56	1,868	27.61
IV	0	365	3.46	586	5.57	249	2.37	367	3.50	480	4.56	194	1.85	265	2.52	420	4.00
	1	4,124	39.13	6,340	60.24	4761	45.29	2,853	27.18	5,306	50.44	1,224	11.68	4,019	38.24	3,388	32.30
	2	988	9.38	885	8.41	2154	20.49	449	4.28	1,048	9.96	1,351	12.89	1,313	12.49	1,887	17.99
	3	2,304	21.86	678	6.44	864	8.22	842	8.02	938	8.92	1,848	17.63	1,111	10.57	3,051	29.08
	4	2,757	26.16	2,036	19.34	2485	23.64	5,984	57.02	2,748	26.12	5,865	55.95	3,802	36.18	1,744	16.63
V	0	244	2.36	353	3.42	253	2.47	244	2.37	288	2.79	421	4.09	319	3.10	366	3.59
	1	2,630	25.45	2,099	20.34	2676	26.11	1,624	15.74	4,099	39.74	3,793	36.85	3,994	38.84	2,195	21.54
	2	841	8.14	548	5.31	1291	12.60	3,077	29.83	1,297	12.57	643	6.25	2,066	20.09	512	5.03
	3	1,042	10.08	1,725	16.72	1284	12.53	1,539	14.92	1,008	9.77	886	8.61	1,602	15.58	797	7.82
	4	5,578	53.97	5,593	54.21	4744	46.29	3,832	37.15	3,623	35.12	4,549	44.20	2,303	22.39	6,318	62.01

Table 6.A.3 Frequency of Operational Task Scores: Science

Science Level	Score on Task	1		2		3		4		5		6		7		8	
		Count	Percent	Count	Percent	Count	Percent	Count	Percent	Count	Percent	Count	Percent	Count	Percent	Count	Percent
I	0	432	12.31	457	13.09	443	12.73	423	12.20	381	10.91	527	15.11	466	13.43	356	10.69
	1	567	16.16	487	13.95	433	12.44	442	12.75	521	14.92	589	16.89	578	16.65	406	12.19
	2	823	23.45	899	25.76	795	22.84	857	24.72	840	24.05	979	28.08	928	26.74	448	13.45
	3	284	8.09	289	8.28	258	7.41	267	7.70	286	8.19	283	8.12	309	8.90	204	6.13
	4	226	6.44	269	7.71	293	8.42	216	6.23	297	8.51	255	7.31	246	7.09	233	7.00
	5	1,177	33.54	1,089	31.20	1,259	36.17	1,262	36.40	1,167	33.42	854	24.49	944	27.20	1,683	50.54
III	0	73	2.18	75	2.24	57	1.71	161	4.82	116	3.47	110	3.30	83	2.49	95	2.87
	1	466	13.91	433	12.95	283	8.48	687	20.58	1,182	35.35	853	25.55	867	26.05	482	14.57
	2	397	11.85	713	21.32	786	23.54	600	17.97	564	16.87	400	11.98	983	29.54	597	18.04
	3	1,603	47.85	696	20.81	1,159	34.71	798	23.90	268	8.01	960	28.76	708	21.27	1,154	34.87
	4	811	24.21	1,427	42.67	1054	31.57	1,093	32.73	1,214	36.30	1,015	30.41	687	20.64	981	29.65
IV	0	79	2.10	83	2.21	104	2.79	127	3.41	114	3.05	102	2.73	136	3.65	128	3.46
	1	1,094	29.10	764	20.37	1,204	32.25	962	25.84	692	18.50	697	18.68	682	18.32	915	24.72
	2	586	15.59	1,109	29.57	660	17.68	1,088	29.22	675	18.05	830	22.24	791	21.25	868	23.45
	3	1,026	27.29	1,025	27.33	948	25.40	746	20.04	871	23.29	1,102	29.53	1,382	37.12	983	26.56
	4	975	25.93	769	20.51	817	21.89	800	21.49	1,388	37.11	1,001	26.82	732	19.66	807	21.80
V	0	111	2.86	194	5.06	117	3.06	199	5.19	129	3.34	132	3.43	137	3.56	121	3.19
	1	1,037	26.75	1,028	26.79	611	15.97	1,276	33.26	478	12.39	546	14.21	1,367	35.54	700	18.44
	2	929	23.97	1,029	26.82	696	18.20	1,046	27.27	855	22.16	637	16.58	879	22.85	637	16.78
	3	1,257	32.43	938	24.45	1,505	39.35	745	19.42	1,224	31.73	1,814	47.20	749	19.47	735	19.36
	4	542	13.98	648	16.89	896	23.42	570	14.86	1,172	30.38	714	18.58	714	18.56	1,603	42.23

Appendix 6.B—Task Statistics Tables

Table 6.B.1 2008 CAPA Task Statistics: Level I

2008 CAPA Task Statistics: Level I English–Language Arts			
Version/Field-Test Form	**Task Position**	**AIS**	**Polyserial**
Operational	1	3.45	.84
1/7 *	2	3.78	.65
Operational	3	2.92	.80
Operational	4	3.32	.86
1/7 *	5	3.93	.75
Operational	6	3.55	.72
Operational	7	3.41	.86
1/7 *	8	2.77	.80
Operational	9	3.46	.88
Operational	10	3.45	.88
1/7 *	11	2.73	.74
Operational	12	3.34	.82
2/8 *	2	3.14	.75
2/8 *	5	3.27	.82
2/8 *	8	3.51	.71
2/8 *	11	2.94	.78
3	2	3.76	.69
3	5	2.95	.80
3	8	3.33	.82
3	11	3.82	.74
4	2	3.17	.79
4	5	2.96	.82
4	8	3.22	.74
4	11	3.44	.75
5	2	3.76	.74
5	5	2.73	.79
5	8	2.59	.74
5	11	2.79	.79
6	2	3.69	.72
6	5	3.58	.81
6	8	2.95	.74
6	11	3.32	.80

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

2008 CAPA Task Statistics: Level I Mathematics			
Version/Field -Test Form	**Task Position**	**AIS**	**Polyserial**
Operational	13	3.27	.85
1/7 *	14	2.50	.74
Operational	15	3.00	.84
Operational	16	2.79	.80
1/7 *	17	3.12	.77
Operational	18	2.86	.82
Operational	19	2.44	.84
1/7 *	20	2.82	.81
Operational	21	2.77	.85
Operational	22	2.52	.83
1/7 *	23	2.80	.82
Operational	24	3.04	.86
2/8 *	14	3.04	.78
2/8 *	17	2.70	.77
2/8 *	20	3.23	.78
2/8 *	23	2.99	.76
3	14	3.05	.76
3	17	3.05	.78
3	20	3.06	.78
3	23	2.58	.76
4	14	2.53	.76
4	17	2.92	.79
4	20	2.75	.81
4	23	2.57	.80
5	14	2.60	.71
5	17	2.78	.82
5	20	2.53	.78
5	23	2.65	.80
6	14	3.06	.77
6	17	2.96	.72
6	20	2.71	.79
6	23	2.93	.81

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

2008 CAPA Task Statistics: Level I Science			
Version/Field-Test Form	**Task Position**	**AIS**	**Polyserial**
Operational	25	2.77	.87
1/3/5/7 *	26	2.52	.69
Operational	27	2.77	.85
Operational	28	2.94	.85
1/3/5/7 *	29	3.35	.79
Operational	30	2.90	.88
Operational	31	2.89	.81
1/3/5/7 *	32	2.86	.83
Operational	33	2.49	.86
Operational	34	2.59	.87
1/3/5/7 *	35	2.99	.86
Operational	36	3.38	.86
2/4/6/8 *	26	3.71	.72
2/4/6/8 *	29	3.09	.82
2/4/6/8 *	32	2.78	.85
2/4/6/8 *	35	3.68	.71

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

Table 6.B.2 2008 CAPA Task Statistics: Level II

2008 CAPA Task Statistics: Level II English–Language Arts				
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	3.81	.66	H
1/5 *	2	3.36	.69	H
Operational	3	2.36	.77	
Operational	4	2.56	.80	
1/5 *	5	2.30	.59	R
Operational	6	2.68	.81	
Operational	7	3.61	.74	H
1/5 *	8	2.99	.60	R
Operational	9	2.51	.77	
Operational	10	3.04	.76	
1/5 *	11	3.22	.63	H
Operational	12	2.32	.69	
2/6 *	2	2.83	.67	
2/6 *	5	2.78	.67	
2/6 *	8	3.20	.64	
2/6 *	11	2.34	.59	R
3/7 *	2	3.01	.67	
3/7 *	5	2.63	.51	R
3/7 *	8	3.50	.65	H
3/7 *	11	2.58	.68	
4/8 *	2	2.89	.64	
4/8 *	5	2.53	.67	
4/8 *	8	2.38	.69	
4/8 *	11	3.51	.61	H

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

2008 CAPA Task Statistics: 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	3.08	.80	
1/5 *	14	3.06	.68	
Operational	15	2.58	.83	
Operational	16	2.56	.80	
1/5 *	17	2.37	.51	R
Operational	18	1.93	.79	
Operational	19	2.46	.67	
1/5 *	20	3.09	.74	
Operational	21	2.25	.83	
Operational	22	2.83	.80	
1/5 *	23	2.93	.68	
Operational	24	2.99	.76	
2/6 *	14	2.72	.65	
2/6 *	17	3.27	.66	H
2/6 *	20	2.61	.66	
2/6 *	23	3.43	.75	H
3/7 *	14	2.67	.59	R
3/7 *	17	2.05	.82	
3/7 *	20	2.36	.68	
3/7 *	23	1.24	.46	R
4/8 *	14	3.37	.76	H
4/8 *	17	2.70	.78	
4/8 *	20	3.03	.65	
4/8 *	23	2.92	.83	

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

Table 6.B.3 2008 CAPA Task Statistics: Level III

2008 CAPA Task Statistics: Level III English–Language Arts				
Flag values are as follows: **A** = low average task score **R** = low correlation with criterion **O** = high percent of omits/not responding **H** = high average task score				
Version/Field-Test Form	**Task Position**	**AIS**	**Polyserial**	**Flag**
Operational	1	3.18	.68	
1/5 *	2	3.28	.72	H
Operational	3	2.48	.81	
Operational	4	2.84	.77	
1/5 *	5	3.20	.66	H
Operational	6	2.22	.74	
Operational	7	3.20	.76	
1/5 *	8	2.44	.75	
Operational	9	2.89	.77	
Operational	10	2.94	.79	
1/5 *	11	2.54	.68	
Operational	12	3.02	.76	
2/6 *	2	3.52	.65	H
2/6 *	5	2.82	.69	
2/6 *	8	3.67	.71	H
2/6 *	11	2.79	.77	
3/7 *	2	2.54	.75	
3/7 *	5	2.93	.68	
3/7 *	8	3.12	.66	
3/7 *	11	3.14	.71	
4/8 *	2	3.12	.59	R
4/8 *	5	3.22	.63	H
4/8 *	8	2.43	.58	R
4/8 *	11	2.93	.73	

