Colorado
Student Assessment Program

Technical Report
2008

Submitted to the

Colorado Department of Education

October 2008

TABLE OF CONTENTS

OVERVIEW 1

PART 1: STANDARDS 2

Reading and Writing 2
The Colorado Model Content Standards 2
The Colorado Model Subcontent Areas 2

Mathematics 3
The Colorado Model Content Standards 3
The Colorado Model Subcontent Areas 4

Science 6
The Colorado Model Content Standards 6
The Colorado Model Subcontent Areas 6

PART 2: TEST DEVELOPMENT 7

Test Development and Content Validity 7

Test Configuration 8

CSAP Content Validity and Alignment Review 8

Universal Design and Plain Language in the Colorado Student Assessment Program 10

Linking Item (Anchor Item) Selection for the 2008 Assessments 11

Items Flagged as for Fit and DIF in Test Assembly 14

PART 3: ADMINISTRATION 16

Test Administration Training 17

Test Sections and Timing 17

PART 4: SCORING AND SCALING DESIGN 19

Test Scores for the Total Test and by Content Standard and Subcontent Area 19

Anchor Paper Review of New Constructed-Response Items 20

Rater Reliability and Severity 21
Interrater Reliability 21
Rater Severity/Leniency Study 22

Scaling Design 23

PART 5: ITEM ANALYSES 25

Third Grade 26
Reading 26
Reading – Spanish 26
Writing 27
Writing – Spanish 27
Mathematics 27

Fourth Grade 28
Reading 28
Reading – Spanish 28
Writing 29
Writing – Spanish 29
Mathematics 30

Fifth Grade 30
Reading 30
Writing 30
Mathematics 31
Science 31

Sixth Grade 32
Reading 32
Writing 32
Mathematics 32

Seventh Grade 33
Reading 33
Writing 33
Mathematics 34

Eighth Grade 34
Reading 34
Writing 35
Mathematics 35
Science 35

Ninth Grade 36
Reading 36
Writing 36
Mathematics 37

Tenth Grade 37
Reading 37
Writing 38
Mathematics 38
Science 38

PART 6: CALIBRATION AND EQUATING 40

Overview of the IRT Models 40

Calibration of the Assessment 41

Model Fit Analyses **42**

Model Fit Analyses Results **43**
Third Grade 43
Fourth Grade 44
Fifth Grade 44
Sixth Grade 44
Seventh Grade 44
Eighth Grade 45
Ninth Grade 45
Tenth Grade 45

Item Local Independence **45**

Evaluation of Item Analysis and Calibration **46**

Equating Procedures **47**
Anchor Items Evaluation Criteria 48
Anchor Items Evaluation Results 50
p-value Comparisons 50
Item Parameter Comparisons 50
Scaling Constants 51
Additional Analyses of Flagged Items 51

Effectiveness of the Equating **51**

PART 7: SCALE SCORE SUMMARY STATISTICS 53

Scale Score Distributions: Student Results **54**
Third Grade 54
Fourth Grade 57
Fifth Grade 59
Sixth Grade 62
Seventh Grade 63
Eighth Grade 65
Ninth Grade 68
Tenth Grade 70

Correlations Among Content Standards and Among Subcontent Areas **72**

PART 8: RELIABILITY AND VALIDITY EVIDENCE 74

Total Test and Subgroup Reliability **74**

Interrater Reliability, Item-to-Total Score Correlation, and DIF **76**

Standard Error of Measurement **77**

Test Validity **78**

Content-Related Validity **78**

Construct Validity **79**
Minimization of Construct-Irrelevant Variance and Under-Representation 79

Minimizing Bias Through DIF Analyses 80
Linn–Harnisch DIF Method 82

Differential Item Functioning Ratings and Results 83

Internal Factor Structure and Unidimensionality of the CSAP Assessment 85

IRT Model to Data Fit as an Evidence of Test Score Validity 85

Divergent (Discriminant) Validity 86

Predictive Validity 86

Classification Consistency and Accuracy 87

Classification Consistency and Accuracy When Pattern Scoring Is Used 88

PART 9: SPECIAL STUDY 91

Writing Trend Study 91

APPENDIX A: DEPTH OF KNOWLEDGE (DOK) LEVELS 93

REFERENCES 95

TABLES 98

FIGURES 483

Overview

This report presents the results of the statewide Spring 2008 administration of the Colorado Student Assessment Program (CSAP). In the spring of 2008, grades 3 through 10 students were assessed on Reading, Writing and Mathematics, and grades 5, 8, and 10 were also assessed on Science. Spanish versions of Reading and Writing tests were also administered in grades 3 and 4. The assessments were developed by CTB/McGraw-Hill, LLC in collaboration with the Colorado Department of Education and were scored and scaled by CTB/McGraw-Hill.

This report is organized in parts, with Part 1 providing an overview of the CSAP assessments, including descriptions of content standards and subcontent areas. Part 2 includes descriptions of test development, content validity, test configuration, and Differential Item Functioning (DIF) and fit in test assembly. Part 3 details test administration. Part 4 describes scoring and scaling design, including a description of test scores for total test and by content standard and subcontent area, interrater reliability, rater severity/leniency, and descriptions of scaling and scoring procedures. Detailed item analyses results including item-to-total score correlations or point-biserial, *p*-values, and omit rates are included in Part 5. Calibration and equating results including an overview of the Item Response Theory (IRT) models, model-to-data fit, item independence, and equating procedures are described in Part 6. Scale score summary statistics and correlations among content standards and subcontent areas are presented in Part 7. Part 8 contains reliability and validity evidence including total and subgroup reliability, test validity, content- and construct-related validity, and minimization of construct irrelevance variance and under-representation. Finally, Part 9 presents results of the Writing subscale trends for paragraph and extended writing.

Part 1: Standards

The CSAP assessments are developed to measure the Colorado content standards. Note that the terms "content standard" and "standard" are used synonymously throughout the text. Beginning in 2001, subcontent reporting categories were added at the request of the Colorado Department of Education to provide additional diagnostic information. Each subcontent area may cover several content standards. Most, but not all, of the items in CSAP are mapped to a subcontent area, whereas all items are mapped to one, and only one, content standard. The various content standards and subcontent areas are listed below for each content area. Table 1 provides an overview of which content standards and subcontent areas assessed in each of the grades.

Reading and Writing

The Colorado Model Content Standards

1) Reading Comprehension – Students read and understand a variety of materials. (Reading)

2) Write for a Variety of Purposes – Students write and speak for a variety of purposes and audiences. (Writing)

3) Write Using Conventions – Students write and speak using conventional grammar, usage, sentence structure, punctuation, capitalization, and spelling. (Writing)

4) Thinking Skills – Students apply thinking skills to their reading, writing, speaking, listening, and viewing. (Reading)

5) Use of Literary Information – Students read to locate, select, and make use of relevant information from a variety of media, reference, and technology source materials. (Reading)

6) Literature – Students read and recognize literature as a record of human experience. (Reading)

The Colorado Model Subcontent Areas

1) Fiction – Students read, predict, summarize, comprehend, and analyze fictional texts; determine the main idea and locate relevant information; and respond to literature that represents different points of view. (Reading)

2) Nonfiction – Students read, predict, summarize, comprehend, and analyze a variety of nonfiction texts including newspaper articles, biographies, and technical writings; locate the main idea and select relevant information; and determine the sequence of steps in technical writings. (Reading)

3) Vocabulary – Students use word recognition skills and resources such as phonics, context clues, word origins, and word order clues; root prefixes and suffixes of words. (Reading)

4) Poetry – Students read, predict, summarize and comprehend poetry; determine the main idea, make inferences, and draw conclusions; and respond to poetry that represents different points of view. (Reading)

5) Paragraph Writing – Students write and edit in a single session. (Writing)

6) Extended Writing – Students plan, organize, and revise writing for an extended essay. (Writing)

7) Grammar and Usage – Students know and use correct grammar in writing including parts of speech, pronouns, conventions, modifiers, sentence structure, and agreement. (Writing)

8) Mechanics – Students know and use conventions correctly including spelling, capitalization, and punctuation. (Writing)

Mathematics

The Colorado Model Content Standards

1) Number Sense – Students develop number sense, use numbers and number relationships in problem-solving situations, and communicate the reasoning used in solving these problems.

2) Algebra, Patterns, and Functions – Students use algebraic methods to explore, model, and describe patterns and functions involving numbers, shapes, data, and graphs in problem-solving situations and communicate the reasoning used in solving these problems.

3) Statistics and Probability – Students use data collection and analysis, statistics, and probability in problem-solving situations and communicate the reasoning used in solving these problems.

4) Geometry – Students use geometric concepts, properties, and relationships in problem-solving situations and communicate the reasoning used in solving these problems.

5) Measurement – Students use a variety of tools and techniques to measure, apply the results in problem-solving situations, and communicate the reasoning used in solving these problems.

6) Computational Techniques – Students link concepts and procedures as they develop and use computational techniques including estimation, mental arithmetic, paper and pencil, calculators, and computers in problem-solving situations, and communicate the reasoning used in solving these problems.

The Colorado Model Subcontent Areas

1) Number and Operation Sense –

- Students demonstrate meanings for whole numbers, commonly used fractions, decimals, and the four basic arithmetic operations through the use of drawings, decomposing and composing numbers, and identify factors, multiples, and prime/composite numbers. (SA 1, grades 4 and 5)

- Students demonstrate an understanding of relationships among benchmark fractions, decimals, and percents and justify the reasoning used. Students add and subtract fractions and decimals in problem-solving solutions. (SA 1, grade 6)

Number Sense – Students demonstrate understanding of the concept of equivalency as related to fractions, decimals, and percents. (SA 1, grade 7)

Linear Pattern Representation – Students represent, describe, and analyze linear patterns using tables, graphs, verbal rules, and standard algebraic notation and solve simple linear equations in problem-solving situations using a variety of methods. (SA 1, grade 8)

Multiple Representations of Linear/Nonlinear Functions – Students represent linear and nonlinear functional relationships modeling real-world phenomena using written explanations, tables, equations, and graphs; describe the connections among these representations; and convert from one representation to another. (SA 1, grade 9)

Multiple Representations of Functions – Students represent functional relationships that model real-world phenomena using written explanations, tables, equations, and graphs; describe the connections among these representations; and convert from one representation to another. (SA 1, grade 10)

2) Patterns –
   - Students reproduce, extend, create, and describe geometric and numeric patterns as problem-solving tools. (SA 2, grade 4)
   - Students represent, describe, and analyze geometric and numeric patterns using tables, graphs, and verbal rules as problem-solving tools. (SA 2, grade 5)
   - Students represent, describe, and analyze geometric and numeric patterns using tables, words, concrete objects, and pictures in problem-solving situations. (SA 2, grade 6)

   Area and Perimeter Relationships – Students demonstrate an understanding of perimeter, circumference, and area and recognize the relationships between them. (SA 2, grade 7)

   Proportional Thinking –
   - Students apply the concepts of ratio, proportion, scale factor, and similarity, including using the relationships among fractions, decimals, and percents, in problem-solving situations. (SA 2, grade 8)
   - Students apply the concepts of ratio and proportion in problem-solving situations. (SA 2, grade 9)

   Probability and Counting Techniques – Students apply organized counting techniques to determine a sample space and the theoretical probability of an identified event which includes differentiating between independent and dependent events and using area models to determine probability. (SA 2, grade 10)

3) Measurement – Students demonstrate knowledge of time, and understand the structure and use of U.S. customary and metric measurement tools and units. (SA 3, grade 4)

   Data Display – Students organize, construct, and interpret displays of data, including tables, charts, pictographs, line plots, bar graphs, and line graphs, and choose the correct graph from possible graph representations of a given scenario. (SA 3, grade 5)

   Geometry –
   - Students will reason informally about the properties of two-dimensional figures and solve problems involving area and perimeter. (SA 3, grade 6)
   - Students describe, analyze, and reason informally about the properties of two- and three-dimensional figures to solve problems. (SA 3, grade 8)

Science

The Colorado Model Content Standards

1) Scientific Investigation– Students apply the processes of scientific investigation and design, conduct, communicate about, and evaluate such investigations

2) Physical Science– Students know and understand common properties, forms, and changes in matter and energy. *(Focus: Physics and Chemistry)*

3) Life Science– Students knows and understand the characteristics and structure of living things, the processes of life, and how living things interact with each other and their environment. *(Focus: Biology-- Anatomy, Physiology, Botany, Zoology, Ecology)*

4) Earth and Space Science– Students know and understand the processes and interactions of Earth's systems and the structure and dynamics of Earth and other objects in space. *(Focus: Geology, Meteorology, Astronomy, Oceanography)*

5) The Nature of Science – Students understand that the nature of science involves a particular way of building knowledge and making meaning of the natural world

The Colorado Model Subcontent Areas

1) Experimental Design and Investigations – Student design, plan, and conduct a variety of investigations; understands and applies scientific questions, hypotheses, variables, and experimental design

2) Results and Data Analysis – Student select and use appropriate technology; organizes, analyzes, interprets, and predicts from scientific data in order to communicate the results of investigations

3) Physics Concepts – Student understands physical forces, the motion of objects, and energy transfer or energy transformation.

4) Chemistry Concepts – Student understands the properties, composition, structure and changes of matter..

5) Life Process – Student understands levels of organization in organisms, cellular structure and processes, and concepts in heredity.

6) Geology and Astronomy – Student understands Earth's composition, energy resources, plate movement, and characteristics of different celestial objects in the universe and how they interact with one another.

Part 2: Test Development

Content-related validity in achievement tests is evidenced by a correspondence between test content and a specification of the content domain. Content-related validity can be demonstrated through consistent adherence to test blueprints and through a high-quality test development process that includes review of items for accessibility by various subgroups including English Language Learner and students with disabilities. Part 2 provides an overview of the CSAP test design and the development of student assessments that assist stakeholders in making informed educational decisions. Specifically, it describes the CSAP test development activities in terms of content validity; test configuration; content revision in terms of sensitivity, bias, and plain language; selection of linking items for maintaining scales; model-to-data fit and Differential Item Functioning (DIF) in 2008 assessments.

Test Development and Content Validity

Content-related validity can be defined as the degree to which elements of an assessment instrument are relevant to and representative of the targeted construct for a particular assessment purpose. In order to ensure the content-related validity of the CSAP assessments, the Colorado Model Content Standards and Assessment Frameworks were studied by CTB's content developers who worked with Colorado content area specialists, teachers, and assessment experts to develop a pool of items that measured Colorado's Assessment Frameworks in each grade and content area. Several sources contributed to the 2008 CSAP items. CTB/McGraw-Hill's extensive pool of previously field-tested Reading passages, Writing prompts, Mathematics, and Science items provided the initial source. Many of these existing items were revised in order to ensure accessibility by different student groups and better measurement of the relevant Colorado standards and benchmarks. Additional items were developed by CTB and the staff at the Colorado Department of Education as needed to complete the alignment of CSAP to the Assessment Frameworks. These items were carefully reviewed under plain language revision and discussed by Content Validity and Alignment Review committees to ensure not only content validity, but also the quality and appropriateness of the items. These committees represented Colorado's diverse population and included Colorado teachers, community members, and State Department of Education staffs. The committees' recommendations were used to select and/or revise items from the item pool to construct the final Reading, Writing, Mathematics, and Science assessments.

Each new form also included a subset of multiple-choice items used in the previous administrations of the CSAP assessments as an anchor set. These repeated items were used to equate the forms across years. Equating is

necessary to account for slight year-to-year differences in test difficulty and to maintain scale comparability across years. Details of the equating are provided later in Part 6 of this report. The assessments that were reported on vertical scales (English Reading, English Writing, and Mathematics) also had items in common between adjacent grades. In grades 3 and 4 Spanish Reading and Writing test, same forms administered in previous administrations were used.

Test Configuration

Tables 2 through 6 provide information regarding the configuration of the CSAP assessments. Table 2 provides the number of multiple-choice (MC) and constructed-response (CR) items on each test, as well as the number of obtainable score points on each CR item. Tables 3 through 6 provide the number of MC and CR items by content standard (CS) and subcontent area (SA). Note that the subcontent areas Fiction (SA 1) and Poetry (SA 4) are combined for grades 3 through 6 Reading. The following content standards are also combined: Algebra, Patterns, and Functions (CS 2) and Statistics and Probability (CS 3) in grade 3 Mathematics; Number Sense (CS 1) and Computational Techniques (CS 6) in grades 7 through 10 Mathematics; Geometry (CS 4) and Measurement (CS 5) in grades 3 through 10 Mathematics; Scientific Investigations and Connections Among Scientific Disciplines (CS 1/6), Physical Science and Its Interrelationship With Technology and Human Activity (CS 2/5), Life Science and Its Interrelationship With Technology and Human Activity (CS 3/5), Earth and Space Science and Its Interrelationship With Technology and Human Activity (CS 4/5) in grades 5, 8, and 10 Science.

Every item is associated with a content standard but not all items are associated with a subcontent area. For this reason, the sum of the subcontent area points may be less than the total number of points for the test.

Appendix A provides the DOK level distribution for the CSAP 2008 assessments. CDE is in the process of specifying suggested DOK distributions for each of the CSAP assessments with input from stakeholders and the Technical Advisory Committee. DOK distribution will be articulated in the blueprint for the 2010 CSAP assessments.

CSAP Content Validity and Alignment Review

The items that appeared in 2008 CSAP tests were carefully reviewed and discussed in June 2007 by Content Validity and Alignment Review committees to ensure content validity, accurate alignment to content standards, and the quality and appropriateness of the items, including review for bias and sensitivity issues. These committees represented Colorado's diverse population and

included Colorado teachers, community members, and State Department of Education staff.

Specific areas of focus of the Content Review committees included:

- alignment of items to assessment objectives under the Colorado Model Content Standards and Assessment Frameworks and depths of knowledge
- accuracy grade-level appropriateness of items
- accessibility of items to all Colorado students, using Universal Design and Plain Language principles
- appropriateness and usability of scoring guides for constructed-response items
- processes for alignment review were designed to ensure that:
    1) reviews resulted in an independent alignment recommendation by each reviewer
    2) thorough discussion of appropriate alignment occurred following the independent reviews
    3) thorough documentation of alignment findings were captured
- processes for bias and sensitivity review were designed to ensure that:
    1) items were neither advantageous nor disadvantageous to a specific group of students
    2) items did not stereotype specific groups
    3) items did not promote personal, moral, or religious values or viewpoints
    4) students' achievement on a given test item is dependent solely on what they know and are able to do

The committees' feedback was reconciled by CDE and CTB staff and used to select and/or revise items from the item pool to construct the final Reading, Writing, Mathematics, and Science assessments.

Universal Design and Plain Language in the Colorado Student Assessment Program

As indicated in the previous section, one of the focuses of the CSAP content review was the application of Universal Design in test assembly. CSAP measures what students know and are able to do as defined in the Colorado Model Content Standards. Assessment must ensure comprehensible access to this content. CDE's and CTB's content experts revised the item pool and removed unnecessary verbiage from the 2008 CSAP tests so that students could show what they know and are able to do. Areas of focus included directions, writing prompts, test questions, and answer choices. New items developed for 2008 were authored using these principles. Items previously developed and administered prior to 2008 were also modified to conform to these principles:

Aspects of Universal Design

- ❑ Precisely Defined Constructs
  - Direct match to objective being measured
- ❑ Accessible, Nonbiased Items
  - Ensure ability to use accommodations from the start (Braille, oral presentation)
  - Ensure that quality is retained in all items
- ❑ Simple, Clear Directions and Procedures
  - Presented in understandable language
  - Consistency in procedures and format in all content areas
- ❑ Maximum Legibility
  - Simple fonts
  - Use of white space
  - Headings and graphic arrangement
    - Direct attention to relative importance
    - Direct attention to the order in which content should be considered
- ❑ Maximum Readability: Plain Language
  - The use of plain language in CSAP
    - Increases validity to the measurement of the construct
    - Increases the accuracy of the inferences made from the resulting data
  - Plain Language in CSAP uses
    - Active instead of passive voice
    - Short sentences
    - Common, everyday words
    - Purposeful graphics to clarify what is being asked

Linking Item (Anchor Item) Selection for the 2008 Assessments

In order to equate current tests to base year scale, a set of 18–23 multiple-choice anchor items was selected for each grade in Reading, Writing, Mathematics, and grade 3 Spanish Reading and Writing of the 2008 assessments. These items demonstrated good classical and IRT statistics and represented the test blueprint. Equating is necessary to account for slight differences in test difficulty and maintain scale comparability across administrations. Details of the equating are provided in Part 6. For grades 5, 8, and 10 Science the standards from previous administrations were reviewed and set with a new scale for each of the grade. Therefore, no linking was necessary for this administration.

Spanish tests for grades 3 and 4 Reading and Writing were constructed with only items that had been previously administered and successfully calibrated (and not changed). All preequated item parameters were used for scoring grade 4 Spanish Reading and Writing. This constraint was imposed because of the diminishing number of students taking the grade 4 Spanish tests in recent years. Grade 3 Spanish tests were recalibrated, as the sample size allowed, in order to obtain postequated item parameters and linked to original scale using a set of anchor items. The following criteria were followed to select anchor items in all content areas:

Content Representation and Item Difficulty – Content representation is one of the two most important criteria for anchor item selection. The items in an anchor set should represent a miniature version of the form. The other critical criterion is the spread of item difficulties across the difficulty range of the test. The item difficulty values for anchor items should cover the item difficulty range in the test, but generally should *not* include extremely easy ($p > 0.90$) or extremely difficult ($p < 0.25$) items. However, a recent study by Sinharay and Holland (2007) indicated that the anchor set difficulty range mirroring the complete form is not necessarily optimal. In any case, one way to think of selecting anchor items is to select "the best items" in the pool.

Number of Anchor Items and Item Format Representation – The 2008 CSAP tests included 18–23 anchor items for each grade and content area. Only multiple-choice items were selected as anchors.[1] For anchor items associated with a passage, all items originally included with the passage were readministered. The length of the passage associated with the anchor items was not extreme relative to the length of other passages in the form in which they served as anchors.

Relative Item Position in a Form – Anchor items were placed in the same relative position in the form as they were previously administered. The position of items can affect their performance. For this reason, the position of each anchor item

[1] When only MC items are used as anchors, it is assumed that the CR items do not measure a significant performance characteristic unique to that item format.

on the new form was as close as possible to its position on the form in which it appeared previously. A minimum requirement was that they be placed in the same third of the form as they were previously administered. Similarly, it was required that the item sets (testlets) with common stimuli be placed on the same side of the two open pages.

It was also required that the anchor items be interspersed throughout the test, not placed at the very beginning or end of a form, and the anchor items associated with passages should not occur within any part of the latter portion of a form or session where speededness effects may occur.

