Transitional Colorado Assessment Program

Technical Report 2012

Submitted to the

Colorado Department of Education

October 2012

CTB McGraw-Hill

TABLE OF CONTENTS

Overview

This report presents the results of the statewide spring 2012 administration of the Transitional Colorado Assessment Program (TCAP). In the spring of 2012 students in grades 3 through 10 were assessed in Reading, Writing, and Mathematics, and students in grades 5, 8, and 10 were also assessed in 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. Part 1 provides an overview of the TCAP 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 the scoring and scaling design (including descriptions of scoring and scaling procedures for the total test and for individual content standards and subcontent areas), as well as interrater reliability and rater severity/leniency. Part 5 includes detailed item analysis results, including item-to-total score correlations, *p*-values, and omit rates. Part 6 describes the calibration and equating results, including an overview of the Item Response Theory (IRT) models, model-to-data fit, item independence, and equating procedures. Part 7 presents scale score summary statistics and correlations among content standards and subcontent areas. 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 the Writing subscale trends for paragraph and extended writing.

Part 1: Standards

The TCAP 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. Some items in TCAP are mapped to a subcontent area, whereas all items are mapped to one, and only one, Colorado Model Content Standard.

The 2012 TCAP assessment represents a transition from the Colorado Model Content Standards (CMCS) to the Colorado Academic Standards (CAS). In order to help facilitate this transition, new items developed for the 2012 TCAP were written to align to both the CMCS and CAS. To the same end, existing items from the pool that align to both standards were selected to fill the blueprint.

The various Colorado Model Content Standards and subcontent areas are listed below for each content area. Table 1 provides an overview of which content standards and subcontent areas are 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 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) Subcontent Area 1 (Varies by Grade).

- Number and Operation Sense (Grades 4 and 5) – Students demonstrate meanings for whole numbers, commonly used fractions, decimals, and the four basic arithmetic operations through the use of drawings, and decomposing and composing numbers; and identify factors, multiples, and prime/composite numbers.

- Number and Operation Sense (Grade 6) – 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 (Grade 7) – Students demonstrate understanding of the concept of equivalency as related to fractions, decimals, and percents.

- Linear Pattern Representation (Grade 8) – 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.

- Multiple Representations of Linear/Nonlinear Functions (Grade 9) – 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.

- Multiple Representations of Functions (Grade 10) – 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.

2) Subcontent Area 2 (Varies by Grade).

- Patterns (Grade 4) – Students reproduce, extend, create, and describe geometric and numeric patterns as problem-solving tools.

- Patterns (Grade 5) – Students represent, describe, and analyze geometric and numeric patterns using tables, graphs, and verbal rules as problem-solving tools.

- Patterns (Grade 6) – Students represent, describe, and analyze geometric and numeric patterns using tables, words, concrete objects, and pictures in problem-solving situations.

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

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

- Proportional Thinking (Grade 9) – Students apply the concepts of ratio and proportion in problem-solving situations.

- Probability and Counting Techniques (Grade 10) – 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.

3) Subcontent Area 3 (Varies by Grade).

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

- Data Display (Grade 5) – 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.

- Geometry (Grade 6) – Students reason informally about the properties of two-dimensional figures and solve problems involving area and perimeter.

- Geometry (Grade 8) – Students describe, analyze, and reason informally about the properties of two- and three-dimensional figures to solve problems.

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 know 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 – Students design, plan, and conduct a variety of investigations; understand and apply scientific questions, hypotheses, variables, and experimental design.

2) Results and Data Analysis – Students select and use appropriate technology; organize, analyze, interpret, and predict from scientific data in order to communicate the results of investigations.

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

4) Chemistry Concepts – Students understand the properties, composition, structure, and changes of matter.

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

6) Geology and Astronomy – Students understand 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 Learners and students with disabilities. Part 2 provides an overview of the TCAP test design and the development of student assessments that assist stakeholders in making informed educational decisions. Specifically, it describes the TCAP test development activities for the 2012 assessments 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).

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 TCAP 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. CTB's content developers studied the Colorado Academic Standards and developed items that aligned to them as well. Several sources contributed to the 2012 TCAP items. CTB/McGraw-Hill's extensive pool of previously 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 Model Content 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 TCAP 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 and alignment to the Colorado Model Content Standards 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 staff. 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 TCAP 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 process are provided later in Part 6 of this report. The assessments that are reported on vertical scales (English Reading, English Writing, and Mathematics) also had items in common between adjacent grades. In grades 3 and 4, the 2012 Spanish Reading and Writing test forms were the same forms that had been administered in previous administrations.

Test Configuration

Tables 2 through 6 provide information regarding the configuration of the TCAP 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; and Scientific Investigation (CS 1) and the Nature of Science (CS 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 is less than the total number of points for the test.

Tables 7 and 8 provide the Depth of Knowledge (DOK) level distribution for the 2012 TCAP assessments. DOK distribution will be articulated in the blueprint for the 2012 TCAP assessments.

TCAP Content Validity and Alignment Review

The items that appeared in 2012 TCAP tests were carefully reviewed and discussed in May 2011 by Content Validity and Alignment Review committees to ensure content validity, accurate alignment to content standards, and the quality and appropriateness of the items. Included was a 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 the following:

- alignment of items to assessment objectives under the Colorado Model Content Standards, the Assessment Frameworks, and Depth of Knowledge
- accuracy and 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

- reviews resulted in an independent alignment recommendation by each reviewer
- thorough discussion of appropriate alignment occurred following the independent reviews; and
- thorough documentation of alignment findings was captured.

Processes for bias and sensitivity review were designed to ensure that

- items were neither advantageous nor disadvantageous to a specific group of students
- items did not stereotype specific groups
- items did not promote personal, moral, or religious values or viewpoints; and
- students' achievement on a given test item would be 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 Transitional Colorado Assessment Program

As indicated in the previous section, one purpose of the TCAP content review was the application of Universal Design in test assembly. The TCAP measures what students know and are able to do as defined in the Colorado Model Content Standards. Assessments must ensure comprehensible access to this content.

CDE's and CTB's content experts revised the item pool and removed unnecessary verbiage from the 2012 TCAP 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 2012 were authored using these principles. Items previously developed and administered prior to 2012 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 TCAP
    - Increases validity to the measurement of the construct
    - Increases the accuracy of the inferences made from the resulting data
  - Plain Language in TCAP 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 2012 Assessments

In order to equate current tests to base year scale, a set of previously administered multiple-choice anchor items was selected for each of the 2012 assessments in Reading, Writing, Mathematics, and Science. 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 process are provided in Part 6.

The following criteria were followed to select anchor items in Reading, Writing, Mathematics, and Science:[1]

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 2012 TCAP tests included 17 to 32 anchor items for each grade and content area. Only multiple-choice items were selected as anchors.[2] 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.

Relative Item Position in a Form – Anchor items were placed in approximately 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 on the new form was as close as possible to its position on the form in which it appeared previously (in most cases within three positions). A minimum requirement was that they be placed in the same third of the form that they appeared in when 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 or session or in any locations where speededness effects may occur.

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

- Point biserials ≤ 0.18 for the correct answer

[1] The 2012 Spanish tests for grades 3 and 4 Reading and Writing were identical to the tests that were administered in previous years. For this reason, and because of the small number of students taking these tests, the Spanish tests were scored using the same item parameters that were used to score the tests in 2011.

[2] 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. Excluding CR items prevents equating error that could occur if raters varied in severity from year to year.

- Positive point biserials for the distractors
- *p*-value ≤ 0.25 or ≥ 0.90
- Omit rates ≥ 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 the 2012 tests 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 (TCCs) 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 as anchors 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 the TCC for all of the MC items on the test. 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.[3]

Items Flagged for Fit and DIF in Test Assembly

The items flagged for poor fit and DIF were avoided as much as possible when assembling the 2012 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 was to 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 9 displays the items with DIF and fit flags from previous administrations across all operational items for the 2012 assembled test forms. For the 1,665 operational English items with available statistics on the TCAP Reading, Writing, Mathematics, and Science assessments, 37 (2.2%) were flagged for marginal poor fit and 49 (2.9%) for DIF for gender and ethnic subgroups. Only five of these items were used as anchors in 2012. Of the 216 previously used Spanish

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

items, 52 (24.1%) were flagged in a previous administration for marginal poor fit and one was flagged for gender DIF.[4] Most of the flagged Spanish items were on the grade 4 Reading and Writing assessments (24 and 18 items, respectively), with very few items flagged in grade 3 (eight Reading items and three Writing items). As mentioned above, the poor fit was marginal for most items, and their inclusion in the tests was essential to meet the test blueprint for content standards.

[4] Because of the very small number of students taking the Spanish assessments each year, the same test forms are readministered each year, and it is not feasible to replace items or create new test forms. DIF statistics for the Spanish tests were not computed after 2008 because of the very small case counts.

Part 3: Administration

The Transitional Colorado Assessment Program (TCAP) is Colorado's large-scale standardized paper-and-pencil achievement test administered every year. In 2012, grade 3 Reading (English and Spanish) assessments were administered between February 13 and March 9. The rest of the English language tests, plus the grade 4 Spanish Reading and grades 3 and 4 Spanish Writing tests, were administered between March 5 and April 13. The purpose of the TCAP is to provide an annual measure of student performance relative to the Colorado Model Content Standards. All TCAP 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 for Reading/Writing and Mathematics and grades 5, 8, and 10 for Science were tested with a single form for each grade. The 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 = Approved nonstandard accommodation
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, with input from parents. The accommodations provided students equal opportunity to access information and 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 2012 test administration manual and the Colorado accommodations manual (Colorado Department of Education, 2012).

The following sections briefly describe the training conducted before the test administration to ensure proper handling of test materials, test administration, and the secure return of materials to the scoring center. That 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 2012 TCAP. 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 TCAP Test Proctor's manuals and the District Assessment Coordinator/School Assessment Coordinator (DAC/SAC) manuals.

The Test Proctor's manuals provided specific instruction on proper administration of the TCAP tests. The manuals provided detailed definitions of the TCAP 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 authorized materials required for testing. Prior to test administration, the TCAP 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 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 TCAP Test Proctor's manual and the TCAP DAC/SAC manual can be found at www.ctb.com/tcap.

Test Sections and Timing

Although the 2012 TCAP tests were administered independently, the TCAP Reading and Writing tests were combined in a single test book 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 TCAP 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 the TCAP 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 TCAP 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 the scaling design for the 2012 TCAP assessments.

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

In the TCAP 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 lowest obtainable scale score (LOSS) and highest obtainable scale score (HOSS) for each grade and content area is provided in Table 10. TCAP reports item pattern scores, and the HOSS increases from grade to grade to allow students' growth to be reflected in the subsequent administrations. The HOSS for grade 3 Reading is markedly different from those for grades 4 through 10 because grade 3 responses were scaled separately when the scale was set, and grade 3 scores were reported earlier than the rest of the grades. The same LOSS and HOSS are maintained over the years in all grades and content areas; however, the Science grades' LOSS/HOSS changed in 2008. 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 a 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 an 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% to 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.

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 2012 TCAP tests. CTB's handscoring supervisors reviewed approximately 300 to 1,000 student written responses to each of the CR items, drawn from the entire set of responses that were available at that time.[5] 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 handscoring supervisors also reviewed anchor papers for CR items that were used in previous years' versions of the tests. If items were revised or if 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 grades and contents to be reviewed with Colorado teachers at a live session in Denver, Colorado, in April 2012.

At the 2012 TCAP anchor paper review, CTB's supervisors distributed numbered packets containing the established scoring guide and the proposed and annotated anchors for all new items in 2012.

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 were 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

[5] While this process did limit the selection to those districts that delivered their materials in time to be included in the sample, there was no attempt to include, exclude, or weight the participation of any particular districts in the sample.

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 handscoring training for the 2012 TCAP at a CTB scoring facility.

Rater Reliability and Severity

The TCAP 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 the TCAP, reliability of handscoring items should be closely examined and documented. Through the ongoing process of training and research analyses, evidence of the reliability of handscoring was continuously gathered. Many training and monitoring techniques were used to ensure handscoring reliability and accuracy. Scoring guides were 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 2012 TCAP administration, approximately 5% of the constructed-response items scored were given a blind double read (i.e., were read by a second reader), and the resulting scores were documented and analyzed. In each case, the response was assigned to a second reader selected at random from all readers except the reader who had provided the first score. The second reader was not aware that this was a second read. For Spanish, approximately 15% of the constructed-response items were a blind double read. Evidence supporting interrater reliability of the TCAP 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.[6] Tables 11 through 16 show the rater reliability indices for all constructed-response items by content area. The results indicate that the kappa is reasonably high for all grades and content areas. Across all items in all grades and content areas, the percentage of exact plus adjacent agreement among raters ranges from 93.3% to 100%.

