Colorado
Student Assessment Program

Technical Report
2003

Submitted to the
Colorado Department of Education

Mc Graw Hill CTB McGraw-Hill

Colorado
Student Assessment Program

Technical Report
2003

Submitted to the
Colorado Department of Education

Mc Graw Hill CTB McGraw-Hill

TABLE OF CONTENTS

This report presents the results of the statewide Spring 2003 administration of the Colorado Student Assessment Program (CSAP). In Spring 2003, 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. <u>Fiction</u> – 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. <u>Nonfiction</u> – 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. <u>Vocabulary</u> – 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. <u>Poetry</u> – 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. <u>Paragraph Writing</u> – Students write and edit in a single session. (Writing)

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

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

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

Mathematics

The Colorado Model Content Standards

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

2. <u>Algebra, Patterns, and Functions</u> – 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. <u>Data Analysis, Probability, and Statistics</u> – Students use data collection and analysis, statistics, and probability in problem-solving situations and communicate the reasoning used in solving these problems.

4. <u>Geometric Concepts</u> – Students use geometric concepts, properties, and relationships in problem-solving situations and communicate the reasoning used in solving these problems.

5. <u>Measurement</u> – 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. <u>Operation and Calculation</u> – 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. <u>Number and Operation Sense</u> – 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. <u>Number Sense</u> – Students demonstrate understanding of the concept of equivalency as related to fractions, decimals, and percents. (SA 1, grade 7)

3. <u>Linear Pattern Representation</u> – 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. <u>Multiple Representations of Linear/Nonlinear Functions</u> – 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. <u>Multiple Representations of Functions</u> – 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. <u>Patterns</u> – 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. <u>Experimental Design and Investigations</u> – Students understand and apply scientific questions, hypotheses, variables, and experimental design.

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

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

4. <u>Chemistry</u> – Students understand the properties, composition, structure, and changes of matter.

5. <u>Earth Science</u> – 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 2003 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 2003 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.

A subset of items used in the 2002 (or 2001) forms of the CSAP Reading, Writing, Mathematics, and Science assessments were also included in the 2003 forms, in order 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. Details of the vertical scaling are provided later in this document. The Spanish version for Writing in grade 3 was administered for the first time in 2003.

Test Configuration

Tables 2 through 6 provide information regarding the configuration of the CSAP assessments. Table 2 provides the number of multiple-choice (MC) items versus 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.

Third Grade

Reading and Writing were assessed using separate forms.

Reading

The third-grade Reading test consisted of 41 items totaling 52 points. There were 34 MC items worth 34 points and 7 CR items worth 18 points.

The numbers of CR items that are worth the various numbers of points are as follows: , 4 items worth 2 points, 2 item worth 3 points, 1 item worth 4 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

    CS 1 has 34 MC and 7 CR (18 points) totaling 52 points.
    SA 1/4 has 10 MC and 7 CR (18 points) totaling 28 points.
    SA 2 has 14 MC and 0 CR (0 points) totaling 14 points.
    SA 3 has 10 MC and 0 CR (0 points) totaling 10 points

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Reading – Spanish Version

The third-grade Spanish Reading test consisted of 40 items totaling 49 points. There were 32 MC items worth 32 points and 8 CR items worth 17 points.

The numbers of CR items that are worth the various numbers of points are as follows: 7 items worth 2 points, 1 item worth 3 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

CS 1 has 32 MC and 8 CR (17 points) totaling 49 points.
SA 1/4 has 12 MC and 7 CR (15 points) totaling 27 points.
SA 2 has 10 MC and 1 CR (2 points) totaling 12 points.
SA 3 has 10 MC and 0 CR (0 points) totaling 10 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Writing

The third-grade Writing test consisted of 53 items totaling 56 points. There were 35 MC items worth 35 points and 18 CR items worth 21 points.

The numbers of CR items that are worth the various numbers of points are as follows:
15 items worth 1 point, 3 items worth 2 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

CS 2 has 7 MC and 9 CR (12 points) totaling 19 points.
CS 3 has 28 MC and 9 CR (9 points) totaling 37 points.
SA 5 has 0 MC and 12 CR (15 points) totaling 15 points.
SA 7 has 16 MC and 3 CR (3 points) totaling 19 points.
SA 8 has 12 MC and 3 CR (3 points) totaling 15 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Writing – Spanish Version

The third-grade Spanish Writing test consisted of 53 items totaling 56 points. There were 35 MC items worth 35 points and 18 CR items worth 21 points.

The numbers of CR items that are worth the various numbers of points are as follows:
15 items worth 1 point, 3 items worth 2 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

CS 2 has 6 MC and 9 CR (12 points) totaling 18 points.
CS 3 has 29 MC and 9 CR (9 points) totaling 38 points.
SA 5 has 0 MC and 12 CR (15 points) totaling 15 points.
SA 7 has 21 MC and 2 CR (2 points) totaling 23 points.
SA 8 has 8 MC and 4 CR (4 points) totaling 12 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Fourth Grade

Reading and Writing were assessed using a single form.

Reading

The fourth-grade Reading test consisted of 70 items totaling 91 points. There were 56 MC items worth 56 points and 14 CR items worth 35 points.

The numbers of CR items that are worth the various numbers of points are as follows:
8 items worth 2 points, 5 items worth 3 points, 1 items worth 4 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

    CS 1 has 25 MC and 4 CR (11 points) totaling 36 points.
    CS 4 has 14 MC and 1 CR (3 points) totaling 17 points.
    CS 5 has 11 MC and 3 CR (7 points) totaling 18 points.
    CS 6 has 6 MC and 6 CR (14 points) totaling 20 points.
    SA 1/4 has 26 MC and 11 CR (28 points) totaling 54 points.
    SA 2 has 11 MC and 0 CR (0 points) totaling 11 points.
    SA 3 has 8 MC and 0 CR (0 points) totaling 8 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Reading – Spanish Version

The fourth-grade Spanish Reading test consisted of 66 items totaling 83 points. There were 52 MC items worth 52 points and 14 CR items worth 31 points.

The numbers of CR items that are worth the various numbers of points are as follows:
2 items worth 1 point, 8 items worth 2 points, 3 items worth 3 points, 1 item worth 4 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

    CS 1 has 27 MC and 1 CR (4 points) totaling 31 points.
    CS 4 has 10 MC and 5 CR (9 points) totaling 19 points.
    CS 5 has 8 MC and 6 CR (14 points) totaling 22 points.
    CS 6 has 7 MC and 2 CR (4 points) totaling 11 points.
    SA 1/4 has 23 MC and 5 CR (9 points) totaling 32 points.
    SA 2 has 14 MC and 3 CR (8 points) totaling 22 points.
    SA 3 has 7 MC and 0 CR (0 points) totaling 7 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Note that these numbers do not include a multiple-choice item that was included on the Spanish Reading assessment but was dropped due to poor fit to the item response model and aberrant student response patterns (discussed further in this report). This was item 107 in session 6 in the operational test booklet.

Writing

The fourth-grade Writing test consisted of 53 items totaling 69 points. There were 40 MC items worth 40 points and 13 CR items worth 29 points.

The numbers of CR items that are worth the various numbers of points are as follows:
7 items worth 1 point, 1 item worth 2 points, 5 items worth 4 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

CS 2 has 13 MC and 6 CR (21 points) totaling 34 points.
CS 3 has 27 MC and 7 CR (8 points) totaling 35 points.
SA 5 has 3 MC and 3 CR (12 points) totaling 15 points.
SA 6 has 0 MC and 4 CR (11 points) totaling 11 points.
SA 7 has 12 MC and 2 CR (2 points) totaling 14 points.
SA 8 has 15 MC and 4 CR (4 points) totaling 19 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Writing – Spanish Version

The fourth-grade Spanish Writing test consisted of 52 items totaling 68 points. There were 39 MC items worth 39 points and 13 CR items worth 29 points.

The numbers of CR items that are worth the various numbers of points are as follows:
7 items worth 1 point, 1 item worth 2 points, 5 items worth 4 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

CS 2 has 5 MC and 6 CR (21 points) totaling 26 points.
CS 3 has 34 MC and 7 CR (8 points) totaling 42 points.
SA 5 has 0 MC and 3 CR (12 points) totaling 12 points.
SA 6 has 0 MC and 4 CR (11 points) totaling 11 points.
SA 7 has 27 MC and 2 CR (2 points) totaling 29 points.
SA 8 has 7 MC and 4 CR (4 points) totaling 11 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Note that these numbers do not include a multiple-choice item that was included on the Spanish Writing assessment but was dropped due to poor fit to the item response model and aberrant student response patterns (discussed further in this report). This was item 54 in session 5 in the operational test booklet.

Fifth Grade

Reading and Writing were assessed using a single form. Mathematics was assessed using a separate form.

Reading

The fifth-grade Reading test consisted of 70 items totaling 90 points. There were 56 MC items worth 56 points and 14 CR items worth 34 points.

The numbers of CR items that are worth the various numbers of points are as follows:
8 items worth 2 points, 6 items worth 3 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

    CS 1 has 17 MC and 2 CR (5 points) totaling 22 points.
    CS 4 has 16 MC and 2 CR (5 points) totaling 21 points.
    CS 5 has 14 MC and 3 CR (8 points) totaling 22 points.
    CS 6 has 9 MC and 7 CR (16 points) totaling 25 points.
    SA 1/4 has 15 MC and 9 CR (21 points) totaling 36 points.
    SA 2 has 18 MC and 2 CR (5 points) totaling 23 points.
    SA 3 has 9 MC and 0 CR (0 points) totaling 9 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Writing

The fifth-grade Writing test consisted of 53 items totaling 69 points. There were 40 MC items worth 40 points and 13 CR items worth 29 points.

The numbers of CR items that are worth the various numbers of points are as follows:
7 items worth 1 point, 1 item worth 2 points, 5 items worth 4 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

    CS 2 has 15MC and 6 CR (21 points) totaling 36 points.
    CS 3 has 25 MC and 7 CR (8 points) totaling 33 points.
    SA 5 has 5 MC and 3 CR (12 points) totaling 17 points.
    SA 6 has 0 MC and 4 CR (11 points) totaling 11 points.
    SA 7 has 18 MC and 2 CR (2 points) totaling 20 points.
    SA 8 has 14 MC and 4 CR (4 points) totaling 18 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Mathematics

The fifth-grade Mathematics test consisted of 69 items totaling 96 points. There were 54 MC items worth 54 points and 15 CR items worth 42 points.

The numbers of CR items that are worth the various numbers of points are as follows:
6 items worth 2 points, 6 items worth 3 points, 3 items worth 4 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

CS 1 has 14 MC and 2 CR (5 points) totaling 19 points.
CS 2 has 8 MC and 4 CR (11 points) totaling 19 points.
CS 3 has 7 MC and 4 CR (12 points) totaling 19 points.
CS 4/5 has 11 MC and 3 CR (9 points) totaling 20 points.
CS 6 has 14 MC and 2 CR (5 points) totaling 19 points.
SA 1 has 12 MC and 3 CR (8 points) totaling 20 points.
SA 2 has 4 MC and 3 CR (9 points) totaling 13 points.
SA 3 has 3 MC and 2 CR (7 points) totaling 10 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Sixth Grade

Reading and Writing were assessed using a single form. Mathematics was assessed using a separate form.

Reading

The sixth-grade Reading test consisted of 69 items totaling 91 points. There were 54 MC items worth 54 points and 15 CR items worth 37 points.

The numbers of CR items that are worth the various numbers of points are as follows:
6 items worth 2 points, 6 items worth 3 points, 3 item worth 4 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

CS 1 has 20 MC and 2 CR (6 points) totaling 26 points.
CS 4 has 14 MC and 3 CR (7 points) totaling 21 points.
CS 5 has 11 MC and 4 CR (9 points) totaling 20 points.
CS 6 has 9 MC and 6 CR (15 points) totaling 24 points.
SA 1/4 has 17 MC and 7 CR (16 points) totaling 33 points.
SA 2 has 13 MC and 4 CR (12 points) totaling 25 points.
SA 3 has 13 MC and 0 CR (0 points) totaling 13 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Writing

The sixth-grade Writing test consisted of 53 items totaling 69 points. There were 40 MC items worth 40 points and 13 CR items worth 29 points.

The numbers of CR items that are worth the various numbers of points are as follows:
7 items worth 1 point, 1 item worth 2 points, 5 items worth 4 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

CS 2 has 17 MC and 6 CR (21 points) totaling 38 points.
CS 3 has 23 MC and 7 CR (8 points) totaling 31 points.
SA 5 has 9 MC and 3 CR (12 points) totaling 21 points.
SA 6 has 0 MC and 4 CR (11 points) totaling 11 points.
SA 7 has 21 MC and 1 CR (1 points) totaling 22 points.
SA 8 has 8 MC and 5 CR (5 points) totaling 13 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Mathematics

The sixth-grade Mathematics test consisted of 60 items totaling 87 points. There were 45 MC items worth 45 points and 15 CR items worth 42 points.

The numbers of CR items that are worth the various numbers of points are as follows:
6 items worth 2 points, 6 items worth 3 points, 3 items worth 4 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

CS 1 has 9 MC and 2 CR (7 points) totaling 16 points.
CS 2 has 7 MC and 3 CR (10 points) totaling 17 points.
CS 3 has 7 MC and 4 CR (11 points) totaling 18 points.
CS 4/5 has 12 MC and 4 CR (10 points) totaling 22 points.
CS 6 has 10 MC and 2 CR (4 points) totaling 14 points.
SA 1 has 12 MC and 4 CR (11 points) totaling 23 points.
SA 2 has 7 MC and 2 CR (6 points) totaling 13 points.
SA 3 has 7 MC and 2 CR (5 points) totaling 12 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Seventh Grade

Reading and Writing were assessed using a single form. Mathematics was assessed using a separate form.

Reading

The seventh-grade Reading test consisted of 71 items totaling 91 points. There were 57 MC items worth 57 points and 14 CR items worth 34 points.

The numbers of CR items that are worth the various numbers of points are as follows:
9 items worth 2 points, 4 items worth 3 points, 1 item worth 4 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

CS 1 has 21 MC and 3 CR (9 points) totaling 30 points.
CS 4 has 13 MC and 3 CR (6 points) totaling 19 points.

CS 5 has 15 MC and 3 CR (6 points) totaling 21 points.
CS 6 has 8 MC and 5 CR (13 points) totaling 21 points.
SA 1 has 12 MC and 4 CR (11 points) totaling 23 points.
SA 2 has 11 MC and 3 CR (8 points) totaling 19 points.
SA 3 has 14 MC and 0 CR (0 points) totaling 14 points.
SA 4 has 5 MC and 4 CR (9 points) totaling 14 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Writing

The seventh-grade Writing test consisted of 53 items totaling 69 points. There were 40 MC items worth 40 points and 13 CR items worth 29 points.

The numbers of CR items that are worth the various numbers of points are as follows:
7 items worth 1 point, 1 item worth 2 points, 5 items worth 4 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

CS 2 has 20 MC and 6 CR (21 points) totaling 41 points.
CS 3 has 20 MC and 7 CR (8 points) totaling 28 points.
SA 5 has 6 MC and 3 CR (12 points) totaling 18 points.
SA 6 has 0 MC and 4 CR (11 points) totaling 11 points.
SA 7 has 13 MC and 2 CR (2 points) totaling 15 points.
SA 8 has 12 MC and 4 CR (4 points) totaling 16 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Mathematics

The seventh-grade Mathematics test consisted of 60 items totaling 87 points. There were 45 MC items worth 45 points and 15 CR items worth 42 points.

The numbers of CR items that are worth the various numbers of points are as follows:
6 items worth 2 points, 6 items worth 3 points, 3 items worth 4 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

CS 1/6 has 15 MC and 4 CR (11 points) totaling 26 points.
CS 2 has 9 MC and 3 CR (8 points) totaling 17 points.
CS 3 has 9 MC and 3 CR (9 points) totaling 18 points.
CS 4/5 has 12 MC and 5 CR (14 points) totaling 26 points.
SA 1 has 6 MC and 2 CR (5 points) totaling 11 points.
SA 2 has 3 MC and 3 CR (9 points) totaling 12 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Eighth Grade

Reading and Writing were assessed using a single form. Mathematics and Science were assessed using two separate forms.

Reading

The eighth-grade Reading test consisted of 69 items totaling 90 points. There were 54 MC items worth 54 points and 15 CR items worth 36 points.

The numbers of CR items that are worth the various numbers of points are as follows: 10 items worth 2 points, 4 items worth 3 points, 1 item worth 4 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

CS 1 has 15MC and 6 CR (14 points) totaling 29 points.
CS 4 has 15 MC and 2 CR (6 points) totaling 21 points.
CS 5 has 12 MC and 3 CR (6 points) totaling 18 points.
CS 6 has 12 MC and 4 CR (10 points) totaling 22 points.
SA 1 has 11 MC and 3 CR (8 points) totaling 19 points.
SA 2 has 16 MC and 5 CR (10 points) totaling 26 points.
SA 3 has 10 MC and 0 CR (0 points) totaling 10 points.
SA 4 has 5 MC and 4 CR (12 points) totaling 17 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Note that these numbers do not include a multiple-choice item that was included on the Reading assessment but was dropped due to poor fit to the item response model and aberrant student response patterns (discussed further in this report). This was item 117 in session 6 in the operational test booklet.

Writing

The eighth-grade Writing test consisted of 53 items totaling 69 points. There were 41 MC items worth 41 points and 12 CR items worth 28 points.

The numbers of CR items that are worth the various numbers of points are as follows: 6 items worth 1 point, 1 item worth 2 points, 5 items worth 4 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

CS 2 has 21 MC and 6 CR (21 points) totaling 42 points.
CS 3 has 20 MC and 6 CR (7 points) totaling 27 points.
SA 5 has 12 MC and 3 CR (12 points) totaling 24 points.
SA 6 has 0 MC and 4 CR (11 points) totaling 11 points.
SA 7 has 18 MC and 1 CR (1 point) totaling 19 points.
SA 8 has 10 MC and 4 CR (4 points) totaling 14 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Note that these numbers do not include a constructed-response item that was included on the Writing assessment but was dropped due to poor fit to the item response model and aberrant student response patterns (discussed further in this report). This was item 2A in session 2 in the operational test booklet.

Mathematics

The eighth-grade Mathematics test consisted of 59 items totaling 86 points. There were 44 MC items worth 44 points and 15 CR items worth 42 points.

The numbers of CR items that are worth the various numbers of points are as follows:
6 items worth 2 points, 6 items worth 3 points, 3 items worth 4 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

    CS 1/6 has 10 MC and 4 CR (11 points) totaling 21 points.
    CS 2 has 11 MC and 4 CR (11 points) totaling 22 points.
    CS 3 has 9 MC and 3 CR (8 points) totaling 17 points.
    CS 4/5 has 14 MC and 4 CR (12 points) totaling 26 points.
    SA 1 has 5 MC and 3 CR (9 points) totaling 14 points.
    SA 2 has 9 MC and 6 CR (15 points) totaling 24 points.
    SA 3 has 5 MC and 2 CR (7 points) totaling 12 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Note that these numbers do not include a multiple-choice item that was included on the Mathematics assessment but was dropped due to poor fit to the item response model and aberrant student response patterns (discussed further in this report). This was item 15 in session 1 in the operational test booklet.

Science

The eighth-grade Science test consisted of 84 items totaling 99 points. There were 64 MC items worth 64 points and 20 CR items worth 35 points.

The numbers of CR items that are worth the various numbers of points are as follows:
8 items worth 1 point, 10 items worth 2 points, 1 item worth 3 points, 1 item worth 4 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

    CS 1 has 13 MC and 8 CR (14 points) totaling 27 points.
    CS 2 has 16 MC and 2 CR (4 points) totaling 20 points.
    CS 3 has 11 MC and 5 CR (9 points) totaling 20 points.

CS 4 has 16 MC and 3 CR (6 points) totaling 22 points.
CS 5/6 has 8 MC and 2 CR (2 points) totaling 10 points.
SA 1 has 6 MC and 3 CR (5 points) totaling 11 points.
SA 2 has 7 MC and 4 CR (7 points) totaling 14 points.
SA 3 has 8 MC and 1 CR (2 point) totaling 10 points.
SA 4 has 8 MC and 1 CR (2 points) totaling 10 points.
SA 5 has 11 MC and 3 CR (6 points) totaling 17 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Ninth Grade

Reading and Writing were assessed using a single form. Mathematics was assessed using a separate form.

Reading

The ninth-grade Reading test consisted of 73 items totaling 96 points. There were 58 MC items worth 58 points and 15 CR items worth 38 points.

The numbers of CR items that are worth the various numbers of points are as follows:
8 items worth 2 points, 6 items worth 3 points, 1 item worth 4 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

CS 1 has 15 MC and 6 CR (14 points) totaling 29 points.
CS 4 has 13 MC and 4 CR (11 points) totaling 24 points.
CS 5 has 16 MC and 1 CR (3 points) totaling 19 points.
CS 6 has 14 MC and 4 CR (10 points) totaling 24 points.
SA 1 has 12 MC and 3 CR (9 points) totaling 21 points.
SA 2 has 12 MC and 6 CR (13 points) totaling 25 points.
SA 3 has 12 MC and 0 CR (0 points) totaling 12 points.
SA 4 has 6 MC and 5 CR (13 points) totaling 19 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Writing

The ninth-grade Writing test consisted of 53 items totaling 69 points. There were 40 MC items worth 40 points and 13 CR items worth 29 points.

The numbers of CR items that are worth the various numbers of points are as follows:
7 items worth 1 point, 1 item worth 2 points, 5 items worth 4 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

CS 2 has 20 MC and 6 CR (21 points) totaling 41 points.
CS 3 has 20 MC and 7 CR (8 points) totaling 28 points.
SA 5 has 10 MC and 3 CR (12 points) totaling 22 points.
SA 6 has 0 MC and 4 CR (11 points) totaling 11 points.
SA 7 has 14 MC and 2 CR (2 points) totaling 16 points.
SA 8 has 10 MC and 4 CR (4 points) totaling 14 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Mathematics

The ninth-grade Mathematics test consisted of 59 items totaling 86 points. There were 44 MC items worth 44 points and 15 CR items worth 42 points.

The numbers of CR items that are worth the various numbers of points are as follows:
 6 items worth 2 points, 6 items worth 3 points, 3 items worth 4 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

CS 1/6 has 10 MC and 3 CR (7 points) totaling 17 points.
CS 2 has 11 MC and 5 CR (14 points) totaling 25 points.
CS 3 has 11 MC and 4 CR (11 points) totaling 22 points.
CS 4/5 has 12 MC and 3 CR (10 points) totaling 22 points.
SA 1 has 3 MC and 4 CR (13 points) totaling 16 points.
SA 2 has 3 MC and 3 CR (10 points) totaling 13 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Note that these numbers do not include a multiple-choice item that was included on the Mathematics assessment but was dropped due to poor fit to the item response model and aberrant student response patterns (discussed further in this report). This was item 51 in session 3 in the operational test booklet.

Tenth Grade

Reading and Writing were assessed using a single form. Mathematics was assessed using a separate form.

Reading

The tenth-grade Reading test consisted of 75 items totaling 101 points. There were 61 MC items worth 61 points and 14 CR items worth 40 points.

The numbers of CR items that are worth the various numbers of points are as follows:
3 items worth 2 points, 10 items worth 3 points, 1 item worth 4 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

CS 1 has 21 MC and 1 CR (3 points) totaling 24 points.
CS 4 has 10 MC and 5 CR (14 points) totaling 24 points.
CS 5 has 12 MC and 3 CR (9 points) totaling 21 points.
CS 6 has 18 MC and 5 CR (14 points) totaling 32 points.
SA 1 has 16 MC and 6 CR (17 points) totaling 33 points.
SA 2 has 15 MC and 3 CR (8 points) totaling 23 points.
SA 3 has 10 MC and 0 CR (0 points) totaling 10 points.
SA 4 has 8 MC and 2 CR (6 points) totaling 14 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Writing

The tenth-grade Writing test consisted of 53 items totaling 69 points. There were 40 MC items worth 40 points and 13 CR items worth 29 points.

The numbers of CR items that are worth the various numbers of points are as follows:
7 items worth 1 point, 1 item worth 2 points, 5 items worth 4 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

CS 2 has 20 MC and 6 CR (21 points) totaling 41 points.
CS 3 has 20 MC and 7 CR (8 points) totaling 28 points.
SA 5 has 9 MC and 3 CR (12 points) totaling 17 points.
SA 6 has 0 MC and 4 CR (11 points) totaling 11 points.
SA 7 has 18 MC and 4 CR (4 points) totaling 22 points.
SA 8 has 8 MC and 2 CR (2 points) totaling 10 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Mathematics

The tenth-grade Mathematics test consisted of 58 items totaling 85 points. There were 43 MC items worth 43 points and 15 CR items worth 42 points.

The numbers of CR items that are worth the various numbers of points are as follows:
6 items worth 2 points, 6 items worth 3 points, 3 items worth 4 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

CS 1/6 has 12 MC and 2 CR (5 points) totaling 17 points.
CS 2 has 11 MC and 5 CR (14 points) totaling 25 points.
CS 3 has 10 MC and 4 CR (11 points) totaling 21 points.
CS 4/5 has 10 MC and 4 CR (12 points) totaling 22 points.

SA 1 has 4 MC and 3 CR (9 points) totaling 13 points.
SA 2 has 5 MC and 2 CR (5 points) totaling 10 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Note that these numbers do not include two multiple-choice items that were included on the Mathematics assessment but was dropped due to poor fit to the item response model and aberrant student response patterns (discussed further in this report). These were items 53 and 57 in session 4 in the operational test booklet.

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 items on the test. 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).

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 2003 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, had been established in 2001. The vertical scales for Writing, spanning grades 3 through 10, and for Mathematics, spanning grades 5

through 10, had been 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.

The 2003 CSAP tests for each grade of Reading, Writing, and Mathematics contained anchor items from the previous administration (2002) for the same grade, that allowed the 2003 tests to be anchored to the previously established scales using the Stocking and Lord (1983) procedure.

By means of these equatings within each grade, the unique metrics of the CSAP Reading, Writing, and Mathematics vertical scales were maintained. The vertical scales had a lowest obtainable scale score (LOSS) of 150 in grade 3 Reading and Writing, a LOSS of 220 in grade 5 Mathematics, a highest obtainable scale score (HOSS) of 999 in grade 10 Reading, and a HOSS of 950 in grade 10 Writing and Mathematics.

The scaling and equating design for Science and for the Spanish versions of the grades 3 and 4 Reading assessment, and the grade 4 Writing assessment, remained unchanged from the previous year. The 2003 CSAP tests for Spanish reading, Spanish Writing, and Science contained anchor items from the previous administration (2002) that allowed the 2003 tests to be anchored to the previously established scale using the Stocking and Lord (1983) procedure.

These scaling and calibration methods are presented in Part 5.

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). New cut scores were set for the Spanish version of the grade 3 Writing assessment. For all other assessments the cut scores from the previous year were used. 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. 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, means as well as 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 2003 third-grade Reading assessment is 568 with a standard deviation of 75.1. The mean scale score for female students is 575 with a standard deviation of 72.8, and the mean scale score for male students is 562 with a standard deviation of 76.7.

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 573 to 606. The median scale scores vary between 566 and 576, and all are close to the median for the total test scale score, 570.

The mean percents of the maximum obtainable score for the sub-content areas range from 72.0 to 83.8. The mean percent of the maximum obtainable score for the total test is 75.8.

Reading – Spanish Version

The mean scale score for the total population of students taking the 2003 third-grade Spanish Reading assessment is 523 with a standard deviation of 41.6. The mean scale score for female students is 530 with a standard deviation of 40.6, and the mean scale score for male students is 516 with a standard deviation of 41.5.

The scale score frequency distribution of the third-grade 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 525 to 538; the median scale scores for the sub-content areas vary between 524 and 526, and all are close to the median for the total test scale score, 525.

The mean percents of the maximum obtainable score for the sub-content areas range from 60.4 to 70.7. The mean percent of the maximum obtainable score for the total test is 66.5.

Writing

The mean scale score for the total population of students taking the 2003 third-grade Writing assessment is 477 with a standard deviation of 57.5. The mean scale score for female students is 485 with a standard deviation of 58.2, and the mean scale score for male students is 468 with a standard deviation of 55.7.

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 mean scale scores for the content standards vary between 481 and 490, with standard deviations of 67.2 and 83.5, respectively. The mean scale scores for the sub-content areas range from 483 to 513. The median scale scores vary between 474 and 497 for the content standards, and between 478 and 484 for the sub-content areas. The median for the total test scale score is 475.

The mean percents of the maximum obtainable score for CS 2 (Write for a Variety of Purposes) and CS 3 (Write Using Conventions) are 80.0 and 85.6, respectively. The mean percent of the maximum obtainable score for the total test is 83.7. The mean percents of the maximum obtainable score for the sub-content areas range from 78.5 to 84.6.

Writing – Spanish Version

The mean scale score for the total population of students taking the 2003 third-grade Spanish Writing assessment is 500 with a standard deviation of 57.8. The mean scale score for female students is 511 with a standard deviation of 55.2, and the mean scale score for male students is 488 with a standard deviation of 58.3.

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 content standards range from 502 to 504; the median scale scores vary between 500 and 501, and all are close to the median for the total test scale score, 499. The mean scale scores for the sub-content areas range from 509 to 513; the median scale

scores for the sub-content areas vary between 500 and 504, and all are close to the median for the total test scale score, 499.

The mean percents of the maximum obtainable score range from 60.2 to 74.6 for the content standards, and from 65.8 to 74.3 for the sub-content areas. The mean percent of the maximum obtainable score for the total test is 70.0.

Fourth Grade

Reading

The mean scale score for the total population of students taking the 2003 fourth-grade Reading assessment is 587 with a standard deviation of 64.6. The mean scale score for female students is 594 with a standard deviation of 61.9, and the mean scale score for male students is 580 with a standard deviation of 66.3.

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 586 to 595. The mean scale scores for the sub-content areas range from 587 to 629. The median scale scores vary between 591 and 595 for the content standards, and between 592 and 604 for the sub-content areas, and all are close to the median for the total test scale score, 592.

The mean percents of the maximum obtainable score for the content standards range from 55.0 on CS 6 (Literature) to 76.0 on CS 1 (Reading Comprehension). The mean percent of the maximum obtainable score for the total test is 66.2. The mean percents for the sub-content areas range from 64.5 to 74.9. The observed ranges indicate that the average item difficulty differed somewhat across the content standards and across the sub-content areas.

Reading – Spanish Version

The mean scale score for the total population of students taking the 2003 fourth-grade Spanish Reading assessment is 527 with a standard deviation of 44.2. The mean scale score for female students is 533 with a standard deviation of 40.2, and the mean scale score for male students is 521 with a standard deviation of 46.4.

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 523 to 528. The mean scale scores for the sub-content areas range from 523 to 531. The median scale scores vary between 528 and 531 for the content standards, and between 530 and 532 for the sub-content areas, and all are close to the median for the total test scale score, 532.

The mean percents of the maximum obtainable score for the content standards range from 42.4 on CS 5 (Use of Literary Information) to 61.4 on CS 1 (Reading Comprehension). The mean percent of the maximum obtainable score for the total test is 52.8. The mean percents for the sub-content areas range from 53.8 to 58.1. The sub-content areas appear to be similar in average item difficulty, whereas the content standards appear to vary somewhat.

Writing

The mean scale score for the total population of students taking the 2003 fourth-grade Writing assessment is 486 with a standard deviation of 55.1. The mean scale score for female students is 497 with a standard deviation of 55.2, and the mean scale score for male students is 477 with a standard deviation of 53.3.

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 486 and 491. The mean scale scores for the sub-content areas range from 486 to 511. The median scale scores vary between 486 and 488 for the content standards, and between 479 and 494 for the sub-content areas, and all are fairly close to the median for the total test scale score, 487.

The mean percents of the maximum obtainable score for CS 2 (Write for a Variety of Purposes) and CS 3 (Write Using Conventions) are 68.7 and 73.3, respectively. The mean percent of the maximum obtainable score for the total test is 71.0. The mean percents of the maximum obtainable score for the sub-content areas range from 61.6 to 80.2. Note that the mean percents are closer for the content standards than for the sub-content areas.

Writing – Spanish Version

The mean scale score for the total population of students taking the 2003 fourth-grade Spanish Writing assessment is 519 with a standard deviation of 43.0. The mean scale score for female students is 525 with a standard deviation of 42.2, and the mean scale score for male students is 513 with a standard deviation of 42.9.

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 scores for the content standards vary between 519 and 523. The mean scale scores for the sub-content areas range from 508 to 524. The median scale scores vary between

516 and 522 for the content standards, and between 515 and 522 for the sub-content areas, and most are close to the median for the total test scale score, 520.

The mean percents of the maximum obtainable score for CS 2 (Write for a Variety of Purposes) and CS 3 (Write Using Conventions) are 54.9 and 54.2, respectively. The mean percent of the maximum obtainable score for the total test is 54.4. The mean percents of the maximum obtainable score for the sub-content areas range from 41.1 to 58.3.

Fifth Grade

Reading

The mean scale score for the total population of students taking the 2003 fifth-grade Reading assessment is 608 with a standard deviation of 68.7. The mean scale score for female students is 615 with a standard deviation of 64.8 and the mean scale score for male students is 601 with a standard deviation of 71.5.

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 608 to 620. The mean scale scores for the sub-content areas range from 609 to 676. The median scale scores vary between 614 and 616 for the content standards, and between 614 and 619 for the sub-content areas, and all are close to the median for the total test scale score, 615.

The mean percents of the maximum obtainable score for content standards range from 59.9 on CS 5 (Use of Literary Information) to 74.0 on CS 1 (Reading Comprehension). The mean percent of the maximum obtainable score for the total test is 64.8. The mean percents for the sub-content areas range from 62.8 to 79.6.

Writing

The mean scale score for the total population of students taking the 2003 fifth-grade Writing assessment is 502 with a standard deviation of 57.2. The mean scale score for female students is 513 with a standard deviation of 57.3, and the mean scale score for male students is 492 with a standard deviation of 55.2.

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 503 and 507. The mean scale scores for the sub-content areas range from 503 to 531. The median scale scores vary between 503 and 504 for the content standards, and between 500 and 505 for the sub-content areas, and all are close to the median for the total test scale score, 504.

The mean percents of the maximum obtainable score for CS 2 (Write for a Variety of Purposes) and CS 3 (Write Using Conventions) are 64.3 and 74.8, respectively. The mean percent of the maximum obtainable score for the total test is 69.3. The mean percents of the maximum obtainable score for the sub-content areas range from 58.3 to 73.8.

Mathematics

The mean scale score for the total population of students taking the 2003 fifth-grade Mathematics assessment is 503 with a standard deviation of 72.2. The mean scale score for female students is 502 with a standard deviation of 69.9, and the mean scale score for male students is 503 with a standard deviation of 74.4.

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 506 to 517. The mean scale scores for the sub-content areas range from 511 to 539. The median scale scores vary between 505 and 506 for the content standards, and between 502 and 505 for the sub-content areas, and all are close to the median for the total test scale score, 505.

The mean percents of the maximum obtainable score for the content standards range from 62.4 on CS 4 (Geometric Concepts) to 72.6 on CS 2 (Algebra, Patterns and Functions). The mean percent of the maximum obtainable score for the total test is 67.9. The mean percents for the sub-content areas range from 65.8 to 75.6.

Sixth Grade

Reading

The mean scale score for the total population of students taking the 2003 sixth-grade Reading assessment is 620 with a standard deviation of 65.3. The mean scale score for female students is 627 with a standard deviation of 60.7, and the mean scale score for male students is 614 with a standard deviation of 68.9.

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 620 to 624. The mean scale scores for the sub-content areas range from 622 to 628. The median scale scores vary between 627 and 629 for the content standards, and between 627 and 629 for the sub-content areas, and all are close to the median for the total test scale score, 628.

The mean percents of the maximum obtainable score for content standards range from 60.0 on CS 6 (Literature) to 67.6 on CS 1 (Reading Comprehension). The mean percent of the maximum obtainable score for the total test is 63.3. The mean percents for the sub-content areas range from 62.7 to 64.2.

Writing

The mean scale score for the total population of students taking the 2003 sixth-grade Writing assessment is 520 with a standard deviation of 62.1. The mean scale score for female students is 533 with a standard deviation of 60.4, and the mean scale score for male students is 507 with a standard deviation of 61.1.

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 521 and 522. The mean scale scores for the sub-content areas range from 521 to 549. The median scale scores vary between 520 and 522 for the content standards, and between 521 and 524 for the sub-content areas, and all are close to the median for the total test scale score, 521.

The mean percents of the maximum obtainable score for CS 2 (Write for a Variety of Purposes) and CS 3 (Write Using Conventions) are 66.6 and 63.7, respectively. The mean percent of the maximum obtainable score for the total test is 65.3. The mean percents of the maximum obtainable score for the sub-content areas range from 59.6 to 72.1. Note that the mean percents are closer for the content standards than for the sub-content areas.

Mathematics

The mean scale score for the total population of students taking the 2003 sixth-grade Mathematics assessment is 520 with a standard deviation of 76.3. The mean scale score for female students is 520 with a standard deviation of 72.1, and the mean scale score for male students is 520 with a standard deviation of 80.0.

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 slightly negatively skewed.

The mean scale scores for the content standards range from 518 to 529. The mean scale scores for sub-content areas range from 520 to 530. The median scale scores vary between 521 and 528 for the content standards, and between 520 and 525 for the sub-content areas, and all are close to the median for the total test scale score, 522.

The mean percents of the maximum obtainable score for the content standards range from 51.8 on CS 1 (Number Sense) to 67.8 on CS 2 (Algebra, Patterns and Functions). The mean percent of the maximum obtainable score for the total test is 58.0. The mean percents of the maximum obtainable score for the sub-content areas range from 55.4 to 67.8. The observed ranges indicate that the average item difficulty varied somewhat across the content standards and across the sub-content areas.

Seventh Grade

Reading

The mean scale score for the total population of students taking the 2003 seventh-grade Reading assessment is 632 with a standard deviation of 66.6. The mean scale score for female students is 641 with a standard deviation of 61.6, and the mean scale score for male students is 624 with a standard deviation of 70.1.