Item Characteristics – Content experts *avoided* using items in the anchor sets with

- Point biserials on the correct answer ≤ 0.18
- Positive point biserials on the distractors
- *p*-value ≤ 0.25 or ≥ 0.90
- Omit rates $\geq 5\%$

For all items, content experts *minimized* the use of items with poor fit statistics (Q1) or significant differential item functioning (DIF) statistics for gender or ethnicity. If it was essential to include an item with DIF, counterbalancing was suggested with an item exhibiting bias in the opposite direction. The number of items flagged for poor fit and DIF in 2008 test are listed and described later in this section, under the heading "Items Flagged for Fit and DIF in Test Assembly."

Form Characteristics – The test characteristic curves (TCC) and standard error (SE) curves of the total test and the anchor set overlaid each other as closely as possible. Since only MC items were used for anchor and the test consisted of both MC and CR items the alignment of the TCCs was difficult for some grades/content areas. In that situation, content developers attempted to match the anchor item TCC with MC items only test TCC. The maximum expected percent difference between TCCs was expected to be less than 0.05. In case this could not be met, content experts met this criterion at cut points. For tests that were vertically scaled, the TCC was sequentially aligned as the grade level increased.[2]

Changes to the Items – The psychometric properties of the anchor items were expected to be stable over various administrations. During 2008 core item review, it became evident to CDE that some anchor items differed in appearance from the other operational items because the Simple Language and Universal Design principles had not been applied to these anchor items. While CDE and

[2] Some overlaps at either the top or bottom end of the TCCs may be permissible. However, a significant overlap in the middle portion is not allowed.

CTB realize that editing all items to comply with Simple Language is gradual, CDE has requested that CTB provide a plan that will accelerate the process.[3] CTB came up with a list of items from the item pool that required attention in order to comply with universal design principles, and CTB's Research and Development staff met to discuss possible strategies to accelerate the application of Simple Language/Universal Design changes to anchors. CDE and CTB then met and classified the types of anchor item revisions into two categories. Revisions that are not expected to alter student response pattern in terms of item performance are considered minor revisions. On the other hand, revisions that are likely to affect the student response pattern are considered major revisions. The minor and major revisions are outlined below:

Minor revisions included some of the following:

1) Line break and line space for question when stem is two or more sentences preceding the question

2) Change proper names to generic names. We already can change proper names to "a student," "a teacher," "a farmer," and so on. We will now be able to expand that to changing names such as "The Denver Museum of Nature and Science" to "a museum."

3) Boldface words and phrases like "best", "most likely", "least likely", and "main"

4) Removing the unnecessary words like "below" and "the following"

Major revisions included some of the following:

1) Reduce reading by substituting "Study the table" (diagram, graph, etc.) in place of sentences describing the table

2) Create bullet list in place of sentences describing the information

3) Wording changes and deleting significant number of words

4) Changes to art.

5) Other case-by-case changes

For each grade and content area, the following procedure was used to identify up to two anchor items that could receive major revisions:

[3] For the details of Universal Design applications, refer to the previous section "Universal Design and Plain Language in the Colorado Student Assessment Program."

1) Identify standards that are represented by more anchors than required by the blueprint. Limiting revisions to items that measure these standards will minimize the impact on representation if an anchor item must be dropped from the anchor set after the administration because of aberrant statistics.

2) Determine the two worst offenders of simple language, but at the same time these two items may minimally effect student response. This parameter is set because the more substantial revisions are, the more risk there is that data will change and affect the performance of the item as an anchor. However, Research was prepared to drop the two anchor items with major revisions if flagged by the evaluation criteria, and the revision of "worst offenders" was a priority for CDE.

Very few, Reading and Writing anchor items required minor revision (fewer than two items per grade and content area), but substantial numbers of Mathematics anchor items received minor revisions. Similarly, Reading and Writing required no major revised anchor items. However, Mathematics anchor set across all grades required the maximum of two major revised anchor items in order to align test structure, match blueprint, and accelerate plain language revision process to have all items under Plain Language revision by 2010.

In order to ensure that the characteristics of the revised anchor items in terms of p-values and IRT parameters for the revised anchors remained statistically similar across the two administrations, three anchor item evaluation methods were put in place: DeltaPlot (Angoff, 1972; Dorans & Holland, 1993), Chi-Square (Lord, 1980), and the TCC method (Stocking & Lord, 1983) procedure. The TCC method was used for initial screening. DeltaPlot and Chi-Square methods were used for further evaluation. If an anchor item is flagged by the four criteria (two evaluation methods, the p-value difference greater than 0.10, and did not violate blueprint representation) the anchor item was dropped from the anchor set.

Only one item (grade 10 Mathematics, item 52) met the criteria and was dropped from the anchor set. Grade 10 Reading, item 114 displayed aberrant characteristics during the item analysis and calibration results review. With the consultation of CTB and CDE content experts the item was suppressed from the test, hence from the anchor set.

Items Flagged as for Fit and DIF in Test Assembly

The items flagged for poor fit and DIF were avoided as much as possible when assembling the 2008 assessments. As a guideline, if it was essential to include an item with poor fit in the test in order to meet the test blueprint, it should be with only marginally poor fit, with *p*-value and item-to-total score correlation in a reasonable range. Similarly, if it was essential to include an item with DIF, content experts were instructed to minimize overall bias by counterbalancing with

an item exhibiting bias in the opposite direction. Moreover, prior to including the item(s) flagged for DIF in the final forms, items were reviewed and judged to be fair by educational community professionals who represent various ethnic groups.

Table 7 displays the items with DIF and fit flags from previous administrations across all operational items for the 2008 assembled test forms. For the 1225 operational English Mathematics, Reading, Writing, and Science test items with available statistics, 41 (3.3%) were flagged for marginal poor fit and 5 (0.4%) as DIF for the gender and ethnic subgroups. Only 5 of these items were used as anchors in 2008. Of the 216 previously used Spanish items, 30 (13.9%) were flagged for marginal poor fit and none were flagged for gender DIF. Only 3 of these items were used as anchors in 2008. As mentioned above, the poor fit was marginal for most items and their inclusion in the tests was essential to meet test blueprint for content standards.

Part 3: Administration

The Colorado Student Assessment Program (CSAP) is Colorado's large-scale standardized paper–pencil achievement test administered every year. In 2008, grade 3 Reading (English and Spanish) assessments were administered between January 28 and February 29. The rest of the English language grades and content areas were administered between March 3 and April 11. In addition, grade 4 Spanish Reading and grades 3 and 4 Spanish Writing tests were administered during the same period. The purpose of the CSAP is to provide an annual measure of student performance relative to the Colorado Model Content Standards. All CSAP forms are timed, standardized assessments administered under standardized conditions to ensure the reliability and validity of the test results. All students in grades 3 through 10 Reading/Writing and Mathematics and grades 5, 8, and 10 Science were tested with a single form for each grade. Following accommodations were allowed to students on the basis of demonstrated need.

1 = Braille version
2 = Large-print version
3 = Teacher-read directions only
4 = Use of manipulatives (Not applicable to Reading and Writing)
5 = Scribe
6 = Signing
7 = Assistive communication device
8 = Extended timing used
9 = Oral script (Not applicable to Reading)
A = Appr. nonstandard accomm
B = Translated oral script (Not applicable to Reading)
C = Word-to-Word dictionary (Not applicable to Reading)

Prior to test administration, accommodation requests were documented in a formal plan created for each individual student by a team of teaching professionals, including the parents. The accommodations provided students equal opportunity to access information and to demonstrate knowledge and skills without affecting the reliability and validity of the assessment. For detailed information regarding the test administration or accommodations, please refer to the 2008 test administration manual and the Colorado accommodations manual (Colorado Department of Education, 2008.)

The sections below briefly describe the training conducted before the test administration to ensure proper handling of test materials, conduct testing, and return the materials securely to the scoring center. This information is followed by the number of sessions in each test and the time given to complete the test.

Test Administration Training

Prior to the actual testing window, CDE, with support from CTB, conducted pretest administration training for the 2008 CSAP. The live training consisted of an overview of CDE policies and procedures for the administration of the CSAP tests. Training included proper use of the CSAP Test Proctor's manuals and the District Assessment Coordinator/School Assessment Coordinator (DAC/SAC) manuals.

The Test Proctor's manuals provided specific detailed instruction on proper administration of the CSAP tests. The manuals provided detailed definitions of the CSAP test proctors' responsibilities, the purpose of the test, security before and during the test, and chain-of-custody guidance to ensure that all students took the tests in a standardized manner (same time, same test, with no student interaction). The manuals also provided a list of test materials authorized and required for testing. Prior to test administration, the CSAP test proctors were responsible for ensuring that an adequate supply of the materials required for testing would be available in testing rooms.

The DAC/SAC manuals provided instruction to the District Assessment Coordinator and the School Assessment Coordinator on how to distribute, safeguard, collect, package, and ship the completed test books to CTB for scoring. Test administrators were instructed to return all test books (both used and unused) to CTB.

CDE scheduled and conducted several regional test administration training sessions. The attendees at these sessions were district assessment coordinators and administrators. CDE stressed policy and procedure guidance as well as test administration training during these sessions. District and school assessment coordinators were required to provide training to all test proctors.

The CSAP Test Proctor's manual and the CSAP DAC/SAC manual can be found at www.ctb.com/csap.

Test Sections and Timing

Although the 2008 CSAP tests were administered independently, the CSAP Reading and Writing tests were combined in a single testbook for grades 4 through 10 with six sections: three sections for Reading and three for Writing. Grade 3 Reading and Writing tests were not combined into one booklet (for both English and Spanish versions) as they were administered at separate times of the year. In grade 3 there were two sections for Reading and two for Writing. Similarly, there were two sections each for grade 3 Spanish Reading and Writing and three sections each for grade 4 Spanish Reading and Writing. For Mathematics, there were three sections for grades 4 through 10 tests and two

sections for the grade 3 test. For Science, grades 5, 8, and 10 each had three sections.

Test developers also considered speededness in the development of the CSAP assessments. CTB believes that achievement tests should not be speeded; little or no useful instructional information can be obtained from the fact that a student did not finish a test, whereas a great deal can be learned from student responses to questions. In CSAP tests, students were allowed a maximum of 60 minutes for each session in Reading/Writing and 65 minutes in Mathematics and Science. The analysis of omit rates of the items showed no indication of speededness in the CSAP assessments. See Part 5 for further details on omit ranges.

Part 4: Scoring and Scaling Design

Part 4 describes scoring procedures for the total test, followed by scoring of constructed-response (CR) items. The succeeding sections describe rater reliability and rater severity. Finally, Part 4 wraps up with a detailed description of scaling design for the 2008 CSAP assessments.

Test Scores for the Total Test and by Content Standard and Subcontent Area

In the CSAP tests, students' total scores are based on their performance on all the scored items on the test. The range of possible scores varies by grade and content area. The highest obtainable scale score (HOSS) and lowest obtainable scale score (LOSS) for each grade and content area is provided in Table 8. CSAP reports item pattern scores and the HOSS is set slightly higher for most grades and content areas for allowing student's growth to be reflected in the subsequent administrations. The HOSS for grade 3 Reading is different from Grades 4 through 10 because grade 3 responses were scaled separately when the scale was set and grade 3 scores have been reported earlier than the rest of the grades. The same LOSS and HOSS are maintained over the years in all grades and content areas. Students also receive a score for each content standard (and for each subcontent area) that is based only on the items that contribute to the given content standard (or subcontent area). Note that every item on the test corresponds to some content standard, but not all items contribute to a subcontent area. The scale scores for the content standards and the subcontent areas are calculated using the item parameters that are obtained when the *total* test is calibrated (see Part 6). For each grade and content area, the minimum and maximum possible scale scores for content standards and subcontent areas are set at the same LOSS and HOSS as the total scale score.

Students were scored at the total test, content standard, and subcontent area levels using item response theory (IRT) item-pattern (IP) scoring procedure. This procedure produces maximum likelihood trait estimates (scale scores) based on students' item response patterns, as described by Lord (1974, 1980, pp. 179–181). Pattern scoring, based on IRT, takes into account which items a student answered correctly and produces better test information, less measurement error, and greater reliability than number-correct scoring. Moreover, pattern scoring produces more accurate scores for individual students. On average, the increase in accuracy is equivalent to approximately a 15%–20% increase in test length (Yen, 1984; Yen & Candell, 1991). Note that score reliability tends to increase with the number of items, and thus the total score is more reliable than the content standard or subcontent area scores.

For the new constructed-response items in the test, an anchor paper review meeting was conducted where content experts from CDE and CTB analyzed student responses by pulling samples and establishing guidelines for assigning different score points based on the writing proficiency. The procedure is described below.

Anchor Paper Review of New Constructed-Response Items

CDE and CTB conducted an "anchor paper" (also called "range finding") review of new constructed-response (CR) items on the 2008 CSAP tests. CTB's hand-scoring supervisors reviewed student written responses to CR items. Using scoring guides and rubrics prepared by CTB's content developers, CTB's supervisors selected responses that they determined were representative of students who demonstrated various levels of proficiency and understanding of the concepts being assessed. Supervisors annotated the sample anchor papers with their comments and logic for assigning scores.

The hand-scoring supervisors also reviewed anchor papers for CR items that were used in previous years' versions of the tests. If items were revised or there was reason to believe that a review should be conducted to obtain fresh anchor papers, the supervisors included sample anchor papers in the review package.

CTB's handscoring supervisors prepared anchor paper review packets for the various grade and contents to be reviewed with Colorado teachers at a live session in Denver, Colorado, in early April 2008.

At the CSAP 2008 anchor paper review, CTB's supervisors distributed numbered packets containing the established scoring guide and proposed and annotated anchors for all new items in 2008.

CTB's supervisors led discussion of each proposed, annotated anchor paper for each reviewed CR item, beginning with the top score point and continuing in reverse order to the lowest score point. Annotations were amended when necessary so that they more closely reflected the teacher-informed scoring stance for the item.

The review participants approved the proposed anchors or selected an alternative anchor for all items reviewed. A Colorado participant, appointed by a CDE consultant, verified the approval of the anchor by signing and dating a copy of each anchor. In the event that one or more anchors for that item are deemed ineffective, participants chose from other sample responses for a replacement. CTB's supervisor, if appropriate, suggested other student responses from additional materials brought to the review.

After the committee of teachers reviewed and approved the scores and annotations of the anchors, members continued to review additional responses

that the supervisor deemed questionable. The approved score, as well as a brief synopsis of the scoring philosophy behind the decision, was recorded by CTB's supervisor.

The reviewed and annotated anchor papers served as the basis for conducting hand-scoring training for CSAP 2008 at CTB scoring facility.

Rater Reliability and Severity

The CSAP test design framework includes a variety of different item types, including short response and extended constructed-response items. Although constructed-response items greatly enhance the construct and instructional validity of CSAP, reliability of hand-scoring items should be closely examined and documented. Through the ongoing process of training and research analyses, evidence of reliability of hand-scoring was continuously gathered. Many training and monitoring techniques were used to ensure scoring reliability and accuracy. Scoring guides are carefully developed and refined; scorers were trained, calibrated, and monitored throughout the scoring process; and rater reliability indices were generated and examined. Reliability for constructed-response items was typically examined by calculating indices of interrater agreement—the reliability with which human raters assign scores to student responses. For this analysis, a certain percentage of student responses are scored by two raters.

Interrater Reliability

To measure interrater reliability *within* the 2008 CSAP administration, approximately 5% of the constructed-response items scored were read by a second reader, a blind double read, and the resulting scores documented and analyzed. For Spanish, approximately 15% of the constructed-response items were a blind double read. Evidence supporting interrater reliability of CSAP assessments is presented in terms of raw score means, raw score standard deviations, and percentages of exact and adjacent agreement between raters. Exact agreement is defined as scores that are exactly the same. Adjacent agreement is defined as scores differing by one point. In addition, Cohen's kappa (Cohen, 1960) is provided as a measure of agreement between the raters and is commonly used to summarize the agreement between raters. It is computed as (Brennan & Prediger, 1981)

$$\kappa = \frac{\sum P_{ii} - \sum P_{i\cdot} P_{\cdot i}}{1 - \sum P_{i\cdot} P_{\cdot i}},$$

where $\sum P_{ii}$ is the observed proportion of agreement and $\sum P_{i\cdot} P_{\cdot i}$ is the chance proportion of agreement. Tables 9 through 14 show the rater reliability indices

for all 2008 constructed-response items by content area. The results indicate that the kappa is reasonably high for all grades and content areas.

Rater Severity/Leniency Study

In addition to examining rater reliability measures within a given administration year, CTB conducts a rater severity study *across* years. Rater severity or leniency is defined as the extent to which scores assigned by raters across years are systematically offset. The study entails sampling student responses from previous administrations, having a representative group of raters from the current administration score them, and comparing the scores against the scores assigned by the previous raters. Table 15 shows the number of rater severity items used in the study by content area and grade. The following specifications describe the rater severity study in detail:

1) In 2008, a rater severity study was done using constructed-response items that were repeated from 2006.

2) Random samples of student responses were selected from the 2006 CSAP tests for tests where repeating items were present:

   - A random sample of approximately 1,000 students was selected for English Reading, English Writing, and Mathematics assessments.

   - A random sample of approximately 250 students was selected for Spanish Reading and Spanish Writing assessments.

   - A random sample of 1000 students was selected for grades 8 and 10 Science. Because of a lack of qualified items in grade 5 Science, this grade/content area was not represented.

3) The samples of papers were administered blindly to the 2008 raters during the second half of 2008 operational scoring; that is, the raters scoring the papers from a previous administration ideally knew neither that the papers had been scored before, nor that they came from the 2006 data. The items to be rescored were shown to the 2008 raters under their 2008 item numbers (see Table 15). Because of minor revision to items stemming from the implementation of Universal Design that occurred for the first time in 2007, raters were able to tell that they were looking at items from multiple years for some constructed-response items, though they were not aware of the previous rating of the item.

4) The scores from the rescore were then compared with the original scores given to the papers by the raters in 2006.

Table 15 shows results of the rater severity study, including mean scores from the 2006 administration; mean scores from the 2008 administration; percent of the scores with exact, adjacent, and discrepant agreement; weighted kappa; correlation; and intraclass correlation.

Weighted kappa, which may be interpreted as the chance-corrected weighted proportional agreement, is reasonably high for the items in most of the content areas (0.58–0.95 with median of 0.84 for Reading, 0.69–0.95 with median of 0.86 for Mathematics, 0.64–0.98 with median of 0.83 for Science). For Writing, the weighted kappa ranged from 0.58 to 0.84 with median value of 0.68 for items without parts. One constructed-response item consisted of multiple-parts each in grade 3 English and Spanish versions and had lower weighted kappas (0.18–0.42 for English and 0.44–0.55 for Spanish version) even though the exact (precise) rater agreements were generally high (approximately 81%–99% for English version and 70%–89% for Spanish version). These items both have multiple parts A–D, with part A having two score points and parts B–D having one score point. Since weighted kappa corrects for chance, with 0 or 1 score points it is likely that higher proportion of students are getting one of the two scores by chance. Notice from Table 15 that the correlations between the two raters for these parts are also low. Similar results were observed last year as well.

Scaling Design

Horizontal equating within each grade was used to place the 2008 forms on the vertical scales that had been established previously for English Reading, Writing, and Mathematics. The vertical scale for English Reading, spanning grades 3 through 10, was established in 2001. The vertical scales for English Writing, spanning grades 3 through 10, and for Mathematics, spanning grades 5 through 10, were established in 2002. Grades 3 and 4 Mathematics were added to the vertical scale in 2005. Stocking and Lord's (1983) procedure was used to place each grade on the vertical scale that had been developed for each content area.

Because of the nonincremental nature of the content standards and the gaps in grade levels, grades 5, 8, and 10 Science were not placed in vertical scale. The Science standards adopted were based on a standard setting review meeting that took place in 2008 (McGraw-Hill, 2008).

Note that the customized versions of the grades 3 and 4 Reading and Writing assessments in Spanish were first administered in 1998. The year before, Supera had been administered to those students eligible for taking a Spanish language version assessment. The customized Spanish version that was first created in 1998 was repeated without modification through 2001. In 2002, new forms were created for the Spanish language assessments, which served as a source for the future tests. Every year thereafter, a new form has been created to meet the Colorado blueprint by selecting psychometrically good quality items

from the existing item pool. Although grades 3 and 4 Spanish tests are designed to measure a student's developmental scale over time, they are not in vertical scale. In 2008, grade 3 Spanish Reading and Writing items were live calibrated in order to estimate item parameters. For grade 4 Spanish Reading and Writing tests, preequated item parameters were used to score student responses because of the diminishing number of students in the grade.

Each 2008 CSAP test contained a set of 18–23 multiple-choice items preselected from the previous administrations for the same grade as anchor set. These repeated multiple-choice items served as anchors in the Stocking and Lord's (1983) equating procedure, which was used to place each test form on the previously established scale. By equating the 2008 CSAP tests within each grade, the unique metrics of the CSAP Reading, Writing, and Mathematics vertical scales as well as grade level scale for Science and Spanish tests were maintained.

These scaling and calibration methods are presented in Part 6 of this report.

Part 5: Item Analyses

All students who participated in the operational administration were scored. For the item analyses and calibration samples, however, student responses from the following categories were excluded as a part of the valid attempt rules:

- Students absent when any items assessing a scale were administered, with out-of-range scores, and/or multiple marks.

- Students who have invalidation flags.

- Students who have the following special accommodation codes:
  1) For Reading no special accommodation codes were excluded
  2) For Writing scribed responses (special code = 5)
  3) For Mathematics all grades and Science grade10 responses where the entire test was presented orally (special code=9) and students who received translated oral script (special code = B)
  4) For Science grades 5 and 8, responses where the entire test was presented orally (special code = 9)

The descriptive statistics of scale scores were based on all valid cases. The frequency distributions by gender, ethnicity, and other subgroups are shown in Tables 16 through 20.

Tables 21 through 82 display the item analysis results for both multiple-choice (MC) and constructed-response (CR) items for each grade and content area. The product–moment correlation coefficient is used to estimate the item-to-total score correlation for each item. The coefficient for each item is based on the item score and the score computed as the total of all *other* items on the test (hence, the item itself is excluded from the total score). For items having only two levels, the product–moment coefficient is the point-biserial correlation. If an item has to be removed from the calibration and the test because of its aberrant characteristics the point-biserial correlation was recomputed with the item dropped from the calculation.

The *p*-value for each multiple-choice item is the percent of students who gave a correct response to the item. The *p*-value for each constructed-response item is the mean percent of the maximum possible score. Any omitted responses to individual items or constructed-response items with condition codes were treated as incorrect for the calculation of the *p*-values and the item-to-total score correlations. This is consistent with how these omits were treated in the computation of the operational scale scores. The item-to-total score correlations or point biserial (these terms are used interchangeably), the *p*-values, the percentage of omits, and the percentages at each score level (for the constructed

response items) are based on the analysis of responses of students who had reported total test scores only.