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

2008 CAPA Task Statistics: Level III Mathematics

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

Version/Field-Test Form	Task Position	AIS	Polyserial	Flag
Operational	13	2.93	.85	
1/5 *	14	2.85	.81	
Operational	15	2.09	.66	
Operational	16	3.00	.79	
1/5 *	17	2.90	.61	
Operational	18	2.71	.85	
Operational	19	2.34	.62	
1/5 *	20	3.34	.77	H
Operational	21	2.37	.85	
Operational	22	3.27	.81	H
1/5 *	23	2.21	.82	
Operational	24	2.21	.79	
2/6 *	14	3.25	.72	H
2/6 *	17	2.09	.44	R
2/6 *	20	2.18	.74	
2/6 *	23	1.69	.46	R
3/7 *	14	2.73	.56	R
3/7 *	17	2.49	.67	
3/7 *	20	2.14	.68	
3/7 *	23	2.48	.58	R
4/8 *	14	2.32	.45	R
4/8 *	17	2.59	.69	
4/8 *	20	2.30	.61	
4/8 *	23	3.25	.63	H

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

2008 CAPA Task Statistics: 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	2.77	.70	
1/3/5/7 *	26	2.20	.71	
Operational	27	2.88	.72	
Operational	28	2.85	.78	
1/3/5/7 *	29	3.11	.70	
Operational	30	2.56	.82	
Operational	31	2.35	.82	
1/3/5/7 *	32	2.98	.65	
Operational	33	2.56	.78	
Operational	34	2.29	.71	
1/3/5/7 *	35	3.13	.66	
Operational	36	2.72	.80	
2/4/6/8 *	26	2.36	.63	
2/4/6/8 *	29	3.43	.73	H
2/4/6/8 *	32	2.48	.67	
2/4/6/8 *	35	2.79	.74	

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

Table 6.B.4 2008 CAPA Task Statistics: Level IV

2008 CAPA Task Statistics: Level IV English–Language Arts				
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	3.09	.76	
1/8 *	2	2.72	.78	
Operational	3	1.63	.73	
Operational	4	3.04	.78	
1/8 *	5	3.49	.66	H
Operational	6	2.52	.82	
Operational	7	2.28	.81	
1/8 *	8	3.32	.57	R, H
Operational	9	2.04	.86	
Operational	10	2.55	.75	
1/8 *	11	3.00	.71	
Operational	12	2.63	.79	
2	2	3.17	.70	
2	5	2.48	.77	
2	8	2.54	.75	
2	11	3.21	.67	H
3	2	2.43	.84	
3	5	1.79	.74	
3	8	2.88	.76	
3	11	2.27	.85	
4	2	2.56	.74	
4	5	3.27	.62	H
4	8	2.51	.74	
4	11	2.28	.63	
5	2	2.43	.75	
5	5	2.68	.75	
5	8	3.09	.70	
5	11	2.64	.66	
6	2	2.70	.80	
6	5	2.86	.75	
6	8	2.65	.68	
6	11	2.50	.65	
7	2	3.24	.69	H
7	5	2.33	.65	
7	8	3.18	.56	R
7	11	2.82	.75	

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

2008 CAPA Task Statistics: 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	2.28	.87	
1/8 *	14	2.09	.69	
Operational	15	1.74	.80	
Operational	16	2.05	.72	
1/8 *	17	2.48	.68	
Operational	18	2.88	.86	
Operational	19	2.01	.86	
1/8 *	20	2.23	.81	
Operational	21	3.15	.65	
Operational	22	2.40	.85	
1/8 *	23	2.75	.70	
Operational	24	2.22	.74	
2	14	1.79	.36	R
2	17	1.82	.73	
2	20	2.72	.84	
2	23	2.15	.60	
3	14	1.88	.50	R
3	17	1.92	.74	
3	20	2.37	.85	
3	23	2.10	.77	
4	14	2.87	.43	R
4	17	3.11	.77	
4	20	2.20	.84	
4	23	2.43	.76	
5	14	2.14	.21	R
5	17	2.00	.65	
5	20	2.93	.80	
5	23	2.13	.86	
6	14	3.20	.56	R, H
6	17	2.95	.83	
6	20	2.85	.79	
6	23	2.38	.81	
7	14	2.72	.76	
7	17	3.12	.62	
7	20	2.99	.80	
7	23	2.04	.71	

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

2008 CAPA Task Statistics: 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	2.46	.75	
1/3/5/7 *	26	3.09	.61	
Operational	27	2.43	.73	
Operational	28	2.31	.73	
1/3/5/7 *	29	2.52	.74	
Operational	30	2.30	.73	
Operational	31	2.78	.69	
1/3/5/7 *	32	2.64	.71	
Operational	33	2.61	.76	
Operational	34	2.54	.78	
1/3/5/7 *	35	3.01	.71	
Operational	36	2.38	.80	
2/4/6/8 *	26	2.44	.56	R
2/4/6/8 *	29	2.51	.68	
2/4/6/8 *	32	2.77	.63	
2/4/6/8 *	35	2.73	.69	

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

Table 6.B.5 2008 CAPA Task Statistics: Level V

2008 CAPA Task Statistics: Level V English–Language Arts

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	3.06	.79	
1/7 *	2	3.25	.72	H
Operational	3	2.63	.84	
Operational	4	2.57	.75	
1/7 *	5	3.21	.54	R H
Operational	6	2.91	.87	
Operational	7	2.86	.75	
1/7 *	8	3.01	.68	
Operational	9	1.89	.82	
Operational	10	2.88	.79	
1/7 *	11	2.68	.81	
Operational	12	2.28	.79	
2/8 *	2	2.84	.39	R
2/8 *	5	3.03	.80	
2/8 *	8	2.89	.71	
2/8 *	11	2.24	.71	
3	2	2.38	.78	
3	5	2.48	.70	
3	8	3.14	.68	
3	11	2.48	.80	
4	2	2.66	.64	
4	5	2.90	.77	
4	8	2.90	.68	
4	11	2.38	.80	
5	2	2.32	.79	
5	5	2.19	.77	
5	8	2.70	.73	
5	11	2.03	.77	
6	2	2.62	.61	
6	5	3.18	.72	
6	8	2.41	.70	
6	11	2.87	.71	

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

2008 CAPA Task Statistics: 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	2.86	.75	
1/7 *	14	3.06	.83	
Operational	15	2.97	.83	
Operational	16	2.73	.74	
1/7 *	17	2.35	.57	R
Operational	18	2.67	.84	
Operational	19	2.35	.78	
1/7 *	20	1.85	.65	
Operational	21	2.52	.76	
Operational	22	2.14	.76	
1/7 *	23	2.51	.78	
Operational	24	3.02	.84	
2/8 *	14	3.13	.81	
2/8 *	17	3.15	.79	
2/8 *	20	3.19	.69	
2/8 *	23	2.24	.74	
3	14	3.09	.79	
3	17	2.96	.83	
3	20	2.24	.71	
3	23	2.42	.76	
4	14	2.26	.73	
4	17	2.67	.78	
4	20	2.58	.74	
4	23	2.93	.83	
5	14	2.35	.65	
5	17	2.57	.69	
5	20	2.02	.67	
5	23	2.72	.76	
6	14	3.22	.73	H
6	17	2.96	.82	
6	20	2.64	.65	
6	23	3.02	.81	

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

2008 CAPA Task Statistics: 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	2.25	.76	
1/3/5/7 *	26	2.65	.71	
Operational	27	2.20	.74	
Operational	28	2.65	.79	
1/3/5/7 *	29	2.34	.71	
Operational	30	2.02	.76	
Operational	31	2.72	.84	
1/3/5/7 *	32	2.54	.62	
Operational	33	2.63	.80	
Operational	34	2.14	.74	
1/3/5/7 *	35	1.91	.61	
Operational	36	2.80	.79	
2/4/6/8 *	26	2.34	.60	
2/4/6/8 *	29	2.44	.67	
2/4/6/8 *	32	3.41	.72	H
2/4/6/8 *	35	2.60	.67	

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

Appendix 6.C— IRT Model Fit Classification Tables

Table 6.C.1 Fit Classifications: Level I Tasks

Fit	ELA Frequency	Mathematics Frequency	Science Frequency
A	1	0	0
B	18	12	3
C	13	20	12
D	0	0	1
F	0	0	0

Table 6.C.2 Fit Classifications: Level II Tasks

Fit	ELA Frequency	Mathematics Frequency
A	0	2
B	15	9
C	9	12
D	0	1
F	0	0

Table 6.C.3 Fit Classifications: Level III Tasks

Fit	ELA Frequency	Mathematics Frequency	Science Frequency
A	1	0	1
B	13	12	11
C	10	11	4
D	0	1	0
F	0	0	0

Table 6.C.4 Fit Classifications: Level IV Tasks

Fit	ELA Frequency	Mathematics Frequency	Science Frequency
A	1	0	1
B	25	10	7
C	10	20	8
D	0	4	0
F	0	2	0

Table 6.C.5 Fit Classifications: Level V Tasks

Fit	ELA Frequency	Mathematics Frequency	Science Frequency
A	4	1	1
B	14	19	3
C	12	12	10
D	2	0	2
F	0	0	0

Chapter 7: Test Fairness

In order to evaluate equity among various subgroups, comprehensive analyses are conducted after test administration. This chapter summarizes the subgroup analyses performed for the CAPA 2008 administration. Because test security is crucial in the sustenance of a fair test, the chapter also briefly describes procedures for maintaining test security.

Demographic Distributions

The demographic variables used in the analyses included gender, ethnicity, and primary disability. Table 7.1 lists the specific subgroups that were used. Sample sizes for the disability subgroups within test level and subject area are presented in Appendix 7.A. Data are based on the P2 data received by ETS Statistical Analysis in late August.

Table 7.1 Subgroup Classifications

DIF Type	Reference Group	Focal Group
Gender	Male	Female
Race/Ethnicity	White	African American
		American Indian
		Asian
		Combined Asian Group (Asian/Pacific Islander/Filipino)
		Filipino
		Hispanic/Latin American
		Pacific Islander
Disability	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 7.2 presents the subgroup sample sizes and percent of total P2 data for each disability classification examined in the CAPA analyses.

Table 7.2 Frequency Distribution by Disability Across All CAPA Levels for 2008

Disability	ELA		Mathematics		Science	
	Frequency	Percent	Frequency	Percent	Frequency	Percent
Mental Retardation	19,383	43.2	19,336	43.2	6,028	46.6
Hard of Hearing	326	0.7	324	0.7	101	0.8
Deafness	431	1.0	429	1.0	127	1.0
Speech or Language Impairment	1,514	3.4	1,509	3.4	292	2.3
Visual Impairment	529	1.2	526	1.2	157	1.2

Disability	ELA		Mathematics		Science	
	Frequency	Percent	Frequency	Percent	Frequency	Percent
Emotional Disturbance	382	0.9	381	0.9	137	1.1
Orthopedic Impairment	4,387	9.8	4,378	9.8	1,234	9.5
Other Health Impairment	1,717	3.8	1,711	3.8	507	3.9
Established Medical Disability	0	0.0	0	0.0	0	0.0
Specific Learning Disability	3,045	6.8	3,043	6.8	886	6.9
Deaf-Blindness	47	0.1	47	0.1	10	0.1
Multiple Disabilities	2,341	5.2	2,333	5.2	634	4.9
Autism	9,611	21.4	9,580	21.4	2,559	19.8
Traumatic Brain Injury	328	0.7	325	0.7	94	0.7
Unknown	846	1.9	847	1.9	177	1.4
TOTAL	44,887	100.0	44,769	100.0	12,943	100.0

The "unknown" category consists of examinees for whom no disability type was marked. The tables in Appendix 7.A provide parallel information for each of the CAPA tests. The tables in Appendix 7.B include the percentage of students in the various proficiency levels for each category for ELA and mathematics as well as the number of students in each demographic category. Statistics for ethnicity by socioeconomic[1] status are included for ethnicity subgroups that contained at least 11 students.

