Colorado
Student Assessment Program

Technical Report
2006

Submitted to the
Colorado Department of Education

November 2006

TABLE OF CONTENTS

PART 1: OVERVIEW OF THE CSAP ASSESSMENTS 4

Reading and Writing: **4**
- The Colorado Model Content Standards 4
- The Colorado Model Sub-Content Areas 4

Mathematics **5**
- The Colorado Model Content Standards 5
- The Colorado Model Sub-Content Areas 6

Science **8**
- The Colorado Model Content Standards 8
- The Colorado Model Sub-Content Areas 8
- Test Development and Content Validity 9

Test Configuration **9**

PART 2: SCALING AND SCORING PROCEDURES 11

Scale Scores for the Total Test and by Content Standard and Sub-Content Area **11**

Scaling Design **11**

PART 3: RESULTS 13

Summary Statistics **13**
- Third Grade 14
- Fourth Grade 17
- Fifth Grade 20
- Sixth Grade 22
- Seventh Grade 24
- Eighth Grade 25
- Ninth Grade 28
- Tenth Grade 29

Correlations Among Content Standards and Among Sub-Content Areas **32**

Test Reliability **33**

PART 4: ITEM ANALYSES 34
- Third Grade 34
- Fourth Grade 36
- Fifth Grade 38
- Sixth Grade 40
- Seventh Grade 41
- Eighth Grade 43

Ninth Grade 44
Tenth Grade 46

PART 5: SCALING AND CALIBRATION 48

Overview of the IRT Models 48

Calibration of the Assessment 49

Model Fit Analyses 49
Model Fit Analyses Results 51

Item Independence 53

Equating Procedures 53
Anchor Set Review Process 55
P-Value Comparisons 55
Item Parameter Comparisons 56
Test Characteristic Curve Comparisons 56
Scaling Constants 56
Additional Analyses of Flagged Items 56

Effectiveness of the Equating 57

PART 6: TOTAL AND SUBGROUP RELIABILITY 58

PART 7: TEST VALIDITY 61

Content-Related Validity 61

Construct Validity 62

Minimization of Construct-irrelevant Variance and Under-representation 62

Minimizing Bias through Differential Item Functioning 62

Linn-Harnisch Differential Item Functioning Analyses (DIF) procedure 65

Differential Item Functioning Ratings 66

Results of the Differential Item Functioning Analyses 67

Internal Structure and Unidimensionality 68

Divergent (Discriminant) Validity 69

Predictive Validity 69

REFERENCES 71

This report presents the results of the statewide Spring 2006 administration of the Colorado Student Assessment Program (CSAP). In the Spring of 2006, students were assessed in Reading grades 3 through 10; Writing grades 3 through 10; Mathematics grades 3 through 10; and Science grades 5, 8, and 10. Spanish versions of Reading and Writing were also administered in grades 3 and 4. The assessments were developed by CTB/McGraw-Hill in collaboration with the Colorado Department of Education and were scored and scaled by CTB/McGraw-Hill.

The report is organized in parts with Part 1 providing an overview of the CSAP assessments including descriptions of the content standards and sub-content areas, test development, and test configuration. Part 2 includes descriptions of scaling and scoring procedures. Summary statistic, correlation, and test reliability results are presented in Part 3. Detailed item analysis results including item-to-total-score correlations, p-values, and omit rates are included in Part 4. Part 5 provides an overview of the item response theory models; calibration procedures; model fit analyses; and a discussion of the equating procedure including anchor set reviews, p-value, parameter, and test characteristic curve comparisons, scaling constants, and additional analyses of flagged items. Total and subgroup reliability and test validity are discussed in Parts 6 and 7, respectively. This report also includes nine appendices. Appendix A provides scale score distribution and plots. Appendix B documents test validity. Appendix C provides classification consistency results. Appendix D includes rater reliability and severity study results. Appendix E includes equating figures. Appendix F provides a study to explore the trends of the Extended Writing sub-content area. Appendix G includes total and subgroup reliability information. Appendix H describes factor analysis figures. Appendix I provides a summary of Fit and DIF as related to test assembly.

Part 1: Overview of the CSAP Assessments

The CSAP assessments are developed to measure the Colorado content standards, which are listed below. Note that the terms "content standard" and "standard" are used synonymously throughout the text. Beginning in 2001, reporting categories were added, at the request of the Colorado Department of Education, to provide additional diagnostic information; these sub-content areas are listed below as well. Each sub-content area may cover several content standards. Most, but not all, of the items in CSAP are assigned to a sub-content area, whereas all items in CSAP are assigned to one, and only one, content standard. The various content standards and sub-content areas are listed below for each content area. Table 1 gives an overview of which content standards and sub-content 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 their reading, writing, speaking, listening, and viewing. (Reading)

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

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

The Colorado Model Sub-Content 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 Sub-Content Areas

1. Number and Operation Sense –

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

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

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

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

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

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

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

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

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

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

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

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

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

Science

The Colorado Model Content Standards

1. Scientific Investigation and Connections Among Scientific Disciplines – Student understands the processes of scientific investigation and design, conducting and evaluating, as well as communicating about, such investigations. Student understands that science involves making connections among disciplines.

2. Physical Science and Its Interrelationship with Technology and Human Activity – Student knows and understands common properties, forms, and changes in matter and energy, as well as interrelationships among physical science, technology, and human activity.

3. Life Science and Its Interrelationship with Technology and Human Activity – Student knows and understands the characteristics and structure of living things, the processes of life, how living things interact with each other and their environment, as well as interrelationships among life science, technology, and human activity.

4. Earth and Space Science and Its Interrelationship with Technology and Human Activity – Student knows and understands the processes and interactions of Earth's systems and the structure and dynamics of Earth and other objects in space, as well as interrelationships among earth and space science, technology, and human activity.

The Colorado Model Sub-Content Areas

1. Experimental Design & Investigations – Student understands and applies scientific questions, hypotheses, variables, and experimental design.

2. Results & Data Analysis – Student organizes, analyzes, interprets, and predicts from scientific data in order to communicate the results of investigations.

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

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

5. Life Processes – Student understands the structure and life processes of organisms.

6. Organisms & Their Interactions – Student uses an understanding of food webs/chains, adaptations, and nonliving factors to explain how organisms interact within ecosystems.

7. Geology and Astronomy – Student understands processes that shape Earth and the structure and interaction of objects in the solar system.

8. Meteorology and Hydrology – Student understands the water cycle and factors affecting the weather.

Test Development and Content Validity

In order to assure the content validity of the CSAP assessments, the Colorado Model Content Standards and Assessment Frameworks were studied by CTB's Content Developers. To develop the 2006 Colorado Student Assessment Program, Colorado content area specialists, teachers, and assessment experts worked with CTB/McGraw-Hill to develop a pool of items that measured Colorado's Assessment Frameworks in each grade and content area. Several sources contributed to the 2006 CSAP items. CTB/McGraw-Hill's extensive pool of previously field-tested reading passages, writing prompts, mathematics and science items provided the initial source. Many of these existing items were revised in order to ensure better measurement of the relevant Colorado standard and benchmark. Additional items were developed by CTB and the staff at the Colorado Department of Education, as needed to complete the alignment of CSAP to the Assessment Frameworks. These items were carefully reviewed and discussed by Content Review, Bias Review, Community Sensitivity Review, and Instructional Impact committees to assure not only content validity, but also the quality and appropriateness of the items. These committees represented Colorado's diverse population and included Colorado teachers, community members, and State Department of Education staff. The committees' recommendations were used to select and/or modify items from the item pool to construct the final Reading, Writing, Mathematics, and Science assessments.

Each new form also included a subset of multiple choice items used in the previous administrations of the CSAP assessments. 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 comparability across years. Details of the equating are provided later in this document. The assessments that were reported on vertical scales (English Reading, English Writing, and Mathematics) also had items in common between adjacent grades.

Test Configuration

Tables 2 through 6 provide information regarding the configuration of the CSAP assessments. Table 2 provides the number of multiple-choice (MC) and constructed-response (CR) items on each test, as well as the number of

obtainable points on each CR item. Tables 3 through 6 provide the number of MC and CR items by content standard (CS) and sub-content area (SA). Note that the sub-content 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, & Functions (CS 2) and Statistics & Probability (CS 3), in Mathematics grade 3; Number Sense (CS 1) and Computational Techniques (CS 6) in Mathematics, grades 7 through 10; Geometry (CS 4) and Measurement (CS 5) in Mathematics, grades 3 through 10; Scientific Investigations and Connections Among Scientific Disciplines (CS 1/6), Physical Science and Its Interrelationship with Technology & Human Activity (CS 2/5), Life Science and Its Interrelationship with Technology & Human Activity (CS 3/5), Earth and Space Science and Its Interrelationship with Technology & Human Activity (CS 4/5) in Science grades 5, 8 and 10.

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

Across all grades and content areas, nine items (eight multiple choice items, and one constructed-response item) were removed from calibration:

- Writing , Grade 5 – Item 60
- Writing, Grade 8 – Item 3C
- Writing, Grade 8 – Item 63
- Mathematics, Grade 9 – Item 57
- Mathematics, Grade 10 – Item 41
- Science, Grade 8 – Item 7
- Science, Grade 8 – Item 18
- Science, Grade 10 – Item 13
- Science, Grade 10 – Item 20

Tables 2--6 indicates the number of items and score points for each test form with suppressed items removed. Part 5 includes an additional discussion of the test blueprint with a description of item and anchor counts by content standard (Tables 145—149).

Part 2: Scaling and Scoring Procedures

Scale Scores for the Total Test and by Content Standard and Sub-Content Area

Students' total scale scores are based on their performance on all the scored items on the test. The range of possible scores varies by grade and content area. The highest obtainable scale score (HOSS) and lowest obtainable scale score (LOSS) for each grade and content area is provided in Table 7. Students also receive a score for each content standard (and for each sub-content area) that is based only on the items that contribute to the given content standard (or sub-content area). Note that every item on the test corresponds to some content standard but not all items contribute to a sub-content area. The scale scores for the content standards and the sub-content areas are calculated using the item parameters that are obtained when the *total* test is calibrated (see Part 5, Scaling and Calibration). For each grade and content area, the minimum and maximum possible scale scores for content standards and sub-content areas are set at the same LOSS and HOSS as the total scale score.

Students were scored at the total test, content standard, and sub-content area levels using item response theory pattern scoring procedures. This procedure produces maximum-likelihood trait estimates (scale scores) based on students' item response patterns, as described by Lord (1974; 1980, pp. 179-181). Item-pattern scoring takes more information into account and is more accurate than number-correct scoring in which all students with the same number correct receive the same score, regardless of how that score is obtained. On average, the increase in accuracy is equivalent to approximately a 15-20% increase in test length (Yen, 1984; Yen & Candell, 1991). Note that score reliability tends to increase with the number of items, and thus the total score is more reliable than the content standard or sub-content area scores.

Scaling Design

Horizontal equating within each grade was used to place the 2006 forms on the vertical scales that had been established previously for English Reading, Writing, and Mathematics. The vertical scale for English Reading, spanning grades 3 through 10, was established in 2001. The vertical scales for English Writing, spanning grades 3 through 10, and for Mathematics, spanning grades 5 through 10, were established in 2002. Grades 3 and 4 Mathematics were added to the vertical scale in 2005. The Stocking and Lord (1983) procedure was used to place each grade on the vertical scale that had been developed for each content area.

In 2006, two new assessments, Science grades 5 and 10, were introduced. Grade 10 Science was scaled to have the same mean and SD as the 2005 Grade 8 Science assessment. The target mean for Grade 5 was shifted to 550 in order to better position the scale in relation to the preset LOSS and HOSS values, i.e., 300 and 790, respectively. The Grade 5 SD was also scaled down to around 55 from the initial target of 60, to better fit the relatively smaller spread of scores observed in raw scores for Grade 5, compared to Grades 8 and 10. Due to the non-incremental nature of the content standards and the gaps in grade levels, a vertical scale in Science was not established in 2006.

Note that the customized versions of the reading and writing assessments in Spanish Grades 3 and 4 were first administered in 1998. The year before, Supera had been administered to those students eligible for taking a Spanish language version assessment. The customized Spanish version that was first created in 1998 was repeated as is through 2001. In 2002 new forms were created for the Spanish language assessments, which served as a source for the future tests. Every year a new form has been created to meet the Colorado blueprint by selecting psychometrically good quality items from the existing item pool. Although grades 3 and 4 Spanish tests are designed to measure student's developmental scale over time, they are not in vertical scale.

Each 2006 CSAP test contained at least 20 multiple choice items from the previous administrations for the same grade. These repeated multiple choice items served as anchors in the Stocking and Lord (1983) equating procedure, which was used to place each test form on the previously established scale. By equating the 2006 tests within each grade, the unique metrics of the CSAP Reading, Writing, and Mathematics vertical scales were maintained.

These scaling and calibration methods are presented in Part 5 of this report.

Part 3: Results

Student results are reported statewide in terms of scale scores and performance levels. The scale score ranges for each grade and content area are listed in Table 7.