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 higher or lower than the scores that would be assigned by an ideal group of objective and unbiased raters. 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 17 shows the number of rater severity/leniency items used in the study by content area and grade. The following specifications describe the rater severity study in detail:

1) In 2012, a rater severity study was done using constructed-response items that were repeated from 2010 or 2011. Random samples of student responses were selected from the CSAP tests in which these repeated items were present: A random sample of approximately 1,000 students was selected for English Reading, English Writing, Mathematics, and Science assessments.

2) The samples of papers were distributed blindly to the 2012 raters during the second half of 2012 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 a previous test administration.

3) The scores from the rescore were then compared with the original scores given to the papers by the raters in 2010 or 2011.

Table 17 shows results of the rater severity study, including mean scores from the previous administration; mean scores from the 2012 administration; percent of the scores with exact, adjacent, and discrepant agreement; correlation; intraclass correlation; and weighted kappa.

[6] The observed proportion of agreement is computed by summing the proportion of agreement across cells; the chance agreement proportion is computed by summing the products of the column and row proportions.

Weighted kappa, which may be interpreted as the chance-corrected weighted proportional agreement, is reasonably high, with the highest values generally found in Mathematics items and the lowest in Writing. Weighted kappa ranged from 0.59 to 0.84 with a median value of 0.67 for Reading items; from 0.52 to 0.77 with a median value of 0.70 for Writing items; 0.75 to 0.97 with a median value of 0.91 for Mathematics items; and from 0.72 to 0.88 with a median value of 0.82 for Science items.

Scaling Design

Horizontal equating within each grade was used to place the 2012 forms on the scales that were established previously for English Reading, Writing, Mathematics, and Science, using Stocking and Lord's (1983) procedure. The vertical scale for English Reading, spanning grades 4 through 10, was established in 2001. The grade 3 reading assessment is sufficiently different from the reading assessments in the higher grades (it assesses only one content standard, whereas the other assessments assess multiple content standards) to warrant it to be treated separately. 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. Grades 5, 8, and 10 Science were placed on scale in 2008. Because of the nonincremental nature of the content standards and the gaps in grade levels, grades 5, 8, and 10 Science were not placed on a vertical scale.

Although the Spanish Reading and Writing tests in grades 3 and 4 are designed to measure a student's development over time, they were built from CTB's Supera assessments and are not on a vertical scale. Note that the customized versions of the grades 3 and 4 Reading and Writing assessments in Spanish were first administered in 1998.[7] The customized Spanish version that was first created in 1998 was repeated without modification through 2001. From 2002 through 2006, new Spanish forms were created by selecting psychometrically sound items from the existing item pool. The 2007 assessments were reprints of 2006, with the exception of a few select items. Because the numbers of students taking these tests are very small, the same Spanish test forms were readministered in 2008, 2009, 2010, 2011, and 2012, and the 2012 tests were scored using the same pre-equated item parameters that were used to score these tests in previous years.

With the exception of the Spanish Reading and Writing tests, each of the new 2012 TCAP tests contained a set of preselected multiple-choice items[8] from a previous administration for the same grade. These repeated multiple-choice items served as anchors in Stocking and Lord's (1983) equating procedure, which was used to place each test form on the previously established scale. By equating the 2012 TCAP tests across years within each grade, the unique metrics of the TCAP scales were maintained. The scaling and calibration methods are presented in Part 6 of this report.

[7] In 1997, Supera had been administered to TCAP students who were eligible to take a Spanish language version of the assessment.

[8] As noted previously in this report, the exclusion of constructed-response items from the anchor set eliminates the possibility of systematic equating error that might otherwise occur if there were shifts in rater severity across administrations.

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 who were absent when any items assessing a scale were administered, and those with multiple marks

- Students who had invalidation flags

- Students who had 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 and Science responses, where the entire test was presented orally (special code = 9) and where students received translated oral script (special code = B)

The descriptive statistics for scale scores were based on all valid cases. The frequency distributions by gender, ethnicity, and other subgroups are shown in Tables 18 through 22.

Tables 23 through 84 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 had 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 the treatment of omits in the computation of the operational scale scores. The item-to-total score correlations or point biserials (these terms may be used interchangeably when referring to multiple-choice items), the *p*-values, the percentages of omits, and the percentages at each score level (for the constructed-response items) are based on the analysis of responses of all students with reported total test scores.

As a part of the evaluation of 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 a reasonable amount of variability in students' responses to most multiple-choice items and a reasonable distribution of score points on most constructed-response items, indicating that these items provided information over the range of student ability. The classical item statistics for all grades and content areas are described briefly in the following sections.

Third Grade

Reading

Table 23 lists the results of the multiple-choice item analyses for the 2012 third-grade Reading assessment. The point biserials for all multiple-choice items ranged from 0.22 to 0.52, with a mean of 0.38. The *p*-values for the multiple-choice items ranged from 0.40 to 0.90, with a mean of 0.70.

Table 24 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items ranged from 0.39 to 0.56, with a mean of 0.48. The *p*-values ranged from 0.28 to 0.64, with a mean of 0.49. More than 50% of the students obtained the highest possible score points for one out of the eight constructed-response items. Scores were generally well distributed across the score points of these items.

The omit rates for the third-grade Reading assessment were generally small, ranging from 0.05% to 2.58% for the multiple-choice items (Table 23) and from 0.51% to 2.70% for the constructed-response items (Table 24).

Reading — Spanish

Table 25 lists the results of the multiple-choice item analyses for the Spanish version of the 2012 third-grade Reading assessment. The point biserials for all multiple-choice items ranged from 0.00 to 0.54, with a mean of 0.33. The *p*-values for the multiple-choice items ranged from 0.23 to 0.93, with a mean of 0.59.

Table 26 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items ranged from 0.42 to 0.62, with a mean of 0.51. The *p*-values ranged from 0.39 to 0.72, with a mean of 0.56. More than 50% of the students obtained the highest possible score points for two out of the eight constructed-response items. Scores were generally well distributed across the score points of the remaining items.

The omit rates for all but one of the multiple-choice items on the Spanish version of the third-grade Reading assessment were small. Omit rates for the multiple-choice items ranged from 0% to 9.25%, with only one item having an omit rate greater than 5% (Table 25). The omit rates for the constructed-response items were small, ranging from 0.76% to 3.20% (Table 26).

Writing

Table 27 lists the results of the multiple-choice item analyses for the 2012 third-grade Writing assessment. The point biserials for all multiple-choice items ranged from 0.21 to 0.49, with a mean of 0.38. The *p*-values for the multiple-choice items ranged from 0.36 to 0.95, with a mean of 0.75.

Table 28 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items ranged from 0.19 to 0.53, with a mean of 0.42. The *p*-values ranged from 0.61 to 0.98, with a mean of 0.78. More than 50% of the students obtained the highest possible score points for 15 of the 18 constructed-response items.

The omit rates for the third-grade Writing assessment were generally small, ranging from 0.03% to 8.79% for the multiple-choice items (Table 27) with only two items having an omit rate greater than 5%. Omit rates for the constructed-response items were small, ranging from 0.10% to 0.66% (Table 28).

Writing — Spanish

Table 29 lists the results of the multiple-choice item analyses for the Spanish version of the 2012 third-grade Writing assessment. The point biserials for all multiple-choice items ranged from 0.20 to 0.45, with a mean of 0.35. The *p*-values for the multiple-choice items ranged from 0.22 to 0.93, with a mean of 0.72.

Table 30 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items ranged from 0.34 to 0.58, with a mean of 0.47. The *p*-values ranged from 0.23 to 0.91, with a mean of 0.69. More than 50% of the students obtained the highest possible score points for 14 of the 18 constructed-response items. Scores were generally well distributed across the score points of the remaining items.

The omit rates for the Spanish version of the third-grade Writing assessment were small, ranging from 0% to 1.25% for the multiple-choice items (Table 29) and from 0.83% to 1.67% for the constructed-response items (Table 30).

Mathematics

Table 31 lists the results of the multiple-choice item analyses for the 2012 third-grade Mathematics assessment. The point biserials for all multiple-choice items ranged from 0.26 to 0.57, with a mean of 0.42. The *p*-values for the multiple-choice items ranged from 0.29 to 0.90, 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 ranged from 0.50 to 0.66, with a mean of 0.59. The *p*-values ranged from 0.36 to 0.74, with a mean of 0.58. More than 50% of the students obtained the highest possible score points for two of the eight constructed-response items. Scores were generally well distributed across the score points of the remaining items.

The omit rates for the third-grade Mathematics assessment were generally small, ranging from 0.14% to 2.06% for the multiple-choice items (Table 31) and from 0.16% to 0.75% for the constructed-response items (Table 32).

Fourth Grade

Reading

Table 33 lists the results of the multiple-choice item analyses for the 2012 fourth-grade Reading assessment. The point biserials for all multiple-choice items ranged from 0.18 to 0.56, with a mean of 0.40. The *p*-values for the multiple-choice items ranged from 0.20 to 0.94, with a mean of 0.69.

Table 34 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items ranged from 0.33 to 0.61, with a mean of 0.50. The *p*-values ranged from 0.36 to 0.78, with a mean of 0.55. More than 50% of the students obtained the highest possible score points for four of the 14 constructed-response items. Scores were generally well distributed across the score points of the remaining items.

The omit rates for the fourth-grade Reading assessment were small, ranging from 0.04% to 1.28% for the multiple-choice items (Table 33) and from 0.48% to 2.43% for the constructed-response items (Table 34).

Reading — Spanish

Table 35 lists the results of the multiple-choice item analyses for the Spanish version of the 2012 fourth-grade Reading assessment. The point biserials for all multiple-choice items ranged from 0.00 to 0.59, with a mean of 0.37. The *p*-values for the multiple-choice items ranged from 0.29 to 0.87, with a mean of 0.56.

Table 36 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items ranged from 0.16 to 0.76, with a mean of 0.54. The *p*-values ranged from 0.18 to 0.63, with a mean of 0.37. More than 50% of the students obtained the highest possible score points for one of the 14 constructed-response items. Scores were generally well distributed across the score points of the remaining items.

The omit rates for the Spanish version of the fourth-grade Reading assessment were generally small, ranging from 0% to 4.40% for the multiple-choice items (Table 35) and from 0% to 5.49% for the constructed-response items (Table 36), with only one item having an omit rate greater than 5%.

Writing

Table 37 lists the results of the multiple-choice item analyses for the 2011 fourth-grade Writing assessment. The point biserials for all multiple-choice items ranged from 0.23 to 0.50, with a mean of 0.37. The *p*-values for the multiple-choice items ranged from 0.35 to 0.94, with a mean of 0.72.

Table 38 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items ranged from 0.12 to 0.65, with a mean of 0.42. The *p*-values ranged from 0.42 to 0.98, with a mean of 0.69. More than 50% of the students obtained the highest possible score points for six of the 13 constructed-response items. Scores were generally well distributed across the score points of the remaining items.

The omit rates for the fourth-grade Writing assessment were small, ranging from 0.13% to 2.99% for the multiple-choice items (Table 37) and from 0% to 4.35% for the constructed-response items (Table 38).

Writing — Spanish

Table 39 lists the results of the multiple-choice item analyses for the Spanish version of the 2012 fourth-grade Writing assessment. The point biserials for all multiple-choice items ranged from -0.06 to 0.54, with a mean of 0.31. The *p*-values for the multiple-choice items ranged from 0.22 to 0.96, with a mean of 0.50.

Table 40 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items ranged from 0.06 to 0.59, with a mean of 0.38. The *p*-values ranged from 0.08 to 0.92, with a mean of 0.54. More than 50% of the students obtained the highest possible score points for five of the seven one-point constructed-response items. Scores were generally well distributed across the score points of the remaining items.

The omit rates for the Spanish version of the fourth-grade Writing assessment were generally small, ranging from 0% to 6.59% for the multiple-choice items (Table 39) and ranging from 0% to 6.59% for the constructed-response items (Table 40). One multiple-choice item and two constructed-response items had omit rates greater than 5%.

Mathematics

Table 41 lists the results of the multiple-choice item analyses for the 2012 fourth-grade Mathematics assessment. The point biserials for all multiple-choice items ranged from 0.22 to 0.58, with a mean of 0.42. The *p*-values for the multiple-choice items ranged from 0.27 to 0.96, with a mean of 0.74.

Table 42 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items ranged from 0.40 to 0.67, with a mean of 0.58. The *p*-values ranged from 0.24 to 0.91, with a mean of 0.62. More than 50% of the students obtained the highest possible score points for seven of the 15 constructed response items. The scores on the remaining constructed-response items were well distributed across the score points in those items.