Note that the statistics in these tables may differ slightly from the statewide statistics reported on the CDE Web site because the P2 data file was used for the analyses in this chapter. In addition, students receiving invalid scores were excluded rather than added into the category of below basic.

DIF Analyses

One of the goals of test development is to assemble a set of tasks that will provide an estimate of a student's ability that is as fair and accurate as possible for all groups within the population. DIF statistics are used to recognize the tasks for which identifiable groups of students with the same underlying level of ability have different probabilities of answering correctly.

If the task is differentially more difficult for an identifiable subgroup when conditioned on ability, the task may be measuring something different from the intended construct. However, it is important to recognize that DIF-flagged tasks might be related to actual differences in relevant knowledge or skills (task impact) or statistical Type 1 error. As a result, DIF statistics are used to identify potential sources of task bias. Tasks with statistically significant differences in performance are flagged so that the tasks can be carefully examined for possible biased or unfair content that was undetected in earlier fairness and bias content review meetings held prior to form construction. Subsequent review by content experts and bias/sensitivity committees are required to determine the source and meaning of performance differences.

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 and Haenszel 1959). The MH ordinal procedure compares the proportions of matched examinees from each group in each polytomous task-response category—that is, the probability of a given task score for the studied groups of interest after matching on total test score. As with dichotomously scored tasks, the common odds ratio is estimated across all categories of matched examinee ability. The resulting estimate is interpreted as the relative likelihood of a given

[1] In this analysis, a student's socioeconomic status was decided by whether or not the student participated in the National School Lunch Program or if both parents/guardians have not received a high school diploma.

task score for members of two groups when matched on ability. As such, the common odds ratio provides an estimated effect size where a value of unity indicates equal odds and thus no DIF (Dorans and 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 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 and Schmitt 1993). This procedure produces a DIF statistic based on the standardized mean difference (SMD) in average task scores between members of two groups who 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:

$$SMD = \sum_m w_m \left(E_{fm} - E_{rm} \right) / \sum_m w_m \quad (7.3)$$

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 and 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 puts tasks into 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 are divided to identify which group is being advantaged. These classifications are displayed in Table 7.3. The categories have been used by all ETS testing programs for more than 13 years.

Table 7.3 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 7.3, 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.

Following standard ETS procedure, tasks classified in Category C are sent for review by test development staff and/or content review committees to consider any identifiable characteristics that may have contributed to the differential task functioning. These tasks might be revised for additional field testing or removed from the task pool.

Test developers have been instructed to avoid selecting field-test tasks flagged as having shown DIF that disadvantage a group (C DIF) for future operational test forms unless their inclusion is deemed essential to meeting test-content specifications.

The groups studied for DIF are based on gender, race/ethnicity, and primary disability. The results of the DIF analyses identifying C-DIF tasks by ethnic group are presented in Table 7.4, and the C-DIF tasks identified for each disability group are given in Table 7.5. There were no C-DIF items identified by gender group.

Table 7.4 Item Exhibiting Significant DIF by Ethnic Group

Content Area	Task No.	Level	Task#	Version	SMD	Comparison	Disadvantaged
English–Language Arts	VC208341	V	12	Operational	0.355	White/Asian	White
	VC208341	IV	12	Operational	0.342	White/Filipino	White
	VC208660	V	12	Operational	0.337	White/ Filipino	White
Mathematics	VC335457	II	20	Field Test	0.346	White/Black	White
	VC203425	II	14	Field Test	–0.292	White/Hispanic	Hispanic
Science *	–	–	–	–	–	–	–

* No science items exhibited significant ethnic DIF.

Table 7.5 Items Exhibiting Significant DIF by Disability Group

Content Area	Task No.	Level	Task#	Version	SMD	Comparison	Disadvantaged
English–Language Arts Operational Tasks	VC205955	I	6	Operational	0.438	MR/VI	MR
	VC208571	IV	4	Operational	–0.546	MR/Autism	AU
	VC208510	IV	1	Operational	0.464	MR/Autism	MR
	VC208470	IV	6	Operational	–0.424	MR/Autism	AU
	VC208476	IV	7	Operational	–0.410	MR/Autism	AU
	VC208341	IV	12	Operational	0.629	MR/Autism	MR
	VC208692	V	4	Operational	–0.322	MR/Autism	AU
	VC208668	V	9	Operational	0.359	MR/Autism	MR
	VC208675	V	10	Operational	–0.569	MR/Autism	AU
	VC208660	V	12	Operational	0.673	MR/Autism	MR
English–Language Arts Field-Test Tasks	VC273005	I	11	3	–0.578	MR/OI	OI
	VC273049	I	11	4	–0.550	MR/OI	OI
	VC273005	I	11	3	–0.525	MR/MD	MD
	VC273049	I	11	4	0.518	MR/Autism	MR
	VC277630	II	5	2, 6	–0.313	MR/Autism	AU
	VC208239	II	8	2, 6	0.434	MR/Autism	MR
	VC277673	II	2	3, 7	–0.322	MR/Autism	AU
	VC334392	III	2	1, 5	–0.269	MR/Autism	AU
	VC334367	III	2	2, 6	–0.254	MR/Autism	AU
	VC334433	III	5	2, 6	–0.361	MR/Autism	AU
	VC334388	III	11	2, 6	–0.505	MR/Autism	AU
	VC334891	IV	5	2	0.492	MR/Autism	MR
	VC335048	IV	2	3	0.341	MR/Autism	MR
	VC334929	IV	8	3	–0.361	MR/Autism	AU
	VC335049	IV	11	3	0.331	MR/Autism	MR
	VC334808	IV	2	4	–0.397	MR/Autism	AU
	VC334788	IV	8	4	–0.286	MR/Autism	AU
	VC334861	IV	2	5	–0.594	MR/Autism	AU
	VC334856	IV	11	5	0.555	MR/Autism	MR
	VC334858	IV	11	7	–0.463	MR/Autism	AU
	VC335246	V	11	1, 7	–0.350	MR/Autism	AU
	VC335269	V	5	2, 8	0.493	MR/Autism	MR
	VC335118	V	8	2, 8	–0.309	MR/Autism	AU
	VC335263	V	11	2, 8	–0.456	MR/Autism	AU
	VC335265	V	2	3	0.505	MR/Autism	MR
	VC335115	V	5	3	–0.415	MR/Autism	AU
	VC335268	V	2	4	0.467	MR/Autism	MR
	VC335110	V	5	4	–0.430	MR/Autism	AU
	VC208642	V	11	4	–0.359	MR/Autism	AU
	VC335277	V	11	5	–0.453	MR/Autism	AU
	VC335276	V	8	6	–0.705	MR/Autism	AU

Content Area	Task No.	Level	Task#	Version	SMD	Comparison	Disadvantaged
Mathematics Operational Tasks	VC207352	III	19	Operational	–0.324	MR/SL	SL
	VC207429	III	15	Operational	–0.317	MR/Autism	AU
	VC207979	V	18	Operational	0.309	MR/SI	MR
	VC208066	V	21	Operational	–0.440	MR/SI	SI
	VC208066	V	21	Operational	–0.433	MR/SL	SL
Mathematics Field-Test Tasks	VC204394	II	14	1, 5	–0.437	MR/Autism	AU
	VC335475	II	23	2, 6	0.337	MR/Autism	MR
	VC205523	II	20	3, 7	–0.325	MR/Autism	AU
	VC203425	II	14	4, 8	0.311	MR/Autism	MR
	VC335538	III	14	2, 6	0.445	MR/Autism	MR
	VC335633	III	17	3, 7	0.324	MR/Autism	MR
	VC207447	III	14	4, 8	–0.314	MR/Autism	AU
	VC335623	III	17	4, 8	0.412	MR/Autism	MR
	VC335889	IV	20	1, 8	0.368	MR/Autism	MR
	VC335730	IV	7	4	0.329	MR/Autism	MR
	VC335725	IV	20	5	0.483	MR/Autism	MR
	VC336031	V	17	1, 7	0.370	MR/SL	MR
	VC207983	V	23	1, 7	0.413	MR/SL	MR
	VC335969	V	23	4	0.495	MR/Autism	MR
	VC335973	V	14	6	0.340	MR/Autism	MR
Science Operational Tasks *	–	–	–	–	–	–	
Science Field-test Tasks	VC331577	VI	26	1,3,5,7	–0.433	MR/Autism	AU
	VC331570	VI	35	1,3,5,7	–0.295	MR/Autism	AU

* There are no items in this category

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).

ETS's Office of Testing Integrity (OTI)

The OTI is a division of ETS that provides quality assurance 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:

- Prevent and minimize any testing security violations that can impact the fairness of testing
- Prevent and 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 consistently follow these 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 whenever they share tasks outside of ETS, including using registered, secure mail and express delivery and tracking records of the sending and receipt of any test materials.

Task Review by ARPs

ETS enforces security measures at ARP meetings to protect the integrity of meeting materials using these 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.
- The use of electronic devices in the meeting rooms is strictly prohibited.

Item Bank for Tasks

When the ARP review is complete, the tasks are placed in the item bank along with their corresponding review information. ETS then delivers the tasks to the CDE via 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. 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. A designated administrator at the CDE and at ETS authorizes the users.

Transfer of Forms and Tasks to the CDE

ETS shares a file transfer protocol (FTP) site with the CDE. FTP is a standard method for exclusive routing of files. It is a password-protected server that only authorized users can access. On that site, ETS posts Word, PDF, and other document files for the CDE to review. ETS sends an e-

mail to the CDE to notify CDE staff that files are posted. Task data are always transmitted in an encrypted format to the FTP site, never via e-mail.

Firewall

A firewall is software that prevents entry to files, e-mail, and other organization-specific programs by unauthorized users or computers. All ETS data exchange and internal e-mail remain within the ETS firewall at all ETS locations, from Princeton, New Jersey, to San Antonio, Texas, to Sacramento, California. The CDE has and will continue to view and approve ETS-developed applications such as those on the STAR Management System at ETS's Sacramento office because the applications remain behind ETS's firewall before release. No hacker has ever broken into ETS's firewall.

Printing

After tasks and test forms are approved, the files, on a CD, are sent for printing using a secure courier system, such as Federal Express. According to established procedures, the OTI pre-approves all printing vendors before they can work on secured confidential and proprietary test material. The printing vendor must submit a completed ETS Printing Plan and Typesetting Facility Security Plan that documents security procedures, access to test materials, work in progress, personnel procedures, and access to the facilities by the employees and visitors. After reviewing the completed plan, members of the OTI visit the printing vendor to conduct an on-site inspection. The secured printing vendor packs and ships printed test booklets to Pearson Educational Measurement for packaging and distribution in a tight and precise way to prevent boxes from opening.

Test Administration

Pearson receives testing materials from printers, packages them, and sends them to districts. After testing, districts return materials to Pearson for scoring. During each of these stages, Pearson takes extraordinary measures to protect 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 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 coordinator, who is responsible for returning them to the STAR Scoring and Processing Centers. More specifically:

- District STAR coordinators must 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 must 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 signs and submits 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 is asked to follow up with a written explanation of the violation or suspected violation.

- Any irregularities in test security may result in invalidation of student test results.

Processing and Scoring

An environment that promotes the security of the test prompts, student responses, data, and employees is of utmost concern to Pearson throughout the project of processing and scoring. 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. The only time they are touched is if there is a dispute of a score. For example, school districts and parents or guardians may request the rescoring of a student's test. In such a case, an answer document is removed from storage, copied, and sent securely to the ETS facility in Concord, California, for hand scoring, after which the copy is destroyed. No school or district personnel are allowed to look at the 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.