The performance level cut scores were adopted by the Colorado State Board of Education, based on the recommendations of standard setting committees composed of qualified Colorado educators, using a variation of the Bookmark standard setting procedure (Lewis, Mitzel, & Green, 1996). The performance standards for Reading were adopted from the 2001 standard setting. The performance standards for Writing and Mathematics were adopted from 2002 standard setting except for grades 3 and 4 Mathematics. The grades 3 and 4 Mathematics assessments were introduced in 2005 and standards were set in the same year. Similarly, performance standards for grade 8 Science were set in 2000, and performance standards for grades 5 and 10 Science were set in 2006. Detailed information about the cut scores and standard setting for the new assessments implemented in 2006 is available in the CSAP Bookmark Standard Setting Technical Report for Grades 5 & 10 Science (2006).

Summary Statistics

Summary statistics are based on the total Colorado student population tested by CSAP. Table 8 presents the mean, median, and standard deviation of the scale scores for the total population and each gender in each grade/content area. Note that the male and female students do not equal the total population because some students did not identify their gender.

Female students scored higher than male students at all grade levels on the Reading and Writing tests, while male students scored slightly higher than female students at all grade levels on the Science assessments. Male students scored one point higher than female students on the Mathematics tests in grades 5, 8, and 9 but both had equivalent scores at grades 6 and 10. In the remaining grade spans, male students scored three points higher in grade 3, five points higher in grade 4 and female students scored one point higher in grade 7.

Tables 9 and 10 contain scale score descriptive statistics for each content standard and sub-content area, respectively. Since the scale scores for content standards and sub-content areas are computed based on 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 11 and 12 contain number-correct descriptive statistics for the total population and the mean percent of the maximum points obtained, for each content standard and sub-content area, respectively.

Note the following particulars: grade 3 Reading measures only one content standard; content standards 2 and 3 are combined for grade 3 Mathematics; content standards 1 and 6 are combined in grades 7 through 10 Mathematics; content standards 4 and 5 are combined in grades 3 through 10 Mathematics; and content standards 1 and 6, 2 and 5, 3 and 5, and 4 and 5 are combined for grades 5, 8 and 10 Science. Similarly, sub-content areas 1 and 4 are combined for grades 3 through 6 Reading. In Tables 9-12, where a content standard or sub-content area is shared (e.g. CS 2/3 for grade 3 Mathematics) the scores are reported under the first CS or SA (e.g. CS 2 for grade 3 Mathematics).

Third Grade

Reading

The mean scale score for the total population of students taking the 2006 third-grade Reading assessment is 554 with a standard deviation of 76.8. The mean scale score for female students is 561 with a standard deviation of 72.9, and the mean scale score for male students is 547 with a standard deviation of 79.9.

The scale score frequency distribution of the third-grade Reading assessment for the total population is shown in Table A-1[1]. Figure A-1 graphically represents the scale score frequency distributions for the total population and for the groups of male and female students separately. The figure shows that the distributions of scale scores for the total population and for each gender are slightly negatively skewed.

The mean scale score for the single content standard is 554, with a standard deviation of 76.8. The mean scale scores for the sub-content areas range from 550 to 565. Although two of the sub-content area scale score medians are within one point of the total test score median of 562 (SA 1 and SA 2), the median score on sub-content area 3 is slightly higher at 568.

The mean percents of the maximum obtainable raw score for the sub-content areas range from 59.4 to 68.8. The mean percent of the maximum obtainable score for the total test is 62.8.

Reading – Spanish Version

The mean scale score for the total population of students taking the 2006 third-grade Spanish Reading assessment is 523 with a standard deviation of 45.8. The mean scale score for female students is 531 with a standard deviation of

[1] All tables and figures referenced in this section (A-1 to A-31) can be found in Appendix A.

41.7, and the mean scale score for male students is 515 with a standard deviation of 48.4.

The scale score frequency distribution of the third-grade Spanish Reading assessment for the total population is shown in Table A-2. Figure A-2 graphically represents the scale score frequency distributions for the total population and for the groups of male and female students separately. The figure shows that the distributions of scale scores for the total population and for each gender are slightly negatively skewed.

The mean scale score for the single content standard is 523, with a standard deviation of 45.8.The mean scale scores for the sub-content areas range from 522 to 525; the median scale scores for the sub-content areas vary between 525 and 526, and all are close to the median for the total test scale score of 524.

The mean percents of the maximum obtainable score for the sub-content areas range from 52.8 to 64.1. The mean percent of the maximum obtainable score for the total test is 59.2.

Writing

The mean scale score for the total population of students taking the 2006 third-grade Writing assessment is 468 with a standard deviation of 53.7. The mean scale score for female students is 476 with a standard deviation of 53.8, and the mean scale score for male students is 460 with a standard deviation of 52.7.

The scale score frequency distribution for the total population is shown in Table A-3. Figure A-3 graphically represents the scale score frequency distributions for the total population and for the groups of male and female students separately. The figure shows that the distributions of scale scores for the total population and for each gender are slightly negatively skewed.

The mean scale scores for the two content standards are 470 and 473, with standard deviations of 60.8 and 63.7, respectively. The mean scale scores for the sub-content areas range from 472 to 495. The median scale scores vary between 469 and 467 for the content standards, and between 471 and 475 for the sub-content areas. The median for the total test scale score is 467.

The mean percents of the maximum obtainable score for CS 2 (Write for a Variety of Purposes) and CS 3 (Write Using Conventions) are 76.2 and 80.1, respectively. The mean percents of the maximum obtainable score for the sub-content areas range from 76.6 to 80.6. The mean percent of the maximum obtainable score for the total test is 78.4.

Writing – Spanish Version

The mean scale score for the total population of students taking the 2006 third-grade Spanish Writing assessment is 497 with a standard deviation of 64.4. The mean scale score for female students is 510 with a standard deviation of 61.3 and the mean scale score for male students is 484 with a standard deviation of 65.1.

The scale score frequency distribution of the third-grade Spanish Writing assessment for the total population is shown in Table A-4. Figure A-4 graphically represents the scale score frequency distributions for the total population and for the groups of male and female students separately. The figure shows that the scale score distributions for the total population and for each gender are approximately normal.

The mean scale scores for the two content standards are 497, and 502 respectively with median scale scores of 497 and 499. The mean scale scores for the sub-content areas range from 503 to 509; the median scale scores for the sub-content areas vary between 496 and 504. The median total scale score is 498.

The mean percents of the maximum obtainable score ranges from 62.1 for CS2 (Write for a Variety of Purposes) to 70.9 for CS3 (Write Using Conventions), and from 60.7 to 72.1 for the sub-content areas. The mean percent of the maximum obtainable score for the total test is 67.2.

Mathematics

The mean scale score for the total population of students taking the 2006 third-grade Mathematics assessment is 464 with a standard deviation of 89.3. The mean scale score for female students is 463 with a standard deviation of 87.9, and the mean scale score for male students is 466 with a standard deviation of 90.7.

The scale score frequency distribution for the total population is shown in Table A-5. Figure A-5 graphically represents the frequency distributions for the total population and for the groups of male and female students separately. The figure shows that the scale score distributions for the total population and for each gender are approximately normal, with a small group of students located at the HOSS (Highest Obtainable Scale Score).

The mean scale scores for the content standards range from 474 to 486. The median scale score is between 460 and 478 for the content standards. The median for the total test scale score is 464.

The mean percents of the maximum obtainable score for the content standards range from 70.1 on CS 6 (Operation and Calculation) to 76.1 on CS 2 (Algebra,

Patterns, and Functions). The mean percent of the maximum obtainable score for the total test is 74.2.

Fourth Grade

Reading

The mean scale score for the total population of students taking the 2006 fourth-grade Reading assessment is 589 with a standard deviation of 62.0. The mean scale score for female students is 596 with a standard deviation of 58.3, and the mean scale score for male students is 583 with a standard deviation of 64.7.

The scale score frequency distribution for the total population is shown in Table A-6. Figure A-6 graphically represents the scale score frequency distributions for the total population and for the groups of male and female students separately. The figure shows that the distributions of scale scores for the total population and for each gender are slightly negatively skewed.

The mean scale scores for the content standards range from 584 to 593. The mean scale scores for the sub-content areas range from 590 to 644. The median scale scores vary between 595 and 600 for the content standards, and between 596 and 597 for the sub-content areas. The median for the total test scale score is 597.

The mean percents of the maximum obtainable score for the content standards range from 55.1 on CS 4 (Thinking Skills) to 72.7 on CS 1 (Reading Comprehension). The mean percent of the maximum obtainable score for the total test is 62.6. The mean percents for the sub-content areas range from 61.7 to 75.9.

Reading – Spanish Version

The mean scale score for the total population of students taking the 2006 fourth-grade Spanish Reading assessment is 521 with a standard deviation of 46.9. The mean scale score for female students is 524 with a standard deviation of 48.0, and the mean scale score for male students is 518 with a standard deviation of 45.8.

The scale score frequency distribution for the total population is shown in Table A-7. Figure A-7 graphically represents the scale score frequency distributions for the total population and for the groups of male and female students separately. The figure shows that the distributions of scale scores for the total population and for each gender are slightly negatively skewed.

The mean scale scores for the content standards range from 514 to 521. The mean scale scores for the sub-content areas range from 517 to 527. The median scale scores vary between 520 and 529 for the content standards, and between 525 and 533 for the sub-content areas, and all are close to the median for the total test scale score, 526.

The mean percents of the maximum obtainable score for the content standards range from 43.6 on CS 6 (Literature) to 59.0 on CS 1 (Reading Comprehension). The mean percent of the maximum obtainable score for the total test is 51.5. The mean percents for the sub-content areas range from 48.5 to 64.4.

Writing

The mean scale score for the total population of students taking the 2006 fourth-grade Writing assessment is 484 with a standard deviation of 51.8. The mean scale score for female students is 493 with a standard deviation of 52.2, and the mean scale score for male students is 476 with a standard deviation of 49.9.

The scale score frequency distribution for the total population is shown in Table A-8. Figure A-8 graphically represents the scale score frequency distributions for the total population and for the groups of male and female students separately. The figure shows that the scale score distributions for the total population and for each gender are approximately normal.

The mean scale scores for the content standards vary between 486 and 488. The mean scale scores for the sub-content areas range from 486 to 510. The median scale scores are 485 for the content standards, and between 486 and 495 for the sub-content areas. The median for the total test scale score is 485.

The mean percents of the maximum obtainable score for CS 2 (Write for a Variety of Purposes) and CS 3 (Write Using Conventions) are 70.6 and 72.9, respectively. The mean percent of the maximum obtainable score for the total test is 71.7. The mean percents of the maximum obtainable score for the sub-content areas range from 59.2 to 80.6

Writing – Spanish Version

The mean scale score for the total population of students taking the 2006 fourth-grade Spanish Writing assessment is 499 with a standard deviation of 50.3. The mean scale score for female students is 506 with a standard deviation of 47.8, and the mean scale score for male students is 493 with a standard deviation of 52.0.

The scale score frequency distribution for the total population is shown in Table A-9. Figure A-9 graphically represents the scale score frequency distributions for

the total population and for the groups of male and female students separately. The figure shows that the distributions of scale scores for the total population and for each gender are slightly negatively skewed.

The mean scale score for each of the two content standards ranges from 489 to 504. The mean scale scores for the sub-content areas range from 465 to 512. The median scale scores for the two content standards are 498 and 510. The median scale scores for the sub-content areas vary between 490 and 511. The median for the total test scale score is 505.

The mean percents of the maximum obtainable score for CS 2 (Write for a Variety of Purposes) and CS 3 (Write Using Conventions) are 41.7 and 52.5, respectively. The mean percent of the maximum obtainable score for the total test is 47.2. The mean percents of the maximum obtainable score for the sub-content areas range from 30.2 to 58.1.

Mathematics

The mean scale score for the total population of students taking the 2006 fourth - grade Mathematics assessment is 489 with a standard deviation of 75.8. The mean scale score for female students is 486 with a standard deviation of 75.1, and the mean scale score for male students is 491 with a standard deviation of 76.3.

The scale score frequency distribution for the total population is shown in Table A-10. Figure A-10 graphically represents the frequency distributions for the total population and for the groups of male and female students separately. The figure shows that the scale score distributions for the total population and for each gender are approximately normal.

The mean scale scores for the content standards range from 491 to 515. The mean scale scores for the sub-content areas range from 492 to 507. The median scale scores vary between 492 and 494 for the content standards, and between 485 and 492 for the sub-content areas. The median for the total test scale score is 492.

The mean percents of the maximum obtainable score for the content standards range from 61.9 on CS 1 (Number Sense) to 78.4 on CS 2 (Algebra, Patterns, & Functions). The mean percent of the maximum obtainable score for the total test is 71.7. The mean percents for the sub-content areas range from 64.4 to 76.3.

Fifth Grade

Reading

The mean scale score for the total population of students taking the 2006 fifth-grade Reading assessment is 612 with a standard deviation of 70.3. The mean scale score for female students is 620 with a standard deviation of 66.6 and the mean scale score for male students is 605 with a standard deviation of 72.9.