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 632 to 639. The mean scale scores for the sub-content areas range from 632 to 658. The median scale scores vary between 639 and 640 for the content standards, and between 639 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 58.4 on CS 6 (Literature) to 70.1 on CS 1 (Reading Comprehension). The mean percent of the maximum obtainable score for the total test is 65.1. The mean percents of the maximum obtainable score for the sub-content areas range from 57.4 to 74.3.

Writing

The mean scale score for the total population of students taking the 2003 seventh-grade Writing assessment is 543 with a standard deviation of 65.9. The mean scale score for female students is 560 with a standard deviation of 63.2, and the mean scale score for male students is 527 with a standard deviation of 64.6.

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 544 and 547. The mean scale scores for the sub-content areas range from 544 to 580. The median scale scores vary between 545 and 546 for the content standards, and between 544 and 554 for the sub-content areas, and most are close to the median for the total test scale score, 545.

The mean percents of the maximum obtainable score for CS 2 (Write for a Variety of Purposes) and CS 3 (Write Using Conventions) are 62.3 and 65.3, respectively. The mean percent of the maximum obtainable score for the total test is 63.5. The mean percents of the maximum obtainable score for the sub-content areas range from 60.0 to 72.0.

Mathematics

The mean scale score for the total population of students taking the 2003 seventh-grade Mathematics assessment is 542 with a standard deviation of 68.2. The mean scale score for female students is 544 with a standard deviation of 65.2. The mean scale score for male students is 541 with a standard deviation of 70.9.

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 slightly negatively skewed for the total population and for each gender.

The mean scale scores for the content standards range from 535 to 545. The mean scale scores for the sub-content areas range from 534 to 541. The median scale scores vary between 546 and 548 for the content standards, and between 546 and 548 for the sub-content areas, and all are close to the median for the total test scale score, 546.

The mean percents of the maximum obtainable score for the content standards range from 44.5 on CS 3 (Data Analysis, Probability, and Statistics) to 55.0 on CS 2 (Algebra, Patterns and Functions). The mean percent of the maximum obtainable score for the total test is 47.8. The mean percents of the maximum obtainable score for the sub-content areas range from 41.2 to 45.8.

Eighth Grade

Reading

The mean scale score for the total population of students taking the 2003 eighth-grade Reading assessment is 653 with a standard deviation of 63.7. The mean scale score for

female students is 662 with a standard deviation of 59.2, and the mean scale score for male students is 644 with a standard deviation of 66.6.

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 649 to 662. The mean scale scores for the sub-content areas range from 649 to 666. The median scale scores vary between 658 and 661 for the content standards, and between 659 and 661 for the sub-content areas, and all are close to the median for the total test scale score, 660.

The mean percents of the maximum obtainable score for the content standards range from 52.9 on CS 6 (Literature) to 72.5 on CS 4 (Thinking Skills). The mean percent of the maximum obtainable score for the total test is 62.6. The mean percents of the maximum obtainable score for the sub-content areas range from 50.8 to 75.9. The observed ranges indicate that the average item difficulty varies somewhat across the content standards and across the sub-content areas.

Writing

The mean scale score for the total population of students taking the 2003 eighth-grade Writing assessment is 554 with a standard deviation of 74.0. The mean scale score for female students is 573 with a standard deviation of 71.9, and the mean scale score for male students is 536 with a standard deviation of 71.3.

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 556 and 558. The mean scale scores for the sub-content areas range from 555 to 590. The median scale scores vary between 555 and 557 for the content standards, and between 554 and 592 for the sub-content areas, and most are close to the median for the total test scale score, 556. The median scale score for SA 6 (Extended Writing), 592, was a bit higher 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 66.4 and 66.9, respectively. The mean percent of the maximum obtainable score for the total test is 66.6. The mean percents of the maximum obtainable score for the sub-content areas range from 60.5 to 77.2.

Mathematics

The mean scale score for the total population of students taking the 2003 eighth-grade Mathematics assessment is 550 with a standard deviation of 73.8. The mean scale score for female students is 553 with a standard deviation of 68.3. The mean scale score for male students is 548 with a standard deviation of 78.8.

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 540 to 548. The mean scale scores for sub-content areas range from 539 to 545. The median scale scores vary between 557 and 560 for the content standards, and between 557 and 558 for the sub-content areas, and all are close to the median for the total test scale score, 558.

The mean percents of the maximum obtainable score for the content standards range from 35.9 on CS 1 (Number Sense) to 52.3 on CS 3 (Data Analysis, Probability, and Statistics). The mean percent of the maximum obtainable score for the total test is 43.7. The mean percents of the maximum obtainable score for the sub-content areas range from 32.1 to 42.0.

Science

The mean scale score for the total population of students taking the 2003 eighth-grade Science assessment is 500 with a standard deviation of 63.1. The mean scale score for female students is 498 with a standard deviation of 58.9, and the mean scale score for male students is 502 with a standard deviation of 66.9.

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 496 to 504. The mean scale scores for the sub-content areas range from 489 to 523. The median scale scores vary between 504 and 508 for the content standards, and between 503 and 508 for the sub-content areas, and all are close to the median for the total test scale score, 506.

The mean percents of the maximum obtainable score for the content standards range from 45.9 on CS 4 (Earth and Space Science) to 63.6 on CS 1 (Scientific Investigation). The mean percent of the obtainable score for the total test is 52.9. The mean percents of the maximum obtainable score for the sub-content areas range from 41.3 to 71.5. The observed ranges

indicate that the average item difficulty varies somewhat across the content standards and across the sub-content areas.

Ninth Grade

Reading

The mean scale score for the total population of students taking the 2003 ninth-grade Reading assessment is 658 with a standard deviation of 62.3. The mean scale score for female students is 670 with a standard deviation of 54.1, and the mean scale score for male students is 648 with a standard deviation of 67.7.

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 655 to 662. The mean scale scores for the sub-content areas range from 650 to 664. The median scale scores vary between 667 and 668 for the content standards, and between 667 and 668 for the sub-content areas, and all are close to the median for the total test scale score, 667.

The mean percents of the maximum obtainable score for the content standards range from 59.9 on CS 1 (Reading Comprehension) to 66.2 on CS 4 (Thinking Skills). The mean percent of the maximum obtainable score for the total test is 62.2. The mean percents of the maximum obtainable score for the sub-content areas range from 52.4 to 69.5.

Writing

The mean scale score for the total population of students taking the 2003 ninth-grade Writing assessment is 566 with a standard deviation of 78.2. The mean scale score for female students is 585 with a standard deviation of 74.7, and the mean scale score for male students is 549 with a standard deviation of 77.6.

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 567 and 569. The mean scale scores for sub-content areas range from 565 to 611. The median scale scores vary between 567 and 569 for the content standards, and between 566 and 593 for the sub-content areas, and most are close to the median for the total test scale score, 568. The median scale score for SA 6 (Extended Writing).

The mean percents of the maximum obtainable score for CS 2 (Write for a Variety of Purposes) and CS 3 (Write Using Conventions) are 69.0 and 63.1, respectively. The mean percent of the maximum obtainable score for the total test is 66.6. The mean percents of the maximum obtainable score for the sub-content areas range from 58.4 to 77.7.

Mathematics

The mean scale score for the total population of students taking the 2003 ninth-grade Mathematics assessment is 564 with a standard deviation of 73.2. The mean scale score for female students is 567 with a standard deviation of 67.2, and the mean scale score for male students is 561 with a standard deviation of 78.4.

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 553 to 564. The mean scale scores for the sub-content areas are 549 and 561. The median scale scores vary between 572 and 574 for the content standards, and between 575 and 576 for 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 31.8 on CS 1 (Number Sense) to 39.7 on CS 3 (Data Analysis, Probability and Statistics). The mean percent of the maximum obtainable score for the total test is 36.5. The mean percents of the maximum obtainable score for the sub-content areas range from 28.3 to 39.4.

Tenth Grade
Reading

The mean scale score for the total population of students taking the 2003 tenth-grade Reading assessment is 681 with a standard deviation of 62.0. The mean scale score for female students is 692 with a standard deviation of 53.4, and the mean scale score for male students is 671 with a standard deviation of 67.8.

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 679 to 690. The mean scale scores for the sub-content areas range from 678 to 700. The median scale scores vary between

689 and 692 for the content standards, and between 690 and 693 for the sub-content areas, and all are close to the median for the total test scale score, 690.

The mean percents of the maximum obtainable score for the content standards range from 56.9 on CS 5 (Use of Literary Information) to 71.5 on CS 4 (Thinking Skills). The mean percent of the maximum obtainable score for the total test is 63.9. The mean percents of the maximum obtainable score for the sub-content areas range from 51.0 to 70.4.

Writing

The mean scale score for the total population of students taking the 2003 tenth-grade Writing assessment is 584 with a standard deviation of 86.0. The mean scale score for female students is 603 with a standard deviation of 83.8, and the mean scale score for male students is 565 with a standard deviation of 83.7.

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 586 and 591. The mean scale scores for the sub-content areas range from 591 to 641. The median scale scores vary between 585 and 586 for the content standards, and between 583 and 590 for the sub-content areas, and all are close to the median for the total test scale score, 586.

The mean percents of the maximum obtainable score for CS 2 (Write for a Variety of Purposes) and CS 3 (Write Using Conventions) are 71.6 and 69.9, respectively. The mean percent of the maximum obtainable score for the total test is 71.0. The mean percents of the maximum obtainable score for the sub-content areas vary from 63.0 to 78.4.

Mathematics

The mean scale score for all students taking the 2003 tenth-grade Mathematics assessment is 582 with a standard deviation of 72.2. The mean scale score for female students is 582 with a standard deviation of 67.6, and the mean scale score for male students is 583 with a standard deviation of 76.5.

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 571 to 582. The mean scale scores for the sub-content areas are 578 and 586. The median scale scores vary between 588

and 591 for the content standards, and between 592 and 598 for the sub-content areas, and all are fairly close to the median for the total test scale score, 589.

The mean percents of the maximum obtainable score for the content standards range from 30.0 on CS 4/5 (Geometry and Measurement) to 43.0 on CS 3 (Data Analysis, Probability and Statistics). The mean percent of the maximum obtainable score for the total test is 38.7. The mean percents of the maximum obtainable score for SA 1 and SA 2 are 53.8 and 35.3, respectively.

The Correlations Between the Content Standards and Between the 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 .70 (grade 8) and .84 (grade 9) for the relationship between the various content standards, and between .58 (grade 10) and .82 (grade 7) for the relationship between the various sub-content areas, respectively.

For the Spanish Reading assessments, the correlations among the various content standards vary between .72 and .79; they vary between .65 and .79 among the various sub-content areas.

For the Writing assessments, the coefficients for the correlation between content standards 2 and 3 vary between .70 (grade 3) and .83 (grade 8). The correlations among the various sub-content areas vary between .44 (grade 4) and .76 (grades 5 and 8).

For the Spanish Writing assessment, the correlation between content standards 2 and 3 varies between .65 and .70; the correlations between the various sub-content areas vary between .41 and .63.

For the Mathematics assessments, the correlations vary between .71 (grade 6) and .82 (grades 7 and 8) for the relationship among the content standards, and between .59 (grade 10) and .77 (grade 5) for the relationship among the sub-content areas.

Finally, for the Science assessment, the correlation coefficients vary between .61 and .71 for the relationship among the content standards, and between .48 and .67 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 for each grade and content area. The alphas for Reading are .90 and .94 for grades 3 and 9 respectively, and .93 for grades 4 through 8 and 10. The alphas for grades 3 and 4 Spanish Reading are .87 and .90, respectively. The alphas for grades 3 through 10 Writing are .90, .91, .91, .92, .91, .93, .92, and .92, respectively. The alphas for grade 3 and 4 Spanish Writing are .90 and .87, respectively. The alphas for Mathematics are .94 for grades 5 through 8 and .93 for grades 9 and 10. The alpha for grade 8 Science is .92. These are all high reliabilities and indicate that the Colorado 2003 assessments had strong internal consistency and that the tests produced relatively stable scores.

Tables 16 and 17 show the estimated reliability index (Cronbach's alpha) for each of the content standards and sub-content areas, respectively, for each grade and content area for the 2003 assessments.

For Reading grades 4 through 10, the alphas for the various content standards vary between .68 (for CS 5 in grade 8) and .86 (for CS 1 in grade 4), with a median of .765. For Reading grades 3 through 10, the alphas for the various sub-content areas vary between .56 (for SA 4 in grade 10) and .88 (for SA 1 in grade 4), with a median of .75.

For the Spanish version of the Reading assessment for grade 4, the alphas for the various content standards vary between .56 and .81 (for CS 6 and 1, respectively), with a median of .67. For the Spanish versions of the Reading assessment for grades 3 and 4, the alphas for the various sub-content areas vary between .43 (for SA 3 in grade 4) and .83 (for SA 1 in grade 4), with a median of .675.

For Writing grades 3 through 10, the alphas for the various content standards vary between .73 (for CS 2 in grade 3) and .88 (for CS 3 in grade 3 and CS 2 in grades 9 and 10) with a median of .855. The alphas for the various sub-content areas vary between .65 and .84 (for SA 8, in grades 10 and SA 7in grade 8, respectively), with a median of .74.

For the Spanish version of the Writing assessment for grade 3 and 4, the alphas for the various content standards vary between .70 (for CS 2 in grade 4) and .85 (for CS 3 in grade 3) with a median of .82. The alphas for the various sub-content areas vary between .55 (for SA 6 in grade 4) and .84 (for SA 5 in grade 3), with a median of .72.

For Mathematics grades 5 through 10, the alphas for the various content standards vary between .70 and .83 (for CS 2 in grade 6 and CS 1 in grade 7, respectively) with a median of .755. The alphas for the various sub-content areas vary between .51 and .81 (for SA 2 in grade 10 and SA 1 in grade 6, respectively), with a median of .69.

For grade 8 Science, the alphas for the various content standards vary between .58 and .81 (for CS 5 and 1, respectively) with a median of .71. The alphas for the various sub-content areas vary between .49 and .70 (for SA 4 and 5, respectively), with a median of .65.

The alpha internal consistency coefficients generally are lower for the various sub-test scores than they are for the total test scores. This is not that surprising, given that the scores for the content standards and the sub-content areas are based on fewer items than those for the total test scores.

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 2003 third-grade Reading assessment. The point-biserials for all multiple-choice items range from .20 to .60 with a mean of .44. The p-values for these items range from .29 to .97 with a mean of .80.

Table 19 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .24 to .57 with a mean of .46. Their p-values range from .34 to .88 with a mean of .67. 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.

For two of the seven constructed-response items, over 70% of the students obtained the highest possible score points. The scores of the remaining students were fairly well distributed across the remaining score points.

The omit rate for the third-grade Reading assessment was small, ranging from .0% to 1.6% for multiple-choice items (Table 18) and .4% to .9% 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 2003 third-grade Reading assessment. The point-biserials for all multiple-choice items range from .18 to .48 with a mean of .35. The p-values for these items range from .26 to .93 with a mean of .70.

Table 21 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .31 to .51 with a mean of .42. Their p-values range from .29 to .87, with a mean of .59. 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 3.4% for multiple-choice items (Table 20) and .3% to 2.9% for constructed-response items (Table 21).

Writing

Table 22 lists the results of the multiple-choice item analyses for the 2003 third-grade Writing assessment. The point-biserials for all multiple-choice items range from .24 to .51 with a mean of .39. The p-values for these items range from .59 to .98 with a mean of .86.

Table 23 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .23 to .51 with a mean of .39. Their p-values range from .54 to .99, with a mean of .83. 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 2.2% for multiple-choice items (Table 22) and from .1% to .3% 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 2003 third-grade Writing assessment. The point-biserials for all multiple-choice items range from .10 to .46 with a mean of .34. The p-values for these items range from .25 to .95 with a mean of .73.

Table 25 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .20 to .54 with a mean of .43. Their p-values range from .33 to .95, with a mean of .67. 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 1.6% for multiple-choice items (Table 24) and .0% to .2% for constructed-response items (Table 25), with only one item having a 5% omit rate.

Fourth Grade

Reading

Table 26 lists the results of the multiple-choice item analyses for the 2003 fourth-grade Reading assessment. The point-biserials for the multiple-choice items range from .14 to .56 with a mean of .43. The p-values for the multiple-choice items range from .43 to .94 with a mean of .71.

Table 27 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .32 to .60 with a mean of .44. Their p-values range from .20 to .86 with a mean of .55. 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 reasonable, ranging from .0% to 3.2% for multiple-choice items (Table 26). The range was .3% to 3.6% 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 2003 fourth-grade Reading assessment. When item 107 is not taken into account, the point-biserials for all multiple-choice items range from .09 to .51 with a mean of .33. The p-values for these items range from .14 to .91 with a mean of .57. MC item 107 did not discriminate between low or 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 .16 to .62 with a mean of .44. Their p-values range from .15 to .79, with a mean of .45. 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 .0% to 2.4% for multiple-choice items (Table 28) and .1% to 3.1% for constructed-response items (Table 29).

Writing

Table 30 lists the results of the multiple-choice item analyses for the 2003 fourth-grade Writing assessment. The point-biserials for all multiple-choice items range from .22 to .48 with a mean of .37. The p-values for these items range from .44 to .99 with a mean of .76.

Table 31 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .05 to .59 with a mean of .42. Their p-values range from .29 to 1.00, with a mean of .65.

The omit rate for the fourth-grade Writing assessment was small, ranging from .1% to 1.8% 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 2003 fourth-grade Writing assessment. When item 54 is not taken into account, the point-biserials for all multiple-choice items range from .13 to .47 with a mean of .31. The p-values for these items range from .26 to .97 with a mean of .59. MC item 54 did not discriminate between low or 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 33 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed response items range from .12 to .59 with a mean of .37. Their p-values range from .03 to .99 with a mean of .49.

The omit rate for the fourth-grade Writing assessment was small, ranging from .0% to 1.1% for multiple-choice items (Table 32) and .0% to 3.4% for constructed-response items (Table 33).

Fifth Grade

Reading

Table 34 lists the results of the multiple-choice item analyses for the 2003 fifth-grade Reading assessment. The point-biserials for the multiple-choice items range from .19 to .52, with a mean of .40. The p-values for the multiple-choice items range from .36 to .95 with a mean of .72.

Table 35 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .33 to .59 with a mean of .48. Their p-values range from .33 to .85 with a mean of .54. 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 2 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 .0% to 2.5% for multiple-choice items (Table 34) and .5% to 5.9% for constructed-response items (Table 35), with only one constructed-response item having an omit rate greater than 5%.

Writing

Table 36 lists the results of the multiple-choice item analyses for the 2003 fifth-grade Writing assessment. The point-biserials for all multiple-choice items range from .13 to .55 with a mean of .38. The p-values for these items range from .26 to .95 with a mean of .70.

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 .46. Their p-values range from .51 to 1.00, with a mean of .73.

The omit rate for the fifth-grade Writing assessment was small, ranging from .1% to 3.8% for multiple-choice items (Table 36) and .0% to 3.8% for constructed-response items (Table 37).

Mathematics

Table 38 lists the results of the multiple-choice item analyses for the 2003 fifth-grade Mathematics assessment. The point-biserials for the multiple-choice items range from .20 to .60, with a mean of .40. The p-values for the multiple-choice items range from .26 to .97 with a mean of .71.

Table 39 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .44 to .68 with a mean of .58. Their p-values range from .41 to .81 with a mean of .64. The distribution of the percent of students obtaining each score point for the constructed-response items shows that there is a good amount of variability in students' responses to most items, indicating that these items work well over the range of student ability.

The omit rate for the fifth-grade Mathematics assessment was reasonable, ranging from .1% to 3.6% for multiple-choice items (Table 38) and .2% to 5.1% for constructed-response items (Table 39), with only one item having an omit rate greater than 5%.

Sixth Grade

Reading

Table 40 lists the results of the multiple-choice item analyses for the 2003 sixth-grade Reading assessment. The point-biserials for the multiple-choice items range from .18 to .59 with a mean of .39. The p-values for the multiple-choice items range from .30 to .89 with a mean of .64.

Table 41 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .32 to .61 with a mean of .51. Their p-values range from .37 to .90 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 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 Reading assessment was small, ranging from .1% to 3.7% for multiple-choice items (Table 40) and .5% to 8.1% for constructed-response items, with only three items having an omit rate greater than 5% (Table 41).

Writing

Table 42 lists the results of the multiple-choice item analyses for the 2003 sixth-grade Writing assessment. The point-biserials for all multiple-choice items range from .16 to .57 with a mean of .40. The p-values for these items range from .27 to .90 with a mean of .68.

Table 43 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .07 to .60 with a mean of .41. Their p-values range from .11 to .99 with a mean of .62.

The omit rate for the sixth-grade Writing assessment was small, ranging from .1% to 5.0% for multiple-choice items (Table 42) and .0% to 7.2% for constructed-response items (Table 43) , with only two items having an omit rate greater than 5% (Tables 42 and 43).

Mathematics

Table 44 lists the results of the multiple-choice item analyses for the 2003 sixth-grade Mathematics assessment. The point-biserials for the multiple-choice items range from .20 to .54, with a mean of .38. The p-values for the multiple-choice items range from .25 to .91 with a mean of .62.

Table 45 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .40 to .71 with a mean of .58. Their p-values range from .11 to .79 with a mean of .52. 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 sixth-grade Mathematics assessment was reasonable, ranging from .1% to 5.8% for multiple-choice items (Table 44) and .2% to 1.6% for constructed-response items (Table 45), with only one item having an omit rate greater than 5%.

Seventh Grade

Reading

Table 46 lists the results of the multiple-choice item analyses for the 2003 seventh-grade Reading assessment. The point-biserials for the multiple-choice items are positive, ranging from .10 to .58 with a mean of .39. The p-values for the multiple-choice items range from .39 to .92 with a mean of .67.

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 .32 to .58 with a mean of .48. The p-values for the constructed-response items range from .27 to .83 with a mean of .61. 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 7 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 2003 grade 7 Reading assessment ranged from .1% to 2.4% (Table 46).   The percent of students who omitted constructed-response items ranged from .8% to 6.2% (Table 47), with only one item having an omit rate equal or greater than 5%.

Writing

Table 48 lists the results of the multiple-choice item analyses for the 2003 seventh-grade Writing assessment. The point-biserials for all multiple-choice items range from .01 to .50 with a mean of .36. The p-values for these items range from .07 to .94 with a mean of .66.

Table 49 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .10 to .62 with a mean of .47. Their p-values range from .28 to .99, with a mean of .57.

The omit rate for the seventh-grade Writing assessment was small, ranging from .1% to 2.0% for multiple-choice items (Table 48) and .0% to 3.9% for constructed-response items (Table 49).

Mathematics

Table 50 lists the results of the multiple-choice item analyses for the 2003 seventh-grade Mathematics assessment. The point-biserials for the multiple-choice items range from .17 to .60, with a mean of .39. The p-values for the multiple-choice items range from .23 to .96 with a mean of .57.

Table 51 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 .58. Their p-values range from .15 to .66 with a mean of .36. 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 reasonable, ranging from .1% to 2.3% for multiple-choice items (Table 50) and .5% to 3.4% for constructed-response items (Table 51).

Eighth Grade

Reading

Table 52 lists the results of the multiple-choice item analyses for the 2003 eighth-grade Reading assessment. When item 117 is not taken into account, the point-biserials for the multiple-choice items range from .11 to .54 with a mean of .37. The p-values for the multiple-choice items range from .15 to .93 with a mean of .66. MC item 117 did not discriminate between low or 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 53 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .34 to .63 with a mean of .51. Their p-values range from .21 to .84 with a mean of .57. 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 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 percent of students who omitted the multiple-choice items in the eighth-grade Reading assessment ranged from .1% to 1.1% (Table 52). The percent of students who omitted the constructed-response items ranged from .9% to 7.7% (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 2003 eighth-grade Writing assessment. The point-biserials for all multiple-choice items range from .03 to .58 with a mean of .42. The p-values for these items range from .08 to .90 with a mean of .68.

Table 55 lists the results of the constructed-response item analyses. When item 2 A is not taken into account, the item-to-total-score correlations for the constructed-response items range from .15 to .65 with a mean of .47. Their p-values range from .14 to .99, with a mean of .65. CR item 2 A did not discriminate between low or 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.

The omit rate for the eighth-grade Writing assessment was small, ranging from .1% to 3.6% for multiple-choice items (Table 54) and .0% to 2.0% for constructed-response items (Table 54).

Mathematics

Table 56 lists the results of the multiple-choice item analyses for the 2003 eighth-grade Mathematics assessment. When item 15 is not taken into account, the point-biserials for the multiple-choice items range from .13 to .56 with a mean of .37. The p-values for these multiple-choice items range from .10 to .88 with a mean of .50. Multiple-choice item 15 did not discriminate between low or 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 57 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .44 to .71 with a mean of .60. Their p-values range from .07 to .59 with a mean of .36. 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 2.2% (Table 56). The percent of students who omitted constructed-response items ranged from .8% to 8.2%, with only one constructed-response item having an omit rate greater than 5% (Table 57).

Science

Table 58 lists the results of the multiple-choice item analyses for the 2003 eighth-grade Science assessment. The point-biserials for the multiple-choice items range from .00 to .49 with a mean of .31. The p-values for the multiple-choice items range from .14 to .94 with a mean of .58.

Table 59 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .30 to .62 with a mean of .46. Their p-values range from .08 to .85 with a mean of .44. The percent of students obtaining each score point for the constructed-response items shows a good amount of variability in students' responses to most items, indicating that these items work well over the range of student ability.

The omit rate for the multiple-choice items for the eighth-grade Science assessment ranged from .1% to 3.7% (Table 58). The omit rate for the constructed-response items ranged from .9% to 9.6% (Table 59), with six 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 2003 ninth-grade Reading assessment. The point-biserials for the multiple-choice items range from .14 to .54 with a mean of .38. The p-values for the multiple-choice items range from .31 to .90 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 .38 to .62 with a mean of .55. Their p-values range from .29 to .76 with a mean of .56. 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 4 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 multiple-choice items for the ninth-grade Reading assessment ranged from .1% to 2.0% (Table 60).   The omit rate for the constructed-response items ranged from 1.7% to 10.1%, with 13 out of the 15 items having an omit rate greater than 5% (Table 61).

Writing

Table 62 lists the results of the multiple-choice item analyses for the 2003 ninth-grade Writing assessment. The point-biserials for all multiple-choice items range from .22 to .55 with a mean of .41. The p-values for these items range from .28 to .91 with a mean of .65.

Table 63 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .16 to .65 with a mean of .44. Their p-values range from .25 to .99, with a mean of .68.

The omit rate for the ninth-grade Writing assessment was small, ranging from .1% to 4.4% for multiple-choice items (Table 62) and .0% to 3.6% for constructed-response items (Table 63).

Mathematics

Table 64 lists the results of the multiple-choice item analyses for the 2003 ninth-grade Mathematics assessment. When item 51 is not taken into account, the point-biserials for the multiple-choice items range from .02 to .52 with a mean of .34. The p-values for these multiple-choice items range from .10 to .87 with a mean of .48. MC item 51 did not discriminate between low or 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 65 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .42 to .71 with a mean of .58. Their p-values range from .04 to .48 with a mean of .24. 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 4.6% (Table 64). The percent of students who omitted constructed-response items ranged from .8% to 8.0% with 3 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 2003 tenth-grade Reading assessment. The point-biserials for the multiple-choice items range from .11 to .58 with a mean of .37. The p-values for the multiple-choice items range from .35 to .90 with a mean of .70.

Table 67 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .43 to .63 with a mean of .54. Their p-values range from .31 to .76 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 5 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 2003 tenth-grade Reading assessment ranged from .1% to 2.5% (Table 66). Omit rates for the constructed-response items ranged from 2.3% to 14.8% (Table 67), with 9 out of the 13 items having an omit rate greater than 5%.

Writing

Table 68 lists the results of the multiple-choice item analyses for the 2003 tenth-grade Writing assessment. The point-biserials for all multiple-choice items range from .20 to .56 with a mean of .41. The p-values for these items range from .46 to .90 with a mean of .73.

Table 69 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .09 to .66 with a mean of .46. Their p-values range from .32 to .98, with a mean of .68.

The omit rate for the tenth-grade Writing assessment was small, ranging from .1% to .9% for multiple-choice items (Table 68) and .0% to 3.7% for constructed-response items (Table 69).

Mathematics

Table 70 lists the results of the multiple-choice item analyses for the 2003 tenth-grade Mathematics assessment. When items 53 and 57 are not taken into account, the point-biserials for the multiple-choice items range from .05 to .51 with a mean of .34. The p-values for the multiple-choice items range from .19 to .83 with a mean of .49. MC items 53 and 57 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.

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 .75 with a mean of .61. The p-values for the constructed-response items range from .07 to .68 with a mean of .28. The percent of students obtaining each score point for the constructed-response items shows a good amount of variability in students' responses to most items.

The omit rate for the multiple-choice items for the tenth-grade Mathematics assessment ranged from .1% to 1.3% (Table 70). The omit rate for the constructed-response items ranged from .4% to 5.9% (Table 71), with 3 out of the 15 items having an omit rate greater than 5%.

Part 5: Scaling and Calibration

Overview of the IRT Models

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

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

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

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

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

where m_j is the number of score levels and

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

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

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

and

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

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

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

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

Reading

The critical Z-value for the third-grade Reading assessment, based on the calibration sample, is 131.5 The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 72 and 73, respectively.

The Z-statistics for all MC and most CR items are small compared to the critical Z-value, indicating that the proportions of students obtaining each item score are close to the proportions that would be predicted by the estimated thetas and item parameters.

CR item 16 in Session 1 exhibited less than optimal fit with a Z-statistic of 213.7, exceeding the critical Z-value of 131.5. A close examination of the ICCs for the item indicated that the

proportion of lower ability students obtaining a "0" was lower than expected and obtaining a "1" was higher than expected. The proportion of close to average ability students obtaining a "0" was slightly higher than expected.

Reading – Spanish Version

The critical Z-value for the third-grade Spanish Reading assessment, based on the calibration sample, is 4.4 The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 74 and 75, respectively.

The Z-statistics for most MC and most CR items are small compared to the critical Z-value, indicating that the proportions of students obtaining each item score are close to the proportions that would be predicted by the estimated thetas and item parameters.

MC item 21 in Session 2 exhibited less than optimal fit with a Z-statistic of 4.6, marginally exceeding the critical Z-value of 4.4. A close examination of the ICC for the item indicated that the proportion of lower ability students obtaining a "1" was higher than expected.

CR item 27 in Session 2 exhibited less than optimal fit with a Z-statistic of 5.8, exceeding the critical Z-value of 4.4. A close examination of the ICCs for the item indicated that the proportion of higher ability students obtaining a "1" was higher than expected.

Writing

The critical Z-value for the third-grade Writing assessment, based on the calibration sample, is 24.5. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 76 and 77, respectively.

The Z-statistics for all MC and CR items are small compared to the critical Z-value, indicating that the proportions of students obtaining each item score are close to the proportions that would be predicted by the estimated thetas and item parameters.

Writing – Spanish Version

The critical Z-value for the third-grade Spanish Writing assessment, based on the calibration sample, is 3.9 The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 78 and 79, respectively.

The Z-statistics for most MC and most CR items are small compared to the critical Z-value, indicating that the proportions of students obtaining each item score are close to the proportions that would be predicted by the estimated thetas and item parameters.

CR item 7E in Session 1 exhibited less than optimal fit with a Z-statistic of 5.5, exceeding the critical Z-value of 3.9. A close examination of the ICCs for the item indicated that the

observed and expected proportion of students obtaining a "1" over the ability continuum were sometimes slightly off; however, there was no clear pattern in these differences.

MC item 8 in Session 12 exhibited less than optimal fit with a Z-statistic of 6.5, exceeding the critical Z-value of 3.9. A close examination of the ICC for the item indicated that the observed and expected proportion of students obtaining a "1" over the ability continuum were sometimes off; however, there was no clear pattern in these differences.

CR item 39A in Session 2 exhibited less than optimal fit with a Z-statistic of 4.2, exceeding the critical Z-value of 3.9. A close examination of the ICCs for the item indicated that the observed and expected proportion of students obtaining the various possible scores over the ability continuum were sometimes off; however, there was no clear pattern in these differences.

Fourth Grade

Reading

The critical Z-value for the fourth-grade Reading assessment, based on the calibration sample, is 27.5. Fit values above this critical value may indicate poor model fit. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 80 and 81, respectively.

The Z-statistics for most CR and MC items are smaller than the critical Z-value, indicating that the proportions of students with each item score are close to the proportions that would be predicted by the estimated thetas and item parameters.

MC item 102 in Session 6 exhibited less than optimal fit with a Z-statistic of 29.8, exceeding the critical Z-value of 27.5. A close examination of the ICC for the item indicated that the proportion of lower ability students obtaining a "1" was lower than expected and the proportion of slightly-below-average-ability students obtaining a "1" was slightly higher than expected.

CR item 38 in Session 4 exhibited less than optimal fit with a Z-statistic of 39.2, exceeding the critical Z-value of 27.5. A close examination of the ICC for the item indicated that the proportion of lower ability students obtaining a "1" was higher than expected and obtaining a "0" was lower than expected.

Reading – Spanish Version

The critical Z-value for the fourth-grade Spanish Reading assessment, based on the calibration sample, is 1.9. Fit values above this critical value may indicate poor model fit. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 82 and 83, respectively.

The Z-statistics for most CR and MC items are smaller than the critical Z-value, indicating that the proportions of students with each item score are close to the proportions that would be predicted by the estimated thetas and item parameters.

CR item 9 in Session 3 exhibited less than optimal fit with a Z-statistic of 2.5, exceeding the critical Z-value of 1.9. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was slightly lower than expected and obtaining a "1" was higher than expected.

MC item 26 in Session 4 exhibited less than optimal fit with a Z-statistic of 2.7, exceeding the critical Z-value of 1.9. A close examination of the ICC for the item indicated that the proportion of average ability students obtaining a "1" was slightly higher than expected.

MC item 31 in Session 4, exhibited less than optimal fit with a Z-statistic of 2.3, exceeding the critical Z-value of 1.9. A close examination of the ICC for the item indicated that the observed and expected proportion of students obtaining a "1" over the ability continuum were sometimes slightly off; however, there was no clear pattern in these differences.

MC item 93 in Session 6 exhibited less than optimal fit with a Z-statistic of 3.3, exceeding the critical Z-value of 1.9. A close examination of the ICC for the item indicated that the observed and expected proportion of students obtaining a "1" over the ability continuum were sometimes slightly off; however, there was no clear pattern in these differences.

MC item 95 in Session 6 exhibited less than optimal fit with a Z-statistic of 2.0, marginally exceeding the critical Z-value of 1.9. A close examination of the ICC for the item indicated that the observed and expected proportion of students obtaining a "1" over the ability continuum were sometimes slightly off; however, there was no clear pattern in these differences.

The item parameter estimation process did not converge for MC item 107 in Session 6. A close examination of the ICC for the item indicated that the student responses do not fit the model well. The item did not discriminate between low or 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.

Writing

The critical Z-value for the <u>fourth-grade Writing</u> assessment, based on the calibration sample, is 26.6. Fit values above this critical value may indicate poor model fit. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 84 and 85, respectively.

The Z-statistics for most MC items and most CR items are small compared to the critical Z-value indicating that the proportions of students with each item score are close to the proportions that would be predicted by the estimated thetas and item parameters.

CR item 3A in Session 2 exhibited less than optimal fit with a Z-statistic of 190.3, exceeding the critical Z-value of 26.6. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was less than expected and obtaining a "1" was higher than expected. The proportion of higher ability students obtaining a "3" was also less than expected and obtaining a "4" was higher than expected.

MC item 79 in Session 5 exhibited less than optimal fit with a Z-statistic of 31.2, exceeding the critical Z-value of 26.6. A close examination of the ICC for the item indicated that the proportion of low ability students obtaining a "1" was lower than expected.

Writing – Spanish Version

The critical Z-value for the fourth-grade Spanish Writing assessment, based on the calibration sample, is 1.9. Fit values above this critical value may indicate poor model fit. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 86 and 87, respectively.

The Z-statistics for most MC and CR items are small compared to the critical Z-value, indicating that the proportions of students with each item score are close to the proportions that would be predicted by the estimated thetas and item parameters.

CR item 2C in Session 2 exhibited less than optimal fit with a Z-statistic of 2.4, exceeding the critical Z-value of 1.9. A close examination of the ICCs for the item indicated that the observed and expected proportion of students obtaining a "1" over the ability continuum were sometimes slightly off; however, there was no clear pattern in these differences.

CR item 3A in Session 2 exhibited less than optimal fit with a Z-statistic of 2.2, exceeding the critical Z-value of 1.9. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "2" was slightly higher than expected.

MC items 62, 69, 76, 78, 79, and 86 in Session 5 exhibited less than optimal fit with Z-statistics of 2.1, 1.9, 3.7, 2.7, 2.0, and 2.4, respectively, all exceeding the critical Z-value of 1.9. A close examination of the ICCs for the items indicated that the observed and expected proportion of students obtaining a "1" over the ability continuum were sometimes slightly off; however, there was no clear pattern in these differences.

The item parameter estimation process did not converge for MC item 54 in Session 5. A close examination of the ICC for the item indicated that the student responses do not fit the model well. The item did not discriminate between low or 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.

Fifth Grade

Reading

The critical Z-value for the <u>fifth-grade Reading</u> assessment, based on the calibration sample, is 27.8. Fit values above this critical value may indicate poor model fit. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 88 and 89, respectively.

The Z-statistics for all MC and CR items are smaller than the critical Z-value, indicating that the proportions of students with each item score are close to the proportions that would be predicted by the estimated thetas and item parameters.

Writing

The critical Z-value for the <u>fifth-grade Writing</u> assessment, based on the calibration sample, is 26.7. Fit values above this critical value may indicate poor model fit. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 90 and 91, respectively.

The Z-statistics for all MC items and most CR items are smaller than the critical Z-value, indicating that the proportions of students with each item score are close to the proportions that would be predicted by the estimated thetas and item parameters.

CR item 3A in Session 2 exhibited less than optimal fit with a Z-statistic of 170.0, exceeding the critical Z-value of 26.7. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was less than expected and obtaining a "1" was higher than expected. The proportion of higher ability students obtaining a "3" was also less than expected and obtaining a "4" was higher than expected.