Transfer of Scores via Secure Data Exchange

After scoring is completed, Pearson sends files to ETS and follows secure data exchange procedures. 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 used. Upon entering the facilities, all Pearson employees are required to display their identification badge, which must be worn at all times while in the facility. Visitors must sign in and out, are assigned a visitor badge, and are escorted by Pearson personnel while at the facility. 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, each of which is done in a strictly controlled Configuration Management environment.

For mainframe processes, Pearson uses 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.

ETS and Pearson have implemented procedures and systems to provide the efficient coordination of secure data exchange, including the established, secure, FTP site that is used for secure data transfers between ETS and Pearson. These well-established procedures provide the timely, efficient, and secure transfer of data. Access to the STAR data files is limited to appropriate personnel who have direct project responsibilities.

Statistical Analysis

ETS systems load the Pearson files in a database. The Data Quality Services department 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 for student results, the files flow in three directions. First, paper reports, some with individual student results and others with summary results, are produced. Second, encrypted files of summary results are also sent to the CDE via FTP. Any summary results for fewer than eleven students are not reported. Third, the statistics from the results are entered into the ETS item bank in San Antonio.

Student Confidentiality

To meet NCLB and state requirements, school districts must collect demographic data about students, such as ethnicity, parent education, disabilities, whether the student qualified for the NSLP, and so forth. ETS takes precautions to prevent any of this information from becoming public or being used for anything other than testing purposes. Such measures are applicable to all documents in which these 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 very limited. 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 an off-site storage facility for safekeeping to permit continued operation in the case of a disaster.

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:

- Scoring
- Transfer of scores via 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 take an ETS-contracted test, such as a STAR exam. The ETS Office of Testing Integrity verifies that these standards are followed throughout the organization, including conducting periodic on-site security audits of departments, and preparing followup reports containing recommendations for improvement.

References

Dorans, N. J. and P. W. Holland 1993. "DIF Detection and Description: Mantel-Haenszel and Standardization," *Differential Item Functioning*. Edited by P. W. Holland and H. Wainer. Hillsdale, NJ: Erlbaum, 35–66.

Dorans, N. J. and E. Kulick 2006. "Differential Item Functioning on the Mini-Mental State Examination: An Application of the Mantel-Haenszel and Standardization Procedures**,"** *Medical Care*, *44*, 107–14.

Dorans, N. J. and A. P. Schmitt 1993. "Constructed Response and Differential Item Functioning: A Pragmatic Approach," in *Construction Versus Choice in Cognitive Measurement*. Edited by R.E. Bennett and W.C. Ward. Hillsdale, NJ: Lawrence Erlbaum Associates, 135–65.

Holland, P. W. and D. T. Thayer 1985. *An Alternative Definition of the ETS Delta Scale of Task Difficulty*, RR-85–43.

Mantel, N. 1963. "Chi-square Tests with One Degree of Freedom; Extensions of the Mantel-Haenszel Procedure," *Journal of the American Statistical Association*, 58: 690–700.

Mantel, N. and W. Haenszel 1959. "Statistical Aspects of the Analyses of Data from Retrospective Studies of Disease," *Journal of the National Cancer Institute*, 22: 719–48.

Appendix 7.A—Frequency Distribution Tables

Table 7.A.1 CAPA Disability Distributions: Level I

Disability	ELA		Mathematics		Science	
	Frequency	Percent	Frequency	Percent	Frequency	Percent
Mental Retardation	4,064	36.5	4,047	36.5	1,148	39.0
Hard of Hearing	70	0.6	70	0.6	21	0.7
Deafness	48	0.4	47	0.4	13	0.4
Speech or Language Impairment	71	0.6	71	0.6	9	0.3
Visual Impairment	284	2.6	284	2.6	74	2.5
Emotional Disturbance	7	0.1	7	0.1	3	0.1
Orthopedic Impairment	2,477	22.2	2,471	22.3	666	22.6
Other Health Impairment	245	2.2	243	2.2	58	2.0
Established Medical Disability	0	0.0	0	0.0	0	0.0
Specific Learning Disability	92	0.8	92	0.8	17	0.6
Deaf-Blindness	30	0.3	30	0.3	6	0.2
Multiple Disabilities	1,188	10.7	1,186	10.7	306	10.4
Autism	2,293	20.6	2,281	20.6	575	19.5
Traumatic Brain Injury	75	0.7	75	0.7	17	0.6
Unknown	192	1.7	192	1.7	33	1.1
TOTAL	11,136	100.0	11,096	100.0	2,946	100.0

Table 7.A.2 CAPA Disability Distributions: Level II

Disability	ELA		Mathematics	
	Frequency	Percent	Frequency	Percent
Mental Retardation	2,362	36.4	2,357	36.5
Hard of Hearing	40	0.6	40	0.6
Deafness	56	0.9	56	0.9
Speech or Language Impairment	592	9.1	589	9.1
Visual Impairment	44	0.7	43	0.7
Emotional Disturbance	45	0.7	45	0.7
Orthopedic Impairment	347	5.4	347	5.4
Other Health Impairment	310	4.8	308	4.8
Established Medical Disability	0	0.0	0	0.0
Specific Learning Disability	486	7.5	486	7.5
Deaf-Blindness	3	0.1	3	0.1
Multiple Disabilities	196	3.0	195	3.0
Autism	1,815	28.0	1,813	28.0
Traumatic Brain Injury	36	0.6	36	0.6
Unknown	150	2.3	148	2.3
TOTAL	6,482	100.0	6,466	100.0

Table 7.A.3 CAPA Disability Distributions: Level III

Disability	ELA		Mathematics		Science	
	Frequency	Percent	Frequency	Percent	Frequency	Percent
Mental Retardation	2,772	42.2	2,766	42.2	1,376	44.1
Hard of Hearing	55	0.8	54	0.8	25	0.8
Deafness	64	1.0	64	1.0	28	0.9
Speech or Language Impairment	349	5.3	347	5.3	147	4.7
Visual Impairment	41	0.6	41	0.6	25	0.8
Emotional Disturbance	59	0.9	59	0.9	31	1.0
Orthopedic Impairment	380	5.8	378	5.8	178	5.7
Other Health Impairment	276	4.2	275	4.2	145	4.6
Established Medical Disability	0	0.0	0	0.0	0	0.0
Specific Learning Disability	598	9.1	597	9.1	270	8.7
Deaf-Blindness	4	0.1	4	0.1	2	0.1
Multiple Disabilities	213	3.2	213	3.3	105	3.4
Autism	1,626	24.7	1,625	24.8	738	23.6
Traumatic Brain Injury	33	0.5	32	0.5	15	0.5
Unknown	107	1.6	108	1.7	38	1.2
TOTAL	6,577	100.0	6,563	100.0	3,123	100.0

Table 7.A.4 CAPA Disability Distributions: Level IV

Disability	ELA		Mathematics		Science	
	Frequency	Percent	Frequency	Percent	Frequency	Percent
Mental Retardation	4,941	47.6	4,942	47.7	1,735	50.5
Hard of Hearing	72	0.7	71	0.7	24	0.7
Deafness	122	1.2	121	1.2	44	1.3
Speech or Language Impairment	316	3.1	315	3.0	81	2.4
Visual Impairment	62	0.6	61	0.6	24	0.7
Emotional Disturbance	93	0.9	93	0.9	33	1.0
Orthopedic Impairment	596	5.8	596	5.8	210	6.1
Other Health Impairment	461	4.4	461	4.5	160	4.7
Established Medical Disability	0	0.0	0	0.0	0	0.0
Specific Learning Disability	964	9.3	964	9.3	288	8.4
Deaf-Blindness	7	0.1	7	0.1	1	0.0
Multiple Disabilities	373	3.6	371	3.6	117	3.4
Autism	2,122	20.5	2,116	20.4	653	19.0
Traumatic Brain Injury	86	0.8	86	0.8	31	0.9
Unknown	157	1.5	157	1.5	35	1.0
TOTAL	10,372	100.0	10,361	100.0	3,436	100.0

Table 7.A.5 CAPA Disability Distributions: Level V

Disability	ELA		Mathematics		Science	
	Frequency	Percent	Frequency	Percent	Frequency	Percent
Mental Retardation	5,244	50.8	5,224	50.8	1,741	51.7
Hard of Hearing	89	0.9	89	0.9	30	0.9
Deafness	141	1.4	141	1.4	41	1.2
Speech or Language Impairment	186	1.8	187	1.8	55	1.6
Visual Impairment	98	1.0	97	0.9	32	1.0
Emotional Disturbance	178	1.7	177	1.7	68	2.0
Orthopedic Impairment	587	5.7	586	5.7	177	5.3
Other Health Impairment	425	4.1	424	4.1	141	4.2
Established Medical Disability	0	0.0	0	0.0	0	0.0
Specific Learning Disability	905	8.8	904	8.8	306	9.1
Deaf-Blindness	3	0.0	3	0.0	1	0.0
Multiple Disabilities	371	3.6	368	3.6	104	3.1
Autism	1,755	17.0	1,745	17.0	571	17.0
Traumatic Brain Injury	98	1.0	96	0.9	29	0.9
Unknown	240	2.3	242	2.4	70	2.1
TOTAL	10,320	100.0	10,283	100.0	3,366	100.0

Appendix 7.B—Proficiency Category Distribution Tables

Table 7.B.1 2008 Proficiency Category Distributions for All Examinees: English–Language Arts

English–Language Arts, All Levels	Subgroup	No. of Students	Percentage of Students				
			Far Below Basic	Below Basic	Basic	Proficient	Advanced
All Examinees *	All Examinees	44,887	8	8	16	30	38
Grade	2	4,759	4	5	16	39	36
	3	4,447	4	4	17	38	37
	4	4,425	6	9	19	28	38
	5	4,260	6	9	16	27	42
	6	4,458	11	11	17	30	32
	7	4,475	10	10	16	26	38
	8	4,577	9	10	15	27	39
	9	4,695	11	9	15	28	37
	10	4,611	11	8	14	27	39
	11	4,180	10	8	14	27	42
Gender	Male	28,672	8	8	16	30	38
	Female	16,057	8	8	16	30	38
	Unknown	158	9	8	12	30	41
Race/Ethnicity	American Indian or Alaska Native	376	6	9	14	28	44
	NSLP †	225	6	9	12	29	44
	Non-NSLP	135	4	10	16	28	42
	Asian	3,136	11	10	17	29	33
	NSLP †	1,406	10	11	16	28	35
	Non-NSLP	1,646	12	10	17	30	31
	Pacific Islander	236	8	11	13	31	37
	NSLP †	127	9	13	16	26	37
	Non-NSLP	98	8	8	9	39	36
	Filipino	1,291	11	11	18	27	33
	NSLP †	418	12	10	19	25	33
	Non-NSLP	829	10	12	17	28	33
	Hispanic or Latino	21,908	8	9	16	30	37
	NSLP †	16,594	7	8	17	30	38
	Non-NSLP	4,796	11	9	15	29	35

