TABLE OF CONTENTS

This report presents the results of the statewide Spring 2004 administration of the Colorado Student Assessment Program (CSAP). In Spring 2004, students were assessed in Reading in grades 3 through 10; Writing in grades 3 through 10; Mathematics in grades 5 through 10; and Science in grade 8. 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.

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, some reporting categories were added at the request of the Colorado Department of Education to provide additional diagnostic information; these reporting categories are called "sub-content areas" and 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. Data Analysis, Probability, and Statistics – Students use data collection and analysis, statistics, and probability in problem-solving situations and communicate the reasoning used in solving these problems.

4. Geometric Concepts – 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. Operation and Calculation – 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, grade 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)

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

3. 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)

4. 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)

5. 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)

6. Patterns –
   - 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)

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

8. 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)

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)

10. 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)

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

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

Science

The Colorado Model Content Standards

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

2. Physical Science – Students know and understand common properties, forms, and changes in matter and energy.

3. Life Science – Students know and understand the characteristics and structure of living things, the processes of life, and how living things interact with each other and their environment.

4. Earth and Space Science – Students know and understand the processes and interactions of Earth's systems and the structure and dynamics of Earth and other objects in space.

5. Science – Students understand that science involves a particular way of knowing and understand common connections among scientific disciplines.

6. Technology – Students know and understand interrelationships among science, technology, and human activity and how they can affect the world.

The Colorado Model Sub-Content Areas

1. Experimental Design and Investigations – Students understand and apply scientific questions, hypotheses, variables, and experimental design.

2. Results and Data Analysis – Students organize, analyze, interpret, and predict from scientific data to communicate the results of investigations.

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

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

5. Earth Science – Students know and understand the composition of the earth, its history, and the processes that shape it.

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 2004 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 2004 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 were composed of 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 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 (Reading, 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: 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 5 through 10; and Science (CS 5) and Technology (CS 6) in Science grade 8.

Every item is associated with a content standard but not all items are associated with a sub-content area, so 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, eight items (all multiple-choice) were removed because of poor statistical performance:

- Reading, Grade 4 – Item 41
- Reading, Grade 10 – Item 11
- Spanish Reading, Grade 4 – Item 99
- Writing, Grade 10 – Item 58
- Spanish Writing, Grade 4 – Item 55
- Spanish Writing, Grade 4 – Item 56
- Mathematics, Grade 9 – Item18
- Science, Grade 8 – Item 42

Table 2 indicates the number of items and score points for each test form with and without the deleted items. The numbers after dropping the eight items listed above are shown in parentheses.

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.

Vertical Scale Design for Reading, Writing and Mathematics

Horizontal equating within each grade was used to place the 2004 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 Stocking and Lord (1983) procedure was used to place each grade on the vertical scale that had been developed for each content area.
Each 2004 CSAP test contained items from the previous administrations for the same grade. These repeated items were used as anchors in a Stocking and Lord (1983) equating procedure, which was used to place each test form on the previously established scale. By equating the 2004 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). Detailed information about the cut scores and standard setting are available in the Colorado CSAP Standard Setting Technical Report (2003).

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' tests did not identify gender.

Girls scored higher than boys at all grade levels on the Reading and Writing tests, while boys scored slightly higher than girls on the Science assessment. Boys scored higher than girls on the Grade 5 Mathematics test, but at all other grades the mean scores of male and female students were no more than one point apart, with males scoring one point higher in grades 6 and 10 and females scoring one point higher in grades 7 and 8.

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 that grade 3 Reading measures only one content standard; content standards 1 and 6 are combined in grades 7 through 10 Mathematics; content standards 4 and 5 are combined in grades 5 through 10 Mathematics; and content standards 5 and 6 are combined for grade 8 Science. Similarly, sub-content areas 1 and 4 are combined for grades 3 through 6 Reading.

Third Grade

Reading

The mean scale score for the total population of students taking the 2004 third-grade Reading assessment is 565 with a standard deviation of 75.5. The mean scale score for female students is 573 with a standard deviation of 72.6, and the mean scale score for male students is 557 with a standard deviation of 77.5. The scale score frequency distribution of the third-grade Reading assessment for the total population is shown in Appendix 1. Figure 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 scale score distributions for the total population and for each gender are slightly negatively skewed.

The mean scale scores for the sub-content areas range from 568 to 618. Although most of the sub-content area scale scores are quite close to the total test score median of 569, the median score on sub-content area 3 is notably higher at 578.

The mean percents of the maximum obtainable raw score for the sub-content areas range from 69.4 to 82.9. The mean percent of the maximum obtainable score for the total test is 75.0.

Reading – Spanish Version

The mean scale score for the total population of students taking the 2004 third-grade Spanish Reading assessment is 523 with a standard deviation of 43.6. The mean scale score for female students is 531 with a standard deviation of 43.2, and the mean scale score for male students is 516 with a standard deviation of 42.9.

The scale score frequency distribution of the third-grade Spanish Reading assessment for the total population is shown in Appendix 2. Figure 2 graphically represents the scale score frequency distributions for the total population and for the groups of male and female students separately.

The mean scale scores for the sub-content areas range from 522 to 524; the median scale scores for the sub-content areas vary between 524 and 527, and all are close to the median for the total test scale score of 525.

The mean percents of the maximum obtainable score for the sub-content areas range from 54.0 to 64.0. The mean percent of the maximum obtainable score for the total test is 59.6.

Writing

The mean scale score for the total population of students taking the 2004 third-grade Writing assessment is 469 with a standard deviation of 54.7. The mean scale score for female students is 477 with a standard deviation of 54.9, and the mean scale score for male students is 462 with a standard deviation of 53.4.

The scale score frequency distribution for the total population is shown in Appendix 3. Figure 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 scale score distributions for the total population and for each gender are slightly negatively skewed.

The mean scale scores for the two content standards are 471 and 476, with standard deviations of 61.0 and 66.1, respectively. The mean scale scores for the sub-content areas range from 472 to 501. The median scale scores vary between 469 and 470 for the content standards, and between 470 and 477 for the sub-content areas. The median for the total test scale score is 468.
The mean percents of the maximum obtainable score for CS 2 (Write for a Variety of Purposes) and CS 3 (Write Using Conventions) are 76.3 and 81.6, respectively.

The mean percent of the maximum obtainable score for the total test is 79.3. The mean percents of the maximum obtainable score for the sub-content areas range from 76.6 to 81.8.

Writing – Spanish Version

The mean scale score for the total population of students taking the 2004 third-grade Spanish Writing assessment is 501 with a standard deviation of 63.2. The mean scale score for female students is 517 with a standard deviation of 64.6 and the mean scale score for male students is 487 with a standard deviation of 58.4.

The scale score frequency distribution of the third-grade Spanish Writing assessment for the total population is shown in Appendix 4. Figure 4 graphically represents the scale score frequency distributions for the total population and for the groups of male and female students separately.