CR item 3B in Session 2 exhibited less than optimal fit with a Z-statistic of 27.1, marginally exceeding the critical Z-value of 26.7. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was slightly less than expected and obtaining a "1" was higher than expected.

CR item 3C in Session 2 exhibited less than optimal fit with a Z-statistic of 68.3, exceeding the critical Z-value of 26.7. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was less than expected and obtaining a "1" was higher than expected.

CR item 27 in Session 3 exhibited less than optimal fit with a Z-statistic of 54.7, exceeding the critical Z-value of 26.7. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was less than expected and obtaining a "1" was higher than expected.

Mathematics

The critical Z-value for the <u>fifth-grade Mathematics</u> assessment, based on the calibration sample, is 26.8. Fit values above this critical value may indicate poor model fit. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 92 and 93, respectively.

The Z-statistics for all MC and CR items are smaller than the critical Z-value, indicating that the proportions of students with each item score are close to the proportions that would be predicted by the estimated thetas and item parameters.

Sixth Grade

Reading

The critical Z-value for the <u>sixth-grade Reading</u> assessment, based on the calibration sample, is 25.3. Fit values above this critical value may indicate poor model fit. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 94 and 95, respectively.

The Z-statistics for most MC items and most CR items are smaller than the critical Z-value, indicating that the proportions of students with each item score are close to the proportions that would be predicted by the estimated thetas and item parameters.

CR item 45 in Session 4 exhibited less than optimal fit with a Z-value of 42.6, exceeding the critical Z-value of 25.3. A detailed examination of the ICC for item 45 indicated that the proportion of lower ability students obtaining a "0" was higher than expected and obtaining a "1" was lower than expected.

MC item 105 in Session 6 exhibited less than optimal fit with a Z-statistic of 40.0, exceeding the critical Z-value of 25.3. A close examination of the ICC for the item indicated that the proportion of slightly below average ability students obtaining a "1" was higher than expected.

MC item 111 in Session 6 exhibited less than optimal fit with a Z-statistic of 30.4, exceeding the critical Z-value of 25.3. A close examination of the ICC for the item indicated that the proportion of lower ability students obtaining a "1" was higher than expected.

Writing

The critical Z-value for the <u>sixth-grade Writing</u> assessment, based on the calibration sample, is 25.1. Fit values above this critical value may indicate poor model fit. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 96 and 97, respectively.

The Z-statistics for all MC items and most CR items are smaller than the critical Z-value, indicating that the proportions of students with each item score are close to the proportions that would be predicted by the estimated thetas and item parameters.

CR item 3A in Session 2 exhibited less than optimal fit with a Z-statistic of 104.8, exceeding the critical Z-value of 25.1. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was less than expected and obtaining a "1" was higher than expected. The proportion of higher ability students obtaining a "3" was also less than expected and obtaining a "4" was higher than expected.

CR item 50 in Session 4 exhibited less than optimal fit with a Z-statistic of 36.8, exceeding the critical Z-value of 25.1. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was less than expected and obtaining a "1" was higher than expected.

CR item 70 in Session 5 exhibited less than optimal fit with a Z-statistic of 36.8, exceeding the critical Z-value of 25.1. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was less than expected and obtaining a "1" was higher than expected.

Mathematics

The critical Z-value for the sixth-grade Mathematics assessment, based on the calibration sample, is 26.5. Fit values above this critical value may indicate poor model fit. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 98 and 99, respectively.

The Z-statistics for all MC items and most CR items are smaller than the critical Z-value, indicating that the proportions of students with each item score are close to the proportions that would be predicted by the estimated thetas and item parameters.

CR item 12 in Session 1 exhibited less than optimal fit with a Z-statistic of 30.4, exceeding the critical Z-value of 26.5. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was lower than expected and obtaining a "1" was higher than expected. The proportion of higher ability students obtaining a score of "1" was lower than expected.

CR item 49 in Session 3 exhibited less than optimal fit with a Z-statistic of 157.2, exceeding the critical Z-value of 26.5. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was lower than expected and obtaining a "1" was higher than expected. The proportion of higher ability students obtaining a score of "1" was lower than expected and obtaining a "2" was higher than expected.

Seventh Grade

Reading

The critical Z-value for the seventh-grade Reading assessment, based on the calibration sample, is 24.9. Fit values above this critical value may indicate poor model fit. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 100 and 101, respectively.

The item responses fit the IRT model reasonably well. The Z-statistics for most MC items and most CR items are small compared to the critical Z-value, indicating that the proportions of students with each item score are close to the proportions that would be expected by the estimated thetas and item parameters.

MC item 28 in Session 3 exhibited less than optimal fit with a Z-value of 46.7, exceeding the critical Z-value of 24.9. A close examination of the ICC for the item indicated that the proportion of below average ability students obtaining a "1" was lower than expected, and the proportion of about average ability students obtaining a "1" higher than expected.

CR item 37 in Session 4 exhibited less than optimal fit with a Z-value of 33.0, exceeding the critical Z-value of 24.9. A detailed examination of the ICC indicates that the proportion of lower ability students obtaining a "1" was lower than expected and obtaining a "0" was higher than expected.

MC item 44 in Session 4 exhibited less than optimal fit with a Z-value of 35.6, exceeding the critical Z-value of 24.9. A close examination of the ICC for the item indicated that the proportion of lower ability students obtaining a "1" was higher than expected.

Writing

The critical Z-value for the seventh-grade Writing assessment, based on the calibration sample, is 25.4. Fit values above this critical value may indicate poor model fit. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 102 and 103, respectively.

The Z-statistics for all MC items and most CR items are smaller than the critical Z-value, indicating that the proportions of students with each item score are close to the proportions that would be predicted by the estimated thetas and item parameters.

CR item 3A in Session 2 exhibited less than optimal fit with a Z-statistic of 64.4, exceeding the critical Z-value of 25.4. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was less than expected and obtaining a "1" was higher than expected. The proportion of higher ability students obtaining a "3" was also less than expected and obtaining a "4" was higher than expected.

CR item 30 in Session 3 exhibited less than optimal fit with a Z-statistic of 73.6, exceeding the critical Z-value of 25.4. A close examination of the ICCs for the item indicated that the

proportion of lower ability students obtaining a "0" was less than expected and obtaining a "1" was higher than expected. The proportion of higher ability students obtaining a "3" was also less than expected and obtaining a "4" was higher than expected.

Mathematics

The critical Z-value for the seventh-grade Mathematics assessment, based on the calibration sample, is 28.6. Fit values above this critical value may indicate poor model fit. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 104 and 105, respectively.

The Z-statistics for most MC and most CR items are smaller than the critical Z-value, indicating that the proportions of students with each item score are close to the proportions that would be predicted by the estimated thetas and item parameters.

MC item 25 in Session 2 exhibited less than optimal fit with a Z-value of 53.7, exceeding the critical Z-value of 28.6. A close examination of the ICC for the item indicated that the proportion of lower ability students obtaining a "1" was higher than expected.

CR item 43 in Session 3 exhibited less than optimal fit with a Z-value of 37.1, exceeding the critical Z-value of 28.6. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "1" was slightly higher than expected and the proportion of higher ability students obtaining a "1" was slightly lower than expected.

CR item 53 in Session 3 exhibited less than optimal fit with a Z-value of 135.8, exceeding the critical Z-value of 28.6. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was lower than expected and obtaining a "1" was higher than expected. The proportion of higher ability students obtaining a "1" was lower than expected and obtaining a "2" was higher than expected.

Eighth Grade
Reading

The critical Z-value for the eighth-grade Reading assessment, based on the calibration sample, is 24.5. Fit values above this critical value may indicate poor model fit. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 106 and 107, respectively. They indicate that the item responses fit the model reasonably well.

The Z-statistics for all CR items and most MC items are small compared to the critical Z-value, indicating that the proportions of students with each item score are close to the proportions that would be expected by the estimated thetas and item parameters.

MC item 31 in Session 4 exhibited less than optimal fit with a Z-value of 35.1, exceeding the critical Z-value of 24.5. A detailed examination of the ICC indicates that the proportion of slightly below average ability students obtaining a "1" for item 31 was slightly higher than expected.

The item parameter estimation process did not converge for MC item 117 in Session 6. A close examination of the ICC for the item indicated that the student responses do not fit the model well. The item did not discriminate between low or 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.

Writing

The critical Z-value for the <u>eighth-grade Writing</u> assessment, based on the calibration sample, is 25.4. Fit values above this critical value may indicate poor model fit. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 108 and 109, respectively.

The Z-statistics for most MC items and most CR items are smaller than the critical Z-value, indicating that the proportions of students with each item score are close to the proportions that would be predicted by the estimated thetas and item parameters.

CR item 3A in Session 2 exhibited less than optimal fit with a Z-statistic of 70.8, exceeding the critical Z-value of 25.4. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was less than expected and obtaining a "1" was higher than expected. The proportion of higher ability students obtaining a "3" was also less than expected and obtaining a "4" was higher than expected.

CR item 25 in Session 3 exhibited less than optimal fit with a Z-statistic of 31.7, exceeding the critical Z-value of 25.4. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was less than expected and obtaining a "1" was higher than expected.

CR item 58 in Session 5 exhibited less than optimal fit with a Z-statistic of 36.2, exceeding the critical Z-value of 25.4. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was less than expected and obtaining a "1" was higher than expected.

MC item 74 in Session 5 exhibited less than optimal fit with a Z-statistic of 27.4, marginally exceeding the critical Z-value of 25.4. A close examination of the ICC for the item indicated that the proportion of lower and higher ability students obtaining a "1" were slightly less than expected.

CR item 89 in Session 5 exhibited less than optimal fit with a Z-statistic of 44.4, exceeding the critical Z-value of 25.4. A close examination of the ICC for the item indicated that the

proportion of lower ability students obtaining a "0" was less than expected, and obtaining a "1" was higher than expected.

Mathematics

The critical Z-value for the eighth-grade Mathematics assessment, based on the calibration sample, is 27.1. Fit values above this critical value may indicate poor model fit. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 110 and 111, respectively.

The Z-statistics for most MC items and most CR items are smaller than the critical Z-value, indicating that the proportions of students with each item score are close to the proportions that would be predicted by the estimated thetas and item parameters.

CR item 9 in Session 1 exhibited a less than optimal fit with a Z-statistic of 32.7, exceeding the critical Z-value of 27.1. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was higher than expected and obtaining a "1" was lower than expected. The proportion of higher ability students obtaining a "2" was also lower than expected.

CR item 14 in Session 1 exhibited a less than optimal fit with a Z-statistic of 56.2, exceeding the critical Z-value of 27.1. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was lower than expected and obtaining a "1" was higher than expected. The proportion of higher ability students obtaining a "1" was lower than expected and obtaining a "2" or "3" was higher than expected. These observations suggest that the item might have functioned better had it been an item with fewer score levels.

The item parameter estimation process did not converge for MC item 15 in Session 1. A close examination of the ICC for the item indicated that the student responses do not fit the model well. The item did not discriminate between low or 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.

MC item 33 in Session 2 exhibited a less than optimal fit with a Z-statistic of 36.9, exceeding the critical Z-value of 27.1. A close examination of the ICC for the item indicated that the proportion of lower ability students obtaining a "1" was higher than expected

CR item 50 in Session 3 exhibited a less than optimal fit with a Z-statistic of 38.9, exceeding the critical Z-value of 27.1. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "1" was more than expected and the proportion of higher ability students obtaining a "1" was lower than expected. These observations suggest that the item might have functioned better had it been an item with fewer score levels.

CR item 53 in Session 3 exhibited a less than optimal fit with a Z-statistic of 28.7, exceeding the critical Z-value of 27.1. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was more than expected and obtaining a "1" was lower than expected.

Science

The critical Z-value for the <u>eight-grade Science</u> assessment, based on the calibration sample, is 26.7. Fit values above this critical value may indicate poor model fit. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 112 and 113, respectively.

The Z-statistics for most MC items and all CR items are smaller than the critical Z-value, indicating that the proportions of students with each item score are close to the proportions that would be predicted by the estimated thetas and item parameters.

MC item 58 in Session 2 exhibited a less than optimal fit with a Z-statistic of 36.3 exceeding the critical Z-value of 26.7. A close examination of the ICC for the item indicated that the proportion of lower ability students obtaining a "1" was higher than expected. The item analysis results also indicated that the item functioned less than is optimal in the assessment. The point-biserial correlation coefficient for one of the incorrect options on this item was also non-negative (r_{itt} = .12) and larger than the value for the correct response (r_{itt} = .00). The item was rather difficult (p-value = .14).

Ninth Grade

Reading

The critical Z-value for the <u>ninth-grade Reading</u> assessment, based on the calibration sample, is 24.1. Fit values above this critical value may indicate poor model fit. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 114 and 115, respectively. They indicate that the item responses fit the model for all items of the ninth-grade Reading assessment reasonably well.

The Z-statistics for all MC items and most CR items are small compared to the critical Z-value, indicating that the proportions of students with each item score are close to the proportions that would be expected by the estimated thetas and item parameters.

CR item 23 in Session 3 exhibited a less than optimal fit with a Z-statistic of 27.8, exceeding the critical Z-value of 24.1. A close examination of the ICC for the item indicated that the proportion of higher ability students obtaining a "2" was lower than expected and obtaining a "1" was higher than expected. The proportion of lower ability students obtaining a "1" was lower than expected and obtaining a "0" was higher than expected.

Writing

The critical Z-value for the ninth-grade Writing assessment, based on the calibration sample, is 24.6. Fit values above this critical value may indicate poor model fit. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 116 and 117, respectively.

The Z-statistics for all MC items and most CR items are smaller than the critical Z-value, indicating that the proportions of students with each item score are close to the proportions that would be predicted by the estimated thetas and item parameters.

CR item 3A in Session 2 exhibited less than optimal fit with a Z-statistic of 141.9, exceeding the critical Z-value of 24.6. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was less than expected and obtaining a "1" was higher than expected.

CR item 48 in Session 4 exhibited less than optimal fit with a Z-statistic of 24.8, marginally exceeding the critical Z-value of 24.6. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was less than expected and obtaining a "1" was higher than expected.

CR item 90 in Session 5 exhibited less than optimal fit with a Z-statistic of 32.1, exceeding the critical Z-value of 24.6. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was less than expected and obtaining a "1" or a "2" was higher than expected.

Mathematics

The critical Z-value for the ninth-grade Mathematics assessment, based on the calibration sample, is 26.7. Fit values above this critical value may indicate poor model fit. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 118 and 119, respectively.

The Z-statistics for most MC items and most CR items are smaller than the critical Z-value, indicating that the proportions of students with each item score are close to the proportions that would be predicted by the estimated thetas and item parameters.

CR item 28 in Session 2 exhibited a less than optimal fit with a Z-statistic of 31.3, exceeding the critical Z-value of 26.7. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was more than expected and obtaining a "1" was lower than expected. The proportion of average ability students obtaining a "3" was also slightly higher than expected.

The item parameter estimation process did not converge for MC item 51 in Session 3. A close examination of the ICC for the item indicated that the student responses do not fit the model well. The item did not discriminate between low or 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.

Tenth Grade

Reading

The critical Z-value for the <u>tenth-grade Reading</u> assessment, based on the calibration sample, is 24.3. Fit values above this critical value may indicate poor model fit. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 120 and 1217, respectively. They indicate that the item responses fit the model reasonably well for most items of the tenth-grade Reading assessment.

The Z-statistics for all MC items and all CR items are small compared to the critical Z-value, indicating that the proportions of students with each item score are close to the proportions that would be expected by the estimated thetas and item parameters.

Writing

The critical Z-value for the <u>tenth-grade Writing</u> assessment, based on the calibration sample, is 24.1. Fit values above this critical value may indicate poor model fit. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 122 and 123, respectively.

The Z-statistics for all MC items and most CR items are smaller than the critical Z-value, indicating that the proportions of students with each item score are close to the proportions that would be predicted by the estimated thetas and item parameters.

CR item 3A in Session 2 exhibited less than optimal fit with a Z-statistic of 25.7, marginally exceeding the critical Z-value of 24.1. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was less than expected, and obtaining a "1" was higher than expected.

CR item 29 in Session 3 exhibited less than optimal fit with a Z-statistic of 56.1, exceeding the critical Z-value of 24.1. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was less than expected, and obtaining a "1" was higher than expected.

CR item 84 in Session 5 exhibited less than optimal fit with a Z-statistic of 41.0, exceeding the critical Z-value of 24.1. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was less than expected, and obtaining a "1" was higher than expected.

CR item 98 in Session 5 exhibited less than optimal fit with a Z-statistic of 59.4, exceeding the critical Z-value of 24.1. A close examination of the ICCs for the item indicated that the

proportion of lower ability students obtaining a "0" was less than expected, and obtaining a "1" was higher than expected.

Mathematics

The critical Z-value for the tenth-grade Mathematics assessment, based on the calibration sample, is 26.5. Fit values above this critical value may indicate poor model fit. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 124 and 125, respectively.

The Z-statistics for most MC items and most CR items are small compared to the critical Z-value, indicating that the proportions of students with each item score are close to the proportions that would be predicted by the estimated thetas and item parameters.

CR item 40 in Session 2 exhibited less than optimal fit with a Z-statistic of 30.0, exceeding the critical Z-value of 26.5. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was lower than expected and obtaining a "1" was higher than expected. The proportion of higher ability students obtaining a "2" or "3" was also lower than expected and obtaining a "4" slightly higher than expected.

CR item 43 in Session 3 exhibited less than optimal fit with a Z-statistic of 84.6, exceeding the critical Z-value of 26.5. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was higher than expected and obtaining a "1" was lower than expected. The proportion of higher ability students obtaining a "1" was also higher than expected and obtaining a "2" lower than expected.

The item parameter estimation process did not converge for MC items 53 and 57 in Session 3. A close examination of the ICCs for the items indicated that the student responses do not fit the model well. The items 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.

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

An example of this procedure for gender bias analyses follows. The parameters for each item (a_i, b_i, and c_i) and the trait or scale score (θ) for each examinee are estimated for the three-parameter logistic model:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

large positive difference in one range of θ and a large negative difference in another range of θ, yet have a small overall difference.

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

Differential Item Functioning Ratings

The DIF 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.

Third Grade
Reading

CR item 15 indicated a sign of DIF disfavoring male students.

Reading – Spanish Version

No items indicated DIF with any of the groups.

Writing

No items indicated DIF with any of the groups.

Writing – Spanish Version

No items indicated DIF with any of the groups.

Fourth Grade

Reading

CR item 36 indicated a sign of DIF favoring female students, disfavoring male students, and favoring Hispanic students. CR item 105 indicated a sign of DIF disfavoring Hispanic students.

Reading – Spanish Version

No items indicated DIF with any of the groups.

Writing

CR items 71 and 90 both indicated signs of DIF favoring African American students.

Writing – Spanish Version

No items indicated DIF with any of the groups.

Fifth Grade

Reading

CR item 8 indicated a sign of DIF favoring Hispanic students.

Writing

No items indicated signs of DIF with any of the groups.

Mathematics

CR item 51 indicated a sign of DIF favoring African American students.

Sixth Grade

Reading

CR item 11 indicated a sign of DIF favoring Hispanic students, and CR item 45 indicated a sign of DIF disfavoring African American students

Writing

CR item 3B indicated a sign of DIF favoring African American students. CR item 70 indicated a sign of DIF favoring female students and disfavoring male students. CR item 92 indicated a sign of DIF disfavoring male students.

Mathematics

CR item 541 indicated a sign of DIF disfavoring African American students.

Seventh Grade

Reading

CR item 22 indicated a sign of DIF favoring female students and disfavoring male students. CR item 25 indicated a sign of DIF disfavoring male students. CR item 98 indicated a sign of DIF favoring female students. CR items 23 and 37 both indicated a sign of DIF disfavoring African American students

Writing

CR items 3A, 3B, and 70 all indicated signs of DIF favoring female students and disfavoring male students.

Mathematics

CR items 24 and 60 both indicated signs of DIF favoring female students, and disfavoring male students.

Eighth Grade

Reading

CR item 17 and MC item 100 both indicated a sign of DIF disfavoring African American students.

Writing

CR item 58 indicated a sign of DIF disfavoring male students.

Mathematics

CR item 28 indicated signs of DIF favoring female students and favoring African American students. CR item 48 indicated a sign of DIF disfavoring African American students.

Science

CR item 20 indicated signs of DIF favoring female students, and disfavoring male students.

Ninth Grade

Reading

No items indicated DIF with any of the groups.

Writing

CR item 69 indicated signs of DIF favoring female students and disfavoring male students. CR items 74 and 76 69 indicated signs of DIF in the reverse direction, disfavoring female students and favoring male students.

Mathematics

CR item 40 indicated signs of DIF favoring female students and disfavoring female students.

Tenth Grade

Reading

CR items 12, 49, 52, and 53 all indicated a sign of DIF favoring African American students. CR item 52 also indicated a sign of DIF favoring Hispanic students. CR item 21 indicated a sign of DIF disfavoring female students. CR item 24 indicated a sign of DIF disfavoring female students and favoring male students. CR items 34 and 35 both indicated signs of DIF favoring female students and disfavoring male students. CR item 121, finally, indicated a sign of DIF favoring female students.

Writing

CR item 84 indicated signs of DIF favoring female students, disfavoring male students and favoring African American students.

Mathematics

No items indicated DIF with any of the groups.

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.

This report presents the results of the statewide Spring 2003 administration of the Colorado Student Assessment Program (CSAP). In Spring 2003, 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. <u>Fiction</u> – 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. <u>Nonfiction</u> – 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. <u>Vocabulary</u> – 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. <u>Poetry</u> – 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. <u>Paragraph Writing</u> – Students write and edit in a single session. (Writing)

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

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

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

Mathematics

The Colorado Model Content Standards

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

2. <u>Algebra, Patterns, and Functions</u> – 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. <u>Data Analysis, Probability, and Statistics</u> – Students use data collection and analysis, statistics, and probability in problem-solving situations and communicate the reasoning used in solving these problems.

4. <u>Geometric Concepts</u> – Students use geometric concepts, properties, and relationships in problem-solving situations and communicate the reasoning used in solving these problems.

5. <u>Measurement</u> – 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. <u>Operation and Calculation</u> – 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. <u>Number and Operation Sense</u> – 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. <u>Number Sense</u> – Students demonstrate understanding of the concept of equivalency as related to fractions, decimals, and percents. (SA 1, grade 7)

3. <u>Linear Pattern Representation</u> – 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. <u>Multiple Representations of Linear/Nonlinear Functions</u> – 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. <u>Multiple Representations of Functions</u> – 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. <u>Patterns</u> – 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. <u>Experimental Design and Investigations</u> – Students understand and apply scientific questions, hypotheses, variables, and experimental design.

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

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

4. <u>Chemistry</u> – Students understand the properties, composition, structure, and changes of matter.

5. <u>Earth Science</u> – 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 2003 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 2003 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.

A subset of items used in the 2002 (or 2001) forms of the CSAP Reading, Writing, Mathematics, and Science assessments were also included in the 2003 forms, in order 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. Details of the vertical scaling are provided later in this document. The Spanish version for Writing in grade 3 was administered for the first time in 2003.

Test Configuration

Tables 2 through 6 provide information regarding the configuration of the CSAP assessments. Table 2 provides the number of multiple-choice (MC) items versus 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.

Third Grade

Reading and Writing were assessed using separate forms.

Reading

The third-grade Reading test consisted of 41 items totaling 52 points. There were 34 MC items worth 34 points and 7 CR items worth 18 points.

The numbers of CR items that are worth the various numbers of points are as follows: , 4 items worth 2 points, 2 item worth 3 points, 1 item worth 4 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

> CS 1 has 34 MC and 7 CR (18 points) totaling 52 points.
> SA 1/4 has 10 MC and 7 CR (18 points) totaling 28 points.
> SA 2 has 14 MC and 0 CR (0 points) totaling 14 points.
> SA 3 has 10 MC and 0 CR (0 points) totaling 10 points

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Reading – Spanish Version

The third-grade Spanish Reading test consisted of 40 items totaling 49 points. There were 32 MC items worth 32 points and 8 CR items worth 17 points.

The numbers of CR items that are worth the various numbers of points are as follows: 7 items worth 2 points, 1 item worth 3 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

CS 1 has 32 MC and 8 CR (17 points) totaling 49 points.
SA 1/4 has 12 MC and 7 CR (15 points) totaling 27 points.
SA 2 has 10 MC and 1 CR (2 points) totaling 12 points.
SA 3 has 10 MC and 0 CR (0 points) totaling 10 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Writing

The third-grade Writing test consisted of 53 items totaling 56 points. There were 35 MC items worth 35 points and 18 CR items worth 21 points.

The numbers of CR items that are worth the various numbers of points are as follows:
15 items worth 1 point, 3 items worth 2 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

CS 2 has 7 MC and 9 CR (12 points) totaling 19 points.
CS 3 has 28 MC and 9 CR (9 points) totaling 37 points.
SA 5 has 0 MC and 12 CR (15 points) totaling 15 points.
SA 7 has 16 MC and 3 CR (3 points) totaling 19 points.
SA 8 has 12 MC and 3 CR (3 points) totaling 15 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Writing – Spanish Version

The third-grade Spanish Writing test consisted of 53 items totaling 56 points. There were 35 MC items worth 35 points and 18 CR items worth 21 points.

The numbers of CR items that are worth the various numbers of points are as follows:
15 items worth 1 point, 3 items worth 2 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

CS 2 has 6 MC and 9 CR (12 points) totaling 18 points.
CS 3 has 29 MC and 9 CR (9 points) totaling 38 points.
SA 5 has 0 MC and 12 CR (15 points) totaling 15 points.
SA 7 has 21 MC and 2 CR (2 points) totaling 23 points.
SA 8 has 8 MC and 4 CR (4 points) totaling 12 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Fourth Grade

Reading and Writing were assessed using a single form.

Reading

The fourth-grade Reading test consisted of 70 items totaling 91 points. There were 56 MC items worth 56 points and 14 CR items worth 35 points.

The numbers of CR items that are worth the various numbers of points are as follows:
8 items worth 2 points, 5 items worth 3 points, 1 items worth 4 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

    CS 1 has 25 MC and 4 CR (11 points) totaling 36 points.
    CS 4 has 14 MC and 1 CR (3 points) totaling 17 points.
    CS 5 has 11 MC and 3 CR (7 points) totaling 18 points.
    CS 6 has 6 MC and 6 CR (14 points) totaling 20 points.
    SA 1/4 has 26 MC and 11 CR (28 points) totaling 54 points.
    SA 2 has 11 MC and 0 CR (0 points) totaling 11 points.
    SA 3 has 8 MC and 0 CR (0 points) totaling 8 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Reading – Spanish Version

The fourth-grade Spanish Reading test consisted of 66 items totaling 83 points. There were 52 MC items worth 52 points and 14 CR items worth 31 points.

The numbers of CR items that are worth the various numbers of points are as follows:
2 items worth 1 point, 8 items worth 2 points, 3 items worth 3 points, 1 item worth 4 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

    CS 1 has 27 MC and 1 CR (4 points) totaling 31 points.
    CS 4 has 10 MC and 5 CR (9 points) totaling 19 points.
    CS 5 has 8 MC and 6 CR (14 points) totaling 22 points.
    CS 6 has 7 MC and 2 CR (4 points) totaling 11 points.
    SA 1/4 has 23 MC and 5 CR (9 points) totaling 32 points.
    SA 2 has 14 MC and 3 CR (8 points) totaling 22 points.
    SA 3 has 7 MC and 0 CR (0 points) totaling 7 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Note that these numbers do not include a multiple-choice item that was included on the Spanish Reading assessment but was dropped due to poor fit to the item response model and aberrant student response patterns (discussed further in this report). This was item 107 in session 6 in the operational test booklet.

Writing

The fourth-grade Writing test consisted of 53 items totaling 69 points. There were 40 MC items worth 40 points and 13 CR items worth 29 points.

The numbers of CR items that are worth the various numbers of points are as follows:
7 items worth 1 point, 1 item worth 2 points, 5 items worth 4 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

CS 2 has 13 MC and 6 CR (21 points) totaling 34 points.
CS 3 has 27 MC and 7 CR (8 points) totaling 35 points.
SA 5 has 3 MC and 3 CR (12 points) totaling 15 points.
SA 6 has 0 MC and 4 CR (11 points) totaling 11 points.
SA 7 has 12 MC and 2 CR (2 points) totaling 14 points.
SA 8 has 15 MC and 4 CR (4 points) totaling 19 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Writing – Spanish Version

The fourth-grade Spanish Writing test consisted of 52 items totaling 68 points. There were 39 MC items worth 39 points and 13 CR items worth 29 points.

The numbers of CR items that are worth the various numbers of points are as follows:
7 items worth 1 point, 1 item worth 2 points, 5 items worth 4 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

CS 2 has 5 MC and 6 CR (21 points) totaling 26 points.
CS 3 has 34 MC and 7 CR (8 points) totaling 42 points.
SA 5 has 0 MC and 3 CR (12 points) totaling 12 points.
SA 6 has 0 MC and 4 CR (11 points) totaling 11 points.
SA 7 has 27 MC and 2 CR (2 points) totaling 29 points.
SA 8 has 7 MC and 4 CR (4 points) totaling 11 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Note that these numbers do not include a multiple-choice item that was included on the Spanish Writing assessment but was dropped due to poor fit to the item response model and aberrant student response patterns (discussed further in this report). This was item 54 in session 5 in the operational test booklet.

Fifth Grade

Reading and Writing were assessed using a single form. Mathematics was assessed using a separate form.

Reading

The fifth-grade Reading test consisted of 70 items totaling 90 points. There were 56 MC items worth 56 points and 14 CR items worth 34 points.

The numbers of CR items that are worth the various numbers of points are as follows:
8 items worth 2 points, 6 items worth 3 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

   CS 1 has 17 MC and 2 CR (5 points) totaling 22 points.
   CS 4 has 16 MC and 2 CR (5 points) totaling 21 points.
   CS 5 has 14 MC and 3 CR (8 points) totaling 22 points.
   CS 6 has 9 MC and 7 CR (16 points) totaling 25 points.
   SA 1/4 has 15 MC and 9 CR (21 points) totaling 36 points.
   SA 2 has 18 MC and 2 CR (5 points) totaling 23 points.
   SA 3 has 9 MC and 0 CR (0 points) totaling 9 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Writing

The fifth-grade Writing test consisted of 53 items totaling 69 points. There were 40 MC items worth 40 points and 13 CR items worth 29 points.

The numbers of CR items that are worth the various numbers of points are as follows:
7 items worth 1 point, 1 item worth 2 points, 5 items worth 4 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

   CS 2 has 15MC and 6 CR (21 points) totaling 36 points.
   CS 3 has 25 MC and 7 CR (8 points) totaling 33 points.
   SA 5 has 5 MC and 3 CR (12 points) totaling 17 points.
   SA 6 has 0 MC and 4 CR (11 points) totaling 11 points.
   SA 7 has 18 MC and 2 CR (2 points) totaling 20 points.
   SA 8 has 14 MC and 4 CR (4 points) totaling 18 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Mathematics

The fifth-grade Mathematics test consisted of 69 items totaling 96 points. There were 54 MC items worth 54 points and 15 CR items worth 42 points.

The numbers of CR items that are worth the various numbers of points are as follows:
6 items worth 2 points, 6 items worth 3 points, 3 items worth 4 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

CS 1 has 14 MC and 2 CR (5 points) totaling 19 points.
CS 2 has 8 MC and 4 CR (11 points) totaling 19 points.
CS 3 has 7 MC and 4 CR (12 points) totaling 19 points.
CS 4/5 has 11 MC and 3 CR (9 points) totaling 20 points.
CS 6 has 14 MC and 2 CR (5 points) totaling 19 points.
SA 1 has 12 MC and 3 CR (8 points) totaling 20 points.
SA 2 has 4 MC and 3 CR (9 points) totaling 13 points.
SA 3 has 3 MC and 2 CR (7 points) totaling 10 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Sixth Grade

Reading and Writing were assessed using a single form. Mathematics was assessed using a separate form.

Reading

The sixth-grade Reading test consisted of 69 items totaling 91 points. There were 54 MC items worth 54 points and 15 CR items worth 37 points.

The numbers of CR items that are worth the various numbers of points are as follows:
6 items worth 2 points, 6 items worth 3 points, 3 item worth 4 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

CS 1 has 20 MC and 2 CR (6 points) totaling 26 points.
CS 4 has 14 MC and 3 CR (7 points) totaling 21 points.
CS 5 has 11 MC and 4 CR (9 points) totaling 20 points.
CS 6 has 9 MC and 6 CR (15 points) totaling 24 points.
SA 1/4 has 17 MC and 7 CR (16 points) totaling 33 points.
SA 2 has 13 MC and 4 CR (12 points) totaling 25 points.
SA 3 has 13 MC and 0 CR (0 points) totaling 13 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Writing

The sixth-grade Writing test consisted of 53 items totaling 69 points. There were 40 MC items worth 40 points and 13 CR items worth 29 points.

The numbers of CR items that are worth the various numbers of points are as follows:
7 items worth 1 point, 1 item worth 2 points, 5 items worth 4 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

CS 2 has 17 MC and 6 CR (21 points) totaling 38 points.
CS 3 has 23 MC and 7 CR (8 points) totaling 31 points.
SA 5 has 9 MC and 3 CR (12 points) totaling 21 points.
SA 6 has 0 MC and 4 CR (11 points) totaling 11 points.
SA 7 has 21 MC and 1 CR (1 points) totaling 22 points.
SA 8 has 8 MC and 5 CR (5 points) totaling 13 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Mathematics

The sixth-grade Mathematics test consisted of 60 items totaling 87 points. There were 45 MC items worth 45 points and 15 CR items worth 42 points.

The numbers of CR items that are worth the various numbers of points are as follows:
6 items worth 2 points, 6 items worth 3 points, 3 items worth 4 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

CS 1 has 9 MC and 2 CR (7 points) totaling 16 points.
CS 2 has 7 MC and 3 CR (10 points) totaling 17 points.
CS 3 has 7 MC and 4 CR (11 points) totaling 18 points.
CS 4/5 has 12 MC and 4 CR (10 points) totaling 22 points.
CS 6 has 10 MC and 2 CR (4 points) totaling 14 points.
SA 1 has 12 MC and 4 CR (11 points) totaling 23 points.
SA 2 has 7 MC and 2 CR (6 points) totaling 13 points.
SA 3 has 7 MC and 2 CR (5 points) totaling 12 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Seventh Grade

Reading and Writing were assessed using a single form. Mathematics was assessed using a separate form.

Reading

The seventh-grade Reading test consisted of 71 items totaling 91 points. There were 57 MC items worth 57 points and 14 CR items worth 34 points.

The numbers of CR items that are worth the various numbers of points are as follows:
9 items worth 2 points, 4 items worth 3 points, 1 item worth 4 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

CS 1 has 21 MC and 3 CR (9 points) totaling 30 points.
CS 4 has 13 MC and 3 CR (6 points) totaling 19 points.

CS 5 has 15 MC and 3 CR (6 points) totaling 21 points.
CS 6 has 8 MC and 5 CR (13 points) totaling 21 points.
SA 1 has 12 MC and 4 CR (11 points) totaling 23 points.
SA 2 has 11 MC and 3 CR (8 points) totaling 19 points.
SA 3 has 14 MC and 0 CR (0 points) totaling 14 points.
SA 4 has 5 MC and 4 CR (9 points) totaling 14 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Writing

The seventh-grade Writing test consisted of 53 items totaling 69 points. There were 40 MC items worth 40 points and 13 CR items worth 29 points.

The numbers of CR items that are worth the various numbers of points are as follows:
7 items worth 1 point, 1 item worth 2 points, 5 items worth 4 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

CS 2 has 20 MC and 6 CR (21 points) totaling 41 points.
CS 3 has 20 MC and 7 CR (8 points) totaling 28 points.
SA 5 has 6 MC and 3 CR (12 points) totaling 18 points.
SA 6 has 0 MC and 4 CR (11 points) totaling 11 points.
SA 7 has 13 MC and 2 CR (2 points) totaling 15 points.
SA 8 has 12 MC and 4 CR (4 points) totaling 16 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Mathematics

The seventh-grade Mathematics test consisted of 60 items totaling 87 points. There were 45 MC items worth 45 points and 15 CR items worth 42 points.

The numbers of CR items that are worth the various numbers of points are as follows:
6 items worth 2 points, 6 items worth 3 points, 3 items worth 4 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

CS 1/6 has 15 MC and 4 CR (11 points) totaling 26 points.
CS 2 has 9 MC and 3 CR (8 points) totaling 17 points.
CS 3 has 9 MC and 3 CR (9 points) totaling 18 points.
CS 4/5 has 12 MC and 5 CR (14 points) totaling 26 points.
SA 1 has 6 MC and 2 CR (5 points) totaling 11 points.
SA 2 has 3 MC and 3 CR (9 points) totaling 12 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Eighth Grade

Reading and Writing were assessed using a single form. Mathematics and Science were assessed using two separate forms.

Reading

The eighth-grade Reading test consisted of 69 items totaling 90 points. There were 54 MC items worth 54 points and 15 CR items worth 36 points.

The numbers of CR items that are worth the various numbers of points are as follows:
10 items worth 2 points, 4 items worth 3 points, 1 item worth 4 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

CS 1 has 15MC and 6 CR (14 points) totaling 29 points.
CS 4 has 15 MC and 2 CR (6 points) totaling 21 points.
CS 5 has 12 MC and 3 CR (6 points) totaling 18 points.
CS 6 has 12 MC and 4 CR (10 points) totaling 22 points.
SA 1 has 11 MC and 3 CR (8 points) totaling 19 points.
SA 2 has 16 MC and 5 CR (10 points) totaling 26 points.
SA 3 has 10 MC and 0 CR (0 points) totaling 10 points.
SA 4 has 5 MC and 4 CR (12 points) totaling 17 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Note that these numbers do not include a multiple-choice item that was included on the Reading assessment but was dropped due to poor fit to the item response model and aberrant student response patterns (discussed further in this report). This was item 117 in session 6 in the operational test booklet.

Writing

The eighth-grade Writing test consisted of 53 items totaling 69 points. There were 41 MC items worth 41 points and 12 CR items worth 28 points.