The scale score frequency distribution for the total population is shown in Table A-11. Figure A-11 graphically represents the frequency distributions for the total population and for the groups of male and female students separately. The figure shows that the scale score distributions for the total population and for each gender are slightly negatively skewed.

The mean scale scores for the content standards range from 612 to 622. The mean scale scores for the sub-content areas range from 613 to 658. The median scale scores vary between 619 and 622 for the content standards, and between 620 and 621 for the sub-content areas, and all are close to the median for the total test scale score, 621.

The mean percents of the maximum obtainable score for content standards range from 52.2 on CS 6 (Literature) to 74.9 on CS 1 (Reading Comprehension). The mean percent of the maximum obtainable score for the total test is 64.4. The mean percents for the sub-content areas range from 61.8 to 75.6.

Writing

The mean scale score for the total population of students taking the 2006 fifth-grade Writing assessment is 511 with a standard deviation of 60.1. The mean scale score for female students is 521 with a standard deviation of 60.5, and the mean scale score for male students is 501 with a standard deviation of 58.0.

The scale score frequency distribution for the total population is shown in Table A-12. Figure A-12 graphically represents the scale score frequency distributions for the total population and for the groups of male and female students separately. The figure shows that the scale score distributions for the total population and for each gender are approximately normal.

The mean scale scores for the content standards vary between 513 and 515. The mean scale scores for the sub-content areas range from 513 to 533. The median scale scores vary between 511 and 512 for the content standards, and between 484 and 513 for the sub-content areas. Most are close to the median for the total test scale score, 512.

The mean percents of the maximum obtainable score for CS 2 (Write for a Variety of Purposes) and CS 3 (Write Using Conventions) are 68.3 and 69.5, respectively. The mean percent of the maximum obtainable score for the total test is 68.9. The mean percents of the maximum obtainable score for the sub-content areas range from 62.2 to 71.6.

Mathematics

The mean scale score for the total population of students taking the 2006 fifth-grade Mathematics assessment is 520 with a standard deviation of 74.5. The mean scale score for female students is 519 with a standard deviation of 72.6, and the mean scale score for male students is 520 with a standard deviation of 76.2.

The scale score frequency distribution for the total population is shown in Table A-13. Figure A-13 graphically represents the frequency distributions for the total population and for the groups of male and female students separately. The figure shows that the scale score distributions for the total population and for each gender are approximately normal.

The mean scale scores for the content standards range from 528 to 539. The mean scale scores for the sub-content areas range from 533 to 561. The median scale scores vary between 521 and 524 for the content standards, and between 520 and 523 for the sub-content areas. The median for the total test scale score is 521.

The mean percents of the maximum obtainable score for the content standards range from 68.6 on CS 4/5 (Geometry and Measurement) to 75.4 on CS 6 (Computational Techniques). The mean percent of the maximum obtainable score for the total test is 72.7. The mean percents for the sub-content areas range from 72.3 to 76.6.

Science

The mean scale score for the total population of students taking the 2006 fifth-grade Science assessment is 548 with a standard deviation of 56.6. The mean scale score for female students is 546 with a standard deviation of 55.3, and the mean scale score for male students is 550 with a standard deviation of 57.6.

The scale score frequency distribution for the total population is shown in Table A-14. Figure A-14 graphically represents the frequency distributions for the total population and for the groups of male and female students separately. The distributions of the scale scores are approximately normal for the total population and for each gender.

The mean scale scores for the content standards range from 550 to 554. The mean scale scores for the sub-content areas range from 552 to 570. The median scale scores vary between 551 and 552 for the content standards and between 551 and 553 for the sub-content areas, and most are very close to the median for the total test scale score, 552.

The mean percents of the maximum obtainable score for the content standards range from 65.0 on CS 1/6 (Scientific Investigations and Connections Among Scientific Disciplines) to 71.4 on CS 4/5 (Earth and Space Science and Its Interrelationship with Technology and Human Activity). The mean percent of the obtainable score for the total test is 67.7. The mean percents of the maximum obtainable score for the sub-content areas range from 60.3 to 73.7.

Sixth Grade

Reading

The mean scale score for the total population of students taking the 2006 sixth-grade Reading assessment is 623 with a standard deviation of 66.5. The mean scale score for female students is 630 with a standard deviation of 61.9, and the mean scale score for male students is 616 with a standard deviation of 70.0.

The scale score frequency distribution for the total population is shown in Table A-15. Figure A-15 graphically represents the frequency distributions for the total population and for the groups of male and female students separately. The figure shows that the scale score distributions for the total population and for each gender are slightly negatively skewed.

The mean scale scores for the content standards range from 618 to 631. The mean scale scores for the sub-content areas range from 623 to 631. The median scale scores vary between 630 and 631 for the content standards, and between 629 and 634 for the sub-content areas, and all are close to the median for the total test scale score, 631.

The mean percents of the maximum obtainable score for content standards range from 49.8 on CS 6 (Literature) to 69.6 on CS 4 (Thinking Skills). The mean percent of the maximum obtainable score for the total test is 63.3. The mean percents for the sub-content areas range from 60.1 to 64.7.

Writing

The mean scale score for the total population of students taking the 2006 sixth-grade Writing assessment is 525 with a standard deviation of 64.1. The mean

scale score for female students is 539 with a standard deviation of 62.6, and the mean scale score for male students is 512 with a standard deviation of 62.9.

The scale score frequency distribution for the total population is shown in Table A-16. Figure A-16 graphically represents the scale score frequency distributions for the total population and for the groups of male and female students separately. The figure shows that the scale score distributions for the total population and for each gender are approximately normal.

The mean scale scores for the content standards vary between 527 and 528. The mean scale scores for the sub-content areas range from 528 to 542. The median scale scores range from 526 to 528 for the content standards and between 490 and 527 for the sub-content areas. The median for the total test scale score is 527.

The mean percents of the maximum obtainable score for CS 2 (Write for a Variety of Purposes) and CS 3 (Write Using Conventions) are 65.0 and 67.7, respectively. The mean percent of the maximum obtainable score for the total test is 66.3. The mean percents of the maximum obtainable score for the sub-content areas range from 62.0 to 70.5.

Mathematics

The mean scale score for the total population of students taking the 2006 sixth-grade Mathematics assessment is 529 with a standard deviation of 76.1. The mean scale score for female students is 529 with a standard deviation of 73.6, and the mean scale score for male students is 529 with a standard deviation of 78.4.

The scale score frequency distribution for the total population is shown in Table A-17. Figure A-17 graphically represents the frequency distributions for the total population and for the groups of male and female students separately. The figure shows that the distributions of scale scores for the total population and for each gender are approximately normal.

The mean scale scores for the content standards range from 525 to 548. The mean scale scores for sub-content areas range from 530 to 537. The median scale scores vary between 532 and 538 for the content standards, and between 533 and 535 for the sub-content areas, and all are close to the median for the total test scale score, 533. The mean percents of the maximum obtainable score for the content standards range from 53.5 on CS 1 (Number Sense) to 74.4 on CS 3 (Statistics and Probability).

The mean percent of the maximum obtainable score for the total test is 65.1. The mean percents of the maximum obtainable score for the sub-content areas range from 60.2 to 61.7.

Seventh Grade

Reading

The mean scale score for the total population of students taking the 2006 seventh-grade Reading assessment is 635 with a standard deviation of 67.1. The mean scale score for female students is 644 with a standard deviation of 62.9, and the mean scale score for male students is 627 with a standard deviation of 69.9.

The scale score frequency distribution for the total population is shown in Table A-18. Figure A-18 graphically represents the frequency distributions for total population and for the groups of male and female students separately. The figure indicates that the distribution of the scale scores for the total population and for each gender is slightly negatively skewed.

The mean scale scores for the content standards range from 632 to 644. The mean scale scores for the sub-content areas range from 632 to 653. The median scale scores vary between 641 and 644 for the content standards, and between 642 and 645 for the sub-content areas, and all are close to the median for the total test scale score, 642.

The mean percents of the maximum obtainable score for the content standards range from 53.0 on CS 1 (Reading Comprehension) to 68.0 on CS 5 (Use of Literary Information). The mean percent of the maximum obtainable score for the total test is 60.9. The mean percents of the maximum obtainable score for the sub-content areas range from 53.5 to 71.6.

Writing

The mean scale score for the total population of students taking the 2006 seventh-grade Writing assessment is 547 with a standard deviation of 71.6. The mean scale score for female students is 561 with a standard deviation of 69.6, and the mean scale score for male students is 534 with a standard deviation of 71.0.

The scale score frequency distribution for the total population is shown in Table A-19. Figure A-19 graphically represents the frequency distributions for the total population and for the groups of male and female students separately. The figure indicates that the scale score distributions are approximately normal for the total population and for each gender.

The mean scale scores for the content standards vary between 548 and 551. The mean scale scores for the sub-content areas range from 549 to 577. The median scale scores vary between 549 and 550 for the content standards, and between 548 and 577 for the sub-content areas, and most

are close to the median for the total test scale score, 549.

The mean percents of the maximum obtainable score for CS 2 (Write for a Variety of Purposes) and CS 3 (Write Using Conventions) are 63.8 and 66.6, respectively. The mean percent of the maximum obtainable score for the total test is 65.1. The mean percents of the maximum obtainable score for the sub-content areas range from 61.0 to 76.3.

Mathematics

The mean scale score for the total population of students taking the 2006 seventh-grade Mathematics assessment is 544 with a standard deviation of 75.8. The mean scale score for female students is 544 with a standard deviation of 72.7. The mean scale score for male students is 543 with a standard deviation of 78.5.

The scale score frequency distribution for the total population is shown in Table A-20. Figure A-20 graphically represents the frequency distributions for the total population and for the groups of male and female students separately. The figure indicates that the scale score distributions are approximately normal for the total population and for each gender.

The mean scale scores for the content standards range from 541 to 546. The mean scale scores for the sub-content areas range from 528 to 549. The median scale scores vary between 548 and 552 for the content standards, are 549 for the sub-content areas, and all are close to the median for the total test scale score, 549.

The mean percents of the maximum obtainable score for the content standards range from 47.4 on CS 4/5 (Geometry and Measurement) to 59.0 on CS 1/6 (Number Sense). The mean percent of the maximum obtainable score for the total test is 54.5. The mean percents of the maximum obtainable score for the sub-content areas range from 34.7 to 53.1.

Eighth Grade

Reading

The mean scale score for the total population of students taking the 2006 eighth-grade Reading assessment is 650 with a standard deviation of 65.1. The mean scale score for female students is 658 with a standard deviation of 60.9, and the mean scale score for male students is 641 with a standard deviation of 67.9.

The scale score frequency distribution for the total population is shown in Table A-21. Figure A-21 graphically represents the frequency distributions for the total

population and for the groups of male and female students separately. The figure shows that the scale score distributions for the total population and for each gender are slightly negatively skewed.

The mean scale scores for the content standards range from 649 to 653. The mean scale scores for the sub-content areas range from 635 to 682. The median scale scores vary between 656 and 660 for the content standards, and between 656 and 663 for the sub-content areas, and all are close to the median for the total test scale score, 658.

The mean percents of the maximum obtainable score for the content standards range from 54.7 on CS 6 (Literature) to 64.9 on CS 4 (Thinking Skills). The mean percent of the maximum obtainable score for the total test is 58.1. The mean percents of the maximum obtainable score for the sub-content areas range from 42.2 to 73.0.

Writing

The mean scale score for the total population of students taking the 2006 eighth-grade Writing assessment is 556 with a standard deviation of 74.1. The mean scale score for female students is 574 with a standard deviation of 71.3, and the mean scale score for male students is 539 with a standard deviation of 72.6.

The scale score frequency distribution for the total population is shown in Table A-22. Figure A-22 graphically represents the frequency distributions for the total population and for the groups of male and female students separately. The figure indicates that the scale score distributions are approximately normal for the total population and for each gender.

The mean scale scores for the content standards vary between 556 and 557. The mean scale scores for the sub-content areas range from 557 to 561. The median scale scores vary between 557 and 560 for the content standards, and between 552 and 561 for the sub-content areas, and most are close to the median for the total test scale score, 559.

The mean percents of the maximum obtainable score for CS 2 (Write for a Variety of Purposes) and CS 3 (Write Using Conventions) are 66.4 and 59.7, respectively. The mean percent of the maximum obtainable score for the total test is 63.3. The mean percents of the maximum obtainable score for the sub-content areas range from 60.2 to 66.9.

Mathematics

The mean scale score for the total population of students taking the 2006 eighth-grade Mathematics assessment is 562 with a standard deviation of 75.5. The

mean scale score for female students is 562 with a standard deviation of 71.6. The mean scale score for male students is 563 with a standard deviation of 79.0.

The scale score frequency distribution for the total population is shown in Table A-23. Figure A-23 graphically represents the frequency distributions for the total population and for the groups of male and female students separately. The scale score distributions are approximately normal (with a small group of students located at the LOSS) for the total population and for each gender.