The mean scale scores for the two content standards are both 505, with median scale scores of 499 (CS3) and 500 (CS2). The mean scale scores for the sub-content areas range from 510 to 522; the median scale scores for the sub-content areas vary between 501 to 507. The median total scale score is 499.

The mean percents of the maximum obtainable score range from 68.3 to 72.4 for the content standards, and from 64.5 to 73.3 for the sub-content areas. The mean percent of the maximum obtainable score for the total test is 70.6.

Fourth Grade

Reading

The mean scale score for the total population of students taking the 2004 fourth-grade Reading assessment is 585 with a standard deviation of 63.2. The mean scale score for female students is 592 with a standard deviation of 60.6, and the mean scale score for male students is 579 with a standard deviation of 65.0.

The scale score frequency distribution for the total population is shown in Appendix 5. Figure 5 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 586 to 643. The median scale scores vary between 589 and 594 for the content standards, and between 590 and 592 for the sub-content areas. The median for the total test scale score is 591. The mean percents of the maximum obtainable score for the content standards range from 56.6 on CS 4 (Thinking Skills) to 75.2 on CS 1 (Reading Comprehension).

The mean percent of the maximum obtainable score for the total test is 65.4. The mean percents for the sub-content areas range from 62.5 to 75.8.

Reading – Spanish Version

The mean scale score for the total population of students taking the 2004 fourth-grade Spanish Reading assessment is 533 with a standard deviation of 40.1. The mean scale score for female students is 540 with a standard deviation of 38.4, and the mean scale score for male students is 526 with a standard deviation of 40.6.

The scale score frequency distribution for the total population is shown in Appendix 6. Figure 6 graphically represents the scale score frequency distributions for the total population and for the groups of male and female students separately.

The mean scale scores for the content standards range from 529 to 532. The mean scale scores for the sub-content areas range from 532 to 538. The median scale scores vary between 533 and 536 for the content standards, and between

535 and 536 for the sub-content areas, and all are close to the median for the total test scale score, 535.

The mean percents of the maximum obtainable score for the content standards range from 48.8 on CS 6 (Literature) to 63.0 on CS 1 (Reading Comprehension). The mean percent of the maximum obtainable score for the total test is 55.9. The mean percents for the sub-content areas range from 53.3 to 67.0.

Writing

The mean scale score for the total population of students taking the 2004 fourth-grade Writing assessment is 489 with a standard deviation of 54.0. The mean scale score for female students is 499 with a standard deviation of 55.2, and the mean scale score for male students is 480 with a standard deviation of 51.3.

The scale score frequency distribution for the total population is shown in Appendix 7. Figure 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 scale score distributions for the total population and for each gender are approximately normal.

The mean scale scores for the content standards vary between 491 and 496. The mean scale scores for the sub-content areas range from 492 to 517. The median scale scores vary between 488 and 489 for the content standards, and between 489 and 505 for the sub-content areas. The median for the total test scale score is 489. The mean percents of the maximum obtainable score for CS 2 (Write for a Variety of Purposes) and CS 3 (Write Using Conventions) are 71.3 and 77.8, respectively.

The mean percent of the maximum obtainable score for the total test is 74.4. The mean percents of the maximum obtainable score for the sub-content areas range from 64.7 to 78.3.

Writing – Spanish Version

The mean scale score for the total population of students taking the 2004 fourth-grade Spanish Writing assessment is 519 with a standard deviation of 43.7. The mean scale score for female students is 529 with a standard deviation of 41.8, and the mean scale score for male students is 510 with a standard deviation of 43.3.

The scale score frequency distribution for the total population is shown in Appendix 8. Figure 8 graphically represents the scale score frequency distributions for the total population and for the groups of male and female students separately.

The mean scale score for each of the two content standards is 519. The mean scale scores for the sub-content areas range from 508 to 520. The median scale scores for the two content standards are 524 and 523. The median scale scores for the sub-content areas vary between 505 and 523. The median for the total test scale score is 523. The mean percents of the maximum obtainable score for CS 2 (Write for a Variety of Purposes) and CS 3 (Write Using Conventions) are 50.5 and 54.3, respectively.

The mean percent of the maximum obtainable score for the total test is 52.3. The mean percents of the maximum obtainable score for the sub-content areas range from 37.7 to 63.7.

Fifth Grade

Reading

The mean scale score for the total population of students taking the 2004 fifth-grade Reading assessment is 611 with a standard deviation of 70.0. The mean scale score for female students is 618 with a standard deviation of 66.3 and the mean scale score for male students is 605 with a standard deviation of 72.8.

The scale score frequency distribution for the total population is shown in Appendix 9. Figure 9 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 601 to 616. The mean scale scores for the sub-content areas range from 610 to 616. The median scale scores vary between 619 and 620 for the content standards, and between 619 and 621 for the sub-content areas, and all are close to the median for the total test scale score, 619.

The mean percents of the maximum obtainable score for content standards range from 52.4 on CS 4 (Thinking Skills) to 68.0 on CS 1 (Reading Comprehension). The mean percent of the maximum obtainable score for the total test is 63.0. The mean percents for the sub-content areas range from 60.2 to 63.6.

Writing

The mean scale score for the total population of students taking the 2004 fifth-grade Writing assessment is 505 with a standard deviation of 57.5. The mean scale score for female students is 515 with a standard deviation of 57.4, and the mean scale score for male students is 496 with a standard deviation of 56.0.

The scale score frequency distribution for the total population is shown in Appendix 10. Figure 10 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 506 and 510. The mean scale scores for the sub-content areas range from 505 to 525. The median scale scores vary between 505 and 506 for the content standards, and between 500 and 508 for the sub-content areas. Most are close to the median for the total test scale score, 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 69.1 and 74.2, respectively. The mean percent of the maximum obtainable score for the total test is 71.5. The mean percents of the maximum obtainable score for the sub-content areas range from 63.8 to 74.9.

Mathematics

The mean scale score for the total population of students taking the 2004 fifth-grade Mathematics assessment is 509 with a standard deviation of 74.7. The mean scale score for female students is 507 with a standard deviation of 72.7, and the mean scale score for male students is 511 with a standard deviation of 76.5.

The scale score frequency distribution for the total population is shown in Appendix 11. Figure 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 approximately normal.

The mean scale scores for the content standards range from 513 to 521. The mean scale scores for the sub-content areas range from 520 to 529. The median scale scores vary between 507 and 514 for the content standards, and between 507 and 516 for the sub-content areas. The median for the total test scale score is 510. The mean percents of the maximum obtainable score for the content standards range from 63.3 on CS 4/5 (Geometry and Measurement) to 73.6 on CS 6 (Computational Techniques).

The mean percent of the maximum obtainable score for the total test is 68.3. The mean percents for the sub-content areas range from 60.7 to 72.1.

Sixth Grade

Reading

The mean scale score for the total population of students taking the 2004 sixth-grade Reading assessment is 623 with a standard deviation of 71.9. The mean scale score for female students is 632 with a standard deviation of 67.5, and the mean scale score for male students is 615 with a standard deviation of 74.8.