The numbers of CR items that are worth the various numbers of points are as follows:
6 items worth 1 point, 1 item worth 2 points, 5 items worth 4 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

CS 2 has 21 MC and 6 CR (21 points) totaling 42 points.
CS 3 has 20 MC and 6 CR (7 points) totaling 27 points.
SA 5 has 12 MC and 3 CR (12 points) totaling 24 points.
SA 6 has 0 MC and 4 CR (11 points) totaling 11 points.
SA 7 has 18 MC and 1 CR (1 point) totaling 19 points.
SA 8 has 10 MC and 4 CR (4 points) totaling 14 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Note that these numbers do not include a constructed-response item that was included on the Writing assessment but was dropped due to poor fit to the item response model and aberrant student response patterns (discussed further in this report). This was item 2A in session 2 in the operational test booklet.

Mathematics

The eighth-grade Mathematics test consisted of 59 items totaling 86 points. There were 44 MC items worth 44 points and 15 CR items worth 42 points.

The numbers of CR items that are worth the various numbers of points are as follows:
6 items worth 2 points, 6 items worth 3 points, 3 items worth 4 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

    CS 1/6 has 10 MC and 4 CR (11 points) totaling 21 points.
    CS 2 has 11 MC and 4 CR (11 points) totaling 22 points.
    CS 3 has 9 MC and 3 CR (8 points) totaling 17 points.
    CS 4/5 has 14 MC and 4 CR (12 points) totaling 26 points.
    SA 1 has 5 MC and 3 CR (9 points) totaling 14 points.
    SA 2 has 9 MC and 6 CR (15 points) totaling 24 points.
    SA 3 has 5 MC and 2 CR (7 points) totaling 12 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Note that these numbers do not include a multiple-choice item that was included on the Mathematics assessment but was dropped due to poor fit to the item response model and aberrant student response patterns (discussed further in this report). This was item 15 in session 1 in the operational test booklet.

Science

The eighth-grade Science test consisted of 84 items totaling 99 points. There were 64 MC items worth 64 points and 20 CR items worth 35 points.

The numbers of CR items that are worth the various numbers of points are as follows:
8 items worth 1 point, 10 items worth 2 points, 1 item worth 3 points, 1 item worth 4 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

    CS 1 has 13 MC and 8 CR (14 points) totaling 27 points.
    CS 2 has 16 MC and 2 CR (4 points) totaling 20 points.
    CS 3 has 11 MC and 5 CR (9 points) totaling 20 points.

CS 4 has 16 MC and 3 CR (6 points) totaling 22 points.
CS 5/6 has 8 MC and 2 CR (2 points) totaling 10 points.
SA 1 has 6 MC and 3 CR (5 points) totaling 11 points.
SA 2 has 7 MC and 4 CR (7 points) totaling 14 points.
SA 3 has 8 MC and 1 CR (2 point) totaling 10 points.
SA 4 has 8 MC and 1 CR (2 points) totaling 10 points.
SA 5 has 11 MC and 3 CR (6 points) totaling 17 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Ninth Grade

Reading and Writing were assessed using a single form. Mathematics was assessed using a separate form.

Reading

The ninth-grade Reading test consisted of 73 items totaling 96 points. There were 58 MC items worth 58 points and 15 CR items worth 38 points.

The numbers of CR items that are worth the various numbers of points are as follows:
8 items worth 2 points, 6 items worth 3 points, 1 item worth 4 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

CS 1 has 15 MC and 6 CR (14 points) totaling 29 points.
CS 4 has 13 MC and 4 CR (11 points) totaling 24 points.
CS 5 has 16 MC and 1 CR (3 points) totaling 19 points.
CS 6 has 14 MC and 4 CR (10 points) totaling 24 points.
SA 1 has 12 MC and 3 CR (9 points) totaling 21 points.
SA 2 has 12 MC and 6 CR (13 points) totaling 25 points.
SA 3 has 12 MC and 0 CR (0 points) totaling 12 points.
SA 4 has 6 MC and 5 CR (13 points) totaling 19 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Writing

The ninth-grade Writing test consisted of 53 items totaling 69 points. There were 40 MC items worth 40 points and 13 CR items worth 29 points.

The numbers of CR items that are worth the various numbers of points are as follows:
7 items worth 1 point, 1 item worth 2 points, 5 items worth 4 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

CS 2 has 20 MC and 6 CR (21 points) totaling 41 points.
CS 3 has 20 MC and 7 CR (8 points) totaling 28 points.
SA 5 has 10 MC and 3 CR (12 points) totaling 22 points.
SA 6 has 0 MC and 4 CR (11 points) totaling 11 points.
SA 7 has 14 MC and 2 CR (2 points) totaling 16 points.
SA 8 has 10 MC and 4 CR (4 points) totaling 14 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Mathematics

The ninth-grade Mathematics test consisted of 59 items totaling 86 points. There were 44 MC items worth 44 points and 15 CR items worth 42 points.

The numbers of CR items that are worth the various numbers of points are as follows:
6 items worth 2 points, 6 items worth 3 points, 3 items worth 4 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

CS 1/6 has 10 MC and 3 CR (7 points) totaling 17 points.
CS 2 has 11 MC and 5 CR (14 points) totaling 25 points.
CS 3 has 11 MC and 4 CR (11 points) totaling 22 points.
CS 4/5 has 12 MC and 3 CR (10 points) totaling 22 points.
SA 1 has 3 MC and 4 CR (13 points) totaling 16 points.
SA 2 has 3 MC and 3 CR (10 points) totaling 13 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Note that these numbers do not include a multiple-choice item that was included on the Mathematics assessment but was dropped due to poor fit to the item response model and aberrant student response patterns (discussed further in this report). This was item 51 in session 3 in the operational test booklet.

Tenth Grade

Reading and Writing were assessed using a single form. Mathematics was assessed using a separate form.

Reading

The tenth-grade Reading test consisted of 75 items totaling 101 points. There were 61 MC items worth 61 points and 14 CR items worth 40 points.

The numbers of CR items that are worth the various numbers of points are as follows:
3 items worth 2 points, 10 items worth 3 points, 1 item worth 4 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

   CS 1 has 21 MC and 1 CR (3 points) totaling 24 points.
   CS 4 has 10 MC and 5 CR (14 points) totaling 24 points.
   CS 5 has 12 MC and 3 CR (9 points) totaling 21 points.
   CS 6 has 18 MC and 5 CR (14 points) totaling 32 points.
   SA 1 has 16 MC and 6 CR (17 points) totaling 33 points.
   SA 2 has 15 MC and 3 CR (8 points) totaling 23 points.
   SA 3 has 10 MC and 0 CR (0 points) totaling 10 points.
   SA 4 has 8 MC and 2 CR (6 points) totaling 14 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Writing

The tenth-grade Writing test consisted of 53 items totaling 69 points. There were 40 MC items worth 40 points and 13 CR items worth 29 points.

The numbers of CR items that are worth the various numbers of points are as follows:
7 items worth 1 point, 1 item worth 2 points, 5 items worth 4 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

   CS 2 has 20 MC and 6 CR (21 points) totaling 41 points.
   CS 3 has 20 MC and 7 CR (8 points) totaling 28 points.
   SA 5 has 9 MC and 3 CR (12 points) totaling 17 points.
   SA 6 has 0 MC and 4 CR (11 points) totaling 11 points.
   SA 7 has 18 MC and 4 CR (4 points) totaling 22 points.
   SA 8 has 8 MC and 2 CR (2 points) totaling 10 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Mathematics

The tenth-grade Mathematics test consisted of 58 items totaling 85 points. There were 43 MC items worth 43 points and 15 CR items worth 42 points.

The numbers of CR items that are worth the various numbers of points are as follows:
6 items worth 2 points, 6 items worth 3 points, 3 items worth 4 points.

The numbers of items and points by Content Standard (CS) and Sub-Content Area (SA) are:

   CS 1/6 has 12 MC and 2 CR (5 points) totaling 17 points.
   CS 2 has 11 MC and 5 CR (14 points) totaling 25 points.
   CS 3 has 10 MC and 4 CR (11 points) totaling 21 points.
   CS 4/5 has 10 MC and 4 CR (12 points) totaling 22 points.

SA 1 has 4 MC and 3 CR (9 points) totaling 13 points.
SA 2 has 5 MC and 2 CR (5 points) totaling 10 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Note that these numbers do not include two multiple-choice items that were included on the Mathematics assessment but was dropped due to poor fit to the item response model and aberrant student response patterns (discussed further in this report). These were items 53 and 57 in session 4 in the operational test booklet.

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 items on the test. 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).

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 2003 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, had been established in 2001. The vertical scales for Writing, spanning grades 3 through 10, and for Mathematics, spanning grades 5

SA 1 has 4 MC and 3 CR (9 points) totaling 13 points.
SA 2 has 5 MC and 2 CR (5 points) totaling 10 points.

Every item is associated with a CS but not all items are associated with a SA, so the points of the SA may not add up to the total of the points for the total test.

Note that these numbers do not include two multiple-choice items that were included on the Mathematics assessment but was dropped due to poor fit to the item response model and aberrant student response patterns (discussed further in this report). These were items 53 and 57 in session 4 in the operational test booklet.

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 items on the test. 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).

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 2003 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, had been established in 2001. The vertical scales for Writing, spanning grades 3 through 10, and for Mathematics, spanning grades 5

through 10, had been 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.

The 2003 CSAP tests for each grade of Reading, Writing, and Mathematics contained anchor items from the previous administration (2002) for the same grade, that allowed the 2003 tests to be anchored to the previously established scales using the Stocking and Lord (1983) procedure.

By means of these equatings within each grade, the unique metrics of the CSAP Reading, Writing, and Mathematics vertical scales were maintained. The vertical scales had a lowest obtainable scale score (LOSS) of 150 in grade 3 Reading and Writing, a LOSS of 220 in grade 5 Mathematics, a highest obtainable scale score (HOSS) of 999 in grade 10 Reading, and a HOSS of 950 in grade 10 Writing and Mathematics.

The scaling and equating design for Science and for the Spanish versions of the grades 3 and 4 Reading assessment, and the grade 4 Writing assessment, remained unchanged from the previous year. The 2003 CSAP tests for Spanish reading, Spanish Writing, and Science contained anchor items from the previous administration (2002) that allowed the 2003 tests to be anchored to the previously established scale using the Stocking and Lord (1983) procedure.

These scaling and calibration methods are presented in Part 5.

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). New cut scores were set for the Spanish version of the grade 3 Writing assessment. For all other assessments the cut scores from the previous year were used. 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. 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

through 10, had been 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.

The 2003 CSAP tests for each grade of Reading, Writing, and Mathematics contained anchor items from the previous administration (2002) for the same grade, that allowed the 2003 tests to be anchored to the previously established scales using the Stocking and Lord (1983) procedure.

By means of these equatings within each grade, the unique metrics of the CSAP Reading, Writing, and Mathematics vertical scales were maintained. The vertical scales had a lowest obtainable scale score (LOSS) of 150 in grade 3 Reading and Writing, a LOSS of 220 in grade 5 Mathematics, a highest obtainable scale score (HOSS) of 999 in grade 10 Reading, and a HOSS of 950 in grade 10 Writing and Mathematics.

The scaling and equating design for Science and for the Spanish versions of the grades 3 and 4 Reading assessment, and the grade 4 Writing assessment, remained unchanged from the previous year. The 2003 CSAP tests for Spanish reading, Spanish Writing, and Science contained anchor items from the previous administration (2002) that allowed the 2003 tests to be anchored to the previously established scale using the Stocking and Lord (1983) procedure.

These scaling and calibration methods are presented in Part 5.

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). New cut scores were set for the Spanish version of the grade 3 Writing assessment. For all other assessments the cut scores from the previous year were used. 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. 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, means as well as 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 2003 third-grade Reading assessment is 568 with a standard deviation of 75.1. The mean scale score for female students is 575 with a standard deviation of 72.8, and the mean scale score for male students is 562 with a standard deviation of 76.7.

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 573 to 606. The median scale scores vary between 566 and 576, and all are close to the median for the total test scale score, 570.

The mean percents of the maximum obtainable score for the sub-content areas range from 72.0 to 83.8. The mean percent of the maximum obtainable score for the total test is 75.8.

Reading – Spanish Version

The mean scale score for the total population of students taking the 2003 third-grade Spanish Reading assessment is 523 with a standard deviation of 41.6. The mean scale score for female students is 530 with a standard deviation of 40.6, and the mean scale score for male students is 516 with a standard deviation of 41.5.

The scale score frequency distribution of the third-grade 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 525 to 538; the median scale scores for the sub-content areas vary between 524 and 526, and all are close to the median for the total test scale score, 525.

The mean percents of the maximum obtainable score for the sub-content areas range from 60.4 to 70.7. The mean percent of the maximum obtainable score for the total test is 66.5.

Writing

The mean scale score for the total population of students taking the 2003 third-grade Writing assessment is 477 with a standard deviation of 57.5. The mean scale score for female students is 485 with a standard deviation of 58.2, and the mean scale score for male students is 468 with a standard deviation of 55.7.

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 mean scale scores for the content standards vary between 481 and 490, with standard deviations of 67.2 and 83.5, respectively. The mean scale scores for the sub-content areas range from 483 to 513. The median scale scores vary between 474 and 497 for the content standards, and between 478 and 484 for the sub-content areas. The median for the total test scale score is 475.

The mean percents of the maximum obtainable score for CS 2 (Write for a Variety of Purposes) and CS 3 (Write Using Conventions) are 80.0 and 85.6, respectively. The mean percent of the maximum obtainable score for the total test is 83.7. The mean percents of the maximum obtainable score for the sub-content areas range from 78.5 to 84.6.

Writing – Spanish Version

The mean scale score for the total population of students taking the 2003 third-grade Spanish Writing assessment is 500 with a standard deviation of 57.8. The mean scale score for female students is 511 with a standard deviation of 55.2, and the mean scale score for male students is 488 with a standard deviation of 58.3.

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 content standards range from 502 to 504; the median scale scores vary between 500 and 501, and all are close to the median for the total test scale score, 499. The mean scale scores for the sub-content areas range from 509 to 513; the median scale

scores for the sub-content areas vary between 500 and 504, and all are close to the median for the total test scale score, 499.

The mean percents of the maximum obtainable score range from 60.2 to 74.6 for the content standards, and from 65.8 to 74.3 for the sub-content areas. The mean percent of the maximum obtainable score for the total test is 70.0.

Fourth Grade

Reading

The mean scale score for the total population of students taking the 2003 fourth-grade Reading assessment is 587 with a standard deviation of 64.6. The mean scale score for female students is 594 with a standard deviation of 61.9, and the mean scale score for male students is 580 with a standard deviation of 66.3.

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 586 to 595. The mean scale scores for the sub-content areas range from 587 to 629. The median scale scores vary between 591 and 595 for the content standards, and between 592 and 604 for the sub-content areas, and all are close to the median for the total test scale score, 592.

The mean percents of the maximum obtainable score for the content standards range from 55.0 on CS 6 (Literature) to 76.0 on CS 1 (Reading Comprehension). The mean percent of the maximum obtainable score for the total test is 66.2. The mean percents for the sub-content areas range from 64.5 to 74.9. The observed ranges indicate that the average item difficulty differed somewhat across the content standards and across the sub-content areas.

Reading – Spanish Version

The mean scale score for the total population of students taking the 2003 fourth-grade Spanish Reading assessment is 527 with a standard deviation of 44.2. The mean scale score for female students is 533 with a standard deviation of 40.2, and the mean scale score for male students is 521 with a standard deviation of 46.4.

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 523 to 528. The mean scale scores for the sub-content areas range from 523 to 531. The median scale scores vary between 528 and 531 for the content standards, and between 530 and 532 for the sub-content areas, and all are close to the median for the total test scale score, 532.

The mean percents of the maximum obtainable score for the content standards range from 42.4 on CS 5 (Use of Literary Information) to 61.4 on CS 1 (Reading Comprehension). The mean percent of the maximum obtainable score for the total test is 52.8. The mean percents for the sub-content areas range from 53.8 to 58.1. The sub-content areas appear to be similar in average item difficulty, whereas the content standards appear to vary somewhat.

Writing

The mean scale score for the total population of students taking the 2003 fourth-grade Writing assessment is 486 with a standard deviation of 55.1. The mean scale score for female students is 497 with a standard deviation of 55.2, and the mean scale score for male students is 477 with a standard deviation of 53.3.

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 486 and 491. The mean scale scores for the sub-content areas range from 486 to 511. The median scale scores vary between 486 and 488 for the content standards, and between 479 and 494 for the sub-content areas, and all are fairly close to the median for the total test scale score, 487.

The mean percents of the maximum obtainable score for CS 2 (Write for a Variety of Purposes) and CS 3 (Write Using Conventions) are 68.7 and 73.3, respectively. The mean percent of the maximum obtainable score for the total test is 71.0. The mean percents of the maximum obtainable score for the sub-content areas range from 61.6 to 80.2. Note that the mean percents are closer for the content standards than for the sub-content areas.

Writing – Spanish Version

The mean scale score for the total population of students taking the 2003 fourth-grade Spanish Writing assessment is 519 with a standard deviation of 43.0. The mean scale score for female students is 525 with a standard deviation of 42.2, and the mean scale score for male students is 513 with a standard deviation of 42.9.

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 scores for the content standards vary between 519 and 523. The mean scale scores for the sub-content areas range from 508 to 524. The median scale scores vary between

516 and 522 for the content standards, and between 515 and 522 for the sub-content areas, and most are close to the median for the total test scale score, 520.

The mean percents of the maximum obtainable score for CS 2 (Write for a Variety of Purposes) and CS 3 (Write Using Conventions) are 54.9 and 54.2, respectively. The mean percent of the maximum obtainable score for the total test is 54.4. The mean percents of the maximum obtainable score for the sub-content areas range from 41.1 to 58.3.

Fifth Grade

Reading

The mean scale score for the total population of students taking the 2003 fifth-grade Reading assessment is 608 with a standard deviation of 68.7. The mean scale score for female students is 615 with a standard deviation of 64.8 and the mean scale score for male students is 601 with a standard deviation of 71.5.

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 608 to 620. The mean scale scores for the sub-content areas range from 609 to 676. The median scale scores vary between 614 and 616 for the content standards, and between 614 and 619 for the sub-content areas, and all are close to the median for the total test scale score, 615.

The mean percents of the maximum obtainable score for content standards range from 59.9 on CS 5 (Use of Literary Information) to 74.0 on CS 1 (Reading Comprehension). The mean percent of the maximum obtainable score for the total test is 64.8. The mean percents for the sub-content areas range from 62.8 to 79.6.

Writing

The mean scale score for the total population of students taking the 2003 fifth-grade Writing assessment is 502 with a standard deviation of 57.2. The mean scale score for female students is 513 with a standard deviation of 57.3, and the mean scale score for male students is 492 with a standard deviation of 55.2.

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 503 and 507. The mean scale scores for the sub-content areas range from 503 to 531. The median scale scores vary between 503 and 504 for the content standards, and between 500 and 505 for the sub-content areas, and all are close to the median for the total test scale score, 504.

The mean percents of the maximum obtainable score for CS 2 (Write for a Variety of Purposes) and CS 3 (Write Using Conventions) are 64.3 and 74.8, respectively. The mean percent of the maximum obtainable score for the total test is 69.3. The mean percents of the maximum obtainable score for the sub-content areas range from 58.3 to 73.8.

Mathematics

The mean scale score for the total population of students taking the 2003 fifth-grade Mathematics assessment is 503 with a standard deviation of 72.2. The mean scale score for female students is 502 with a standard deviation of 69.9, and the mean scale score for male students is 503 with a standard deviation of 74.4.

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 506 to 517. The mean scale scores for the sub-content areas range from 511 to 539. The median scale scores vary between 505 and 506 for the content standards, and between 502 and 505 for the sub-content areas, and all are close to the median for the total test scale score, 505.

The mean percents of the maximum obtainable score for the content standards range from 62.4 on CS 4 (Geometric Concepts) to 72.6 on CS 2 (Algebra, Patterns and Functions). The mean percent of the maximum obtainable score for the total test is 67.9. The mean percents for the sub-content areas range from 65.8 to 75.6.

Sixth Grade

Reading

The mean scale score for the total population of students taking the 2003 sixth-grade Reading assessment is 620 with a standard deviation of 65.3. The mean scale score for female students is 627 with a standard deviation of 60.7, and the mean scale score for male students is 614 with a standard deviation of 68.9.

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 620 to 624. The mean scale scores for the sub-content areas range from 622 to 628. The median scale scores vary between 627 and 629 for the content standards, and between 627 and 629 for the sub-content areas, and all are close to the median for the total test scale score, 628.

The mean percents of the maximum obtainable score for content standards range from 60.0 on CS 6 (Literature) to 67.6 on CS 1 (Reading Comprehension). The mean percent of the maximum obtainable score for the total test is 63.3. The mean percents for the sub-content areas range from 62.7 to 64.2.

Writing

The mean scale score for the total population of students taking the 2003 sixth-grade Writing assessment is 520 with a standard deviation of 62.1. The mean scale score for female students is 533 with a standard deviation of 60.4, and the mean scale score for male students is 507 with a standard deviation of 61.1.

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 521 and 522. The mean scale scores for the sub-content areas range from 521 to 549. The median scale scores vary between 520 and 522 for the content standards, and between 521 and 524 for the sub-content areas, and all are close to the median for the total test scale score, 521.

The mean percents of the maximum obtainable score for CS 2 (Write for a Variety of Purposes) and CS 3 (Write Using Conventions) are 66.6 and 63.7, respectively. The mean percent of the maximum obtainable score for the total test is 65.3. The mean percents of the maximum obtainable score for the sub-content areas range from 59.6 to 72.1. Note that the mean percents are closer for the content standards than for the sub-content areas.

Mathematics

The mean scale score for the total population of students taking the 2003 sixth-grade Mathematics assessment is 520 with a standard deviation of 76.3. The mean scale score for female students is 520 with a standard deviation of 72.1, and the mean scale score for male students is 520 with a standard deviation of 80.0.

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 slightly negatively skewed.

The mean scale scores for the content standards range from 518 to 529. The mean scale scores for sub-content areas range from 520 to 530. The median scale scores vary between 521 and 528 for the content standards, and between 520 and 525 for the sub-content areas, and all are close to the median for the total test scale score, 522.

The mean percents of the maximum obtainable score for the content standards range from 51.8 on CS 1 (Number Sense) to 67.8 on CS 2 (Algebra, Patterns and Functions). The mean percent of the maximum obtainable score for the total test is 58.0. The mean percents of the maximum obtainable score for the sub-content areas range from 55.4 to 67.8. The observed ranges indicate that the average item difficulty varied somewhat across the content standards and across the sub-content areas.

Seventh Grade

Reading

The mean scale score for the total population of students taking the 2003 seventh-grade Reading assessment is 632 with a standard deviation of 66.6. The mean scale score for female students is 641 with a standard deviation of 61.6, and the mean scale score for male students is 624 with a standard deviation of 70.1.

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 632 to 639. The mean scale scores for the sub-content areas range from 632 to 658. The median scale scores vary between 639 and 640 for the content standards, and between 639 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 58.4 on CS 6 (Literature) to 70.1 on CS 1 (Reading Comprehension). The mean percent of the maximum obtainable score for the total test is 65.1. The mean percents of the maximum obtainable score for the sub-content areas range from 57.4 to 74.3.

Writing

The mean scale score for the total population of students taking the 2003 seventh-grade Writing assessment is 543 with a standard deviation of 65.9. The mean scale score for female students is 560 with a standard deviation of 63.2, and the mean scale score for male students is 527 with a standard deviation of 64.6.

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 544 and 547. The mean scale scores for the sub-content areas range from 544 to 580. The median scale scores vary between 545 and 546 for the content standards, and between 544 and 554 for the sub-content areas, and most are close to the median for the total test scale score, 545.

The mean percents of the maximum obtainable score for CS 2 (Write for a Variety of Purposes) and CS 3 (Write Using Conventions) are 62.3 and 65.3, respectively. The mean percent of the maximum obtainable score for the total test is 63.5. The mean percents of the maximum obtainable score for the sub-content areas range from 60.0 to 72.0.

Mathematics

The mean scale score for the total population of students taking the 2003 seventh-grade Mathematics assessment is 542 with a standard deviation of 68.2. The mean scale score for female students is 544 with a standard deviation of 65.2. The mean scale score for male students is 541 with a standard deviation of 70.9.

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 slightly negatively skewed for the total population and for each gender.

The mean scale scores for the content standards range from 535 to 545. The mean scale scores for the sub-content areas range from 534 to 541. The median scale scores vary between 546 and 548 for the content standards, and between 546 and 548 for the sub-content areas, and all are close to the median for the total test scale score, 546.

The mean percents of the maximum obtainable score for the content standards range from 44.5 on CS 3 (Data Analysis, Probability, and Statistics) to 55.0 on CS 2 (Algebra, Patterns and Functions). The mean percent of the maximum obtainable score for the total test is 47.8. The mean percents of the maximum obtainable score for the sub-content areas range from 41.2 to 45.8.

Eighth Grade

Reading

The mean scale score for the total population of students taking the 2003 eighth-grade Reading assessment is 653 with a standard deviation of 63.7. The mean scale score for

female students is 662 with a standard deviation of 59.2, and the mean scale score for male students is 644 with a standard deviation of 66.6.

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 649 to 662. The mean scale scores for the sub-content areas range from 649 to 666. The median scale scores vary between 658 and 661 for the content standards, and between 659 and 661 for the sub-content areas, and all are close to the median for the total test scale score, 660.

The mean percents of the maximum obtainable score for the content standards range from 52.9 on CS 6 (Literature) to 72.5 on CS 4 (Thinking Skills). The mean percent of the maximum obtainable score for the total test is 62.6. The mean percents of the maximum obtainable score for the sub-content areas range from 50.8 to 75.9. The observed ranges indicate that the average item difficulty varies somewhat across the content standards and across the sub-content areas.

Writing

The mean scale score for the total population of students taking the 2003 eighth-grade Writing assessment is 554 with a standard deviation of 74.0. The mean scale score for female students is 573 with a standard deviation of 71.9, and the mean scale score for male students is 536 with a standard deviation of 71.3.

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 556 and 558. The mean scale scores for the sub-content areas range from 555 to 590. The median scale scores vary between 555 and 557 for the content standards, and between 554 and 592 for the sub-content areas, and most are close to the median for the total test scale score, 556. The median scale score for SA 6 (Extended Writing), 592, was a bit higher 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 66.4 and 66.9, respectively. The mean percent of the maximum obtainable score for the total test is 66.6. The mean percents of the maximum obtainable score for the sub-content areas range from 60.5 to 77.2.

Mathematics

The mean scale score for the total population of students taking the 2003 eighth-grade Mathematics assessment is 550 with a standard deviation of 73.8. The mean scale score for female students is 553 with a standard deviation of 68.3. The mean scale score for male students is 548 with a standard deviation of 78.8.

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 540 to 548. The mean scale scores for sub-content areas range from 539 to 545. The median scale scores vary between 557 and 560 for the content standards, and between 557 and 558 for the sub-content areas, and all are close to the median for the total test scale score, 558.

The mean percents of the maximum obtainable score for the content standards range from 35.9 on CS 1 (Number Sense) to 52.3 on CS 3 (Data Analysis, Probability, and Statistics). The mean percent of the maximum obtainable score for the total test is 43.7. The mean percents of the maximum obtainable score for the sub-content areas range from 32.1 to 42.0.

Science

The mean scale score for the total population of students taking the 2003 eighth-grade Science assessment is 500 with a standard deviation of 63.1. The mean scale score for female students is 498 with a standard deviation of 58.9, and the mean scale score for male students is 502 with a standard deviation of 66.9.

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 496 to 504. The mean scale scores for the sub-content areas range from 489 to 523. The median scale scores vary between 504 and 508 for the content standards, and between 503 and 508 for the sub-content areas, and all are close to the median for the total test scale score, 506.

The mean percents of the maximum obtainable score for the content standards range from 45.9 on CS 4 (Earth and Space Science) to 63.6 on CS 1 (Scientific Investigation). The mean percent of the obtainable score for the total test is 52.9. The mean percents of the maximum obtainable score for the sub-content areas range from 41.3 to 71.5. The observed ranges

indicate that the average item difficulty varies somewhat across the content standards and across the sub-content areas.

Ninth Grade

Reading

The mean scale score for the total population of students taking the 2003 ninth-grade Reading assessment is 658 with a standard deviation of 62.3. The mean scale score for female students is 670 with a standard deviation of 54.1, and the mean scale score for male students is 648 with a standard deviation of 67.7.

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 655 to 662. The mean scale scores for the sub-content areas range from 650 to 664. The median scale scores vary between 667 and 668 for the content standards, and between 667 and 668 for the sub-content areas, and all are close to the median for the total test scale score, 667.

The mean percents of the maximum obtainable score for the content standards range from 59.9 on CS 1 (Reading Comprehension) to 66.2 on CS 4 (Thinking Skills). The mean percent of the maximum obtainable score for the total test is 62.2. The mean percents of the maximum obtainable score for the sub-content areas range from 52.4 to 69.5.

Writing

The mean scale score for the total population of students taking the 2003 ninth-grade Writing assessment is 566 with a standard deviation of 78.2. The mean scale score for female students is 585 with a standard deviation of 74.7, and the mean scale score for male students is 549 with a standard deviation of 77.6.

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 567 and 569. The mean scale scores for sub-content areas range from 565 to 611. The median scale scores vary between 567 and 569 for the content standards, and between 566 and 593 for the sub-content areas, and most are close to the median for the total test scale score, 568. The median scale score for SA 6 (Extended Writing).

The mean percents of the maximum obtainable score for CS 2 (Write for a Variety of Purposes) and CS 3 (Write Using Conventions) are 69.0 and 63.1, respectively. The mean percent of the maximum obtainable score for the total test is 66.6. The mean percents of the maximum obtainable score for the sub-content areas range from 58.4 to 77.7.

Mathematics

The mean scale score for the total population of students taking the 2003 ninth-grade Mathematics assessment is 564 with a standard deviation of 73.2. The mean scale score for female students is 567 with a standard deviation of 67.2, and the mean scale score for male students is 561 with a standard deviation of 78.4.

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 553 to 564. The mean scale scores for the sub-content areas are 549 and 561. The median scale scores vary between 572 and 574 for the content standards, and between 575 and 576 for 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 31.8 on CS 1 (Number Sense) to 39.7 on CS 3 (Data Analysis, Probability and Statistics). The mean percent of the maximum obtainable score for the total test is 36.5. The mean percents of the maximum obtainable score for the sub-content areas range from 28.3 to 39.4.

Tenth Grade
Reading

The mean scale score for the total population of students taking the 2003 tenth-grade Reading assessment is 681 with a standard deviation of 62.0. The mean scale score for female students is 692 with a standard deviation of 53.4, and the mean scale score for male students is 671 with a standard deviation of 67.8.

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 679 to 690. The mean scale scores for the sub-content areas range from 678 to 700. The median scale scores vary between

689 and 692 for the content standards, and between 690 and 693 for the sub-content areas, and all are close to the median for the total test scale score, 690.

The mean percents of the maximum obtainable score for the content standards range from 56.9 on CS 5 (Use of Literary Information) to 71.5 on CS 4 (Thinking Skills). The mean percent of the maximum obtainable score for the total test is 63.9. The mean percents of the maximum obtainable score for the sub-content areas range from 51.0 to 70.4.

Writing

The mean scale score for the total population of students taking the 2003 tenth-grade Writing assessment is 584 with a standard deviation of 86.0. The mean scale score for female students is 603 with a standard deviation of 83.8, and the mean scale score for male students is 565 with a standard deviation of 83.7.

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 586 and 591. The mean scale scores for the sub-content areas range from 591 to 641. The median scale scores vary between 585 and 586 for the content standards, and between 583 and 590 for the sub-content areas, and all are close to the median for the total test scale score, 586.

The mean percents of the maximum obtainable score for CS 2 (Write for a Variety of Purposes) and CS 3 (Write Using Conventions) are 71.6 and 69.9, respectively. The mean percent of the maximum obtainable score for the total test is 71.0. The mean percents of the maximum obtainable score for the sub-content areas vary from 63.0 to 78.4.

Mathematics

The mean scale score for all students taking the 2003 tenth-grade Mathematics assessment is 582 with a standard deviation of 72.2. The mean scale score for female students is 582 with a standard deviation of 67.6, and the mean scale score for male students is 583 with a standard deviation of 76.5.

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 571 to 582. The mean scale scores for the sub-content areas are 578 and 586. The median scale scores vary between 588

and 591 for the content standards, and between 592 and 598 for the sub-content areas, and all are fairly close to the median for the total test scale score, 589.

The mean percents of the maximum obtainable score for the content standards range from 30.0 on CS 4/5 (Geometry and Measurement) to 43.0 on CS 3 (Data Analysis, Probability and Statistics). The mean percent of the maximum obtainable score for the total test is 38.7. The mean percents of the maximum obtainable score for SA 1 and SA 2 are 53.8 and 35.3, respectively.

The Correlations Between the Content Standards and Between the 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 .70 (grade 8) and .84 (grade 9) for the relationship between the various content standards, and between .58 (grade 10) and .82 (grade 7) for the relationship between the various sub-content areas, respectively.

For the Spanish Reading assessments, the correlations among the various content standards vary between .72 and .79; they vary between .65 and .79 among the various sub-content areas.

For the Writing assessments, the coefficients for the correlation between content standards 2 and 3 vary between .70 (grade 3) and .83 (grade 8). The correlations among the various sub-content areas vary between .44 (grade 4) and .76 (grades 5 and 8).

For the Spanish Writing assessment, the correlation between content standards 2 and 3 varies between .65 and .70; the correlations between the various sub-content areas vary between .41 and .63.

For the Mathematics assessments, the correlations vary between .71 (grade 6) and .82 (grades 7 and 8) for the relationship among the content standards, and between .59 (grade 10) and .77 (grade 5) for the relationship among the sub-content areas.

Finally, for the Science assessment, the correlation coefficients vary between .61 and .71 for the relationship among the content standards, and between .48 and .67 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 for each grade and content area. The alphas for Reading are .90 and .94 for grades 3 and 9 respectively, and .93 for grades 4 through 8 and 10. The alphas for grades 3 and 4 Spanish Reading are .87 and .90, respectively. The alphas for grades 3 through 10 Writing are .90, .91, .91, .92, .91, .93, .92, and .92, respectively. The alphas for grade 3 and 4 Spanish Writing are .90 and .87, respectively. The alphas for Mathematics are .94 for grades 5 through 8 and .93 for grades 9 and 10. The alpha for grade 8 Science is .92. These are all high reliabilities and indicate that the Colorado 2003 assessments had strong internal consistency and that the tests produced relatively stable scores.

Tables 16 and 17 show the estimated reliability index (Cronbach's alpha) for each of the content standards and sub-content areas, respectively, for each grade and content area for the 2003 assessments.

For Reading grades 4 through 10, the alphas for the various content standards vary between .68 (for CS 5 in grade 8) and .86 (for CS 1 in grade 4), with a median of .765. For Reading grades 3 through 10, the alphas for the various sub-content areas vary between .56 (for SA 4 in grade 10) and .88 (for SA 1 in grade 4), with a median of .75.

For the Spanish version of the Reading assessment for grade 4, the alphas for the various content standards vary between .56 and .81 (for CS 6 and 1, respectively), with a median of .67. For the Spanish versions of the Reading assessment for grades 3 and 4, the alphas for the various sub-content areas vary between .43 (for SA 3 in grade 4) and .83 (for SA 1 in grade 4), with a median of .675.

For Writing grades 3 through 10, the alphas for the various content standards vary between .73 (for CS 2 in grade 3) and .88 (for CS 3 in grade 3 and CS 2 in grades 9 and 10) with a median of .855. The alphas for the various sub-content areas vary between .65 and .84 (for SA 8, in grades 10 and SA 7in grade 8, respectively), with a median of .74.

For the Spanish version of the Writing assessment for grade 3 and 4, the alphas for the various content standards vary between .70 (for CS 2 in grade 4) and .85 (for CS 3 in grade 3) with a median of .82. The alphas for the various sub-content areas vary between .55 (for SA 6 in grade 4) and .84 (for SA 5 in grade 3), with a median of .72.

For Mathematics grades 5 through 10, the alphas for the various content standards vary between .70 and .83 (for CS 2 in grade 6 and CS 1 in grade 7, respectively) with a median of .755. The alphas for the various sub-content areas vary between .51 and .81 (for SA 2 in grade 10 and SA 1 in grade 6, respectively), with a median of .69.

For grade 8 Science, the alphas for the various content standards vary between .58 and .81 (for CS 5 and 1, respectively) with a median of .71. The alphas for the various sub-content areas vary between .49 and .70 (for SA 4 and 5, respectively), with a median of .65.

The alpha internal consistency coefficients generally are lower for the various sub-test scores than they are for the total test scores. This is not that surprising, given that the scores for the content standards and the sub-content areas are based on fewer items than those for the total test scores.

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 2003 third-grade Reading assessment. The point-biserials for all multiple-choice items range from .20 to .60 with a mean of .44. The p-values for these items range from .29 to .97 with a mean of .80.

Table 19 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .24 to .57 with a mean of .46. Their p-values range from .34 to .88 with a mean of .67. 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.

For two of the seven constructed-response items, over 70% of the students obtained the highest possible score points. The scores of the remaining students were fairly well distributed across the remaining score points.

For grade 8 Science, the alphas for the various content standards vary between .58 and .81 (for CS 5 and 1, respectively) with a median of .71. The alphas for the various sub-content areas vary between .49 and .70 (for SA 4 and 5, respectively), with a median of .65.

The alpha internal consistency coefficients generally are lower for the various sub-test scores than they are for the total test scores. This is not that surprising, given that the scores for the content standards and the sub-content areas are based on fewer items than those for the total test scores.

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 2003 third-grade Reading assessment. The point-biserials for all multiple-choice items range from .20 to .60 with a mean of .44. The p-values for these items range from .29 to .97 with a mean of .80.

Table 19 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .24 to .57 with a mean of .46. Their p-values range from .34 to .88 with a mean of .67. 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.

For two of the seven constructed-response items, over 70% of the students obtained the highest possible score points. The scores of the remaining students were fairly well distributed across the remaining score points.