English–Language Arts, All Levels	Subgroup	No. of Students	Percentage of Students				
			Far Below Basic	Below Basic	Basic	Proficient	Advanced
Race/Ethnicity (cont.)	African American	4,906	8	7	16	30	40
	NSLP †	3,207	6	7	16	30	42
	Non-NSLP	1,553	11	8	16	29	37
	White (not Hispanic origin)	12,318	8	8	15	30	40
	NSLP †	4,202	6	7	14	31	42
	Non-NSLP	7,725	9	9	15	29	39
	Unknown	716	8	7	14	31	40
	NSLP †	224	7	9	13	32	39
	Non-NSLP	287	7	8	13	34	39
Language Fluency	English Only	27,669	8	8	15	30	39
	Initially–Fluent English Proficient	1,255	11	11	17	28	33
	English Learner	14,306	8	9	16	30	37
	Reclassified–Fluent English Proficient	915	7	9	17	29	39
	Unknown	742	8	7	15	33	37
Economic Disadvantage	Yes	26,403	7	8	16	30	39
	No	17,069	10	9	16	29	37
	Unknown	1,415	7	7	14	33	40
Primary Disability	Mental Retardation	19,383	7	10	19	30	34
	Hard of Hearing	326	7	10	16	30	37
	Deafness	431	3	7	22	38	30
	Speech/Language Impairment	1,514	1	2	8	35	55
	Visual Impairment	529	11	9	16	29	35
	Emotional Disturbance	382	3	3	6	27	61
	Orthopedic Impairment	4,387	14	8	13	30	35
	Other Health Impairment	1,717	4	5	14	30	47
	Specific Learning Impairment	3,045	1	1	5	27	66
	Deaf Blindness	47	23	6	19	30	21
	Multiple Group	2,341	16	10	14	30	30
	Autism	9,611	10	10	16	29	35
	Traumatic Brain Injury	328	7	7	13	25	48
	Unknown	846	7	7	13	31	42

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

† National School Lunch Program

Table 7.B.2 2008 Proficiency Category Distributions for All Examinees: Mathematics

Mathematics, All Levels—Group	Subgroup	N	Percent				
			Far Below Basic	Below Basic	Basic	Proficient	Advanced
All Examinees *	All	44,769	12	11	18	30	29
Grade	2	4,740	7	7	16	34	36
	3	4,439	6	7	15	32	41
	4	4,416	6	11	14	32	37
	5	4,248	6	10	13	31	40
	6	4,448	17	16	21	24	21
	7	4,464	15	15	19	24	26
	8	4,573	14	14	19	24	29
	9	4,677	17	12	22	31	19
	10	4,592	16	10	21	32	22
	11	4,172	15	11	19	32	24
Gender	Male	28,598	11	11	17	29	32
	Female	16,013	13	12	19	30	26
	Unknown	158	15	4	16	29	35
Race/Ethnicity	American Indian or Alaska Native	374	9	7	18	33	32
	NSLP †	223	9	8	15	32	36
	Non-NSLP	135	7	7	25	33	27
	Asian	3,129	14	12	19	29	26
	NSLP †	1,403	14	11	18	29	27
	Non-NSLP	1,642	15	12	20	29	24
	Pacific Islander	237	11	12	19	31	27
	NSLP †	127	11	14	17	31	26
	Non-NSLP	99	11	11	20	30	27
	Filipino	1,289	14	12	19	31	24
	NSLP †	417	15	11	21	30	23
	Non-NSLP	828	13	13	18	32	25
	Hispanic or Latino	21,855	12	12	18	29	30
	NSLP †	16,556	11	11	17	29	31
	Non-NSLP	4,783	15	13	18	28	26
	African American	4,898	12	10	17	31	30
	NSLP †	3,201	10	10	17	32	32
	Non-NSLP	1,551	15	11	18	30	25

Mathematics, All Levels—Group	Subgroup	N	Percent				
			Far Below Basic	Below Basic	Basic	Proficient	Advanced
Race/Ethnicity (cont.)	White (not Hispanic origin)	12,273	12	11	18	30	29
	NSLP †	4,192	10	10	16	30	34
	Non-NSLP	7,692	12	11	20	30	27
	Unknown	714	11	9	18	29	34
	NSLP †	223	13	10	17	26	33
	Non-NSLP	287	10	9	19	29	33
Language Fluency	English Only	27,589	12	11	18	30	29
	Initially–Fluent English Proficient	1,252	15	15	21	31	19
	English Learner	14,276	11	11	17	29	31
	Reclassified–Fluent English Proficient	914	10	12	19	28	31
	Unknown	738	11	7	20	31	31
Economic Disadvantage	Yes	26,342	11	11	17	30	31
	No	17,017	14	12	19	29	26
	Unknown	1,410	9	8	18	30	34
Primary Disability	Mental Retardation	19,336	12	13	21	31	24
	Hard of Hearing	324	10	7	17	31	35
	Deafness	429	4	6	17	29	44
	Speech/Language Impairment	1,509	1	2	6	26	65
	Visual Impairment	526	17	19	19	28	17
	Emotional Disturbance	381	5	3	12	26	54
	Orthopedic Impairment	4,378	20	17	18	27	18
	Other Health Impairment	1,711	6	9	17	29	38
	Specific Learning Impairment	3,043	1	2	7	24	66
	Deaf Blindness	47	28	28	19	11	15
	Multiple group	2,333	24	15	18	28	15
	Autism	9,580	12	10	18	31	29
	Traumatic Brain Injury	325	10	10	14	30	35

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

† National School Lunch Program

Chapter 8: Reliability

This chapter summarizes the evidence of reliability for the spring 2008 CAPA administration.

Test Score Reliability

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 an 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 the state of health of the examinee or the 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 is measured by coefficient alpha (Cronbach, 1951). Coefficient alpha, α, can be thought of as a lower bound to a theoretical reliability and is reported below.

$$\alpha \geq \frac{k}{k-1}\left(1-\frac{\sum \sigma_i^2}{\sigma_X^2}\right), \tag{8.1}$$

where:

k is the number of tasks on the test,

$\sum \sigma_i^2$ is the task score variance summed over all tasks, and

σ_X^2 is the test-score variance.

The reliabilities for the CAPA tests are displayed in Table 8.1 on page 87—the number of examinees, the means, standard deviation and the standard error of measurement (SEM) that will be explained in the following section. The reliabilities are given for both the raw and scale scores for ELA and mathematics and for the raw scores for science.

Standard Error of Measurement (SEM)

The SEM is an estimate of error score variance, σ_E^2. The SEM is in the metric of the scale and is estimated on the basis of the standard deviation of observed scores and the test reliability coefficient:

$$\text{SEM} = s_x \sqrt{1-\alpha}, \tag{8.2}$$

where:

SEM = standard error of measurement,

s_x = standard deviation of observed scores, and

α = coefficient of reliability (alpha).

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 with ±1.96 SEM around the observed score would contain an examinee's true score (Crocker and 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 ± 3.77 rounded to the nearest integer).

SEMs for the CAPAs are displayed in Table 8.1.

Table 8.1 Reliabilities and Standard Errors of Measurement for the CAPA

					Scale Score			Raw Score		
Subject Area	Level	No. of Items	No. of Examinees	Reliab.	Mean	Std. Dev.	SEM	Mean	Std. Dev.	SEM
English–Language Arts	I	8	11,136	0.93	45.95	13.34	3.53	27.00	11.83	3.13
	II	8	6,482	0.84	38.16	7.51	3.00	22.87	6.17	2.47
	III	8	6,577	0.87	38.16	9.83	3.54	22.79	6.47	2.33
	IV	8	10,372	0.88	36.32	9.11	3.16	19.74	7.25	2.51
	V	8	10,320	0.89	37.57	9.50	3.15	21.07	7.29	2.42
Mathematics	I	8	11,096	0.91	34.88	11.42	3.43	22.75	11.04	3.31
	II	8	6,466	0.88	40.11	8.70	3.01	20.73	7.57	2.62
	III	8	6,563	0.88	40.85	9.32	3.23	21.02	7.25	2.51
	IV	8	10,361	0.88	35.11	10.18	3.53	18.68	7.66	2.65
	V	8	10,283	0.88	35.21	9.15	3.17	21.22	7.89	2.73
*Science **	I	8	2,964	0.93	–	–	–	22.66	11.86	3.14
	III	8	3,123	0.87	–	–	–	21.06	6.79	2.45
	IV	8	3,436	0.85	–	–	–	19.70	6.50	2.52
	V	8	3,366	0.88	–	–	–	19.31	6.62	2.29

* There are no scale scores for science.

Inter-Rater Reliability

Inter-rater reliability addresses the consistency of the implementation of a rating system. For the CAPA, approximately 10 percent of students received two ratings, one by the primary examiner and a second independent rating by a trained observer. Consistency between the two ratings is evaluated with the following statistics:

- Number and percentage of exact agreement between raters
- Number and percentage of adjacent agreement between raters
- Number and percentage of nonadjacent scores assigned by raters
- Mean absolute difference between the ratings assigned by the examiner and the observer
- Correlation between the ratings assigned by the examiner and the observer

Inter-rater reliabilities for the operational tasks are presented by level in Appendix 8.A.

Reliability of Classification and Decision Accuracy

The methodology used for estimating the reliability of performance-level classification decisions as described in Livingston and Lewis (1995) provides estimates of decision accuracy and classification consistency:

> The term *accuracy* … refers to the extent to which the actual classifications of test takers (on the basis of their single-form scores) agree with those that would be made on the basis of their true scores, if their true scores could somehow be known. The term *consistency* refers to the agreement between the classifications based on two non-overlapping, equally difficult forms of the test. (Livingston and Lewis 1995, p. 178)

For the CAPA, the estimation of reliability of performance-level classification decisions is implemented through the use of the ETS-proprietary computer program RELCLASS-COMP (Version 4.12). For each test level and subject area, RELCLASS-COMP estimates true scores and

single-form scores on forms parallel to the one actually given. RELCLASS-COMP estimates decision accuracy using an estimated joint distribution of reported performance-level classifications on the current form of the exam and the performance-level classifications based on an all-forms average (true score). RELCLASS-COMP estimates decision consistency using an estimated joint distribution of reported performance-level classifications on the current form of the exam and performance-level classifications on the alternate (parallel) form.

In each case, the proportion of performance-level classifications with exact agreement is the sum of the entries in the diagonal of the contingency table representing the joint distribution. Reliability of classification at each performance-level cut score is estimated by collapsing the joint distribution at the passing score boundary into a two-by-two table and summing the two entries in the diagonal. The reliability of classification and decision accuracies is presented for each test level and subject area in Appendix 8.B.

References

Crocker, L. and J. Algina 1986. *Introduction to Classical and Modern Test Theory*. New York: Holt.

Livingston, S. A., and C. Lewis 1995. "Estimating the Consistency and Accuracy of Classification Based on Test Scores," *Journal on Educational Measurement*, 32: 179–97.