The scale score frequency distribution for the total population is shown in Appendix 12. Figure 12 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 625 to 629. The mean scale scores for the sub-content areas range from 624 to 636. The median scale scores vary between 630 and 631 for the content standards, and between 630 and 631 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 62.5 on CS 5 (Use of Literary Information) to 65.8 on CS 1 (Reading Comprehension).

The mean percent of the maximum obtainable score for the total test is 64.2. The mean percents for the sub-content areas range from 63.5 to 67.6

Writing

The mean scale score for the total population of students taking the 2004 sixth-grade Writing assessment is 523 with a standard deviation of 63.5. The mean scale score for female students is 537 with a standard deviation of 62.3, and the mean scale score for male students is 510 with a standard deviation of 61.7.

The scale score frequency distribution for the total population is shown in Appendix 13. Figure 13 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 524 and 528. The mean scale scores for the sub-content areas range from 523 to 559. The median scale scores vary between 522 and 525 for the content standards, and between 522 and 547 for the sub-content areas. The median for the total test scale score is 523. The mean percents of the maximum obtainable score for CS 2 (Write for a Variety of Purposes) and CS 3 (Write Using Conventions) are 67.5 and 70.2, respectively.

The mean percent of the maximum obtainable score for the total test is 68.7. The mean percents of the maximum obtainable score for the sub-content areas range from 62.2 to 76.6.

Mathematics

The mean scale score for the total population of students taking the 2004 sixth-grade Mathematics assessment is 523 with a standard deviation of 77.4. The mean scale score for female students is 522 with a standard deviation of 74.8, and the mean scale score for male students is 523 with a standard deviation of 79.7.

The scale score frequency distribution for the total population is shown in Appendix 14. Figure 14 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 very slightly negatively skewed.

The mean scale scores for the content standards range from 517 to 539. The mean scale scores for sub-content areas range from 525 to 526. The median scale scores vary between 524 and 530 for the content standards, and between 525 and 527 for the sub-content areas, and all are close to the median for the total test scale score, 527. The mean percents of the maximum obtainable score for the content standards range from 52.2 on CS 1 (Number Sense) to 70.6 on CS 6 (Computational Techniques).

The mean percent of the maximum obtainable score for the total test is 62.9. The mean percents of the maximum obtainable score for the sub-content areas range from 57.5 to 62.5.

Seventh Grade

Reading

The mean scale score for the total population of students taking the 2004 seventh-grade Reading assessment is 631 with a standard deviation of 68.1. The mean scale score for female students is 641 with a standard deviation of 63.2, and the mean scale score for male students is 621 with a standard deviation of 71.1.

The scale score frequency distribution for the total population is shown in Appendix 15. Figure 15 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 629 to 636. The mean scale scores for the sub-content areas range from 632 to 638. The median scale scores vary between 638 and 640 for the content standards, and between 638 and 641 for the sub-content areas, and all are close to the median for the total test scale score, 639.

The mean percents of the maximum obtainable score for the content standards range from 57.5 on CS 5 (Use of Literary Information) to 69.6 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 59.6 to 68.8.

Writing

The mean scale score for the total population of students taking the 2004 seventh-grade Writing assessment is 542 with a standard deviation of 72.0. The mean scale score for female students is 558 with a standard deviation of 69.9, and the mean scale score for male students is 528 with a standard deviation of 70.8.

The scale score frequency distribution for the total population is shown in Appendix 16. Figure 16 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 543 and 545. The mean scale scores for the sub-content areas range from 543 to 574. The median scale scores vary between 542 and 545 for the content standards, and between 543 and 557 for the sub-content areas, and most are close to the median for the total test scale score, 544.

The mean percents of the maximum obtainable score for CS 2 (Write for a Variety of Purposes) and CS 3 (Write Using Conventions) are 67.3 and 62.2, respectively. The mean percent of the maximum obtainable score for the total test is 64.9. The mean percents of the maximum obtainable score for the sub-content areas range from 61.6 to 78.4.

Mathematics

The mean scale score for the total population of students taking the 2004 seventh-grade Mathematics assessment is 539 with a standard deviation of 74.5. The mean scale score for female students is 539 with a standard deviation of

71.6. The mean scale score for male students is 538 with a standard deviation of 77.1.

The scale score frequency distribution for the total population is shown in Appendix 17. Figure 17 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 very slightly negatively skewed for the total population and for each gender.

The mean scale scores for the content standards range from 532 to 542. The mean scale scores for the sub-content areas range from 529 to 534. The median scale scores vary between 542 and 547 for the content standards, and between 542 and 545 for the sub-content areas, and all are close to the median for the total test scale score, 544. The mean percents of the maximum obtainable score for the content standards range from 44.9 on CS 4/5 (Geometry and Measurement) to 58.0 on CS 3 (Statistics and Probability).

The mean percent of the maximum obtainable score for the total test is 53.1. The mean percents of the maximum obtainable score for the sub-content areas range from 35.1 to 48.0.

Eighth Grade

Reading

The mean scale score for the total population of students taking the 2004 eighth-grade Reading assessment is 649 with a standard deviation of 65.2. The mean scale score for female students is 659 with a standard deviation of 60.7, and the mean scale score for male students is 639 with a standard deviation of 67.9.

The scale score frequency distribution for the total population is shown in Appendix 18. Figure 18 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 650 to 653. The mean scale scores for the sub-content areas range from 645 to 663. The median scale scores vary between 654 and 656 for the content standards, and between 654 and 660 for the sub-content areas, and all are fairly close to the median for the total test scale score, 656. The mean percents of the maximum obtainable score for the content standards range from 60.5 on CS 6 (Literature) to 67.4 on CS 1 (Reading Comprehension). The mean percent of the maximum obtainable score for the total test is 64.3. The mean percents of the maximum obtainable score for the sub-content areas range from 50.0 to 73.0.

Writing

The mean scale score for the total population of students taking the 2004 eighth-grade Writing assessment is 553 with a standard deviation of 75.3. The mean scale score for female students is 572 with a standard deviation of 72.6, and the mean scale score for male students is 535 with a standard deviation of 73.3.

The scale score frequency distribution for the total population is shown in Appendix 19. Figure 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 554 and 557. The mean scale scores for the sub-content areas range from 556 to 576. The median scale scores vary between 553 and 557 for the content standards, and between 555 and 583 for the sub-content areas, and most are close to the median for the total test scale score, 556. Both the mean and median scale scores for SA 6 (Extended Writing),576 and 583 respectively, were a bit higher than the mean and 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 68.1 and 64.5, respectively. The mean percent of the maximum obtainable score for the total test is 66.4. The mean percents of the maximum obtainable score for the sub-content areas range from 62.7 to 76.4.

Mathematics

The mean scale score for the total population of students taking the 2004 eighth-grade Mathematics assessment is 555 with a standard deviation of 75.7. The mean scale score for female students is 556 with a standard deviation of 72.1. The mean scale score for male students is 555 with a standard deviation of 79.0.