As a part of evaluating item analysis results, the percent of students obtaining each score point for the constructed-response items across all grades and content areas was examined. The results indicated that there was a reasonable amount of variability in students' responses to most multiple-choice items and reasonable distribution of score points to most constructed-response items, indicating that these items work well over the range of student ability. The classical item statistics for all grades and content areas are described briefly below.

Third Grade

Reading

Table 21 lists the results of the multiple-choice item analyses for the 2008 third-grade Reading assessment. The point biserials for all multiple-choice items range from 0.24 to 0.59 with a mean of 0.44. The *p*-values for the multiple-choice items range from 0.27 to 0.88 with a mean of 0.70.

Table 22 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items range from 0.51 to 0.70 with a mean of 0.60. The *p*-values range from 0.36 to 0.65 with a mean of 0.51. The percentage of students was well distributed across the score points in all the constructed-response items.

The omit rate for the third-grade Reading assessment was small, ranging from 0.05% to 2.06% for the multiple-choice items (Table 21) and from 0.70% to 2.61% for the constructed-response items (Table 22).

Reading – Spanish

Table 23 lists the results of the multiple-choice item analyses for the Spanish version of the 2008 third-grade Reading assessment. The point biserials for all multiple-choice items range from 0.09 to 0.59 with a mean of 0.37. The *p*-values for the multiple-choice items range from 0.25 to 0.94 with a mean of 0.60.

Table 24 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items range from 0.50 to 0.66 with a mean of 0.58. The *p*-values range from 0.40 to 0.69 with a mean of 0.56. More than 50% of the students obtained the highest possible score points for 2 out of the 8 constructed-response items. The scores of the remaining students were well distributed across the score points in those items.

The omit rate for the Spanish version of the third-grade Reading assessment ranged from 0% to 8.72% for the multiple-choice items, with one item having an omit rate greater than 5% (Table 23). The omit rate for the constructed-response items was small, ranging from 0.47% to 3.29% (Table 24).

Writing

Table 25 lists the results of the multiple-choice item analyses for the 2008 third-grade Writing assessment. The point biserials for all multiple-choice items range from 0.29 to 0.53 with a mean of 0.42. The *p*-values for the multiple-choice items range from 0.60 to 0.97 with a mean of 0.82.

Table 26 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items range from 0.14 to 0.55 with a mean of 0.39. The *p*-values range from 0.13 to 0.98 with a mean of 0.73. More than 50% of the students obtained the highest possible score points for 13 out of the 18 constructed-response items. The percentage of students was well distributed across the score points in all the constructed-response items.

The omit rate for the third-grade Writing assessment was small, ranging from 0.02% to 1.03% for the multiple-choice items (Table 25) and from 0.08% to 0.56% for the constructed-response items (Table 26).

Writing – Spanish

Table 27 lists the results of the multiple-choice item analyses for the Spanish version of the 2008 third-grade Writing assessment. The point biserials for all multiple-choice items range from 0.15 to 0.50 with a mean of 0.41. The *p*-values for the multiple-choice items range from 0.23 to 0.94 with a mean of 0.71.

Table 28 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items range from 0.37 to 0.62 with a mean of 0.49. The *p*-values range from 0.24 to 0.89 with a mean of 0.63. More than 50% of the students obtained the highest possible score points for 12 out of the 18 constructed-response items. The percentage of students was well distributed across the score points of the remaining items.

The omit rate for the Spanish version of the third-grade Writing assessment was small, ranging from 0% to 1.62% for the multiple-choice items (Table 27) and from 0.54% to 0.95% for the constructed-response items (Table 28).

Mathematics

Table 29 lists the results of the multiple-choice item analyses for the 2008 third-grade Mathematics assessment. The point biserials for all multiple-choice items range from 0.22 to 0.56 with a mean of 0.44. The *p*-values for the multiple-choice items range from 0.48 to 0.97 with a mean of 0.75.

Table 30 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items range from 0.50 to 0.69 with a mean of 0.59. The *p*-values range from 0.48 to 0.79 with a mean of 0.62. More than 50% of the students obtained the highest possible score points for 3 out of the 8 constructed-response items. The scores of the remaining students were well distributed across the score points in those items.

The omit rate for the third-grade Mathematics assessment ranged from 0.05% to 1.50% for the multiple-choice items, with one item having an omit rate greater than 5% (Table 29). The omit rate for the constructed-response items was small, ranging from 0.07% to 0.90% (Table 30).

Fourth Grade

Reading

Table 31 lists the results of the multiple-choice item analyses for the 2008 fourth-grade Reading assessment. The point biserials for all multiple-choice items range from 0.16 to 0.58 with a mean of 0.43. The p-values for the multiple-choice items range from 0.34 to 0.89 with a mean of 0.69.

Table 32 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items range from 0.38 to 0.59 with a mean of 0.49. The p-values range from 0.33 to 0.86 with a mean of 0.52. More than 50% of the students obtained the highest possible score points for 3 out of the 14 constructed-response items. The percentage of students was well distributed across the score points of the remaining items.

The omit rate for the fourth-grade Reading assessment was small, ranging from 0.04% to 2.84% for the multiple-choice items (Table 31) and from 0.41% to 1.69% for the constructed-response items (Table 32).

Reading – Spanish

Table 33 lists the results of the multiple-choice item analyses for the Spanish version of the 2008 fourth-grade Reading assessment. The point biserials for all multiple-choice items range from 0.02 to 0.62 with a mean of 0.35. The *p*-values for the multiple-choice items range from 0.28 to 0.93 with a mean of 0.57.

Table 34 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items range from 0.29 to 0.73 with a mean of 0.53. The *p*-values range from 0.25 to 0.69 with a mean of 0.40. More than 50% of the students obtained the highest possible score points for 2 out of the 14 constructed-response items. The scores of the remaining students were well distributed across the score points in that item.

The omit rate for the Spanish version of the fourth-grade Reading assessment was small, ranging from 0% to 4.55% for the multiple-choice items (Table 33). The omit rate for the constructed-response items ranged from 0% to 8.52% with 3 of the 14 items having an omit rate greater than 5% (Table 34).

Writing

Table 35 lists the results of the multiple-choice item analyses for the 2008 fourth-grade Writing assessment. The point biserials for all multiple-choice items range from 0.25 to 0.57 with a mean of 0.42. The *p*-values for the multiple-choice items range from 0.35 to 0.94 with a mean of 0.77.

Table 36 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items range from 0.10 to 0.68 with a mean of 0.48. The *p*-values range from 0.30 to 0.98 with a mean of 0.66. More than 50% of the students obtained the highest possible score points for 6 out of the 13 constructed-response items. The percentage of students was well distributed across the score points of the remaining items.

The omit rate for the fourth-grade Writing assessment was small, ranging from 0.03% to 2.05% for the multiple-choice items (Table 35) and from 0% to 2.59% for the constructed-response items (Table 36).

Writing – Spanish

Table 37 lists the results of the multiple-choice item analyses for the Spanish version of the 2008 fourth-grade Writing assessment. The point biserials for all multiple-choice items range from 0.09 to 0.54 with a mean of 0.32. The *p*-values for the multiple-choice items range from 0.26 to 0.95 with a mean of 0.50.

Table 38 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items range from 0.03 to 0.59 with a mean of 0.40. The *p*-values range from 0.13 to 0.92 with a mean of 0.53. More than 50% of the students obtained the highest possible score points for 5 out of the 13 constructed-response items. The percentage of students was well distributed across the score points of the remaining items.

The omit rate for the Spanish version of the fourth-grade Writing assessment ranged from 0% to 6.32% for the multiple-choice items, with 2 out of the 40 items having an omit rate greater than 5% (Table 37). The omit rate for the constructed-response items was small, ranging from 0% to 6.90%, with one item having an omit rate greater than 5% (Table 38).

Mathematics

Table 39 lists the results of the multiple-choice item analyses for the 2008 fourth-grade Mathematics assessment. The point biserials for all multiple-choice items range from 0.20 to 0.59 with a mean of 0.42. The *p*-values for the multiple-choice items range from 0.47 to 0.97 with a mean of 0.77.

Table 40 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items range from 0.41 to 0.76 with a mean of 0.58. The *p*-values range from 0.33 to 0.85 with a mean of 0.64. More than 50% of the students obtained the highest possible score points for 4 out of the 15 constructed-response items. The scores of the remaining students were well distributed across the score points in those items.

The omit rate for the fourth-grade Mathematics assessment was small, ranging from 0.05% to 1.35%% for the multiple-choice items (Table 39) and from 0.08% to 1.51% for the constructed-response items (Table 40).

Fifth Grade

Reading

Table 41 lists the results of the multiple-choice item analyses for the 2008 fifth-grade Reading assessment. The point biserials for all multiple-choice items range from 0.14 to 0.57 with a mean of 0.41. The *p*-values for the multiple-choice items range from 0.34 to 0.92 with a mean of 0.71.

Table 42 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items range from 0.21 to 0.65 with a mean of 0.48. The *p*-values range from 0.12 to 0.84 with a mean of 0.50. More than 50% of the students obtained the highest possible score points for 5 out of the 14 constructed-response items. The percentage of students was well distributed across the score points of the remaining items.

The omit rate for the fifth-grade Reading assessment ranged from 0.04% to 1.97% for the multiple-choice items (Table 41). The omit rate for the constructed-response items was small, ranging from 0.30% to 2.45% (Table 42).

Writing

Table 43 lists the results of the multiple-choice item analyses for the 2008 fifth-grade Writing assessment. The point biserials for all multiple-choice items range from 0.24 to 0.56 with a mean of 0.42. The *p*-values for the multiple-choice items range from 0.24 to 0.92 with a mean of 0.73.

Table 44 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items range from 0.12 to 0.60 with a mean of 0.46. The *p*-values range from 0.31 to 0.99 with a mean of 0.66. More than 50% of the students obtained the highest possible score points for 6 out of the 13 constructed-response items. The percentage of students was well distributed across the score points of the remaining items.

The omit rate for the fifth-grade Writing assessment was small, ranging from 0.07% to 1.35% for the multiple-choice items (Table 43) and from 0% to 1.86% for the constructed-response items (Table 44).

Mathematics

Table 45 lists the results of the multiple-choice item analyses for the 2008 fifth-grade Mathematics assessment. The point biserials for all multiple-choice items range from 0.24 to 0.60 with a mean of 0.42. The *p*-values for the multiple-choice items range from 0.28 to 0.94 with a mean of 0.71.

Table 46 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items range from 0.46 to 0.72 with a mean of 0.61. The *p*-values range from 0.38 to 0.87 with a mean of 0.64. More than 50% of the students obtained the highest possible score points for 5 out of the 15 constructed-response items. The scores of the remaining students were well distributed across the score points in those items.

The omit rate for the fifth-grade Mathematics assessment was small, ranging from 0.02% to 1.76% for the multiple-choice items (Table 45) and from 0.10% to 4.82% for the constructed-response items (Table 46).

Science

Table 47 lists the results of the multiple-choice item analyses for the 2008 fifth-grade Science assessment. The point biserials for all multiple-choice items range from 0.02 to 0.59 with a mean of 0.36. Item 13, which had the 0.02 point biserial, was removed from the calibration along with item 34. The *p*-values for the multiple-choice items range from 0.34 to 0.98 with a mean of 0.72.

Table 48 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items range from 0.18 to 0.60 with a mean of 0.47. The *p*-values range from 0.06 to 0.87 with a mean of 0.58. More than 50% of the students obtained the highest possible score points for 9 out of the 18 constructed-response items. The percentage of students was well distributed across the score points of the remaining items.

The omit rate for the fifth-grade Science assessment was small, ranging from 0.02% to 0.93% for the multiple-choice items (Table 47) and from 0.07% to 1.74% for the constructed-response items (Table 48).

Sixth Grade

Reading

Table 49 lists the results of the multiple-choice item analyses for the 2008 sixth-grade Reading assessment. The point biserials for all multiple-choice items range from 0.06 to 0.59 with a mean of 0.39. The *p*-values for the multiple-choice items range from 0.33 to 0.94 with a mean of 0.71.

Table 50 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items range from 0.34 to 0.61 with a mean of 0.50. The *p*-values range from 0.23 to 0.87 with a mean of 0.47. More than 50% of the students obtained the highest possible score points for 3 out of the 14 constructed-response items. The scores of the remaining students were well distributed across the score points in that item.

The omit rate for the sixth-grade Reading assessment ranged from 0.05% to 3.18% for the multiple-choice items (Table 49) and from 0.30% to 2.66% for the constructed-response items (Table 50).

Writing

Table 51 lists the results of the multiple-choice item analyses for the 2008 sixth-grade Writing assessment. The point biserials for all multiple-choice items range from 0.25 to 0.58 with a mean of 0.42. The *p*-values for the multiple-choice items range from 0.38 to 0.91 with a mean of 0.71.

Table 52 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items range from 0.12 to 0.63 with a mean of 0.44. The *p*-values range from 0.24 to 0.99 with a mean of 0.71. More than 50% of the students obtained the highest possible score points for 7 out of the 13 constructed-response items. The percentage of students was well distributed across the score points of the remaining items.

The omit rate for the sixth-grade Writing assessment was small, ranging from 0.06% to 1.07% for the multiple-choice items (Table 51) and from 0% to 2.38% for the constructed-response items (Table 52).

Mathematics

Table 53 lists the results of the multiple-choice item analyses for the 2008 sixth-grade Mathematics assessment. The point biserials for all multiple-choice items range from 0.17 to 0.62 with a mean of 0.42. The *p*-values for the multiple-choice items range from 0.36 to 0.97 with a mean of 0.69.

Table 54 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items range from 0.40 to

0.76 with a mean of 0.60. The *p*-values range from 0.34 to 0.87 with a mean of 0.60. More than 50% of the students obtained the highest possible score points for 6 out of the 15 constructed-response items. The scores of the remaining students were well distributed across the score points in those items

The omit rate for the sixth-grade Mathematics assessment was small, ranging from 0.04% to 1.17% for the multiple-choice items (Table 53) and from 0.15% to 1.71% for the constructed-response items (Table 54).

Seventh Grade

Reading

Table 55 lists the results of the multiple-choice item analyses for the 2008 seventh-grade Reading assessment. The point biserials for all multiple-choice items range from –0.01 to 0.52 with a mean of 0.37. Item 24, which has the only negative point biserial, was removed from the calibration and the test. The *p*-values for the multiple-choice items range from 0.28 to 0.95 with a mean of 0.68.

Table 56 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items range from 0.38 to 0.62 with a mean of 0.51. The *p*-values for the constructed-response items range from 0.22 to 0.76 with a mean of 0.50. More than 50% of the students obtained the highest possible score points for 3 out of the 14 constructed-response items. The percentage of students was well distributed across the score points of the remaining items.

The omit rate for the seventh-grade Reading assessment was small, ranging from 0.08% to 3.47% for the multiple-choice items (Table 55) and from 0.66% to 3.12% for the constructed-response items (Table 56).

Writing

Table 57 lists the results of the multiple-choice item analyses for the 2008 seventh-grade Writing assessment. The point biserials for all multiple-choice items range from 0.07 to 0.57 with a mean of 0.42. The *p*-values for the multiple-choice items range from 0.13 to 0.90 with a mean of 0.67.

Table 58 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items range from 0.12 to 0.64 with a mean of 0.43. The *p*-values range from 0.35 to 0.99 with a mean of 0.61. More than 50% of the students obtained the highest possible score points for 3 out of the 13 constructed-response items. The percentage of students was well distributed across the score points of the remaining items.

The omit rate for the seventh-grade Writing assessment was small, ranging from 0.04% to 2.67% for the multiple-choice items (Table 57) and from 0% to 1.83% for the constructed-response items (Table 58).

Mathematics

Table 59 lists the results of the multiple-choice item analyses for the 2008 seventh-grade Mathematics assessment. The point biserials for all multiple-choice items range from 0.11 to 0.54 with a mean of 0.38. The *p*-values for the multiple-choice items range from 0.20 to 0.97 with a mean of 0.61.

Table 60 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items range from 0.37 to 0.75 with a mean of 0.61. The *p*-values range from 0.15 to 0.64 with a mean of 0.42. More than 50% of the students obtained the highest possible score points for 1 out of the 15 constructed-response items. The scores of the remaining students were well distributed across the score points in that item.

The omit rate for the seventh-grade Mathematics assessment was small, ranging from 0.04% to 0.86% for the multiple-choice items (Table 59) and from 0.35% to 3.11% for the constructed-response items (Table 60).

Eighth Grade

Reading

Table 61 lists the results of the multiple-choice item analyses for the 2008 eighth-grade Reading assessment. The point biserials for all multiple-choice items range from 0.06 to 0.56 with a mean of 0.39. The *p*-values for the multiple-choice items range from 0.22 to 0.88 with a mean of 0.67.

Table 62 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items range from 0.47 to 0.68 with a mean of 0.56. The *p*-values range from 0.26 to 0.87 with a mean of 0.53. More than 50% of the students obtained the highest possible score points for 2 out of the 14 constructed-response items. The percentage of students was well distributed across the score points of the remaining items.

The omit rate for the eighth-grade Reading assessment was small, ranging from 0.05% to 3.05% multiple-choice items (Table 61). The omit rate for the constructed-response items ranged from 0.56% to 5.90% with 1 out of the 14 items having an omit rate greater than 5% (Table 62).

Writing

Table 63 lists the results of the multiple-choice item analyses for the 2008 eighth-grade Writing assessment. The point biserials for all multiple-choice items range from 0.27 to 0.54 with a mean of 0.43. The *p*-values for the multiple-choice items range from 0.31 to 0.96 with a mean of 0.68.

Table 64 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items range from 0.12 to 0.59 a mean of 0.43. The *p*-values range from 0.24 to 0.99 with a mean of 0.63. More than 50% of the students obtained the highest possible score points for 6 out of the 13 constructed-response items. The percentage of students was well distributed across the score points of the remaining items.

The omit rate for the eighth-grade Writing assessment ranged from 0.02% to 5.65% for multiple-choice items, with one item having an omit rate greater than 5% (Table 63). The omit rate for the constructed-response items was small, ranging from 0% to 1.36% (Table 64).

Mathematics

Table 65 lists the results of the multiple-choice item analyses for the 2008 eighth-grade Mathematics assessment. The point biserials for all multiple-choice items range from 0.12 to 0.62 with a mean of 0.40. The *p*-values for the multiple-choice items range from 0.13 to 0.82 with a mean of 0.53.

Table 66 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items range from 0.38 to 0.73 with a mean of 0.62. The *p*-values range from 0.23 to 0.73 with a mean of 0.48. More than 50% of the students obtained the highest possible score points for 2 of the 15 constructed-response items. The scores of the remaining students were well distributed across the score points in those items.

The omit rate for the eighth-grade Mathematics assessment was small, ranging from 0.06% to 0.75 for the multiple-choice items (Table 65) and from 0.18% to 3.79% for the constructed-response items (Table 66).

Science

Table 67 lists the results of the multiple-choice item analyses for the 2008 eighth-grade Science assessment. The point biserials for all multiple-choice items range from 0.05 to 0.50 with a mean of 0.33. The *p*-values for the multiple-choice items range from 0.18 to 0.77 with a mean of 0.54.

Table 68 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items range from 0.25 to 0.68 with a mean of 0.48. The *p*-values range from 0.11 to 0.83 with a mean of

0.42. More than 50% of the students obtained the highest possible score points for 4 out of the 23 constructed-response items. The scores of the remaining students were well distributed across the score points in those items.

The omit rate for the eighth-grade Science assessment was small, ranging from 0.02% to 2.00% for the multiple-choice items (Table 67). The omit rate for the constructed-response items ranged from 0.60% to 7.57% with 2 of the 23 items having an omit rate greater than 5% (Table 68).

Ninth Grade

Reading

Table 69 lists the results of the multiple-choice item analyses for the 2008 ninth-grade Reading assessment. The point biserials for all multiple-choice items range from −0.20 to 0.63 with a mean of 0.41. Item 32, which has the only negative point biserial, was removed from the calibration and the test. The *p*-values for the multiple-choice items range from 0.17 to 0.90 with a mean of 0.66.

Table 70 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items range from 0.35 to 0.65 with a mean of 0.54. The *p*-values range from 0.33 to 0.91 with a mean of 0.54. More than 50% of the students obtained the highest possible score points for 1 out of the 14 constructed-response items. The percentage of students was well distributed across the score points of the remaining items.

The omit rate for the ninth-grade Reading assessment was small, ranging from 0.07% to 0.78% for the multiple-choice items (Table 69). The omit rate for the constructed-response items ranged from 1.01% to 6.16% with 1 out of the 14 items having an omit rate greater than 5% (Table 70).

Writing

Table 71 lists the results of the multiple-choice item analyses for the 2008 ninth-grade Writing assessment. The point biserials for all multiple-choice items range from 0.08 to 0.57 with a mean of 0.43. The *p*-values for the multiple-choice items range from 011 to 0.88 with a mean of 0.66.

Table 72 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items range from 0.13 to 0.64 with a mean of 0.45. The *p*-values range from 0.37 to 0.98 with a mean of 0.71. More than 50% of the students obtained the highest possible score points for 6 out of the 13 constructed-response items. The percentage of students was well distributed across the score points of the remaining items.

The omit rate for the ninth-grade Writing assessment was small, ranging from 0.04% to 5.41% for the multiple-choice items with one item having an omit rate greater than 5% (Table 71). The omit rate for the constructed-response items ranged from 0% to 2.40% (Table 72).

Mathematics

Table 73 lists the results of the multiple-choice item analyses for the 2008 ninth-grade Mathematics assessment. The point biserials for all multiple-choice items range from 0.06 to 0.55 with a mean of 0.36. Item 29, which had the 0.06 point biserial, was removed from the calibration. The *p*-values for the multiple-choice items range from 0.10 to 0.91 with a mean of 0.54.

Table 74 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items range from 0.56 to 0.73 with a mean of 0.65. The *p*-values range from 0.13 to 0.52 with a mean of 0.34. The percentage of students was well distributed across the score points in all the constructed-response items.

The omit rate for the ninth-grade Mathematics assessment was small, ranging from 0.05% to 0.96% for the multiple-choice items (Table 73). The omit rate for the constructed-response items ranged from 1.12% to 5.53% with 1 out of the 15 items having an omit rate greater than 5% (Table 74).

Tenth Grade

Reading

Table 75 lists the results of the multiple-choice item analyses for the 2008 tenth-grade Reading assessment. The point biserials for all multiple-choice items range from −0.15 to 0.58 with a mean of 0.35. Item 45, which had the only negative point biserial, was removed from the calibration and the test meaning that the item did not contribute to the students' total scores. The *p*-values for the multiple-choice items range from 0.24 to 0.91 with a mean of 0.65.