The mean scale scores for the content standards range from 554 to 564. The mean scale scores for sub-content areas range from 537 to 567. The median scale scores vary between 566 and 569 for the content standards, and between 563 and 569 for the sub-content areas, and all are fairly close to the median for the total test scale score, 568.

The mean percents of the maximum obtainable score for the content standards range from 39.9 on CS 4/5 (Geometry and Measurement) to 54.7 on CS 2 (Algebra, Patterns, and Functions). The mean percent of the maximum obtainable score for the total test is 47.8. The mean percents of the maximum obtainable score for the sub-content areas range from 32.5 to 55.9.

Science

The mean scale score for the total population of students taking the 2006 eighth-grade Science assessment is 500 with a standard deviation of 60.7. The mean scale score for female students is 498 with a standard deviation of 57.0, and the mean scale score for male students is 502 with a standard deviation of 64.0.

The scale score frequency distribution for the total population is shown in Table A-24. Figure A-24 graphically represents the frequency distributions for the total population and for the groups of male and female students separately. The distributions of the scale scores are approximately normal (with a small group of students at the LOSS) for the total population and for each gender.

The mean scale scores for the content standards range from 498 to 501. The mean scale scores for the sub-content areas range from 487 to 515. The median scale scores vary between 505 and 507 for the content standards, and between 504 and 513 for the sub-content areas, and most are very close to the median for the total test scale score, 507.

The mean percents of the maximum obtainable score for the content standards range from 40.9 on CS 4/5 (Earth and Space Science and Its Interrelationship with Technology & Human Activity) to 61.8 on CS 1/6 (Scientific Investigations and Connections Among Scientific Disciplines). The mean percent of the obtainable score for the total test is 53.2. The mean percents of the maximum obtainable score for the sub-content areas range from 34.5 to 66.4.

Ninth Grade

Reading

The mean scale score for the total population of students taking the 2006 ninth-grade Reading assessment is 658 with a standard deviation of 62.3. The mean scale score for female students is 668 with a standard deviation of 56.3, and the mean scale score for male students is 649 with a standard deviation of 66.2.

The scale score frequency distribution for the total population is shown in Table A-25. Figure A-25 graphically represents the frequency distributions for the total population and for the groups of male and female students separately. The figure shows that the scale score distributions for the total population and for each gender are approximately normal, with a small group of students at the LOSS.

The mean scale scores for the content standards range from 653 to 660. The mean scale scores for the sub-content areas range from 643 to 662. The median scale scores vary between 666 and 667 for the content standards, and between 665 and 667 for the sub-content areas, and all are close to the median for the total test scale score, 666.

The mean percents of the maximum obtainable score for the content standards range from 51.3 on CS 6 (Literature) to 66.4 on CS 4 (Thinking Skills). The mean percent of the maximum obtainable score for the total test is 61.3. The mean percents of the maximum obtainable score for the sub-content areas range from 51.6 to 68.1.

Writing

The mean scale score for the total population of students taking the 2006 ninth-grade Writing assessment is 565 with a standard deviation of 79.6. The mean scale score for female students is 581 with a standard deviation of 76.3, and the mean scale score for male students is 550 with a standard deviation of 79.8.

The scale score frequency distribution for the total population is shown in Table A-26. Figure A-26 graphically represents the frequency distributions for the total population and for the groups of male and female students separately. The figure indicates that the scale score distributions are approximately normal for the total population and for each gender.

The mean scale scores for the content standards vary between 565 and 569. The mean scale scores for sub-content areas range from 564 to 580. The median scale scores vary between 568 and 569 for the content standards, and

between 536 and 570 for the sub-content areas, and most, with the exception of SA 6 with a median of 536, are close to the median for the total test scale score, 568. The median scale score for SA 6 (Extended Writing) was somewhat lower than the median for the total test score. It should be noted that the score for this sub-content area is computed based on the four scores a student gets for his/her response to the extended writing prompt. Consequently, the scale score variable for this sub-content area is rather discrete.

The mean percents of the maximum obtainable score for CS 2 (Write for a Variety of Purposes) and CS 3 (Write Using Conventions) are 63.5 and 64.9, respectively. The mean percent of the maximum obtainable score for the total test is 64.2. The mean percents of the maximum obtainable score for the sub-content areas range from 58.5 to 68.4.

Mathematics

The mean scale score for the total population of students taking the 2006 ninth-grade Mathematics assessment is 576 with a standard deviation of 72.9. The mean scale score for female students is 575 with a standard deviation of 68.2, and the mean scale score for male students is 576 with a standard deviation of 77.2.

The scale score frequency distribution for the total population is shown in Table A-27. Figure A-27 graphically represents the frequency distributions for the total population and for the groups of male and female students separately. The scale score distributions are approximately normal (with a small group of students at the LOSS) for the total population and for each gender.

The mean scale scores for the content standards range from 566 to 575. The mean scale scores for the sub-content areas are 563 and 575. The median scale scores vary between 582 and 583 for the content standards, and 581 and 583 for the sub-content areas, and all are close to the median for the total test scale score, 583.

The mean percents of the maximum obtainable score for the content standards range from 35.6 on CS 4/5 (Geometry and Measurement) to 51.8 on CS 2 (Algebra, Patterns, and Functions). The mean percent of the maximum obtainable score for the total test is 45.7. The mean percents of the maximum obtainable score for the sub-content areas range from 44.0 to 46.4.

Tenth Grade

Reading

The mean scale score for the total population of students taking the 2006 tenth-grade Reading assessment is 683 with a standard deviation of 62.4. The mean

scale score for female students is 693 with a standard deviation of 56.8, and the mean scale score for male students is 674 with a standard deviation of 66.1.

The scale score frequency distribution for the total population is shown in Table A-28. Figure A-28 graphically represents the frequency distributions for total population and for the groups of male and female students separately. The figure shows that the scale score distributions for the total population and for each gender are negatively skewed.

The mean scale scores for the content standards range from 680 to 688. The mean scale scores for the sub-content areas range from 677 to 685. The median scale scores vary between 690 and 691 for the content standards, and between 690 and 692 for the sub-content areas, and all are close to the median for the total test scale score, 691.

The mean percents of the maximum obtainable score for the content standards range from 48.4 on CS 6 (Literature) to 66.5 on CS 5 (Use of Literary Information). The mean percent of the maximum obtainable score for the total test is 58.2. The mean percents of the maximum obtainable score for the sub-content areas range from 44.9 to 61.2.

Writing

The mean scale score for the total population of students taking the 2006 tenth-grade Writing assessment is 578 with a standard deviation of 82.3. The mean scale score for female students is 597 with a standard deviation of 79.0, and the mean scale score for male students is 560 with a standard deviation of 81.4.

The scale score frequency distribution for the total population is shown in Table A-29. Figure A-29 graphically represents the frequency distributions for the total population and for the groups of male and female students separately. The figure shows that the scale score distributions for the total population and for each gender are approximately normal.

The mean scale scores for the content standards vary between 579 and 583. The mean scale scores for the sub-content areas range from 583 to 586. The median scale scores vary between 578 and 582 for the content standards, and between 523 and 584 for the sub-content areas, and most, with the exception of SA 6 with a median of 523, are close to the median for the total test scale score, 581.

The mean percents of the maximum obtainable score for CS 2 (Write for a Variety of Purposes) and CS 3 (Write Using Conventions) are 66.9 and 64.3 respectively. The mean percent of the maximum obtainable score for the total test is 65.6. The mean percents of the maximum obtainable score for the sub-content areas vary from 61.7 to 70.5.

Mathematics

The mean scale score for total population of students taking the 2006 tenth-grade Mathematics assessment is 586 with a standard deviation of 73.6. The mean scale score for female students is 586 with a standard deviation of 69.5, and the mean scale score for male students is 586 with a standard deviation of 77.4.

The scale score frequency distribution for the total population is shown in Table A-30. Figure A-30 graphically represents the frequency distributions for the total population and for the groups of male and female students separately. The figure shows that the scale score distributions for the total population and for each gender are approximately normal with a group of students at the LOSS.

The mean scale scores for the content standards range from 576 to 588. The mean scale scores for the sub-content areas are 585 and 598. The median scale scores vary between 594 and 596 for the content standards, and between 594 and 600 for the sub-content areas, and most are close to the median for the total test scale score, 595.

The mean percents of the maximum obtainable score for the content standards range from 33.9 on CS 4/5 (Geometry and Measurement) to 51.0 on CS 2 (Algebra, Patterns, and Functions). The mean percent of the maximum obtainable score for the total test is 45.4. The mean percents of the maximum obtainable score for the sub-content areas vary from 47.8 to 52.5.

Science

The mean scale score for the total population of students taking the 2006 tenth-grade Science assessment is 501 with a standard deviation of 61.9. The mean scale score for female students is 498 with a standard deviation of 57.9, and the mean scale score for male students is 504 with a standard deviation of 65.4.

The scale score frequency distribution for the total population is shown in Table A-31. Figure A-31 graphically represents the frequency distributions for the total population and for the groups of male and female students separately. The distributions of the scale scores are approximately normal (with a group of students at the LOSS) the total population and for each gender.

The mean scale scores for the content standards range from 497 to 503. The mean scale scores for the sub-content areas range from 489 to 511. The median scale scores vary between 508 and 510 for the content standards, and between 507 and 510 for the sub-content areas, and most are very close to the median for the total test scale score, 510.

The mean percents of the maximum obtainable score for the content standards range from 42.2 on CS 3/5 (Life Science and Its Interrelationship with Technology & Human Activity) to 59.0 on CS 1/6 (Scientific Investigations and Connections Among Scientific Disciplines). The mean percent of the obtainable score for the total test is 49.8. The mean percents of the maximum obtainable score for the sub-content areas range from 39.2 to 66.4.

Correlations Among Content Standards and Among Sub-Content Areas

Tables 13 through 16 show the correlations between the scale scores for the total test and for the various content standards and sub-content areas, for each grade and content area. All content standards and sub-content areas are positively correlated, as would be expected.

For the Reading assessments, the correlation coefficients vary between .59 (grade 4) and .77 (grade 8) for the relationship between the various content standards, and between .50 (grades 4 and 9) and .75 (grade 5) for the relationship between the various sub-content areas, respectively.

For the Grade 3 Spanish Reading assessments, correlations among sub-content areas vary between .56 and .63. For the Grade 4 Spanish Reading assessments, the correlations among the various content standards vary between .57 and .70 and correlations among sub-content areas vary between .55 and .73.

For the Writing assessments, the coefficients for the correlation between content standards 2 and 3 vary between .69 (grade 3) and .79 (grade 9). The correlations among the various sub-content areas vary between .36 (grade 4) and .74 (grade 10).

For the Spanish Writing assessments, the correlation between content standards 2 and 3 varies between .67 (grade 4) and .74 (grade 3); the correlations between the various sub-content areas vary between .33 (grade 4) and .55 (grade 3).

For the Mathematics assessments, the correlations vary between .58 (grade 4) and .80 (grades 7, 9, and 10) for the relationship among the content standards, and between .54 (grade 5) and .70 (grade 9) for the relationship among the sub-content areas.

Finally, for the Science assessments, the correlation coefficients vary between .65 (grade 5) and .77 (grade 8) for the relationship among the content standards, and between .43 (grade 8) and .65 (grade 10) for the relationship among the sub-content areas.

Test 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. Based on 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).

Table 17 shows the estimated reliability index (Cronbach's alpha) for the total test and for each content standard at each grade. Total score reliability coefficients are all .86 or greater. These reliability coefficients indicate that the Colorado 2006 assessments had strong internal consistency and that the tests produce relatively stable scores.

Table 17 also shows the reliability coefficients for individual standards. Table 18 provides similar information for all of the sub-content areas. These coefficients tend to be somewhat lower than the coefficients for the total test scores. These results are consistent with the smaller numbers of items that contribute to each standard and sub-content area.

Sub-group reliability coefficients are shown in Appendix G.

Part 4: Item Analyses

Tables 19 through 80 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. The p-value for each MC item is the percent of students who gave a correct response to the item. The p-value for each CR item is the mean percent of the maximum possible score. The item-to-total-score correlations, the p-values, the percentage of omits, and the percentages at each score level (for the CR items) are based on the analysis of responses of students who had valid total test scores only. Any omitted responses to individual items were treated as incorrect for the calculation of the p-values and the item-to-total-score correlations. This was consistent with how these omits are treated in the computation of the operational scale scores.

Third Grade

Reading

Table 19 lists the results of the multiple-choice item analyses for the 2006 third-grade Reading assessment. The point-biserials for all multiple-choice items range from .15 to .58 with a mean of .42. The p-values for the multiple-choice items range from .25 to .96 with a mean of .70.

Table 20 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .50 to .70 with a mean of .59. The p-values range from .39 to .65 with a mean of .51. An examination of the percent of students obtaining each score point for the constructed-response items shows that there is a reasonable amount of variability in students' responses to most items, indicating that these items work well over the range of student ability.