The scale score frequency distribution for the total population is shown in Appendix 20. Figure 20 graphically represents the frequency distributions for the total population and for the groups of male and female students separately. The scale score distributions are slightly negatively skewed for the total population and for each gender.

The mean scale scores for the content standards range from 550 to 557. The mean scale scores for sub-content areas range from 531 to 560. The median scale scores vary between 561 and 563 for the content standards, and between

558 and 564 for the sub-content areas, and all are fairly close to the median for the total test scale score, 562.

The mean percents of the maximum obtainable score for the content standards range from 36.8 on CS 4/5 (Geometry and Measurement) to 54.3 on CS 2 (Algebra, Patterns, and Functions). The mean percent of the maximum obtainable score for the total test is 46.3. The mean percents of the maximum obtainable score for the sub-content areas range from 25.7 to 53.5.

Science

The mean scale score for the total population of students taking the 2004 eighth-grade Science assessment is 501 with a standard deviation of 63.5. The mean scale score for female students is 499 with a standard deviation of 59.2, and the mean scale score for male students is 503 with a standard deviation of 67.3.

The scale score frequency distribution for the total population is shown in Appendix 21. Figure 21 graphically represents the frequency distributions for the total population and for the groups of male and female students separately. The distributions of the scale scores are slightly negatively skewed for the total population and for each gender.

The mean scale scores for the content standards range from 497 to 511. The mean scale scores for the sub-content areas range from 497 to 521. The median scale scores vary between 507 and 510 for the content standards, and between 499 and 510 for the sub-content areas, and most are very close to the median for the total test scale score, 509.

The mean percents of the maximum obtainable score for the content standards range from 40.0 on CS 4 (Earth and Space Science) to 64.8 on CS 5/6 (Science and Technology). The mean percent of the obtainable score for the total test is 55.2. The mean percents of the maximum obtainable score for the sub-content areas range from 39.1 to 68.4.

Ninth Grade

Reading

The mean scale score for the total population of students taking the 2004 ninth-grade Reading assessment is 661 with a standard deviation of 60.0. The mean scale score for female students is 672 with a standard deviation of 53.3, and the mean scale score for male students is 651 with a standard deviation of 64.0.

The scale score frequency distribution for the total population is shown in Appendix 22. Figure 22 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 659 to 665. The mean scale scores for the sub-content areas range from 662 to 688. The median scale scores vary between 668 and 669 for the content standards, and between 668 and 670 for the sub-content areas, and all are close to the median for the total test scale score, 668.

The mean percents of the maximum obtainable score for the content standards range from 56.6 on CS 5 (Use of Literary Information) to 66.1 on CS 6 (Literature). The mean percent of the maximum obtainable score for the total test is 62.9. The mean percents of the maximum obtainable score for the sub-content areas range from 60.2 to 72.3.

Writing

The mean scale score for the total population of students taking the 2004 ninth-grade Writing assessment is 570 with a standard deviation of 73.4. The mean scale score for female students is 587 with a standard deviation of 70.4, and the mean scale score for male students is 553 with a standard deviation of 72.3.

The scale score frequency distribution for the total population is shown in Appendix 23. Figure 23 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 570 and 574. The mean scale scores for sub-content areas range from 571 to 580. The median scale scores vary between 571 and 573 for the content standards, and between 545 and 574 for the sub-content areas, and most are close to the median for the total test scale score, 572. 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 64.5 and 65.6, respectively. The mean percent of the maximum obtainable score for the total test is 65.0. The mean percents of the maximum obtainable score for the sub-content areas range from 63.1 to 67.5.

Mathematics

The mean scale score for the total population of students taking the 2004 ninth-grade Mathematics assessment is 565 with a standard deviation of 74.6. The mean scale score for female students is 565 with a standard deviation of 70.5, and the mean scale score for male students is 565 with a standard deviation of 78.3.

The scale score frequency distribution for the total population is shown in Appendix 24. Figure 24 graphically represents the frequency distributions for the total population and for the groups of male and female students separately. The scale score distributions are slightly negatively skewed for the total population and for each gender.

The mean scale scores for the content standards range from 557 to 562. The mean scale scores for the sub-content areas are 554 and 561. The median scale scores vary between 573 and 574 for the content standards, and 573 for both of the sub-content areas, and all are close to the median for the total test scale score, 574.

The mean percents of the maximum obtainable score for the content standards range from 34.1 on CS 4/5 (Geometry and Measurement) to 49.2 on CS 2 (Algebra, Patterns, and Functions). The mean percent of the maximum obtainable score for the total test is 40.6. The mean percents of the maximum obtainable score for the sub-content areas range from 38.1 to 42.6.

Tenth Grade

Reading

The mean scale score for the total population of students taking the 2004 tenth-grade Reading assessment is 680 with a standard deviation of 59.6. The mean scale score for female students is 690 with a standard deviation of 53.5, and the mean scale score for male students is 671 with a standard deviation of 63.6.

The scale score frequency distribution for the total population is shown in Appendix 25. Figure 25 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 674 to 687. The mean scale scores for the sub-content areas range from 672 to 685. The median scale scores vary between 685 and 687 for the content standards, and between 685 and 690 for the sub-content areas, and all are close to the median for the total test scale score, 687.

The mean percents of the maximum obtainable score for the content standards range from 48.2 on CS 6 (Literature) to 72.1 on CS 4 (Thinking Skills). The mean percent of the maximum obtainable score for the total test is 63.0. The mean percents of the maximum obtainable score for the sub-content areas range from 46.4 to 70.4.

Writing

The mean scale score for the total population of students taking the 2004 tenth-grade Writing assessment is 579 with a standard deviation of 76.4. The mean scale score for female students is 596 with a standard deviation of 73.1, and the mean scale score for male students is 562 with a standard deviation of 76.0.

The scale score frequency distribution for the total population is shown in Appendix 26. Figure 26 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 580 and 583. The mean scale scores for the sub-content areas range from 581 to 588. The median scale scores vary between 578 and 583 for the content standards, and between 537 and 584 for the sub-content areas, and most are close to the median for the total test scale score, 581. 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 67.5 and 63.8, respectively. The mean percent of the maximum obtainable score for the total test is 65.7. The mean percents of the maximum obtainable score for the sub-content areas vary from 58.1 to 69.9.

Mathematics

The mean scale score for all students taking the 2004 tenth-grade Mathematics assessment is 580 with a standard deviation of 73.1. The mean scale score for female students is 579 with a standard deviation of 69.9, and the mean scale score for male students is 580 with a standard deviation of 76.2.

The scale score frequency distribution for the total population is shown in Appendix 27. Figure 27 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 570 to 580. The mean scale scores for the sub-content areas are 570 and 578. The median scale scores vary between 585 and 590 for the content standards, and between 583 and 588 for the sub-content areas, and most are close to the median for the total test scale score, 588.

The mean percents of the maximum obtainable score for the content standards range from 33.5 on CS 4/5 (Geometry and Measurement) to 48.7 on CS 2 (Algebra, Patterns, and Functions). The mean percent of the maximum obtainable score for the total test is 43.1. The mean percents of the maximum obtainable score for SA 1 and SA 2 are 44.1 and 37.2, respectively.