The omit rates for the fourth-grade Mathematics assessment were generally small, ranging from 0.04% to 1.05% for the multiple-choice items (Table 41) and from 0.11% to 1.41% for the constructed-response items (Table 42).

Fifth Grade

Reading

Table 43 lists the results of the multiple-choice item analyses for the 2012 fifth-grade Reading assessment. The point biserials for all multiple-choice items ranged from 0.20 to 0.56, with a mean of 0.39. The *p*-values for the multiple-choice items ranged from 0.29 to 0.93, with a mean of 0.68.

Table 44 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items ranged from 0.33 to 0.62, with a mean of 0.50. The *p*-values ranged from 0.26 to 0.84, with a mean of 0.45. More than 50% of the students obtained the highest possible score points for one of the 14 constructed-response items. Scores were generally well distributed across the score points of the remaining items.

The omit rates for the fifth-grade Reading assessment ranged from 0.03% to 1.57% for the multiple-choice items (Table 43) and from 0.56% to 2.85% for the constructed-response items (Table 44).

Writing

Table 45 lists the results of the multiple-choice item analyses for the 2012 fifth-grade Writing assessment. The point biserials for all multiple-choice items ranged from 0.18 to 0.51, with a mean of 0.40. The *p*-values for the multiple-choice items ranged from 0.38 to 0.92, with a mean of 0.72.

Table 46 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items ranged from 0.11 to 0.62, with a mean of 0.43. The *p*-values ranged from 0.46 to 0.99, with a mean of 0.74. More than 50% of the students obtained the highest possible score points for seven of the 13 constructed-response items. Scores were generally well distributed across the score points of the remaining items.

The omit rates for the fifth-grade Writing assessment were small, ranging from 0.04% to 2.04% for the multiple-choice items (Table 45) and from 0% to 3.59% for the constructed-response items (Table 46).

Mathematics

Table 47 lists the results of the multiple-choice item analyses for the 2012 fifth-grade Mathematics assessment. The point biserials for all multiple-choice items ranged from 0.17 to 0.57, with a mean of 0.39. The *p*-values for the multiple-choice items ranged from 0.25 to 0.92, with a mean of 0.68.

Table 48 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items ranged from 0.52 to 0.70, with a mean of 0.61. The *p*-values ranged from 0.34 to 0.78, with a mean of 0.58. More than 50% of the students obtained the highest possible score points for one of the 15 constructed-response items. Scores were generally well distributed across the score points of the remaining scored items.

The omit rates for the fifth-grade Mathematics assessment were small, ranging from 0.03% to 1.45% for the multiple-choice items (Table 47) and from 0.08% to 0.51% for the constructed-response items (Table 48).

Science

Table 49 lists the results of the multiple-choice item analyses for the 2012 fifth-grade Science assessment. The point biserials for all multiple-choice items ranged from 0.12 to 0.58, with a mean of 0.33. The *p*-values for the multiple-choice items ranged from 0.26 to 0.87, with a mean of 0.62.

Table 50 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items ranged from 0.18 to 0.60, with a mean of 0.47. The *p*-values ranged from 0.22 to 0.80, with a mean of 0.54. More than 50% of the students obtained the highest possible score points for six of the 18 constructed-response items. Scores were generally well distributed across the score points of the remaining scored items.

The omit rates for the fifth-grade Science assessment were small, ranging from 0.02% to 1.42% for the multiple-choice items (Table 49) and from 0.10% to 1.19% for the constructed-response items (Table 50).

Sixth Grade

Reading

Table 51 lists the results of the multiple-choice item analyses for the 2012 sixth-grade Reading assessment. The point biserials for all multiple-choice items ranged from 0.16 to 0.57, with a mean of 0.40. The *p*-values for the multiple-choice items ranged from 0.32 to 0.94, with a mean of 0.72.

Table 52 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items ranged from 0.39 to 0.62, with a mean of 0.49. The *p*-values ranged from 0.32 to 0.76, with a mean of 0.47. More than 50% of the students obtained the highest possible score points for two of the 14 constructed-response items. Scores were generally well distributed across the score points of the remaining items.

The omit rates for the sixth-grade Reading assessment ranged from 0.04% to 2.71% for the multiple-choice items (Table 51) and from 0.61% to 4.92% for the constructed-response items (Table 52).

Writing

Table 53 lists the results of the multiple-choice item analyses for the 2012 sixth-grade Writing assessment. The point biserials for all multiple-choice items ranged from 0.22 to 0.52, with a mean of 0.39. The *p*-values for the multiple-choice items ranged from 0.42 to 0.92, 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 ranged from 0.11 to 0.63, with a mean of 0.41. The *p*-values ranged from 0.09 to 0.99, with a mean of 0.63. More than 50% of the students obtained the highest possible score points for five of the 13 constructed-response items (four of the seven one-point items and the only two-point item). Scores were generally well distributed across the score points of the remaining items.

The omit rates for the sixth-grade Writing assessment ranged from 0.08% to 2.35% for the multiple-choice items (Table 53) and from 0% to 4.93% for the constructed-response items (Table 54).

Mathematics

Table 55 lists the results of the multiple-choice item analyses for the 2012 sixth-grade Mathematics assessment. The point biserials ranged from 0.13 to 0.61, with a mean of 0.40. The *p*-values for the multiple-choice items ranged from 0.25 to 0.90, with a mean of 0.59.

Table 56 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items ranged from 0.40 to 0.72, with a mean of 0.58. The *p*-values ranged from 0.30 to 0.80, with a mean of 0.55. More than 50% of the students obtained the highest possible score points for four of the 15 constructed-response items. Scores were generally well distributed across the score points of the remaining items.

The omit rates for the sixth-grade Mathematics assessment ranged from 0.07% to 1.42% for the multiple-choice items (Table 55) and from 0.14% to 2.42% for the constructed-response items (Table 56).

Seventh Grade

Reading

Table 57 lists the results of the multiple-choice item analyses for the 2012 seventh-grade Reading assessment. The point biserials ranged from 0.10 to 0.50, with a mean of 0.37. The *p*-values for the multiple-choice items ranged from 0.34 to 0.91, with a mean of 0.71.

Table 58 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items ranged from 0.38 to 0.65, with a mean of 0.52. The *p*-values for the constructed-response items ranged from 0.33 to 0.80, with a mean of 0.56. More than 50% of the students obtained the highest possible score points for three of the 14 constructed-response items. Scores were generally well distributed across the score points of the remaining items.

The omit rates for the seventh-grade Reading assessment ranged from 0.05% to 4.02% for the multiple-choice items (Table 57) and from 0.85% to 3.25% for the constructed-response items (Table 58).

Writing

Table 59 lists the results of the multiple-choice item analyses for the 2012 seventh-grade Writing assessment. The point biserials for all multiple-choice items ranged from 0.19 to 0.57, with a mean of 0.41. The *p*-values for the multiple-choice items ranged from 0.27 to 0.91, with a mean of 0.74.

Table 60 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items ranged from 0.11 to 0.60, with a mean of 0.39. The *p*-values ranged from 0.21 to 0.99, with a mean of 0.68. More than 50% of the students obtained the highest possible score points for five of the 13 constructed-response items (four of the seven one-point items and the only two-point item). Scores were generally well distributed across the score points of the remaining items.

The omit rates for the seventh-grade Writing assessment ranged from 0.09% to 1.80% for the multiple-choice items (Table 59) and from 0% to 3.56% for the constructed-response items (Table 60).

Mathematics

Table 61 lists the results of the multiple-choice item analyses for the 2012 seventh-grade Mathematics assessment. The point biserials for all multiple-choice items ranged from 0.14 to 0.60, with a mean of 0.39. The *p*-values for the multiple-choice items ranged from 0.27 to 0.87, with a mean of 0.51.

Table 62 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items ranged from 0.31 to 0.73, with a mean of 0.56. The *p*-values ranged from 0.17 to 0.83, with a mean of 0.44. More than 50% of the students obtained the highest possible score points for two of the 15 constructed-response items. Scores were generally well distributed across the score points of the items.

The omit rates for the seventh-grade Mathematics assessment ranged from 0.06% to 2.19% for the multiple-choice items (Table 61) and from 0.12% to 2.78% for the constructed-response items (Table 62).

Eighth Grade

Reading

Table 63 lists the results of the multiple-choice item analyses for the 2012 eighth-grade Reading assessment. The point biserials for the multiple-choice items ranged from 0.12 to 0.50, with a mean of 0.35. The *p*-values for the scored multiple-choice items ranged from 0.37 to 0.90, 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 ranged from 0.35 to 0.61, with a mean of 0.49. The *p*-values ranged from 0.21 to 0.84, with a mean of 0.53. More than 50% of the students obtained the highest possible score points for two of the 14 constructed-response items. Scores were generally well distributed across the score points of the remaining items.

The omit rates for the eighth-grade Reading assessment ranged from 0.04% to 7.25% for multiple-choice items (Table 63) and from 1.07% to 10.51% for constructed-response items (Table 64). There were four multiple-choice items and two constructed-response items with an omit rate greater than 5%.

Writing

Table 65 lists the results of the multiple-choice item analyses for the 2012 eighth-grade Writing assessment. The point biserials for all multiple-choice items ranged from 0.10 to 0.52, with a mean of 0.38. The *p*-values for the multiple-choice items ranged from 0.38 to 0.94, with a mean of 0.71.

Table 66 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items ranged from 0.10 to 0.61, with a mean of 0.41. The *p*-values ranged from 0.60 to 0.99, with a mean of 0.74. More than 50% of the students obtained the highest possible score points for eight of the 13 constructed-response items (all seven of the one-point items and on the only two-point item in the test). Scores were generally well distributed across the score points of the remaining items.

The omit rates for the eighth-grade Writing assessment ranged from 0.08% to 1.05% for multiple-choice items (Table 65), and from 0% to 3.55% for constructed-response items (Table 66).

Mathematics

Table 67 lists the results of the multiple-choice item analyses for the 2012 eighth-grade Mathematics assessment. The point biserials for all multiple-choice items ranged from 0.11 to 0.60, with a mean of 0.39. The *p*-values for the multiple-choice items ranged from 0.14 to 0.82, with a mean of 0.50.

Table 68 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items ranged from 0.44 to 0.71, with a mean of 0.60. The *p*-values ranged from 0.19 to 0.77, with a mean of 0.41. Scores were generally well distributed across the score points of all of the constructed-response items.

The omit rates for the eighth-grade Mathematics ranged from 0.08% to 2.10% for the multiple-choice items (Table 67) and from 0.47% to 6.26% for the constructed-response items (Table 68). There was one constructed-response item with an omit rate greater than 5%.

Science

Table 69 lists the results of the multiple-choice item analyses for the 2012 eighth-grade Science assessment. The point biserials for all multiple-choice items ranged from 0.13 to 0.56, with a mean of 0.34. The *p*-values for the multiple-choice items ranged from 0.30 to 0.80, with a mean of 0.58.

Table 70 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items ranged from 0.20 to

0.69, with a mean of 0.45. The *p*-values ranged from 0.18 to 0.69, with a mean of 0.43. More than 50% of the students obtained the highest possible score points for three of the 23 constructed-response items. Scores were generally well distributed across the score points of the remaining constructed-response items.

The omit rates for the eighth-grade Science assessment ranged from 0.02% to 0.85% for the multiple-choice items (Table 69) and from 0.52% to 6.32% for the constructed-response items (Table 70). There was one constructed-response item with an omit rate greater than 5%.

Ninth Grade

Reading

Table 71 lists the results of the multiple-choice item analyses for the 2012 ninth-grade Reading assessment. The point biserials ranged from -0.04 to 0.51, with a mean of 0.33. The *p*-values for these multiple-choice items ranged from 0.12 to 0.98, 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 ranged from 0.35 to 0.62, with a mean of 0.52. The *p*-values ranged from 0.09 to 0.90, with a mean of 0.52. More than 50% of the students obtained the highest possible score points for one of the 14 constructed-response items. Scores were generally well distributed across the score points of the constructed-response items.

The omit rates for the ninth-grade Reading assessment ranged from 0.08% to 2.47% for the multiple-choice items (Table 71). The omit rates for the constructed-response items ranged from 0.84% to 7.64%, with three items having an omit rate greater than 5% (Table 72).

Writing

Table 73 lists the results of the multiple-choice item analyses for the 2012 ninth-grade Writing assessment. The point biserials for all multiple-choice items ranged from 0.19 to 0.58, with a mean of 0.42. The *p*-values for the multiple-choice items ranged from 0.41 to 0.90, with a mean of 0.70.

Table 74 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items ranged from 0.15 to 0.65, with a mean of 0.42. The *p*-values ranged from 0.22 to 0.98, with a mean of 0.66. More than 50% of the students obtained the highest possible score points for seven of the 13 constructed-response items. Scores were generally well distributed across the score points of the remaining constructed-response items.