The omit rate for the third-grade Reading assessment was small, ranging from .1% to 1.7% for multiple-choice items (Table 19) and .8% to 3.0% for constructed-response items (Table 20).

Reading – Spanish Version

Table 21 lists the results of the multiple-choice item analyses for the Spanish version of the 2006 third-grade Reading assessment. The point-biserials for all

multiple-choice items range from .06 to .53 with a mean of .35. The p-values for the multiple-choice items range from .19 to .94 with a mean of .59.

Table 22 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .46 to .64 with a mean of .56. The p-values range from .37 to .83, with a mean of .60. An examination of the percent of students obtaining each score point for the constructed-response items shows that there is a reasonable amount of variability in students' responses to most items, indicating that these items work reasonably well over the range of student ability.

The omit rate for the Spanish version of the third-grade Reading assessment was small, ranging from .1% to 3.2% for multiple-choice items (Table 21) and .7% to 2.6% for constructed-response items (Table 22).

Writing

Table 23 lists the results of the multiple-choice item analyses for the 2006 third-grade Writing assessment. The point-biserials for all multiple-choice items range from .25 to .50 with a mean of .42. The p-values for the multiple-choice items range from .59 to .95 with a mean of .81.

Table 24 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .16 to .56 with a mean of .43. The p-values range from .05 to .98, with a mean of .78. For 12 out of the 18 constructed-response items, over 80% of the students obtained the highest possible score points.

The omit rate for the third-grade Writing assessment was small, ranging from .0% to 1.9% for multiple-choice items (Table 23) and from .2% to .4% for constructed-response items (Table 24).

Writing – Spanish Version

Table 25 lists the results of the multiple-choice item analyses for the Spanish version of the 2006 third-grade Writing assessment. The point-biserials for all multiple-choice items range from .05 to .49 with a mean of .38. The p-values for the multiple-choice items range from .31 to .97 with a mean of .71.

Table 26 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .38 to .58 with a mean of .50. The p-values range from .24 to .86, with a mean of .61. An examination of the percent of students obtaining each score point for the constructed-response items shows that there is a reasonable amount of variability in students' responses to most items, indicating that these items work reasonably well over the range of student ability.

The omit rate for the Spanish version of the third-grade Writing assessment was small, ranging from .0% to 1.3% for multiple-choice items (Table 25) and .4% to .9% for constructed-response items (Table 26).

Mathematics

Table 27 lists the results of the multiple-choice item analyses for the 2006 third-grade Mathematics assessment. The point-biserials for the multiple-choice items range from .20 to .57, with a mean of .41. The p-values for the multiple-choice items range from .57 to .97 with a mean of .81.

Table 28 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .48 to .72 with a mean of .61. The p-values range from .46 to .78 with a mean of .64. The distribution of the percent of students obtaining each score point for the constructed-response items shows that there is a good amount of variability in students' responses to most items, indicating that these items work well over the range of student ability.

The omit rate for the third-grade Mathematics assessment was low, ranging from .2% to 2.5% for multiple-choice items (Table 27) and .1% to .5% for constructed-response items (Table 28).

Fourth Grade

Reading

Table 29 lists the results of the multiple-choice item analyses for the 2006 fourth-grade Reading assessment. The point-biserials for the multiple-choice items range from .15 to .57 with a mean of .41. The p-values for the multiple-choice items range from .33 to .96 with a mean of .71.

Table 30 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .42 to .58 with a mean of .50. The p-values range from .31 to .79 with a mean of .48. An examination of the percent of students obtaining each score point for the constructed-response items shows that there is a good amount of variability in students' responses to most items, indicating that these items work well over the range of student ability.

The omit rate for the fourth-grade Reading assessment was low, ranging from .1% to 3.3% for multiple-choice items (Table 29). The range was .6% to 2.4% for constructed-response items (Table 30).

Reading – Spanish Version

Table 31 lists the results of the multiple-choice item analyses for the Spanish version of the 2006 fourth-grade Reading assessment. The point-biserials for all multiple-choice items range from .11 to .54 with a mean of .34. The p-values for the multiple-choice items range from .23 to .92 with a mean of .57.

Table 32 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .31 to .69 with a mean of .52. The p-values range from .18 to .75, with a mean of .42. An examination of the percent of students obtaining each score point for the constructed-response items shows that there is a reasonable amount of variability in students' responses to most items, indicating that these items work reasonably well over the range of student ability.

The omit rate for the Spanish version of the fourth-grade Reading assessment was low, ranging from .0% to 4.6% for multiple-choice items (Table 31) and .2% to 6.5% for constructed-response items (Table 32).

Writing

Table 33 lists the results of the multiple-choice item analyses for the 2006 fourth-grade Writing assessment. The point-biserials for all multiple-choice items range from .28 to .58 with a mean of .42. The p-values for the multiple-choice items range from .49 to .97 with a mean of .77.

Table 34 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .12 to .65 with a mean of .47. The p-values range from .28 to .97, with a mean of .63.

The omit rate for the fourth-grade Writing assessment was small, ranging from .1% to 1.8% for multiple-choice items (Table 33) and .0% to 2.1% for constructed-response items (Table 34).

Writing – Spanish Version

Table 35 lists the results of the multiple-choice item analyses for the Spanish version of the 2006 fourth-grade Writing assessment. The point-biserials for all multiple-choice items range from .17 to .48 with a mean of .33. The p-values for the items range from .27 to .95 with a mean of .52.

Table 36 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed response items range from .17 to .68 with a mean of .45. The p-values range from .14 to .88 with a mean of .52.

The omit rate for the Spanish version of the fourth-grade Writing assessment was small, ranging from .0% to 7.2% for multiple-choice items, with one multiple choice item having a omit rate of greater than or equal to 5% (Table 35), and .4% to 3.8% for constructed-response items (Table 36).

Mathematics

Table 37 lists the results of the multiple-choice item analyses for the 2006 fourth-grade Mathematics assessment. The point-biserials for the multiple-choice items range from .16 to .59, with a mean of .41. The p-values for the multiple-choice items range from .48 to .98 with a mean of .79.

Table 38 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .40 to .76 with a mean of .57. The p-values range from .33 to .84 with a mean of .63. The distribution of the percent of students obtaining each score point for the constructed-response items shows that there is a good amount of variability in students' responses to most items, indicating that these items work well over the range of student ability.

The omit rate for the fourth-grade Mathematics assessment was low, ranging from .1% to 1.0 % for multiple-choice items (Table 37) and .1% to 1.3% for constructed-response items (Table 38).

Fifth Grade

Reading

Table 39 lists the results of the multiple-choice item analyses for the 2006 fifth-grade Reading assessment. The point-biserials for the multiple-choice items range from .12 to .59, with a mean of .43. The p-values for the multiple-choice items range from .38 to .92 with a mean of .73.

Table 40 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .31 to .67 with a mean of .49. The p-values range from .20 to .85 with a mean of .52. The distribution of the percent of students obtaining score level for the constructed-response items shows that there is a good amount of variability in students' responses to most items, indicating that these items work well over the range of student ability. More than 50% of the students obtained the highest possible score points for 4 out of the 15 constructed-response items. The scores of the remaining students were well distributed across the score points in that item, indicating that they produced a reasonable amount of variability.

The omit rate for the fifth-grade Reading assessment was small, ranging from .1% to 2.9% for multiple-choice items (Table 39) and .5% to 2.1% for constructed-response items (Table 40).

Writing

Table 41 lists the results of the multiple-choice item analyses for the 2006 fifth-grade Writing assessment. The point-biserials for all multiple-choice items range from .23 to .57 with a mean of .42. The p-values for the multiple-choice items range from .37 to .94 with a mean of .74. Item 60 in the fifth-grade Writing assessment was removed from calibration.

Table 42 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .10 to .68 with a mean of .47. The p-values range from .07 to .98, with a mean of .60.

The omit rate for the fifth-grade Writing assessment was small, ranging from .1% to 2.7% for multiple-choice items (Table 41) and .1% to 1.5% for constructed-response items (Table 42).

Mathematics

Table 43 lists the results of the multiple-choice item analyses for the 2006 fifth-grade Mathematics assessment. The point-biserials for the multiple-choice items range from .26 to .57, with a mean of .43. The p-values for the multiple-choice items range from .38 to .94 with a mean of .75.

Table 44 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .52 to .71 with a mean of .61. The p-values range from .47 to .87 with a mean of .69. The distribution of the percent of students obtaining each score point for the constructed-response items shows that there is a good amount of variability in students' responses to most items, indicating that these items work well over the range of student ability.

The omit rate for the fifth-grade Mathematics assessment was low, ranging from .1% to 1.8% for multiple-choice items (Table 43) and .1% to 3.1% for constructed-response items (Table 44).

Science

Table 45 lists the results of the multiple-choice item analyses for the 2006 fifth-grade Science assessment. The point-biserials for the multiple-choice items range from .16 to .53, with a mean of .35. The p-values for the multiple-choice items range from .50 to .98 with a mean of .76.

Table 46 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .17 to .57 with a mean of .43. The p-values range from .12 to .92 with a mean of .55. The distribution of the percent of students obtaining each score point for the constructed-response items shows that there is a good amount of variability in students' responses to most items, indicating that these items work well over the range of student ability.

The omit rate for the fifth-grade Science assessment was low, ranging from .0% to .6% for multiple-choice items (Table 45) and .1% to 1.2% for constructed-response items (Table 46).

Sixth Grade

Reading

Table 47 lists the results of the multiple-choice item analyses for the 2006 sixth-grade Reading assessment. The point-biserials for the multiple-choice items range from .19 to .59 with a mean of .40. The p-values for the multiple-choice items range from .29 to .94 with a mean of .69.

Table 48 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .34 to .62 with a mean of .50. The p-values range from .30 to .87 with a mean of .55. The distribution of the percent of students obtaining each score point for the constructed-response items shows that there is a good amount of variability in students' responses to most items, indicating that these items work well over the range of student ability. Over 50% of the students obtained the highest possible score points in 3 out of the 14 constructed-response items. The scores of the remaining students were well distributed across the score points, indicating that these items produced a reasonable amount of variability.

The omit rate for the sixth-grade Reading assessment was low, ranging from .1% to 4.6% for multiple-choice items (Table 47) and .5% to 3.5% for constructed-response items (Table 48).

Writing

Table 49 lists the results of the multiple-choice item analyses for the 2006 sixth-grade Writing assessment. The point-biserials for all multiple-choice items range from .19 to .58 with a mean of .41. The p-values for the multiple-choice items range from .25 to .89 with a mean of .68.

Table 50 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .10 to .62 with a mean of .45. The p-values range from .24 to .99 with a mean of .67.

The omit rate for the sixth-grade Writing assessment was small, ranging from .1% to 2.7% for multiple-choice items (Table 49) and .1% to 6.6% for constructed-response items (Table 50).

Mathematics

Table 51 lists the results of the multiple-choice item analyses for the 2006 sixth-grade Mathematics assessment. The point-biserials for the multiple-choice items range from .22 to .61 with a mean of .43. The p-values for the multiple-choice items range from .34 to .94 with a mean of .69.

Table 52 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .32 to .76 with a mean of .59. The p-values range from .29 to .88 with a mean of .61. The distribution of the percent of students obtaining each score point for the constructed-response items shows that there is a good amount of variability in students' responses to most items, indicating that these items work well over the range of student ability. Over 50% of the students obtained the highest possible score points in 3 out of the 15 constructed-response items. The scores of the remaining students were well distributed across the score points, indicating that these items produced a reasonable amount of variability.

The omit rate for the sixth-grade Mathematics assessment was low, ranging from .1% to 1.5% for multiple-choice items (Table 51) and .2% to 2.1% for constructed-response items (Table 52).

Seventh Grade

Reading

Table 53 lists the results of the multiple-choice item analyses for the 2006 seventh-grade Reading assessment. The point-biserials for the multiple-choice items range from .16 to .56 with a mean of .39. The p-values for the multiple-choice items range from .20 to .95 with a mean of .67.

Table 54 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items are positive, ranging from .39 to .58 with a mean of .50. The p-values for the constructed-response items range from .26 to .83 with a mean of .53. The distribution of the percent of students obtaining each score point for the Reading constructed-

response items shows that there is a good amount of variability in students' responses to most items, indicating that these items work well over the range of student ability. Over 50% of the students obtained the highest possible score points in three out of the 14 constructed-response items. The scores of the remaining students are well distributed across the score points, indicating that these items produced a reasonable amount of variability.

The percent of students who omitted the multiple-choice items in the 2006 seventh-grade Reading assessment ranged from .1% to 2.5% (Table 53). The percent of students who omitted constructed-response items ranged from .8% to 3.4% (Table 54).

Writing

Table 55 lists the results of the multiple-choice item analyses for the 2006 seventh-grade Writing assessment. The point-biserials for all multiple-choice items range from .24 to .53 with a mean of .42. The p-values for the multiple-choice items range from .27 to .92 with a mean of .66.