Correlations Among Content Standards and Among Sub-Content Areas

Tables 13 through 15 show the correlations between the *raw* 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 .68 (grade 10) and .83 (grade 9) for the relationship between the various content standards, and between .57 (grade 9) and .82 (grade 4) for the relationship between the various sub-content areas, respectively.

For the Grade 3 Spanish Reading assessments, correlations among sub-content areas vary between .57 and .65. For the Grade 4 Spanish Reading assessments, the correlations among the various content standards vary between .63 and .75 and correlations among sub-content areas vary between .65 and .76.

For the Writing assessments, the coefficients for the correlation between content standards 2 and 3 vary between .78 (grade 3) and .80 (grades 5 and 9). The correlations among the various sub-content areas vary between .49 (grade 4) and .77 (grade 7).

For the Spanish Writing assessment, 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 .42 and .58.

For the Mathematics assessments, the correlations vary between .71 (grade 6) and .82 (grades 7 and 10) for the relationship among the content standards, and between .63 (grade 10) and .75 (grade 8) for the relationship among the sub-content areas.

Finally, for the Science assessment, the correlation coefficients vary between .63 and .75 for the relationship among the content standards, and between .53 and .73 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 16 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 greater than .85, and all except two (Grade 3 Spanish Reading and Grade 4 Spanish Writing) are greater than .90. These reliability coefficients indicate that the Colorado 2004 assessments had strong internal consistency and that the tests produce relatively stable scores.

Table 16 also shows the reliability coefficients for individual standards. Table 17 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.

Part 4: Item Analyses

Tables 18 through 71 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 (r_itt) 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 18 lists the results of the multiple-choice item analyses for the 2004 third-grade Reading assessment. The point-biserials for all multiple-choice items range from .08 to .55 with a mean of .42. The p-values for these items range from .25 to .98 with a mean of .77.

Table 19 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .44 to .62 with a mean of .56. Their p-values range from .53 to .81 with a mean of .71. 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 .0% to 2.0% for multiple-choice items (Table 18) and .4% to 1.8% for constructed-response items (Table 19).

Reading – Spanish Version

Table 20 lists the results of the multiple-choice item analyses for the Spanish version of the 2004 third-grade Reading assessment. The point-biserials for all multiple-choice items range from .06 to .50 with a mean of .32. The p-values for these items range from .19 to .94 with a mean of .60.

Table 21 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .39 to .56 with a mean of .46. Their p-values range from .38 to .81, 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 3.5% for multiple-choice items (Table 20) and .4% to 2.9% for constructed-response items (Table 21).

Writing

Table 22 lists the results of the multiple-choice item analyses for the 2004 third-grade Writing assessment. The point-biserials for all multiple-choice items range from .25 to .53 with a mean of .38. The p-values for these items range from .60 to .97 with a mean of .82.

Table 23 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .14 to .51 with a mean of .39. Their p-values range from .06 to .95, 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 1.2% for multiple-choice items (Table 22) and from .1% to .2% for constructed-response items (Table 23).

Writing – Spanish Version

Table 24 lists the results of the multiple-choice item analyses for the Spanish version of the 2004 third-grade Writing assessment. The point-biserials for all multiple-choice items range from .02 to .43 with a mean of .34. The p-values for these items range from .28 to .97 with a mean of .72.

Table 25 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .31 to .54 with a mean of .45. Their p-values range from .26 to .94, with a mean of .69. 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 1.4% for multiple-choice items (Table 24) and .3% to 1.1% for constructed-response items (Table 25).

Fourth Grade

Reading

Table 26 lists the results of the multiple-choice item analyses for the 2004 fourth-grade Reading assessment. The point-biserials for the multiple-choice items range from -.08 to .53 with a mean of .40. The p-values for the multiple-choice items range from .16 to .94 with a mean of .72. Item 41, with a point-biserial of -.08 and a p-value of .16, was dropped prior to scoring.

Table 27 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .31 to .61 with a mean of .48. Their p-values range from .20 to .84 with a mean of .56. 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. Over 60% of the students obtained the highest possible score points in 4 out of 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 fourth-grade Reading assessment was low, ranging from .1% to 2.4% for multiple-choice items (Table 26). The range was .3% to 2.3% for constructed-response items (Table 27).

Reading – Spanish Version

Table 28 lists the results of the multiple-choice item analyses for the Spanish version of the 2004 fourth-grade Reading assessment. The point-biserials for all multiple-choice items range from .05 to .53 with a mean of .33. The p-values for these items range from .32 to .95 with a mean of .61. MC item 99, with a point-biserial of .05, did not discriminate between low and high ability students, and thus did not provide useful information on student ability. This item was removed from the operational test results for this reason.

Table 29 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .25 to .61 with a mean of .45. Their p-values range from .25 to .80, 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 reasonable amount of variability in students' responses to most items, indicating that these items work reasonably well over the range of student ability. However, only 0.8% of students obtained the maximum possible score on CR item 44.

The omit rate for the Spanish version of the third-grade Reading assessment was somewhat higher than for most other grades and content areas, ranging from .0% to 7.4% for multiple-choice items (Table 28) and .2% to 8.2% for constructed-response items (Table 29).

Writing

Table 30 lists the results of the multiple-choice item analyses for the 2004 fourth-grade Writing assessment. The point-biserials for all multiple-choice items range from .24 to .49 with a mean of .38. The p-values for these items range from .52 to .98 with a mean of .79.

Table 31 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .06 to .62 with a mean of .41. Their p-values range from .49 to 1.00, with a mean of .74.

The omit rate for the fourth-grade Writing assessment was small, ranging from .1% to 5.9% for multiple-choice items (Table 30) and .0% to 1.5% for constructed-response items (Table 31).

Writing – Spanish Version

Table 32 lists the results of the multiple-choice item analyses for the Spanish version of the 2004 fourth-grade Writing assessment. The point-biserials for all multiple-choice items range from .03 to .50 with a mean of .32. The p-values for the items range from .22 to .95 with a mean of .56. MC items 55 and 56, with point-biserials of .03 and .04, did not discriminate between low or high ability students, and thus did not provide useful information on student ability. These items were removed from the operational test results for this reason. When MC items 55 and 56 are excluded, the point-biserial range is .12 to .50, and the p-value range is .30 to 95.

Table 33 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed response items range from .02 to .57 with a mean of .33. Their p-values range from .06 to .99 with a mean of .51.

The omit rate for the fourth-grade Writing assessment was small, ranging from .0% to 3.4% for multiple-choice items (Table 32) and .0% to 3.2% for constructed-response items (Table 33).

Fifth Grade

Reading

Table 34 lists the results of the multiple-choice item analyses for the 2004 fifth-grade Reading assessment. The point-biserials for the multiple-choice items range from .02 to .53, with a mean of .37. The p-values for the multiple-choice items range from .17 to .93 with a mean of .69.