The omit rate for the third-grade Reading assessment was small, ranging from .0% to 1.6% for multiple-choice items (Table 18) and .4% to .9% 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 2003 third-grade Reading assessment. The point-biserials for all multiple-choice items range from .18 to .48 with a mean of .35. The p-values for these items range from .26 to .93 with a mean of .70.

Table 21 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .31 to .51 with a mean of .42. Their p-values range from .29 to .87, with a mean of .59. 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 3.4% for multiple-choice items (Table 20) and .3% to 2.9% for constructed-response items (Table 21).

Writing

Table 22 lists the results of the multiple-choice item analyses for the 2003 third-grade Writing assessment. The point-biserials for all multiple-choice items range from .24 to .51 with a mean of .39. The p-values for these items range from .59 to .98 with a mean of .86.

Table 23 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .23 to .51 with a mean of .39. Their p-values range from .54 to .99, with a mean of .83. 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 2.2% for multiple-choice items (Table 22) and from .1% to .3% 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 2003 third-grade Writing assessment. The point-biserials for all multiple-choice items range from .10 to .46 with a mean of .34. The p-values for these items range from .25 to .95 with a mean of .73.

Table 25 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .20 to .54 with a mean of .43. Their p-values range from .33 to .95, with a mean of .67. 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 1.6% for multiple-choice items (Table 24) and .0% to .2% for constructed-response items (Table 25), with only one item having a 5% omit rate.

Fourth Grade

Reading

Table 26 lists the results of the multiple-choice item analyses for the 2003 fourth-grade Reading assessment. The point-biserials for the multiple-choice items range from .14 to .56 with a mean of .43. The p-values for the multiple-choice items range from .43 to .94 with a mean of .71.

Table 27 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .32 to .60 with a mean of .44. Their p-values range from .20 to .86 with a mean of .55. 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 reasonable, ranging from .0% to 3.2% for multiple-choice items (Table 26). The range was .3% to 3.6% 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 2003 fourth-grade Reading assessment. When item 107 is not taken into account, the point-biserials for all multiple-choice items range from .09 to .51 with a mean of .33. The p-values for these items range from .14 to .91 with a mean of .57. MC item 107 did not discriminate between low or 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 .16 to .62 with a mean of .44. Their p-values range from .15 to .79, with a mean of .45. 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 .0% to 2.4% for multiple-choice items (Table 28) and .1% to 3.1% for constructed-response items (Table 29).

Writing

Table 30 lists the results of the multiple-choice item analyses for the 2003 fourth-grade Writing assessment. The point-biserials for all multiple-choice items range from .22 to .48 with a mean of .37. The p-values for these items range from .44 to .99 with a mean of .76.

Table 31 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .05 to .59 with a mean of .42. Their p-values range from .29 to 1.00, with a mean of .65.

The omit rate for the fourth-grade Writing assessment was small, ranging from .1% to 1.8% 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 2003 fourth-grade Writing assessment. When item 54 is not taken into account, the point-biserials for all multiple-choice items range from .13 to .47 with a mean of .31. The p-values for these items range from .26 to .97 with a mean of .59. MC item 54 did not discriminate between low or 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 33 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed response items range from .12 to .59 with a mean of .37. Their p-values range from .03 to .99 with a mean of .49.

The omit rate for the fourth-grade Writing assessment was small, ranging from .0% to 1.1% for multiple-choice items (Table 32) and .0% to 3.4% for constructed-response items (Table 33).

Fifth Grade

Reading

Table 34 lists the results of the multiple-choice item analyses for the 2003 fifth-grade Reading assessment. The point-biserials for the multiple-choice items range from .19 to .52, with a mean of .40. The p-values for the multiple-choice items range from .36 to .95 with a mean of .72.

Table 35 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .33 to .59 with a mean of .48. Their p-values range from .33 to .85 with a mean of .54. 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 2 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 .0% to 2.5% for multiple-choice items (Table 34) and .5% to 5.9% for constructed-response items (Table 35), with only one constructed-response item having an omit rate greater than 5%.

Writing

Table 36 lists the results of the multiple-choice item analyses for the 2003 fifth-grade Writing assessment. The point-biserials for all multiple-choice items range from .13 to .55 with a mean of .38. The p-values for these items range from .26 to .95 with a mean of .70.

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 .46. Their p-values range from .51 to 1.00, with a mean of .73.

The omit rate for the fifth-grade Writing assessment was small, ranging from .1% to 3.8% for multiple-choice items (Table 36) and .0% to 3.8% for constructed-response items (Table 37).

Mathematics

Table 38 lists the results of the multiple-choice item analyses for the 2003 fifth-grade Mathematics assessment. The point-biserials for the multiple-choice items range from .20 to .60, with a mean of .40. The p-values for the multiple-choice items range from .26 to .97 with a mean of .71.

Table 39 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .44 to .68 with a mean of .58. Their p-values range from .41 to .81 with a mean of .64. The distribution of the percent of students obtaining each score point for the constructed-response items shows that there is a good amount of variability in students' responses to most items, indicating that these items work well over the range of student ability.

The omit rate for the fifth-grade Mathematics assessment was reasonable, ranging from .1% to 3.6% for multiple-choice items (Table 38) and .2% to 5.1% for constructed-response items (Table 39), with only one item having an omit rate greater than 5%.

Sixth Grade

Reading

Table 40 lists the results of the multiple-choice item analyses for the 2003 sixth-grade Reading assessment. The point-biserials for the multiple-choice items range from .18 to .59 with a mean of .39. The p-values for the multiple-choice items range from .30 to .89 with a mean of .64.

Table 41 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .32 to .61 with a mean of .51. Their p-values range from .37 to .90 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 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 Reading assessment was small, ranging from .1% to 3.7% for multiple-choice items (Table 40) and .5% to 8.1% for constructed-response items, with only three items having an omit rate greater than 5% (Table 41).

Writing

Table 42 lists the results of the multiple-choice item analyses for the 2003 sixth-grade Writing assessment. The point-biserials for all multiple-choice items range from .16 to .57 with a mean of .40. The p-values for these items range from .27 to .90 with a mean of .68.

Table 43 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .07 to .60 with a mean of .41. Their p-values range from .11 to .99 with a mean of .62.

The omit rate for the sixth-grade Writing assessment was small, ranging from .1% to 5.0% for multiple-choice items (Table 42) and .0% to 7.2% for constructed-response items (Table 43) , with only two items having an omit rate greater than 5% (Tables 42 and 43).

Mathematics

Table 44 lists the results of the multiple-choice item analyses for the 2003 sixth-grade Mathematics assessment. The point-biserials for the multiple-choice items range from .20 to .54, with a mean of .38. The p-values for the multiple-choice items range from .25 to .91 with a mean of .62.

Table 45 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .40 to .71 with a mean of .58. Their p-values range from .11 to .79 with a mean of .52. 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 sixth-grade Mathematics assessment was reasonable, ranging from .1% to 5.8% for multiple-choice items (Table 44) and .2% to 1.6% for constructed-response items (Table 45), with only one item having an omit rate greater than 5%.

Seventh Grade
Reading

Table 46 lists the results of the multiple-choice item analyses for the 2003 seventh-grade Reading assessment. The point-biserials for the multiple-choice items are positive, ranging from .10 to .58 with a mean of .39. The p-values for the multiple-choice items range from .39 to .92 with a mean of .67.

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 .32 to .58 with a mean of .48. The p-values for the constructed-response items range from .27 to .83 with a mean of .61. 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 7 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 2003 grade 7 Reading assessment ranged from .1% to 2.4% (Table 46). The percent of students who omitted constructed-response items ranged from .8% to 6.2% (Table 47), with only one item having an omit rate equal or greater than 5%.

Writing

Table 48 lists the results of the multiple-choice item analyses for the 2003 seventh-grade Writing assessment. The point-biserials for all multiple-choice items range from .01 to .50 with a mean of .36. The p-values for these items range from .07 to .94 with a mean of .66.

Table 49 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .10 to .62 with a mean of .47. Their p-values range from .28 to .99, with a mean of .57.

The omit rate for the seventh-grade Writing assessment was small, ranging from .1% to 2.0% for multiple-choice items (Table 48) and .0% to 3.9% for constructed-response items (Table 49).

Mathematics

Table 50 lists the results of the multiple-choice item analyses for the 2003 seventh-grade Mathematics assessment. The point-biserials for the multiple-choice items range from .17 to .60, with a mean of .39. The p-values for the multiple-choice items range from .23 to .96 with a mean of .57.

Table 51 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 .58. Their p-values range from .15 to .66 with a mean of .36. 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 reasonable, ranging from .1% to 2.3% for multiple-choice items (Table 50) and .5% to 3.4% for constructed-response items (Table 51).

Eighth Grade

Reading

Table 52 lists the results of the multiple-choice item analyses for the 2003 eighth-grade Reading assessment. When item 117 is not taken into account, the point-biserials for the multiple-choice items range from .11 to .54 with a mean of .37. The p-values for the multiple-choice items range from .15 to .93 with a mean of .66. MC item 117 did not discriminate between low or 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 53 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .34 to .63 with a mean of .51. Their p-values range from .21 to .84 with a mean of .57. 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 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 percent of students who omitted the multiple-choice items in the eighth-grade Reading assessment ranged from .1% to 1.1% (Table 52). The percent of students who omitted the constructed-response items ranged from .9% to 7.7% (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 2003 eighth-grade Writing assessment. The point-biserials for all multiple-choice items range from .03 to .58 with a mean of .42. The p-values for these items range from .08 to .90 with a mean of .68.

Table 55 lists the results of the constructed-response item analyses. When item 2 A is not taken into account, the item-to-total-score correlations for the constructed-response items range from .15 to .65 with a mean of .47. Their p-values range from .14 to .99, with a mean of .65. CR item 2 A did not discriminate between low or 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.

The omit rate for the eighth-grade Writing assessment was small, ranging from .1% to 3.6% for multiple-choice items (Table 54) and .0% to 2.0% for constructed-response items (Table 54).

Mathematics

Table 56 lists the results of the multiple-choice item analyses for the 2003 eighth-grade Mathematics assessment. When item 15 is not taken into account, the point-biserials for the multiple-choice items range from .13 to .56 with a mean of .37. The p-values for these multiple-choice items range from .10 to .88 with a mean of .50. Multiple-choice item 15 did not discriminate between low or 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 57 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .44 to .71 with a mean of .60. Their p-values range from .07 to .59 with a mean of .36. 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 2.2% (Table 56). The percent of students who omitted constructed-response items ranged from .8% to 8.2%, with only one constructed-response item having an omit rate greater than 5% (Table 57).

Science

Table 58 lists the results of the multiple-choice item analyses for the 2003 eighth-grade Science assessment. The point-biserials for the multiple-choice items range from .00 to .49 with a mean of .31. The p-values for the multiple-choice items range from .14 to .94 with a mean of .58.

Table 59 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .30 to .62 with a mean of .46. Their p-values range from .08 to .85 with a mean of .44. The percent of students obtaining each score point for the constructed-response items shows a good amount of variability in students' responses to most items, indicating that these items work well over the range of student ability.

The omit rate for the multiple-choice items for the eighth-grade Science assessment ranged from .1% to 3.7% (Table 58). The omit rate for the constructed-response items ranged from .9% to 9.6% (Table 59), with six 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 2003 ninth-grade Reading assessment. The point-biserials for the multiple-choice items range from .14 to .54 with a mean of .38. The p-values for the multiple-choice items range from .31 to .90 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 .38 to .62 with a mean of .55. Their p-values range from .29 to .76 with a mean of .56. 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 4 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 multiple-choice items for the ninth-grade Reading assessment ranged from .1% to 2.0% (Table 60).   The omit rate for the constructed-response items ranged from 1.7% to 10.1%, with 13 out of the 15 items having an omit rate greater than 5% (Table 61).

Writing

Table 62 lists the results of the multiple-choice item analyses for the 2003 ninth-grade Writing assessment. The point-biserials for all multiple-choice items range from .22 to .55 with a mean of .41. The p-values for these items range from .28 to .91 with a mean of .65.

Table 63 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .16 to .65 with a mean of .44. Their p-values range from .25 to .99, with a mean of .68.

The omit rate for the ninth-grade Writing assessment was small, ranging from .1% to 4.4% for multiple-choice items (Table 62) and .0% to 3.6% for constructed-response items (Table 63).

Mathematics

Table 64 lists the results of the multiple-choice item analyses for the 2003 ninth-grade Mathematics assessment. When item 51 is not taken into account, the point-biserials for the multiple-choice items range from .02 to .52 with a mean of .34. The p-values for these multiple-choice items range from .10 to .87 with a mean of .48. MC item 51 did not discriminate between low or 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 65 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .42 to .71 with a mean of .58. Their p-values range from .04 to .48 with a mean of .24. 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 4.6% (Table 64). The percent of students who omitted constructed-response items ranged from .8% to 8.0% with 3 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 2003 tenth-grade Reading assessment. The point-biserials for the multiple-choice items range from .11 to .58 with a mean of .37. The p-values for the multiple-choice items range from .35 to .90 with a mean of .70.

Table 67 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .43 to .63 with a mean of .54. Their p-values range from .31 to .76 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 5 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 2003 tenth-grade Reading assessment ranged from .1% to 2.5% (Table 66). Omit rates for the constructed-response items ranged from 2.3% to 14.8% (Table 67), with 9 out of the 13 items having an omit rate greater than 5%.

Writing

Table 68 lists the results of the multiple-choice item analyses for the 2003 tenth-grade Writing assessment. The point-biserials for all multiple-choice items range from .20 to .56 with a mean of .41. The p-values for these items range from .46 to .90 with a mean of .73.

Table 69 lists the results of the constructed-response item analyses. The item-to-total-score correlations for the constructed-response items range from .09 to .66 with a mean of .46. Their p-values range from .32 to .98, with a mean of .68.

The omit rate for the tenth-grade Writing assessment was small, ranging from .1% to .9% for multiple-choice items (Table 68) and .0% to 3.7% for constructed-response items (Table 69).

Mathematics

Table 70 lists the results of the multiple-choice item analyses for the 2003 tenth-grade Mathematics assessment. When items 53 and 57 are not taken into account, the point-biserials for the multiple-choice items range from .05 to .51 with a mean of .34. The p-values for the multiple-choice items range from .19 to .83 with a mean of .49. MC items 53 and 57 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.

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 .75 with a mean of .61. The p-values for the constructed-response items range from .07 to .68 with a mean of .28. The percent of students obtaining each score point for the constructed-response items shows a good amount of variability in students' responses to most items.

The omit rate for the multiple-choice items for the tenth-grade Mathematics assessment ranged from .1% to 1.3% (Table 70). The omit rate for the constructed-response items ranged from .4% to 5.9% (Table 71), with 3 out of the 15 items having an omit rate greater than 5%.

Part 5: Scaling and Calibration

Overview of the IRT Models

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

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

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

Mathematics

Table 70 lists the results of the multiple-choice item analyses for the 2003 tenth-grade Mathematics assessment. When items 53 and 57 are not taken into account, the point-biserials for the multiple-choice items range from .05 to .51 with a mean of .34. The p-values for the multiple-choice items range from .19 to .83 with a mean of .49. MC items 53 and 57 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.

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 .75 with a mean of .61. The p-values for the constructed-response items range from .07 to .68 with a mean of .28. The percent of students obtaining each score point for the constructed-response items shows a good amount of variability in students' responses to most items.

The omit rate for the multiple-choice items for the tenth-grade Mathematics assessment ranged from .1% to 1.3% (Table 70). The omit rate for the constructed-response items ranged from .4% to 5.9% (Table 71), with 3 out of the 15 items having an omit rate greater than 5%.

Part 5: Scaling and Calibration

Overview of the IRT Models

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

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

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

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

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

where m_j is the number of score levels and

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

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

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

and

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

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

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

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

Reading

The critical Z-value for the third-grade Reading assessment, based on the calibration sample, is 131.5 The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 72 and 73, respectively.

The Z-statistics for all MC and most CR items are small compared to the critical Z-value, indicating that the proportions of students obtaining each item score are close to the proportions that would be predicted by the estimated thetas and item parameters.

CR item 16 in Session 1 exhibited less than optimal fit with a Z-statistic of 213.7, exceeding the critical Z-value of 131.5. A close examination of the ICCs for the item indicated that the

proportion of lower ability students obtaining a "0" was lower than expected and obtaining a "1" was higher than expected. The proportion of close to average ability students obtaining a "0" was slightly higher than expected.

Reading – Spanish Version

The critical Z-value for the <u>third-grade Spanish Reading</u> assessment, based on the calibration sample, is 4.4 The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 74 and 75, respectively.

The Z-statistics for most MC and most CR items are small compared to the critical Z-value, indicating that the proportions of students obtaining each item score are close to the proportions that would be predicted by the estimated thetas and item parameters.

MC item 21 in Session 2 exhibited less than optimal fit with a Z-statistic of 4.6, marginally exceeding the critical Z-value of 4.4. A close examination of the ICC for the item indicated that the proportion of lower ability students obtaining a "1" was higher than expected.

CR item 27 in Session 2 exhibited less than optimal fit with a Z-statistic of 5.8, exceeding the critical Z-value of 4.4. A close examination of the ICCs for the item indicated that the proportion of higher ability students obtaining a "1" was higher than expected.

Writing

The critical Z-value for the <u>third-grade Writing</u> assessment, based on the calibration sample, is 24.5. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 76 and 77, respectively.

The Z-statistics for all MC and CR items are small compared to the critical Z-value, indicating that the proportions of students obtaining each item score are close to the proportions that would be predicted by the estimated thetas and item parameters.

Writing – Spanish Version

The critical Z-value for the <u>third-grade Spanish Writing</u> assessment, based on the calibration sample, is 3.9 The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 78 and 79, respectively.

The Z-statistics for most MC and most CR items are small compared to the critical Z-value, indicating that the proportions of students obtaining each item score are close to the proportions that would be predicted by the estimated thetas and item parameters.

CR item 7E in Session 1 exhibited less than optimal fit with a Z-statistic of 5.5, exceeding the critical Z-value of 3.9. A close examination of the ICCs for the item indicated that the

observed and expected proportion of students obtaining a "1" over the ability continuum were sometimes slightly off; however, there was no clear pattern in these differences.

MC item 8 in Session 12 exhibited less than optimal fit with a Z-statistic of 6.5, exceeding the critical Z-value of 3.9. A close examination of the ICC for the item indicated that the observed and expected proportion of students obtaining a "1" over the ability continuum were sometimes off; however, there was no clear pattern in these differences.

CR item 39A in Session 2 exhibited less than optimal fit with a Z-statistic of 4.2, exceeding the critical Z-value of 3.9. A close examination of the ICCs for the item indicated that the observed and expected proportion of students obtaining the various possible scores over the ability continuum were sometimes off; however, there was no clear pattern in these differences.

Fourth Grade

Reading

The critical Z-value for the fourth-grade Reading assessment, based on the calibration sample, is 27.5. Fit values above this critical value may indicate poor model fit. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 80 and 81, respectively.

The Z-statistics for most CR and MC items are smaller than the critical Z-value, indicating that the proportions of students with each item score are close to the proportions that would be predicted by the estimated thetas and item parameters.

MC item 102 in Session 6 exhibited less than optimal fit with a Z-statistic of 29.8, exceeding the critical Z-value of 27.5. A close examination of the ICC for the item indicated that the proportion of lower ability students obtaining a "1" was lower than expected and the proportion of slightly-below-average-ability students obtaining a "1" was slightly higher than expected.

CR item 38 in Session 4 exhibited less than optimal fit with a Z-statistic of 39.2, exceeding the critical Z-value of 27.5. A close examination of the ICC for the item indicated that the proportion of lower ability students obtaining a "1" was higher than expected and obtaining a "0" was lower than expected.

Reading – Spanish Version

The critical Z-value for the fourth-grade Spanish Reading assessment, based on the calibration sample, is 1.9. Fit values above this critical value may indicate poor model fit. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 82 and 83, respectively.

The Z-statistics for most CR and MC items are smaller than the critical Z-value, indicating that the proportions of students with each item score are close to the proportions that would be predicted by the estimated thetas and item parameters.

CR item 9 in Session 3 exhibited less than optimal fit with a Z-statistic of 2.5, exceeding the critical Z-value of 1.9. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was slightly lower than expected and obtaining a "1" was higher than expected.

MC item 26 in Session 4 exhibited less than optimal fit with a Z-statistic of 2.7, exceeding the critical Z-value of 1.9. A close examination of the ICC for the item indicated that the proportion of average ability students obtaining a "1" was slightly higher than expected.

MC item 31 in Session 4, exhibited less than optimal fit with a Z-statistic of 2.3, exceeding the critical Z-value of 1.9. A close examination of the ICC for the item indicated that the observed and expected proportion of students obtaining a "1" over the ability continuum were sometimes slightly off; however, there was no clear pattern in these differences.

MC item 93 in Session 6 exhibited less than optimal fit with a Z-statistic of 3.3, exceeding the critical Z-value of 1.9. A close examination of the ICC for the item indicated that the observed and expected proportion of students obtaining a "1" over the ability continuum were sometimes slightly off; however, there was no clear pattern in these differences.

MC item 95 in Session 6 exhibited less than optimal fit with a Z-statistic of 2.0, marginally exceeding the critical Z-value of 1.9. A close examination of the ICC for the item indicated that the observed and expected proportion of students obtaining a "1" over the ability continuum were sometimes slightly off; however, there was no clear pattern in these differences.

The item parameter estimation process did not converge for MC item 107 in Session 6. A close examination of the ICC for the item indicated that the student responses do not fit the model well. The item did not discriminate between low or 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.

Writing

The critical Z-value for the fourth-grade Writing assessment, based on the calibration sample, is 26.6. Fit values above this critical value may indicate poor model fit. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 84 and 85, respectively.

The Z-statistics for most MC items and most CR items are small compared to the critical Z-value indicating that the proportions of students with each item score are close to the proportions that would be predicted by the estimated thetas and item parameters.

CR item 3A in Session 2 exhibited less than optimal fit with a Z-statistic of 190.3, exceeding the critical Z-value of 26.6. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was less than expected and obtaining a "1" was higher than expected. The proportion of higher ability students obtaining a "3" was also less than expected and obtaining a "4" was higher than expected.

MC item 79 in Session 5 exhibited less than optimal fit with a Z-statistic of 31.2, exceeding the critical Z-value of 26.6. A close examination of the ICC for the item indicated that the proportion of low ability students obtaining a "1" was lower than expected.

Writing – Spanish Version

The critical Z-value for the <u>fourth-grade Spanish Writing</u> assessment, based on the calibration sample, is 1.9. Fit values above this critical value may indicate poor model fit. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 86 and 87, respectively.

The Z-statistics for most MC and CR items are small compared to the critical Z-value, indicating that the proportions of students with each item score are close to the proportions that would be predicted by the estimated thetas and item parameters.

CR item 2C in Session 2 exhibited less than optimal fit with a Z-statistic of 2.4, exceeding the critical Z-value of 1.9. A close examination of the ICCs for the item indicated that the observed and expected proportion of students obtaining a "1" over the ability continuum were sometimes slightly off; however, there was no clear pattern in these differences.

CR item 3A in Session 2 exhibited less than optimal fit with a Z-statistic of 2.2, exceeding the critical Z-value of 1.9. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "2" was slightly higher than expected.

MC items 62, 69, 76, 78, 79, and 86 in Session 5 exhibited less than optimal fit with Z-statistics of 2.1, 1.9, 3.7, 2.7, 2.0, and 2.4, respectively, all exceeding the critical Z-value of 1.9. A close examination of the ICCs for the items indicated that the observed and expected proportion of students obtaining a "1" over the ability continuum were sometimes slightly off; however, there was no clear pattern in these differences.

The item parameter estimation process did not converge for MC item 54 in Session 5. A close examination of the ICC for the item indicated that the student responses do not fit the model well. The item did not discriminate between low or 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.

Fifth Grade
Reading

The critical Z-value for the <u>fifth-grade Reading</u> assessment, based on the calibration sample, is 27.8. Fit values above this critical value may indicate poor model fit. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 88 and 89, respectively.

The Z-statistics for all MC and CR items are smaller than the critical Z-value, indicating that the proportions of students with each item score are close to the proportions that would be predicted by the estimated thetas and item parameters.

Writing

The critical Z-value for the <u>fifth-grade Writing</u> assessment, based on the calibration sample, is 26.7. Fit values above this critical value may indicate poor model fit. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 90 and 91, respectively.

The Z-statistics for all MC items and most CR items are smaller than the critical Z-value, indicating that the proportions of students with each item score are close to the proportions that would be predicted by the estimated thetas and item parameters.

CR item 3A in Session 2 exhibited less than optimal fit with a Z-statistic of 170.0, exceeding the critical Z-value of 26.7. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was less than expected and obtaining a "1" was higher than expected. The proportion of higher ability students obtaining a "3" was also less than expected and obtaining a "4" was higher than expected.

CR item 3B in Session 2 exhibited less than optimal fit with a Z-statistic of 27.1, marginally exceeding the critical Z-value of 26.7. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was slightly less than expected and obtaining a "1" was higher than expected.

CR item 3C in Session 2 exhibited less than optimal fit with a Z-statistic of 68.3, exceeding the critical Z-value of 26.7. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was less than expected and obtaining a "1" was higher than expected.

CR item 27 in Session 3 exhibited less than optimal fit with a Z-statistic of 54.7, exceeding the critical Z-value of 26.7. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was less than expected and obtaining a "1" was higher than expected.

Mathematics

The critical Z-value for the <u>fifth-grade Mathematics</u> assessment, based on the calibration sample, is 26.8. Fit values above this critical value may indicate poor model fit. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 92 and 93, respectively.

The Z-statistics for all MC and CR items are smaller than the critical Z-value, indicating that the proportions of students with each item score are close to the proportions that would be predicted by the estimated thetas and item parameters.

Sixth Grade
Reading

The critical Z-value for the <u>sixth-grade Reading</u> assessment, based on the calibration sample, is 25.3. Fit values above this critical value may indicate poor model fit. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 94 and 95, respectively.

The Z-statistics for most MC items and most CR items are smaller than the critical Z-value, indicating that the proportions of students with each item score are close to the proportions that would be predicted by the estimated thetas and item parameters.

CR item 45 in Session 4 exhibited less than optimal fit with a Z-value of 42.6, exceeding the critical Z-value of 25.3. A detailed examination of the ICC for item 45 indicated that the proportion of lower ability students obtaining a "0" was higher than expected and obtaining a "1" was lower than expected.

MC item 105 in Session 6 exhibited less than optimal fit with a Z-statistic of 40.0, exceeding the critical Z-value of 25.3. A close examination of the ICC for the item indicated that the proportion of slightly below average ability students obtaining a "1" was higher than expected.

MC item 111 in Session 6 exhibited less than optimal fit with a Z-statistic of 30.4, exceeding the critical Z-value of 25.3. A close examination of the ICC for the item indicated that the proportion of lower ability students obtaining a "1" was higher than expected.

Writing

The critical Z-value for the <u>sixth-grade Writing</u> assessment, based on the calibration sample, is 25.1. Fit values above this critical value may indicate poor model fit. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 96 and 97, respectively.

The Z-statistics for all MC items and most CR items are smaller than the critical Z-value, indicating that the proportions of students with each item score are close to the proportions that would be predicted by the estimated thetas and item parameters.

CR item 3A in Session 2 exhibited less than optimal fit with a Z-statistic of 104.8, exceeding the critical Z-value of 25.1. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was less than expected and obtaining a "1" was higher than expected. The proportion of higher ability students obtaining a "3" was also less than expected and obtaining a "4" was higher than expected.

CR item 50 in Session 4 exhibited less than optimal fit with a Z-statistic of 36.8, exceeding the critical Z-value of 25.1. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was less than expected and obtaining a "1" was higher than expected.

CR item 70 in Session 5 exhibited less than optimal fit with a Z-statistic of 36.8, exceeding the critical Z-value of 25.1. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was less than expected and obtaining a "1" was higher than expected.

Mathematics

The critical Z-value for the sixth-grade Mathematics assessment, based on the calibration sample, is 26.5. Fit values above this critical value may indicate poor model fit. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 98 and 99, respectively.

The Z-statistics for all MC items and most CR items are smaller than the critical Z-value, indicating that the proportions of students with each item score are close to the proportions that would be predicted by the estimated thetas and item parameters.

CR item 12 in Session 1 exhibited less than optimal fit with a Z-statistic of 30.4, exceeding the critical Z-value of 26.5. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was lower than expected and obtaining a "1" was higher than expected. The proportion of higher ability students obtaining a score of "1" was lower than expected.

CR item 49 in Session 3 exhibited less than optimal fit with a Z-statistic of 157.2, exceeding the critical Z-value of 26.5. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was lower than expected and obtaining a "1" was higher than expected. The proportion of higher ability students obtaining a score of "1" was lower than expected and obtaining a "2" was higher than expected.

Seventh Grade

Reading

The critical Z-value for the <u>seventh-grade Reading</u> assessment, based on the calibration sample, is 24.9. Fit values above this critical value may indicate poor model fit. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 100 and 101, respectively.

The item responses fit the IRT model reasonably well. The Z-statistics for most MC items and most CR items are small compared to the critical Z-value, indicating that the proportions of students with each item score are close to the proportions that would be expected by the estimated thetas and item parameters.

MC item 28 in Session 3 exhibited less than optimal fit with a Z-value of 46.7, exceeding the critical Z-value of 24.9. A close examination of the ICC for the item indicated that the proportion of below average ability students obtaining a "1" was lower than expected, and the proportion of about average ability students obtaining a "1" higher than expected.

CR item 37 in Session 4 exhibited less than optimal fit with a Z-value of 33.0, exceeding the critical Z-value of 24.9. A detailed examination of the ICC indicates that the proportion of lower ability students obtaining a "1" was lower than expected and obtaining a "0" was higher than expected.

MC item 44 in Session 4 exhibited less than optimal fit with a Z-value of 35.6, exceeding the critical Z-value of 24.9. A close examination of the ICC for the item indicated that the proportion of lower ability students obtaining a "1" was higher than expected.

Writing

The critical Z-value for the <u>seventh-grade Writing</u> assessment, based on the calibration sample, is 25.4. Fit values above this critical value may indicate poor model fit. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 102 and 103, respectively.

The Z-statistics for all MC items and most CR items are smaller than the critical Z-value, indicating that the proportions of students with each item score are close to the proportions that would be predicted by the estimated thetas and item parameters.

CR item 3A in Session 2 exhibited less than optimal fit with a Z-statistic of 64.4, exceeding the critical Z-value of 25.4. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was less than expected and obtaining a "1" was higher than expected. The proportion of higher ability students obtaining a "3" was also less than expected and obtaining a "4" was higher than expected.

CR item 30 in Session 3 exhibited less than optimal fit with a Z-statistic of 73.6, exceeding the critical Z-value of 25.4. A close examination of the ICCs for the item indicated that the

proportion of lower ability students obtaining a "0" was less than expected and obtaining a "1" was higher than expected. The proportion of higher ability students obtaining a "3" was also less than expected and obtaining a "4" was higher than expected.

Mathematics

The critical Z-value for the seventh-grade Mathematics assessment, based on the calibration sample, is 28.6. Fit values above this critical value may indicate poor model fit. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 104 and 105, respectively.

The Z-statistics for most MC and most CR items are smaller than the critical Z-value, indicating that the proportions of students with each item score are close to the proportions that would be predicted by the estimated thetas and item parameters.

MC item 25 in Session 2 exhibited less than optimal fit with a Z-value of 53.7, exceeding the critical Z-value of 28.6. A close examination of the ICC for the item indicated that the proportion of lower ability students obtaining a "1" was higher than expected.

CR item 43 in Session 3 exhibited less than optimal fit with a Z-value of 37.1, exceeding the critical Z-value of 28.6. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "1" was slightly higher than expected and the proportion of higher ability students obtaining a "1" was slightly lower than expected.

CR item 53 in Session 3 exhibited less than optimal fit with a Z-value of 135.8, exceeding the critical Z-value of 28.6. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was lower than expected and obtaining a "1" was higher than expected. The proportion of higher ability students obtaining a "1" was lower than expected and obtaining a "2" was higher than expected.

Eighth Grade

Reading

The critical Z-value for the eighth-grade Reading assessment, based on the calibration sample, is 24.5. Fit values above this critical value may indicate poor model fit. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 106 and 107, respectively. They indicate that the item responses fit the model reasonably well.

The Z-statistics for all CR items and most MC items are small compared to the critical Z-value, indicating that the proportions of students with each item score are close to the proportions that would be expected by the estimated thetas and item parameters.

MC item 31 in Session 4 exhibited less than optimal fit with a Z-value of 35.1, exceeding the critical Z-value of 24.5. A detailed examination of the ICC indicates that the proportion of slightly below average ability students obtaining a "1" for item 31 was slightly higher than expected.

The item parameter estimation process did not converge for MC item 117 in Session 6. A close examination of the ICC for the item indicated that the student responses do not fit the model well. The item did not discriminate between low or 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.

Writing

The critical Z-value for the eighth-grade Writing assessment, based on the calibration sample, is 25.4. Fit values above this critical value may indicate poor model fit. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 108 and 109, respectively.

The Z-statistics for most MC items and most CR items are smaller than the critical Z-value, indicating that the proportions of students with each item score are close to the proportions that would be predicted by the estimated thetas and item parameters.

CR item 3A in Session 2 exhibited less than optimal fit with a Z-statistic of 70.8, exceeding the critical Z-value of 25.4. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was less than expected and obtaining a "1" was higher than expected. The proportion of higher ability students obtaining a "3" was also less than expected and obtaining a "4" was higher than expected.

CR item 25 in Session 3 exhibited less than optimal fit with a Z-statistic of 31.7, exceeding the critical Z-value of 25.4. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was less than expected and obtaining a "1" was higher than expected.

CR item 58 in Session 5 exhibited less than optimal fit with a Z-statistic of 36.2, exceeding the critical Z-value of 25.4. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was less than expected and obtaining a "1" was higher than expected.

MC item 74 in Session 5 exhibited less than optimal fit with a Z-statistic of 27.4, marginally exceeding the critical Z-value of 25.4. A close examination of the ICC for the item indicated that the proportion of lower and higher ability students obtaining a "1" were slightly less than expected.

CR item 89 in Session 5 exhibited less than optimal fit with a Z-statistic of 44.4, exceeding the critical Z-value of 25.4. A close examination of the ICC for the item indicated that the

proportion of lower ability students obtaining a "0" was less than expected, and obtaining a "1" was higher than expected.

Mathematics

The critical Z-value for the <u>eighth-grade Mathematics</u> assessment, based on the calibration sample, is 27.1. Fit values above this critical value may indicate poor model fit. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 110 and 111, respectively.

The Z-statistics for most MC items and most CR items are smaller than the critical Z-value, indicating that the proportions of students with each item score are close to the proportions that would be predicted by the estimated thetas and item parameters.

CR item 9 in Session 1 exhibited a less than optimal fit with a Z-statistic of 32.7, exceeding the critical Z-value of 27.1. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was higher than expected and obtaining a "1" was lower than expected. The proportion of higher ability students obtaining a "2" was also lower than expected.

CR item 14 in Session 1 exhibited a less than optimal fit with a Z-statistic of 56.2, exceeding the critical Z-value of 27.1. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was lower than expected and obtaining a "1" was higher than expected. The proportion of higher ability students obtaining a "1" was lower than expected and obtaining a "2" or "3" was higher than expected. These observations suggest that the item might have functioned better had it been an item with fewer score levels.

The item parameter estimation process did not converge for MC item 15 in Session 1. A close examination of the ICC for the item indicated that the student responses do not fit the model well. The item did not discriminate between low or 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.

MC item 33 in Session 2 exhibited a less than optimal fit with a Z-statistic of 36.9, exceeding the critical Z-value of 27.1. A close examination of the ICC for the item indicated that the proportion of lower ability students obtaining a "1" was higher than expected

CR item 50 in Session 3 exhibited a less than optimal fit with a Z-statistic of 38.9, exceeding the critical Z-value of 27.1. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "1" was more than expected and the proportion of higher ability students obtaining a "1" was lower than expected. These observations suggest that the item might have functioned better had it been an item with fewer score levels.

CR item 53 in Session 3 exhibited a less than optimal fit with a Z-statistic of 28.7, exceeding the critical Z-value of 27.1. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was more than expected and obtaining a "1" was lower than expected.

Science

The critical Z-value for the eight-grade Science assessment, based on the calibration sample, is 26.7. Fit values above this critical value may indicate poor model fit. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 112 and 113, respectively.

The Z-statistics for most MC items and all CR items are smaller than the critical Z-value, indicating that the proportions of students with each item score are close to the proportions that would be predicted by the estimated thetas and item parameters.

MC item 58 in Session 2 exhibited a less than optimal fit with a Z-statistic of 36.3 exceeding the critical Z-value of 26.7. A close examination of the ICC for the item indicated that the proportion of lower ability students obtaining a "1" was higher than expected. The item analysis results also indicated that the item functioned less than is optimal in the assessment. The point-biserial correlation coefficient for one of the incorrect options on this item was also non-negative (r_{itt} = .12) and larger than the value for the correct response (r_{itt} = .00). The item was rather difficult (p-value = .14).

Ninth Grade

Reading

The critical Z-value for the ninth-grade Reading assessment, based on the calibration sample, is 24.1. Fit values above this critical value may indicate poor model fit. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 114 and 115, respectively. They indicate that the item responses fit the model for all items of the ninth-grade Reading assessment reasonably well.

The Z-statistics for all MC items and most CR items are small compared to the critical Z-value, indicating that the proportions of students with each item score are close to the proportions that would be expected by the estimated thetas and item parameters.

CR item 23 in Session 3 exhibited a less than optimal fit with a Z-statistic of 27.8, exceeding the critical Z-value of 24.1. A close examination of the ICC for the item indicated that the proportion of higher ability students obtaining a "2" was lower than expected and obtaining a "1" was higher than expected. The proportion of lower ability students obtaining a "1" was lower than expected and obtaining a "0" was higher than expected.

Writing

The critical Z-value for the ninth-grade Writing assessment, based on the calibration sample, is 24.6. Fit values above this critical value may indicate poor model fit. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 116 and 117, respectively.

The Z-statistics for all MC items and most CR items are smaller than the critical Z-value, indicating that the proportions of students with each item score are close to the proportions that would be predicted by the estimated thetas and item parameters.