Table 56 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .11 to .66 with a mean of .42. The p-values range from .36 to .99, with a mean of .62.

The omit rate for the seventh-grade Writing assessment was small, ranging from .1% to 2.3% for multiple-choice items (Table 55) and .1% to 3.0% for constructed-response items (Table 56).

Mathematics

Table 57 lists the results of the multiple-choice item analyses for the 2006 seventh-grade Mathematics assessment. The point-biserials for the multiple-choice items range from .18 to .54, with a mean of .39. The p-values for the multiple-choice items range from .33 to .95 with a mean of .65.

Table 58 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .31 to .76 with a mean of .61. The p-values range from .14 to .86 with a mean of .45. The distribution of the percent of students obtaining each score point for the constructed-response items shows that there is a good amount of variability in students' responses to most items, indicating that these items work well over the range of student ability.

The omit rate for the seventh-grade Mathematics assessment was generally low, ranging from .1% to 3.9% for multiple-choice items (Table 57) and .3% to 4.8% for constructed-response items (Table 58).

Eighth Grade

Reading

Table 59 lists the results of the multiple-choice item analyses for the 2006 eighth-grade Reading assessment. The point-biserials for the multiple-choice items range from .16 to .58 with a mean of .39. The p-values for the multiple-choice items range from .23 to .93 with a mean of .65.

Table 60 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .38 to .65 with a mean of .55. The p-values range from .18 to .85 with a mean of .47. The distribution of the percent of students obtaining each score point for the constructed-response items shows a good amount of variability in students' responses to most items. Over 50% of the students obtained the highest possible score points in three out of the 14 constructed-response items. The scores of the remaining students were well distributed across the score points, indicating that these items produced a reasonable amount of variability.

The percent of students who omitted the multiple-choice items in the eighth-grade Reading assessment ranged from .1% to 3.6% (Table 59). The percent of students who omitted the constructed-response items ranged from 1.3% to 9.3% (Table 60), with three items having an omit rate greater than 5%.

Writing

Table 61 lists the results of the multiple-choice item analyses for the 2006 eighth-grade Writing assessment. The point-biserials for all multiple-choice items range from .17 to .56 with a mean of .42. The p-values for the multiple-choice items range from .24 to .91 with a mean of .65. Item 63 was removed from calibration.

Table 62 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .11 to .63 with a mean of .46. The p-values range from .25 to .99, with a mean of .61. Item 3C was removed from calibration.

The omit rate for the eighth-grade Writing assessment was reasonable, ranging from .1% to 6.0% for multiple-choice items, with two items with omit rates greater than 5% (Table 61) and .1% to 2.2% for constructed-response items (Table 62).

Mathematics

Table 63 lists the results of the multiple-choice item analyses for the 2006 eighth-grade Mathematics assessment. The point-biserials for the multiple-choice items

range from .15 to .61 with a mean of .39. The p-values for the multiple-choice items range from .11 to .85 with a mean of .53.

Table 64 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .41 to .76 with a mean of .62. The p-values range from .08 to .69 with a mean of .41. The distribution of the percent of students obtaining each score point for the Mathematics constructed-response items shows a good amount of variability in students' responses to most items, indicating that these items work well over the range of student ability.

The percent of students who omitted multiple-choice items in the eighth-grade Mathematics assessment ranged from .1% to 10.0% with only one multiple-choice item having an omit rate greater than 5% (Table 63). The percent of students who omitted constructed-response items ranged from .4% to 4.6%, (Table 64).

Science

Table 65 lists the results of the multiple-choice item analyses for the 2006 eighth-grade Science assessment. The point-biserials for the multiple-choice items range from .09 to .51 with a mean of .34. The p-values for the multiple-choice items range from .16 to .93 with a mean of .62. Items 7 and 18 were removed from calibration.

Table 66 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .22 to .68 with a mean of .46. The p-values range from .06 to .84 with a mean of .44. The percent of students obtaining each score point for the constructed-response items shows a good amount of variability in students' responses to most items, indicating that these items work well over the range of student ability.

The omit rate for the multiple-choice items for the eighth-grade Science assessment ranged from .0% to 3.4% (Table 65). The omit rate for the constructed-response items ranged from .6% to 11.0% (Table 66), with only one item having an omit rate greater than 5%.

Ninth Grade

Reading

Table 67 lists the results of the multiple-choice item analyses for the 2006 ninth-grade Reading assessment. The point-biserials for the multiple-choice items

range from .18 to .59 with a mean of .40. The p-values for the multiple-choice items range from .23 to .90 with a mean of .66.

Table 68 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .40 to .69 with a mean of .55. The p-values range from .21 to .89 with a mean of .53. The distribution of the percent of students obtaining each score point for the constructed-response items shows a good amount of variability in students' responses to most items, indicating that these items work well over the range of student ability

The omit rate for the multiple-choice items for the ninth-grade Reading assessment ranged from .1% to 2.2% (Table 67). The omit rate for the constructed-response items ranged from 1.2% to 6.0%, with seven out of the 14 items having an omit rate greater than 5% (Table 68).

Writing

Table 69 lists the results of the multiple-choice item analyses for the 2006 ninth-grade Writing assessment. The point-biserials for all multiple-choice items range from .28 to .56 with a mean of .45. The p-values for the multiple-choice items range from .29 to .88 with a mean of .64.

Table 70 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .16 to .64 with a mean of .47. The p-values range from .37 to .99, with a mean of .69.

The omit rate for the ninth-grade Writing assessment was small, ranging from .1% to 1.4% for multiple-choice items (Table 69) and .0% to 2.8% for constructed-response items (Table 70).

Mathematics

Table 71 lists the results of the multiple-choice item analyses for the 2006 ninth-grade Mathematics assessment. The point-biserials for the multiple-choice items range from .01 (on item 57 which was removed from calibration) to .56 with a mean of .37. The p-values for the multiple-choice items range from .27 to .93 with a mean of .56.

Table 72 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .55 to .72 with a mean of .64. The p-values range from .15 to .54 with a mean of .35. The distribution of the percent of students obtaining each score point for the Mathematics constructed-response items shows a fair amount of variability in students' responses to most items, indicating that these items work well over the range of student ability.

The percent of students who omitted multiple-choice items in the ninth-grade Mathematics assessment ranged from .1% to .8% (Table 71). The percent of students who omitted constructed-response items ranged from 1.2% to 4.0% (Table 72).

Tenth Grade

Reading

Table 73 lists the results of the multiple-choice item analyses for the 2006 tenth-grade Reading assessment. The point-biserials for the multiple-choice items range from .06 to .52 with a mean of .36. The p-values for the multiple-choice items range from .28 to .92 with a mean of .66.

Table 74 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .40 to .65 with a mean of .54. The p-values range from .30 to .74 with a mean of .47. The distribution of the percent of students obtaining each score point for the constructed-response items shows a good amount of variability in students' responses to most items. Over 50% of the students obtained the highest possible score points in two out of the 14 constructed-response items. The scores of the remaining students were well distributed across the score points, indicating that these items produced a reasonable amount of variability.

The omit rates for the multiple-choice items for the 2006 tenth-grade Reading assessment ranged from .1% to 1.8% (Table 73). Omit rates for the constructed-response items ranged from 2.4% to 11.6% (Table 74), with six out of the 14 items having an omit rate greater than 5%.

Writing

Table 75 lists the results of the multiple-choice item analyses for the 2006 tenth-grade Writing assessment. The point-biserials for all multiple-choice items range from .06 to .59 with a mean of .41. The p-values for the multiple-choice items range from .26 to .91 with a mean of .67.

Table 76 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .17 to .65 with a mean of .49. The p-values range from .34 to .99, with a mean of .64. The percent of students obtaining each score point for the constructed-response items shows a good amount of variability in students' responses to most items.

The omit rate for the tenth-grade Writing assessment was small, ranging from .1% to .7% for multiple-choice items (Table 75) and .0% to 3.7% for constructed-response items (Table 76).

Mathematics

Table 77 lists the results of the multiple-choice item analyses for the 2006 tenth-grade Mathematics assessment. The point-biserials for the multiple-choice items range from -.19 (for item 41 that was removed from calibration) to .59 with a mean of .36. The p-values for the multiple-choice items range from .18 to .89 with a mean of .54.

Table 78 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .49 to .74 with a mean of .62. The p-values for the constructed-response items range from .07 to .71 with a mean of .38. The distribution of the percent of students obtaining each score point for the constructed-response items shows a good amount of variability in students' responses to most items.

The omit rate for the multiple-choice items for the tenth-grade Mathematics assessment ranged from .1% to 1.4% (Table 77). The omit rate for the constructed-response items ranged from 1.4% to 5.3% (Table 78) with one out of the 15 items having an omit rate greater than 5%.

Science

Table 79 lists the results of the multiple-choice item analyses for the 2006 tenth-grade Science assessment. The point-biserials for the multiple-choice items range from .04 to .55 with a mean of .36. The p-values for the multiple-choice items range from .21 to .95 with a mean of .58. Items 13 and 20 were removed from calibration.

Table 80 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .18 to .58 with a mean of .44. The p-values for the constructed-response items range from .05 to .74 with a mean of .36. The distribution of the percent of students obtaining each score point for the constructed-response items shows a good amount of variability in students' responses to most items.

The omit rate for the multiple-choice items for the tenth-grade Science assessment ranged from .1% to .6% (Table 79). The omit rate for the constructed-response items ranged from 1.9% to 16.0% (Table 80) with nine out of the 23 items having an omit rate greater than or equal to 5%.

Part 5: Scaling and Calibration

Overview of the IRT Models

CTB uses item response theory (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 & Novick, 1968; Lord, 1980) 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

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

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:

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

and

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

where α_j and γ_{ji} are the parameters to be estimated from the data. The first constraint implies that higher item scores reflect higher ability levels and that items can vary in their discriminations. For the 2PPC model, for each item, there

are m_j –1 independent γ_{ji} parameters and one α_j parameter; a total of m_j independent item parameters are estimated.

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

Calibration of the Assessment

The items within 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 parameter estimates output by PARDUX are in two different parameterizations, corresponding to the two item response models (3PL and 2PPC). The location (i.e., difficulty) and discrimination parameters for the multiple-choice items are in the traditional 3PL metric and are labeled 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 MC and CR 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 scored items and using a generalization of this statistic for the multi-level (OE) items. As described by Yen, Q_1 is a Pearson chi-square of the form

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

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

$$E_{ji} = \frac{1}{N_{ji}} \sum_{a \varepsilon i}^{N_{ji}} P_j\left(\hat{\theta}_a\right).$$

The generalization of Q_1 for multi-level items 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

$$E_{jki} = \frac{1}{N_{ji}} \sum_{a \varepsilon i}^{N_{j1}} P_{jk}\left(\hat{\theta}_a\right).$$

O_{jki} is the observed proportion of examinees in cell i who performed at the *k-th* score level.

Chi-squared statistics are affected by sample size and extreme expectations (Stone, Ankenmann, Lane & Lia, 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 about 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

$$Z_c = \left(\frac{\text{Calibration Sample Size}}{1500}\right) 4.00.$$

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

Model Fit Analyses Results

The model fit statistics and item parameter results are based on the analysis of a sample data set used for item calibration and scaling. The summary fit statistics for the multiple-choice and constructed-response items for different grades and content areas are shown in Tables 81 through 142.

Detailed summaries of the model fit results are presented below.

Third Grade

The third grade item parameters and fit statistics are shown in Tables 81 to 90. The critical Z-values for these tables are 138.1 for Reading, 4.6 for Spanish Reading, 106.5 for Writing, 4.1 for Spanish Writing and 117.1 for Mathematics.

Across all content areas, three items exceeded these critical Z-values and exhibited less than optimal fit: one Reading item (CR item 7), and two Spanish Reading items (MC item 3 and CR item 26).

Fourth Grade

The fourth-grade item parameters and fit statistics are shown in Tables 91 to 100. The critical Z-values for these tables are 93.6 for Reading, 92.3 for Writing, and 97.3 for Mathematics. Spanish Reading and Spanish Writing both had critical Z-values of 1.4 for items that originated from the 2004 administration, and 1.3 for items that originated in the 2005 administration, and 2.7 for the CR items from the 2002 administration.

Across all content areas, five items exceeded these critical Z-values and exhibited less than optimal fit: three Reading items (MC item 100 and CR items 4 and 33), and two Writing items (CR items 3A and 53). Because of low sample sizes (approximately 500 students in 2004 and 2005) in pre-equated Spanish Grade 4 Reading and Writing, a slightly higher number of items (6 items in Spanish Reading and 5 in Spanish Writing) were flagged for misfit.

Fifth Grade

The fifth-grade item parameters and fit statistics are shown in Tables 101 to 108. The critical Z-values for these tables are 90.1 for Reading, 88.7 for Writing, 96.0 for Mathematics, and 110.3 for Science.

Across all content areas, five items exceeded these critical Z-values and exhibited less than optimal fit: one Reading item (MC item 95), two Writing items (CR items 3A and 92), one Mathematics item (CR item 35), and one Science item (CR item 38).