The omit rates for the ninth-grade Writing assessment ranged from 0.07% to 0.93% for the multiple-choice items (Table 73). The omit rates for the constructed-response items ranged from 0% to 3.72% (Table 74).

Mathematics

Table 75 lists the results of the multiple-choice item analyses for the 2012 ninth-grade Mathematics assessment. The point biserials for all multiple-choice items ranged from 0.14 to 0.61, with a mean of 0.36. The *p*-values for the multiple-choice items ranged from 0.17 to 0.88, with a mean of 0.49.

Table 76 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items ranged from 0.44 to 0.74, with a mean of 0.61. The *p*-values ranged from 0.13 to 0.61, with a mean of 0.32. Scores were generally well distributed across the score points of the 15 constructed-response items.

The omit rates for the ninth-grade Mathematics assessment ranged from 0.07% to 0.71% for the multiple-choice items (Table 75). The omit rates for the constructed-response items ranged from 0.58% to 11.95%, with only two items greater than 5% (Table 76).

Tenth Grade

Reading

Table 77 lists the results of the multiple-choice item analyses for the 2012 tenth-grade Reading assessment. The point biserials for all multiple-choice items ranged from 0.04 to 0.56, with a mean of 0.35. The *p*-values for the multiple-choice items ranged from 0.29 to 0.95, 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 ranged from 0.38 to 0.67, with a mean of 0.55. The *p*-values ranged from 0.12 to 0.67, with a mean of 0.49. Scores were generally well distributed across the score points of the 14 constructed-response items.

The omit rates for the tenth-grade Reading assessment were small for the multiple-choice items but large for the constructed-response items. The omit rates for multiple-choice items ranged from 0.06% to 1.81% (Table 77). The omit rates for the constructed-response items ranged from 2.34% to 10.72%, with five out of the 14 items having an omit rate greater than 5% (Table 78).

Writing

Table 79 lists the results of the multiple-choice item analyses for the 2012 tenth-grade Writing assessment. The point biserials for all multiple-choice items ranged from 0.18 to 0.56, with a mean of 0.41. The *p*-values for the multiple-choice items ranged from 0.33 to 0.94, with a mean of 0.72.

Table 80 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items ranged from 0.14 to 0.66, with a mean of 0.41. The *p*-values ranged from 0.38 to 0.98, with a mean of 0.71. More than 50% of the students obtained the highest possible score points for six of the seven one-point items and for the only two-point item. Scores were generally well distributed across the score points of the remaining constructed-response items.

The omit rates for the tenth-grade Writing assessment ranged from 0.08% to 0.77% for the multiple-choice items (Table 79). The omit rates for the constructed-response items ranged from 0% to 3.89% (Table 80).

Mathematics

Table 81 lists the results of the multiple-choice item analyses for the 2012 tenth-grade Mathematics assessment. The point biserials for the multiple-choice items ranged from 0.14 to 0.58, with a mean of 0.37. The *p*-values for the multiple-choice items ranged from 0.24 to 0.87, with a mean of 0.48.

Table 82 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items ranged from 0.32 to 0.72, with a mean of 0.60. The *p*-values for the constructed-response items ranged from 0.17 to 0.65, with a mean of 0.38. Scores were generally well distributed across the score points of the 15 constructed-response items.

The omit rates for the tenth-grade Mathematics assessment ranged from 0.06% to 0.70% for the multiple-choice items (Table 81) and from 0.93% to 11.26% for the constructed-response items (Table 82). There was one constructed-response item with an omit rate greater than 5%.

Science

Table 83 lists the results of the multiple-choice item analyses for the 2012 tenth-grade Science assessment. The point biserials for the multiple-choice items ranged from -0.01 to 0.50, with a mean of 0.30. The *p*-values for these items ranged from 0.18 to 0.87, with a mean of 0.55.

Table 84 lists the results of the constructed-response item analyses. The item-to-total score correlations for the constructed-response items ranged from 0.11 to 0.53, with a mean of 0.39. The *p*-values for the constructed-response items ranged from 0.06 to 0.63, with a mean of 0.38. More than 50% of the students obtained the highest possible score points for three of the 23 constructed-response items. Scores were generally well distributed across the score points of the remaining constructed-response items.

The omit rates for the tenth-grade Science assessment ranged from 0.03% to 0.72% for the multiple-choice items (Table 83). The omit rates for the constructed-response items ranged from 1.18% to 18.72%, with four out of the 23 items having an omit rate greater than 5% (Table 84).

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 each 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

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

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:

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

and

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

and 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 where 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 92% 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 (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-square 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 TCAP 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.[9] The summary fit statistics for the multiple-choice and constructed-response items for all grades and content areas are shown in Tables 85 through 146.

Detailed summaries of the model fit results are presented below.

Third Grade

The third-grade item parameters and fit statistics are shown in Tables 85 through 94. The critical *Z*-values for these tables are 139.04 for Reading, 3.60 for Spanish Reading, 167.82 for Writing, 3.92 for Spanish Writing, and 170.46 for Mathematics.

[9] Results for the Spanish tests are based on previous years' data because these four tests were not recalibrated in 2012. The grade 4 Spanish tests were pre-equated using item parameters from several different prior administrations.

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

Fourth Grade

The fourth-grade item parameters and fit statistics are shown in Tables 95 through 104. The critical *Z*-values for these tables are 167.73 for Reading, 167.54 for Writing, and 167.69 for Mathematics. The pre-equated Spanish Reading test had a critical *Z*-value of 1.39 for items that originated in the 2004 administration, 1.30 for items that originated in the 2005 administration, and 0.70 for items that originated in the 2007 administration. The pre-equated Spanish Writing test had a critical *Z*-value of 1.40 for items that originated in the 2004 administration and 1.31 for items that originated in the 2005 administration. Spanish Writing grade 4 had 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, six items exceeded the critical *Z*-values and exhibited less than optimal fit: three Reading items (CR items 29, 108, and 109), two Writing items (CR items 3A and 92), and one Mathematics item (CR item 30).

Fifth Grade

The fifth-grade item parameters and fit statistics are shown in Tables 105 through 112. The critical *Z*-values for these tables are 165.93 for Reading, 166.00 for Writing, 166.83 for Mathematics, and 167.09 for Science.

Across all content areas, seven items exceeded these critical *Z*-values and exhibited less than optimal fit: two Reading items (MC item 102, CR item 40), two Writing items (CR items 3A and 93), and three Mathematics items (MC item 52, CR items 23 and 26).

Sixth Grade

The sixth-grade item parameters and fit statistics are shown in Tables 113 through 118. The critical *Z*-values for these tables are 164.81 for Reading, 164.62 for Writing, and 163.77 for Mathematics.

Across all content areas, 10 items exceeded these critical *Z*-values and exhibited less than optimal fit: three Reading items (MC item 45, CR items 12 and 32), two Writing items (CR items 3A and 94), and five Mathematics items (MC item 29, CR items 4, 17, 36, and 40).

Seventh Grade

The seventh-grade item parameters and fit statistics are shown in Tables 119 through 124. The critical *Z*-values for these tables are 161.72 for Reading, 161.52 for Writing, and 161.38 for Mathematics.

Across all content areas, six items exceeded these critical *Z*-values and exhibited less than optimal fit: two Writing items (CR items 3A and 93), and four Mathematics items (CR items 11, 20, 26, and 51).

Eighth Grade

The eighth-grade item parameters and fit statistics are shown in Tables 125 through 132. The critical *Z*-values for these tables are 158.95 for Reading, 158.33 for Writing, 159.14 for Mathematics, and 146.46 for Science.

Across all content areas, nine items exceeded these critical *Z*-values and exhibited less than optimal fit: two Reading items (MC item 47, CR item 46), two Writing items (CR items 3A and 98), four Mathematics items (CR items 25, 47, 53, and 60), and one Science item (CR item 75).

Ninth Grade

The ninth-grade item parameters and fit statistics are shown in Tables 133 through 138. The critical *Z*-values for these tables are 158.14 for Reading, 157.76 for Writing, and 158.71 for Mathematics.

Across all content areas, seven items exceeded these critical *Z*-values and exhibited less than optimal fit: two Writing items (CR items 3A and 116), and five Mathematics items (MC items 5, 15, and 47, CR items 2 and 60).

Tenth Grade

The tenth-grade item parameters and fit statistics are shown in Tables 139 through 146. The critical *Z*-values for these tables are 150.55 for Reading, 150.52 for Writing, 151.45 for Mathematics, and 151.15 for Science.

Across all content areas, six items exceeded these critical *Z*-values and exhibited less than optimal fit: two Reading items (MC item 97, CR item 27) and four Mathematics items (MC items 6 and 51, CR items 12 and 17).

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 within a test form 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 for each test form was quite small, ranging from zero to four pairs across grades and content areas. For Reading, two item pairs were flagged across all grades (grade 3 items 7 and 24 and grade 9 items 100 and 101). For Writing, 23 pairs were flagged (grade 4 items 3A and 3B; grade 4 items 3B and 3C; grade 4 items 69 and 116; grade 4 items 92 and 116; grade 5 items 3A and 3B; grade 5 items 93 and 116; grade 6 items 3A and 3B; grade 7 items 3A and 3B; grade 7 items 3B and 3C; grade 7 items 70 and 116; grade 8 items 3A and 3B; grade 8 items 75 and 98; grade 8 items 75 and 116; grade 8 items 98 and 116; grade 9 items 3A and 3B; grade 9 items 72 and 94; grade 9 items 72 and 116; grade 9 items 94 and 116; grade 10 items 3A and 3B; grade 10 items 3B and 3C; grade 10 items 71 and 95; grade 10 items 71 and 116; and grade 10 items 95 and 116). For Mathematics, one pair was flagged (grade 10 items 1 and 19). For Science, no item pairs were flagged. When compared to grades 3 and 4 English Writing items, a relatively larger number of items in the Spanish tests were flagged[10] but for lower Q3 values ranging from 0.33 to 0.56 (12 pairs in all of the 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).

Evaluation of Item Analysis and Calibration

After the evaluation of item analyses and calibration outputs across all grades and content areas, eight multiple-choice items exhibited aberrant characteristics (non-convergence where the item parameters could not be estimated, poor model fit, negative point biserials for the correct choice, or positive point biserials

[10] Note that the item dependency statistics for the Spanish assessments are based upon results from a previous administration because these tests were not recalibrated in 2012.

for distractor(s)). After consulting with CTB content experts and CDE, the following MC items were removed from the final calibration:

- Reading, grade 4—Item 22
- Reading, grade 7—Item 50
- Reading, grade —9 Items 21 and 105
- Reading, grade 10—Items 9 and 20
- Mathematics, grade 8—Item 57
- Science, grade 10—Item 60

Tables 2 through 6 indicate the number of items and score points for each test form after suppressed items were removed.

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. 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 in the same third of the test as their 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 2012 TCAP to link the tests across years. The anchor parameters were not fixed during calibration and were used in the equating procedures defined by Stocking and Lord (1983). The anchor parameters were used to place the estimated parameters for the spring 2012 TCAP 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 procedure (1983), 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 re-estimation of item parameters using current data and transformed to the previous scale.

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

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—the TCC method (Stocking & Lord, 1983), the Delta Plot method (Angoff, 1972; Dorans & Holland, 1993), and the 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 (100*(1– *p*))th percentiles. For example, for a *p*-value of 0.90, the corresponding *Z*-score will be at the 10^{th} percentile (100*(1 – 0.90)), 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

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

and

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

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.

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 details). 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*-value differences.
3) The IRT 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 results included the following:

6) The *p*-values of the common anchor items between the two administrations were compared to show that changes in the *p*-values were consistent with changes in the scale scores.
7) The full distribution of scale scores was compared for reasonableness across administrations and results verified to ensure that any observed differences were consistent with the differences 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 quality-check steps were followed during equating for all grades and content areas.

Anchor Items Evaluation Results

The Colorado Department of Education had the final responsibility for determining which items would or would not be removed from the anchor sets. The primary 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 0.10, it would be dropped from the anchor set. Items that did not meet these criteria but exhibited other serious statistical problems or content-related issues also were carefully reviewed in making this determination.

After a careful review of the 31 grade/content areas in the 2012 TCAP administration, three items met all of the criteria for removal from the anchor sets (delta, chi-square, and *p*-value difference). Of these three items, one item (grade 4 Mathematics item 42) was removed from the anchor set, and two were retained in order to maintain appropriate blueprint representation in the anchor set. An additional 11 items met only the delta and chi-square criteria. Three of these were dropped from the anchor set because they exhibited other signs of poor statistical performance (grade 5 Science item 52, grade 7 Writing item 62, and grade 9 Mathematics item 17), and the remaining eight were retained in order to maintain appropriate blueprint representation. Five other items (grade 8 Reading items 27, 28, 29, 31, and 34) were subsequently removed from the anchor set because their item positions in the 2012 test books were too far away from their previous positions.[11] The *p*-values and item parameter comparison results for the final anchor sets are presented below.