Table 76 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items range from 0.35 to 0.67 with a mean of 0.58. The *p*-values range from 0.27 to 0.77 with a mean of 0.50. More than 50% of the students obtained the highest possible score points for 1 out of the 14 constructed-response items. The scores of the remaining students were well distributed across the score points in that item.

The omit rate for the tenth-grade Reading assessment was small, ranging from 0.10% to 2.01% for the multiple-choice items (Table 75). The omit rate for the constructed-response items ranged from 3.18% to 11.90% with 6 out of the 14 items having an omit rate greater than 5% (Table 76).

Writing

Table 77 lists the results of the multiple-choice item analyses for the 2008 tenth-grade Writing assessment. The point biserials for all multiple-choice items range from 0.21 to 0.56 with a mean of 0.42. The *p*-values for the multiple-choice items range from 0.26 to 0.88 with a mean of 0.67.

Table 78 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items range from 0.17 to 0.66 with a mean of 0.48. The *p*-values range from 0.37 to 0.98 with a mean of 0.65. More than 50% of the students obtained the highest possible score points for 6 out of the 13 constructed-response items. The scores of the remaining students were well distributed across the score points in those items.

The omit rate for the tenth-grade Writing assessment was small, ranging from 0.06% to 6.84% for the multiple-choice items, with one item having an omit rate greater than 5% (Table 77). The omit rate for the constructed-response items ranged from 0% to 3.45% (Table 78).

Mathematics

Table 79 lists the results of the multiple-choice item analyses for the 2008 tenth-grade Mathematics assessment. The point biserials for all multiple-choice items range from 0.17 to 0.53 with a mean of 0.35. The *p*-values for the multiple-choice items range from 0.22 to 0.88 with a mean of 0.51.

Table 80 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items range from 0.18 to 0.75 with a mean of 0.60. The *p*-values for the constructed-response items range from 0.03 to 0.72 with a mean of 0.34. More than 50% of the students obtained the highest possible score points for 1 out of the 15 constructed-response items. The scores of the remaining students were well distributed across the score points in those items.

The omit rate for the tenth-grade Mathematics assessment was small, ranging from 0.04% to 0.71% for the multiple-choice items (Table 79) and from 0.79% to 2.84% for the constructed-response items (Table 80).

Science

Table 81 lists the results of the multiple-choice item analyses for the 2008 tenth-grade Science assessment. The point biserials for all multiple-choice items range from 0.05 to 0.51 with a mean of 0.35. The *p*-values for the multiple-choice items range from 0.11 to 0.87 with a mean of 0.56.

Table 82 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items range from 0.25 to

0.59 with a mean of 0.46. The *p*-values for the constructed-response items range from 0.07 to 0.72 with a mean of 0.43. More than 50% of the students obtained the highest possible score points for 7 out of the 23 constructed-response items. The percentage of students was well distributed across the score points of the remaining items.

The omit rate for the tenth-grade Science assessment was small, ranging from 0.04% to 1.80% for the multiple-choice items (Table 81). The omit rate for the constructed-response items ranged from 1.07% to 12.60% with 5 out of the 23 items having an omit rate greater than or equal to 5% (Table 82).

Part 6: Calibration and Equating

Part 6 describes item response theory (IRT) models used for calibration and equating, fit criterion for model-to-data fit, and items flagged for poor model fit for all grades and content areas. It also briefly presents the number of item pairs correlated within a grade and content area measured by Yen's Q3 statistic (Yen, 1984), followed by equating design and methods for evaluating anchor items. The test characteristic curves for the total test and anchor set are presented as evidence that the anchor set was representative of the total test and linking was reasonable. Finally, the scaling constants resulting from the linking are presented.

Overview of the IRT Models

CTB uses IRT to place multiple-choice and constructed-response items on the same scale. Because the characteristics of selected-response (multiple-choice) and constructed-response (open-ended) items are different, two-item response theory models are used in the analysis of test forms containing both item types. The three-parameter logistic (3PL) model (Lord, 1980; Lord & Novick, 1968) is used for the analysis of selected-response items. In this model, the probability that a student with scale score θ responds correctly to item *i* is

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + \exp[-1.7a_i(\theta - b_i)]},$$

where a_i is the item discrimination, b_i is the item difficulty, and c_i is the probability of a correct response by a very low scoring student. These three parameters are estimated from the item response data.

For analysis of constructed-response items, the two-parameter partial credit model (2PPC) (Muraki, 1992; Yen, 1993) is used. The 2PPC model is a special case of Bock's (1972) nominal model. Bock's model states that the probability of an examinee with ability θ having a score at the *k*th level of the *j*th item is

$$P_{jk}(\theta) = P(x_j = k-1 \mid \theta) = \frac{\exp Z_{jk}}{\sum_{i=1}^{m_j} \exp Z_{ji}}, \quad k = 1, \text{ K }, m_j,$$

where m_j is the number of score levels and

$$Z_{jk} = A_{jk}\theta + C_{jk}.$$

For the special case of the 2PPC model used here, the following constraints are used:

$$A_{jk} = \alpha_j (k-1), \quad k = 1,\ 2,\ \mathrm{K}\ ,\ m_j,$$

and

$$C_{jk} = -\sum_{i=0}^{k-1} \gamma_{ji}, \quad \text{where } \gamma_{j0} = 0,$$

where α_j and γ_{ji} are the parameters to be estimated from the data. The first constraint implies that higher item scores reflect higher ability levels and that the items can vary in their discriminations. For the 2PPC model, for each item there are $m_j - 1$ independent γ_{ji} parameters and one α_j parameter; a total of m_j independent item parameters are estimated.

The IRT models are implemented using CTB's PARDUX computer program (Burket, 1993). PARDUX estimates parameters simultaneously for dichotomous (multiple-choice) and polytomous (constructed-response) items using marginal maximum likelihood procedures implemented via the EM algorithm (Bock & Aitkin, 1981; Thissen, 1982).

Calibration of the Assessment

The items within a grade in each content area were calibrated using CTB's computer program PARDUX (Burket, 1993), and all items were evaluated for model fit and local independence. The calibration sample ranged from 89.4% to 100% of the total tested population for all grades and content areas.

The parameters estimated by PARDUX are in two different parameterizations, corresponding to the two item response theory models (3PL and 2PPC). The location (i.e., difficulty) and discrimination (characteristics of an item to differentiate students with different abilities) parameters for the multiple-choice items are in the traditional 3PL metric and are designated as *b* and *a*, respectively. The location and discrimination parameters for the constructed-response items are in the 2PPC metric, designated *g* (gamma) and *f* (alpha), respectively. Because of the different metrics used, the 3PL (multiple-choice) parameters (*a* and *b*) are not directly comparable to the 2PPC (constructed-response) parameters (*f* and *g*). However, they can be converted to a common metric. The two metrics are related by $b = g/f$ and $a = f/1.7$ (see Burket, 1993). As a result of this procedure, the multiple-choice and constructed-response items are placed on the same scale. Note that for the 2PPC model there are $m_j - 1$ (where m_j is the number of score levels for item *j*) independent *g*'s and one *f*, for a total of m_j independent parameters estimated for each item. For the 3PL

model, there is one "*a*" parameter, one "*b*" parameter, and one pseudo-guessing parameter, "*c*," for each item.

Model Fit Analyses

During the calibration process, each item is reviewed for how well the item parameters in the model fit the observed data. Item fit was assessed using the Q_1 statistic described by Yen (1981) for the dichotomously (multiple-choice) scored items and using a generalization of this statistic for the multilevel (constructed-response) items. As described by Yen, Q_1 is a Pearson chi-square of the form in each cell

$$Q_{1j} = \sum_{i=1}^{I} \frac{N_{ji}\left(O_{ji} - E_{ji}\right)^2}{E_{ji}} + \sum_{i=1}^{I} \frac{N_{ji}\left[\left(1-O_{ji}\right) - \left(1-E_{ji}\right)\right]^2}{1-E_{ji}},$$

where N_{ji} is the number of examinees in cell i for item j. O_{ji} and E_{ji} are the observed and predicted proportions of examinees in cell i that attain the maximum possible score on item j, where

$$E_{ji} = \frac{1}{N_{ji}} \sum_{a \in cell\, i}^{N_{ji}} P_j(\hat{\theta}_a).$$

The generalization of Q_1 for multilevel (constructed-response) items in each cell can be stated as

$$Q_{1j} = \sum_{i=1}^{I} \sum_{k=1}^{m_j} \frac{N_{ji}\left(O_{jki} - E_{jki}\right)^2}{E_{jki}},$$

where

$$E_{jki} = \frac{1}{N_{ji}} \sum_{a \in cell\, i}^{N_{ji}} P_{jk}(\hat{\theta}_a).$$

O_{jki} and E_{jki} are the observed and expected proportion of examinees in cell i who performed at the *kth* score level.

Chi-squared statistics are affected by sample size and extreme expectations (Stone, Ankenmann, Lane, & Liu, 1993), and their degrees of freedom are a function of the number of independent observations entering into the calculation minus the number of parameters estimated. Items with more score levels have more degrees of freedom, making it awkward to compare fit for items that differ in

the number of score levels. To facilitate this comparison, the following standardization of the Q_1 statistic was used:

$$Z_{Q_1 j} \equiv \frac{Q_{1j} - df}{\sqrt{(2df)}}.$$

The value of Z still will increase with sample size, all else being equal. To use this standardized statistic to flag items for potential misfit, it has been CTB's practice to vary the critical value for Z as a function of sample size. When piloting multiple-choice items for new tests, CTB typically has used the flagging criterion $Z \geq 4.00$ with sample sizes of approximately 1,000 students. For the operational tests, which have larger calibration sample sizes, the criterion Z_c used to flag items was calculated using the expression

$$Z_c = \left(\frac{\text{Calibration Sample Size}}{1{,}500} \right) * 4.00.$$

This criterion was used to flag operational CSAP items for potential misfit. Item characteristic curves (ICCs) of all flagged items were visually inspected in order to decide whether their high Z's resulted from poor model-data fit or from irrelevant variables such as extreme expectations that often accompany unusually easy or hard items. Only those items judged to be poorly fit by the model were defined as misfitting items.

Model Fit Analyses Results

The model fit statistics and item parameter results are based on the analysis of a sample data set used for item calibration and scaling. The summary fit statistics for the multiple-choice and constructed-response items for different grades and content areas are shown in Tables 83 through 144.

Detailed summaries of the model fit results are presented below.

Third Grade

The third-grade item parameters and fit statistics are shown in Tables 83 through 92. The critical *Z*-values for these tables are 152.5 for Reading, 4.0 for Spanish Reading, 150.9 for Writing, 3.9 for Spanish Writing, and 139.0 for Mathematics.

Across all content areas, seven items exceeded these critical *Z*-values and exhibited less than optimal fit: two Reading item (CR items 6 and 26), four Spanish Reading items (MC items 24 and CR items 1, 32, and 36), and one Spanish Writing item (MC item 6).

Fourth Grade

The fourth-grade item parameters and fit statistics are shown in Tables 93 through 102. The critical *Z*-values for these tables are 153.4 for Reading, 150.6 for Writing, and 141.4 for Mathematics. Spanish Reading had a critical *Z*-value of 1.39 for items that originated from the 2004 administration, 1.30 for items that originated in the 2005 administration, and 0.70 for items that originated in the 2007 administration. Spanish Writing had a critical *Z*-value of 1.40 for items that originated from the 2004 administration and 1.31 for items that originated in the 2005 administration. Spanish Writing grade 4 has a critical *Z*-value of 2.67 for constructed-response items that originated in 2002, 1.31 for items that originated in the 2005 administration, and 0.70 for items that originated in the 2007 administration.

Across all English content areas, three items exceeded these critical *Z*-values and exhibited less than optimal fit: one Reading item (CR item 30), one Writing item (CR item 3A) and one Mathematics item (CR item 65).

Fifth Grade

The fifth-grade item parameters and fit statistics are shown in Tables 103 through 110. The critical *Z*-values for these tables are 148.8 for Reading, 146.5 for Writing, 139.0 for Mathematics, and 133.7 for Science.

Across all content areas, six items exceeded these critical *Z*-values and exhibited less than optimal fit: one Reading item (CR item 12), one Writing item (CR item 3A), three Mathematics items (CR items 27, 32 and 66), and one Science item (CR item 53).

Sixth Grade

The sixth-grade item parameters and fit statistics are shown in Tables 111 through 116. The critical *Z*-values for these tables are 149.2 for Reading, 147.9 for Writing, and 142.2 for Mathematics.

Across all content areas, four items exceeded these critical *Z*-values and exhibited less than optimal fit: one Writing item (CR item 27) and three Mathematics items (MC item 9 and CR items 29 and 46).

Seventh Grade

The seventh-grade item parameters and fit statistics are shown in Tables 117 through 122. The critical *Z*-values for these tables are 142.0 for Reading, 141.0 for Writing, and 144.2 for Mathematics.

Across all content areas, seven items exceeded these critical *Z*-values and exhibited less than optimal fit: one Reading item (CR item 6), two Writing items

(MC item 75 and CR item 2F) and four Mathematics items (MC item 18 and CR items 20, 35, and 40).

Eighth Grade

The eighth-grade item parameters and fit statistics are shown in Tables 123 through 130. The critical *Z*-values for these tables are 150.3 for Reading, 149.5 for Writing, 144.7 for Mathematics, and 131.7 for Science.

Across all content areas, 10 items exceeded these critical *Z*-values and exhibited less than optimal fit: two Reading items (MC item 25 and CR item 29), one Writing item (CR item 88), five Mathematics items (CR items 3, 34, 40, 46, and 60), and two Science items (MC item 37 and CR item 56).

Ninth Grade

The ninth-grade item parameters and fit statistics are shown in Tables 131 through 136. The critical *Z*-values for these tables are 154.3 for Reading, 153.9 for Writing, and 151.7 for Mathematics.

Across all content areas, eight items exceeded these critical *Z*-values and exhibited less than optimal fit: four Reading items (MC items 31, 36,100, and 107), one Writing item (CR item 26), and three Mathematics items (MC item 57 and CR items 20 and 46).

Tenth Grade

The tenth-grade item parameters and fit statistics are shown in Tables 137 through 144. The critical *Z*-values for these tables are 147.4 for Reading, 147.0 for Writing, 142.1 for Mathematics, and 136.2 for Science.

Across all content areas, 11 items exceeded these critical *Z*-values and exhibited less than optimal fit: six Reading items (MC items 11, 15, 44, 100,101, and 105), one Writing item (CR item 96) and three Mathematics items (MC items 12 and 13 and CR item 7).

Item Local Independence

In using IRT models, one of the assumptions made is that the items are locally independent. That is, a student's response to one item is not dependent on the response to another item. Statistically speaking, when a student's ability is accounted for, the response to each item is statistically independent.

One way to measure the statistical local independence of items within a test is via the Q3 statistic (Yen, 1984). This statistic was obtained by correlating differences between students' observed and expected responses for pairs of

items after taking into account overall test performance. If a substantial number of items in the test demonstrate local dependence, these items may need to be calibrated separately. Pairs of items with Q3 values greater than 0.30 were classified as locally dependent. The maximum value for this index is 1.00. The number of item pairs flagged under the criterion was quite small and varied across forms and content areas. For English Reading, Science, and Spanish Reading, no item pairs were flagged. For Mathematics, there were six item pairs flagged across all grades and content areas (grade 3 items 8 and 23; grade 3 items 21 and 4; grade 5 items 37 and 15; grade 6 items 22 and 2; grade 6 items 28 and 23; and grade 6 items 49 and 2). In contrast, 18 pairs were flagged for the Writing tests, with one to three pairs at each grade level (grade 3 items 2 and 38; grade 3 items 2 and 50; grade 4 items 2 and 8; grade 5 items 2 and 8; grade 6 items 2 and 8; grade 6 items 2 and 40; grade 6 items 16 and 38; grade 7 items 2 and 8; grade 7 items 22 and 11; grade 7 items 22 and 32; grade 8 items 2 and 8; grade 8 items 2 and 28; grade 8 items 16 and 38; grade 9 items 2 and 8; grade 9 items 22 and 32; grade 10 items 2 and 8; grade 10 items 23 and 30; and grade 10 items 24 and 30.) Overall in the English assessments, these 24 pairs exhibited dependency across all possible item pair combinations for which Q3 ranged from 0.30 to 0.92. When compared to grades 3 and 4 English Writing items, a relatively larger number of items in the Spanish tests (12 pairs in all Spanish assessments - grade 3 items 2 and 21; grade 3 items 2 and 37; grade 3 items 2 and 50; grade 3 items 2 and 52; grade 3 items 3 and 28; grade 3 items 3 and 35; grade 3 items 3 and 50; grade 3 items 4 and 21; grade 3 items 4 and 35; grade 3 items 4 and 50; grade 4 items 2 and 8; and grade 4 items 2 and 9) were flagged, but for lower Q3 values ranged from 0.33 to 0.56.

Evaluation of Item Analysis and Calibration

After the evaluation of item analysis and calibration outputs across all grades and content areas, 12 new items (10 multiple-choice and 2 constructed-response) exhibited aberrant characteristics (mainly non-convergence where the item parameters could not be estimated, negative point biserials for the correct choice, and positive point biserials for distractor(s)). After consulting with CTB content experts and CDE, the following items were removed from the final calibration:

- Reading, grade 7 – Item 24
- Reading, grade 9 – Item 32
- Reading, grade 10 – Item 45
- Reading, grade 10 – Item 112
- Reading, grade 10 – Item 114
- Mathematics, grade 3 – Item 24
- Mathematics, grade 4 – Item 27
- Mathematics, grade 9 – Item 29
- Science, grade 5 – Item 13
- Science, grade 5 – Item 34

- Science, grade 10 – Item 15
- Science, grade 10 – Item 70

Tables 2 through 6 indicate the number of items and score points for each test form with suppressed items removed. Writing item 3, part C, across all grades, with a maximum of two score points, measures a writing trait (writing using conventions). This item in most grades was flagged for non-convergence. A further investigation showed that most students (98% or higher) at each grade received the maximum score on this item type. Item parameters for these items were reestimated first by providing different Bayesian priors during the calibration. However, some items still did not converge and were provided larger item difficulty range to reestimate the item parameters for the items. In order to avoid the non-convergence issue in the future administrations on this item, one of the options could be expanding the maximum score point higher than the current one.

Equating Procedures

Through a common item equating design, the calibrated/scaled item parameters for each test were placed onto a vertical (cross-grade) or grade-specific scale. A set of previously selected common or anchor multiple-choice items that had been used in previous operational tests were among the items administered in each grade and content area. Some of the anchor items, especially in Mathematics, were considered minor revised under Universal Design Plain Language principles in order to align format with other items in the tests across administrations. (See page 13 for more information on minor revisions under the Universal Design Plan).Three statistical methods were in place to evaluate the differential performance of these anchor items. The methods are described in the next section. These items were given in approximately the same location or same third of the original administration location. The items were operational in previous administrations and maintained original starting parameter values. These multiple-choice items were used as anchors in the Spring 2008 CSAP to link the tests across years. The anchor parameters were not fixed during calibration, and were used during the equating procedures defined by Stocking and Lord (1983). The anchor parameters were used to place the parameters estimated for all the Spring 2008 CSAP items on the original scales.

As mentioned previously, equating is a statistical procedure that allows adjusting scores on test forms so that the scores are comparable. The Stocking and Lord (1983) procedure, also called the test characteristic curve (TCC) method, was used to place each grade on the vertical scale that had been developed previously for each content area. It minimizes the mean squared difference between the two characteristics curves, one based on estimates from the previous calibration and the other on transformed estimates from the current calibration. Let $\hat{\psi}_j$ be a true score for an examinee, *j*, with ability θ_j based on

item parameter estimates (a_j, b_j, c_j) from the previous calibration and $\hat{\psi}_j^*$ be the estimated true score obtained after the reestimation of item parameters using current data and transformed to the previous scale.

$$\hat{\psi}_j = \hat{\psi}(\theta_j) = \sum_{i=1}^{n} P_i(\theta_j; a_i, b_i, c_i)$$

$$\hat{\psi}_j^* = \hat{\psi}(\theta_j) = \sum_{i=1}^{n} P_i(\theta_j; \frac{a_i}{M_1}, M_1 b_i + M_2, c_i)$$

The TCC method determines the scaling constants (multiplicative – M1 and additive – M2) by minimizing the following quadratic loss function (*F*).

$$F = \frac{1}{N} \sum_{a=1}^{N} (\hat{\psi}_j - \hat{\psi}_j^*)^2$$

where N is the number of examinees in the arbitrary group.

Anchor Items Evaluation Criteria

The multiple-choice anchor items were carefully reviewed to ensure they were performing very similarly in both current and reference years. Three statistical methods—test characteristic curve or TCC method (Stocking & Lord, 1983), Delta Plot method (Angoff, 1972; Dorans & Holland, 1993), and Chi-Square method (Lord, 1980)—were applied to evaluate the anchor items. A description of the TCC method can be seen in the previous section (Equating Procedures). The Delta Plot and Lord's Chi-Square methods are described briefly below.

The Delta Plot method relies only on the differences in the probability of responding to the item correctly (*p*-value). For example, *p*-values of the anchor items based on the previous and current year's population will be calculated. The *p*-values then will be converted to standard normal distribution, *Z*-scores, that correspond to the (1– *p*)th percentiles. For example, for a *p*-value of 0.90, the corresponding *Z*-score will be the (1 – 0.90)th percentile, which is –1.2816. A simple rule to identify outlier items that are functioning differentially between the two groups with respect to the level of difficulty is to draw perpendicular distance to the line of best fit. The fitted line is chosen so as to minimize the sum of squared perpendicular distances of the points to the line. The perpendicular distance is given by

$$D = \frac{AZ_{old} - Z_{new} + B}{\sqrt{A^2 + 1}},$$

where

$$A = \frac{(SD_{Znew}^2 - SD_{Zold}^2) + \sqrt{(SD_{Znew}^2 - SD_{Zold}^2)^2 + 4r_{(Zold)(Znew)}^2 SD_{Zold}^2 SD_{Znew}^2}}{2r_{(Znew)(Zold)} SD_{Zold} SD_{Znew}}$$

and

$$B = \text{Mean}(Z_{new}) - A*\text{Mean}(Z_{old})$$

The standard deviation (SD) of the perpendicular distance is given by

$$SD_D = [(SD_{Znew} + SD_{Zold})/2] * \sqrt{1 - r_{(Zold)(Znew)}} \ .$$

As a rule of thumb, any items lying more than three standard deviations of the distances away from the fitted line are flagged as outliers.

The Lord's Chi-Square criterion involves significance testing of both item difficulty and discrimination parameters simultaneously for each item and evaluating the result based on the chi-square distribution table (see Divgi, 1985, and Lord, 1980, for detail). If the null hypotheses that the item difficulty and discrimination parameters are equal are true, the χ^2 follows chi-square distribution with 2 degrees of freedom.