CR item 3A in Session 2 exhibited less than optimal fit with a Z-statistic of 141.9, exceeding the critical Z-value of 24.6. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was less than expected and obtaining a "1" was higher than expected.

CR item 48 in Session 4 exhibited less than optimal fit with a Z-statistic of 24.8, marginally exceeding the critical Z-value of 24.6. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was less than expected and obtaining a "1" was higher than expected.

CR item 90 in Session 5 exhibited less than optimal fit with a Z-statistic of 32.1, exceeding the critical Z-value of 24.6. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was less than expected and obtaining a "1" or a "2" was higher than expected.

Mathematics

The critical Z-value for the ninth-grade Mathematics assessment, based on the calibration sample, is 26.7. Fit values above this critical value may indicate poor model fit. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 118 and 119, respectively.

The Z-statistics for most MC items and most CR items are smaller than the critical Z-value, indicating that the proportions of students with each item score are close to the proportions that would be predicted by the estimated thetas and item parameters.

CR item 28 in Session 2 exhibited a less than optimal fit with a Z-statistic of 31.3, exceeding the critical Z-value of 26.7. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was more than expected and obtaining a "1" was lower than expected. The proportion of average ability students obtaining a "3" was also slightly higher than expected.

The item parameter estimation process did not converge for MC item 51 in Session 3. A close examination of the ICC for the item indicated that the student responses do not fit the model well. The item did not discriminate between low or 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.

Tenth Grade

Reading

The critical Z-value for the <u>tenth-grade Reading</u> assessment, based on the calibration sample, is 24.3. Fit values above this critical value may indicate poor model fit. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 120 and 1217, respectively. They indicate that the item responses fit the model reasonably well for most items of the tenth-grade Reading assessment.

The Z-statistics for all MC items and all CR items are small compared to the critical Z-value, indicating that the proportions of students with each item score are close to the proportions that would be expected by the estimated thetas and item parameters.

Writing

The critical Z-value for the <u>tenth-grade Writing</u> assessment, based on the calibration sample, is 24.1. Fit values above this critical value may indicate poor model fit. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 122 and 123, respectively.

The Z-statistics for all MC items and most CR items are smaller than the critical Z-value, indicating that the proportions of students with each item score are close to the proportions that would be predicted by the estimated thetas and item parameters.

CR item 3A in Session 2 exhibited less than optimal fit with a Z-statistic of 25.7, marginally exceeding the critical Z-value of 24.1. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was less than expected, and obtaining a "1" was higher than expected.

CR item 29 in Session 3 exhibited less than optimal fit with a Z-statistic of 56.1, exceeding the critical Z-value of 24.1. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was less than expected, and obtaining a "1" was higher than expected.

CR item 84 in Session 5 exhibited less than optimal fit with a Z-statistic of 41.0, exceeding the critical Z-value of 24.1. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was less than expected, and obtaining a "1" was higher than expected.

CR item 98 in Session 5 exhibited less than optimal fit with a Z-statistic of 59.4, exceeding the critical Z-value of 24.1. A close examination of the ICCs for the item indicated that the

proportion of lower ability students obtaining a "0" was less than expected, and obtaining a "1" was higher than expected.

Mathematics

The critical Z-value for the <u>tenth-grade Mathematics</u> assessment, based on the calibration sample, is 26.5. Fit values above this critical value may indicate poor model fit. The item parameters and fit statistics for the multiple-choice and constructed-response items are shown in Tables 124 and 125, respectively.

The Z-statistics for most MC items and most CR items are small compared to the critical Z-value, indicating that the proportions of students with each item score are close to the proportions that would be predicted by the estimated thetas and item parameters.

CR item 40 in Session 2 exhibited less than optimal fit with a Z-statistic of 30.0, exceeding the critical Z-value of 26.5. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was lower than expected and obtaining a "1" was higher than expected. The proportion of higher ability students obtaining a "2" or "3" was also lower than expected and obtaining a "4" slightly higher than expected.

CR item 43 in Session 3 exhibited less than optimal fit with a Z-statistic of 84.6, exceeding the critical Z-value of 26.5. A close examination of the ICCs for the item indicated that the proportion of lower ability students obtaining a "0" was higher than expected and obtaining a "1" was lower than expected. The proportion of higher ability students obtaining a "1" was also higher than expected and obtaining a "2" lower than expected.

The item parameter estimation process did not converge for MC items 53 and 57 in Session 3. A close examination of the ICCs for the items indicated that the student responses do not fit the model well. The items 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.

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

An example of this procedure for gender bias analyses follows. The parameters for each item (a_i, b_i, and c_i) and the trait or scale score (θ) for each examinee are estimated for the three-parameter logistic model:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

large positive difference in one range of θ and a large negative difference in another range of θ, yet have a small overall difference.

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

Differential Item Functioning Ratings

The DIF 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.

Third Grade
Reading

CR item 15 indicated a sign of DIF disfavoring male students.

Reading – Spanish Version

No items indicated DIF with any of the groups.

Writing

No items indicated DIF with any of the groups.

Writing – Spanish Version

No items indicated DIF with any of the groups.

Fourth Grade

Reading

CR item 36 indicated a sign of DIF favoring female students, disfavoring male students, and favoring Hispanic students. CR item 105 indicated a sign of DIF disfavoring Hispanic students.

Reading – Spanish Version

No items indicated DIF with any of the groups.

Writing

CR items 71 and 90 both indicated signs of DIF favoring African American students.

Writing – Spanish Version

No items indicated DIF with any of the groups.

Fifth Grade

Reading

CR item 8 indicated a sign of DIF favoring Hispanic students.

Writing

No items indicated signs of DIF with any of the groups.

Mathematics

CR item 51 indicated a sign of DIF favoring African American students.

Sixth Grade

Reading

CR item 11 indicated a sign of DIF favoring Hispanic students, and CR item 45 indicated a sign of DIF disfavoring African American students

Writing

CR item 3B indicated a sign of DIF favoring African American students. CR item 70 indicated a sign of DIF favoring female students and disfavoring male students. CR item 92 indicated a sign of DIF disfavoring male students.

Mathematics

CR item 541 indicated a sign of DIF disfavoring African American students.

Seventh Grade

Reading

CR item 22 indicated a sign of DIF favoring female students and disfavoring male students. CR item 25 indicated a sign of DIF disfavoring male students. CR item 98 indicated a sign of DIF favoring female students. CR items 23 and 37 both indicated a sign of DIF disfavoring African American students

Writing

CR items 3A, 3B, and 70 all indicated signs of DIF favoring female students and disfavoring male students.

Mathematics

CR items 24 and 60 both indicated signs of DIF favoring female students, and disfavoring male students.

Eighth Grade

Reading

CR item 17 and MC item 100 both indicated a sign of DIF disfavoring African American students.

Writing

CR item 58 indicated a sign of DIF disfavoring male students.

Mathematics

CR item 28 indicated signs of DIF favoring female students and favoring African American students. CR item 48 indicated a sign of DIF disfavoring African American students.

Science

CR item 20 indicated signs of DIF favoring female students, and disfavoring male students.

Ninth Grade

Reading

No items indicated DIF with any of the groups.

Writing

CR item 69 indicated signs of DIF favoring female students and disfavoring male students. CR items 74 and 76 69 indicated signs of DIF in the reverse direction, disfavoring female students and favoring male students.

Mathematics

CR item 40 indicated signs of DIF favoring female students and disfavoring female students.

Tenth Grade

Reading

CR items 12, 49, 52, and 53 all indicated a sign of DIF favoring African American students. CR item 52 also indicated a sign of DIF favoring Hispanic students. CR item 21 indicated a sign of DIF disfavoring female students. CR item 24 indicated a sign of DIF disfavoring female students and favoring male students. CR items 34 and 35 both indicated signs of DIF favoring female students and disfavoring male students. CR item 121, finally, indicated a sign of DIF favoring female students.

Writing

CR item 84 indicated signs of DIF favoring female students, disfavoring male students and favoring African American students.

Mathematics

No items indicated DIF with any of the groups.

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.

Table 1. The Colorado Model Content Standards and Sub-Content Areas

Reading		
Content Standards		
CS	Grades	
1	3-10	Reading Comprehension
4	4-10	Thinking Skills
5	4-10	Use of Literary Information
6	4-10	Literature
Sub-Content Standards		
SA	Grades	
1	7-10	Fiction
2	3-10	Nonfiction
3	3-10	Vocabulary
4	7-10	Poetry
1/4	3-6	Fiction and Poetry
Writing		
Content Standards		
CS	Grades	
2	3-10	Write for a Variety of Purposes
3	3-10	Write Using Conventions
Sub-Content Standards		
SA	Grades	
5	3-10	Paragraph Writing
6	4-10	Extended Writing
7	3-10	Grammar and Usage
8	3-10	Mechanics
Mathematics		
Content Standards		
CS	Grades	
1	5-6	Number Sense
2	5-10	Algebra, Patterns, and Functions
3	5-10	Satistics and Probability
4/5	5-10	Geometry and Measurement
6	5-6	Computational Techniques
1/6	7-10	Number Sense and Computational Techniques
Sub-Content Standards		
SA	Grades	
1	5-6	Number and Operational Sense
1	7	Number Sense
1	8	Linear Pattern Representation
0	9	Multiple Representations of Linear/Nonlinear
1	10	Multiple Representations of Functions
2	5-6	Patterns
2	7	Area and Perimeter Relationships
2	8-9	Proportional Thinking
2	10	Probability and Counting Techniques
3	5	Data Displays
3	6 and 8	Geometry

Table1. The Colorado Model Content Standards and Sub-Content Areas (Continued)		
Science		
Content Standards		
CS	Grades	
1	8	Scientific Investigations
2	8	Physical Science
3	8	Life Science
4	8	Earth and Space Science
5/6	8	Science and Technology
Sub-Content Standards		
SA	Grades	
1	8	Experimental Design and Investigations
2	8	Results and Data Analysis
3	8	Physics
4	8	Cemistry
5	8	Earth Science

Table 2. Configuration of 2003 CSAP Assessments

Content Area	Grade	Maximum Possible Points	Total # of Items			Frequency of CR Items with the Given Number of Maximum Points			
			MC	CR	Total	1	2	3	4
Reading	3	52	34	7	41		4	2	1
Reading	4	91	56	14	70		8	5	1
Reading	5	90	56	14	70		8	6	
Reading	6	91	54	15	69		10	3	2
Reading	7	91	57	14	71		9	4	1
Reading	8	90	54	15	69		10	4	1
Reading	9	96	58	15	73		8	6	1
Reading	10	101	61	14	75		3	10	1
Writing	3	56	35	18	53	15	3		
Writing	4	69	40	13	53	7	1		5
Writing	5	69	40	13	53	7	1		5
Writing	6	69	40	13	53	7	1		5
Writing	7	69	40	13	53	7	1		5
Writing	8	69	41	12	53	6	1		5
Writing	9	69	40	13	53	7	1		5
Writing	10	69	40	13	53	7	1		5
Spanish Reading	3	49	32	8	40		7	1	
Spanish Reading	4	83	52	14	66	2	8	3	1
Spanish Writing	3	56	35	18	53	15	3		
Spanish Writing	4	68	39	13	52	7	1		5
Mathematics	5	96	54	15	69		6	6	3
Mathematics	6	87	45	15	60		6	6	3
Mathematics	7	87	45	15	60		6	6	3
Mathematics	8	86	44	15	59		6	6	3
Mathematics	9	86	44	15	59		6	6	3
Mathematics	10	85	43	15	58		6	6	3
Science	8	99	64	20	84	8	10	1	1

Table 3. Configuration of 2003 CSAP Reading Assessments by Grade, Content Standard, Sub-Content Area, and Item Type

Grade	Content Area	Item Type	CS 1 No. of Items	CS 1 Total Pts.	CS 4 No. of Items	CS 4 Total Pts.	CS 5 No. of Items	CS 5 Total Pts.	CS 6 No. of Items	CS 6 Total Pts.	SC 1 No. of Items	SC 1 Total Pts.	SC 1/4 No. of Items	SC 1/4 Total Pts.	SC 2 No. of Items	SC 2 Total Pts.	SC 3 No. of Items	SC 3 Total Pts.	SC 4 No. of Items	SC 4 Total Pts.
3	Reading	MC	34	34							10	10			14	14	10	10		
		CR	7	18							7	18								
4	Reading	MC	25	25	14	14	11	11	6	6	26	26			11	11	8	8		
		CR	4	11	1	3	3	7	6	14	11	28								
5	Reading	MC	17	17	16	16	14	14	9	9	15	15			18	18	9	9		
		CR	2	5	2	5	3	8	7	16	9	21			2	5				
6	Reading	MC	20	20	14	14	11	11	9	9	17	17			13	13	13	13		
		CR	2	6	3	7	4	9	6	15	7	16			4	12				
7	Reading	MC	21	21	13	13	15	15	8	8	12	12			11	11	14	14	5	5
		CR	3	9	3	6	3	6	5	13	4	11			3	8			4	9
8	Reading	MC	15	15	15	15	12	12	12	12	11	11			16	16	10	10	5	5
		CR	6	14	2	6	3	6	4	10	3	8			5	10			4	12
9	Reading	MC	15	15	13	13	16	16	14	14	12	12			12	12	12	12	6	6
		CR	6	14	4	11	1	3	4	10	3	9			6	13			5	13
10	Reading	MC	21	21	10	10	12	12	18	18	16	16			15	15	10	10	8	8
		CR	1	3	5	14	3	9	5	14	6	17			3	8			2	6
3	Spanish Reading	MC	32	32							12	12			10	10	10	10		
		CR	8	17							7	15			1	2				
4	Spanish Reading	MC	27	27	10	10	8	8	7	7	23	23			14	14	7	7		
		CR	1	4	5	9	6	14	2	4	5	9			3	8				

Table 4. Configuration of 2003 CSAP Writing Assessments by Grade, Content Standard, Sub-Content Area, and Item Type

Grade	Content Area	Item Type	Content Standard 2 No. of Items	Total Pts.	Content Standard 3 No. of Items	Total Pts.	Sub-Content 5 No. of Items	Total Pts.	Sub-Content 6 No. of Items	Total Pts.	Sub-Content 7 No. of Items	Total Pts.	Sub-Content 8 No. of Items	Total Pts.
3	Writing	MC	7	7	28	28					16	16	12	12
		CR	9	12	9	9	12	15			3	3	3	3
4	Writing	MC	13	13	27	27	3	3			12	12	15	15
		CR	6	21	7	8	3	12	4	11	2	2	4	4
5	Writing	MC	15	15	25	25	5	5			18	18	14	14
		CR	6	21	7	8	3	12	4	11	2	2	4	4
6	Writing	MC	17	17	23	23	9	9			21	21	8	8
		CR	6	21	7	8	3	12	4	11	1	1	5	5
7	Writing	MC	20	20	20	20	6	6			13	13	12	12
		CR	6	21	7	8	3	12	4	11	2	2	4	4
8	Writing	MC	21	21	20	20	12	12			18	18	10	10
		CR	6	21	6	7	3	12	4	11	1	1	4	4
9	Writing	MC	20	20	20	20	10	10			14	14	10	10
		CR	6	21	7	8	3	12	4	11	2	2	4	4
10	Writing	MC	20	20	20	20	9	9			18	18	8	8
		CR	6	21	7	8	3	12	4	11	4	4	2	2
3	Spanish Writing	MC	6	6	29	29					21	21	8	8
		CR	9	12	9	9	12	15			2	2	4	4
4	Spanish Writing	MC	5	5	34	34					27	27	7	7
		CR	6	21	7	8	3	12	4	11	2	2	4	4

Table 5. Configuration of 2003 CSAP Mathematics Assessments by Grade, Content Standard, Sub-Content Area, and Item Type

| Grade | Content Area | Item Type | Content Standard | | | | | | | | | | | | | Sub-Content Area | | | | | |
| | | | 1 | | 1/6 | | 2 | | 3 | | 4/5 | | 6 | | 1 | | 2 | | 3 | |
			No. of Items	Total Pts.	No. of Items	Total Pts.	No. of Items	Total Pts.	No. of Items	Total Pts.	No. of Items	Total Pts.	No. of Items	Total Pts.	No. of Items	Total Pts.	No. of Items	Total Pts.	No. of Items	Total Pts.
5	Math	MC	14	14			8	8	7	7	11	11	14	14	12	12	4	4	3	3
		CR	2	5			4	11	4	12	3	9	2	5	3	8	3	9	2	7
6	Math	MC	9	9			7	7	7	7	12	12	10	10	12	12	7	7	7	7
		CR	2	7			3	10	4	11	4	10	2	4	4	11	2	6	2	5
7	Math	MC			15	15	9	9	9	9	12	12			6	6	3	3		
		CR			4	11	3	8	3	9	5	14			2	5	3	9		
8	Math	MC			10	10	11	11	9	9	14	14			5	5	9	9	5	5
		CR			4	11	4	11	3	8	4	12			3	9	6	15	2	7
9	Math	MC			10	10	11	11	11	11	12	12			3	3	3	3		
		CR			3	7	5	14	4	11	3	10			4	13	3	10		
10	Math	MC			12	12	11	11	10	10	10	10			4	4	5	5		
		CR			2	5	5	14	4	11	4	12			3	9	2	5		

Table 6. Configuration of 2003 CSAP Science Assessments by Grade, Content Standard, Sub-Content Area, and Item Type

Grade	Content Area	Item Type	Content Standard									
			1		**2**		**3**		**4**		**5/6**	
			No. of Items	Total Pts.	No. of Items	Total Pts.	No. of Items	Total Pts.	No. of Items	Total Pts.	No. of Items	Total Pts.
8	Science	MC	13	13	16	16	11	11	16	16	8	8
		CR	8	14	2	4	5	9	3	6	2	2

Grade	Content Area	Item Type	Sub-Content Area									
			1		**2**		**3**		**4**		**5**	
			No. of Items	Total Pts.	No. of Items	Total Pts.	No. of Items	Total Pts.	No. of Items	Total Pts.	No. of Items	Total Pts.
8	Science	MC	6	6	7	7	8	8	8	8	11	11
		CR	3	5	4	7	1	2	1	2	3	6

Table 7. Lowest and Highest Obtainable Scale Scores for the 2003 CSAP Assessments

Grade	Content Area	LOSS	HOSS
3	Reading	150	795
4	Reading	180	940
5	Reading	220	955
6	Reading	260	970
7	Reading	300	980
8	Reading	330	990
9	Reading	350	995
10	Reading	370	999
3	Spanish Reading	300	740
4	Spanish Reading	300	790
3	Writing	150	680
4	Writing	190	730
5	Writing	220	780
6	Writing	230	840
7	Writing	240	890
8	Writing	250	910
9	Writing	260	930
10	Writing	270	950
3	Spanish Writing	250	750
4	Spanish Writing	300	790
5	Mathematics	220	800
6	Mathematics	240	830
7	Mathematics	280	860
8	Mathematics	310	890
9	Mathematics	340	920
10	Mathematics	370	950
8	Science	300	790

Table 8. Scale Score Descriptive Statistics for the Total Test

Content Area	Grade	Group	N*	Mean SS	SD	Median SS
Reading	3	Female	25863	575	72.8	575
		Male	27024	562	76.7	565
		Total	52888	568	75.1	570
Reading	4	Female	26868	594	61.9	598
		Male	28215	580	66.3	586
		Total	55085	587	64.6	592
Reading	5	Female	27871	615	64.8	621
		Male	28908	601	71.5	610
		Total	56787	608	68.7	615
Reading	6	Female	27633	627	60.7	633
		Male	29008	614	68.9	624
		Total	56642	620	65.3	628
Reading	7	Female	27563	641	61.6	645
		Male	28969	624	70.1	633
		Total	56532	632	66.6	639
Reading	8	Female	26896	662	59.2	668
		Male	27878	644	66.6	653
		Total	54778	653	63.7	660
Reading	9	Female	27132	670	54.1	676
		Male	28419	648	67.7	659
		Total	55554	658	62.3	667
Reading	10	Female	24508	692	53.4	697
		Male	25145	671	67.8	684
		Total	49654	681	62.0	690
Spanish Reading	3	Female	853	530	40.6	532
		Male	810	516	41.5	519
		Total	1663	523	41.6	525
Spanish Reading	4	Female	312	533	40.2	536
		Male	396	521	46.4	526
		Total	708	527	44.2	532
Writing	3	Female	25888	485	58.2	483
		Male	27067	468	55.7	467
		Total	52956	477	57.5	475
Writing	4	Female	26889	497	55.2	497
		Male	28217	477	53.3	478
		Total	55108	486	55.1	487
Writing	5	Female	27889	513	57.3	514
		Male	28925	492	55.2	494
		Total	56822	502	57.2	504
Writing	6	Female	27642	533	60.4	533
		Male	29030	507	61.1	510
		Total	56674	520	62.1	521

Table 8. Scale Score Descriptive Statistics for the Total Test
(Continued)

Content Area	Grade	Group	N*	Mean SS	SD	Median SS
Writing	**7**	Female	27600	560	63.2	561
		Male	29009	527	64.6	530
		Total	56609	543	65.9	545
Writing	**8**	Female	26954	573	71.9	574
		Male	27936	536	71.3	539
		Total	54894	554	74.0	556
Writing	**9**	Female	27148	585	74.7	584
		Male	28389	549	77.6	552
		Total	55540	566	78.2	568
Writing	**10**	Female	24517	603	83.8	603
		Male	25140	565	83.7	569
		Total	49658	584	86.0	586
Spanish Writing	**3**	Female	853	511	55.2	510
		Male	791	488	58.3	487
		Total	1645	500	57.8	499
Spanish Writing	**4**	Female	313	525	42.2	524
		Male	396	513	42.9	517
		Total	709	519	43.0	520
Mathematics	**5**	Female	27928	502	69.9	504
		Male	29002	503	74.4	506
		Total	56938	503	72.2	505
Mathematics	**6**	Female	27636	520	72.1	522
		Male	29089	520	80.0	522
		Total	56731	520	76.3	522
Mathematics	**7**	Female	27614	544	65.2	546
		Male	29060	541	70.9	545
		Total	56679	542	68.2	546
Mathematics	**8**	Female	26937	553	68.3	559
		Male	27918	548	78.8	557
		Total	54859	550	73.8	558
Mathematics	**9**	Female	27190	567	67.2	576
		Male	28513	561	78.4	572
		Total	55706	564	73.2	574
Mathematics	**10**	Female	24551	582	67.6	588
		Male	25166	583	76.5	591
		Total	49721	582	72.2	589
Science	**8**	Female	26903	498	58.9	503
		Male	27881	502	66.9	509
		Total	54787	500	63.1	506

Table 9. Scale Score Descriptive Statistics by Content Standard

		Content Standard																			
		1			2			3			4			5			6			Total		
	Grade	Mean	SD	Median	Mean	SD	Median	Mean	SD	Median	Mean	SD	Median	Mean	SD	Median	Mean	SD	Median	Mean	SD	Median
Reading	3	568	75.1	570																568	75.1	570
Reading	4	595	89.5	592							588	116.6	595	586	96.0	592	586	87.5	591	587	64.6	592
Reading	5	620	105.3	614							611	97.6	615	608	88.6	615	610	81.9	616	608	68.7	615
Reading	6	624	88.7	627							623	92.5	629	620	91.8	628	622	82.3	629	620	65.3	628
Reading	7	637	87.4	640							639	92.2	640	633	82.3	639	632	87.7	639	632	66.6	639
Reading	8	656	78.3	661							662	87.9	658	649	86.5	659	650	82.4	660	653	63.7	660
Reading	9	659	72.0	668							662	81.0	667	655	92.0	667	658	74.5	667	658	62.3	667
Reading	10	685	82.5	690							690	91.3	692	679	81.6	689	682	71.4	691	681	62.0	690
Spanish Reading	3	523	41.6	525																523	41.6	525
Spanish Reading	4	526	53.6	531							523	60.6	531	524	60.5	531	528	72.0	528	527	44.2	532
Writing	3				481	67.2	497	490	83.5	474										477	57.5	475
Writing	4				486	60.5	486	491	67.2	488										486	55.1	487
Writing	5				503	61.0	504	507	70.0	503										502	57.2	504
Writing	6				521	65.8	522	522	72.9	520										520	62.1	521
Writing	7				544	68.3	546	547	78.4	545										543	65.9	545
Writing	8				556	77.1	557	558	86.6	555										554	74.0	556
Writing	9				569	84.5	569	567	88.9	567										566	78.2	568
Writing	10				586	92.4	585	591	103.1	586										584	86.0	586
Spanish Writing	3				502	74.8	500	504	67.0	501										500	57.8	499
Spanish Writing	4				523	54.7	516	519	44.7	522										519	43.0	520
Mathematics	5	511	97.7	506	517	111.0	506	507	88.8	505	506	94.6	506				514	107.1	505	503	72.2	505
Mathematics	6	518	98.2	521	529	95.6	528	522	92.2	524	519	91.6	523				524	108.6	523	520	76.3	522
Mathematics	7	540	78.8	546	545	88.2	548	535	87.3	547	539	82.9	547							542	68.2	546
Mathematics	8	540	97.1	557	548	83.1	557	548	89.7	559	548	86.2	560							550	73.8	558
Mathematics	9	554	95.1	574	556	90.2	574	564	78.8	572	553	97.1	574							564	73.2	574
Mathematics	10	573	92.7	588	579	82.4	591	582	78.2	588	571	90.9	588							582	72.2	589
Science	8	504	72.6	507	496	78.2	506	500	75.8	508	496	74.4	504	503	87.2	504				500	63.1	506

Table 10. Scale Score Descriptive Statistics by Sub-Content Area

	Grade	Sub-Content Area 1 Mean	SD	Median	2 Mean	SD	Median	3 Mean	SD	Median	4 Mean	SD	Median	5 Mean	SD	Median	6 Mean	SD	Median	7 Mean	SD	Median	8 Mean	SD	Median	Total Mean	SD	Median
Reading	3	573	87.6	570	606	129.4	566	581	114.1	576																568	75.1	570
Reading	4	587	69.7	592	629	151.7	593	624	150.7	604																587	64.6	592
Reading	5	609	77.2	616	611	93.2	614	676	178.8	619																608	68.7	615
Reading	6	622	76.0	629	623	85.4	628	628	118.0	627																620	65.3	628
Reading	7	637	88.3	640	658	129.0	639	639	112.3	641	632	100.5	640													632	66.6	639
Reading	8	656	87.9	661	666	97.7	660	661	117.0	659	649	86.5	660													653	63.7	660
Reading	9	662	80.1	667	664	87.2	668	663	95.8	668	650	94.4	667													658	62.3	667
Reading	10	686	74.5	691	683	83.5	690	700	132.3	693	678	91.7	692													681	62.0	690
Spanish Reading	3	525	46.4	526	526	63.7	525	538	85.5	524																523	41.6	525
Spanish Reading	4	526	51.0	532	523	63.6	531	531	96.1	530																527	44.2	532
Writing	3													483	68.4	481				504	102.5	478	513	111.3	484	477	57.5	475
Writing	4													486	73.5	479	497	91.7	479	511	102.9	494	492	79.4	489	486	55.1	487
Writing	5													503	72.3	504	531	121.1	500	510	81.0	505	517	98.2	505	502	57.2	504
Writing	6													521	71.2	521	549	123.8	524	522	89.9	521	521	95.7	522	520	62.1	521
Writing	7													544	78.9	546	580	141.4	554	550	92.8	545	544	96.3	544	543	65.9	545
Writing	8													555	82.4	556	590	136.5	592	562	101.7	555	558	102.7	554	554	74.0	556
Writing	9													572	98.0	570	611	157.0	593	573	116.1	569	565	109.3	566	566	78.2	568
Writing	10													591	107.7	583	641	185.8	590	593	113.9	588	592	141.3	587	584	86.0	586
Spanish Writing	3													512	88.1	500				509	82.8	501	513	98.2	504	500	57.8	499
Spanish Writing	4													521	75.8	515	524	61.6	518	518	50.0	522	508	85.3	520	519	43.0	520
Mathematics	5	511	96.9	505	539	141.2	505	522	120.5	502																503	72.2	505
Mathematics	6	520	90.3	520	530	103.6	522	524	113.6	525																520	76.3	522
Mathematics	7	541	107.5	546	534	100.9	548																			542	68.2	546
Mathematics	8	545	90.0	558	539	98.8	557	545	96.6	557																550	73.8	558
Mathematics	9	561	89.3	575	549	104.5	576																			564	73.2	574
Mathematics	10	586	86.2	592	578	89.7	598																			582	72.2	589
Science	8	509	101.7	504	523	106.3	508	499	91.8	505	489	98.8	505	495	75.8	503										500	63.1	506

Table 11. Raw Score Descriptive Statistics by Content Standard

	Grade	Mean	SD	Median	Mean % of Max.	Mean	SD	Median	Mean % of Max.	Mean	SD	Median	Mean % of Max.	Mean	SD	Median	Mean % of Max.	Mean	SD	Median	Mean % of Max.	Mean	SD	Median	Mean % of Max.	Mean	SD	Median	Mean % of Max.
		1				**2**				**3**				**4**				**5**				**6**				**Total**			
Reading	3	39	9.0	42.0	75.8																					39	9.0	42.0	75.8
Reading	4	27	6.5	29.0	76.0									11	4.0	11.0	62.0	11	3.8	12.0	63.2	11	4.0	11.0	55.0	60	16.6	64.0	66.2
Reading	5	16	4.3	17.0	74.0									14	4.3	14.0	65.2	13	4.5	14.0	59.9	15	5.0	16.0	60.6	58	16.4	62.0	64.8
Reading	6	18	5.4	18.0	67.6									13	4.3	14.0	62.2	13	4.2	13.0	62.8	14	5.1	15.0	60.0	58	17.4	61.0	63.3
Reading	7	21	6.0	22.0	70.1									13	3.9	14.0	69.4	13	3.9	13.0	60.5	12	4.3	13.0	58.4	59	16.5	63.0	65.1
Reading	8	20	5.7	21.0	67.6									15	3.9	16.0	72.5	10	3.4	10.0	54.9	12	4.1	12.0	52.9	56	15.5	59.0	62.6
Reading	9	17	6.0	18.0	59.9									16	5.1	17.0	66.2	11	3.8	12.0	60.2	15	4.6	16.0	62.6	60	17.9	63.0	62.2
Reading	10	16	4.8	17.0	68.2									17	5.4	19.0	71.5	12	4.3	12.0	56.9	19	5.7	20.0	60.0	65	18.3	69.0	63.9
Spanish Reading	3	33	8.4	34.0	66.5																						33	8.4	34.0	66.5
Spanish Reading	4	19	6.3	20.0	61.4									9	3.8	9.0	49.2	9	3.9	9.0	42.4	6	2.3	6.0	55.4	44	14.4	45.0	52.8
Writing	3					15	2.1	16.0	80.0	32	5.4	33.0	85.6													47	7.1	49.0	83.7
Writing	4					23	4.7	24.0	68.7	26	6.0	27.0	73.3													49	10.0	50.0	71.0
Writing	5					23	5.9	24.0	64.3	25	5.7	26.0	74.8													48	10.9	50.0	69.3
Writing	6					25	6.3	26.0	66.6	20	5.5	20.0	63.7													45	11.2	47.0	65.3
Writing	7					26	6.9	26.0	62.3	18	4.7	19.0	65.3													44	11.0	45.0	63.5
Writing	8					28	6.8	29.0	66.4	18	5.1	19.0	66.9													46	11.3	48.0	66.6
Writing	9					28	7.0	29.0	69.0	18	5.1	18.0	63.1													46	11.6	48.0	66.6
Writing	10					29	7.2	31.0	71.6	20	5.1	21.0	69.9													49	11.8	51.0	71.0
Spanish Writing	3					11	3.9	11.0	60.2	28	6.1	29.0	74.6													39	9.2	41.0	70.0
Spanish Writing	4					14	4.2	14.0	54.9	23	6.9	23.0	54.2													37	10.1	37.0	54.4
Mathematics	5	13	4.0	13.0	67.6	14	4.2	15.0	72.6	12	4.0	13.0	65.4	12	4.1	13.0	62.4					14	4.0	15.0	71.9	65	18.4	68.0	67.9
Mathematics	6	8	3.8	8.0	51.8	12	3.3	12.0	67.8	10	4.0	10.0	54.4	12	4.3	12.0	56.1					9	3.1	9.0	61.0	50	16.6	51.0	58.0
Mathematics	7	12	5.5	12.0	46.3	9	3.7	9.0	55.0	8	3.9	8.0	44.5	12	5.6	12.0	46.8									42	17.2	40.0	47.8
Mathematics	8	8	4.7	7.0	35.9	10	4.7	9.0	43.8	9	3.8	9.0	52.3	12	5.6	11.0	44.4									38	17.3	35.0	43.7
Mathematics	9	5	3.5	5.0	31.8	9	5.3	9.0	37.4	9	3.6	9.0	39.7	8	4.2	7.0	36.0									31	15.0	29.0	36.5
Mathematics	10	7	3.5	7.0	41.6	10	5.4	9.0	40.6	9	3.6	9.0	43.0	7	4.0	6.0	30.0									33	15.0	30.0	38.7
Science	8	17	5.6	18.0	63.6	9	3.8	9.0	46.4	10	3.3	10.0	49.9	10	3.7	10.0	45.9	6	2.0	6.0	57.7					52	16.0	53.0	52.9

Table 12. Raw Score Descriptive Statistics by Sub-Content Area

Content	Grade	Sub-Content Area 1 Mean	SD	Median	Mean % of Max.	2 Mean	SD	Median	Mean % of Max.	3 Mean	SD	Median	Mean % of Max.	4 Mean	SD	Median	Mean % of Max.	Total Mean	SD	Median	Mean % of Max.
Reading	3	20	5.2	21.0	72.0	12	2.5	13.0	83.8	7	2.1	8.0	75.0					39	9.0	42.0	75.8
Reading	4	35	10.1	37.0	64.5	8	2.4	9.0	74.9	6	1.6	6.0	73.4					60	16.6	64.0	66.2
Reading	5	23	7.0	24.0	62.8	15	4.7	16.0	66.8	7	1.8	8.0	79.6					58	16.4	62.0	64.8
Reading	6	21	6.7	22.0	63.7	16	5.2	17.0	62.7	8	2.9	9.0	64.2					58	17.4	61.0	63.3
Reading	7	16	4.5	17.0	67.5	14	4.2	15.0	74.3	9	3.1	9.0	63.2	8	3.1	8.0	57.4	59	16.5	63.0	65.1
Reading	8	12	3.5	13.0	64.4	20	5.1	21.0	75.9	6	2.3	6.0	58.6	9	3.7	9.0	50.8	56	15.5	59.0	62.6
Reading	9	14	4.0	14.0	65.1	17	5.6	19.0	69.5	7	2.5	7.0	60.2	10	4.4	10.0	52.4	60	17.9	63.0	62.2
Reading	10	23	6.8	25.0	70.4	15	4.6	16.0	67.2	7	2.2	7.0	69.4	7	2.8	7.0	51.0	65	18.3	69.0	63.9
Spanish Reading	3	18	5.0	19.0	67.6	7	2.4	7.0	60.4	7	2.1	7.0	70.7					33	8.4	34.0	66.5
Spanish Reading	4	19	6.2	19.0	58.1	12	4.7	12.0	53.8	4	1.6	4.0	57.8					44	14.4	45.0	52.8

Content	Grade	Sub-Content Area 5 Mean	SD	Median	Mean % of Max.	6 Mean	SD	Median	Mean % of Max.	7 Mean	SD	Median	Mean % of Max.	8 Mean	SD	Median	Mean % of Max.	Total Mean	SD	Median	Mean % of Max.
Writing	3	12	1.7	12.0	78.5					16	3.1	17.0	84.4	13	2.6	14.0	84.6	47	7.1	49.0	83.7
Writing	4	9	2.6	9.0	61.6	8	1.9	8.0	69.7	11	2.3	12.0	80.2	13	4.0	13.0	67.2	49	10.0	50.0	71.0
Writing	5	10	2.9	10.0	58.3	8	2.0	8.0	73.8	15	3.6	16.0	73.6	13	3.8	14.0	71.9	48	10.9	50.0	69.3
Writing	6	13	3.6	14.0	62.8	8	1.9	8.0	72.1	15	4.4	15.0	66.1	8	2.7	8.0	59.6	45	11.2	47.0	65.3
Writing	7	11	3.2	11.0	60.0	8	2.1	8.0	72.0	10	2.8	10.0	64.4	10	3.2	10.0	61.5	44	11.0	45.0	63.5
Writing	8	15	4.0	15.0	61.4	8	1.7	9.0	77.2	14	3.9	15.0	71.5	8	3.0	9.0	60.5	46	11.3	48.0	66.6
Writing	9	15	4.0	16.0	67.9	9	1.8	9.0	77.7	11	3.4	11.0	66.7	8	2.9	8.0	58.4	46	11.6	48.0	66.6
Writing	10	14	4.0	15.0	68.8	9	2.0	9.0	78.4	16	4.2	17.0	71.4	6	2.3	7.0	63.0	49	11.8	51.0	71.0
Spanish Writing	3	10	3.4	10.0	65.8					17	3.9	18.0	74.3	8	2.5	9.0	70.4	39	9.2	41.0	70.0
Spanish Writing	4	5	2.5	5.0	41.1	6	1.9	7.0	58.3	17	5.4	17.0	57.6	5	2.4	5.0	45.9	37	10.1	37.0	54.4

Table 12. Raw Score Descriptive Statistics by Sub-Content Area (Continued)