Figures 1 through 9 show the item characteristic curves for the anchor items removed from the equating of the 2012 TCAP operational tests.

p-value Comparisons

The analysis of *p*-values across administrations indicated that the values were aligned closely, with correlations at or above 0.97 for all grades and content areas (Table 147). 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 across years.

[11] One of these items (grade 8 Reading item 31) was also flagged by the delta plot and chi-square methods; but was not removed during the initial anchor review because it did not show a large difference in *p*-values.

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 exhibit the greatest fluctuation and were not considered in the evaluation criteria.

Results indicate that the parameter correlations for item difficulty (*b*) and discrimination (*a*) are high (see Table 147). This indicates that the items were performing similarly in the two administrations and provides further evidence that the equating results are reasonable and accurate. The *b*-parameter correlations ranged from 0.91 to 0.99. The *a*-parameter correlations ranged from 0.90 to 0.98.

As noted previously, anchor item functioning in terms of item characteristic curves from the two administrations was also evaluated using Lord's Chi-Square and Delta Plot methods. In addition to the items that were removed from the anchor sets, nine other items were flagged using Lord's Chi Square and Delta Plot methods but were not removed. This group was made up of two items in grade 8 Reading, one in grade 5 Writing, one in grade 6 Writing, two in grade 8 Writing, one in grade 7 Mathematics, one in grade 8 Mathematics, and one in Grade 5 Science.

Scaling Constants

The scaling constants (linear transformation parameters that were used to place scores onto the equated scale score metric) were examined to determine whether the ability levels of students in the calibration and equating samples varied over time or were similar across years. Since the calibration "centers" the raw IRT scale close to the average ability of the sample, differences in these scaling constants would indicate differences in the ability distributions of the calibration samples from reference to new form administrations. The scaling constants for the TCAP grades and content areas are displayed in Table 148 for the 2011 and 2012 administrations.

Table 148 indicates that for most grades and content areas the scaling constants are fairly similar across the two administrations.

Analyses after removing the Flagged Items

Review of the content balance for the final anchor sets after removing the flagged items indicated that these anchors were reasonably representative of the blueprint for the total tests. Tables 149 through 153 show the number and percentage of items by content standard for the total test and the anchor set.

Effectiveness of the Equating

Figures 10 through 36 show the TCC and SEM plots for the spring 2012 operational tests in grades 3 through 10 Reading (Figures 10 through 17), Writing (Figures 18 through 25), Mathematics (Figures 26 through 33), and grades 5, 8, and 10 Science (Figures 34 through 36) 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 similarity of the plots of the TCCs (the S-shaped curves) and the SEM curves (the U-shaped curves) for each subject area and grade indicates that the test forms administered in 2011 and 2012 strongly resembled each other in terms of item difficulty, discrimination, and accuracy. Note that because the Spanish Reading and Writing tests were not post-equated this year, the plots for these tests are not included.

After the tests were equated, the final scaled parameters were used to derive each student's scale score. The TCAP uses item-pattern scoring for all tests. During item-pattern scoring, the pattern of student responses and the attributes of each item contribute to the student's final scale score. 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. This enhances the comparability of scores across years.

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 10.

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 the 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 the CSAP. Table 154 presents the mean, median, and standard deviation of the scale scores for the total population and for each gender in each grade/content area. Note that the male and female students do not necessarily equal the total population because some students may not identify their gender.

On average, female students scored higher than male students at all grade levels on the Reading and Writing tests. For Mathematics, male students scored slightly higher than females in grades 3, 4, 9, and 10, female students scored slightly higher than males in grades 6, 7, and 8, and male and female students scored the same in grade 5. For Science, male students scored slightly higher than females in grades 5 and 10, and female students scored slightly higher than males in grade 8.

Tables 155 and 156 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 157 and 158 contain raw score descriptive statistics for the total population, including the mean percent of the maximum points obtained for each content standard and subcontent area, respectively.

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 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 155 through 158, where content standards or subcontent areas are combined (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 2012 third-grade Reading assessment are 563 and 569, respectively, with a standard deviation of 74.5. The mean scale score for female students is 572, with a standard deviation of 70.8, and the mean scale score for male students is 554, with a standard deviation of 77.1.

The scale score frequency distribution of the third-grade Reading assessment for the total population is shown in Table 159. Figure 37 graphically represents the scale score frequency distributions for the total population and for the groups of female and male students separately. The figure shows that the distributions of scale scores for the total population and for each gender are slightly negatively skewed with some floor effects.

The mean scale score for the single content standard is 563, and the median is 569 (Table 155). The mean scale scores for the subcontent areas range from 556 to 594, and the median scale scores range from 569 to 571 (Table 156).

The mean percentages of the maximum obtainable raw score for the subcontent areas range from 56.3% to 75.9% (Table 158). The mean percentage of the maximum obtainable raw score for the total test is 61.2%.

Reading—Spanish

The mean scale score for the total population of students taking the 2012 third-grade Spanish Reading assessment is 518, with a standard deviation of 50.1. The mean scale score for female students is 527, with a standard deviation of 44.2, and the mean scale score for male students is 510, with a standard deviation of 53.8.

The scale score frequency distribution of the third-grade Spanish Reading assessment for the total population is shown in Table 160. Figure 38 graphically represents the scale score frequency distributions for the total population and for the groups of female and male students separately. The figure shows that the distributions of scale scores for the total population and for each gender are essentially normal, but with some floor effects.[12]

The mean scale score for the single content standard is 518, and the median is 521. The mean scale scores for all the subcontent areas range from 518 to 520; the median scale scores for the subcontent areas range from 521 to 524, close to the median scale score of 521 for the total test.

The mean percentages of the maximum obtainable raw score for the subcontent areas range from 54.0% to 60.8%. The mean percentage of the maximum obtainable raw score for the total test is 57.7%.

Writing

The mean and median scale scores for the total population of students taking the 2012 third-grade Writing assessment are both 467, with a standard deviation of 50.0. The mean scale score for female students is 475, with a standard deviation of 48.7, and the mean scale score for male students is 459, with a standard deviation of 49.9.

The scale score frequency distribution for the total population is shown in Table 161. Figure 39 graphically represents the scale score frequency distributions for the total population and for the groups of female and male students separately. The figure shows that the distributions of scale scores for the total population and for each gender are approximately normal, but with some slight ceiling effects.[13]

The mean scale scores for the two content standards are 468 and 476. The mean scale scores for the subcontent areas range from 471 to 491. The median scale scores range from 467 to 468 for the content standards and from 468 to 469 for the subcontent areas.

The mean percentages of the maximum obtainable raw score for the content standards range from 71.5% to 78.5%. The mean percentages of the maximum obtainable raw score for the subcontent areas range from 73.4% to 79.7%. The mean percentage of the maximum obtainable raw score for the total test is 75.4%.

[12] Floor effects are indicated by a pileup of scores at the bottom of the scale and suggest that the true ability of some of the tested students was lower than the lowest obtainable scale score.

[13] Ceiling effects are indicated by a pileup of scores at the top of the scale and suggest that the true ability of some of the tested students was higher than the highest obtainable scale score.

Writing—Spanish

The mean and median scale scores for the total population of students taking the 2012 third-grade Spanish Writing assessment are both 511, with a standard deviation of 72.1. The mean scale score for female students is 525, with a standard deviation of 70.1, and the mean scale score for male students is 498, with a standard deviation of 71.6.

The scale score frequency distribution of the third-grade Spanish Writing assessment for the total population is shown in Table 162. Figure 40 graphically represents the scale score frequency distributions for the total population and for the groups of female and male 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 521 and 507, with median scale scores of 524 and 500. The mean scale scores for the subcontent areas range from 510 to 535, and the median scale scores for the subcontent areas vary between 503 and 551.

The mean percentages of the maximum obtainable raw scores range from 67.0% to 72.5% for the content standards and from 67.7% to 73.5% for the subcontent areas. The mean percentage of the maximum obtainable raw score for the total test is 70.1%.

Mathematics

The mean and median scale scores for the total population of students taking the 2012 third-grade Mathematics assessment are 464 and 468, respectively, with a standard deviation of 90.8. The mean scale score for female students is 460, with a standard deviation of 87.7, and the mean scale score for male students is 467, with a standard deviation of 93.6.

The scale score frequency distribution for the total population is shown in Table 163. Figure 41 graphically represents the frequency distributions for the total population and for the groups of female and male students separately. The figure shows that the scale score distributions for the total population and for each gender are approximately normal, with some slight floor and ceiling effects.

The mean scale scores for the content standards range from 462 to 469, and the medians range from 466 to 469. Subcontent area scores are not computed for the grade 3 Mathematics test.

The mean percentages of the maximum obtainable raw score for the content standards range from 59.6% to 68.9%. The mean percentage of the maximum obtainable raw score for the total test is 65.3%.

Fourth Grade

Reading

The mean and median scale scores for the total population of students taking the 2012 fourth-grade Reading assessment are 587 and 595, respectively, with a standard deviation of 63.4. The mean scale score for female students is 594, with a standard deviation of 57.9, and the mean scale score for male students is 579, with a standard deviation of 67.4.

The scale score frequency distribution for the total population is shown in Table 164. Figure 42 graphically represents the scale score frequency distributions for the total population and for the groups of female and male students separately. The figure shows that the distributions of scale scores for the total population and for each gender are slightly negatively skewed with some slight floor effects.

The mean scale scores for the content standards range from 583 to 590. The mean scale scores for the subcontent areas range from 582 to 624. The median scale scores are 595 for all of the content standards and range from 595 to 596 for the subcontent areas.

The mean percentages of the maximum obtainable raw score for the content standards range from 57.2% to 67.4%. The mean percentage of the maximum obtainable raw score for the total test is 63.7%. The mean percentages of the maximum raw score for the subcontent areas range from 57.0% to 75.6%.

Reading—Spanish

The mean and median scale scores for the total population of students taking the 2012 fourth-grade Spanish Reading assessment are 515 and 517, respectively, with a standard deviation of 46.7. The mean scale score for female students is 525, with a standard deviation of 51.0, and the mean scale score for male students is 506, with a standard deviation of 40.3.

The scale score frequency distribution for the total population is shown in Table 165. Figure 43 graphically represents the scale score frequency distributions for the total population and for the groups of female and male 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 505 to 515. The mean scale scores for the subcontent areas range from 506 to 525. The median scale scores vary between 507 and 518 for the content standards and between 514 and 526 for the subcontent areas. The median for the total test scale score is 517.

The mean percentages of the maximum obtainable raw score for the content standards range from 43.4% to 54.9%. The mean percentage of the maximum obtainable score for the total test is 48.9%. The mean percentages of the maximum raw score for the subcontent areas range from 45.4% to 63.4%.

Writing

The mean and median scale scores for the total population of students taking the 2012 fourth-grade Writing assessment are 483 and 484, respectively, with a standard deviation of 50.3. The mean scale score for female students is 493, with a standard deviation of 49.6, and the mean scale score for male students is 475, with a standard deviation of 49.4.

The scale score frequency distribution for the total population is shown in Table 166. Figure 44 graphically represents the scale score frequency distributions for the total population and for the groups of female and male 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 484 and 489. The mean scale scores for the subcontent areas range from 485 to 503. The median scale scores are 484 for both of the content standards and range from 479 to 487 for the subcontent areas.

The mean percentages of the maximum obtainable raw score for the content standards range from 65.4% to 75.1%. The mean percentage of the maximum obtainable raw score for the total test is 70.0%. The mean percentages of the maximum raw score for the subcontent areas range from 63.0% to 79.2%.

Writing—Spanish

The mean and median scale scores for the total population of students taking the 2012 fourth-grade Spanish Writing assessment are 497 and 495, respectively, with a standard deviation of 46.3. The mean scale score for female students is 509, with a standard deviation of 47.0, and the mean scale score for male students is 486, with a standard deviation of 43.2.

The scale score frequency distribution for the total population is shown in Table 167. Figure 45 graphically represents the scale score frequency distributions for the total population and for the groups of female and male students separately. The figure shows that the distributions of scale scores are approximately normal.

The mean scale scores for the two content standards are 490 and 500. The mean scale scores for the subcontent areas range from 469 to 502. The median scale scores for the two content standards are 495 and 503. The median scale scores for the subcontent areas vary between 482 and 508.