Table 35 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .36 to .60 with a mean of .49. Their p-values range from .33 to .85 with a mean of .56. The distribution of 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 14 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 2.6% for multiple-choice items (Table 34) and .5% to 4.1% for constructed-response items (Table 35).

Writing

Table 36 lists the results of the multiple-choice item analyses for the 2004 fifth-grade Writing assessment. The point-biserials for all multiple-choice items range from .18 to .54 with a mean of .39. The p-values for these items range from .28 to .99 with a mean of .76.

Table 37 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .05 to .64 with a mean of .40. Their p-values range from .32 to 1.00, with a mean of .66.

The omit rate for the fifth-grade Writing assessment was small, ranging from .1% to 3.1% for multiple-choice items (Table 36) and .0% to 1.0% for constructed-response items (Table 37).

Mathematics

Table 38 lists the results of the multiple-choice item analyses for the 2004 fifth-grade Mathematics assessment. The point-biserials for the multiple-choice items range from .23 to .56, with a mean of .40. The p-values for the multiple-choice items range from .36 to .93 with a mean of .73.

Table 39 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .48 to .68 with a mean of .59. Their p-values range from .46 to .85 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 fifth-grade Mathematics assessment was low, ranging from .0% to 3.9% for multiple-choice items (Table 38) and .2% to 1.1% for constructed-response items (Table 39).

Sixth Grade

Reading

Table 40 lists the results of the multiple-choice item analyses for the 2004 sixth-grade Reading assessment. The point-biserials for the multiple-choice items range from .17 to .58 with a mean of .39. The p-values for the multiple-choice items range from .30 to .94 with a mean of .67.

Table 41 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .39 to .61 with a mean of .51. Their p-values range from .40 to .80 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 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 4 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 small, ranging from .1% to 3.3% for multiple-choice items (Table 40) and .7% to 3.2% for constructed-response items (Table 41).

Writing

Table 42 lists the results of the multiple-choice item analyses for the 2004 sixth-grade Writing assessment. The point-biserials for all multiple-choice items range from .20 to .59 with a mean of .39. The p-values for these items range from .36 to .90 with a mean of .69.

Table 43 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .08 to .66 with a mean of .45. Their p-values range from .32 to .99 with a mean of .73.

The omit rate for the sixth-grade Writing assessment was small, ranging from .1% to 4.0% for multiple-choice items (Table 42) and .0% to 1.4% for constructed-response items (Table 43).

Mathematics

Table 44 lists the results of the multiple-choice item analyses for the 2004 sixth-grade Mathematics assessment. The point-biserials for the multiple-choice items range from .28 to .58, with a mean of .46. The p-values for the multiple-choice items range from .31 to .94 with a mean of .68.

Table 45 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .49 to .72 with a mean of .61. Their p-values range from .22 to .88 with a mean of .56. 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 5 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 reasonable, ranging from 2.4% to 4.1% for multiple-choice items (Table 44) and 2.6% to 4.7% for constructed-response items (Table 45).

Seventh Grade

Reading

Table 46 lists the results of the multiple-choice item analyses for the 2004 seventh-grade Reading assessment. The point-biserials for the multiple-choice items range from .10 to .55 with a mean of .38. The p-values for the multiple-choice items range from .21 to .95 with a mean of .66.

Table 47 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items are positive, ranging from .33 to .59 with a mean of .46. The p-values for the constructed-response items range from .27 to .82 with a mean of .54. An examination 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 only one 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 2004 grade 7 Reading assessment ranged from .1% to 4.4% (Table 46). The percent of students who omitted constructed-response items ranged from .8% to 5.1% (Table 47), with two items having an omit rate equal to or greater than 5%.

Writing

Table 48 lists the results of the multiple-choice item analyses for the 2004 seventh-grade Writing assessment. The point-biserials for all multiple-choice items range from .22 to .57 with a mean of .38. The p-values for these items range from .26 to .92 with a mean of .65.

Table 49 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .08 to .61 with a mean of .40. Their p-values range from .35 to .99, with a mean of .61.

The omit rate for the seventh-grade Writing assessment was small, ranging from .1% to 3.7% for multiple-choice items (Table 48) and .0% to 1.8% for constructed-response items (Table 49).

Mathematics

Table 50 lists the results of the multiple-choice item analyses for the 2004 seventh-grade Mathematics assessment. The point-biserials for the multiple-choice items range from .15 to .53, with a mean of .36. The p-values for the multiple-choice items range from .33 to .94 with a mean of .64.

Table 51 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .31 to .72 with a mean of .57. Their p-values range from .14 to .89 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 .0% to 7.6% for multiple-choice items (Table 50) and .4% to 5.2% for constructed-response items (Table 51). There was only one multiple-choice item and one constructed response item with an omit rate greater than or equal to 5.0.

Eighth Grade

Reading

Table 52 lists the results of the multiple-choice item analyses for the 2004 eighth-grade Reading assessment. The point-biserials for the multiple-choice items range from .18 to .56 with a mean of .37. The p-values for the multiple-choice items range from .25 to .93 with a mean of .56.

Table 53 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .43 to .70 with a mean of .54. Their p-values range from .38 to .79 with a mean of .59. 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 six 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 percent of students who omitted the multiple-choice items in the eighth-grade Reading assessment ranged from .1% to 4.8% (Table 52). The percent of students who omitted the constructed-response items ranged from 1.0% to 9.4% (Table 53), with four items having an omit rate greater than 5%.

Writing

Table 54 lists the results of the multiple-choice item analyses for the 2004 eighth-grade Writing assessment. The point-biserials for all multiple-choice items range from .16 to .52 with a mean of .40. The p-values for these items range from .27 to .92 with a mean of .68.

Table 55 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .08 to .63 with a mean of .42. Their p-values range from .23 to .99, with a mean of .62.

The omit rate for the eighth-grade Writing assessment was small, ranging from .1% to 3.0% for multiple-choice items (Table 54) and .0% to 1.5% for constructed-response items (Table 54).

Mathematics

Table 56 lists the results of the multiple-choice item analyses for the 2004 eighth-grade Mathematics assessment. The point-biserials for the multiple-choice items range from .12 to .54 with a mean of .36. The p-values for these multiple-choice items range from .12 to .82 with a mean of .53.

Table 57 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .37 to .71 with a mean of .58. Their p-values range from .07 to .71 with a mean of .38. An examination 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 6.3% (Table 56). The percent of students who omitted constructed-response items ranged from .4% to 5.0%, with only one multiple-choice item and constructed-response item having an omit rate greater than or equal to 5% (Table 57).

Science

Table 58 lists the results of the multiple-choice item analyses for the 2004 eighth-grade Science assessment. The point-biserials for the multiple-choice items range from .03 to .57 with a mean of .40. The p-values for the multiple-choice items range from .15 to .93 with a mean of .61. Multiple-choice item 42, with a point-biserial of .03 did not discriminate between high and low ability students, so this item was dropped before scoring. After removing this item, the range of point-biserials was .12 to .57.