		Sub-Content Area																					
		\multicolumn{4}{c}{1}				\multicolumn{4}{c}{2}				\multicolumn{4}{c}{3}				\multicolumn{4}{c}{4}				\multicolumn{4}{c}{5}				\multicolumn{4}{c}{Total}			
Content	Grade	Mean	SD	Med	Mean % of Max.	Mean	SD	Med	Mean % of Max.	Mean	SD	Med	Mean % of Max.	Mean	SD	Med	Mean % of Max.	Mean	SD	Med	Mean % of Max.	Mean	SD	Median	Mean % of Max.
Mathematics	5	14	4.3	14.0	67.6	10	2.9	11.0	75.6	7	2.4	7.0	65.8									65	18.4	68.0	67.9
Mathematics	6	13	5.2	12.0	55.4	9	2.6	9.0	67.8	7	2.8	7.0	59.2									50	16.6	51.0	58.0
Mathematics	7	5	2.8	5.0	45.8	5	2.9	5.0	41.2													42	17.2	40.0	47.8
Mathematics	8	5	3.1	5.0	38.4	8	5.2	6.0	32.1	5	2.6	5.0	42.0									38	17.3	35.0	43.7
Mathematics	9	6	3.7	6.0	39.4	4	2.9	3.0	28.3													31	15.0	29.0	36.5
Mathematics	10	7	2.7	7.0	53.8	4	1.7	3.0	35.3													33	15.0	30.0	38.7
Science	8	7	2.7	7.0	60.3	10	3.1	11.0	71.5	5	2.2	5.0	51.5	4	2.0	4.0	41.3	7	3.0	7.0	42.3	52	16.0	53.0	52.9

Table 13. Reading Raw Score Correlations by Content Standard and Sub-Content Area

Grade 3 Reading

	SA 1/4	SA 2	SA 3	Total
CS 1	.96	.88	.84	1.00
SA 1/4		.75	.71	.96
SA 2			.70	.88
SA 3				.84

Grade 3 Spanish Reading

	SA 1/4	SA 2	SA 3	Total
CS 1	.95	.83	.80	1.00
SA 1/4		.68	.64	.95
SA 2			.57	.83
SA 3				.80

Grade 4 Reading

	CS 4	CS 5	CS 6	SA 1/4	SA 2	SA 3	Total
CS 1	.78	.79	.77	.93	.81	.78	.94
CS 4		.75	.72	.88	.84	.65	.89
CS 5			.72	.81	.75	.66	.89
CS 6				.90	.70	.67	.88
SA 1/4					.79	.70	.98
SA 2						.65	.86
SA 3							.77

Grade 4 Spanish Reading

	CS 4	CS 5	CS 6	SA 1/4	SA 2	SA 3	Total
CS 1	.75	.72	.67	.87	.88	.70	.94
CS 4		.68	.64	.84	.82	.57	.88
CS 5			.61	.71	.69	.53	.86
CS 6				.79	.66	.59	.79
SA 1/4					.73	.57	.92
SA 2						.60	.90
SA 3							.69

Grade 5 Reading

	CS 4	CS 5	CS 6	SA 1/4	SA 2	SA 3	Total
CS 1	.79	.78	.79	.84	.89	.87	.92
CS 4		.76	.76	.85	.91	.70	.90
CS 5			.76	.78	.79	.68	.90
CS 6				.96	.78	.68	.91
SA 1/4					.81	.70	.95
SA 2						.73	.92
SA 3							.80

Grade 6 Reading

	CS 4	CS 5	CS 6	SA 1/4	SA 2	SA 3	Total
CS 1	.80	.80	.80	.87	.85	.91	.93
CS 4		.76	.78	.88	.88	.73	.91
CS 5			.77	.79	.78	.74	.90
CS 6				.92	.86	.75	.92
SA 1/4					.80	.76	.94
SA 2						.73	.92
SA 3							.86

Grade 7 Reading

	CS 4	CS 5	CS 6	SA 1	SA 2	SA 3	SA 4	Total
CS 1	.82	.78	.76	.84	.89	.90	.69	.94
CS 4		.76	.75	.82	.88	.73	.76	.91
CS 5			.72	.74	.75	.71	.66	.89
CS 6				.87	.75	.66	.86	.89
SA 1					.78	.70	.67	.90
SA 2						.70	.67	.91
SA 3							.60	.84
SA 4								.81

Grade 8 Reading

	CS 4	CS 5	CS 6	SA 1	SA 2	SA 3	SA 4	Total
CS 1	.81	.73	.77	.80	.91	.78	.78	.94
CS 4		.70	.75	.80	.89	.62	.80	.91
CS 5			.71	.66	.70	.63	.68	.85
CS 6				.86	.76	.67	.84	.90
SA 1					.77	.59	.68	.87
SA 2						.61	.72	.92
SA 3							.59	.76
SA 4								.86

Grade 9 Reading

	CS 4	CS 5	CS 6	SA 1	SA 2	SA 3	SA 4	Total
CS 1	.84	.76	.80	.81	.90	.84	.82	.94
CS 4		.74	.80	.80	.91	.66	.87	.93
CS 5			.73	.69	.74	.68	.70	.87
CS 6				.92	.80	.66	.83	.91
SA 1					.78	.63	.73	.88
SA 2						.65	.77	.92
SA 3							.64	.79
SA 4								.89

Grade 10 Reading

	CS 4	CS 5	CS 6	SA 1	SA 2	SA 3	SA 4	Total
CS 1	.76	.77	.78	.84	.83	.88	.66	.91
CS 4		.74	.78	.88	.92	.63	.64	.91
CS 5			.75	.76	.75	.67	.63	.89
CS 6				.90	.77	.68	.87	.92
SA 1					.80	.68	.66	.94
SA 2						.66	.64	.91
SA 3							.58	.79
SA 4								.78

Table 14. Writing Raw Score Correlations by Content Standard and Sub-Content Area

Grade 3 Writing

	CS 3	SA 5	SA 7	SA 8	Total
CS 2	.70	.91	.65	.62	.84
CS 3		.61	.94	.91	.98
SA 5			.55	.55	.74
SA 7				.71	.91
SA 8					.88

Grade 3 Spanish Writing

	CS 3	SA 5	SA 7	SA 8	Total
CS 2	.70	.94	.63	.62	.88
CS 3		.64	.93	.85	.95
SA 5			.54	.59	.82
SA 7				.63	.88
SA 8					.82

Grade 4 Writing

	CS 3	SA 5	SA 6	SA 7	SA 8	Total
CS 2	.74	.89	.75	.65	.68	.92
CS 3		.62	.55	.86	.95	.95
SA 5			.52	.53	.59	.79
SA 6				.44	.49	.68
SA 7					.68	.82
SA 8						.89

Grade 4 Spanish Writing

	CS 3	SA 5	SA 6	SA 7	SA 8	Total
CS 2	.65	.88	.79	.59	.50	.85
CS 3		.50	.56	.95	.72	.95
SA 5			.48	.45	.41	.70
SA 6				.49	.43	.70
SA 7					.46	.89
SA 8						.69

Grade 5 Writing

	CS 3	SA 5	SA 6	SA 7	SA 8	Total
CS 2	.79	.89	.77	.77	.80	.95
CS 3		.66	.59	.93	.91	.94
SA 5			.56	.63	.63	.82
SA 6				.53	.54	.72
SA 7					.76	.90
SA 8						.90

Grade 6 Writing

	CS 3	SA 5	SA 6	SA 7	SA 8	Total
CS 2	.78	.94	.73	.78	.75	.95
CS 3		.73	.52	.96	.85	.94
SA 5			.58	.72	.66	.89
SA 6				.48	.45	.67
SA 7					.73	.91
SA 8						.84

Grade 7 Writing

	CS 3	SA 5	SA 6	SA 7	SA 8	Total
CS 2	.78	.91	.73	.75	.75	.96
CS 3		.69	.56	.89	.90	.92
SA 5			.59	.64	.65	.87
SA 6				.47	.49	.70
SA 7					.69	.85
SA 8						.86

Grade 8 Writing

	CS 3	SA 5	SA 6	SA 7	SA 8	Total
CS 2	.83	.95	.75	.83	.80	.97
CS 3		.77	.60	.94	.90	.94
SA 5			.61	.76	.71	.91
SA 6				.56	.55	.72
SA 7					.76	.92
SA 8						.88

Grade 9 Writing

	CS 3	SA 5	SA 6	SA 7	SA 8	Total
CS 2	.80	.95	.75	.79	.74	.96
CS 3		.75	.59	.92	.91	.93
SA 5			.64	.73	.68	.91
SA 6				.54	.53	.72
SA 7					.71	.89
SA 8						.85

Grade 10 Writing

	CS 3	SA 5	SA 6	SA 7	SA 8	Total
CS 2	.82	.95	.78	.80	.77	.97
CS 3		.76	.62	.97	.82	.94
SA 5			.65	.75	.67	.91
SA 6				.59	.53	.75
SA 7					.72	.92
SA 8						.83

Table 15. Mathematics and Science Raw Score Correlations by Content Standard and Sub-Content Area

Grade 5 Mathematics

	CS 2	CS 3	CS 4/5	CS 6	SA 1	SA 2	SA 3	Total
CS 1	.79	.79	.76	.79	.95	.73	.76	.91
CS 2		.79	.74	.80	.80	.96	.76	.91
CS 3			.76	.78	.80	.74	.92	.91
CS 4/5				.74	.77	.68	.72	.88
CS 6					.85	.74	.74	.91
SA 1						.74	.77	.92
SA 2							.71	.85
SA 3								.86

Grade 8 Mathematics

	CS 2	CS 3	CS 4/5	SA 1	SA 2	SA 3	Total
CS 1/6	.80	.76	.82	.76	.91	.74	.92
CS 2		.76	.81	.94	.80	.72	.92
CS 3			.78	.71	.77	.69	.89
CS 4/5				.77	.85	.90	.94
SA 1					.76	.68	.87
SA 2						.73	.91
SA 3							.84

Grade 6 Mathematics

	CS 2	CS 3	CS 4/5	CS 6	SA 1	SA 2	SA 3	Total
CS 1	.73	.79	.76	.76	.93	.70	.72	.91
CS 2		.73	.72	.71	.76	.94	.67	.87
CS 3			.76	.74	.80	.72	.72	.91
CS 4/5				.72	.78	.69	.94	.90
CS 6					.91	.68	.69	.87
SA 1						.72	.74	.94
SA 2							.64	.83
SA 3								.85

Grade 9 Mathematics

	CS 2	CS 3	CS 4/5	SA 1	SA 2	Total
CS 1/6	.78	.72	.76	.73	.81	.89
CS 2		.78	.78	.93	.77	.94
CS 3			.74	.80	.75	.89
CS 4/5				.74	.88	.91
SA 1					.73	.89
SA 2						.88

Grade 10 Mathematics

	CS 2	CS 3	CS 4/5	SA 1	SA 2	Total
CS 1/6	.79	.73	.74	.68	.62	.89
CS 2		.78	.79	.84	.65	.94
CS 3			.73	.80	.85	.89
CS 4/5				.68	.62	.90
SA 1					.59	.84
SA 2						.75

Grade 7 Mathematics

	CS 2	CS 3	CS 4/5	SA 1	SA 2	Total
CS 1/6	.80	.78	.82	.92	.76	.94
CS 2		.74	.78	.74	.72	.90
CS 3			.77	.73	.71	.89
CS 4/5				.76	.91	.93
SA 1					.71	.87
SA 2						.86

Grade 8 Science

	CS 2	CS 3	CS 4	CS5/6	SA 1	SA 2	SA 3	SA 4	SA 5	Total
CS 1	.70	.71	.69	.64	.90	.92	.65	.58	.66	.91
CS 2		.67	.71	.61	.66	.59	.89	.87	.69	.86
CS 3			.68	.61	.65	.62	.62	.56	.66	.85
CS 4				.63	.65	.59	.65	.58	.96	.86
CS 5/6					.60	.56	.57	.50	.61	.77
SA 1						.67	.61	.55	.63	.84
SA 2							.56	.48	.56	.80
SA 3								.55	.63	.79
SA 4									.57	.72
SA 5										.83

Table 16. Cronbach Alpha Reliability Coefficients for Content Standards

| | | Content Standard | | | | | | |
| | | 1 | 2 | 3 | 4 | 5 | 6 | Total |
Content Area	Grade	Alpha	Alpha	Alpha	Alpha	Alpha	Alpha	Alpha
Reading	3	.90						.90
Reading	4	.86			.78	.77	.69	.93
Reading	5	.82			.74	.76	.77	.93
Reading	6	.82			.75	.77	.76	.93
Reading	7	.83			.79	.74	.72	.93
Reading	8	.82			.78	.68	.75	.93
Reading	9	.82			.82	.73	.75	.94
Reading	10	.79			.82	.71	.79	.93
Spanish Reading	3	.87						.87
Spanish Reading	4	.81			.70	.64	.56	.90
Writing	3		.73	.88				.90
Writing	4		.80	.87				.91
Writing	5		.84	.86				.91
Writing	6		.87	.83				.92
Writing	7		.86	.81				.91
Writing	8		.87	.85				.93
Writing	9		.88	.82				.92
Writing	10		.88	.84				.92
Spanish Writing	3		.80	.85				.90
Spanish Writing	4		.70	.84				.87
Mathematics	5	.78	.78	.75	.73		.81	.94
Mathematics	6	.74	.70	.74	.78		.74	.94
Mathematics	7	.83	.76	.73	.81			.94
Mathematics	8	.80	.78	.73	.81			.94
Mathematics	9	.73	.82	.73	.75			.93
Mathematics	10	.74	.82	.73	.77			.93
Science	8	.81	.71	.65	.73	.58		.92

Table 17. Cronbach Alpha Reliability Coefficients for Sub-Content Areas

Content Area	Grade	Sub-Content Area							
		1 Alpha	**2** Alpha	**3** Alpha	**4** Alpha	**5** Alpha	**6** Alpha	**7** Alpha	**8** Alpha
Reading	3	.80	.79	.71					
Reading	4	.88	.73	.59					
Reading	5	.83	.80	.68					
Reading	6	.84	.75	.75					
Reading	7	.77	.77	.76	.66				
Reading	8	.71	.85	.67	.70				
Reading	9	.72	.86	.66	.75				
Reading	10	.84	.78	.69	.56				
Spanish Reading	3	.78	.60	.64					
Spanish Reading	4	.83	.71	.43					
Writing	3					.73		.80	.79
Writing	4					.72	.72	.70	.81
Writing	5					.69	.70	.78	.81
Writing	6					.79	.71	.81	.70
Writing	7					.74	.76	.70	.73
Writing	8					.79	.70	.84	.75
Writing	9					.81	.71	.77	.69
Writing	10					.80	.74	.82	.65
Spanish Writing	3					.84		.77	.72
Spanish Writing	4					.60	.55	.82	.64
Mathematics	5	.79	.67	.63					
Mathematics	6	.81	.67	.69					
Mathematics	7	.71	.68						
Mathematics	8	.70	.80	.56					
Mathematics	9	.78	.71						
Mathematics	10	.67	.52						
Science	8	.65	.67	.60	.49	.70			

Table 18. Multiple-Choice Item Analyses, Grade 3 Reading

Item #	Session	CS	SA	R_ITT	Omit %	p-Value	Item #	Session	CS	SA	R_ITT	Omit %	p-Value
1	S1	1	1/4	.37	.0	.97	22	S2	1	2	.40	.7	.56
2	S1	1	3	.25	.1	.94	23	S2	1	2	.45	.1	.79
3	S1	1	3	.51	.1	.89	24	S2	1	3	.45	.3	.58
6	S1	1	3	.35	.8	.77	25	S2	1	2	.60	.1	.81
7	S1	1	2	.45	.1	.95	26	S2	1	2	.45	.2	.72
8	S1	1	2	.41	.1	.96	27	S2	1	3	.20	.1	.29
9	S1	1	2	.34	.1	.89	28	S2	1	2	.49	.2	.80
10	S1	1	2	.39	.1	.96	29	S2	1	3	.47	.5	.81
11	S1	1	2	.41	.1	.96	30	S2	1	1/4	.45	.1	.72
12	S1	1	2	.42	.2	.87	31	S2	1	1/4	.50	.2	.70
13	S1	1	2	.43	.8	.79	35	S2	1	1/4	.49	.3	.92
14	S1	1	3	.47	1.1	.74	36	S2	1	1/4	.40	.3	.85
17	S1	1	1/4	.45	1.6	.76	37	S2	1	3	.55	.4	.79
18	S1	1	1/4	.48	.8	.65	38	S2	1	3	.38	.6	.88
19	S1	1	1/4	.54	.9	.85	39	S2	1	1/4	.41	.6	.62
20	S2	1	2	.44	.0	.87	40	S2	1	1/4	.51	.6	.80
21	S2	1	2	.42	.1	.78	41	S2	1	3	.47	.6	.82

Mean R_ITT: .44
Mean p-value: .80

Table 19. Constructed-Response Item Analyses, Grade 3 Reading

Item #	Session	CS	SA	R_ITT	Percent of Students Obtaining Score Level					Omit %	p-Value
					0	1	2	3	4		
4	S1	1	1/4	.24	38.2	39.4	21.9			.4	.42
5	S1	1	1/4	.45	1.4	4.6	20.8	72.7		.6	.88
15	S1	1	1/4	.48	8.1	11.3	21.4	23.7	34.6	.9	.66
16	S1	1	1/4	.30	54.2	21.6	23.4			.8	.34
32	S2	1	1/4	.54	7.0	25.7	66.8			.5	.80
33	S2	1	1/4	.57	9.3	13.4	15.4	61.2		.8	.76
34	S2	1	1/4	.57	8.7	17.7	72.8			.9	.82

Mean R_ITT: .46
Mean p-value: .67

Table 20. Multiple-Choice Item Analyses, Grade 3 Spanish Reading

Item #	Session	CS	SA	R_ITT	Omit %	p-Value	Item #	Session	CS	SA	R_ITT	Omit %	p-Value
1	S1	1	1/4	.23	.1	.93	21	S2	1	1/4	.40	.1	.84
2	S1	1	1/4	.21	.1	.32	22	S2	1	1/4	.40	.4	.93
5	S1	1	3	.35	1.3	.79	23	S2	1	1/4	.37	.1	.68
6	S1	1	2	.25	.2	.76	24	S2	1	1/4	.44	.4	.87
7	S1	1	2	.37	.2	.70	25	S2	1	3	.40	.4	.87
8	S1	1	2	.38	.4	.89	28	S2	1	2	.33	.5	.86
9	S1	1	2	.35	.7	.53	30	S2	1	3	.38	.7	.73
10	S1	1	2	.31	.5	.66	31	S2	1	3	.27	.8	.74
11	S1	1	3	.30	.2	.89	33	S2	1	1/4	.48	1.0	.81
12	S1	1	1/4	.18	.5	.26	34	S2	1	3	.39	1.4	.66
14	S1	1	3	.28	1.4	.45	35	S2	1	1/4	.42	1.6	.67
15	S1	1	3	.37	1.5	.66	36	S2	1	3	.29	1.4	.53
16	S1	1	1/4	.38	2.1	.85	37	S2	1	2	.46	1.4	.76
18	S1	1	1/4	.35	3.4	.77	38	S2	1	2	.26	1.6	.40
19	S1	1	1/4	.41	2.9	.75	39	S2	1	2	.40	1.6	.70
20	S1	1	3	.45	3.0	.75	40	S2	1	2	.26	1.6	.42

Mean R_ITT: .35
Mean p-value: .70

Table 21. Constructed-Response Item Analyses, Grade 3 Spanish Reading

Item #	Session	CS	SA	R_ITT	Percent of Students Obtaining Score Level					Omit %	p-Value
					0	1	2	3	4		
3	S1	1	1/4	.48	10.9	9.1	18.2	59.2		2.5	.74
4	S1	1	1/4	.31	42.8	37.5	19.0			.7	.38
13	S1	1	1/4	.51	22.7	28.7	47.6			1.0	.62
17	S1	1	1/4	.44	5.9	9.4	81.8			2.9	.87
26	S2	1	1/4	.42	9.1	30.1	60.5			.3	.76
27	S2	1	1/4	.39	46.3	23.1	30.0			.6	.42
29	S2	1	2	.31	59.3	18.8	19.3			2.6	.29
32	S2	1	1/4	.51	18.3	34.6	45.9			1.2	.63

Mean R_ITT: .42
Mean p-value: .59

Table 22. Multiple-Choice Item Analyses, Grade 3 Writing

Item #	Session	CS	SA	R_ITT	Omit %	p-Value	Item #	Session	CS	SA	R_ITT	Omit %	p-Value
1	S1	3	7	.37	.0	.97	21	S2	3	8	.40	.1	.93
2	S1	3	7	.48	.1	.90	22	S2	3	8	.47	.1	.90
3	S1	3	7	.24	.1	.61	23	S2	3	8	.41	.2	.93
4	S1	3	8	.37	.2	.87	24	S2	3	7	.44	.1	.87
6	S1	2		.39	.0	.98	25	S2	3	7	.43	.1	.95
7	S1	2		.43	.1	.96	26	S2	3	7	.37	.4	.90
8	S1	2		.29	.4	.84	27	S2	3	7	.41	.1	.91
9	S1	3	7	.51	1.0	.87	28	S2	3	7	.39	.4	.94
10	S1	3	7	.27	.2	.75	30	S2	3	8	.31	.1	.59
11	S1	3	7	.43	.2	.85	31	S2	3	7	.35	.1	.88
12	S1	3	7	.34	.3	.85	32	S2	3	7	.46	.2	.89
14	S1	2		.37	.1	.94	33	S2	3	8	.51	.1	.92
15	S1	2		.30	.2	.98	34	S2	3	8	.41	.2	.80
16	S1	2		.38	.1	.92	35	S2	3	8	.34	.4	.92
17	S1	2		.29	2.2	.73	36	S2	3	8	.44	.2	.87
18	S1	3	7	.48	.2	.76	37	S2	3	8	.40	.3	.75
19	S1	3	7	.36	.3	.67	38	S2	3	8	.36	.3	.96
20	S2	3	8	.37	.0	.87							

Mean R_ITT: .39
Mean p-value: .86

Table 23. Constructed-Response Item Analyses, Grade 3 Writing

Item #	Part	Session	CS	SA	R_ITT	Percent of Students Obtaining Score Level					Omit %	p-Value
						0	1	2	3	4		
5	A	S1	3	7	.51	14.1	85.8				.2	.86
5	B	S1	3	8	.50	21.6	78.2				.2	.78
5	C	S1	3	8	.48	14.1	85.8				.2	.86
5	D	S1	3	7	.23	7.4	92.5				.2	.93
5	E	S1	3	8	.47	26.2	73.7				.2	.74
5	F	S1	3	7	.50	34.8	65.0				.2	.65
13	A	S1	2	5	.31	3.1	80.7	16.0			.1	.56
13	B	S1	2	5	.34	4.8	95.0				.1	.95
13	C	S1	2	5	.45	14.4	85.4				.1	.85
13	D	S1	3	5	.33	3.1	96.8				.1	.97
29	A	S2	2	5	.35	5.0	79.9	14.9			.3	.55
29	B	S2	2	5	.33	1.9	97.9				.2	.98
29	C	S2	2	5	.48	10.7	89.0				.2	.89
29	D	S2	3	5	.32	1.4	98.4				.2	.98
39	A	S2	2	5	.29	1.8	87.6	10.3			.3	.54
39	B	S2	2	5	.30	2.2	97.6				.3	.98
39	C	S2	2	5	.47	12.5	87.3				.3	.87
39	D	S2	3	5	.30	1.0	98.7				.3	.99

Mean R_ITT: .39
Mean p-value: .83

Table 24. Multiple-Choice Item Analyses, Grade 3 Spanish Writing

Item #	Session	CS	SA	R_ITT	Omit %	p-Value	Item #	Session	CS	SA	R_ITT	Omit %	p-Value
1	S1	3	7	.33	.1	.92	21	S2	3	8	.31	.2	.71
2	S1	3	7	.46	.2	.80	22	S2	3	8	.40	.2	.52
3	S1	3	7	.23	.2	.95	23	S2	3	8	.38	.4	.79
4	S1	3	7	.36	.8	.69	24	S2	3	7	.38	.2	.71
5	S1	3	8	.39	.5	.88	25	S2	3	7	.36	.2	.91
6	S1	3	8	.29	1.0	.80	26	S2	3	7	.38	.6	.85
8	S1	3	7	.25	.0	.75	27	S2	3	8	.43	.2	.73
9	S1	3	7	.29	.1	.57	28	S2	3	8	.28	.4	.84
10	S1	3	7	.35	.7	.69	30	S2	2		.35	.1	.72
11	S1	3	7	.30	.9	.75	31	S2	2		.10	.2	.25
12	S1	3	7	.34	1.1	.56	32	S2	3	7	.40	.5	.81
13	S1	3	7	.30	1.2	.55	33	S2	3	7	.43	1.0	.79
14	S1	3	7	.32	.2	.70	34	S2	3	7	.29	1.0	.56
15	S1	3	7	.32	.4	.89	35	S2	3	7	.33	1.6	.54
16	S1	3	7	.35	.4	.92	36	S2	2		.41	.1	.70
17	S1	3	7	.35	.2	.92	37	S2	2		.26	.3	.50
18	S1	2		.36	.2	.81	38	S2	2		.26	.3	.79
20	S2	3	8	.38	.0	.58							

Mean R_ITT: .34
Mean p-value: .73

Table 25. Constructed-Response Item Analyses, Grade 3 Spanish Writing

Item #	Part	Session	CS	SA	R_ITT	Percent of Students Obtaining Score Level					Omit %	p-Value
						0	1	2	3	4		
7	A	S1	3	8	.34	17.0	82.9				.1	.83
7	B	S1	3	8	.46	14.3	85.6				.1	.86
7	C	S1	3	8	.35	66.8	33.1				.1	.33
7	D	S1	3	7	.20	30.3	69.6				.1	.70
7	E	S1	3	8	.44	41.8	58.1				.1	.58
7	F	S1	3	7	.28	44.8	55.1				.1	.55
19	A	S1	2	5	.45	8.7	60.4	30.8			.1	.61
19	B	S1	2	5	.46	39.1	60.8				.1	.61
19	C	S1	2	5	.53	37.2	62.7				.1	.63
19	D	S1	3	5	.40	4.7	95.2				.1	.95
29	A	S2	2	5	.48	6.7	62.4	30.9			.0	.62
29	B	S2	2	5	.51	30.3	69.7				.0	.70
29	C	S2	2	5	.54	39.6	60.4				.0	.60
29	D	S2	3	5	.38	5.4	94.6				.0	.95
39	A	S2	2	5	.45	27.7	53.3	18.9			.2	.46
39	B	S2	2	5	.48	39.1	60.7				.2	.61
39	C	S2	2	5	.52	46.0	53.9				.2	.54
39	D	S2	3	5	.42	8.1	91.7				.2	.92

Mean R_ITT: .43
Mean p-value: .67

Table 26. Multiple-Choice Item Analyses, Grade 4 Reading

Item #	Session	CS	SA	R_ITT	Omit %	p-Value	Item #	Session	CS	SA	R_ITT	Omit %	p-Value
4	S3	1	2	.37	.1	.93	41	S4	1	1/4	.48	.4	.90
5	S3	1	2	.42	.1	.90	42	S4	6	1/4	.50	.4	.89
6	S3	4	2	.51	.2	.74	43	S4	4	1/4	.56	.5	.74
7	S3	1	2	.40	2.9	.86	44	S4	4	2	.48	.4	.90
10	S3	1	1/4	.41	.3	.69	45	S4	1	2	.41	.6	.77
11	S3	6	1/4	.41	.2	.66	46	S4	5		.44	.5	.68
12	S3	6	3	.14	.3	.44	47	S4	1	2	.53	.8	.68
13	S3	4	1/4	.47	.2	.67	91	S6	1	1/4	.48	.1	.83
14	S3	1	1/4	.36	.6	.46	92	S6	1	1/4	.50	.1	.80
15	S3	5		.43	.6	.86	95	S6	1	1/4	.35	.9	.94
16	S3	5		.31	.4	.52	96	S6	1	1/4	.37	.2	.91
21	S3	4	1/4	.47	1.7	.64	97	S6	1	3	.51	.3	.88
22	S3	6	3	.48	1.7	.90	98	S6	1	3	.29	.5	.87
23	S3	1	3	.50	1.9	.75	99	S6	1	1/4	.44	.4	.77
24	S3	5		.43	2.8	.65	100	S6	1	1/4	.47	.2	.89
25	S4	4	2	.44	.0	.73	101	S6	6	1/4	.49	.3	.75
26	S4	5		.40	.1	.80	102	S6	1	3	.37	.3	.88
27	S4	4	2	.35	.2	.43	103	S6	4	2	.40	.2	.57
28	S4	1	1/4	.45	.1	.73	104	S6	4	2	.50	.3	.71
29	S4	4	1/4	.47	.1	.80	106	S6	1	3	.38	.6	.47
30	S4	6	1/4	.38	.6	.74	107	S6	5		.43	.7	.64
31	S4	4	1/4	.35	.4	.51	108	S6	5		.43	.3	.60
32	S4	4	1/4	.49	1.1	.76	109	S6	5		.39	.4	.69
33	S4	1	3	.42	.2	.65	110	S6	5		.44	.3	.61
34	S4	5		.43	.3	.75	111	S6	5		.47	.4	.78
37	S4	4	1/4	.32	3.2	.51	112	S6	1	1/4	.50	.5	.77
39	S4	1	1/4	.40	.7	.78	113	S6	1	1/4	.46	.9	.60
40	S4	1	1/4	.50	.9	.72	115	S6	4	1/4	.51	1.5	.62

Mean R_ITT: .43
Mean p-value: .71

Table 27. Constructed-Response Item Analyses, Grade 4 Reading

Item #	Session	CS	SA	R_ITT	Percent of Students Obtaining Score Level					Omit %	p-Value
					0	1	2	3	4		
8	S3	6	1/4	.46	27.2	23.1	49.0			.8	.61
9	S3	6	1/4	.32	46.0	30.6	22.2			1.3	.38
17	S3	5		.48	37.0	41.6	21.0			.3	.42
18	S3	5		.33	49.1	46.2	4.2			.5	.27
19	S3	4	1/4	.44	48.1	.3	31.4	18.9		1.3	.40
20	S3	6	1/4	.40	66.1	10.0	9.7	10.6		3.6	.20
35	S4	6	1/4	.41	17.8	6.4	74.8			1.1	.78
36	S4	1	1/4	.42	5.8	7.6	17.7	22.5	45.3	1.1	.73
38	S4	1	1/4	.42	47.4	21.1	31.0			.5	.42
93	S6	1	1/4	.41	4.0	20.2	75.4			.4	.86
94	S6	1	1/4	.50	5.7	9.3	15.7	68.5		.8	.82
105	S6	5		.60	8.2	6.6	12.6	69.4		3.2	.80
114	S6	6	1/4	.52	17.3	44.8	17.0	19.8		1.2	.46
116	S6	6	1/4	.48	27.3	33.9	37.5			1.3	.54

Mean R_ITT: .44
Mean p-value: .55

Table 28. Multiple-Choice Item Analyses, Grade 4 Spanish Reading

Item #	Session	CS	SA	R_ITT	Omit %	p-Value	Item #	Session	CS	SA	R_ITT	Omit %	p-Value
4	S3	1	2	.41	1.3	.42	41	S4	1	1/4	.25	.6	.36
5	S3	1	3	.10	.6	.49	42	S4	6	3	.35	.7	.79
6	S3	6	2	.29	.7	.66	43	S4	6	1/4	.31	.7	.44
7	S3	5		.32	.3	.37	44	S4	1	1/4	.09	1.0	.19
8	S3	5		.35	.4	.50	45	S4	6	1/4	.41	1.0	.82
10	S3	1	2	.35	.4	.57	46	S4	1	1/4	.47	1.4	.71
12	S3	4	2	.25	.4	.62	47	S4	1	1/4	.51	1.4	.62
15	S3	4	2	.40	.7	.49	48	S4	4	1/4	.44	1.4	.39
17	S3	4	1/4	.22	1.0	.67	92	S6	1	2	.36	.1	.84
18	S3	1	1/4	.43	1.0	.68	93	S6	1	3	.35	.1	.67
19	S3	4	1/4	.24	1.4	.76	94	S6	1	3	.36	.3	.50
20	S3	1	1/4	.41	2.0	.70	95	S6	1	2	.24	.3	.75
21	S3	6	1/4	.27	2.1	.91	96	S6	1	2	.41	.3	.81
22	S3	1	1/4	.17	2.3	.50	97	S6	1	2	.35	.6	.53
23	S3	1	1/4	.45	2.4	.54	98	S6	1	1/4	.40	.0	.80
24	S3	1	1/4	.43	2.4	.54	99	S6	1	1/4	.42	.0	.90
25	S3	6	1/4	.28	2.4	.53	100	S6	4	1/4	.23	.0	.45
26	S4	1	2	.24	.3	.53	102	S6	1	1/4	.45	.0	.60
27	S4	1	3	.42	.1	.77	104	S6	5		.29	.3	.31
28	S4	4	2	.36	.3	.65	105	S6	5		.34	.0	.50
30	S4	4	2	.30	.7	.41	106	S6	1	3	.31	.3	.31
31	S4	1	2	.37	.4	.63	107	S6	1	3	.06	.1	.19
33	S4	4	2	.14	2.4	.31	108	S6	5		.16	.3	.14
34	S4	1	1/4	.41	.6	.80	109	S6	5		.35	.3	.55
35	S4	1	1/4	.42	.6	.69	110	S6	5		.15	.3	.57
36	S4	6	3	.26	.8	.51	111	S6	5		.10	.1	.29
38	S4	4	1/4	.36	.7	.49							

Mean R_ITT (item 107 excluded): .33
Mean p-value (item 107 excluded): .57

Table 29. Constructed-Response Item Analyses, Grade 4 Spanish Reading

Item #	Session	CS	SA	R_ITT	Percent of Students Obtaining Score Level					Omit %	p-Value
					0	1	2	3	4		
9	S3	5		.51	18.8	16.4	62.7			2.1	.71
11	S3	1	2	.62	26.6	3.2	5.9	7.1	55.5	1.7	.65
13	S3	5		.24	71.0	14.1	7.9	4.8		2.1	.15
14	S3	4	2	.38	66.5	16.7	13.7			3.1	.22
16	S3	5		.53	39.7	25.4	32.6			2.3	.45
29	S4	5		.59	10.2	10.6	11.0	67.9		.3	.79
32	S4	4	2	.51	61.2	16.7	20.3			1.8	.29
37	S4	6	1/4	.48	44.9	26.3	17.5	9.7		1.6	.30
39	S4	6	1/4	.46	48.7	49.0				2.3	.49
40	S4	4	1/4	.48	50.3	47.0				2.7	.47
101	S6	4	1/4	.40	8.6	37.9	53.2			.3	.72
103	S6	4	1/4	.49	35.5	9.9	53.8			.8	.59
112	S6	5		.22	64.4	13.4	22.0			.1	.29
113	S6	5		.16	75.3	11.7	12.6			.4	.18

Mean R_ITT: .44
Mean p-value: .45

Table 30. Multiple-Choice Item Analyses, Grade 4 Writing

Item #	Session	CS	SA	R_ITT	Omit %	p-Value	Item #	Session	CS	SA	R_ITT	Omit %	p-Value
49	S5	2		.22	.1	.99	69	S5	3	8	.38	.9	.66
50	S5	2		.27	.1	.98	70	S5	3	8	.40	1.5	.77
51	S5	2		.42	.3	.89	72	S5	2	5	.34	.2	.76
52	S5	3	7	.31	.9	.83	73	S5	2	5	.30	.2	.54
53	S5	2		.25	1.1	.85	74	S5	2	5	.42	.3	.59
54	S5	3	7	.28	1.6	.95	75	S5	3	7	.40	.5	.85
55	S5	3	7	.44	.1	.92	76	S5	3	8	.43	.4	.62
56	S5	3	7	.31	.3	.80	77	S5	3	8	.45	.9	.56
57	S5	2		.31	.2	.94	78	S5	3	7	.41	.3	.90
58	S5	2		.31	.5	.80	79	S5	3	7	.35	.3	.96
59	S5	3	8	.37	.2	.72	80	S5	3	7	.37	.3	.92
60	S5	3	8	.28	.2	.80	81	S5	3	7	.33	.8	.49
61	S5	3	8	.30	.3	.90	82	S5	3	7	.39	1.2	.58
62	S5	2		.42	.2	.44	83	S5	3	7	.47	1.3	.81
63	S5	2		.34	.2	.69	84	S5	3	7	.44	1.8	.86
64	S5	3	8	.26	.3	.95	85	S5	2		.33	1.1	.75
65	S5	3	8	.45	.1	.70	86	S5	2		.39	1.1	.78
66	S5	3	8	.42	.2	.64	87	S5	3	8	.32	.4	.54
67	S5	3	8	.40	.4	.64	88	S5	3	8	.48	.7	.59
68	S5	3	8	.35	.7	.64	89	S5	3	8	.43	1.1	.83

Mean R_ITT: .37
Mean p-value: .76

Table 31. Constructed-Response Item Analyses, Grade 4 Writing

Item #	Part	Session	CS	SA	R_ITT	Percent of Students Obtaining Score Level					Omit %	p-Value
						0	1	2	3	4		
1		S1	2	6	.05	.5	99.5				.0	1.00
2	A	S2	3	8	.44	43.0	56.1				.9	.56
2	B	S2	3	8	.38	70.1	29.1				.9	.29
2	C	S2	3	7	.32	54.8	44.3				.9	.44
2	D	S2	3	8	.36	7.9	91.2				.9	.91
2	E	S2	3	7	.10	7.4	91.8				.9	.92
2	F	S2	3	8	.43	58.3	40.8				.9	.41
3	A	S2	2	6	.50	4.3	4.7	40.4	41.1	9.4	.1	.62
3	B	S2	2	6	.59	.4	7.3	42.7	40.4	9.1	.1	.63
3	C	S2	3	6	.55	.6	30.7	68.7			.1	.84
48		S4	2	5	.58	.8	8.4	44.4	34.7	10.2	1.5	.61
71		S5	2	5	.54	.2	5.8	42.9	39.0	11.6	.4	.64
90		S5	2	5	.53	.3	8.3	50.4	32.9	7.5	.7	.59