Sixth Grade

The sixth-grade item parameters and fit statistics are shown in Tables 109 to 114. The critical Z-values for these tables are 88.9 for Reading, 87.9 for Writing, and 89.5 for Mathematics.

Across all content areas, six items exceeded these critical Z-values and exhibited less than optimal fit: two Writing items (CR item 3A and 32) and four Mathematics items (MC item 9 and CR items 29, 47, 57).

Seventh Grade

The seventh-grade item parameters and fit statistics are shown in Tables 115 to 120. The critical Z-values for these tables are 74.9 for Reading, 74.6 for Writing, and 74.6 for Mathematics.

Across all content areas, five items exceeded these critical Z-values and exhibited less than optimal fit: three Writing items (CR items 2F, 3A and, 118) and two Mathematics items (CR items 35 and 44).

Eighth Grade

The eighth-grade item parameters and fit statistics are shown in Tables 121 to 128. The critical Z-values for these tables are 77.7 for Reading, 77.5 for Writing, 77.6 for Mathematics, and 125.1 for Science.

Across all content areas, 8 items exceeded these critical Z-values and exhibited less than optimal fit: two Reading items (MC item 110 and CR item 38), two Writing items (CR items 69 and 86), two Mathematics items (CR items 35 and 40), and two Science items (MC item 48 and CR item 36).

Ninth Grade

The ninth-grade item parameters and fit statistics are shown in Tables 129 to 134. The critical Z-values for these tables are 84.9 for Reading, 84.7 for Writing, and 73.7 for Mathematics.

Across all content areas, seven items exceeded these critical Z-values and exhibited less than optimal fit: two Reading items (MC items 39 and 102), two Writing items (CR items 2C and 3A), and three Mathematics items (CR items 12, 20, and 34).

Tenth Grade

The tenth-grade item parameters and fit statistics are shown in Tables 135 to 142. The critical Z-values for these tables are 72.7 for Reading, 72.5 for Writing, and 73.3 for Mathematics, and 121.0 for Science.

Across all content areas, seven items exceeded these critical Z-values and exhibited less than optimal fit: four Reading items (MC items 99, 100 104 and CR item 20), one Writing item (CR item 95), and two Mathematics items (MC item 13 and CR item 15).

Item Independence

In using IRT models, one of the assumptions made is that the items are locally independent. That is, a student's response on one item is not dependent upon 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 independence of items within a test is via the Q3 statistic (Yen, 1984). This statistic was obtained by correlating differences between students' observed and expected responses for pairs of items after taking into account overall test performance. If a substantial number of items in the test demonstrate local dependence, these items may need to be calibrated separately. Pairs of items with Q3 values greater than 0.30 were classified as locally dependent. The maximum value for this index is 1.00.

The number of item pairs flagged under the criterion was quite small and varied across forms and content areas. For the Reading, Mathematics, and Science, there were only 5 item pairs flagged across all grades and content areas. In contrast, 22 pairs were flagged for the Writing tests, with nine of those pairs flagged in Grade 3 and one to three pairs per remaining grades. The largest Q3 was found in Grade 10 Writing (.89). Overall, the few items exhibiting dependency (27 pairs across all possible item pair combinations for which Q3 ranged .31 - .89) were not of sufficient magnitude to warrant concern. Compared to Reading Grades 3 and 4 English items, relatively larger number of items in the Spanish tests are flagged for higher Q3 values ranging from .30 to .93.

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. Using the data from the calibration sample, the equating resulted in parameters expressed on the vertical scales for each content area.

A set of 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. These items remain unchanged across administrations and are given in approximately the same location or same third of the original administration location. In addition, these anchor items maintained original starting parameter values; that is, parameters expressed on the vertical scale for all content areas. These multiple choice items were used as anchors in the Spring 2006 CSAP to link the tests across years one to another. The anchor parameters were not fixed during calibration, and were used during the equating procedures defined by Stocking and Lord (1983). The anchor parameters were used to place the parameters estimated for all the Spring 2006 CSAP items on the scales described.

Equating is a statistical procedure that allows adjusting scores on test forms so that the scores are comparable. Horizontal equating within each grade was used to place the 2006 forms on the vertical scales that had been established previously for Reading, Writing, and Mathematics. The vertical scale for Reading, spanning grades 3 through 10, was established in 2001. The vertical scales for Writing, spanning grades 3 through 10, and for Mathematics, spanning grades 5 through 10, were established in 2002. The vertical scale for Math grades 3 and 4 was established in 2005.

The Stocking and Lord (1983) procedure, also called test characteristic curve (TCC) method, was used to place each grade on the vertical scale that had been developed 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 the test characteristic curve based on estimates from previous calibration and $\hat{\psi}_j^*$ be the test characteristic curve based on transformed estimates from the current calibration.

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

The TCC method determines the scaling constants (M1 and M2) by minimizing

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

the following quadratic loss function (F):

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

Anchor Set Review Process

The multiple choice anchor item set is carefully reviewed to ensure that it is performing very similarly in both current and reference years. The following verifications were performed to ensure the quality and accuracy of the equating.

1. P-values of the anchor items are compared. The anchor items should be similar in difficulty in both new and reference administrations. The estimated new form and the reference form p-values should be aligned on the regression line. If the samples are similar in ability, this regression line will be the identity line.
2. IRT item parameters are compared. The correlation coefficients between the reference and equated item parameters should be very high (.90 – 1.00).
3. Test Characteristic Curves for the anchor items are compared before and after the equating transformation is applied. The reference and equated anchor item TCCs should be closely overlapping.
4. The linear transformation parameters (aka, scaling constants) should be fairly stable across administrations.

Additional analyses of the equating include:

5. The p-values of the common anchor items between the two administrations show the same direction and magnitude of change as do the scale scores.
6. The full distribution of scale scores is reasonably comparable across administrations and reflects any differences in ability that are indicated by the anchor items.
7. The pass rates are reasonable across administrations, given any noted ability changes.

These routine CTB Research quality check steps were followed during equating for all grades and content areas. If an item was flagged for performing differentially from the Stocking and Lord (1983) procedures outlined above, it was further evaluated using Delta plot (Angoff, 1972; Dorans and Holland, 1993) and Lord's Chi-square (Lord, 1980) procedures. If all the statistical criteria suggest dropping the anchor item from the anchor set, it was further evaluated in terms of blueprint representation. See Appendix E.

P-Value Comparisons

The p-values were aligned closely with correlation higher than 0.95 for all grades and content areas (Table 143). This indicates that in a plot of reference and estimated new form P-values that the items fall closely to the identity line, i.e., the estimated proportion correct (p-value) for the reference and estimated new form item parameters are very similar. This indicates that the anchor items performed similarly in the two populations (2005 and 2006).

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 item difficulty (b), discrimination (a), and proportion correct (p-value) values as well as their plots. Guessing (c) parameters are most fluctuating and was not considered in the evaluation criteria.

Results indicate that the correlations for the discrimination (a) and difficulty (b) parameters are high, ranging from 0.86 to 0.99 for “a” and 0.88 to 1.00 for “b” (see Table 143.) These high correlations indicate that the items were performing essentially similarly between the two administrations. And, this is further evidence that the equating results are reasonable and accurate.

Test Characteristic Curve Comparisons

The observed and estimated anchor item TCCs were overlapped in most cases indicating that the equating transformation worked very well. Comparisons of TCCs between years are presented in Appendix E.

Scaling Constants

The scaling constants, or the linear transformation parameters, were examined to determine whether the ability differences were similar across years. Since the calibration “centers” the raw IRT scale close to the average ability of the test-takers, differences in these scaling constants would indicate differences in the ability from reference to new form administrations. The scaling constants for the CSAP grades and content areas are displayed in Table 144 for the two administrations (2005 and 2006).

Table 144 indicates that the scaling constants are fairly similar across the two administrations.

Additional Analyses of Flagged Items

As mentioned above, all items flagged as outliers in the standard Stocking & Lord process were evaluated further to determine whether they should be removed from the anchor item set.

For each flagged item, the Item Characteristic Curves (ICCs) for the reference and new form administration are compared. Root Mean Squared Differences between these curves are calculated, as is the chi-square. Appendix E displays the ICCs for all items that were removed from the anchor sets for all grades and content areas.

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 in all grades and content areas (see Tables 145 through 149).

Effectiveness of the Equating

Figures E-16 through E-44 in Appendix E show the TCC and SEM plots for the Spring 2006 operational tests Grades 3-10 Reading, Mathematics, Writing, Grade 8 Science, and Grades 3-4 Spanish Reading and Writing, compared to the previous year's plots. These plots illustrate the effectiveness of the equatings. The plots of the TCCs (the S shaped curves) and the SEM curves (the U shaped curves) indicate that the 2005 and 2006 for a given subject area and grade strongly resembled each other (in that they lay close to or even on top of another) in terms of difficulty, discrimination, and accuracy. Note that due to limited sample size for Spanish Grade 4 Reading and Writing, the tests were pre-equated using items from 2002, 2004, and 2005.

Once the tests are equated, final parameter tables are developed into scoring tables from which each student's scale score is derived. CSAP uses pattern-scoring for all items. During pattern scoring, the pattern of student responses and the attributes of each item contribute to the student's final scale score. This enhances the comparability of scores across years. For example, two students who respond correctly to a total of 20 questions have the same number correct raw score of 20. However, if one student answers the 20 most difficult questions while the other, the 20 easiest, the pattern-scoring is able to take those responses and item attributes into account and provide a scale score that better represents the students' abilities.

Part 6: Total and Subgroup Reliability

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

Item specific reliability statistics include inter-rater reliability, item total correlation, and DIF. The inter-rater reliability across CR items in terms of the kappa and intraclass correlations is one way to measure the consistency of the hand score. Appendix D provides the results of both rater reliability measures, which assess the agreement rates within a given administration, and rater severity analyses, which compare the scoring leniency across years. These results demonstrate that the CSAP tests have relatively high reader reliability. The kappa correlation for Mathematics tests ranged from 0.58 – 0.97, with a median value of 0.90. For Reading (English), the range was 0.42 – 0.99 with a median value of 0.83. And, for Reading (Spanish), the kappa ranged from 0.53 to 0.96 with a median of 0.89. For Science, the range was .0.54 – 0.97, and the median was 0.85. Writing (English) kappa values had a wider range, from 0.22 – 0.98 (median = 0.72), as did Writing (Spanish), which ranged from 0.51 – 0.97 (median = 0.72). Although 1 of the grade 5 Reading (English), and 8 of the Writing (English) (1 each for grades 4, 6, 7, and 8 and 4 for grade 3) items had low kappa values (below .45), the adjacent agreement for all Writing items was nearly 100%. Given the high adjacent agreement rates, these values are well within acceptable limits.

Additionally, Table D-7 in Appendix D displays the high consistency of the rating scales that were used from year to the next. This is an indication that the standards applied in the scoring of the CR items are quite stable within and administration and over time.

The item total correlation type of internal consistency measure is one measure of the correlation between each item and the overall test. This provides a source of how consistent the item measures information similar to the other items. Tables 19-80 in the main table document display the item total correlations (and p-values) for each grade and content area. Below each table is displayed the average values for each statistic. Review of these tables shows that the range of item total correlations is .05 to as high as .76. Item total correlations are calculated and thus dependent upon the number of items answered correctly divided by the number of items answered incorrectly. Thus, the p-values of the items are important to consider when reviewing the item total correlations.

According to a study cited in Croker & Algina (1986), if the average biserial correlation is in a range of about .30 to .40, the average p-value should ideally be between .40 and .60. Given that the mean item total correlations for CSAP assessments range from .42 to .64 across test forms and that the average p-values range from .35 to .78 across forms, the range of item total correlations is acceptable and close to the very rough rule of thumb cited.

DIF provides a measure about the systematic errors found within subgroups, specifically attributed to some bias or systematic over or under representation of subgroup performance compared to total group performance. Items exhibiting DIF have been avoided as much as possible when operational test forms are selected.