The following verifications were performed to ensure the quality and accuracy of the equating:

1) The IRT item parameters (a, b, and c), and p-values between reference and current anchor sets were plotted for preliminary screening.
2) *The p*-values of the anchor items were compared to make sure that the anchor items were similar in difficulty in both new and reference administrations. A regression line was drawn for the p-values between the estimated new form and the reference form. If the samples are similar in ability, this regression line will be the identity line. The Delta Plot method (Angoff, 1972; Dorans & Holland, 1993) was used to evaluate the significant *p*-values differences.
3) The RT item parameters for each anchor item were compared. Lord's Chi-Square (Lord, 1980) method was used for flagging items with significantly differential item characteristic curves.
4) The reference and equated anchor item set TCCs were compared to make sure that they were closely overlapping. Similarly, the correlation coefficients between the reference and equated item parameters were compared.
5) The linear transformation parameters (also known as scaling constants) were compared to make sure that they were fairly stable across administrations.

Additional analyses of the equating include the following:

6) The *p*-values of the common anchor items between the two administrations were compared to show that the p-values were in the same direction and magnitude of change as do the scale scores.
7) The full distribution of scale scores was compared for reasonableness across administrations and made sure that the reflected any differences could be the difference in ability that were indicated by the anchor items.
8) The pass rates were compared for reasonableness across administrations, given any noted ability changes.

These routine CTB Research's quality check steps were followed during equating for all grades and content areas.

Anchor Items Evaluation Results

Although a few items were flagged using both Delta Plot and Lord's Chi-Square methods in grade 5 English Reading and Writing and grades 4 and 10 Mathematics, the criteria for removing an anchor item from the anchor set were as follows: If an anchor item was flagged by both Delta Plot and Lord's Chi-Square methods *and* had a *p*-value difference of greater than 0.1, it would be dropped from the anchor set. In all of the 31 grade/content areas in the 2008 CSAP administration, only one item met these criteria (grade 10 Mathematics, item 52) and was removed from the anchor set. The *p*-value and item parameters comparison results are presented below.

Figure 1 shows the item characteristic curve for the anchor item removed from the equating of the 2008 CSAP operational tests (grade 10 Mathematics, item 52).

p-value Comparisons

The differential anchor item functioning between the two administrations in terms of *p*-values indicated that they were aligned closely, with correlations at or higher than 0.98 for all grades and content areas (Table 145). This indicates that the estimated *p*-values for the reference and estimated new form item parameters are very similar, suggesting that the anchor items performed similarly in the two populations (2006 and 2008).

Item Parameter Comparisons

The differential anchor item functioning between the two administrations was evaluated by comparing the correlations between the reference and estimated new form items for difficulty (*b*) and discrimination (*a*) values as well as their plots. Guessing (*c*) parameters are the most fluctuating and were not considered in the evaluation criteria.

Results indicate that the correlations for the discrimination (*a*) and difficulty (*b*) parameters are high, ranging from 0.81 to 0.99 for "*a*" and from 0.87 to 0.99 for "*b*" (see Table 145). These high correlations indicate that the items were performing essentially similarly between the two administrations. This is further evidence that the equating results are reasonable and accurate.

Similarly, the differential item functioning in terms of item characteristic curves between the two administrations for the anchor items was also evaluated using Lord's Chi-Square method. One item each in grades 4, 8, 9, and 10 Mathematics; grades 5, 6, 7, and 10 Reading; and grades 3, 5, and 6 Writing was flagged for performing significantly differentially However, only one anchor item from grade 10 Mathematics was dropped from the anchor set because the item was flagged by both Lord's Chi-Square and DeltaPlot methods and the p-value difference was greater than 0.10.

Scaling Constants

The scaling constants, or the linear transformation parameters, were examined to determine whether the ability differences in terms of consistent trends in ability over time were similar across years. Since the calibration "centers" the raw IRT scale close to the average ability of the test takers, differences in these scaling constants would indicate differences in the ability from reference to new form administrations. The scaling constants for the CSAP grades and content areas are displayed in Table 146 for the two administrations (2007 and 2008). Table 146 indicates that the scaling constants are fairly similar across the two administrations.

Additional Analyses of Flagged Items

Review of the content balance for the final anchor sets in grade 10 Mathematics after removing one flagged item (item 52) indicated that these anchors were reasonably representative of the blueprint for the total tests. Tables 147 through 151 indicate number and percentage of items by content standard for total test and anchor set.

Effectiveness of the Equating

Figures 2 through 27 show the TCC and SEM plots for the Spring 2008 operational tests grades 3 through 10 Reading (Figures 2 through 9), Writing (Figures 10 through 17), and Mathematics (Figures 18 through 25), and grade 3 Spanish Reading and Writing (Figures 29 and 31), compared to the previous year's plots based on census data. Each figure included in this section displays four comparison curves: (a) test characteristic curves, (b) standard errors of measurement, (c) test information curves, and (d) cumulative frequency distributions. These plots illustrate the effectiveness of the equating. The plots

of the TCCs (the S-shaped curves) and the SEM curves (the U-shaped curves) indicate that 2007 and 2008 for a given subject area and grade strongly resembled each other (in that they lay close to or even on top of another) in terms of difficulty, discrimination, and accuracy. Note that because of a new scale being established in 2008 for Science and the limited sample size for grade 4 Spanish Reading and Writing requiring preequated item parameters to be used, only 2008 plots are included for grades 5, 8, and 10 Science (Figures 26 through 28) and grade 4 Spanish Reading and Writing (Figures 30 and 32).

Once the tests were equated, the final parameters were used for deriving each student's scale score. CSAP uses item pattern scoring for all tests. During pattern scoring, the pattern of student responses and the attributes of each item contribute to the student's final scale score. This enhances the comparability of scores across years. For example, two students who respond correctly to a total of 20 questions obtain the same scale score in number-correct scoring. Depending upon the difficulty and discrimination of the items the students answered correctly they may receive different scale scores in item-pattern scoring. The item-pattern scoring is able to take those responses and item attributes into account and provide a scale score that better represents the students' abilities.

Part 7: Scale Score Summary Statistics

Student results are reported statewide in terms of scale scores and performance levels. All valid cases were used for the computation. The scale score ranges (LOSS and HOSS) for each grade and content area are listed in Table 8.

The performance level cut scores were adopted by the Colorado State Board of Education on the basis of the recommendations of standard setting committees composed of qualified Colorado educators, using a variation of the Bookmark standard setting procedure (Lewis, Mitzel, & Green, 1996). As mentioned previously, the performance standards for Reading were adopted from the 2001 standard setting. The performance standards for Writing and Mathematics were adopted from 2002 standard setting except for grades 3 and 4 Mathematics. The grades 3 and 4 Mathematics assessments were introduced in 2005 and standards were set in the same year. Similarly, performance standards for grades 5, 8, and 10 Science were reviewed and set in 2008.

Summary statistics are based on the total Colorado student population tested by CSAP. Table 152 presents the mean, median, and standard deviation of the scale scores for the total population and each gender in each grade/content area. Note that the male and female students do not equal the total population because some students did not identify their gender.

On average, female students scored higher than male students at all grade levels on the Reading and Writing tests, while male students scored slightly higher than female students at all grade levels on the Science assessments. Although the mean difference was less than five points, male students scored slightly higher than female students on the Mathematics tests in grades 3–5, 8, and 9, and both male and female students had equivalent scores at grade 10.

Tables 153 and 154 contain scale score descriptive statistics for each content standard and subcontent area, respectively. Since the scale scores for content standards and subcontent areas are computed on the basis of fewer items, students more easily get the highest obtainable score or the lowest obtainable score on these than on the total test, causing the scale score distributions to be skewed in some cases. For that reason, both means and medians are reported. Tables 155 and 156 contain number-correct descriptive statistics for the total population and the mean percent of the maximum points obtained for each content standard and subcontent area, respectively. The mean percent of the maximum points reflects the relative difficulty of the test. One can compare the relative difficulty of the test for the current administration and previous administrations by comparing these values.

Note the following particulars for reporting purposes: grade 3 Reading measures only one content standard; content standards 2 and 3 are combined for grade 3

Mathematics; content standards 1 and 6 are combined in grades 7 through 10 Mathematics; content standards 4 and 5 are combined in grades 3 through 10 Mathematics; and content standards 1 and 6, 2 and 5, 3 and 5, and 4 and 5 are combined for grades 5, 8, and 10 Science. Similarly, subcontent areas 1 and 4 are combined for grades 3 through 6 Reading. In Tables 153 through 156, where a content standard or subcontent area is shared (e.g., CS 2/3 for grade 3 Mathematics) the scores are reported under the first content standard or subcontent area (e.g., CS 2 for grade 3 Mathematics).

Scale Score Distributions: Student Results

Third Grade

Reading

The mean and median scale scores for the total population of students taking the 2008 third-grade Reading assessment are 554 and 563, respectively, with a standard deviation of 83.2. The mean scale score for female students is 561 with a standard deviation of 77.8, and the mean scale score for male students is 547 with a standard deviation of 87.6.

The scale score frequency distribution of the third-grade Reading assessment for the total population is shown in Table 157. Figure 33 graphically represents the scale score frequency distributions for the total population and for the groups of male and female students separately. The figure shows that the distributions of scale scores for the total population and for each gender are slightly negatively skewed.

The mean scale score for the single content standard is 554 and a median of 563. The mean scale scores for the subcontent areas range from 545 to 572 and the median scale scores range between 563 and 564 (Table 154).

The mean proportions of the maximum obtainable raw score for the subcontent areas range from 55.6 to 72.4. The mean proportion of the maximum obtainable raw score for the total test is 62.0.

Reading – Spanish

The mean scale score for the total population of students taking the 2008 third-grade Spanish Reading assessment is 519 with a standard deviation of 45.8. The mean scale score for female students is 526 with a standard deviation of 41.2, and the mean scale score for male students is 511 with a standard deviation of 48.6.

The scale score frequency distribution of the third-grade Spanish Reading assessment for the total population is shown in Table 158. Figure 34 graphically represents the scale score frequency distributions for the total population and for the groups of male and female students separately. The figure shows that the distributions of scale scores for the total population and for each gender are slightly negatively skewed.

The mean scale score for the single content standard is 519 and a median of 523. The mean scale score for all the subcontent areas is 518; the median scale scores for the subcontent areas vary between 521 and 523, and all are close to the median for the total test scale score of 523.

The mean proportions of the maximum obtainable raw score for the subcontent areas range from 54.9 to 60.7. The mean proportion of the maximum obtainable raw score for the total test is 58.3.

Writing

The mean and median scale scores for the total population of students taking the 2008 third-grade Writing assessment are 466 and 465, respectively, with a standard deviation of 52.3. The mean scale score for female students is 474 with a standard deviation of 52.1, and the mean scale score for male students is 458 with a standard deviation of 51.2.

The scale score frequency distribution for the total population is shown in Table 159. Figure 35 graphically represents the scale score frequency distributions for the total population and for the groups of male and female students separately. The figure shows that the distributions of scale scores for the total population and for each gender are approximately normal.

The mean scale scores for the two content standards are 468 and 472. The mean scale scores for the subcontent areas range from 469 to 487. The median scale score ranges from 465 to 466 for the content standards, and from 466 to 470 for the subcontent areas.

The mean proportions of the maximum obtainable score for the content standards range from 76.4 on CS 2 (Write for a Variety of Purposes) to 77.9 on CS 3 (Write Using Conventions). The mean proportions of the maximum obtainable raw score for the subcontent areas range from 73.9 to 79.7. The mean proportion of the maximum obtainable raw score for the total test is 77.2.

Writing – Spanish

The mean and median scale scores for the total population of students taking the 2008 third-grade Spanish Writing assessment are 501 and 502, respectively, with a standard deviation of 72.2. The mean scale score for female students is 517

with a standard deviation of 70.8, and the mean scale score for male students is 487 with a standard deviation of 70.5.

The scale score frequency distribution of the third-grade Spanish Writing assessment for the total population is shown in Table 160. Figure 36 graphically represents the scale score frequency distributions for the total population and for the groups of male and female students separately. The figure shows that the scale score distributions for the total population and for each gender are approximately normal.

The mean scale scores for the two content standards are 502 and 506, respectively, with median scale scores of 502 and 504. The mean scale scores for the subcontent areas range from 507 to 514, and the median scale scores for the subcontent areas vary between 505 and 512.

The mean proportions of the maximum obtainable raw score for the content standards range from 62.0 on CS 2 (Write for a Variety of Purposes) to 72.2 on CS 3 (Write Using Conventions), and from 61.9 to 73.3 for the subcontent areas. The mean proportion of the maximum obtainable raw score for the total test is 67.8.

Mathematics

The mean and median scale scores for the total population of students taking the 2008 third-grade Mathematics assessment are 464 and 465, respectively, with a standard deviation of 92.2. The mean scale score for female students is 462 with a standard deviation of 90.7, and the mean scale score for male students is 465 with a standard deviation of 93.5.

The scale score frequency distribution for the total population is shown in Table 161. Figure 37 graphically represents the frequency distributions for the total population and for the groups of male and female students separately. The figure shows that the scale score distributions for the total population and for each gender are approximately normal, with a small group of students located at the HOSS.

The mean scale scores for the content standards range from 470 to 483. The median scale score is between 464 and 472 for the content standards.

The mean proportions of the maximum obtainable raw score for the content standards range from 66.6 on CS 2 (Algebra, Patterns, and Functions) to 73.0 on CS 4 (Geometry and Measurement). The mean proportion of the maximum obtainable raw score for the total test is 70.7.

Fourth Grade

Reading

The mean and median scale scores for the total population of students taking the 2008 fourth-grade Reading assessment are 586 and 594, respectively, with a standard deviation of 61.8. The mean scale score for female students is 592 with a standard deviation of 58.1, and the mean scale score for male students is 580 with a standard deviation of 64.6.

The scale score frequency distribution for the total population is shown in Table 162. Figure 38 graphically represents the scale score frequency distributions for the total population and for the groups of male and female students separately. The figure shows that the distributions of scale scores for the total population and for each gender are slightly negatively skewed.

The mean scale scores for the content standards range from 585 to 588. The mean scale scores for the subcontent areas range from 586 to 646. The median scale scores vary between 592 and 594 for the content standards and between 591 and 594 for the subcontent areas.

The mean proportions of the maximum obtainable raw score for the content standards range from 50.5 on CS 6 (Literature) to 69.2 on CS1 (Reading Comprehension). The mean proportion of the maximum obtainable raw score for the total test is 62.0. The mean proportions of the maximum raw score for the subcontent areas range from 58.2 to 76.9.

Reading – Spanish

The mean scale score for the total population of students taking the 2008 fourth-grade Spanish Reading assessment is 519 with a standard deviation of 43.1. The mean scale score for female students is 525 with a standard deviation of 39.4, and the mean scale score for male students is 513 with a standard deviation of 46.0.

The scale score frequency distribution for the total population is shown in Table 163. Figure 39 graphically represents the scale score frequency distributions for the total population and for the groups of male and female students separately. The figure shows that the distributions of scale scores for the total population and for each gender are relatively normal.

The mean scale scores for the content standards range from 511 to 516. The mean scale scores for the subcontent areas range from 515 to 518. The median scale scores vary between 514 and 527 for the content standards and between 521 and 528 for the subcontent areas, and all are close to the median for the total test scale score of 523.

The mean proportions of the maximum obtainable raw score for the content standards range from 43.6 on CS 6 (Literature) to 56.2 on CS 1 (Reading Comprehension). The mean proportion of the maximum obtainable score for the total test is 50.7. The mean proportions of the maximum raw score for the subcontent areas range from 47.3 to 62.2.

Writing

The mean and median scale scores for the total population of students taking the 2008 fourth-grade Writing assessment are 488 and 487 respectively with a standard deviation of 51.7. The mean scale score for female students is 496 with a standard deviation of 52.9, and the mean scale score for male students is 479 with a standard deviation of 48.9.

The scale score frequency distribution for the total population is shown in Table 164. Figure 40 graphically represents the scale score frequency distributions for the total population and for the groups of male and female students separately. The figure shows that the scale score distributions for the total population and for each gender are approximately normal.

The mean scale scores for the content standards vary between 489 and 492. The mean scale scores for the subcontent areas range from 489 to 517. The median scale score is 487 for both content standards, and between 487 and 495 for the subcontent areas.

The mean proportions of the maximum obtainable raw score for the content standards range from 68.8 on CS 2 (Write for a Variety of Purposes) to 75.4 on CS 3 (Write Using Conventions). The mean proportion of the maximum obtainable raw score for the total test is 71.9. The mean proportions of the maximum raw score for the subcontent areas range from 59.4 to 79.8.

Writing – Spanish

The mean and median scale scores for the total population of students taking the 2008 fourth-grade Spanish Writing assessment is 500 for both, with a standard deviation of 38.6. The mean scale score for female students is 508 with a standard deviation of 36.3, and the mean scale score for male students is 490 with a standard deviation of 38.9.

The scale score frequency distribution for the total population is shown in Table 165. Figure 41 graphically represents the scale score frequency distributions for the total population and for the groups of male and female students separately. The figure shows that the distributions of scale scores for the total population and for each gender are approximately normal.

The mean scale score for each of the two content standards (Write for a Variety of Purposes, Write Using Conventions) ranges from 496 to 500. The mean scale scores for the subcontent areas range from 487 to 499. The median scale scores for the two content standards are 494 and 504. The median scale scores for the subcontent areas vary between 490 and 507.

The mean proportions of the maximum obtainable raw score for the content standards range from 48.7 on CS 2 (Write for a Variety of Purposes) to 50.2 on CS 3 (Write Using Conventions). The mean proportion of the maximum obtainable raw score for the total test is 49.5. The mean proportions of the maximum raw score for the subcontent areas range from 43.1 to 54.2.

Mathematics

The mean and median scale scores for the total population of students taking the 2008 fourth-grade Mathematics assessment are 489 and 491, respectively, with a standard deviation of 78.4. The mean scale score for female students is 487 with a standard deviation of 78.0, and the mean scale score for male students is 490 with a standard deviation of 79.3.

The scale score frequency distribution for the total population is shown in Table 166. Figure 42 graphically represents the frequency distributions for the total population and for the groups of male and female students separately. The figure shows that the scale score distributions for the total population and for each gender are approximately normal.

The mean scale scores for the content standards range from 494 to 511. The mean scale scores for the subcontent areas range from 498 to 513. The median scale scores range from 489 to 492 for the content standards and from 482 to 493 for the subcontent areas.

The mean proportions of the maximum obtainable raw score for the content standards range from 62.7 on CS 1 (Number Sense) to 77.1 on CS 4 (Geometry and Measurement). The mean proportion of the maximum obtainable raw score for the total test is 71.2. The mean proportions of the maximum raw score for the subcontent areas range from 69.0 to 75.0.

Fifth Grade

Reading

The mean scale score for the total population of students taking the 2008 fifth-grade Reading assessment is 613 with a standard deviation of 69.7. The mean scale score for female students is 620 with a standard deviation of 66.4, and the mean scale score for male students is 607 with a standard deviation of 72.1.

The scale score frequency distribution for the total population is shown in Table 167. Figure 43 graphically represents the frequency distributions for the total population and for the groups of male and female students separately. The figure shows that the scale score distributions for the total population and for each gender are slightly negatively skewed.

The mean scale scores for the content standards range from 611 to 619. The mean scale scores for the subcontent areas range from 614 to 649. The median scale scores vary from 620 to 622 for the content standards and from 621 to 622 for the subcontent areas, and all are close to the median for the total test scale score of 621.

The mean proportions of the maximum obtainable raw score for content standards range from 54.5 on CS 6 (Literature) to 70.4 on CS 1 (Reading Comprehension). The mean proportion of the maximum obtainable raw score for the total test is 62. The mean proportions of the maximum raw score for the subcontent areas range from 57.2 to 74.6.

Writing

The mean scale score for the total population of students taking the 2008 fifth-grade Writing assessment is 510 with a standard deviation of 58.4. The mean scale score for female students is 519 with a standard deviation of 58.0, and the mean scale score for male students is 501 with a standard deviation of 57.5.

The scale score frequency distribution for the total population is shown in Table 168. Figure 44 graphically represents the scale score frequency distributions for the total population and for the groups of male and female students separately. The figure shows that the scale score distributions for the total population and for each gender are approximately normal.

The mean scale scores for the content standards vary between 512 and 513. The mean scale scores for the subcontent areas range from 512 to 535. The median scale score is 511 for both the content standards and between 510 and 536 for the subcontent areas. Most median scale scores for the content standards and subcontent areas are close to the median for the total test scale score of 511.

The mean proportions of the maximum obtainable raw score for content standards range from 68.5 on CS 2 (Write for a Variety of Purposes) to 72.5 on CS 3 (Write Using Conventions). The mean proportion of the maximum obtainable raw score for the total test is 70.4. The mean proportions of the maximum raw score for the subcontent areas range from 61.0 to 77.1.

Mathematics

The mean and median scale scores for the total population of students taking the 2008 fifth-grade Mathematics assessment are 520 and 522, respectively, with a standard deviation of 72.5. The mean scale score for female students is 519 with a standard deviation of 69.9, and the mean scale score for male students is 521 with a standard deviation of 74.8.

The scale score frequency distribution for the total population is shown in Table 169. Figure 45 graphically represents the frequency distributions for the total population and for the groups of male and female students separately. The figure shows that the scale score distributions for the total population and for each gender are approximately normal.

The mean scale scores for the content standards range from 525 to 538. The mean scale scores for the subcontent areas range from 529 to 547. The median scale scores vary from 522 to 524 for the content standards and from 521 to 526 for the subcontent areas.

The mean proportions of the maximum obtainable raw score for the content standards range from 66.0 on CS 4/5 (Geometry and Measurement) to 73.1 on CS 1 (Number Sense). The mean proportion of the maximum obtainable raw score for the total test is 68.3. The mean proportions of the maximum raw score for the subcontent areas range from 66.2 to 73.6.

Science

The mean scale score for the total population of students taking the 2008 fifth-grade Science assessment is 493 with a standard deviation of 63.6. The mean scale score for female students is 490 with a standard deviation of 61.7, and the mean scale score for male students is 497 with a standard deviation of 65.1.

The scale score frequency distribution for the total population is shown in Table 170. Figure 46 graphically represents the frequency distributions for the total population and for the groups of male and female students separately. The distributions of the scale scores are approximately normal for the total population and for each gender.

The mean scale scores for the content standards range from 493 to 502. The mean scale scores for the subcontent areas range from 512 to 523. The median scale scores vary from 497 to 499 for the content standards and from 498 to 499 for the subcontent areas, and most are very close to the median for the total test scale score of 498.