Appendix 8.A—Inter-Rater Reliabilities

Table 8.A.1 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	1,346	3.65	1.75	1,346	3.63	1.76	92.86	4.46	2.68	0.13	0.95
	3	1,346	3.13	1.74	1,346	3.11	1.76	89.37	7.21	3.42	0.16	0.95
	4	1,346	3.54	1.78	1,346	3.52	1.78	90.86	5.72	3.42	0.16	0.93
	6	1,346	3.80	1.72	1,346	3.76	1.74	89.90	5.94	4.16	0.19	0.91
	7	1,346	3.65	1.72	1,346	3.64	1.72	90.49	6.69	2.82	0.15	0.94
	9	1,346	3.67	1.75	1,346	3.64	1.77	92.80	4.68	2.52	0.13	0.95
	10	1,346	3.67	1.76	1,346	3.65	1.76	90.49	6.39	3.12	0.16	0.94
	12	1,346	3.60	1.73	1,346	3.58	1.74	91.59	5.65	2.76	0.14	0.94
Mathematics	1	1,321	3.59	1.72	1,321	3.59	1.71	93.10	4.62	2.28	0.11	0.95
	3	1,321	3.23	1.69	1,321	3.20	1.70	91.00	5.68	3.32	0.15	0.94
	4	1,321	2.98	1.67	1,321	2.99	1.67	88.72	7.72	3.56	0.18	0.93
	6	1,321	3.12	1.70	1,321	3.10	1.70	91.75	5.53	2.72	0.13	0.95
	7	1,321	2.51	1.46	1,321	2.51	1.46	91.52	6.59	1.89	0.11	0.95
	9	1,321	3.02	1.76	1,321	3.00	1.75	90.01	6.51	3.48	0.17	0.94
	10	1,321	2.68	1.63	1,321	2.66	1.63	89.33	8.25	2.42	0.15	0.94
	12	1,321	3.37	1.75	1,321	3.39	1.74	91.52	5.30	3.18	0.15	0.93
Science	1	350	2.87	1.8	350	2.8	1.80	87.42	8.57	4.01	0.21	0.92
	3	350	2.96	1.8	350	2.9	1.77	86.00	9.43	4.57	0.23	0.91
	4	350	3.08	1.8	350	3.1	1.79	86.57	8.86	4.57	0.25	0.90
	6	350	2.91	1.8	350	2.9	1.75	88.00	7.71	4.29	0.20	0.92
	7	350	3.02	1.7	350	3.0	1.71	89.43	7.14	3.43	0.18	0.93
	9	350	2.63	1.7	350	2.6	1.72	86.57	9.14	4.29	0.21	0.92
	10	350	2.64	1.7	350	2.7	1.73	88.57	7.14	4.29	0.20	0.92
	12	350	3.36	1.9	350	3.4	1.82	89.71	6.29	4.00	0.20	0.91

* Mean absolute difference between first and second ratings

† Pearson correlation between first and second ratings

Table 8.A.2 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	1,078	3.84	0.53	1,078	3.84	0.54	98.14	1.48	0.37	0.02	0.94
	3	1,078	2.20	1.19	1,078	2.20	1.19	92.30	6.22	1.49	0.09	0.95
	4	1,078	2.44	1.36	1,078	2.45	1.36	92.67	5.57	0.65	0.10	0.95
	6	1,078	2.68	1.32	1,078	2.68	1.32	94.62	4.08	1.30	0.07	0.97
	7	1,078	3.64	0.75	1,078	3.64	0.78	96.47	2.50	1.03	0.05	0.92
	9	1,078	2.52	1.17	1,078	2.53	1.18	95.73	3.43	0.83	0.05	0.97
	10	1,078	3.14	1.08	1,078	3.14	1.08	94.34	4.73	0.93	0.07	0.96
	12	1,078	2.22	1.09	1,078	2.23	1.09	92.39	6.03	1.58	0.10	0.93
Mathematics	1	1,068	3.07	1.21	1,068	3.07	1.22	96.25	3.09	0.65	0.05	0.98
	3	1,068	2.60	1.39	1,068	2.61	1.40	96.54	2.53	0.93	0.05	0.97
	4	1,068	2.60	1.21	1,068	2.61	1.20	95.41	3.37	1.22	0.06	0.97
	6	1,068	1.82	1.27	1,068	1.82	1.28	95.69	3.28	1.03	0.06	0.97
	7	1,068	2.47	1.03	1,068	2.48	1.02	94.57	4.31	1.12	0.07	0.96
	9	1,068	2.21	1.34	1,068	2.21	1.34	96.25	2.81	0.94	0.05	0.97
	10	1,068	2.90	1.24	1,068	2.91	1.23	95.97	2.81	1.21	0.06	0.96
	12	1,068	3.05	1.21	1,068	3.06	1.19	94.85	3.28	1.88	0.09	0.93

* Mean absolute difference between first and second ratings

† Pearson correlation between first and second ratings

Table 8.A.3 Inter-Rater Reliabilities for Operational Tasks: Level III

Level III		First Rating			Second Rating			% Agreement			MAD *	Corr †
Subject	Task	N	Mean	SD	N	Mean	SD	Exact	Adjacent	Neither		
English–Language Arts	1	1,023	3.14	0.99	1,023	3.14	0.98	93.06	6.16	0.78	0.08	0.95
	3	1,023	2.49	1.09	1,023	2.49	1.09	93.06	6.35	0.59	0.08	0.96
	4	1,023	2.94	1.05	1,023	2.94	1.05	96.19	2.93	0.88	0.06	0.95
	6	1,023	2.25	0.89	1,023	2.28	0.89	92.18	6.74	1.08	0.09	0.92
	7	1,023	3.20	0.87	1,023	3.19	0.88	95.31	3.81	0.89	0.06	0.93
	9	1,023	2.94	1.31	1,023	2.93	1.31	95.70	3.13	1.18	0.06	0.96
	10	1,023	2.98	1.15	1,023	2.97	1.17	94.13	5.08	0.79	0.07	0.96
	12	1,023	3.02	1.19	1,023	3.03	1.19	95.70	3.23	1.08	0.07	0.95
Mathematics	1	1,024	2.99	1.27	1,024	2.99	1.28	96.19	3.42	0.40	0.04	0.98
	3	1,024	2.13	0.92	1,024	2.15	0.92	91.41	7.81	0.79	0.10	0.93
	4	1,024	3.05	1.17	1,024	3.04	1.18	95.80	3.42	0.78	0.06	0.96
	6	1,024	2.79	1.40	1,024	2.78	1.41	96.48	2.15	1.37	0.06	0.97
	7	1,024	2.39	1.05	1,024	2.38	1.05	95.90	3.03	1.08	0.05	0.96
	9	1,024	2.44	1.35	1,024	2.44	1.36	96.78	2.64	0.59	0.05	0.97
	10	1,024	3.34	1.12	1,024	3.33	1.13	97.95	1.07	0.98	0.04	0.96
	12	1,024	2.23	1.29	1,024	2.24	1.28	95.31	3.81	0.88	0.06	0.97
Science	1	492	2.74	1.01	492	2.76	1.00	95.12	3.66	1.22	0.07	0.95
	3	492	2.79	1.16	492	2.80	1.14	92.48	6.91	0.61	0.08	0.96
	4	492	2.87	0.96	492	2.87	0.96	96.14	3.66	0.20	0.04	0.98
	6	492	2.55	1.24	492	2.55	1.27	92.07	6.71	1.22	0.10	0.96
	7	492	2.34	1.37	492	2.33	1.37	95.12	4.07	0.81	0.06	0.97
	9	492	2.46	1.22	492	2.51	1.21	94.31	4.47	1.22	0.07	0.96
	10	492	2.36	1.08	492	2.36	1.06	95.33	4.07	0.61	0.06	0.96
	12	492	2.73	1.05	492	2.77	1.02	91.46	6.71	1.83	0.12	0.90

* Mean absolute difference between first and second ratings

† Pearson correlation between first and second ratings

Table 8.A.4 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	1,352	3.15	1.10	1,352	3.15	1.12	96.67	2.66	0.66	0.04	0.97
	3	1,352	1.49	1.08	1,352	1.49	1.09	91.94	6.07	1.99	0.11	0.92
	4	1,352	3.09	1.22	1,352	3.07	1.24	91.35	6.21	2.44	0.13	0.92
	6	1,352	2.56	1.19	1,352	2.54	1.22	89.50	9.10	1.41	0.12	0.94
	7	1,352	2.29	1.19	1,352	2.29	1.19	87.57	10.87	1.55	0.14	0.93
	9	1,352	2.02	1.23	1,352	2.02	1.24	93.71	5.10	1.19	0.08	0.96
	10	1,352	2.52	1.13	1,352	2.54	1.13	94.23	4.59	1.19	0.07	0.95
	12	1,352	2.64	1.37	1,352	2.67	1.36	93.05	5.10	1.85	0.11	0.94
Mathematics	1	1,351	2.25	1.27	1,351	2.26	1.28	95.19	3.70	1.11	0.06	0.97
	3	1,351	1.70	1.23	1,351	1.70	1.23	96.30	2.89	0.81	0.05	0.98
	4	1,351	1.93	1.20	1,351	1.92	1.20	93.86	4.89	1.25	0.08	0.95
	6	1,351	2.94	1.39	1,351	2.92	1.39	94.89	3.85	1.26	0.07	0.96
	7	1,351	1.95	1.32	1,351	1.96	1.33	95.93	3.18	0.89	0.06	0.97
	9	1,351	3.21	1.06	1,351	3.19	1.09	95.11	3.63	1.26	0.07	0.94
	10	1,351	2.38	1.35	1,351	2.37	1.36	94.52	3.70	1.78	0.09	0.95
	12	1,351	2.23	1.14	1,351	2.24	1.15	91.64	6.37	2.00	0.11	0.93
Science	1	444	2.44	1.18	444	2.48	1.18	94.59	3.83	1.58	0.07	0.96
	3	444	2.39	1.03	444	2.41	1.03	90.99	7.88	1.13	0.10	0.94
	4	444	2.23	1.15	444	2.25	1.16	87.39	9.46	3.16	0.17	0.90
	6	444	2.31	1.14	444	2.30	1.17	90.99	7.66	1.36	0.11	0.94
	7	444	2.90	1.16	444	2.93	1.16	94.14	4.28	1.59	0.09	0.94
	9	444	2.57	1.11	444	2.56	1.11	89.64	8.56	1.80	0.14	0.90
	10	444	2.58	1.05	444	2.58	1.07	91.89	5.86	2.26	0.12	0.89
	12	444	2.34	1.12	444	2.35	1.14	87.39	10.14	2.48	0.17	0.88

* Mean absolute difference between first and second ratings

† Pearson correlation between first and second ratings

Table 8.A.5 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	989	3.05	1.30	989	3.05	1.30	93.53	4.35	2.12	0.10	0.94
	3	989	2.65	1.12	989	2.61	1.14	89.38	9.40	1.21	0.13	0.92
	4	989	2.53	1.10	989	2.53	1.11	91.00	6.67	2.31	0.13	0.90
	6	989	3.01	1.07	989	2.98	1.10	91.51	6.98	1.51	0.11	0.92
	7	989	2.91	1.12	989	2.92	1.12	91.41	6.37	2.22	0.12	0.92
	9	989	1.83	1.13	989	1.82	1.14	89.08	8.09	2.83	0.14	0.91
	10	989	2.97	1.24	989	2.93	1.27	89.89	7.68	2.43	0.14	0.92
	12	989	2.34	1.28	989	2.33	1.29	92.82	4.65	2.53	0.12	0.92
Mathematics	1	985	2.92	1.34	985	2.90	1.35	96.04	2.03	1.93	0.07	0.96
	3	985	3.03	1.26	985	3.01	1.28	94.52	3.96	1.52	0.09	0.94
	4	985	2.78	1.34	985	2.78	1.35	91.68	5.18	3.15	0.14	0.91
	6	985	2.76	1.15	985	2.76	1.16	94.52	3.55	1.93	0.08	0.94
	7	985	2.35	1.38	985	2.32	1.38	92.99	3.86	3.14	0.12	0.94
	9	985	2.39	1.46	985	2.38	1.46	94.52	3.05	2.43	0.09	0.95
	10	985	2.18	1.24	985	2.19	1.24	94.01	3.96	2.03	0.09	0.94
	12	985	3.07	1.35	985	3.10	1.31	94.82	2.84	2.33	0.10	0.92
Science	1	344	2.19	1.06	344	2.23	1.03	88.08	8.72	3.20	0.16	0.88
	3	344	2.29	1.11	344	2.30	1.13	87.50	9.30	3.20	0.16	0.90
	4	344	2.71	1.04	344	2.74	1.05	89.83	7.85	2.32	0.13	0.90
	6	344	2.02	1.15	344	2.04	1.17	90.41	8.14	1.45	0.12	0.94
	7	344	2.77	0.98	344	2.76	1.00	88.95	8.14	2.90	0.15	0.86
	9	344	2.75	0.91	344	2.77	0.88	93.02	4.94	2.03	0.09	0.91
	10	344	2.27	1.17	344	2.26	1.18	91.28	6.10	2.61	0.12	0.93
	12	344	2.93	1.21	344	2.97	1.18	92.15	4.65	3.19	0.13	0.91