Table 59 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .29 to .65 with a mean of .48. Their p-values range from .11 to .70 with a mean of .46. 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 0.0% to 4.4% (Table 58). The omit rate for the constructed-response items ranged from .8% to 10.8% (Table 59), with three of the items having an omit rate greater than 5%.

Ninth Grade

Reading

Table 60 lists the results of the multiple-choice item analyses for the 2004 ninth-grade Reading assessment. The point-biserials for the multiple-choice items range from .09 to .57 with a mean of .37. The p-values for the multiple-choice items range from .27 to .89 with a mean of .65.

Table 61 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .28 to .65 with a mean of .54. Their p-values range from .29 to .79 with a mean of .60. 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. Over 50% of the students obtained the highest possible score points in six 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 multiple-choice items for the ninth-grade Reading assessment ranged from .1% to 4.3% (Table 60). The omit rate for the constructed-response items ranged from 2.2% to 10.8%, with seven out of the 14 items having an omit rate greater than 5% (Table 61).

Writing

Table 62 lists the results of the multiple-choice item analyses for the 2004 ninth-grade Writing assessment. The point-biserials for all multiple-choice items range from .14 to .54 with a mean of .41. The p-values for these items range from .29 to .89 with a mean of .67.

Table 63 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .15 to .60 with a mean of .43. Their p-values range from .40 to .99, with a mean of .68.

The omit rate for the ninth-grade Writing assessment was small, ranging from .1% to 4.5% for multiple-choice items (Table 62) and .0% to 4.9% for constructed-response items (Table 63).

Mathematics

Table 64 lists the results of the multiple-choice item analyses for the 2004 ninth-grade Mathematics assessment. The point-biserials for the multiple-choice items range from .07 to .53 with a mean of .35. The p-values for these multiple-choice items range from .11 to .82 with a mean of .53. MC item 18, with a point-biserial of .09, did not discriminate between low and high ability students, and thus did not provide useful information on student ability. This item was removed from the operational test results for this reason. After removing this item, the range of MC point-biserials is .14 to .53.

Table 65 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .47 to .71 with a mean of .60. Their p-values range from .10 to .49 with a mean of .27. An examination 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 1.5% (Table 64). The percent of students who omitted constructed-response items ranged from 1.7% to 5.2% with two out of the 15 constructed-response items having an omit rate greater than 5% (Table 65).

Tenth Grade

Reading

Table 66 lists the results of the multiple-choice item analyses for the 2004 tenth-grade Reading assessment. The point-biserials for the multiple-choice items range from .01 to .55 with a mean of .32. The p-values for the multiple-choice items range from .18 to .93 with a mean of .64. Item 11, with a point-biserial of .01, did not discriminate between high and low ability students, and therefore was removed before scoring. After removing this item, the point-biserials range from .11 to .55.

Table 67 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .36 to .66 with a mean of .55. Their p-values range from .37 to .80 with a mean of .61. 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 five 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 2004 tenth-grade Reading assessment ranged from .1% to 1.8% (Table 66). Omit rates for the constructed-response items ranged from 2.3% to 10.1% (Table 67), with six out of the 14 items having an omit rate greater than or equal to 5%.

Writing

Table 68 lists the results of the multiple-choice item analyses for the 2004 tenth-grade Writing assessment. The point-biserials for all multiple-choice items range from -.08 to .56 with a mean of .37. The p-values for these items range from .05 to .91 with a mean of .67. Item 58, with a point-biserial of -.08 and a p-value of .05, did not discriminate between high and low ability students, and therefore was removed prior to scoring. After removing this item, the point-biserials range from .13 to .56 and the p-values range from .26 to .91.

Table 69 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .13 to .63 with a mean of .46. Their p-values range from .41 to .99, with a mean of .65. The omit rate for the tenth-grade Writing assessment was small, ranging from .1% to 2.8% for multiple-choice items (Table 68) and .0% to 5.1% for constructed-response items (Table 69), with only one item with a p-value greater than or equal to 5%.

Mathematics

Table 70 lists the results of the multiple-choice item analyses for the 2004 tenth-grade Mathematics assessment. The point-biserials for the multiple-choice items range from .20 to .57 with a mean of .40. The p-values for the multiple-choice items range from .20 to .89 with a mean of .52.

Table 71 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .46 to .70 with a mean of .61. The p-values for the constructed-response items range from .08 to .60 with a mean of .34. 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 4.2% to 5.3% (Table 70), with three items having an omit rate greater than or equal to 5%. The omit rate for the constructed-response items ranged from 5.2% to 8.7% (Table 71).

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

$$P_{jk}(\theta) = P\left(x_j = k-1\middle|\theta\right) = \frac{\exp Z_{jk}}{\sum_{i=1}^{m_j}\exp Z_{ji}}, k = 1,\ldots,m_{j,}$$

where m_j is the number of score levels and

$$Z_{jk} \;=\; A_{jk}\,\theta \;+\; C_{jk}$$

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

$$A_{jk} = \alpha_j\,(k-1)$$
$$\mathrm{k} = 1, 2, \ldots, \mathrm{m}_{\mathrm{j}}$$

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).

Scaling and Calibration of the Assessment

The items within each content area were scaled using CTB's computer program PARDUX (Burket, 1993), and a linear transformation was used to translate the PARDUX calibration scale to a unique Colorado scale. The parameter estimates 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.

Summary output tables from the PARDUX program present information on model fit for each item. Model fit information is obtained from the Z-statistic. The Z-statistic is a transformation of the chi-square (Q1) statistic that takes into account differing numbers of score levels as well as sample size:

$$Z_j = \frac{(Q_{1j} - DF)}{\sqrt{2DF}},$$

for the *j*th item. The Z-statistic is an index of the degree to which obtained proportions of students with each item score are close to the proportions that would be predicted by the estimated thetas and item parameters. These values, along with their associated chi-squares (Q1), are computed for ten intervals corresponding to deciles of the theta distribution (Burket, 1991). The chi-square statistic is used to characterize item fit as "good" or "poor."
The estimated item parameters will be used to score the student responses in a given test.

Model Fit

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 70 through 121.

The relationship, Z=N*4/1500, gives the approximate critical Z-value for the CSAP assessments, where N is the sample size for the calibration sample. Fit statistics above this critical Z-value may indicate poor model fit.

Third Grade

The third grade item parameters and fit statistics are shown in Tables 72 to 79. The critical Z-values for these tables are 140.0 for Reading, 4.4 for Spanish Reading, 17.2 for Writing, 3.5 for Spanish Writing.

Across all content areas, only four items exceeded these critical Z-values and exhibited less than optimal fit: 2 Spanish Reading items (MC items 3 and 7), 1 Writing item (CR item 30A), and 1 Spanish Writing item (MC item 34).

Fourth Grade

The fourth-grade item parameters and fit statistics are shown in Tables 80 to 87. The critical Z-values for these tables are 21.1 for Reading, 1.4 for Spanish Reading, 20.1 for Writing, and1.4 for Spanish Writing.