The mean proportions of the maximum obtainable raw score for the content standards range from 56.8 on CS 2 (Physical Science) to 72.2 on CS 1/5 (Scientific Investigations and the Nature of Science). The mean proportion of the

maximum obtainable raw score for the total test is 66.9. The mean proportions of the maximum raw score for the subcontent areas range from 71.7 to 73.0.

Sixth Grade

Reading

The mean scale score for the total population of students taking the 2008 sixth-grade Reading assessment is 628 with a standard deviation of 65.0. The mean scale score for female students is 636 with a standard deviation of 61.7, and the mean scale score for male students is 620 with a standard deviation of 67.1.

The scale score frequency distribution for the total population is shown in Table 171. Figure 47 graphically represents the frequency distributions for the total population and for the groups of male and female students separately. The figure shows that the scale score distributions for the total population and for each gender are slightly negatively skewed.

The mean scale scores for the content standards range from 622 to 649. The mean scale scores for the subcontent areas range from 621 to 658. The median scale scores vary from 633 to 635 for both the content standards and the subcontent areas, and all are close to the median for the total test scale score of 634.

The mean proportions of the maximum obtainable raw score for content standards range from 48.6 on CS 6 (Literature) to 76.4 on CS 5 (Use of Literary Information).The mean proportion of the maximum obtainable raw score for the total test is 61.6. The mean proportions of the maximum raw score for the subcontent areas range from 49.0 to 76.5.

Writing

The mean and median scale scores for the total population of students taking the 2008 sixth-grade Writing assessment is 526 for both with a standard deviation of 63.3. The mean scale score for female students is 538 with a standard deviation of 62.1, and the mean scale score for male students is 515 with a standard deviation of 62.4.

The scale score frequency distribution for the total population is shown in Table 172. Figure 48 graphically represents the scale score frequency distributions for the total population and for the groups of male and female students separately. The figure shows that the scale score distributions for the total population and for each gender are approximately normal.

The mean scale scores for the content standards vary between 528 and 530. The mean scale scores for the subcontent areas range from 532 to 558. The median scale scores range from 526 to 527 for both of the content standards and from 527 to 538 for the subcontent areas.

The mean proportions of the maximum obtainable raw score for content standards range from 70.7 on CS 3 (Write Using Conventions) to 70.9 on CS 2 (Write for a Variety of Purposes). The mean proportion of the maximum obtainable raw score for the total test is 70.8. The mean proportions of the maximum raw score for the subcontent areas range from 66.0 to 79.3.

Mathematics

The mean scale score for the total population of students taking the 2008 sixth-grade Mathematics assessment is 537 with a standard deviation of 75.4. The mean scale score for female students is 538 with a standard deviation of 72.6, and the mean scale score for male students is 537 with a standard deviation of 78.0.

The scale score frequency distribution for the total population is shown in Table 173. Figure 49 graphically represents the frequency distributions for the total population and for the groups of male and female students separately. The figure shows that the distributions of scale scores for the total population and for each gender are approximately normal.

The mean scale scores for the content standards range from 538 to 567. The mean scale scores for subcontent areas range from 539 to 555. The median scale scores vary between 540 and 544 for the content standards and between 539 and 541 for the subcontent areas, and all are close to the median for the total test scale score of 540.

The mean proportions of the maximum obtainable raw score for the content standards range from 58.4 on CS 2 (Algebra, Patterns, and Functions) to 73.5 on CS 6 (Computational Techniques).The mean proportion of the maximum obtainable raw score for the total test is 64.3. The mean proportions of the maximum raw score for the subcontent areas range from 56.7 to 67.5.

Seventh Grade

Reading

The mean scale score for the total population of students taking the 2008 seventh-grade Reading assessment is 638 with a standard deviation of 64.3. The mean scale score for female students is 649 with a standard deviation of

61.8, and the mean scale score for male students is 628 with a standard deviation of 64.9.

The scale score frequency distribution for the total population is shown in Table 174. Figure 50 graphically represents the frequency distributions for total population and for the groups of male and female students separately. The figure indicates that the distribution of the scale scores for the total population and for each gender is slightly negatively skewed.

The mean scale scores for the content standards range from 638 to 640. The mean scale scores for the subcontent areas range from 637 to 667. The median scale scores vary from 643 to 645 for the content standards and from 643 to 646 for the subcontent areas, and all are close to the median for the total test scale score of 643

The mean proportions of the maximum obtainable raw score for the content standards range from 58.2 on CS 6 (Literature) to 63.7 on CS 1 (Reading Comprehension). The mean proportion of the maximum obtainable raw score for the total test is 61.0. The mean proportions of the maximum raw score for the subcontent areas range from 53.9 to 70.6.

Writing

The mean scale score for the total population of students taking the 2008 seventh-grade Writing assessment is 551 with a standard deviation of 70.3. The mean scale score for female students is 567 with a standard deviation of 68.2, and the mean scale score for male students is 535 with a standard deviation of 69.0.

The scale score frequency distribution for the total population is shown in Table 175. Figure 51 graphically represents the frequency distributions for the total population and for the groups of male and female students separately. The figure indicates that the scale score distributions are approximately normal for the total population and for each gender.

The mean scale score for the content standards ranges from 552 to 553. The mean scale scores for the subcontent areas range from 552 to 580. The median scale scores vary from 552 to 553 for the content standards and from 550 to 572 for the subcontent areas. Most of the median scale scores for content standards and subcontent areas are close to the median for the total test scale score of 552.

The mean proportions of the maximum obtainable raw score for content standards range from 65.6 on CS 3 (Write Using Conventions) to 66.9 on CS 2 (Write for a Variety of Purposes). The mean proportion of the maximum

obtainable raw score for the total test is 66.2. The mean proportions of the maximum raw score for the subcontent areas range from 62.2 to 78.5.

Mathematics

The mean scale score for the total population of students taking the 2008 seventh-grade Mathematics assessment is 548 with a standard deviation of 73.4. The mean scale score for female students is 549 with a standard deviation of 69.7, and the mean scale score for male students is 548 with a standard deviation of 76.8.

The scale score frequency distribution for the total population is shown in Table 176. Figure 52 graphically represents the frequency distributions for the total population and for the groups of male and female students separately. The figure indicates that the scale score distributions are approximately normal for the total population and for each gender.

The mean scale scores for the content standards range from 545 to 550. The mean scale scores for the subcontent areas range from 532 to 547. The median scale scores vary from 551 to 553 for both the content standards and the subcontent areas, and all are close to the median for the total test scale score of 552.

The mean proportions of the maximum obtainable raw score for the content standards range from 44.8 on CS 4/5 (Geometry and Measurement) to 55.7 on CS 2 (Algebra, Patterns, and Functions). The mean proportion of the maximum obtainable raw score for the total test is 51.1. The mean proportions of the maximum raw score for the subcontent areas range from 40.1 to 46.3.

Eighth Grade

Reading

The mean scale score for the total population of students taking the 2008 eighth-grade Reading assessment is 652 with a standard deviation of 62.8. The mean scale score for female students is 661 with a standard deviation of 59.6, and the mean scale score for male students is 643 with a standard deviation of 64.5.

The scale score frequency distribution for the total population is shown in Table 177. Figure 53 graphically represents the frequency distributions for the total population and for the groups of male and female students separately. The figure shows that the scale score distributions for the total population and for each gender are slightly negatively skewed.

The mean scale scores for the content standards range from 652 to 654. The mean scale scores for the subcontent areas range from 648 to 702. The median scale scores vary from 657 to 660 for the content standards and from 658 to 674 for the subcontent areas. Most of the median scale scores for content standards and subcontent areas are close to the median for the total test scale score of 658.

The mean proportions of the maximum obtainable raw score for the content standards range from 54.9 on CS 6 (Literature) to 65.1 on CS 4 (Thinking Skills). The mean proportion of the maximum obtainable raw score for the total test is 61.1. The mean proportions of the maximum raw score for the subcontent areas range from 49.3 to 70.4.

Writing

The mean scale score for the total population of students taking the 2008 eighth-grade Writing assessment is 560 with a standard deviation of 75.8. The mean scale score for female students is 577 with a standard deviation of 73.8, and the mean scale score for male students is 544 with a standard deviation of 74.1.

The scale score frequency distribution for the total population is shown in Table 178. Figure 54 graphically represents the frequency distributions for the total population and for the groups of male and female students separately. The figure indicates that the scale score distributions are approximately normal for the total population and for each gender.

The mean scale score for both the content standards is 562. The mean scale scores for the subcontent areas range from 559 to 580. The median scale scores vary from 561 to 562 for the content standards and from 561 to 612 for the subcontent areas, and most median scale scores are close to the median for the total test scale score of 562

The mean proportions of the maximum obtainable raw score for content standards range from 64.9 on CS 3 (Write Using Conventions) to 68.7 on CS 2 (Write for a Variety of Purposes). The mean proportion of the maximum obtainable raw score for the total test is 66.8. The mean proportions of the maximum raw score for the subcontent areas range from 58.0 to 76.3.

Mathematics

The mean scale score for the total population of students taking the 2008 eighth-grade Mathematics assessment is 568 with a standard deviation of 74.1. The mean scale score for female students is 566 with a standard deviation of 70.8, and the mean scale score for male students is 570 with a standard deviation of 77.0.

The scale score frequency distribution for the total population is shown in Table 179. Figure 55 graphically represents the frequency distributions for the total population and for the groups of male and female students separately. The scale score distributions are slightly negatively skewed (with a small group of students located at the LOSS) for the total population and for each gender.

The mean scale scores for the content standards range from 564 to 570. The mean scale scores for subcontent areas range from 557 to 572. The median scale scores vary between 571 and 574 for the content standards and between 569 and 572 for the subcontent areas, and all are fairly close to the median for the total test scale score of 572.

The mean proportions of the maximum obtainable raw score for the content standards range from 44.7 on CS 4 (Geometry and Measurement) to 56.8 on CS 2 (Algebra, Patterns, and Functions).The mean proportion of the maximum obtainable raw score for the total test is 50.3. The mean proportions of the maximum raw score for the subcontent areas range from 38.9 to 56.5.

Science

The mean scale score for the total population of students taking the 2008 eighth-grade Science assessment is 495 with a standard deviation of 63.5. The mean scale score for female students is 492 with a standard deviation of 60.6, and the mean scale score for male students is 498 with a standard deviation of 66.0.

The scale score frequency distribution for the total population is shown in Table 180. Figure 56 graphically represents the frequency distributions for the total population and for the groups of male and female students separately. The distributions of the scale scores are slightly negatively skewed (with a small group of students at the LOSS) for the total population and for each gender.

The mean scale scores for the content standards range from 486 to 495. The mean scale scores for the subcontent areas range from 483 to 500. The median scale scores vary between 501 and 502 for the content standards and between 500 and 503 for the subcontent areas, and most are very close to the median for the total test scale score of 502.

The mean proportions of the maximum obtainable raw score for the content standards range from 38.2 on CS 4 (Earth & Space Science) to 56.5 on CS 1/5 (Scientific Investigations and the Nature of Science). The mean proportion of the maximum obtainable raw score for the total test is 48.7. The mean proportions of the maximum raw score for the subcontent areas range from 39.5 to 58.2.

Ninth Grade

Reading

The mean scale score for the total population of students taking the 2008 ninth-grade Reading assessment is 661 with a standard deviation of 57.8. The mean scale score for female students is 669 with a standard deviation of 53.5, and the mean scale score for male students is 652 with a standard deviation of 60.5.

The scale score frequency distribution for the total population is shown in Table 181. Figure 57 graphically represents the frequency distributions for the total population and for the groups of male and female students separately. The figure shows that the scale score distributions for the total population and for each gender are slightly negatively skewed, with a small group of students at the LOSS.

The mean scale scores for the content standards range from 659 to 667. The mean scale scores for the subcontent areas range from 655 to 665. The median scale scores vary between 664 and 667 for both the content standards and the subcontent areas, and all are close to the median for the total test scale score of 666.

The mean proportions of the maximum obtainable raw score for the content standards range from 51.6 on CS 6 (Literature) to 65.9 on CS 5 (Use of Literary Information). The mean proportion of the maximum obtainable raw score for the total test is 61.4. The mean proportions of the maximum raw score for the subcontent areas range from 49.1 to 66.4.

Writing

The mean scale score for the total population of students taking the 2008 ninth-grade Writing assessment is 561 with a standard deviation of 78.7. The mean scale score for female students is 577 with a standard deviation of 76.7, and the mean scale score for male students is 545 with a standard deviation of 77.3.

The scale score frequency distribution for the total population is shown in Table 182. Figure 58 graphically represents the frequency distributions for the total population and for the groups of male and female students separately. The figure indicates that the scale score distributions are approximately normal for the total population and for each gender.

The mean scale score for the content standards ranges from 562 to 564. The mean scale scores for subcontent areas range from 562 to 579. The median scale scores vary between 562 and 563 for the content standards and between 498 and 565 for the subcontent areas, and most, with the exception of SA 6 with a median of 498, are close to the median for the total test scale score of 562.

The median scale score for SA 6 (Extended Writing) was somewhat lower than the median for the total test score. It should be noted that the score for this subcontent area is computed on the basis of the four scores a student gets for his or her response to the extended writing prompt. Consequently, the scale score variable for this subcontent area is rather discrete.

The mean proportions of the maximum obtainable raw score for the content standards range from 65.1 on CS 2 (Write for a Variety of Purposes) to 67.6 on CS 3 (Write Using Conventions). The mean proportion of the maximum obtainable raw score for the total test is 66.3. The mean proportions of the maximum raw score for the subcontent areas range from 61.6 to 70.8.

Mathematics

The mean scale score for the total population of students taking the 2008 ninth-grade Mathematics assessment is 577 with a standard deviation of 71.3. The mean scale score for female students is 575 with a standard deviation of 68.2, and the mean scale score for male students is 578 with a standard deviation of 74.0.

The scale score frequency distribution for the total population is shown in Table 183. Figure 59 graphically represents the frequency distributions for the total population and for the groups of male and female students separately. The scale score distributions are slightly negatively skewed (with a small group of students at the LOSS) for the total population and for each gender.

The mean scale scores for the content standards range from 555 to 578. The mean scale scores for the subcontent areas are 564 to 576. The median scale scores vary between 582 and 583 for the content standards, and between 581 and 582 for the subcontent areas, and all are close to the median for the total test scale score of 582.

The mean proportions of the maximum obtainable raw score for the content standards range from 31.6 on CS 4 (Geometry and Measurement) to 49.6 on CS 2 (Algebra, Patterns, and Functions). The mean proportion of the maximum obtainable raw score for the total test is 44.1. The mean proportions of the maximum raw score for the subcontent areas range from 37.8 to 43.9.

Tenth Grade

Reading

The mean scale score for the total population of students taking the 2008 tenth-grade Reading assessment is 682 with a standard deviation of 61.1. The mean scale score for female students is 693 with a standard deviation of 55.7, and the mean scale score for male students is 670 with a standard deviation of 64.0.

The scale score frequency distribution for the total population is shown in Table 184. Figure 60 graphically represents the frequency distributions for total population and for the groups of male and female students separately. The figure shows that the scale score distributions for the total population and for each gender are negatively skewed.

The mean scale scores for the content standards range from 679 to 687. The mean scale scores for the subcontent areas range from 680 to 692. The median scale scores vary from 687 to 689 for the content standards and from 688 to 693 for the subcontent areas, and all are close to the median for the total test scale score of 689.

The mean proportions of the maximum obtainable raw score for the content standards range from 48.6 on CS 6 (Literature) to 67.7 on CS 5 (Use of Literary Information). The mean proportion of the maximum obtainable raw score for the total test is 60.3. The mean proportions of the maximum raw score for the subcontent areas range from 50.7 to 64.9.

Writing

The mean scale score for the total population of students taking the 2008 tenth-grade Writing assessment is 572 with a standard deviation of 87.3. The mean scale score for female students is 590 with a standard deviation of 83.3, and the mean scale score for male students is 554 with a standard deviation of 87.5.

The scale score frequency distribution for the total population is shown in Table 185. Figure 61 graphically represents the frequency distributions for the total population and for the groups of male and female students separately. The figure shows that the scale score distributions for the total population and for each gender are approximately normal.

The mean scale scores for the content standards vary between 574 and 575. The mean scale scores for the subcontent areas range from 576 to 590. The median scale scores vary between 573 and 575 for the content standards and between 572 and 635 for the subcontent areas, and most, with the exception of SA 6 with a median of 635, are close to the median for the total test scale score of 575.

The mean proportions of the maximum obtainable raw score for the content standards range from 65.1 on CS 3 (Write Using Conventions) to 67.3 on CS 2 (Write for a Variety of Purposes). The mean proportion of the maximum obtainable raw score for the total test is 66.3. The mean proportions of the maximum raw score for the subcontent areas range from 63.7 to 74.5.

Mathematics

The mean scale score for total population of students taking the 2008 tenth-grade Mathematics assessment is 586 with a standard deviation of 74.0. The mean scale score for female students is 586 with a standard deviation of 69.8, and the mean scale score for male students is 586 with a standard deviation of 77.9.

The scale score frequency distribution for the total population is shown in Table 186. Figure 62 graphically represents the frequency distributions for the total population and for the groups of male and female students separately. The figure shows that the scale score distributions for the total population and for each gender are approximately normal with a group of students at the LOSS.

The mean scale scores for the content standards range from 574 to 586. The mean scale scores for the subcontent areas range from 587 to 588. The median scale scores vary between 593 and 595 for the content standards and between 593 and 602 for the subcontent areas, and most are close to the median for the total test scale score of 594.

The mean proportions of the maximum obtainable raw score for the content standards range from 34.9 on CS 4 (Geometry and Measurement) to 50.0 on CS 2 (Algebra, Patterns, and Functions) The mean proportion of the maximum obtainable raw score for the total test is 42.3. The mean proportions of the maximum raw score for the subcontent areas range from 37.6 to 48.6.

Science

The mean scale score for the total population of students taking the 2008 tenth-grade Science assessment is 497 with a standard deviation of 62.0. The mean scale score for female students is 496 with a standard deviation of 58.2, and the mean scale score for male students is 498 with a standard deviation of 65.4.

The scale score frequency distribution for the total population is shown in Table 187. Figure 63 graphically represents the frequency distributions for the total population and for the groups of male and female students separately. The distributions of the scale scores are slightly negatively skewed (with a group of students at the LOSS) for the total population and for each gender.

The mean scale scores for the content standards range from 490 to 498. The mean scale scores for the subcontent areas range from 487 to 509. The median scale scores vary from 504 to 506 for the content standards and from 502 to 506 for the subcontent areas, and most are very close to the median for the total test scale score of 505.

The mean proportions of the maximum obtainable raw score for the content standards range from 39.8 on CS 2 (Physical Science) to 55.2 on CS 1/5(Scientific Investigations and the Nature of Science). The mean proportion of the maximum obtainable raw score for the total test is 51.3. The mean proportions of the maximum raw score for the subcontent areas range from 39.1 to 60.4.

Correlations Among Content Standards and Among Subcontent Areas

Tables 188 through 218 show the correlations between the scale scores for the total test and for the various content standards and subcontent areas for each grade and content area. All content standards and subcontent areas are positively correlated, as would be expected.

For the Reading assessments, the correlation coefficients vary between 0.58 (grade 6) and 0.77 (grades 8 and 9) for the relationship between the various content standards and between 0.47 (grade 10) and 0.76 (grade 4) for the relationship between the various subcontent areas, respectively.

For the third-grade Spanish Reading assessments, correlations among subcontent areas vary between 0.60 and 0.65. For the fourth-grade Spanish Reading assessments, the correlations among the various content standards vary between 0.53 and 0.72 and the correlations among subcontent areas vary between 0.59 and 0.75.

For the Writing assessments, the coefficients for the correlation between content standards 2 and 3 vary between 0.68 (grade 3) and 0.78 (grade 9). The correlations among the various subcontent areas vary between 0.35 (grade 5) and 0.73 (grade 10).

For the Spanish Writing assessments, the correlation between content standards 2 and 3 varies between 0.66 (grade 4) and 0.76 (grade 3) and the correlations between the various subcontent areas vary between 0.18 (grade 4) and 0.59 (grade 3).

For the Mathematics assessments, the correlations vary between 0.57 (grade 3) and 0.80 (grade 9) for the relationship among the content standards and between 0.54 (grade 4) and 0.71 (grade 9) for the relationship among the subcontent areas.

Finally, for the Science assessments, the correlation coefficients vary between 0.59 (grade 5) and 0.76 (grade 8) for the relationship among the content standards and between 0.50 (grade 5) and 0.70 (grade 8) for the relationship among the subcontent areas.

Part 8: Reliability and Validity Evidence

Part 8 describes reliability and validity evidence for the 2008 CSAP assessments. First, the total test and subgroup reliability coefficients are presented, measured by Cronbach's alpha, as an index of the internal consistency, followed by interrater reliability of constructed-response items, item-to-total score correlation, and items functioning differentially in the CSAP tests. The section further discusses the reliability in terms of standard error of measurement of scale scores.

Second, the test validity in terms of content-related validity, construct-related validity, factor structures, fit and DIF, divergent or discriminant validity, and predictive validity of the CSAP tests are described. Finally, the section is concluded by presenting results from classification consistency and accuracy analyses.

Total Test and Subgroup Reliability

Reliability is an index of the consistency of test results. A reliable test is one that produces scores that are expected to be relatively stable if the test is administered repeatedly under similar conditions. Cronbach's alpha is a frequently used measure of internal consistency. On the basis of a single administration of a test, Cronbach's alpha provides a reliability estimate that equals the average of all split-half coefficients that would be obtained on all possible divisions of the test into halves. Such a split-half coefficient would be obtained by correlating one half of the test with the other half and then adjusting the correlation with the Spearman–Brown formula so that it applies to the whole test (see Allen & Yen, 1979, pp. 83–88).

Total test reliability coefficients (in this case measured by Cronbach's alpha) may range from 0.00 to 1.00, where 1.00 refers to a perfectly consistent test. The data are based on representative samples from each grade (the calibration sample), and they are typical of the results obtained for all CSAP operational tests. The total test reliabilities of the operational forms were evaluated first by Cronbach's alpha (Cronbach, 1951) calculated as

$$\hat{\alpha} = \frac{k}{k-1}\left(1 - \frac{\sum \hat{\sigma}_i^2}{\hat{\sigma}_X^2}\right),$$

where k is the number of items on the test form, $\hat{\sigma}_i^2$ is the variance of item i, and $\hat{\sigma}_X^2$ is the total test variance. Achievement tests are typically considered of sound reliability when their reliability coefficients are in the range of 0.80 and above. Tables 219 and 220 show the Cronbach's alpha for the grades and

content areas involved in the Spring 2008 operational CSAP test administration for content standards and subcontent areas. At the state level, the reliabilities ranged between 0.87 (grade 4 Spanish Writing) and 0.94 (grades 4, 5, 6 and 8 Mathematics) with a median value of 0.92. Such a reliability coefficient range is indicative of high internal consistency and signifies that the CSAP tests produce relatively stable scores. The median coefficients for each content area are as follows:

Test	Median	Range
Reading (English)	0.920	(0.90–0.93)
Reading (Spanish)	0.895	(0.87–0.92)
Mathematics	0.935	(0.91–0.94)
Writing (English)	0.910	(0.90–0.92)
Writing (Spanish)	0.885	(0.85–0.92)
Science	0.930	(0.92–0.93)

Table 219 also shows the reliability coefficients for content standards. Table 220 provides similar information for all of the subcontent areas. These coefficients tend to be somewhat lower than the coefficients for the total test scores. These results are consistent with the smaller numbers of items that contribute to each standard and subcontent area.