* Mean absolute difference between first and second ratings

† Pearson correlation between first and second ratings

Appendix 8.B—Decision Accuracy and Decision Consistency

Table 8.B.1 Decision Accuracy and Decision Consistency: Level I English–Language Arts

	Placement Score	Advanced	Proficient	Basic	Below Basic	Far Below Basic	Category Total †
Decision Accuracy	28–40	0.53	0.04	0.00	0.00	0.00	0.57
	16–27	0.04	0.17	0.02	0.00	0.00	0.23
	12–15	0.00	0.02	0.02	0.01	0.00	0.05
All-forms Average *	9–11	0.00	0.00	0.01	0.01	0.01	0.04
	0–8	0.00	0.00	0.01	0.02	0.08	0.11
	Estimated Proportion Correctly Classified: Total = 0.81, Proficient & Above = 0.96						
Decision Consistency	28–40	0.51	0.05	0.00	0.00	0.00	0.57
	16–27	0.05	0.14	0.03	0.01	0.00	0.23
	12–15	0.00	0.02	0.01	0.01	0.01	0.05
Alternate Form *	9–11	0.00	0.01	0.01	0.01	0.01	0.04
	0–8	0.00	0.01	0.01	0.01	0.08	0.11
	Estimated Proportion Correctly Classified: Total =0.75, Proficient & Above = 0.92						

* Values in table are proportions of the total sample.

† Inconsistencies with category cell entries are due to rounding.

Table 8.B.2 Decision Accuracy and Decision Consistency: Level I Mathematics

	Placement Score	Advanced	Proficient	Basic	Below Basic	Far Below Basic	Category Total †
Decision Accuracy	33–40	0.18	0.03	0.00	0.00	0.00	0.21
	23–32	0.04	0.23	0.05	0.01	0.00	0.34
	18–22	0.00	0.04	0.07	0.04	0.00	0.15
All-forms Average *	9–17	0.00	0.00	0.02	0.13	0.01	0.15
	0–8	0.00	0.00	0.00	0.04	0.10	0.15
	Estimated Proportion Correctly Classified: Total = 0.71, Proficient & Above = 0.90						
Decision Consistency	33–40	0.18	0.04	0.00	0.00	0.00	0.21
	23–32	0.07	0.20	0.05	0.02	0.00	0.34
	18–22	0.00	0.05	0.05	0.04	0.00	0.15
Alternate Form *	9–17	0.00	0.00	0.02	0.10	0.02	0.15
	0–8	0.00	0.00	0.01	0.04	0.10	0.15
	Estimated Proportion Correctly Classified: Total = 0.63, Proficient & Above = 0.88						

* Values in table are proportions of the total sample.

† Inconsistencies with category cell entries are due to rounding.

Table 8.B.3 Decision Accuracy and Decision Consistency: Level II English–Language Arts

	Placement Score	Advanced	Proficient	Basic	Below Basic	Far Below Basic	Category Total †
Decision Accuracy	28–32	0.20	0.07	0.00	0.00	0.00	0.27
	20–27	0.04	0.36	0.04	0.00	0.00	0.45
	14–19	0.00	0.06	0.13	0.02	0.00	0.21
All-forms Average *	8–13	0.00	0.00	0.02	0.03	0.00	0.05
	0–7	0.00	0.00	0.00	0.01	0.01	0.02
	Estimated Proportion Correctly Classified: Total = 0.73 , Proficient & Above = 0.90						
Decision Consistency	28–32	0.19	0.08	0.00	0.00	0.00	0.27
	20–27	0.08	0.30	0.06	0.00	0.00	0.45
	14–19	0.00	0.07	0.10	0.03	0.00	0.21
Alternate Form *	8–13	0.00	0.00	0.02	0.03	0.00	0.05
	0–7	0.00	0.00	0.00	0.01	0.01	0.02
	Estimated Proportion Correctly Classified: Total =0.63, Proficient & Above = 0.87						

* Values in table are proportions of the total sample.

† Inconsistencies with category cell entries are due to rounding.

Table 8.B.4 Decision Accuracy and Decision Consistency: Level II Mathematics

	Placement Score	Advanced	Proficient	Basic	Below Basic	Far Below Basic	Category Total †
Decision Accuracy	23–32	0.40	0.05	0.00	0.00	0.00	0.45
	15–22	0.05	0.23	0.04	0.00	0.00	0.31
	10–14	0.00	0.05	0.09	0.02	0.01	0.16
All-forms Average *	8–9	0.00	0.00	0.02	0.01	0.01	0.04
	0–7	0.00	0.00	0.00	0.01	0.02	0.04
	Estimated Proportion Correctly Classified: Total = 0.75, Proficient & Above = 0.91						
Decision Consistency	23–32	0.39	0.07	0.00	0.00	0.00	0.45
	15–22	0.07	0.19	0.05	0.01	0.00	0.31
	10–14	0.00	0.05	0.07	0.02	0.02	0.16
Alternate Form *	8–9	0.00	0.00	0.02	0.01	0.01	0.04
	0–7	0.00	0.00	0.01	0.01	0.02	0.04
	Estimated Proportion Correctly Classified: Total =0.68, Proficient & Above =0.89						

* Values in table are proportions of the total sample.

† Inconsistencies with category cell entries are due to rounding.

Table 8.B.5 Decision Accuracy and Decision Consistency: Level III English–Language Arts

	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	0.35
	22–26	0.05	0.18	0.05	0.00	0.00	0.28
	17–21	0.00	0.05	0.12	0.03	0.00	0.21
All-forms Average *	11–16	0.00	0.00	0.03	0.07	0.00	0.11
	0–10	0.00	0.00	0.00	0.02	0.03	0.05
	Estimated Proportion Correctly Classified: Total = 0.68, Proficient & Above = 0.90						
Decision Consistency	27–32	0.27	0.07	0.01	0.00	0.00	0.35
	22–26	0.07	0.14	0.06	0.01	0.00	0.28
	17–21	0.01	0.06	0.09	0.04	0.00	0.21
Alternate Form *	11–16	0.00	0.00	0.03	0.06	0.01	0.11
	0–10	0.00	0.00	0.00	0.02	0.03	0.05
	Estimated Proportion Correctly Classified: Total = 0.59, Proficient & Above = 0.85						

* Values in table are proportions of the total sample.

† Inconsistencies with category cell entries are due to rounding.

Table 8.B.6 Decision Accuracy and Decision Consistency: Level III Mathematics

	Placement Score	Advanced	Proficient	Basic	Below Basic	Far Below Basic	Category Total †
Decision Accuracy	24–32	0.38	0.06	0.00	0.00	0.00	0.44
	16–23	0.04	0.23	0.02	0.00	0.00	0.30
	12–15	0.01	0.04	0.06	0.02	0.00	0.13
All-forms Average *	8–11	0.00	0.01	0.03	0.04	0.01	0.09
	0–7	0.00	0.00	0.00	0.01	0.02	0.03
	Estimated Proportion Correctly Classified: Total =0.73 , Proficient & Above = 0.92						
Decision Consistency	24–32	0.36	0.08	0.00	0.00	0.00	0.44
	16–23	0.07	0.19	0.04	0.01	0.00	0.30
	12–15	0.01	0.05	0.04	0.03	0.01	0.13
Alternate Form *	8–11	0.00	0.01	0.03	0.03	0.02	0.09
	0–7	0.00	0.00	0.00	0.01	0.03	0.03
	Estimated Proportion Correctly Classified: Total = 0.65, Proficient & Above = 0.88						

* Values in table are proportions of the total sample.

† Inconsistencies with category cell entries are due to rounding.

Table 8.B.7 Decision Accuracy and Decision Consistency: Level IV English–Language Arts

	Placement Score	Advanced	Proficient	Basic	Below Basic	Far Below Basic	Category Total †
Decision Accuracy	25–32	0.25	0.06	0.00	0.00	0.00	0.31
	19–24	0.04	0.20	0.05	0.00	0.00	0.29
	14–18	0.00	0.05	0.11	0.03	0.00	0.19
All-forms Average *	10–13	0.00	0.00	0.04	0.06	0.02	0.12
	0–9	0.00	0.00	0.00	0.02	0.07	0.09
	Estimated Proportion Correctly Classified: Total = 0.69, Proficient & Above = 0.90						
Decision Consistency	25–32	0.24	0.06	0.00	0.00	0.00	0.31
	19–24	0.06	0.16	0.06	0.01	0.00	0.29
	14–18	0.00	0.06	0.09	0.04	0.01	0.19
Alternate Form *	10–13	0.00	0.01	0.04	0.05	0.03	0.12
	0–9	0.00	0.00	0.00	0.02	0.07	0.09
	Estimated Proportion Correctly Classified: Total = 0.61, 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.8 Decision Accuracy and Decision Consistency: Level IV Mathematics

	Placement Score	Advanced	Proficient	Basic	Below Basic	Far Below Basic	Category Total †
Decision Accuracy	25–32	0.23	0.04	0.00	0.00	0.00	0.27
	19–24	0.03	0.15	0.04	0.00	0.00	0.22
	14–18	0.00	0.05	0.12	0.04	0.01	0.21
All-forms Average *	11–13	0.00	0.00	0.05	0.05	0.04	0.14
	0–10	0.00	0.00	0.00	0.03	0.12	0.15
	Estimated Proportion Correctly Classified: Total =0.67, Proficient & Above = 0.91						
Decision Consistency	25–32	0.22	0.05	0.01	0.00	0.00	0.27
	19–24	0.05	0.12	0.05	0.00	0.00	0.22
	14–18	0.01	0.05	0.09	0.04	0.02	0.21
Alternate Form *	11–13	0.00	0.01	0.04	0.04	0.04	0.14
	0–10	0.00	0.00	0.01	0.03	0.11	0.15
	Estimated Proportion Correctly Classified: Total = 0.58, 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.9 Decision Accuracy and Decision Consistency: Level V English–Language Arts

	Placement Score	Advanced	Proficient	Basic	Below Basic	Far Below Basic	Category Total †
Decision Accuracy	26–32	0.28	0.06	0.00	0.00	0.00	0.34
	20–25	0.04	0.20	0.05	0.00	0.00	0.30
	15–19	0.00	0.04	0.10	0.02	0.00	0.17
All-forms Average *	11–14	0.00	0.00	0.03	0.05	0.01	0.10
	0–10	0.00	0.00	0.00	0.02	0.07	0.10
	Estimated Proportion Correctly Classified: Total = 0.70, Proficient & Above = 0.91						
Decision Consistency	26–32	0.26	0.07	0.00	0.00	0.00	0.34
	20–25	0.07	0.16	0.06	0.01	0.00	0.30
	15–19	0.00	0.05	0.08	0.03	0.01	0.17
Alternate Form *	11–14	0.00	0.01	0.03	0.04	0.02	0.10
	0–10	0.00	0.00	0.01	0.02	0.07	0.10
	Estimated Proportion Correctly Classified: Total = 0.61 , 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 V Mathematics

	Placement Score	Advanced	Proficient	Basic	Below Basic	Far Below Basic	Category Total †
Decision Accuracy	29–32	0.15	0.05	0.00	0.00	0.00	0.21
	23–28	0.04	0.21	0.06	0.01	0.00	0.31
	16–22	0.00	0.04	0.16	0.02	0.00	0.23
All-forms Average *	12–15	0.00	0.00	0.04	0.05	0.02	0.10
	0–11	0.00	0.00	0.01	0.03	0.11	0.16
	Estimated Proportion Correctly Classified: Total = 0.68 , Proficient & Above = 0.89						
Decision Consistency	29–32	0.15	0.06	0.00	0.00	0.00	0.21
	23–28	0.06	0.16	0.07	0.01	0.00	0.31
	16–22	0.00	0.06	0.13	0.03	0.01	0.23
Alternate Form *	12–15	0.00	0.00	0.04	0.03	0.03	0.10
	0–11	0.00	0.00	0.02	0.03	0.10	0.16
	Estimated Proportion Correctly Classified: Total = 0.57 , Proficient & Above = 0.86						

* Values in table are proportions of the total sample.