The mean percentages of the maximum obtainable raw score for the content standards range from 47.9% to 50.3%. The mean percentage of the maximum obtainable raw score for the total test is 49.1%. The mean percentages of the maximum raw score for the subcontent areas range from 42.5% to 55.8%.

Mathematics

The mean and median scale scores for the total population of students taking the 2012 fourth-grade Mathematics assessment are 495 and 500, respectively, with a standard deviation of 80.2. The mean scale score for female students is 493, with a standard deviation of 77.9, and the mean scale score for male students is 497, with a standard deviation of 82.2.

The scale score frequency distribution for the total population is shown in Table 168. Figure 46 graphically represents the frequency distributions for the total population and for the groups of female and male 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 498 to 506. The mean scale scores for the subcontent areas range from 501 to 551. The median scale scores range from 499 to 502 for the content standards and from 497 to 500 for the subcontent areas.

The mean percentages of the maximum obtainable raw score for the content standards range from 64.5% to 71.7%. The mean percentage of the maximum obtainable raw score for the total test is 68.7%. The mean percentages of the maximum raw score for the subcontent areas range from 65.2% to 79.3%.

Fifth Grade

Reading

The mean and median scale scores for the total population of students taking the 2012 fifth-grade Reading assessment are 611 and 620, respectively, with a standard deviation of 69.8. The mean scale score for female students is 619, with a standard deviation of 63.4, and the mean scale score for male students is 603, with a standard deviation of 74.7.

The scale score frequency distribution for the total population is shown in Table 169. Figure 47 graphically represents the frequency distributions for the total population and for the groups of female and male students separately. The figure shows that the scale score distributions for the total population and for each gender are slightly negatively skewed with some floor effects.

The mean scale scores for the content standards range from 605 to 612. The mean scale scores for the subcontent areas range from 606 to 659. The median scale scores range from 619 to 621 for the content standards and are 620 for all of the subcontent areas, with a median of 620 for the total test.

The mean percentages of the maximum obtainable raw score for content standards range from 56.6% to 61.4%. The mean percentage of the maximum obtainable raw score for the total test is 58.7%. The mean percentages of the maximum raw score for the subcontent areas range from 53.0% to 75.9%.

Writing

The mean and median scale scores for the total population of students taking the 2012 fifth-grade Writing assessment are both 509, with a standard deviation of 58.2. The mean scale score for female students is 519, with a standard deviation of 57.2, and the mean scale score for male students is 498, with a standard deviation of 57.2.

The scale score frequency distribution for the total population is shown in Table 170. Figure 48 graphically represents the scale score frequency distributions for the total population and for the groups of female and male 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 510 and 516. The mean scale scores for the subcontent areas range from 512 to 534. The median scale scores range from 508 to 509 for the content standards and from 509 to 541 for the subcontent areas. Most of the median scale scores for the content standards and subcontent areas are at or near the median of 509 for the total test.

The mean percentages of the maximum obtainable raw score for content standards range from 68.4% to 75.0%. The mean percentage of the maximum obtainable raw score for the total test is 71.5%. The mean percentages of the maximum raw score for the subcontent areas range from 65.4% to 74.4%.

Mathematics

The mean and median scale scores for the total population of students taking the 2012 fifth-grade Mathematics assessment are 518 and 521, respectively, with a standard deviation of 75.4. The mean scale score for female students is 518, with a standard deviation of 72.2, and the mean scale score for male students is 518, with a standard deviation of 78.3.

The scale score frequency distribution for the total population is shown in Table 171. Figure 49 graphically represents the frequency distributions for the total population and for the groups of female and male 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 520 to 530. The mean scale scores for the subcontent areas range from 523 to 527. The median scale scores vary from 520 to 522 for the content standards and from 519 to 521 for the subcontent areas.

The mean percentages of the maximum obtainable raw score for the content standards range from 61.1% to 70.2%. The mean percentage of the maximum obtainable raw score for the total test is 63.6%. The mean percentages of the maximum raw score for the subcontent areas range from 62.1% to 65.7%.

Science

The mean and median scale scores for the total population of students taking the 2012 fifth-grade Science assessment are 499 and 505, respectively, with a standard deviation of 66.6. The mean scale score for female students is 498, with a standard deviation of 65.0, and the mean scale score for male students is 500, with a standard deviation of 68.1.

The scale score frequency distribution for the total population is shown in Table 172. Figure 50 graphically represents the frequency distributions for the total population and for the groups of female and male students separately. The distributions of the scale scores are approximately normal with some floor effects for the total population and for each gender.

The mean scale scores for the content standards range from 499 to 503. The mean scale scores for the subcontent areas range from 500 to 515. The median scale scores vary from 504 to 506 for the content standards and were 505 for both of the subcontent areas, and all are very close to the median scale score of 505 for the total test.

The mean percentages of the maximum obtainable raw score for the content standards range from 53.3% to 61.8%. The mean percentage of the maximum obtainable raw score for the total test was 58.8%. The mean percentages of the maximum raw score for the subcontent areas range from 58.0% to 61.9%.

Sixth Grade

Reading

The mean and median scale scores for the total population of students taking the 2012 sixth-grade Reading assessment are 629 and 636, respectively, with a standard deviation of 63.8. The mean scale score for female students is 638, with a standard deviation of 58.9, and the mean scale score for male students is 620, with a standard deviation of 67.0.

The scale score frequency distribution for the total population is shown in Table 173. Figure 51 graphically represents the frequency distributions for the total population and for the groups of female and male students separately. The figure shows that the scale score distributions for the total population and for each gender are slightly negatively skewed with some slight floor effects.

The mean scale scores for the content standards range from 625 to 635. The mean scale scores for the subcontent areas range from 625 to 678. The median scale scores vary from 635 to 637 for the content standards, from 636 to 640 for the subcontent areas, and all are close to the median scale score of 636 for the total test.

The mean percentages of the maximum obtainable raw score for content standards range from 52.5% to 68.0%. The mean percentage of the maximum obtainable raw score for the total test is 62.5%. The mean percentages of the maximum raw score for the subcontent areas range from 55.2% to 76.3%.

Writing

The mean and median scale scores for the total population of students taking the 2012 sixth-grade Writing assessment are both 521, with a standard deviation of 61.8. The mean scale score for female students is 534, with a standard deviation of 59.6, and the mean scale score for male students is 509, with a standard deviation of 61.5.

The scale score frequency distribution for the total population is shown in Table 174. Figure 52 graphically represents the scale score frequency distributions for the total population and for the groups of female and male 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 are both 523. The mean scale scores for the subcontent areas range from 523 to 597. The median scale scores are 522 for both of the content standards and range from 511 to 528 for the subcontent areas.

The mean percentages of the maximum obtainable raw score for content standards range from 68.3% to 68.9%. The mean percentage of the maximum obtainable raw score for the total test is 68.6%. The mean percentages of the maximum raw score for the subcontent areas range from 64.9% to 82.1%.

Mathematics

The mean and median scale scores for the total population of students taking the 2012 sixth-grade Mathematics assessment are 537 and 541, respectively, with a standard deviation of 73.9. The mean scale score for female students is 537, with a standard deviation of 70.5, and the mean scale score for male students is 536, with a standard deviation of 77.1.

The scale score frequency distribution for the total population is shown in Table 175. Figure 53 graphically represents the frequency distributions for the total population and for the groups of female and male students separately. The figure shows that the distributions of scale scores for the total population and for each gender are approximately normal, with some slight floor effects.

The mean scale scores for the content standards range from 531 to 554. The mean scale scores for the subcontent areas range from 513 to 541. The median scale scores vary between 541 and 543 for the content standards and between 537 and 544 for the subcontent areas, and all are close to the median total test scale score of 541.

The mean percentages of the maximum obtainable raw score for the content standards range from 45.5% to 65.7%.The mean percentage of the maximum obtainable raw score for the total test is 56.8%. The mean percentages of the maximum raw score for the subcontent areas range from 38.6% to 62.3%.

Seventh Grade

Reading

The mean and median scale scores for the total population of students taking the 2012 seventh-grade Reading assessment are 641 and 649, respectively, with a standard deviation of 65.7. The mean scale score for female students is 649, with a standard deviation of 61.0, and the mean scale score for male students is 632, with a standard deviation of 68.9.

The scale score frequency distribution for the total population is shown in Table 176. Figure 54 graphically represents the frequency distributions for total population and for the groups of female and male students separately. The figure indicates that the distribution of the scale scores for the total population and for each gender is slightly negatively skewed with some slight floor effects.

The mean scale scores for the content standards range from 639 to 648. The mean scale scores for the subcontent areas range from 640 to 710. The median scale scores vary from 647 to 649 for the content standards and from 648 to 653 for the subcontent areas, and all are close to the median total test scale score of 649.

The mean percentages of the maximum obtainable raw score for the content standards range from 57.6% to 70.7%. The mean percentage of the maximum obtainable raw score for the total test is 65.1%. The mean percentages of the maximum raw score for the subcontent areas range from 59.2% to 79.3%.

Writing

The mean and median scale scores for the total population of students taking the 2012 seventh-grade Writing assessment are 559 and 558, respectively, with a standard deviation of 70.3. The mean scale score for female students is 575, with a standard deviation of 69.1, and the mean scale score for male students is 543, with a standard deviation of 67.8.

The scale score frequency distribution for the total population is shown in Table 177. Figure 55 graphically represents the frequency distributions for the total population and for the groups of female and male 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 560 to 564. The mean scale scores for the subcontent areas range from 562 to 582. The median scale scores range from 556 to 559 for the content standards and from 526 to 560 for the subcontent areas. Most of the median scale scores for content standards and subcontent areas are close to the median total test scale score of 558.

The mean percentages of the maximum obtainable raw score for content standards range from 67.2% to 73.8%. The mean percentage of the maximum obtainable raw score for the total test is 70.3%. The mean percentages of the maximum raw score for the subcontent areas range from 61.1% to 78.5%.

Mathematics

The mean and median scale scores for the total population of students taking the 2012 seventh-grade Mathematics assessment are 561 and 565, respectively, with a standard deviation of 75.8. The mean scale score for female students is 562, with a standard deviation of 71.8, and the mean scale score for male students is 560, with a standard deviation of 79.5.

The scale score frequency distribution for the total population is shown in Table 178. Figure 56 graphically represents the frequency distributions for the total population and for the groups of female and male students separately. The figure indicates that the scale score distributions are approximately normal with some slight floor effects for the total population and for each gender.

The mean scale scores for the content standards ranged from 550 to 561. The mean scale scores were 548 for both of the subcontent areas. The median scale scores vary from 564 to 566 for the content standards and vary from 560 to 566 for the subcontent areas. All are close to the median total test scale score of 565.

The mean percentages of the maximum obtainable raw score for the content standards range from 39.3% to 52.1%. The mean percentage of the maximum obtainable raw score for the total test is 47.8%. The mean percentages of the maximum raw score for the subcontent areas range from 34.4% to 38.6%.

Eighth Grade

Reading

The mean and median scale scores for the total population of students taking the 2012 eighth-grade Reading assessment are 651 and 657, respectively, with a standard deviation of 60.3. The mean scale score for female students is 660, with a standard deviation of 56.6, and the mean scale score for male students is 642, with a standard deviation of 62.4.

The scale score frequency distribution for the total population is shown in Table 179. Figure 57 graphically represents the frequency distributions for the total population and for the groups of female and male students separately. The figure shows that the scale score distributions are very slightly negatively skewed.

The mean scale scores for the content standards range from 647 to 654. The mean scale scores for the subcontent areas range from 639 to 685. The median scale scores vary from 656 to 658 for the content standards and for the subcontent areas. All of the median scale scores for content standards and subcontent areas are close to the median total test scale score of 657.

The mean percentages of the maximum obtainable raw score for the content standards range from 51.7% to 69.2%. The mean percentage of the maximum obtainable raw score for the total test is 61.4%. The mean percentages of the maximum raw score for the subcontent areas range from 49.8% to 74.8%.

Writing

The mean and median scale scores for the total population of students taking the 2012 eighth-grade Writing assessment are both 564, with a standard deviation of 70.8. The mean scale score for female students is 581, with a standard deviation of 69.7, and the mean scale score for male students is 549, with a standard deviation of 68.3.

The scale score frequency distribution for the total population is shown in Table 180. Figure 58 graphically represents the frequency distributions for the total population and for the groups of female and male 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 565 to 573. The mean scale scores for the subcontent areas range from 567 to 589. The median scale scores vary from 563 to 565 for the content standards and from 562 to 609 for the subcontent areas, and most are close to the median total test scale score of 564.

The mean percentages of the maximum obtainable raw score for the content standards range from 66.3% to 75.2%. The mean percentage of the maximum obtainable raw score for the total test is 70.5%. The mean percentages of the maximum raw score for the subcontent areas range from 63.5% to 74.6%.