As evidence that a test is performing similarly across various subgroups, the reliability values for these subgroups to those for the total population can be examined. The reliability measures are impacted by the population distribution and can be lowered when the subgroup is considerably less variable than the total population. However, one would expect the subgroup reliabilities to be adequately high for all groups. Tables 221 through 226 show the reliability estimates by gender, ethnicity, free lunch eligibility, immigrant status, disability, accommodation and language proficiency. Even at the subgroup level, the ranges were quite similar and the lowest reliability (0.74) was found for the language proficiency – FEP group, in grade 3 Reading (Spanish). All reliabilities are well within acceptable ranges.

The performance of accommodated and nonaccommodated students with and without reported disabilities is summarized in Table 227. Overall, nonaccommodated students scored higher than accommodated students in every grade and content area. As shown in this table, the mean scores of students with reported disabilities were lower than the scores of students without reported disabilities in every grade and content area.

Among students with reported disabilities, the mean scores of students who did not receive accommodations were higher than the scores of students who received accommodations. However, this should not be interpreted as an indication that the testing accommodations were unhelpful, since it is likely that

the disabilities of students receiving accommodations were more severe than those of students who were able to complete the test without accommodations.

It is noteworthy that the difference between the mean scores of students with and without reported disabilities was lower in the accommodated groups than in the nonaccommodated groups for every grade and content area except for grade 3 and grade 4 Reading; for these latter two groups, the score differences between students with and without reported disabilities were similar in the accommodated and nonaccommodated groups.

Interrater Reliability, Item-to-Total Score Correlation, and DIF

Test scores always contain some amount of measurement error. This kind of error can be random or systematic. Standardization of assessments is meant to minimize random error that occurs because of random factors that affect a student's performance on the test. Systematic errors are inherent to examinees and are typically specific to some subgroup characteristic (i.e., students who need accommodations but are not offered them). Reliability refers to the degree to which students' scores are free from such effects and provides a measure of consistency. In other words, reliability helps to describe how consistent students' performance would be if the assessment were given over multiple occasions.

Item specific reliability statistics include interrater reliability, item-to-total score correlation, and DIF. As discussed in Part 4, the interrater reliability across CR items in terms of the kappa and intraclass correlations is one way to measure the consistency of the hand score. Tables 9 through 15 provide the results of both rater reliability measures, which assess the agreement rates within a given administration, and rater severity analyses, which compare the scoring leniency across years. As previously mentioned, these results demonstrate that the CSAP tests have relatively high reader reliability. As shown in rater reliability Tables 9 through 14, the kappa for Mathematics tests ranged from 0.57 to 0.96 with a median value of 0.86. For English Reading, the range was 0.49 to 0.96 with a median value of 0.72. For Spanish Reading, the kappa ranged from 0.58 to 0.99 with a median of 0.86. For Science, the range was 0.41 to 0.95 and the median was 0.76. English Writing kappa values had a wider range, from 0.40 to 0.97 (median = 0.73), as did Spanish Writing, which ranged from 0.0 to 1.00 (median = 0.74). The lower kappa values for some writing items are associated with lower maximum score point(s).

Additionally, Table 15 displays the high consistency of the rating scales that were used from one year to the next. The kappa for Mathematics tests ranged from 0.69 to 0.95 with a median value of 0.86. For English Reading, the range was 0.58–0.95 with a median value of 0.84. For Spanish Reading, the kappa ranged from 0.80 to 0.93 with a median of 0.93. For Science grades 5, 8, and 10, the range was 0.64–0.98 and the median was 0.83. English Writing kappa values had a wider range, from 0.18 to 0.75 (median = 0.59), as did Spanish Writing,

which ranged from 0.44 to 0.84 (median = 0.46). As in the rater reliability the smallest weighted kappa for rater leniency in writing was also observed in the items with parts and lower maximum score point of one.

The reasonable range of weighted kappa for rater leniency for most items is an indication that the standards applied in the scoring of the constructed-response items are quite stable within an administration and over time.

The item-to-total score correlation type of internal consistency measure is one measure of the correlation between each item and the overall test. This provides a source of how consistent the item measures information similar to the other items. As discussed in Part 5, Tables 21 through 82 display the item-to-total score correlations (and *p*-values) for each grade and content area. Below each table are displayed the average values for each statistic. Item-to-total score correlations are calculated and thus dependent on the number of items answered correctly divided by the number of items answered incorrectly. Thus, the *p*-values of the items are important to consider when reviewing the item-to-total score correlations. According to a study cited in Crocker and Algina (1986), if the average biserial correlation is in a range of about 0.30–0.40, the average *p*-value should ideally be between 0.40 and 0.60. Given that the mean item-to-total score correlations for CSAP assessments range from 0.32 to 0.65 across test forms and that the average *p*-values range from 0.34 to 0.82 across forms, the item-to-total score correlations and *p*-values are in a reasonable range.

DIF provides a measure about the systematic error found within subgroups, specifically attributed to some bias or systematic over- or under-representation of subgroup performance compared to total group performance. Items exhibiting DIF have been avoided as much as possible when operational test forms are selected. The CSAP 2008 DIF results are presented in a later section in this part.

Standard Error of Measurement

Another measure of reliability is a direct estimate of the degree of measurement error in a student's total score on a test. In the case of the CSAP, this total score is a scale score. This score is produced by the statistical IRT models that are used to scale, equate, and pattern score the CSAP, as described in the CSAP Equating and Calibration Procedures. This second measure is called a standard error of measurement (SEM). This represents the number of score points about which a given score can vary, similar to the standard deviation of a score: the smaller the SEM, the smaller the variability and higher the reliability. The SEMs are computed with the following formula:

$$\text{SEM} = \text{SD\_SS}(\sqrt{1-\hat{\alpha}})$$

where SD_SS is the standard deviation of the scale score and $\hat{\alpha}$ is the result of the calculation of Cronbach's alpha above. The SEMs represent the total standard error of measurement in the scale score metric. The overall estimates of SEM are shown in Table 228. The scale scores and associated standard errors by content area and grade are shown in Tables 229 through 233. Tables 221 through 226 provide the SEM values for various subgroups by content area and grade. All SEMs are within reasonable limits.

It is most important to note the specific scale score SEM for each cut score. Table 234 shows the cut scores used for the proficiency levels at each grade and content area. Comparison of the SEMs at the proficient cut to the SEMs associated with other CSAP scale scores for each test reveal that these values near the cut score are among the lowest for most grades and content areas, meaning that the CSAP tests tend to measure most accurately near the cut score. This is a desirable quality when cut scores are used to classify examinees.

Test Validity

> Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing and evaluating tests. The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations (AERA, APA, NCME, 1999)

The purpose of test validation is not to validate the test itself but to validate interpretations of the test scores for particular purposes or uses. Test validation is not a quantifiable property but an ongoing process, beginning at initial conceptualization and continuing throughout the lifetime of an assessment. Every aspect of an assessment provides evidence in support of its validity (or evidence to the contrary), including design, content specifications, item development, and psychometric quality.

Content-Related Validity

Content-related validity in achievement tests is evidenced by a correspondence between test content and a specification of the content domain. To ensure such correspondence, the Colorado Department of Education conducted a comprehensive curriculum review. They met with educational experts to determine common educational goals and the knowledge and skills emphasized in curricula. The Colorado Model Content Standards and Assessment Frameworks are the outcomes of the process.

The Colorado Model Content Standards and Assessment Frameworks are the foundation for the CSAP assessments. All CSAP items are developed to

measure the content standards and are subject to numerous levels of scrutiny, both internal and external, before their operational use. All items are closely examined according to the "Criteria for Item Acceptability"[4] to ensure the adequacy and relevancy of each item with respect to content, theme, wording, format, and style prior to formal review by Content and Bias Review panels. Through this process all efforts are made to ensure test items are tightly aligned with the Colorado Model Content Standards. Tables 235 through 238 show for each content area test the number of score reporting categories (SRCs), the number of performance indicators (PIs) in each SRC, the number of items measuring each SRC, the number of PIs assessed by the current test, and finally the percentage of all PIs assessed. It may not be feasible to assess all PIs in a single test; however, as appropriate, efforts are made to assess all measurable PIs across years.

Construct Validity

Construct validity—the meaning of test scores and the inferences they support—is the central concept underlying the CSAP validation process. Evidence for construct validity is comprehensive and integrates evidence from both content- and criterion-related validity. For example, to demonstrate comprehensiveness, CSAP tests must contain items that represent essential instructional objectives. The following sections present evidence supporting content- and criterion-related validity.

Minimization of Construct-Irrelevant Variance and Under-Representation

Minimization of construct-irrelevant variance and construct under-representation is addressed in the following steps of the test development process: (1) specification, (2) item writing, (3) review, (4) field testing, (5) test construction, and (6) calibration. While CSAP does not field test, the quality of the item pool used in the construction of the CSAP assessments is evidenced by the item analysis results and the low number of items suppressed during calibration.

Construct-irrelevant variance refers to error variance that is caused by factors unrelated to the constructs measured by the test. For example, when tests are not administered under standardized conditions (e.g., one administration may be timed, but another administration may be untimed), differences in student performance related to different administration conditions may result. Careful specification of content and review of the items under plain language representing that content are first steps in minimizing construct-irrelevant variance. Then, empirical evidence, especially item-level data, is used to infer construct irrelevance.

[4] This checklist is used to train item writers and when reviewing items for test construction.

Construct under-representation occurs when the content of the assessment does not reflect the full range of content that the assessment is expected to cover. CSAP is designed to represent the Colorado Model Content Standards. Specification and review, in which test blueprints are developed and reviewed, are primary steps in the development process designed to ensure that content is equitably represented.

Minimizing Bias Through DIF Analyses

The position of CTB/McGraw-Hill concerning test bias is based on two general propositions. First, students may differ in their background knowledge, cognitive and academic skills, language, attitudes, and values. To the degree that these differences are large, no one curriculum and no one set of instructional materials will be equally suitable for all. Therefore, no one test will be equally appropriate for all. Furthermore, it is difficult to specify what amount of difference can be called large and to determine how these differences will affect the outcome of a particular test.

Second, schools have been assigned the tasks of developing certain basic cognitive skills and supporting development of these skills equitably among all students. Therefore, there is a need for tests that measure the common skills and bodies of knowledge that are common to all learners. The test publisher's task is to develop assessments that measure these key cognitive skills without introducing extraneous or construct-irrelevant elements in the performances on which the measurement is based. If these tests require that students have cultural specific knowledge and skills not taught in school, differences in performance among students can occur because of differences in student background and out-of-school learning. Such tests are measuring different things for different groups and can be called biased (Camilli & Shepard, 1994; Green, 1975). In order to lessen this bias, CTB/McGraw-Hill strives to minimize the role of the extraneous elements, thereby increasing the number of students for whom the test is appropriate. Careful attention is taken in the test construction process to lessen the influence of these elements for large numbers of students. Unfortunately, in some cases these elements may continue to play a substantial role.

Four measures were taken to minimize bias in the CSAP assessments. The first was based on the premise that careful editorial attention to validity is an essential step in keeping bias to a minimum. Bias can occur only if the test is measuring different things for different groups. If the test entails irrelevant skills or knowledge, however common, the possibility of bias is increased. Thus, careful attention was paid to content validity during the item-writing and item-selection process.

The second way to minimize bias was to follow the McGraw-Hill guidelines designed to reduce or eliminate bias. Item writers were directed to the following published guidelines: *Guidelines for bias-free publishing* (MacMillan/McGraw-Hill,

1993a) and *Reflecting diversity: Multicultural guidelines for educational publishing professionals* (MacMillan/McGraw-Hill, 1993b). Developers reviewed CSAP Assessment materials with these considerations in mind. Such internal editorial reviews were conducted by at least three different people or groups of people: a content editor, who directly supervised the item writers; a style editor; and a content supervisor. The final test was again reviewed by at least these same people, as well as being given an independent review by a quality assurance editor.

As part of the test assembly process, attempts are made to avoid using items with poor statistical fit or distractors with positive item-to-total score correlations, since this may indicate that an item is tapping ability irrelevant to the construct being measured. Differential item functioning with respect to subgroups might also indicate construct irrelevance. Items with these attributes are not selected or are given a lower priority for selection during the test construction stage. For CSAP, particular scrutiny is given to the equating (or "anchor") sets in each form, since these items impact the resulting scale scores developed for the entire test. Including DIF items in this equating set could have a greater impact on the overall fairness of the reported scores. The number of fit and DIF flagged items, including anchor items, included in 2008 test assembly is presented in Table 7.

In the third effort to minimize bias, educational community professionals who represent various ethnic groups reviewed all new materials. They were asked to consider and comment on the appropriateness of language, subject matter, and representation of groups of people.

The fourth procedure, an empirical approach, involves statistical procedures referred to as DIF analyses. A procedure suggested by Linn and Harnisch (1981) was used for the CSAP DIF evaluation.

For all CSAP tests, DIF studies are conducted. DIF studies include a systematic item analysis to determine if examinees with the same underlying level of ability have the same probability of getting the item correct. The inclusion of the items flagged is minimized in the test development process. DIF studies have been routinely done for all major test batteries published by CTB/McGraw-Hill after 1970. Differential item functioning of the CSAP test items was assessed for gender, ethnicity, and students with disabilities.

Because CSAP tests were built using IRT, DIF analyses that capitalized on the information and item statistics provided by this theory were implemented. There are several IRT-based DIF procedures, including those that assess the equality of item parameters across groups (Lord, 1980) and those that assess area differences between item characteristic curves (Camilli & Shepard, 1994; Linn, Levine, Hastings, & Wardrop, 1981). However, these procedures require a minimum of 800–1,000 cases in each group of comparison to produce reliable and consistent results. In contrast, the Linn–Harnisch procedure (Linn &

Harnisch, 1981) utilizes the information provided by the three-parameter IRT model but requires fewer cases. This procedure was used to complete the DIF studies for the 2008 CSAP tests.

After the administration of new forms, all items were evaluated for poor item statistics, fit, and DIF. The items flagged for fit and DIF were noted in the item analyses report and item pool so that content experts will be able to reevaluate the items for future selection.

Linn–Harnisch DIF Method

An example of Linn–Harnisch procedure for gender DIF analyses for multiple-choice items is described below.

The parameters for each item (a_i, b_i, and c_i) and the trait or scale score (θ) for each examinee are estimated for the three-parameter logistic model:

$$P_{ij}(\theta) = c_i + \frac{1-c_i}{1+\exp[-1.7a_i(\theta_j - b_i)]},$$

where $P_{ij}(\theta)$ is the probability that examinee *j*, with a given value of θ, will obtain a correct score on item *i*. Note that the item parameter estimates are based on data from the total sample of valid examinees. The sample is then divided into gender groups, and the members in each group are sorted into 10 equal score categories (deciles) based on their location on the score scale (θ). The expected proportion correct for each group based on the model prediction is compared to the observed (actual) proportion correct obtained by the group.

The proportion of people in decile *g* who are expected to answer item *i* correctly is

$$P_{ij} = P_{ig}(\theta) = \frac{1}{n_g}\sum_{j\,\varepsilon\,g} P_{ij}(\theta),$$

where n_g is the number of examinees in decile *g*. The formula to compute the proportion of students expected to answer item *i* correctly (over all deciles) for a group (e.g., female) is given by

$$P_{i.} = P_i(\theta). = \frac{\sum_{g=1}^{10} n_g P_{ig}(\theta)}{\sum_{g=1}^{10} n_g}.$$

The corresponding observed proportion correct for examinees in a decile (O_{ig}) is the number of examinees in decile *g* who answered item *i* correctly divided by the number of people in the decile (n_g). That is,

$$O_{ig} = \frac{\sum_{j \varepsilon g} u_{ij}}{n_g},$$

where u_{ij} is the dichotomous score for item *i* for examinee *j*.

The corresponding formula to compute the observed proportion answering each item correctly (over all deciles) for a complete gender group is given by

$$O_{i\cdot} = \frac{\sum_{g=1}^{10} n_g O_{ig}}{\sum_{g=1}^{10} n_g}.$$

After the values are calculated for these variables, the difference between the observed proportion correct (for gender) and expected proportion correct can be computed. The decile group difference (D_{ig}) for observed and expected proportion correctly answering item *i* in decile *g* is

$$D_{ig} = O_{ig} - P_{ig},$$

and the overall group difference (D_i) between observed and expected proportion correct for item *i* in the complete group (over all deciles) is

$$D_{i\cdot} = O_{i\cdot} - P_{i\cdot}.$$

These indices are indicators of the degree to which members of gender groups perform better or worse than expected on each item, based on the parameter estimates from all subsamples. Differences for decile groups provide an index for each of the 10 regions on the score (θ) scale. The decile group difference (D_{ig}) can be either positive or negative. Use of the decile group differences as well as the overall group difference allows one to detect items that give a large positive difference in one range of θ and a large negative difference in another range of θ, yet have a small overall difference.

A generalization of the Linn and Harnisch's (1981) procedure was used to measure DIF for constructed-response items.

Differential Item Functioning Ratings and Results

Differential item functioning is defined in terms of the decile group and total target subsample differences, the D_{i-} (sum of the negative group differences) and D_{i+} (sum of the positive group differences) values, and the corresponding standardized difference (Z_i) for the subsample (see Linn & Harnisch, 1981, p. 112). Items for which $|D_i| \geq 0.10$ and $|Z_i| \geq 2.58$ are identified as possibly biased.

If D_i is positive, the item is functioning differentially in favor of the target subsample. If D_i is negative, the item is functioning differentially against the target subsample.

The DIF analyses were conducted for all grades and content areas for African Americans, Hispanics, Asians, males, and females. Table 239 provides an overview of items flagged for gender and ethnicity DIF in the various assessments based on the entire student population. The results for each assessment are briefly described below.

On the Reading assessments, DIF was most evident at the higher grades. DIF was observed in one grade 3 Reading item, one grade 4 Reading item, and one grade 5 Reading item, compared to two grade 6 Reading items, three grade 7 Reading items, four grade 8 Reading items, four grade 9 Reading items, and five grade 10 Reading items. Across all grades, the Reading items that exhibited DIF tended to favor Asian students (15 items) and to disfavor males (four items). One item disfavored African American students, no item disfavored Hispanic students, and no item disfavored females. Two items favored African American students, three items favored Hispanic students, and three items favored females.

On the Writing assessments, DIF was observed in all grades 3 through 10. Across all grades, no item disfavored female students, 10 disfavored males, eight disfavored Asian students, and one item disfavored Hispanic students. In addition, two Writing items favored Hispanic students, six items favored Asian students, one item favored African American students, no items favored males, and six items favored females.

On the Mathematics assessments, DIF was observed in grades 4, 5, 7, 9 and 10. No DIF was observed in grades 3, 6, and 8. Across the grades showing DIF, three items disfavored Asians, two items disfavored African American students, one item disfavored Hispanic students, one item disfavored females, and two items disfavored males. In addition, one Mathematics item favored females, no item favored African American or Hispanic students, and one item favored Asian students.

On the Science assessments, items exhibited DIF in grades 8 and 10. Two items favored females, four items favored Asians, three items favored Hispanic students, and one item favored African American students. Two items disfavored males, no items disfavored African American students, Asian students, or Hispanic students.

Additional DIF analyses are presented in Tables 240 (Gender), 241 (Accommodations), 242 (Primary Disability State), 243 (Enrollment), 244 (Language Proficiency), 245 (Education Plan), and 246 (Focal Group: Immigrant, Migrant, Homeless).

Internal Factor Structure and Unidimensionality of the CSAP Assessment

Analyses of the internal structure of a test can indicate the extent to which the relationships among test items and components conform to the construct the test purports to measure. Educational assessments are usually designed to measure a single overall construct or domain (e.g., Reading achievement). CSAP test items are calibrated using unidimensional IRT models, which posits the presence of an essentially unidimensional construct underlying a group of test items and components. Unless tests are designed to have a complex internal structure, a measure of item homogeneity is relevant to validity. The internal consistency coefficient is a measure of item homogeneity. In order for a group of items to be homogeneous, they must measure the same construct (construct validity) or represent the same content domain (content validity).

To assess the overall factor structure of the CSAP assessments, exploratory factor analyses were conducted for each content and grade. Polychoric correlations were obtained, and a principal components factor analysis was conducted. The resulting Eigen values for each factor are an indication of the relative proportion of variance accounted for by each successive factor. Figures 64 through 94 contain plots of the Eigen values (part a) and proportions of variance (part b) for each factor identified in these analyses. All CSAP tests (English) demonstrated a strong single factor, accounting for 25%–56% of the overall variance, providing evidence that the items in each test are measuring a single construct. The variance accounted for by the single factor for grades 3 and 4 Spanish Reading and Writing tests was slightly lower range, from 19% to 32%.

IRT Model to Data Fit as an Evidence of Test Score Validity

When IRT models are used to calibrate test items and to report student scores, demonstrating item fit is also relevant to construct validity. That is, the extent to which test items function as the IRT model in use prescribes is relevant to the validation of test scores. As part of the scaling process, all CSAP items were examined closely with respect to classical (i.e., *p*-value and item-to-total score correlation) and IRT (Q1) fit indices. Items judged to be poorly fit by the model were visually inspected to decide whether the misfit was substantive in origin or from irrelevant sources such as extreme expectations that often accompany extremely easy or hard items. Very few items (4%) on the 2008 assessments were flagged for poor model fit, indicating that the test items were adequately scaled by the unidimensional IRT models and the resulting scores are interpretable and valid. IRT fit statistics are discussed in greater detail in Part 6 of this Technical Report. Summaries of the IRT fit statistics are presented in Tables 83 through 144.

Divergent (Discriminant) Validity

Measures of different constructs should not be highly correlated with each other. Divergent validity is a subtype of construct validity that can be estimated by the extent to which measures of constructs that theoretically should not be related to each other are, in fact, observed as not related to each other. Typically, correlation coefficients among measures are examined in support of divergent validity.