Across all content areas, 23 items exceeded these critical Z-values and exhibited less than optimal fit: 1 Reading item (MC item 100), 12 Spanish Reading items (MC items 23, 29, 31, 34, 47, 91, 92, 113, and CR items 28, 43, 100 and 116), 2 Writing items (CR items 3A and 52), and 8 Spanish Writing items (MC items 65, 71, 74,79, 83, and CR items 1, 3B, and 24).

Fifth Grade

The fifth-grade item parameters and fit statistics are shown in Tables 88 to 93. The critical Z-values for these tables are 18.2 for Reading, 20.5 for Writing, and 19.7 for Mathematics.

Across all content areas, only two items exceeded these critical Z-values and exhibited less than optimal fit. These were constructed-response Writing items 3A and 69.

Sixth Grade

The sixth-grade item parameters and fit statistics are shown in Tables 94 to 99. The critical Z-values for these tables are 20.0 for Reading, 17.5 for Writing, and 20.0 for Mathematics.

Across all content areas, only five items exceeded these critical Z-values and exhibited less than optimal fit: 2 Writing items (CR items 3A and 98) and 3 Mathematics items (CR items 30, 53, and 58).

Seventh Grade

The seventh-grade item parameters and fit statistics are shown in Tables 100 to 105. The critical Z-values for these tables are 20.0 for Reading, 18.1 for Writing, and 20.0 for Mathematics.

Across all content areas, eight items exceeded these critical Z-values and exhibited less than optimal fit: 2 Reading items (MC item 109 and CR item 31), 4 Writing items (CR items 3A, 67, 90, and 118) and 2 Mathematics items (CR items 35 and 44).

Eighth Grade

The eighth-grade item parameters and fit statistics are shown in Tables 106 to 113. The critical Z-values for these tables are 20.0 for Reading, 17.8 for Writing, 20.0 for Mathematics, and 19.9 for Science.

Across all content areas, 11 items exceeded these critical Z-values and exhibited less than optimal fit: 2 Reading items (MC item 114 and CR item 37), 4 Writing items (MC item 67 and CR items 3A, 23, and 94), 3 Mathematics items (MC items 13, 32, and CR item 40), and 2 Science items (MC item 41 and CR item 50).

Ninth Grade

The ninth-grade item parameters and fit statistics are shown in Tables 114 to 119. The critical Z-values for these tables are 20.0 for Reading, 19.7 for Writing, and 20.2 for Mathematics.

Across all content areas, five items exceeded these critical Z-values and exhibited less than optimal fit: 1 Reading item (MC item 105), 3 Writing items (CR items 2C, 53, and 95), and 1 Mathematics item (CR item 33).

Tenth Grade

The tenth-grade item parameters and fit statistics are shown in Tables 120 to 125. The critical Z-values for these tables are 16.9 for Reading, 19.6 for Writing, and 20.0 for Mathematics.

Across all content areas, nine items exceeded these critical Z-values and exhibited less than optimal fit: 2 Reading items (MC items 21 and 103), 4 Writing items (CR items 3A, 76, 99, and 116), and 3 Mathematics items (MC items 3, 52, and CR item14).

Procedures for Detecting and Reducing Bias in CSAP

Four procedures were used to reduce bias in the CSAP Assessments. The first procedure is 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 is paid to content validity.

The second step is to follow the McGraw-Hill guidelines designed to reduce or eliminate bias. Item writers are directed to the following published guidelines: Guidelines for Bias-Free Publishing (McGraw-Hill, 1983) and Reflecting Diversity: Multicultural Guidelines for Educational Publishing Professionals (Macmillan/McGraw-Hill, 1993). Developers review the materials with these considerations in mind.

In the third procedure, educational professionals and community members in the state who represent various gender and ethnic groups review all items. They are asked to consider and comment on the appropriateness of language, subject matter, and representation of people.

It is believed that these three procedures both improve the quality of CSAP and reduce bias. Current evidence, however, suggests that expertise in this area is no substitute for data; reviewers are often wrong about which items work to the disadvantage of a group, apparently because some of their ideas about how students will react to items may be faulty (Sandoval & Mille, 1980; Jensen, 1980; Scheuneman, 1987).

The fourth procedure, an empirical approach, involves statistical procedures referred to as differential item functioning (DIF) analyses.

Differential Item Functioning Analyses

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:

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

where P_{ij} (θ) 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, Males, and Females. Table 126 provides an overview of items flagged for DIF in the various assessments. The results for each assessment are briefly described below.

On the Reading assessments, DIF was most evident at the higher grades. DIF was evident in only one grade 3 Reading item and one grade 4 Reading item, compared with 4 grade 9 and 8 grade 10 Reading items. Across all grades, the Reading items that exhibited DIF tended to favor Hispanic students (7 items), Black students (7 items), and female students (10 items). Two items disfavored Black students, one item disfavored females, and eight disfavored males.

On the Writing assessments, DIF was observed at all grades except grades 5 and 10. Across all grades, two items disfavored Hispanic students, one disfavored Black students, and three disfavored males. In addition, one Writing item favored Hispanic students, and three favored females.

On the Mathematics assessments, DIF was observed at all levels except grade 10. Across all grades, and two items disfavored Black students. In addition, two Mathematics items favored females.

On the Science assessment, two items exhibited DIF. One of these items favored Black students, while the other favored female and Hispanic students and disfavored males and Black students.

None of the Spanish Reading or Spanish Writing items exhibited significant DIF.

Standard Errors of Measurement

Measurement error is associated with every test score. A student’s true score is the hypothetical average score that would result if the test could be administered repeatedly without the effects of practice or fatigue. The standard error of measurement (SEM) can be used to obtain a range within which a student’s true score is likely to fall. The fact that the score for a single test may not represent an individual’s true status gives rise to the need for the standard error.
For example, an obtained score should be regarded not as an absolute value but as a point within a range that probably includes a student’s true score. It is expected that 68% of the time a student’s score obtained from a single testing would fall within one SEM of that student’s true score and that 95% of the time the obtained score would fall within two standard errors of the true score.

Table 127 gives as an overall indication of the standard error of measurement for the scale scores of the CSAP assessments for each grade/content area, the square root of the average value of the variances of the error of measurement associated with each of the scale scores. Tables 128 through 131 provide estimates, based on item response theory, of standard errors of measurement for selected pattern scale scores for each of the CSAP assessments. The tables show that scores closer to the lowest and the highest obtainable scale scores for a particular grade, have higher measurement errors than scores closer to the mean.

References

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

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.

Jensen, A. R. (1980). Bias in mental testing. Free Press, New York.

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.

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.

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.

McGraw-Hill (1983). Guidelines for bias-free publishing.

McMillan/McGraw-Hill (1993). Multicultural guidelines for educational publishing professionals.

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

Sandoval, J. & Mille, M. P. W. (1980). Accuracy of judgements 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.

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

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., Burket, G. R., & Sykes, R. C. (1988). Non-unique solutions to the likelihood equation for the three-parameter logistic model. Paper presented at the meeting of the Psychometric Society, Los Angeles.

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.