Mathematics

The mean and median scale scores for the total population of students taking the 2012 eighth-grade Mathematics assessment are 576 and 580, respectively, with a standard deviation of 71.6. The mean scale score for female students is 577, with a standard deviation of 67.6, and the mean scale score for male students is 575, with a standard deviation of 75.1.

The scale score frequency distribution for the total population is shown in Table 181. Figure 59 graphically represents the frequency distributions for the total population and for the groups of female and male students separately. The scale score distributions are very slightly negatively skewed with some slight floor effects for the total population and for each gender.

The mean scale scores for the content standards range from 569 to 578. The mean scale scores for subcontent areas range from 539 to 579. The median scale scores vary between 578 and 581 for the content standards and between 577 and 580 for the subcontent areas, and all are close to the median total test scale score of 580.

The mean percentages of the maximum obtainable raw score for the content standards range from 41.8% to 51.9%.The mean percentage of the maximum obtainable raw score for the total test is 46.7%. The mean percentages of the maximum raw score for the subcontent areas range from 36.2% to 51.1%.

Science

The mean and median scale scores for the total population of students taking the 2012 eighth-grade Science assessment are 499 and 505, respectively, with a standard deviation of 60.0. The mean scale score for female students is 501, with a standard deviation of 57.7, and the mean scale score for male students is 498, with a standard deviation of 62.2.

The scale score frequency distribution for the total population is shown in Table 182. Figure 60 graphically represents the frequency distributions for the total population and for the groups of female and male students separately. The distributions of the scale scores are approximately normal with some slight floor effects for the total population and for each gender.

The mean scale scores for the content standards range from 496 to 500. The mean scale scores for the subcontent areas range from 495 to 502. The median scale scores vary between 504 and 506 for the content standards and between 504 and 508 for the subcontent areas, all are close to the median total test scale score of 505.

The mean percentages of the maximum obtainable raw score for the content standards range from 47.7% to 56.6%. The mean percentage of the maximum obtainable raw score for the total test is 52.7%. The mean percentages of the maximum raw score for the subcontent areas range from 48.3% to 57.5%.

Ninth Grade

Reading

The mean and median scale scores for the total population of students taking the 2012 ninth-grade Reading assessment are 660 and 664, respectively, with a standard deviation of 50.2. The mean scale score for female students is 668, with a standard deviation of 48.0, and the mean scale score for male students is 653, with a standard deviation of 51.0.

The scale score frequency distribution for the total population is shown in Table 183. Figure 61 graphically represents the frequency distributions for the total population and for the groups of female and male students separately. The figure shows that the scale score distributions for the total population and for each gender are very slightly negatively skewed.

The mean scale scores for the content standards range from 655 to 668. The mean scale scores for the subcontent areas range from 654 to 685. The median scale scores were all 664 for the content standards, vary between 663 and 666 for the subcontent areas, and almost all are close to the median total test scale score of 664.

The mean percentages of the maximum obtainable raw score for the content standards range from 41.9% to 68.6%. The mean percentage of the maximum obtainable raw score for the total test is 60.6%. The mean percentages of the maximum raw score for the subcontent areas range from 50.1% to 72.4%.

Writing

The mean and median scale scores for the total population of students taking the 2012 ninth-grade Writing assessment are 565 and 566 respectively, with a standard deviation of 75.4. The mean scale score for female students is 581, with a standard deviation of 72.7, and the mean scale score for male students is 550, with a standard deviation of 74.8.

The scale score frequency distribution for the total population is shown in Table 184. Figure 62 graphically represents the frequency distributions for the total population and for the groups of female and male 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 566 to 574. The mean scale scores for the subcontent areas range from 566 to 592. The median scale scores range from 566 to 567 for the content standards and range from 564 to 600 for the subcontent areas. Most are close to the median scale score of 566 for the total test. The median scale score for SA 6 (Extended Writing) is somewhat higher 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 for this subcontent area is rather discrete.

The mean percentages of the maximum obtainable raw score for the content standards range from 64.3% to 72.9%. The mean percentage of the maximum obtainable raw score for the total test is 68.4%. The mean percentages of the maximum raw score for the subcontent areas range from 60.6% to 76.0%.

Mathematics

The mean and median scale scores for the total population of students taking the 2012 ninth-grade Mathematics assessment are 575 and 581, respectively, with a standard deviation of 73.4. The mean scale score for female students is 575, with a standard deviation of 71.2, and the mean scale score for male students is 576, with a standard deviation of 75.5.

The scale score frequency distribution for the total population is shown in Table 185. Figure 63 graphically represents the frequency distributions for the total population and for the groups of female and male students separately. The scale score distributions are very slightly negatively skewed with some floor effects for the total population and for each gender.

The mean scale scores for the content standards range from 563 to 575. The mean scale scores for the subcontent areas range from 556 to 578. The median scale scores vary between 580 and 583 for the content standards and the subcontent areas, with all of the medians very close to the median total test scale score of 581.

The mean percentages of the maximum obtainable raw score for the content standards range from 36.0% to 44.3%. The mean percentage of the maximum obtainable raw score for the total test is 40.6%. The mean percentages of the maximum raw score for the subcontent areas range from 34.2% to 46.7%.

Tenth Grade

Reading

The mean and median scale scores for the total population of students taking the 2012 tenth-grade Reading assessment are 682 and 687, respectively, with a standard deviation of 53.5. The mean scale score for female students is 692, with a standard deviation of 48.5, and the mean scale score for male students is 672, with a standard deviation of 56.2.

The scale score frequency distribution for the total population is shown in Table 186. Figure 64 graphically represents the frequency distributions for total population and for the groups of female and male students separately. The figure shows that the scale score distributions for the total population and for each gender are slightly negatively skewed with some very slight floor effects.

The mean scale scores for the content standards range from 680 to 686. The mean scale scores for the subcontent areas range from 677 to 699. The median scale scores vary from 686 to 688 for the content standards and from 686 to 689 for the subcontent areas, and all are close to the median total test scale score of 687.

The mean percentages of the maximum obtainable raw score for the content standards range from 50.2% to 66.6%. The mean percentage of the maximum obtainable raw score for the total test is 60.0%. The mean percentages of the maximum raw score for the subcontent areas range from 52.2% to 69.8%.

Writing

The mean and median scale scores for the total population of students taking the 2012 tenth-grade Writing assessment are 576 and 577, respectively, with a standard deviation of 76.9. The mean scale score for female students is 593, with a standard deviation of 74.5, and the mean scale score for male students is 560, with a standard deviation of 75.7.

The scale score frequency distribution for the total population is shown in Table 187. Figure 65 graphically represents the frequency distributions for the total population and for the groups of female and male students separately. The figure shows that the scale score distributions for the total population and for each gender are approximately normal with some very slight floor effects.

The mean scale scores for the content standards range from 578 to 583. The mean scale scores for the subcontent areas range from 572 to 625. The median scale scores vary between 575 and 578 for the content standards and between 579 and 606 for the subcontent areas, with most very close to the median scale score of 577 for the total test.

The mean percentages of the maximum obtainable raw score for the content standards range from 67.1% to 74.5%. The mean percentage of the maximum obtainable raw score for the total test is 70.6%. The mean percentages of the maximum raw score for the subcontent areas range from 56.3% to 79.8%.

Mathematics

The mean and median scale scores for the total population of students taking the 2012 tenth-grade Mathematics assessment are 594 and 600, respectively, with a standard deviation of 70.8. The mean scale score for female students is 593, with a standard deviation of 66.2, and the mean scale score for male students is 594, with a standard deviation of 75.0.

The scale score frequency distribution for the total population is shown in Table 188. Figure 66 graphically represents the frequency distributions for the total population and for the groups of female and male students separately. The figure shows that the scale score distributions for the total population and for each gender are slightly negatively skewed with some floor effects.

The mean scale scores for the content standards range from 576 to 595. The mean scale scores for the subcontent areas range from 592 to 633. The median scale scores vary between 599 and 600 for the content standards and between

584 and 600 for the subcontent areas, and most are close to the median total test scale score of 600.

The mean percentages of the maximum obtainable raw score for the content standards range from 30.6% to 56.7%. The mean percentage of the maximum obtainable raw score for the total test is 43.4%. The mean percentages of the maximum raw score for the subcontent areas range from 42.1% to 57.7%.

Science

The mean and median scale scores for the total population of students taking the 2012 tenth-grade Science assessment are 500 and 507, respectively, with a standard deviation of 62.2. The mean scale score for female students is 499, with a standard deviation of 58.0, and the mean scale score for male students is 500, with a standard deviation of 66.0.

The scale score frequency distribution for the total population is shown in Table 189. Figure 67 graphically represents the frequency distributions for the total population and for the groups of female and male students separately. The distributions of the scale scores are slightly negatively skewed with some floor effects for the total population and for each gender.

The mean scale scores for the content standards range from 494 to 500. The mean scale scores for the subcontent areas range from 495 to 512. The median scale scores vary from 505 to 508 for the content standards and from 506 to 508 for the subcontent areas, and all are very close to the median total test scale score of 507.

The mean percentages of the maximum obtainable raw score for the content standards range from 41.7% to 52.9%. The mean percentage of the maximum obtainable raw score for the total test is 49.0%. The mean percentages of the maximum raw score for the subcontent areas range from 44.4% to 54.9%.

Correlations among Content Standards and among Subcontent Areas

Tables 190 through 220 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 moderately to highly correlated, as would be expected.

For the English Reading assessments, the correlations among the various content standards range from 0.59 (in grade 4) to 0.78 (in grade 5). The correlations among the various English Reading subcontent areas range from 0.45 (in grade 9) to 0.77 (in grade 4).

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 0.58 and 0.72, and the correlations among the subcontent areas vary between 0.61 and 0.65.

For the English Writing assessments, the correlation between content standards 2 and 3 range from 0.66 (grade 3) to 0.76 (grade 6). The correlations among the various English Writing subcontent areas vary between 0.378 (grade 6) and 0.67 (grade 6).

For the Spanish Writing assessments, the correlation between content standards 2 and 3 is 0.77 in grade 3 and 0.71 in grade 4. The correlations among the various Spanish Writing subcontent areas range from 0.42 (grade 4) to 0.57 (grade 3).

For the Mathematics assessments, the correlations among content standards range from 0.66 (grade 6) to 0.782 (grade 7). Correlations among the Mathematics subcontent areas range from 0.546 (grade 10) to 0.71 (grade 6).

Finally, for the Science assessments, the correlations among content standards range from 0.65 (grade 8) to 0.78 (grade 8). Correlations among the Science subcontent areas range from 0.40 (grade 10) to 0.72 (grade 8).

Part 8: Reliability and Validity Evidence

Part 8 describes reliability and validity evidence for the 2012 TCAP assessments. First, the total test and subgroup reliability coefficients, measured by Cronbach's alpha, are presented as an index of the internal consistency. This is followed by interrater reliability of constructed-response items, item-to-total score correlations, and items functioning differentially in the TCAP 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 TCAP 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 samples, which ranged from 92% to 100% of the total tested population for all grades and content areas) and are typical of the results obtained for all TCAP 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 to be of sound reliability when their reliability coefficients are in the range of 0.80 and above. Tables 221 and 222 show Cronbach's coefficient alpha for all content standards and subcontent areas. At the state level, the total reliability coefficients for the content areas range between 0.87 (grade 3 Spanish Reading) and 0.95 (grade 4 Mathematics), with a median value of 0.93. Such a reliability coefficient range is indicative of high internal consistency and signifies that the TCAP tests produce relatively stable scores. The median coefficients for each content area and the ranges across grade levels are as follows:

Test	Median	Range
Reading (English)	0.93	(0.89–0.94)
Writing (English)	0.92	(0.91–0.92)
Mathematics	0.94	(0.92–0.95)
Science	0.93	(0.91–0.93)
Reading (Spanish)	0.91	(0.87–0.94)
Writing (Spanish)	0.90	(0.88–0.91)

Table 221 also shows the individual reliability coefficients for content standards at each grade level. Table 222 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 content standard and subcontent area.

As evidence that a test is performing similarly across various subgroups, the reliability values for these subgroups can be compared to those for the total population. 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 223 through 228 show the total test reliability estimates for each content area by disability, accommodation, free lunch eligibility, gender, language proficiency, and immigrant status. Even at the subgroup level, the ranges are generally quite similar. Of the 760 reliability coefficients in Tables 223 through 228, only two are lower than 0.80. These are for migrant/immigrant students on the Grade 10 Mathematics test (alpha = 0.72) and for non-English proficient (NEP) students on the Grade 10 Science test (alpha = 0.73).