To assess the divergent validity of CSAP tests, scale scores were obtained and correlated for students who took various CSAP content area tests in 2008. Tables 247 and 248 show the intercorrelation among content areas (scale scores and percentile ranks) in different content areas by grade level. The correlation coefficients among scale scores ranged from 0.722 (between Reading and Mathematics in grade 3) to 0.864 (between Reading and Writing in grade 9). The correlation coefficients suggest that individual student scores for Reading, Mathematics, Writing, and Science are moderately to highly related. These coefficients are not so low as to call into question whether these tests are tapping into achievement constructs, and not so high as to arouse suspicion that the intended constructs are not distinct.

It is worth noting that the correlation coefficients between Reading and Writing were generally higher than those between Mathematics and Reading and between Mathematics and Writing. It is also interesting to note that Science is correlated with Reading and Mathematics to a similar degree; however, the correlation between Science and Writing was relatively lower. A similar pattern of correlations has been observed in *TerraNova* (CTB/McGraw-Hill, 2001d).

Additional evidence of divergent validity can be obtained by evaluating the correlations of test scores with extraneous demographic variables. Correlations were computed between total scale scores and age, gender, and ethnic group. Overall, these correlations were found to be somewhat small, ranging from nearly −0.38 to 0.09 (see Table 249). The fact that these correlations are generally greater than zero in absolute term can be attributed to differences in the overall ability of the various groups.

Predictive Validity

Predictive validity is a type of criterion validity that refers to the degree to which test scores predict criterion measurements that will be made at some point in the future (Crocker & Algina, 1986). In the context of annual assessment of student proficiency in a content area, the extent to which test scores in a year are predictive of those in the subsequent year can provide evidence for predictive validity. Colorado Model Content Standards in Mathematics, Reading, and Writing are designed to be incremental and progressive from lower to higher grade level, which is the basis for vertical scaling and measuring student growth

across years on a common scale. Table 250 shows predictive validity coefficients measured as the correlation between test scores for two adjacent years (2007 and 2008) on the basis of group of students matched on student ID data.

Factors affecting the measures of predictive validity include the time interval between assessments, reliability of assessments, differential individual and school effects, and so on. The correlation coefficients reported in Table 250 indicate strong predictability of test scores between two adjacent years. The validity coefficients (corrected for attenuation) are closer to or greater than 0.85 for all grades and content areas indicating a high degree of determination of performance from one year to next. The lowest validity coefficients are associated with grades 3 and 4. This may be attributed to the relatively short test length at grade 3 and differences in content standards between the grades.

Classification Consistency and Accuracy

One of the cornerstones of the No Child Left Behind Act of 2001 (2002) is the measurement of adequate yearly progress (AYP) with respect to the percentage of students at or above performance standards set by states. Because of this heavy emphasis on the classification of student performance, a psychometric property of particular interest is how consistently and accurately assessment instruments can classify students into performance categories.

Classification consistency is defined conceptually as the extent to which the performance classifications of students agree given two independent administrations of the same test or two parallel test forms. That is, if students are tested twice on the same test or on two parallel tests, what is the likelihood of classifying the students into the same performance categories? It is, however, virtually impractical to obtain data from repeated administrations of the same or parallel forms because of cost, testing burden, and effects of student memory or practice. Therefore, a common practice is to estimate classification consistency from a single administration of a test.

When a method to estimate decision consistency is applied, a contingency table of $(H + 1) \times (H + 1)$ is constructed, where H is the number of cut scores. For example, with three cut scores, a 4×4 contingency table can be built as follows:

Contingency Table With Three Cut Scores

	Level 1	Level 2	Level 3	Level 4	Sum
Level 1	P_{11}	P_{21}	P_{31}	P_{41}	$P_{\cdot 1}$
Level 2	P_{12}	P_{22}	P_{32}	P_{42}	$P_{\cdot 2}$
Level 3	P_{13}	P_{23}	P_{33}	P_{43}	$P_{\cdot 3}$
Level 4	P_{14}	P_{24}	P_{34}	P_{44}	$P_{\cdot 4}$
Sum	$P_{1\cdot}$	$P_{2\cdot}$	$P_{3\cdot}$	$P_{4\cdot}$	1.0

It is common to report two indices of classification consistency: the classification agreement *P* and coefficient kappa. Hambleton and Novick (1973) proposed *P* as a measure of classification consistency, where *P* is defined as the sum of diagonal values of the contingency table:

$$P = P_{11} + P_{22} + P_{33} + P_{44}$$

To reflect statistical chance agreement, Swaminathan, Hambleton, and Algina (1974) suggest using Cohen's kappa (Cohen, 1960):

$$\text{kappa} = \frac{P - P_c}{1 - P_c},$$

where P_c is the chance probability of a consistent classification under two completely random assignments. This probability P_c is the sum of the probabilities obtained by multiplying the marginal probability of the first administration and the corresponding marginal probability of the second administration:

$$P_c = (P_{1.} \times P_{.1}) + (P_{2.} \times P_{.2}) + (P_{3.} \times P_{.3}) + (P_{4.} \times P_{.4}).$$

Classification accuracy is defined as the extent to which the actual classifications of test takers agree with those that would be made on the basis of their true scores (Livingston & Lewis, 1995). That is, classification consistency refers to the agreement between two observed scores, while classification accuracy refers to the agreement between observed and true scores. Since true scores are unobservable, a psychometric model is typically used to estimate them on the basis of observed scores and the parameters of the model being used.

Classification Consistency and Accuracy When Pattern Scoring Is Used

A variety of IRT scoring procedures are available for estimating student proficiency scores. Of the most popular score estimation techniques are item-pattern (IP) scoring and number-correct (NC) scoring under the IRT framework. NC scoring considers only how many items a student answered correctly (or the sum of item scores) in determining his or her score. In contrast, the IP scoring method takes into account not only a student's total raw score, but also which items he or she got right.

Several methods have been proposed to measure classification consistency and accuracy on the basis of number-correct (summed) scores. However, few studies have proposed methods for IP scoring. Kolen and Kim (2004) developed a method to estimate classification consistency and accuracy when IP scoring is used. The following describes the Kolen–Kim method:

Step 1: Obtain ability distribution weight ($\hat{g}(\theta)$) at each quadrature (θ_j) point *j*.

Step 2: At each quadrature point θ_j, generate two sets of item responses using the item parameters from a test form, assuming that the same test form was administered twice to examinees with the true ability θ_j.

θ_j → (1, 1, 0, 0, …: Item response from the first administration) → $\hat{\theta}_{j1}$

θ_j → (0, 1, 1, 0, …: Item response from the second administration) → $\hat{\theta}_{j2}$

If two parallel (or alternative) forms were used, the two response patterns can be generated on the basis of the item parameters from the two forms. Estimate $\hat{\theta}_{j1}$ and $\hat{\theta}_{j2}$ for the two sets of item responses.

Step 3: Construct a classification matrix (as shown in the example below) at each quadrature point (θ_j). Determine the joint probability for the cells in the example below using the two ability estimates obtained from Step 2.

Classification Table for One Cut Point (C_1)[5]

		First administration or Form 1	
		$\hat{\theta}_{ji} \geq C_1$	$\hat{\theta}_{j1} < C_1$
Second administration, or Form 2	$\hat{\theta}_{j2} \geq C_1$		
	$\hat{\theta}_{j2} < C_1$		

Step 4: Repeat Steps 2 and 3 *r* times and compute average values over *r* replications. *r* should be a large number, for example, 500, to obtain stable results.

Step 5: Multiply the distribution weight ($\hat{g}(\theta)$) by the average values obtained in Step 4 for each quadrature point, and sum the results across all quadrature points. From these results a final contingency table can be constructed and classification consistency indices, such as kappa, can be computed. In addition, because examinees' abilities are estimated at each quadrature point, this quadrature point can be considered the true score. Therefore, classification accuracy may be computed using both examinees' estimated abilities (observed scores) and quadrature point (true score).

Table 251 (composed of two tables) includes the classification consistency and accuracy measures for CSAP grade 3 Mathematics. The first table is a

[5] This table is constructed for each quadrature point and replication. One, and only one, cell will have a value of 1, and zeros elsewhere.

contingency table with all three cut scores prepared using the Kolen–Kim procedure. The rows represent the first administration of an assessment, and the columns represent the second administration of the same assessment to the same students. As mentioned above, in the procedure by Kolen and Kim, the score distributions for the first administration and the second administration are estimated using simulation. So, the value in each cell represents the probability of belonging to certain performance levels in two hypothetical administrations. For example, 0.0629 represents the probability of belonging to "Unsatisfactory" in both first and second administrations. The 0.0002 represents the probability of belonging to "Unsatisfactory" in the first administration and "Proficient" in the second administration. "Sum" is obtained simply by adding the four row values or the four column values. The "Observed Score Dist" row shows the distribution of real data belonging to each performance level. In general, it is expected that the sum values and the distribution of observed scores from real data agree.

The second table shows indices for classification consistency and classification accuracy. Each index was described above. Because there are three cut scores for CSAP, four performance levels exist. The values in "All cuts" were obtained by applying all three cuts simultaneously during analysis. From Table 251 for grade 3 Mathematics, classification agreement (*P*) is 0.7437, chance probability is 0.3009, kappa is 0.6334, and classification accuracy is 0.8155, when all three cuts were used for computation. The values for cut 1 were obtained by applying only the first cut score, 335. Therefore, there are two levels whenever only one cut is applied. It is clear that the values for *P*, decision accuracy, obtained with all three cuts are smaller than those obtained with only one cut. This explanation is the same for tables for all other grade levels and content areas (Tables 251 through 281).

Part 9: Special Study

Part 9 presents results from a special study that investigated the reasons for unstable scale scores in the Extended Writing subcontent area in Writing tests, which were composed of a small number of constructed-response items.

Writing Trend Study

CSAP incorporates the philosophy of multiple measures of a construct. All CSAP assessments are composed of multiple-choice item types. CSAP Writing assessments consist of a mixture of multiple-choice (MC) and constructed-response (CR) items measuring the total writing proficiency and skills at various content standards and subcontent areas (e.g., Write Using Conventions, Paragraph Writing, Extend Writing, and Grammar and Usage). CR items in CSAP take different forms and solicit varying response lengths. Compared to other statewide Writing assessments, for example, single-prompt extended writing, the CSAP Writing assessment taps into a variety of writing skills using various item formats.

In addition to providing an overall measure of writing ability, CSAP provides subscores at various content standards and subcontent areas to provide more diagnostic information on the examinee's writing ability. The subscores are derived on the basis of the examinee's performance on subsets of items, typically composed of a mixture of MC and CR items of various lengths. One exception is the Extended Writing subcontent area, which is measured only by a small number of CR items. It has been observed historically that the score in the Extended Writing subcontent area is unstable across administrations. That is, the historical trends on this subcontent area have fluctuated more radically than the overall construct, the other content standards and subcontent areas. Furthermore, the trends on the subcontent area did not coincide with those on the overall test or other subcontent areas.

At the request of the CSAP Technical Advisory Committee (TAC), a study in English Writing was conducted to explore the unstable trends of the Extended Writing subcontent area in 2008 also. Grade 3 Writing does not include the Extended Writing subcontent area so the study was conducted on grades 4 through 10. To conduct this study, the Paragraph Writing (SA 5) and Extended Writing (SA 6) subcontent areas were combined. That is, a new subcontent area was formed by collapsing the two subcontent areas and the items contributing to them. Scores for this new combined SA 5/SA 6 subcontent area were generated for the past seven years (2002 through 2008). The results in mean and median scale scores are presented in Table 282. Median scores were examined because subcontent scores tend to be affected unduly by extreme scores. Median scale scores are also presented in Figures 95 through 101.

The Extend Writing/Paragraph Writing combined is more stable with the total and other subcontent median scale score trends. Because of the increased number of items in the combined subcontent area, the stability of the subcontent is improved considerably. Furthermore, the fluctuation in difficulty for the Extended Writing subcontent area has been stabilized.

Appendix A: Depth of Knowledge (DOK) Levels

Table A-1. Percentage of Items in Depth of Knowledge Levels

Tests	Grade	DOK Level								Number of Items in Test
		1		2		3		4		
		MC	CR	MC	CR	MC	CR	MC	CR	
Reading	3	2.5%	0.0%	77.5%	20.0%	0.0%	0.0%	0.0%	0.0%	40
	4	27.1%	2.9%	45.7%	11.4%	5.7%	5.7%	1.4%	0.0%	70
	5	28.6%	4.3%	40.0%	10.0%	11.4%	5.7%	0.0%	0.0%	70
	6	37.1%	7.1%	32.9%	7.1%	10.0%	5.7%	0.0%	0.0%	70
	7	20.3%	2.9%	36.2%	13.0%	23.2%	4.4%	0.0%	0.0%	69
	8	12.9%	0.0%	44.3%	12.9%	22.9%	7.1%	0.0%	0.0%	70
	9	11.6%	2.9%	36.2%	10.1%	31.9%	7.3%	0.0%	0.0%	69
	10	22.4%	1.5%	37.3%	3.0%	19.4%	16.4%	0.0%	0.0%	67
Writing	3	32.1%	11.3%	9.4%	7.6%	24.5%	15.1%	0.0%	0.0%	53
	4	64.2%	11.3%	11.3%	5.7%	0.0%	7.6%	0.0%	0.0%	53
	5	56.6%	11.3%	18.9%	3.8%	0.0%	9.4%	0.0%	0.0%	53
	6	47.2%	0.0%	24.5%	5.7%	3.8%	18.9%	0.0%	0.0%	53
	7	45.3%	11.3%	30.2%	1.9%	0.0%	11.3%	0.0%	0.0%	53
	8	50.9%	11.3%	5.7%	0.0%	18.9%	13.2%	0.0%	0.0%	53
	9	34.0%	0.0%	24.5%	11.3%	17.0%	13.2%	0.0%	0.0%	53
	10	20.8%	0.0%	26.4%	11.3%	28.3%	13.2%	0.0%	0.0%	53
Mathematics	3	48.7%	0.0%	28.2%	10.3%	5.1%	7.7%	0.0%	0.0%	39
	4	32.4%	0.0%	45.6%	8.8%	1.5%	11.8%	0.0%	0.0%	68
	5	39.1%	0.0%	34.8%	8.7%	4.4%	13.0%	0.0%	0.0%	69
	6	26.7%	1.7%	48.3%	13.3%	0.0%	10.0%	0.0%	0.0%	60
	7	21.7%	0.0%	50.0%	11.7%	3.3%	13.3%	0.0%	0.0%	60
	8	25.0%	0.0%	46.7%	8.3%	3.3%	16.7%	0.0%	0.0%	60
	9	15.3%	0.0%	52.5%	17.0%	6.8%	8.5%	0.0%	0.0%	59
	10	23.3%	1.7%	41.7%	16.7%	10.0%	6.7%	0.0%	0.0%	60
Science	5	43.8%	9.6%	30.1%	11.0%	1.4%	4.1%	0.0%	0.0%	73
	8	31.3%	8.4%	34.9%	7.2%	6.0%	12.1%	0.0%	0.0%	83
	10	19.8%	4.9%	51.9%	18.5%	0.0%	4.9%	0.0%	0.0%	81
Spanish Reading	3	0.0%	0.0%	80.0%	20.0%	0.0%	0.0%	0.0%	0.0%	40
	4	34.3%	1.4%	38.6%	17.1%	7.1%	1.4%	0.0%	0.0%	70
Spanish Writing	3	66.0%	11.3%	0.0%	22.6%	0.0%	0.0%	0.0%	0.0%	53
	4	64.2%	11.3%	11.3%	11.3%	0.0%	1.9%	0.0%	0.0%	53

Table A-2. Percentage of Score Points in Depth of Knowledge Levels

Tests	Grade	DOK Level 1 MC	DOK Level 1 CR	DOK Level 2 MC	DOK Level 2 CR	DOK Level 3 MC	DOK Level 3 CR	DOK Level 4 MC	DOK Level 4 CR	Total Number of Points in Test
Reading	3	1.9%	0.0%	59.6%	38.5%	0.0%	0.0%	0.0%	0.0%	52
	4	20.9%	4.4%	35.2%	23.1%	4.4%	11.0%	1.1%	0.0%	91
	5	22.0%	7.7%	30.8%	18.7%	8.8%	12.1%	0.0%	0.0%	91
	6	28.6%	14.3%	25.3%	15.4%	7.7%	8.8%	0.0%	0.0%	91
	7	15.6%	5.6%	27.8%	25.6%	17.8%	7.8%	0.0%	0.0%	90
	8	9.9%	0.0%	34.1%	24.2%	17.6%	14.3%	0.0%	0.0%	91
	9	8.5%	6.4%	26.6%	19.1%	23.4%	16.0%	0.0%	0.0%	94
	10	16.3%	3.3%	27.2%	5.4%	14.1%	33.7%	0.0%	0.0%	92
Writing	3	30.4%	10.7%	8.9%	8.9%	23.2%	17.9%	0.0%	0.0%	56
	4	49.3%	8.7%	8.7%	17.4%	0.0%	15.9%	0.0%	0.0%	69
	5	43.5%	8.7%	14.5%	11.6%	0.0%	21.7%	0.0%	0.0%	69
	6	36.2%	0.0%	18.8%	4.3%	2.9%	37.7%	0.0%	0.0%	69
	7	34.8%	8.7%	23.2%	1.4%	0.0%	31.9%	0.0%	0.0%	69
	8	39.1%	8.7%	4.3%	0.0%	14.5%	33.3%	0.0%	0.0%	69
	9	26.1%	0.0%	18.8%	8.7%	13.0%	33.3%	0.0%	0.0%	69
	10	15.9%	0.0%	20.3%	8.7%	21.7%	33.3%	0.0%	0.0%	69
Mathematics	3	39.6%	0.0%	22.9%	18.8%	4.2%	14.6%	0.0%	0.0%	48
	4	23.4%	0.0%	33.0%	16.0%	1.1%	26.6%	0.0%	0.0%	94
	5	28.1%	0.0%	25.0%	13.5%	3.1%	30.2%	0.0%	0.0%	96
	6	18.4%	2.3%	33.3%	23.0%	0.0%	23.0%	0.0%	0.0%	87
	7	14.9%	0.0%	34.5%	18.4%	2.3%	29.9%	0.0%	0.0%	87
	8	17.2%	0.0%	32.2%	12.6%	2.3%	35.6%	0.0%	0.0%	87
	9	10.5%	0.0%	36.0%	30.2%	4.7%	18.6%	0.0%	0.0%	86
	10	16.1%	2.3%	28.7%	32.2%	6.9%	13.8%	0.0%	0.0%	87
Science	5	37.2%	12.8%	25.6%	16.3%	1.2%	7.0%	0.0%	0.0%	86
	8	26.0%	10.0%	29.0%	10.0%	5.0%	20.0%	0.0%	0.0%	100
	10	16.3%	6.1%	42.9%	26.5%	0.0%	8.2%	0.0%	0.0%	98
Spanish Reading	3	0.0%	0.0%	61.5%	38.5%	0.0%	0.0%	0.0%	0.0%	52
	4	26.4%	2.2%	29.7%	34.1%	5.5%	2.2%	0.0%	0.0%	91
Spanish Writing	3	62.5%	10.7%	0.0%	26.8%	0.0%	0.0%	0.0%	0.0%	56
	4	49.3%	8.7%	8.7%	31.9%	0.0%	1.4%	0.0%	0.0%	69

References

AERA, APA, and NCME (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education) (1999). *Standards for educational and psychological testing.* Washington, DC: APA.

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.

Angoff, W. H. (1972). A technique for the investigation of cultural differences. Paper presented at the annual meeting of the American Psychological Association, Honolulu.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29–51.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443–459.

Brennan, R. L., & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses and alternatives. *Educational and Psychological Measurement*, *41*, 687–699.

Burket, G. R. (1993). PARDUX [computer program], Version 1.7.

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased items*. Newbury Park, CA: Sage.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.

Colorado Department of Education. (2008). *Colorado accommodations manual: Selecting and using accommodations.* First Edition, August 2007.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando, FL: Harcourt Brace Jovanovich College Publisher.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

CTB/McGraw-Hill (2008). *Technical Report for the Cut Score Review 2008 for Grades 5, 8, and 10 Science. Monterey CA: Author*

CTB/McGraw-Hill (2001d). TerraNova Technical Report. Monterey, CA: Author

Divgi, D. R. (1985). A minimum chi-square method for developing a common metric in item response theory. *Applied Psychological Measurement*, *9(4)*, 413–415.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel–Haenszel and standardization. In P. W. Holland and H. Wainer (Eds.), *Differential item function* (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum.

Green, D. R. (1975). Procedures for assessing bias in achievement tests. Presented at the National Institute of Education Conference on Test Bias, Annapolis, MD.

Hambleton, R. K., & Novick. M. R. (1973). Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement*, *10*, 159–196.

Kolen, M., & Kim, D. (2004). Personal Communication.

Lewis, D. M., Mitzel, H. C., & Green, D. R. (June 1996). Standard setting: A Bookmark approach. In D. R. Green (Chair), *IRT-based standard-setting procedures utilizing behavioral anchoring*. Symposium conducted at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix, AZ.

Linn, R. L., & Harnisch, D. L. (1981). Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement*, *18(2)*, 109–118.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). Item bias in a test of reading comprehension. *Applied Psychological Measurement*, *5*, 159–173.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, *32*, 179–197.

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika*, *39*, 247–264.

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

Lord, F. M., & Novick M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

MacMillan/McGraw-Hill (1993a). *Guidelines for bias-free publishing*.

MacMillan/McGraw-Hill (1993b). *Reflecting diversity: Multicultural guidelines for educational publishing professionals*.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–176.

No Child Left Behind Act of 2001 (2002). Pub. I, No. 107–110, 115 Stat 1425.

Sinharay, S. & Holland, P. W. (2007). Is it necessary to make anchor tests mini-versions of the tests being equated or can some restrictions be relaxed? *Journal of Educational Measurement*, *44(3)*, 249–275.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*, 201–210.

Stone, C. A., Ankenmann, R. D., Lane, S., & Liu, M. (April 1993). Scaling Quasar's performance assessments. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Swaminathan, H., Hambleton, R. K., & Algina, J. (1974). Reliability of criterion referenced tests: A decision theoretic formulation. *Journal of Educational Measurement*, *11*, 263–268.

Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika*, *47*, 175–186.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, *5*, 245–262.

Yen, W. M. (1984). Obtaining maximum likelihood trait estimates from number-correct scores for the three-parameter logistic model. *Journal of Educational Measurement*, *21*, 93–111.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item independence. *Journal of Educational Measurement*, *30*, 187–213.

Yen, W. M., & Candell, G. L. (1991). Increasing score reliability with item-pattern scoring: An empirical study in five score metrics. *Applied Measurement in Education*, *4*, 209–228.