Total test reliability measures (alpha and SEMs) consider the level of consistency (reliability) of performance over all test questions in a given form, the results of which imply how well the questions measure the content domain and could continue to do so over repeated administrations. Total test reliability coefficients (in this case measured by Cronbach's alpha) may range from 0.00 to 1.00, where 1.00 refers to a perfectly consistent test. The data are based on representative samples from each grade (the calibration sample), and they are typical of the results obtained for all CSAP operational tests. The total test reliabilities of the operational forms were evaluated first by Cronbach's α (Cronbach, 1951) index of internal consistency. The specific calculation for Cronbach's α is 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, and $\hat{\sigma}_i^2$ is the variance of item *i* and $\hat{\sigma}_X^2$ is the total test variance. Achievement tests are typically considered of sound reliability when their reliability coefficients are in the range of .80 and above. Tables 17 and 18 in the main table document shows the reliability coefficients for the grades and subject areas involved in the Spring 2006 operational CSAP test administration for content standards and subcontent areas. At the state level, the reliabilities ranged between .86 (Grade 3 Spanish Reading) and .94 (Grades 4-6 Mathematics) with a median value of 0.92. Such a range is indicative of high reliability of the CSAP tests. The median coefficients for each content area are as follows:

Test	Median	Range
Reading (English)	.930	(0.89 – 0.93)
Reading (Spanish)	.885	(0.86 – 0.91)
Mathematics	.930	(0.90 – 0.94)
Writing (English)	.910	(0.91 – 0.93)
Writing (Spanish)	.895	(0.88 – 0.91)
Science	.930	(0.92 – 0.93)

As evidence that a test is performing similarly across various subgroups, the reliability values for these subgroups to those for the total population can be examined. The reliability measures are impacted by the population distribution, and can be lowered when the subgroup is considerably less variable than the total population. However, one would expect the subgroup reliabilities to be adequately high for all groups. Tables G-6 through G-11 show the reliability estimates by Gender, Ethnicity, Free Lunch Eligibility, Immigrant Status, Disability, and Language Proficiency. Even at the subgroup level, the ranges were quite similar and the lowest reliability (.77) was found for the Language Proficiency NEP group, in Reading grade 8. All reliabilities are well within acceptable ranges.

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

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

where SD_SS is the standard deviation of the scale score and $\hat{\alpha}$ is the result of the calculation of Cronbach's α above. The SEMs represent the total standard error of measurement in the scale score metric across all items. The overall estimates of SEM are shown in Table G-1. The SEM for total test and by content area and grade are shown in Tables G-2 through G-5. Tables G-6 through G-11 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 G-12 shows the cut scores used for the proficiency levels at each grade and content area. Comparison of the SEMs at the Proficient cut to the SEMs associated with other CSAP scale scores for each test reveal that these values are among the lowest, meaning that the CSAP tests tend to measure most accurately near the cut score. This is a desirable quality when cut scores are used to classify examinees.

Part 7: Test Validity

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

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

Content-Related Validity

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

The Colorado Model Content Standards and Assessment Frameworks are the foundation for the *CSAP* assessments. All *CSAP* items are developed to measure the content standards and are subject to numerous levels of scrutiny, both internal and external, before their operational use. All items are closely examined according to the "Criteria for Item Acceptability[2]" 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 B-1 through B-4 show for each content area test the number of score reporting categories (SRCs), the number of performance indicators (PIs) in each SRC, the number of items measuring each SRC, the number of PIs assessed by the current test, and finally the percentage of all PIs assessed. It may not be feasible to assess all PIs in a

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

single test; however, as appropriate, efforts are made to assess all measurable PIs across years.

Construct Validity

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

Minimization of Construct-irrelevant Variance and Under-representation

Minimization of construct-irrelevant variance and construct under-representation is addressed in the following steps of the test development process: 1) specification, 2) item writing, 3) review, 4) field testing, 5) test construction, and 6) calibration.

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

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

Minimizing Bias through Differential Item Functioning

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

called large and to determine how these differences will affect the outcome of a particular test.

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

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

The second way bias was minimized was to follow the McGraw-Hill guidelines designed to reduce or eliminate bias. Item writers were directed to the following published guidelines: Guidelines for Bias-Free Publishing (MacMillan/McGraw-Hill, 1993a) and Reflecting Diversity: Multicultural Guidelines for Educational Publishing Professionals (Macmillan/McGraw-Hill, 1993b). Developers reviewed CSAP Assessment materials with these considerations in mind. Such internal editorial reviews were conducted by at least three different people or groups of people: a content editor, who directly supervised the item writers; a style editor, and a content supervisor. The final test was again reviewed by at least these same people, as well as being given an independent review by a quality assurance editor.

As part of the test assembly process, attempts are made to avoid using items with poor statistical fit or distractors with positive item-total correlations, since this may indicate that an item is tapping an ability irrelevant to the construct being measured. Differential item functioning with respect to subgroups might also indicate construct irrelevance. Items with these attributes are not selected or are

given a lower priority for selection during the test construction stage. For CSAP, particular scrutiny is given to the equating (or "anchor") sets in each form, since these items impact the resulting scale scores developed for the entire test. Including DIF items in this equating set could have a greater impact on the overall fairness of the reported scores. More detailed Fit and DIF information for 2006 test assembly are presented in Appendix I.

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

It is believed that these three procedures both improve the quality of an assessment and reduce item and test bias. However, current evidence suggests that expertise in this area is no substitute for data. Reviewers are often wrong about which items perform differently between specific subgroups of students, apparently because some of their ideas about how students will react to items may be inaccurate (Camilli & Shepard, 1994; Sandoval & Mille, 1980; Scheuneman, 1987). Thus, the fourth method for minimizing bias, an empirical approach, was also used to identify potential sources of item bias. For all CSAP tests, differential item functioning (DIF) studies are conducted. DIF studies include a systematic item analysis to determine if examinees with the same underlying level of ability have the same probability of getting the item correct. Items identified with DIF are then examined to determine if item performance differences between identifiable subgroups of the population are due to extraneous or construct-irrelevant information, making the items unfairly difficult. The inclusion of these items is minimized in the test development process. DIF studies have been routinely done for all major test batteries published by CTB/McGraw-Hill after 1970. Differential item functioning of the CSAP assessment items was assessed for both gender and ethnic comparisons.

Because CSAP tests were built using item response theory, 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 (Linn, Levine, Hastings, & Wardrop, 1981; Camilli & Shepard, 1994). However, these procedures require a minimum of 800 to 1000 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 was the procedure used to complete the gender and ethnic DIF studies for the 2005 CSAP operational data.

After the administration of new forms, all items are evaluated for poor item statistics, fit, and DIF. The items flagged for the fit and DIF were noted in the item

analyses report and item pool so that Content experts will be able to reevaluate the items for future selection.

Linn-Harnisch Differential Item Functioning Analyses (DIF) procedure

Because the tests were scored using item response theory, the appropriate procedure for examining DIF is one that reflects that use. A procedure suggested by Linn and Harnisch (1981) was used for the CSAP DIF studies.

An example of this procedure for gender bias analyses follows.

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 ten equal score categories (deciles) based upon 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*. To compute the proportion of students expected to answer item *i* correctly (over all deciles) for a group (e.g., female) the formula 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:

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

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 sub-samples. 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 (1981) procedure was used to measure DIF for constructed-response items.

Differential Item Functioning Ratings

Differential item functioning is defined in terms of the decile group and total target sub-sample 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 sub-sample (see Linn and 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 sub-sample. If D_i is negative, the item is functioning differentially against the target sub-sample.

Results of the Differential Item Functioning Analyses

The DIF analyses were conducted for all grades and content areas for African Americans, Hispanics, Asians, Caucasians, Males, and Females. Table B-5 provides an overview of items flagged for gender and ethnicity DIF in the various assessments based on the entire student population. The results for each assessment are briefly described below.

On the Reading assessments, DIF was most evident at the higher grades. DIF was observed in only one grade 3 Reading item, one grade 4 Reading item and two grade 5 Reading items, compared to four grade 6 Reading items, four grade 7 Reading items, four grade 8 Reading items, five grade 9 Reading items and eight grade 10 Reading items. Across all grades, the Reading items that exhibited DIF tended to favor Asian students (20 items), and disfavored males (nine items). Four items disfavored African American students, one item disfavored Hispanic students, and two items disfavored females.

On the Writing assessments, DIF was observed in grades 4 through10. Across all grades, two items disfavored female students, eight disfavored males, and ten disfavored Asian students. In addition, two Writing items favored Hispanic students, two items favored Asian students, two items favored males, and four items favored females.

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

On the Science assessments, items exhibited DIF in Grades 5, 8 and 10. One item favored females and three items favored Asians. Two items disfavored males, two items disfavored African American students, and one item disfavored Asians.

Book item 89 of grade 7 Reading did not flag for DIF for any of the categories with calibration data. However, it flagged as DIF for every category with the population data. This item also flagged as poor fit with the population data. This item will be removed from the future item pool.

Additional DIF analyses are presented in Tables B-6 (Accommodations), B-7 (Primary Disability State), B-8 (Enrollment), B-9 (Language Proficiency), B-10 (Education Plan), and B-11 (Focal group: Immigrant, Migrant, Homeless).

Internal Structure and Unidimensionality

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

The internal consistency measures, computed as coefficient alpha, for the 2006 *CSAP* tests ranged from 0.86 (Grade 3 Spanish Reading test) to 0.94 (Grades 4, 5, and 6 Mathematics) with a median of 0.92. Coefficient alphas for each test are provided in Table 17 (in the main text tables). Values of 0.90 and above provide strong evidence of internal consistency on the tests. 2006 *CSAP* assessments had strong internal consistency—28 out of 31 tests had coefficient alpha of 0.90 or greater. See Appendix G titled, *Total and Subgroup Reliability* for detailed information about the reliability of the CSAP tests.

When IRT models are used to calibrate test items and to report student scores, demonstrating item fit is also relevant to construct validity. That is, the extent to which test items function as the IRT model in use prescribes is relevant to the validation of test scores. As part of the scaling process, all CSAP items were examined closely with respect to classical (i.e., p-value and item total 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 (3%) on the 2006 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 5 of this Technical Report. Summaries of the IRT fit statistics can be found at the end of Part 5, and detailed lists of these statistics are presented in Tables 81 through 142 of the main text document.

Finally, to assess the overall factor structure of the CSAP assessments, exploratory factor analyses were conducted for each content and grade. Polychoric correlations were obtained, and a principal components factor analysis was conducted. The resulting eigenvalues for each factor are an indication of the relative proportion of variance accounted for by each successive factor. *Appendix H – Factor Analysis Results* contains plots of the eigenvalues and proportions of variance for each factor identified in these analyses. All CSAP

tests demonstrated a strong single factor, accounting for 35-40% of the overall variance, providing evidence that the items in each test are measuring a single construct.

Divergent (Discriminant) Validity

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

To assess the divergent validity of *CSAP* tests, scale scores were obtained and correlated for students who took various *CSAP* content area tests in 2006. Tables B-12 and B-13 show the correlation coefficients among scale scores (and percentile ranks) in different content areas by grade level. The correlation coefficients among scale scores ranged from 0.721 (between Reading and Mathematics in Grade 3) to 0.889 (between Reading and Writing in Grade 9). The correlation coefficients suggest that individual student scores for Reading, Mathematics, Writing, and Science are moderately to highly related. These coefficients are not so low as to call into question whether these tests are tapping into achievement constructs, and not so high as to arouse suspicion that the intended constructs are not distinct.

It is worth noting that the correlation coefficients between Reading and Writing were generally higher than those between Mathematics and Reading and between Mathematics and Writing. It is also interesting to note that Science is correlated with Reading and Mathematics to a similar degree; however the correlation between Science and Writing was relatively lower. A similar pattern of correlations have been observed in *TerraNova*.

Additional evidence of divergent validity can be obtained by evaluating the correlations of test scores with extraneous variables. Correlations were computed between total scale scores and Age, Gender, and Ethnic group. Overall, these correlations were found to be somewhat small, ranging in absolute value from nearly 0 to .39 (see Table B-14). The fact that these correlations are generally greater than zero can be attributed to differences in the overall ability of the various groups.

Predictive Validity

Predictive validity is a type of criterion validity that refers to the degree to which test scores predict criterion measurements that will be made at some point in the future (Crocker & Algina, 1986). In the context of annual assessment of student

proficiency in a content area, the extent to which test scores in a year are predictive of those in the subsequent year can provide evidence for predictive validity. Colorado Model Content Standards in Mathematics, Reading, and Writing are designed to be incremental and progressive from lower to higher grade level, which is the basis for vertical scaling and measuring student growth across years on a common scale. Table B-15 shows predictive validity coefficients measured as the correlation between test scores for two adjacent years (2005-2006) based on matched group of students.

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 B-15 indicate strong predictability of test scores between two adjacent years. The validity coefficients (corrected for attenuation) are generally the highest in Mathematics, indicating a high degree of determination of performance from one year to next. The lowest validity coefficients are between grades 3 and 4. This may be attributed to the relatively short test length at grade 3, differences in content standards between the grades, and relatively large within-student variability across years.

References

AERA, APA, NCME, 1999 American Educational Research Association, American Psychological Association, & 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.

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

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

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

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrica, 16, 297-334.

Dorans and Holland (1993). DIF detection and description: Mantel-Haenszel and standardization. In P.W. Holland and H Wainer (Eds.) Differential Item Function (p 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.

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.

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

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

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

MacMillan/McGraw-Hill (1993a). Guidelines for Bias-Free Publishing.

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

McGraw-Hill (2006). CSAP Bookmark Standard Setting Technical Report for Grades 5 & 10 Science.

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

Sandoval, J. & Mille, M. P. W. (1980). Accuracy of judgments of WISC-R item difficulty for minority groups. Journal of Consulting and Clinical Psychology 48 (2) 249-253.

Scheuneman, J. D. (1987). An experimental, exploratory study of causes of bias in test items. Journal of Educational Measurement 24 (2) 970118.

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

Stone, Ankenmann, Lane & Lia (1993).

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.