The performance of accommodated and non-accommodated students with and without reported disabilities is summarized in Table 229. Overall, non-accommodated students scored higher than accommodated students in every grade and content area except for grade 4 Spanish Writing, where the mean score of students without a disability and without an IEP was slightly higher for accommodated than for non-accommodated students.[14] As shown in the 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.

[14] It should be noted that the small numbers of students taking the Spanish tests make it difficult to draw any meaningful conclusions about group differences.

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 for all grades and content areas except grade 4 Spanish Writing. 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 generally lower in the accommodated groups than in the non-accommodated 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 (e.g., students who need accommodations but are not offered them). Reliability refers to the degree to which students' scores are free from such effects and it 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 handscoring. Tables 11 through 16 provide the results of rater reliability measures, which assess the agreement rates within a given administration, and Table 17 provides the results of rater severity analyses, which compare the scoring leniency across years. As previously mentioned, these results demonstrate that the TCAP tests have relatively high interrater reliability.

As shown in rater reliability Tables 11 through 16, the kappa for English Reading items ranges from 0.49 to 0.90 with a median value of 0.69. English Writing kappa values have a wider range, from 0.37 to 0.97, with a median of 0.80. (The lower kappa values for some writing items are associated with lower maximum score points.) The kappa values for Mathematics items range from 0.56 to 0.95, with a median value of 0.86, and Science items range from 0.46 to 0.93, with a median value of 0.76. On the Spanish versions, kappa ranges from 0.55 to 0.93, with a median of 0.64 for Reading and from 0.34 to 0.92, with a median of 0.57 for Writing.

Table 17 displays the consistency of the ratings assigned to the same papers in 2012 and when they were previously administered. The values of kappa for Reading items range from 0.59 to 0.84, with a median value of 0.67. For Writing items, the range is from 0.52 to 0.77, with a median value of 0.70. For Mathematics items, the range is from 0.75 to 0.97, with a median value of 0.91. For Science items, the range is from 0.72 to 0.88, with a median value of 0.82. 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 is an indication of the relationship between each item and the overall test. As discussed in Part 5 of this report, Tables 23 through 84 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 these two statistics. Item-to-total score correlations are limited by the response distributions, and, therefore, tend to be lower among very easy and very difficult items. 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 test forms range from 0.30 to 0.42 for MC items and from 0.38 to 0.61 for CR items, with average *p*-values from 0.48 to 0.75 and from 0.32 to 0.78, respectively, the item-to-total score correlations and *p*-values are in a reasonable range.

The differential item functioning (DIF) statistic provides a measure of the systematic over- or under-performance of selected subgroups on individual test items. Items exhibiting DIF were avoided as much as possible when operational test forms were created. The TCAP 2012 DIF results are presented in a later section of Part 8.

Standard Error of Measurement

Another measure of reliability is the standard error of measurement (SEM). This statistic is a direct estimate of the degree of measurement error in a student's total score on a test. The SEM 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 the higher the reliability. The SEMs are computed with the following formula:

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

where SD_SS is the standard deviation of the scale score, and $\hat{\alpha}$ is the result of the calculation of Cronbach's alpha. The SEMs represent the total standard error of measurement in the scale score metric. The overall estimates of SEM are shown in Table 230. The scale scores and associated SEMs by content area and grade are shown in Tables 231 through 235. Tables 223 through 228 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 236 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 TCAP 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 TCAP 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, and 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 TCAP assessments. All TCAP 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"[15] 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 237 through 240 show for each content area test the number of score reporting categories (SRCs),[16] 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 TCAP validation process. Evidence for construct validity is comprehensive and integrates evidence from both content- and criterion-related validity. For example, to demonstrate comprehensiveness, TCAP 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 the TCAP does not field test, the quality of the item pool used in the construction of the TCAP 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, while 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.

[15] The *Criteria for Item Acceptability* is a checklist that is used to train item writers and to guide the development and review of items during test construction. A copy of the checklist is provided as an appendix to this report.

[16] These score reporting categories correspond to the Colorado Model Content Standards and subcontent areas listed in Table 1.

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. The TCAP 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 into the performances on which the measurement is based. If these tests require that students have culture-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 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 TCAP 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 bias was minimized was by following specific McGraw-Hill guidelines designed to reduce or eliminate bias. Item writers were directed to the following published guidelines: *Guidelines for bias-free publishing* (McGraw-Hill,

1983) and *Reflecting diversity: Multicultural guidelines for educational publishing professionals* (MacMillan/McGraw-Hill, 1993). Developers reviewed the TCAP 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 independently reviewed by a quality assurance editor.

As part of the standard TCAP test assembly process, items with poor statistical fit, or distractors with positive item-to-total score correlations are avoided insofar as practicable since these item characteristics may indicate that an item is tapping ability irrelevant to the construct being measured. DIF 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 the TCAP, 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 fit and DIF flagged items, including anchor items, in the 2012 test assembly are presented in Table 9.

The third strategy for minimizing bias is to involve educational community professionals who represent various ethnic groups in the review of all new materials. These reviewers are asked to consider and comment on the appropriateness of language, subject matter, and representation of groups of people.

The fourth procedure for minimizing bias involves statistical procedures referred to as DIF analyses to evaluate differential item functioning in all of the TCAP tests. 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 use of items that have been flagged for DIF is minimized in the test development process. DIF studies have been done routinely for all major test batteries published by CTB/McGraw-Hill after 1970. All TCAP test items are analyzed for DIF in subgroups identified by gender, ethnicity, and disabilities.

Because the TCAP 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 2012 TCAP 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 to enable content experts 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

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

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 ten 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

DIF 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[17] were conducted for African Americans, Hispanics, Asians, Native Americans, males, and females. Table 241 provides an overview of items flagged for ethnicity DIF in the various assessments based on the entire student population, and Table 242 presents an overview of items flagged for gender DIF. The results for each assessment are briefly described below.

On the Reading assessments, DIF for gender or ethnicity was observed in every grade except for grade 5. Across all grades, one item favored and two items disfavored American Indian/Alaska Native students; 25 items favored and three items disfavored Asian students; three items favored and one item disfavored African American students; four items favored Hispanic students; four items favored and four items disfavored Hawaiian/Pacific Islander students; two items favored and one item disfavored female students; and one item favored and four items disfavored male students.

On the Writing assessments, DIF for gender or ethnicity was observed in every grade. Across all grades, one item disfavored American Indian/Alaska Native students; two items favored and two items disfavored Asian students; one item favored African American students; two items favored Hispanic students; four items favored and five items disfavored Hawaiian/Pacific Islander students; 10 items favored female students; and 11 items disfavored male students.

On the Mathematics assessments, DIF for gender or ethnicity was observed in every grade except for grades 3 and 7. Across the grades showing DIF, one item favored and one item disfavored American Indian/Alaska Native students; three items favored and three items disfavored Asian students; two items favored and five items disfavored African American students; one item disfavored Hispanic students; and two items favored and two items disfavored Hawaiian/Pacific Islander students; and one item disfavored male students.

On the Science assessments, items exhibited DIF for gender or ethnicity in grades 8 and 10. Three items favored Asian students; two items favored Hispanic students; two items favored and one item disfavored Hawaiian/Pacific Islander students; two items favored female students; and one item disfavored male students.

Additional DIF analyses are presented in Tables 243 (Accommodations), 244 (Primary Disability State), 245 (Enrollment), 246 (Language Proficiency), 247 (Education Plan), and 248 (Homeless, Immigrant, Migrant, and Free Lunch Eligible).

[17] DIF analyses are not reported for the Spanish Reading and Writing assessments because of small case counts and relative homogeneity of the examinees for these tests.

Internal Factor Structure and Unidimensionality of the TCAP 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). TCAP test items are calibrated using a unidimensional IRT model, 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 TCAP assessments, exploratory factor analyses were conducted for each content and grade. Polychoric correlations were obtained, and a principal components analysis was conducted. The resulting eigenvalues for each factor are an indication of the relative proportion of variance accounted for by each successive factor. Figures 68 through 98 contain plots of the eigenvalues (part a) and proportions of variance (part b) for each factor identified in these analyses. These figures show that each of the TCAP tests (English versions) demonstrated a strong single factor, accounting for approximately 28% to 51% 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, ranging from 23% to 35%. However, the number of examinees taking the Spanish tests was so small that the factor analyses should be interpreted with extreme caution.

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 prescribes is relevant to the validation of test scores. As part of the scaling process, all TCAP 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 (fewer than 5%) on the 2012 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 85 through 146.

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 the TCAP tests, scale scores were obtained and correlated for students who took various TCAP content area tests in 2012. Tables 249 and 250 show the intercorrelations among content areas (scale scores and percentile ranks) by grade level. The correlation coefficients among scale scores range from 0.74 (between Reading and Mathematics in grade 3) to 0.86 (between Mathematics and Science in grade 8 and 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 consistently higher than those between Mathematics and Reading and between Mathematics and Writing. It is also interesting to note that Science was 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, 2001).

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 –0.35 to 0.09 (Table 251). The fact that these correlations are generally greater than zero in absolute terms can be attributed to differences in the overall ability of the various groups.

Predictive Validity

Predictive validity is a type of criterion-related 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 252 shows predictive validity coefficients measured as the correlation between test scores for two adjacent years (2011 and 2012) on the basis of a group of students matched on student ID data. Spanish tests are excluded from this table because of the very small number of matched students (N<40).

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 252 indicate strong predictability of test scores between two adjacent years. The validity coefficients (corrected for attenuation) range from 0.83 to 0.96 for all English content areas and grades indicating a high level of prediction from one year to the next.

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-by-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	$P_{1.}$	$P_{2.}$	$P_{3.}$	$P_{4.}$	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. Two 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*$_1$)[18]

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

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

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

examinees' estimated abilities (observed scores) and quadrature points (true scores).

Table 253 (composed of two tables) includes the classification consistency and accuracy measures for TCAP grade 3 Reading. The first table is a contingency table with all three cut scores prepared using the Kolen-Kim method. 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, 0.0409 represents the probability of belonging to “Unsatisfactory” in both the first and second administrations. The 0.0112 represents the probability of belonging to “Unsatisfactory” in the first administration and “Partially 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 will be similar to one another. For example, the absolute differences between the sum values and the corresponding observed scores in Table 253 for the Proficient level are 0.0385 (0.6670 vs. 0.6285) and 0.0345 (0.6670 vs. 0.6325). The largest differences were found in the Advanced level, with absolute differences of 0.1370 (0.0720 vs. 0.2090) and 0.1323 (0.0720 vs. 0.2043).

The second table shows indices for classification consistency and classification accuracy. Each index was described above. The values in “All Cuts” were obtained by applying all three cut points simultaneously during analysis. From Table 253, classification agreement (*P*) for grade 3 Reading is 0.7805, chance probability is 0.4552, kappa is 0.5972, and classification accuracy is 0.8412, when all three cuts were used for computation. Because there are only two levels of classification when only one cut is applied, 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 grade levels and content areas (Tables 253 through 283).

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

The TCAP incorporates the philosophy of multiple measures of a construct. All TCAP assessments are composed of multiple-choice item types. The TCAP 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, Extended Writing, and Grammar and Usage). CR items in the TCAP take different forms and solicit varying response lengths. Compared to other statewide writing assessments—for example, single-prompt extended writing—the TCAP Writing assessment taps into a variety of writing skills using various item formats.

In addition to providing an overall measure of writing ability, the TCAP 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 the other 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 TCAP Technical Advisory Committee (TAC), a study in English Writing was conducted to explore the unstable trends of the Extended Writing subcontent area in 2012. Grade 3 Writing does not include the Extended Writing subcontent area, so the study was conducted on grades 4 through 10. In this study, Extended Writing (SA6) was examined as a single subcontent area and was also combined into a new subcontent area that included the more stable Paragraph Writing (SA 5); 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 11 years (2002 through 2012). The resulting mean and median scale scores are presented in Table 284. Median scores were examined because subcontent scores tend to be affected unduly by extreme scores. Median scale scores are also presented in Figures 99 through 105. Although the median scale scores on the Extended Writing subcontent area differ markedly from scores in the other areas and show much greater fluctuation across years, the combined score on Extended Writing and Paragraph Writing is much more stable. Because of the increased number of items in the combined subcontent area, the stability of the scores across years is improved considerably, and the fluctuations in difficulty are reduced at every grade level. As shown in Figures 99 through 105, the median scores on the combined Extended Writing/Paragraph Writing subcontent area are quite stable across years and very similar to the scores on the total test and on the other subcontent areas.

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 for instruction and assessment.* Second Edition, August 2008.

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 (2001). *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.

McGraw-Hill (1983). *Guidelines for bias-free publishing*. Monterey, CA: Author.

MacMillan/McGraw-Hill (1993). *Reflecting diversity: Multicultural guidelines for educational publishing professionals*. New York, NY: Author.

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. L, 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.