† Inconsistencies with category cell entries are due to rounding.

Appendix 8.C—Score Conversions Based on 2008 Standard Setting

In fall 2008, a CAPA standard setting was held to establish performance-level cut scores for Levels I through V in English–language arts and mathematics and Levels I and Levels III through V in science. These cut scores will be implemented for the spring 2009 operational administration. For the purpose of creating impact data, data from the spring 2008 operational administration was used for estimation of all levels except Level I. Level I impact data and scoring conversions were not included because of the scoring rubric change to be implemented in the spring 2009 operational administration.

The tables in Appendix 8.C show, for Levels II through V in English–language arts and mathematics and Levels III through V in science, the raw-score-to-scale-score conversions, the CSEMs, and percent at each performance level. The information shown is the result of applying the cut scores and performance levels from the fall 2008 standard setting to the data from the spring 2008 operational administration of CAPA.

Table 8.C.1 Score Conversions: English–Language Arts Level II—Standard Setting, 2008

Raw Score	Scale Score	CSEM	Performance Level	% Students at Performance Level
32	60	17		
31	48	5		
30	45	3		
29	43	3	**Advanced**	**37.83**
28	42	2		
27	41	2		
26	40	2		
25	39	2		
24	38	2		
23	38	2		
22	37	2	**Proficient**	**39.21**
21	36	2		
20	36	2		
19	35	2		
18	34	2		
17	33	2		
16	33	2	**Basic**	**17.31**
15	32	2		
14	31	2		
13	30	2		
12	29	2		
11	28	2		
10	27	2		
9	26	2		
8	25	2	**Below Basic**	**5.07**
7	24	2		
6	23	3		
5	21	3		
4	19	3		
3	17	3		
2	15	4	**Far Below Basic**	**0.59**
1	15	4		
0	15	4		

Table 8.C.2 Score Conversions: English–Language Arts Level III—Standard Setting, 2008

Raw Score	Scale Score	CSEM	Performance Level	% Students at Performance Level
32	60	16		
31	48	4		
30	45	3		
29	43	3	**Advanced**	**40.59**
28	42	2		
27	41	2		
26	40	2		
25	39	2		
24	38	2		
23	38	2		
22	37	1	**Proficient**	**39.94**
21	37	1		
20	36	1		
19	36	1		
18	35	1		
17	34	1		
16	34	1		
15	33	2		
14	33	2	**Basic**	**15.08**
13	32	2		
12	31	2		
11	31	2		
10	30	2		
9	29	2		
8	28	2		
7	27	2	**Below Basic**	**3.30**
6	26	2		
5	25	2		
4	23	3		
3	22	3		
2	20	3	**Far Below Basic**	**1.10**
1	16	4		
0	15	5		

Table 8.C. 3 Score Conversions: English–Language Arts Level IV—Standard Setting, 2008

Raw Score	Scale Score	CSEM	Performance Level	% Students at Performance Level
32	60	11		
31	53	5		
30	50	4		
29	48	3		
28	46	3	**Advanced**	**35.13**
27	45	3		
26	44	2		
25	43	2		
24	42	2		
23	41	2		
22	40	2		
21	40	2		
20	39	2		
19	38	2	**Proficient**	**39.49**
18	37	2		
17	37	2		
16	36	2		
15	35	2		
14	34	2		
13	33	2		
12	32	2	**Basic**	**15.84**
11	31	3		
10	30	3		
9	29	3		
8	27	3		
7	25	3	**Below Basic**	**7.56**
6	23	4		
5	20	4		
4	18	4		
3	15	4		
2	15	4	**Far Below Basic**	**1.99**
1	15	4		
0	15	4		

Table 8.C. 4 Score Conversions: English–Language Arts Level V—Standard Setting, 2008

Raw Score	Scale Score	CSEM	Performance Level	% Students at Performance Level
32	60	17		
31	48	4		
30	45	3		
29	44	2	**Advanced**	**39.01**
28	42	2		
27	42	2		
26	41	2		
25	40	2		
24	39	2		
23	39	2		
22	38	1		
21	38	1		
20	37	1	**Proficient**	**38.04**
19	37	1		
18	36	1		
17	36	1		
16	35	1		
15	34	1		
14	34	1		
13	33	2		
12	33	2	**Basic**	**16.37**
11	32	2		
10	31	2		
9	30	2		
8	29	2		
7	27	2		
6	26	2	**Below Basic**	**4.64**
5	24	3		
4	23	3		
3	21	3		
2	19	3	**Far Below Basic**	**1.96**
1	15	4		
0	15	4		

Table 8.C.5 Score Conversions: Mathematics Level II—Standard Setting, 2008

Raw Score	Scale Score	CSEM	Performance Level	% Students at Performance Level
32	60	15		
31	50	7		
30	46	4	**Advanced**	**28.19**
29	44	4		
28	42	3		
27	41	3		
26	40	3		
25	39	3		
24	38	2		
23	37	2	**Proficient**	**28.54**
22	37	2		
21	36	2		
20	35	2		
19	34	2		
18	33	2		
17	33	2	**Basic**	**23.82**
16	32	2		
15	31	3		
14	30	3		
13	29	3		
12	28	3		
11	26	3		
10	25	3	**Below Basic**	**16.24**
9	23	4		
8	21	4		
7	18	4		
6	15	5		
5	15	5		
4	15	5		
3	15	5	**Far Below Basic**	**3.21**
2	15	5		
1	15	5		
0	15	5		

Table 8.C.6 Score Conversions: Mathematics Level III—Standard Setting, 2008

Raw Score	Scale Score	CSEM	Performance Level	% Students at Performance Level
32	60	16		
31	48	5		
30	44	3	**Advanced**	**28.57**
29	42	3		
28	41	3		
27	40	2		
26	39	2		
25	38	2		
24	37	2		
23	37	2	**Proficient**	**32.95**
22	36	2		
21	36	2		
20	35	2		
19	34	2		
18	34	2		
17	33	2		
16	33	2	**Basic**	**25.99**
15	32	2		
14	31	2		
13	31	2		
12	30	2		
11	29	2		
10	28	2		
9	27	3		
8	25	3	**Below Basic**	**11.05**
7	23	3		
6	21	3		
5	19	3		
4	16	3		
3	15	3		
2	15	3	**Far Below Basic**	**1.44**
1	15	3		
0	15	3		

Table 8.C.7 Score Conversions: Mathematics Level IV—Standard Setting, 2008

Raw Score	Scale Score	CSEM	Performance Level	% Students at Performance Level
32	60	12		
31	53	7		
30	49	5		
29	46	3	**Advanced**	**27.15**
28	45	3		
27	44	3		
26	43	3		
25	42	3		
24	41	3		
23	40	2		
22	39	2		
21	38	2	**Proficient**	**30.00**
20	37	2		
19	37	2		
18	36	2		
17	35	2		
16	34	3		
15	33	3		
14	32	3	**Basic**	**22.16**
13	31	3		
12	30	3		
11	29	3		
10	27	4		
9	25	4	**Below Basic**	**16.88**
8	22	5		
7	18	5		
6	15	5		
5	15	5		
4	15	5		
3	15	5	**Far Below Basic**	**3.82**
2	15	5		
1	15	5		
0	15	5		

Table 8.C.8 Score Conversions: Mathematics Level V—Standard Setting, 2008

Raw Score	Scale Score	CSEM	Performance Level	% Students at Performance Level
32	60	21		
31	47	6		
30	43	3		
29	42	3	**Advanced**	**36.70**
28	40	2		
27	39	2		
26	39	2		
25	38	2		
24	37	2		
23	37	2	**Proficient**	**26.58**
22	36	2		
21	36	2		
20	35	2		
19	34	2		
18	34	2		
17	33	2		
16	33	2	**Basic**	**20.91**
15	32	2		
14	32	2		
13	31	2		
12	30	2		
11	29	3		
10	28	3		
9	26	4	**Below Basic**	**12.84**
8	24	4		
7	21	4		
6	18	4		
5	15	4		
4	15	4		
3	15	4	**Far Below Basic**	**2.94**
2	15	4		
1	15	4		
0	15	4		

Table 8.C.9 Score Conversions: Science Level III—Standard Setting, 2008

Raw Score	Scale Score	CSEM	Performance Level	% Students at Performance Level
32	60	30		
31	45	4		
30	42	3	**Advanced**	**19.32**
29	41	2		
28	40	2		
27	39	2		
26	38	2		
25	37	1		
24	37	1		
23	36	1	**Proficient**	**46.93**
22	36	1		
21	35	1		
20	35	1		
19	35	1		
18	34	1		
17	34	1		
16	33	1		
15	33	1	**Basic**	**26.23**
14	32	1		
13	31	2		
12	31	2		
11	30	2		
10	29	2		
9	28	2		
8	27	2		
7	26	2	**Below Basic**	**6.04**
6	25	2		
5	23	2		
4	22	2		
3	20	3		
2	18	3	**Far Below Basic**	**1.46**
1	15	4		
0	15	4		

Table 8.C.10 Score Conversions: Science Level IV—Standard Setting, 2008

Raw Score	Scale Score	CSEM	Performance Level	% Students at Performance Level
32	60	21		
31	47	4		
30	44	3	**Advanced**	**15.86**
29	42	2		
28	41	2		
27	40	2		
26	39	2		
25	38	2		
24	38	2		
23	37	2	**Proficient**	**43.09**
22	37	2		
21	36	1		
20	36	1		
19	35	1		
18	34	1		
17	34	1		
16	33	2		
15	33	2	**Basic**	**32.28**
14	32	2		
13	32	2		
12	31	2		
11	30	2		
10	29	2		
9	28	2		
8	26	2		
7	25	3	**Below Basic**	**7.38**
6	23	3		
5	21	3		
4	20	3		
3	18	3		
2	15	3	**Far Below Basic**	**1.40**
1	15	3		
0	15	3		

Table 8.C.11 Score Conversions: Science Level V—Standard Setting, 2008

Raw Score	Scale Score	CSEM	Performance Level	% Students at Performance Level
32	60	23		
31	47	4		
30	44	3		
29	42	2	**Advanced**	**21.96**
28	41	2		
27	40	2		
26	39	2		
25	39	2		
24	38	2		
23	37	1		
22	37	1	**Proficient**	**36.96**
21	36	1		
20	36	1		
19	35	1		
18	34	1		
17	34	1		
16	33	1		
15	33	1	**Basic**	**30.34**
14	32	2		
13	32	2		
12	31	2		
11	30	2		
10	29	2		
9	28	2		
8	27	2		
7	25	2	**Below Basic**	**8.20**
6	24	2		
5	23	2		
4	21	2		
3	19	3		
2	17	3	**Far Below Basic**	**2.57**
1	15	4		
0	15	4		