Mean R_ITT: .42
Mean p-value: .65

Table 32. Multiple-Choice Item Analyses, Grade 4 Spanish Writing

Item #	Session	CS	SA	R_ITT	Omit %	p-Value	Item #	Session	CS	SA	R_ITT	Omit %	p-Value
50	S5	3	7	.35	.1	.61	70	S5	3	7	.36	.1	.44
51	S5	3	7	.34	.3	.85	71	S5	3	7	.36	.6	.62
52	S5	3	7	.37	.3	.80	72	S5	3	7	.36	.6	.48
53	S5	2		.33	.1	.85	73	S5	3	7	.32	.3	.55
54	S5	2		.12	.1	.19	75	S5	3	7	.23	.4	.60
55	S5	2		.29	.6	.47	76	S5	3	7	.31	1.1	.59
56	S5	2		.27	.3	.92	77	S5	3	7	.21	.3	.41
57	S5	3	7	.37	.3	.63	78	S5	3	7	.30	.0	.37
58	S5	2		.32	.3	.77	79	S5	3	7	.32	.1	.53
59	S5	3	7	.28	.6	.94	80	S5	3	7	.17	.6	.40
60	S5	3	7	.32	.3	.85	81	S5	3	7	.27	.3	.28
61	S5	3	7	.20	.7	.97	82	S5	3	7	.23	.7	.31
62	S5	2		.34	.6	.90	83	S5	3	7	.28	.4	.50
63	S5	3	7	.39	.4	.41	84	S5	3	8	.24	.1	.36
64	S5	3	7	.39	.4	.76	85	S5	3	8	.31	.1	.56
65	S5	3	7	.33	.4	.60	86	S5	3	8	.28	.3	.50
66	S5	3	7	.44	1.0	.64	87	S5	3	8	.38	.1	.58
67	S5	3	7	.38	.3	.72	88	S5	3	8	.13	.4	.26
68	S5	3	7	.33	.7	.46	89	S5	3	8	.31	.4	.54
69	S5	3	7	.47	.8	.68	90	S5	3	8	.32	.7	.45

Mean R_ITT (item 54 excluded): .31
Mean p-value (item 54 excluded): .59

Table 33. Constructed-Response Item Analyses, Grade 4 Spanish Writing

Item #	Part	Session	CS	SA	R_ITT	Percent of Students Obtaining Score Level					Omit %	p-Value
						0	1	2	3	4		
1		S1	2	6	.12	1.4	98.6				.0	.99
2	A	S2	3	8	.38	30.2	69.3				.6	.69
2	B	S2	3	7	.34	66.3	33.1				.6	.33
2	C	S2	3	8	.33	64.0	35.4				.6	.35
2	D	S2	3	8	.16	96.9	2.5				.6	.03
2	E	S2	3	7	.35	60.2	39.2				.6	.39
2	F	S2	3	8	.22	26.8	72.6				.6	.73
3	A	S2	2	6	.51	17.6	7.9	29.8	41.9	2.8	.0	.51
3	B	S2	2	6	.59	2.0	9.9	39.6	46.0	2.5	.0	.59
3	C	S2	3	6	.39	5.5	91.1	3.4			.0	.49
49		S4	2	5	.54	11.1	17.5	35.8	29.1	3.1	3.4	.47
74		S5	2	5	.34	61.2	3.8	15.5	16.5	2.4	.6	.23
91		S5	2	5	.46	6.8	12.1	46.7	28.8	4.8	.8	.53

Mean R_ITT: .37
Mean p-value: .49

Table 34. Multiple-Choice Item Analyses, Grade 5 Reading

Item #	Session	CS	SA	R_ITT	Omit %	p-Value	Item #	Session	CS	SA	R_ITT	Omit %	p-Value
4	S3	1	2	.45	.2	.73	43	S4	5		.46	.3	.78
5	S3	1	2	.31	.2	.71	44	S4	6	2	.44	.4	.75
6	S3	4	2	.32	.3	.57	45	S4	4	2	.46	1.1	.88
9	S3	6	1/4	.45	.2	.85	46	S4	5		.20	.5	.21
10	S3	1	3	.32	.5	.66	47	S4	4	1/4	.35	.5	.62
11	S3	6	1/4	.52	.4	.78	48	S4	4	1/4	.41	.6	.71
12	S3	1	3	.49	.4	.82	49	S4	6	1/4	.44	.6	.51
15	S3	6	3	.42	.4	.95	50	S4	6	1/4	.51	.6	.84
16	S3	1	3	.48	.5	.82	93	S6	1	1/4	.34	.1	.82
17	S3	1	3	.46	.6	.92	94	S6	6	1/4	.51	.1	.90
18	S3	1	3	.35	1.0	.95	95	S6	4	1/4	.38	.2	.87
21	S3	4	2	.23	1.8	.65	96	S6	4	1/4	.29	.5	.77
22	S3	4	2	.19	2.1	.51	97	S6	1	3	.35	.5	.72
23	S3	4	2	.20	2.0	.56	100	S6	5		.43	.2	.54
24	S3	1	2	.44	2.1	.80	101	S6	5		.41	.3	.86
25	S3	1	2	.44	2.2	.79	102	S6	5		.31	.4	.50
26	S3	1	3	.48	2.5	.76	103	S6	5		.49	.2	.79
28	S4	6	1/4	.24	.1	.44	104	S6	5		.26	.4	.47
29	S4	6	1/4	.37	.2	.66	105	S6	4	2	.43	.1	.67
30	S4	5		.40	.0	.89	106	S6	4	2	.51	.3	.76
31	S4	5		.29	.1	.76	107	S6	1	3	.43	.3	.55
34	S4	4	2	.48	.1	.80	108	S6	5		.45	.5	.68
35	S4	5		.38	.2	.86	109	S6	5		.37	.2	.72
36	S4	4	2	.43	.3	.54	110	S6	5		.44	.2	.67
37	S4	1	2	.47	.5	.76	111	S6	5		.20	2.1	.36
38	S4	4	2	.51	.3	.77	112	S6	1	1/4	.50	.2	.83
41	S4	1	2	.45	.7	.87	113	S6	1	1/4	.48	.4	.69
42	S4	4	2	.49	1.6	.83	115	S6	4	1/4	.52	.9	.69

Mean R_ITT: .40
Mean p-value: .72

Table 35. Constructed-Response Item Analyses, Grade 5 Reading

Item #	Session	CS	SA	R_ITT	Percent of Students Obtaining Score Level					Omit %	p-Value
					0	1	2	3	4		
7	S3	1	2	.58	22.0	33.3	27.5	16.7		.5	.46
8	S3	1	1/4	.46	6.2	17.0	76.0			.7	.85
13	S3	4	1/4	.43	40.3	.3	33.1	25.9		.5	.48
14	S3	6	1/4	.46	50.6	14.0	14.4	18.7		2.3	.33
19	S3	5		.54	16.3	20.5	30.0	31.3		2.0	.58
20	S3	5		.47	27.0	47.9	19.2			5.9	.43
32	S4	6	1/4	.33	21.5	32.2	45.4			1.0	.62
33	S4	6	1/4	.48	17.2	29.7	51.5			1.5	.66
39	S4	4	2	.46	34.4	25.6	38.6			1.5	.51
40	S4	5		.59	24.5	25.5	23.8	25.0		1.2	.49
98	S6	6	1/4	.40	14.1	45.7	39.3			.9	.62
99	S6	6	1/4	.46	30.4	46.2	22.6			.9	.46
114	S6	6	1/4	.53	14.4	37.0	19.9	27.8		.9	.53
116	S6	6	1/4	.45	23.2	33.9	42.1			.8	.59

Mean R_ITT: .48
Mean p-value: .54

Table 36. Multiple-Choice Item Analyses, Grade 5 Writing

Item #	Session	CS	SA	R_ITT	Omit %	p-Value	Item #	Session	CS	SA	R_ITT	Omit %	p-Value
51	S5	3	7	.41	.1	.67	72	S5	3	7	.51	.3	.77
52	S5	3	7	.38	.1	.80	73	S5	3	7	.47	.6	.82
53	S5	3	7	.33	.2	.87	74	S5	3	8	.42	.3	.66
54	S5	2	5	.27	.2	.36	75	S5	3	8	.50	.5	.73
55	S5	2		.42	.2	.93	76	S5	3	8	.39	.9	.89
56	S5	3	7	.30	.3	.92	77	S5	3	8	.40	1.2	.71
57	S5	3	7	.31	.7	.83	78	S5	3	8	.43	1.4	.74
58	S5	2		.28	.8	.37	79	S5	3	8	.40	1.9	.78
59	S5	2		.43	1.0	.82	80	S5	2	5	.40	.4	.67
61	S5	3	7	.27	.2	.68	81	S5	3	8	.42	.4	.61
62	S5	3	7	.30	.2	.81	82	S5	3	8	.50	.8	.66
63	S5	3	7	.15	.3	.26	83	S5	2	5	.33	.3	.60
64	S5	3	7	.13	3.8	.44	84	S5	2	5	.28	.4	.48
65	S5	3	7	.29	.3	.40	85	S5	2	5	.41	.4	.68
66	S5	2	8	.32	.3	.63	86	S5	2	8	.47	.5	.73
67	S5	3	7	.28	.3	.85	87	S5	2	8	.41	.3	.69
68	S5	3	7	.48	.4	.82	88	S5	2	8	.39	.4	.71
69	S5	3	7	.42	.8	.88	89	S5	2	8	.37	.6	.63
70	S5	3	7	.35	.3	.75	90	S5	2	7	.55	.4	.78
71	S5	3	8	.27	.3	.95	91	S5	2	7	.54	.5	.71

Mean R_ITT: .38
Mean p-value: .70

Table 37. Constructed-Response Item Analyses, Grade 5 Writing

Item #	Part	Session	CS	SA	R_ITT	Percent of Students Obtaining Score Level					Omit %	p-Value
						0	1	2	3	4		
1		S1	2	6	.05	.5	99.5				.0	1.00
2	A	S2	3	8	.43	19.7	79.2				1.0	.79
2	B	S2	3	7	.48	21.0	78.0				1.0	.78
2	C	S2	3	8	.26	8.0	91.0				1.0	.91
2	D	S2	3	7	.42	15.2	83.7				1.0	.84
2	E	S2	3	8	.44	41.2	57.8				1.0	.58
2	F	S2	3	8	.40	43.2	55.7				1.0	.56
3	A	S2	2	6	.57	2.4	6.1	37.4	34.9	19.1	.1	.66
3	B	S2	2	6	.64	.8	7.1	38.2	34.7	19.1	.1	.66
3	C	S2	3	6	.52	1.1	13.6	85.3			.1	.92
27		S3	2	5	.51	1.3	8.7	39.8	38.4	8.1	3.8	.59
60		S5	2	5	.52	.5	5.2	34.0	42.0	17.9	.4	.68
92		S5	2	5	.57	.8	21.3	52.4	21.1	3.7	.7	.51

Mean R_ITT: .46
Mean p-value: .73

Table 38. Multiple-Choice Item Analyses, Grade 5 Mathematics

Item #	Session	CS	SA	R_ITT	Omit %	p-Value	Item #	Session	CS	SA	R_ITT	Omit %	p-Value
1	S1	1	1	.26	.5	.34	35	S2	2		.55	.2	.49
2	S1	6		.39	.2	.56	36	S2	1	1	.45	.6	.74
3	S1	1	1	.23	.9	.86	38	S2	6		.31	.2	.58
4	S1	1		.38	.1	.83	39	S2	6		.32	.9	.90
6	S1	1		.26	.6	.97	40	S2	2		.50	.2	.75
7	S1	4/5		.44	.2	.79	41	S2	6	1	.50	.2	.66
8	S1	6		.42	.2	.79	43	S2	6	1	.60	.7	.73
10	S1	1	1	.47	1.0	.83	44	S2	6		.45	.7	.79
11	S1	2	2	.44	.6	.79	45	S2	4/5		.36	1.0	.59
12	S1	4/5		.34	.5	.65	47	S3	4/5		.30	.1	.83
13	S1	3	3	.44	.5	.76	48	S3	6		.39	.4	.77
14	S1	6		.43	.7	.70	49	S3	1	1	.30	.1	.96
15	S1	2		.47	3.6	.75	50	S3	4/5		.37	.1	.84
16	S1	2	2	.46	.7	.70	52	S3	1	1	.36	.2	.55
17	S1	6		.37	.4	.58	53	S3	3	3	.52	.2	.88
19	S1	3		.31	.7	.26	55	S3	4/5		.28	.3	.50
20	S1	3	3	.46	.7	.79	56	S3	4/5		.31	.3	.68
22	S1	6		.26	1.7	.87	57	S3	1		.47	.3	.59
24	S2	2	2	.29	.1	.91	59	S3	4/5		.50	.3	.69
25	S2	3		.35	.3	.54	60	S3	4/5		.37	.3	.60
26	S2	1	1	.45	.2	.66	62	S3	1	1	.37	.2	.60
27	S2	1	1	.37	.1	.78	63	S3	1		.48	.4	.73
28	S2	6		.56	.6	.71	64	S3	1		.45	.8	.74
30	S2	6	1	.44	.5	.81	65	S3	3		.46	.7	.87
31	S2	4/5		.46	.3	.72	66	S3	2	2	.54	.2	.57
32	S2	4/5		.20	.3	.76	67	S3	2		.38	.2	.81
34	S2	3		.21	.2	.35	68	S3	6		.43	.7	.72

Mean R_ITT: .40
Mean p-value: .71

Table 39. Constructed-Response Item Analyses, Grade 5 Mathematics

Item #	Session	CS	SA	R_ITT	Percent of Students Obtaining Score Level					Omit %	p-Value
					0	1	2	3	4		
5	S1	4/5		.47	20.9	35.0	43.4			.6	.61
9	S1	3	3	.66	2.6	32.2	28.1	36.8		.2	.66
18	S1	1	1	.68	30.2	25.1	22.9	21.1		.8	.45
21	S1	3		.44	10.6	14.3	73.7			1.4	.81
23	S1	2	2	.61	3.2	5.5	17.3	15.8	57.1	1.1	.79
29	S2	1	1	.63	14.9	28.5	51.4			5.1	.66
33	S2	3		.60	6.9	19.8	18.2	54.5		.5	.73
37	S2	2		.58	23.7	36.4	39.2			.7	.57
42	S2	2	2	.54	9.0	13.8	21.6	55.4		.2	.74
46	S2	4/5		.61	1.8	37.8	10.7	19.0	30.4	.2	.59
51	S3	2	2	.49	13.5	25.4	60.9			.2	.74
54	S3	6	1	.55	4.5	11.9	27.1	56.0		.5	.78
58	S3	4/5		.53	22.7	41.5	24.2	10.7		1.0	.41
61	S3	6		.62	21.9	41.8	35.7			.5	.57
69	S3	3	3	.67	7.0	35.7	8.1	30.9	17.9	.4	.54

Mean R_ITT: .58
Mean p-value: .64

Table 40. Multiple-Choice Item Analyses, Grade 6 Reading

Item #	Session	CS	SA	R_ITT	Omit %	p-Value	Item #	Session	CS	SA	R_ITT	Omit %	p-Value
5	S3	6	1/4	.41	.1	.89	37	S4	4	2	.46	.1	.77
6	S3	6	1/4	.45	.6	.85	38	S4	1	3	.42	.1	.66
7	S3	1	3	.45	.1	.88	39	S4	1	3	.50	.1	.82
8	S3	5		.46	.4	.69	42	S4	6	1/4	.44	.3	.87
9	S3	5		.25	.3	.32	43	S4	6	3	.25	.7	.59
10	S3	5		.24	1.1	.39	47	S4	4	2	.40	.3	.80
13	S3	4	2	.35	.5	.78	48	S4	4	2	.34	.8	.66
14	S3	5		.37	.7	.57	49	S4	4	2	.45	.6	.73
15	S3	6	2	.37	.4	.59	95	S6	6	1/4	.36	.2	.69
16	S3	1	3	.30	.5	.41	96	S6	4	1/4	.28	.4	.49
17	S3	1	3	.46	.8	.62	97	S6	4	1/4	.42	.4	.81
18	S3	1	3	.43	1.4	.55	98	S6	1	3	.31	.3	.48
20	S3	5		.36	1.1	.68	99	S6	1	3	.37	.5	.78
21	S3	1	1/4	.46	.9	.64	100	S6	1	3	.59	.5	.80
22	S3	1	1/4	.47	1.2	.76	101	S6	5		.48	.2	.61
23	S3	1	1/4	.39	1.3	.77	102	S6	5		.40	.3	.89
24	S3	4	1/4	.40	1.1	.68	103	S6	5		.33	.5	.52
25	S3	6	1/4	.22	1.2	.64	104	S6	1	1/4	.33	.4	.59
26	S3	1	3	.36	1.2	.32	105	S6	6	1/4	.44	.9	.71
29	S3	4	2	.25	2.9	.59	106	S6	5		.32	.2	.52
30	S3	4	2	.26	3.2	.54	107	S6	5		.54	.4	.74
31	S3	4	2	.18	3.4	.52	108	S6	1	3	.50	.7	.64
32	S3	1	2	.46	3.5	.79	110	S6	5		.42	.5	.73
33	S3	1	2	.44	3.6	.79	111	S6	1	1/4	.12	.5	.30
34	S3	1	3	.51	3.7	.79	112	S6	6	1/4	.46	.6	.63
35	S4	4	2	.42	.1	.85	113	S6	4	1/4	.49	.8	.64
36	S4	4	2	.37	.2	.68	114	S6	1	1/4	.45	1.2	.57

Mean R_ITT: .39
Mean p-value: .64

Table 41. Constructed-Response Item Analyses, Grade 6 Reading

Item #	Session	CS	SA	R_ITT	Percent of Students Obtaining Score Level					Omit %	p-Value
					0	1	2	3	4		
4	S3	1	1/4	.40	3.5	12.7	83.2			.5	.90
11	S3	1	2	.61	13.4	4.4	15.5	16.1	48.9	1.7	.70
12	S3	4	2	.54	35.6	18.7	44.8			.9	.54
19	S3	6	1/4	.60	26.3	12.7	56.1			5.0	.62
27	S3	5		.58	13.2	15.9	30.4	37.0		3.5	.63
28	S3	5		.50	20.2	54.0	17.8			8.1	.45
40	S4	6	2	.32	9.2	80.6	9.0			1.2	.49
41	S4	6	2	.57	33.4	6.4	19.7	18.9	17.3	4.2	.43
44	S4	4	1/4	.55	14.0	39.6	45.0			1.5	.65
45	S4	5		.46	11.7	25.1	62.1			1.1	.75
46	S4	5		.56	11.8	9.3	77.5			1.5	.82
93	S6	6	1/4	.55	10.1	16.5	18.9	53.4		1.1	.72
94	S6	6	1/4	.58	36.0	24.1	37.2			2.7	.49
109	S6	6	1/4	.38	30.4	43.9	19.9			5.8	.42
115	S6	4	1/4	.45	40.6	17.0	20.9	17.8		3.7	.37

Mean R_ITT: .51
Mean p-value: .60

Table 42. Multiple-Choice Item Analyses, Grade 6 Writing

Item #	Session	CS	SA	R_ITT	Omit %	p-Value	Item #	Session	CS	SA	R_ITT	Omit %	p-Value
51	S5	3	8	.33	.1	.77	72	S5	2	5	.46	.2	.90
52	S5	3	8	.44	.1	.73	73	S5	2	5	.50	.2	.68
53	S5	2		.43	.7	.77	74	S5	2	5	.36	.3	.76
54	S5	2		.40	1.0	.87	75	S5	2	5	.29	.4	.49
55	S5	3	7	.39	.2	.70	76	S5	2	5	.32	.3	.44
56	S5	3	7	.32	.3	.49	77	S5	2	5	.42	.4	.72
57	S5	3	7	.26	1.0	.44	78	S5	2	5	.47	.4	.74
58	S5	3	7	.42	.3	.65	79	S5	3	8	.43	.5	.69
59	S5	3	7	.46	.7	.69	80	S5	3	8	.24	.9	.34
60	S5	3	7	.40	.9	.62	81	S5	2	8	.40	.3	.71
61	S5	3	7	.43	.2	.84	82	S5	2	8	.41	.5	.68
62	S5	3	7	.32	.2	.89	83	S5	2	8	.53	.4	.76
63	S5	3	7	.21	.4	.27	84	S5	2	8	.46	5.0	.64
64	S5	3	7	.16	.5	.55	85	S5	2	7	.55	.4	.82
65	S5	3	7	.40	.2	.73	86	S5	2	7	.56	.6	.76
66	S5	3	7	.35	.3	.67	87	S5	3	7	.57	.6	.76
67	S5	3	7	.33	.5	.80	88	S5	3	7	.52	1.5	.75
68	S5	3	7	.31	1.6	.40	89	S5	3	7	.30	.3	.63
69	S5	2	5	.43	.1	.83	90	S5	3	7	.40	.6	.75
71	S5	2	5	.46	.2	.74	91	S5	3	7	.42	.4	.65

Mean R_ITT: .40
Mean p-value: .68

Table 43. Constructed-Response Item Analyses, Grade 6 Writing

Item #	Part	Session	CS	SA	R_ITT	Percent of Students Obtaining Score Level					Omit %	p-Value
						0	1	2	3	4		
1		S1	2	6	.07	.7	99.3				.0	.99
2	A	S2	3	8	.24	86.0	11.4				2.7	.11
2	B	S2	3	8	.34	41.1	55.7				3.1	.56
2	C	S2	3	8	.43	41.5	55.5				3.0	.56
2	D	S2	3	7	.26	27.7	69.0				3.3	.69
2	E	S2	3	8	.21	57.4	37.9				4.7	.38
2	F	S2	3	8	.18	11.7	81.0				7.2	.81
3	A	S2	2	6	.53	2.0	7.3	37.5	38.5	14.5	.1	.64
3	B	S2	2	6	.58	.3	8.5	43.1	34.8	13.2	.1	.63
3	C	S2	3	6	.50	.4	15.6	83.9			.1	.92
50		S4	2	5	.58	2.1	8.9	42.2	41.8	3.9	1.1	.59
70		S5	2	5	.60	.8	10.2	44.7	38.7	4.8	.8	.59
92		S5	2	5	.60	.6	16.8	48.1	30.0	3.6	.9	.54

Mean R_ITT: .41
Mean p-value: .62

Table 44. Multiple-Choice Item Analyses, Grade 6 Mathematics

Item #	Session	CS	SA	R_ITT	Omit %	p-Value	Item #	Session	CS	SA	R_ITT	Omit %	p-Value
1	S1	1		.38	.3	.41	30	S2	3		.43	1.3	.72
2	S1	4/5	3	.40	.2	.62	32	S2	1		.54	.4	.52
3	S1	3	2	.35	.2	.77	33	S2	2	2	.38	.3	.43
4	S1	1		.31	.2	.87	34	S2	3		.42	.2	.52
6	S1	4/5	3	.34	.4	.48	36	S2	4/5		.38	.3	.64
8	S1	6	1	.20	1.0	.88	37	S2	6		.30	.3	.66
9	S1	2	2	.44	.2	.79	38	S2	2	2	.29	.2	.91
10	S1	1		.34	.6	.25	41	S3	6		.50	.2	.57
11	S1	6	1	.36	.9	.66	42	S3	6	1	.47	.2	.40
13	S1	3		.41	.6	.37	43	S3	1	1	.38	.4	.40
14	S1	2	2	.42	.4	.81	45	S3	4/5		.38	.2	.69
16	S1	2	2	.47	.7	.34	46	S3	6		.36	.4	.82
17	S1	4/5	3	.37	.7	.70	48	S3	1	1	.34	.2	.69
18	S1	1	1	.42	.9	.79	50	S3	1	1	.35	.2	.69
19	S1	6	1	.35	.9	.82	51	S3	4/5	3	.36	.4	.60
21	S2	4/5	3	.42	.2	.69	52	S3	4/5		.31	.2	.88
22	S2	1	1	.42	.1	.62	53	S3	4/5	3	.45	.4	.69
23	S2	6	1	.41	5.8	.62	55	S3	2		.23	.4	.45
25	S2	6	1	.43	.7	.31	56	S3	3		.32	.4	.35
26	S2	4/5		.35	1.0	.68	57	S3	2	2	.43	.4	.84
27	S2	3		.26	.7	.39	58	S3	6	1	.39	2.3	.69
28	S2	4/5	3	.41	.8	.63	59	S3	4/5		.35	.6	.84
29	S2	3		.33	.8	.53							

Mean R_ITT: .38
Mean p-value: .62

Table 45. Constructed-Response Item Analyses, Grade 6 Mathematics

Item #	Session	CS	SA	R_ITT	Percent of Students Obtaining Score Level					Omit %	p-Value
					0	1	2	3	4		
5	S1	3		.64	1.6	23.5	27.7	47.0		.2	.73
7	S1	4/5	3	.51	10.0	33.2	56.1			.6	.73
12	S1	3		.56	26.7	31.7	40.6			1.0	.56
15	S1	1	1	.71	23.3	22.7	24.1	29.0		.9	.53
20	S1	2		.62	2.5	6.7	17.2	28.6	43.9	1.0	.76
24	S2	3		.44	72.1	11.2	15.8			.9	.21
31	S2	4/5		.54	21.2	40.1	24.4	12.7		1.6	.42
35	S2	2	2	.48	9.0	10.2	15.7	64.8		.3	.79
39	S2	4/5		.40	79.1	17.2	2.4			1.3	.11
40	S2	1	1	.70	20.4	42.9	14.8	11.7	9.6	.6	.37
44	S3	6	1	.64	11.7	41.2	46.6			.5	.67
47	S3	2	2	.54	9.9	33.3	48.6	7.8		.4	.51
49	S3	6	1	.53	47.3	26.3	25.5			1.0	.39
54	S3	4/5	3	.58	30.0	24.2	35.5	9.6		.6	.41
60	S3	3		.70	4.2	29.5	12.3	28.1	25.3	.7	.60

Mean R_ITT: .58
Mean p-value: .52

Table 46. Multiple-Choice Item Analyses, Grade 7 Reading

Item #	Session	CS	SA	R_ITT	Omit %	p-Value	Item #	Session	CS	SA	R_ITT	Omit %	p-Value
4	S3	6	1	.34	.1	.96	43	S4	1	3	.41	.7	.55
5	S3	1	1	.40	.1	.82	44	S4	5		.17	1.0	.11
6	S3	6	1	.34	.2	.74	46	S4	4	2	.33	.5	.72
7	S3	5		.38	.1	.90	47	S4	1	2	.39	.5	.83
10	S3	4	2	.34	.3	.81	48	S4	4	2	.39	.6	.65
11	S3	5		.37	.4	.62	49	S4	4	2	.42	.6	.75
12	S3	6	2	.43	.2	.69	50	S4	1	1	.38	.5	.57
13	S3	1	3	.38	.2	.51	51	S4	6	1	.54	.5	.81
14	S3	1	3	.48	.3	.72	52	S4	4	1	.36	.9	.43
15	S3	1	3	.43	2.1	.63	53	S4	1	1	.49	.9	.85
16	S3	1	3	.30	2.4	.40	54	S4	5		.47	.7	.64
17	S3	5		.51	.3	.85	55	S4	5		.21	1.0	.59
18	S3	6	4	.22	.2	.41	100	S6	6	1	.39	.1	.75
19	S3	4	4	.43	.3	.60	101	S6	4	1	.39	.2	.60
20	S3	5		.50	.5	.74	102	S6	4	1	.41	.2	.85
21	S3	5		.37	.7	.81	103	S6	1	3	.33	.2	.54
24	S3	4	2	.48	.7	.77	104	S6	1	3	.37	.3	.82
26	S3	5		.50	1.0	.67	105	S6	1	3	.58	.5	.85
28	S3	5		.38	1.7	.45	106	S6	1	2	.49	.1	.92
29	S3	5		.33	2.4	.44	107	S6	1	3	.55	.2	.79
32	S4	1	3	.41	.2	.70	108	S6	1	2	.43	.5	.72
33	S4	4	1	.44	.1	.75	109	S6	5		.37	1.2	.84
34	S4	1	3	.37	.1	.45	112	S6	4	4	.50	.6	.75
35	S4	1	3	.42	.2	.92	113	S6	5		.46	.2	.59
36	S4	4	1	.44	.4	.91	114	S6	5		.18	.2	.67
39	S4	4	2	.56	.3	.83	115	S6	1	3	.42	.8	.60
40	S4	1	2	.49	1.4	.73	116	S6	6	4	.25	.9	.48
41	S4	5		.10	.6	.42	117	S6	6	4	.33	.9	.51
42	S4	1	3	.10	.6	.39							

Mean R_ITT: .39
Mean p-value: .67

Table 47. Constructed-Response Item Analyses, Grade 7 Reading

Item #	Session	CS	SA	R_ITT	Percent of Students Obtaining Score Level					Omit %	p-Value
					0	1	2	3	4		
8	S3	1	2	.58	9.8	4.5	14.0	15.9	54.5	1.2	.75
9	S3	4	2	.55	28.3	19.5	51.4			.8	.61
22	S3	6	4	.49	22.8	22.1	25.7	25.6		3.8	.50
23	S3	6	4	.49	20.4	26.4	49.7			3.6	.63
25	S3	5		.38	55.4	26.7	13.2			4.7	.27
27	S3	1	2	.50	9.2	27.6	58.7			4.6	.73
31	S4	5		.47	23.8	28.7	45.5			2.0	.60
37	S4	6	1	.35	33.8	35.7	13.7	10.6		6.2	.32
38	S4	1	1	.50	11.0	.5	28.5	58.0		2.0	.77
45	S4	5		.39	12.6	10.8	75.6			1.0	.81
98	S6	6	1	.53	10.1	16.9	18.8	53.1		1.2	.71
99	S6	6	1	.57	31.5	23.8	41.8			3.0	.54
110	S6	4	4	.54	8.4	15.2	74.9			1.5	.83
111	S6	4	4	.32	24.2	59.5	13.6			2.7	.43

Mean R_ITT: .48
Mean p-value: .61

Table 48. Multiple-Choice Item Analyses, Grade 7 Writing

Item #	Session	CS	SA	R_ITT	Omit %	p-Value	Item #	Session	CS	SA	R_ITT	Omit %	p-Value
56	S5	3	7	.15	.1	.45	77	S5	3	8	.35	.4	.78
57	S5	3	7	.44	.1	.86	78	S5	3	8	.20	1.6	.65
58	S5	3	7	.37	.5	.87	79	S5	3	8	.28	1.8	.94
59	S5	3	8	.34	1.3	.78	80	S5	3	8	.34	2.0	.86
60	S5	3	8	.44	1.6	.73	81	S5	2	8	.31	.2	.44
61	S5	3	8	.21	.2	.45	82	S5	2	8	.40	.4	.50
62	S5	3	7	-.01	.5	.07	83	S5	2	8	.48	.6	.75
63	S5	3	7	.47	.6	.63	84	S5	3	7	.40	.3	.71
64	S5	3	7	.31	.5	.69	85	S5	2	5	.28	.3	.51
65	S5	2		.08	.6	.30	86	S5	2	5	.50	.3	.61
66	S5	2		.47	.4	.82	87	S5	2	5	.28	.4	.60
67	S5	3	7	.43	.2	.86	88	S5	2	5	.34	.2	.49
68	S5	3	7	.35	.5	.91	89	S5	2	5	.41	.3	.74
69	S5	2	7	.44	.7	.80	90	S5	2		.41	.2	.72
71	S5	2	5	.34	.2	.72	91	S5	2		.49	.3	.84
72	S5	2	7	.36	.2	.46	92	S5	2		.43	.3	.54
73	S5	3	7	.30	.2	.66	93	S5	2		.43	.2	.87
74	S5	3	7	.40	.4	.79	94	S5	2		.47	.2	.68
75	S5	3	8	.39	.2	.78	95	S5	2		.43	.4	.55
76	S5	3	8	.33	.3	.53	96	S5	2		.26	.4	.30

Mean R_ITT: .36
Mean p-value: .67

Table 49. Constructed-Response Item Analyses, Grade 7 Writing

Item #	Part	Session	CS	SA	R_ITT	Percent of Students Obtaining Score Level					Omit %	p-Value
						0	1	2	3	4		
1		S1	2	6	.10	.9	99.1				.0	.99
2	A	S2	3	8	.36	70.9	27.9				1.2	.28
2	B	S2	3	8	.30	55.4	43.4				1.2	.43
2	C	S2	3	8	.41	57.2	41.6				1.2	.42
2	D	S2	3	7	.44	42.5	56.3				1.2	.56
2	E	S2	3	8	.49	50.1	48.8				1.2	.49
2	F	S2	3	7	.30	65.3	33.6				1.1	.34
3	A	S2	2	6	.59	1.5	9.4	33.8	39.2	15.9	.1	.65
3	B	S2	2	6	.62	.2	11.2	35.4	37.7	15.4	.1	.64
3	C	S2	3	6	.51	.2	23.6	76.1			.1	.88
30		S3	2	5	.58	.6	13.2	39.6	37.4	5.4	3.9	.57
70		S5	2	5	.62	1.2	8.8	33.4	46.1	9.6	.9	.63
97		S5	2	5	.61	.3	12.5	41.0	39.8	5.1	1.2	.59

Mean R_ITT: .47
Mean p-value: .57

Table 50. Multiple-Choice Item Analyses, Grade 7 Mathematics

Item #	Session	CS	SA	R_ITT	Omit %	p-Value	Item #	Session	CS	SA	R_ITT	Omit %	p-Value
1	S1	3		.38	.2	.57	32	S2	1/6	1	.39	2.3	.70
2	S1	4/5		.34	.1	.62	33	S2	2		.41	.3	.48
4	S1	4/5		.38	.4	.66	35	S2	1/6		.39	.4	.66
5	S1	2		.33	.2	.58	36	S2	1/6		.34	1.2	.38
6	S1	4/5	2	.40	.3	.54	37	S2	4/5		.43	.5	.72
8	S1	3		.36	.5	.46	38	S2	4/5		.34	.6	.61
9	S1	1/6		.36	.4	.42	39	S2	1/6		.42	1.2	.30
10	S1	1/6		.20	.8	.89	41	S3	2		.41	.1	.83
12	S1	2		.31	.7	.40	42	S3	1/6	1	.43	.1	.64
13	S1	1/6		.33	.4	.53	44	S3	3		.33	.2	.53
14	S1	4/5	2	.49	1.2	.54	45	S3	4/5		.37	.9	.60
16	S1	1/6	1	.55	.8	.59	46	S3	2		.40	.3	.53
17	S1	4/5		.29	.7	.52	47	S3	1/6		.50	.4	.45
18	S1	3		.38	.6	.42	48	S3	3		.45	.3	.64
19	S1	1/6		.33	.8	.83	50	S3	1/6	1	.41	.4	.44
21	S2	2		.40	.2	.72	51	S3	1/6	1	.48	.5	.72
22	S2	2		.39	.2	.64	52	S3	2		.51	.5	.40
23	S2	3		.39	.2	.60	54	S3	4/5		.19	.4	.80
25	S2	4/5		.18	.3	.30	55	S3	4/5	2	.38	.3	.39
27	S2	3		.36	.6	.23	56	S3	3		.31	.5	.64
28	S2	4/5		.40	.7	.53	58	S3	2		.46	.4	.60
29	S2	3		.17	.3	.96	59	S3	1/6		.60	.8	.53
30	S2	1/6	1	.54	.3	.62							

Mean R_ITT: .39
Mean p-value: .56

Table 51. Constructed-Response Item Analyses, Grade 7 Mathematics

Item #	Session	CS	SA	R_ITT	Percent of Students Obtaining Score Level					Omit %	p-Value
					0	1	2	3	4		
3	S1	2		.50	61.8	12.3	22.5			3.4	.29
7	S1	1/6	1	.51	69.8	17.1	3.8	7.6		1.7	.16
11	S1	3		.49	74.8	15.4	8.4			1.4	.16
15	S1	4/5	2	.59	22.9	21.2	40.9	14.3		.8	.49
20	S1	1/6		.56	10.3	60.3	19.8	4.2	4.4	1.0	.33
24	S2	3		.66	28.4	22.0	24.7	23.1		1.8	.47
26	S2	4/5	2	.47	70.8	23.8	3.5			1.9	.15
31	S2	4/5		.56	67.5	9.0	20.6			2.9	.25
34	S2	2		.61	8.1	19.1	37.6	34.5		.7	.66
40	S2	4/5	2	.71	14.0	35.1	22.0	16.9	10.4	1.6	.43
43	S3	1/6		.51	61.0	6.0	30.8			2.2	.34
49	S3	4/5		.65	33.6	16.6	22.4	24.6		2.8	.45
53	S3	1/6	1	.58	45.1	21.9	31.7			1.3	.43
57	S3	2		.56	8.9	30.9	47.9	11.8		.5	.54
60	S3	3		.63	34.9	24.4	18.6	17.5	2.3	2.4	.31

Mean R_ITT: .58
Mean p-value: .36

Table 52. Multiple-Choice Item Analyses, Grade 8 Reading

Item #	Session	CS	SA	R_ITT	Omit %	p-Value	Item #	Session	CS	SA	R_ITT	Omit %	p-Value
4	S3	6	1	.21	.1	.66	41	S4	4	2	.33	.5	.76
5	S3	6	1	.29	.2	.55	42	S4	6	4	.33	.2	.86
6	S3	6	1	.42	.1	.85	43	S4	6	4	.39	.3	.63
7	S3	6	1	.43	.1	.77	44	S4	6	3	.44	.5	.66
8	S3	5		.38	.1	.77	46	S4	1	3	.34	.4	.42
9	S3	5		.28	.2	.49	91	S6	1	2	.32	.1	.69
10	S3	5		.35	.2	.52	92	S6	1	2	.43	.1	.84
11	S3	1	2	.37	.1	.93	93	S6	1	2	.47	.5	.72
12	S3	1	2	.44	.2	.92	94	S6	1	3	.34	.6	.45
13	S3	4	2	.43	.4	.92	95	S6	5		.14	.3	.23
14	S3	5		.39	.6	.50	96	S6	5		.44	.3	.89
19	S3	4	2	.41	.4	.90	99	S6	5		.11	.7	.29
20	S3	4	2	.52	.4	.86	100	S6	4	4	.32	.3	.52
21	S3	4	2	.45	.9	.92	101	S6	6	4	.25	.2	.43
22	S3	4	2	.26	.4	.86	102	S6	6	4	.21	.3	.36
23	S3	6	2	.46	.5	.73	103	S6	5		.20	.3	.34
24	S3	1	3	.44	.6	.72	105	S6	4	2	.44	.2	.54
27	S4	1	3	.43	.2	.74	106	S6	4	2	.33	.3	.52
28	S4	4	1	.45	.1	.79	107	S6	1	3	.20	.5	.38
29	S4	1	3	.42	.2	.53	108	S6	5		.45	.3	.81
30	S4	1	3	.42	.2	.93	109	S6	5		.40	.3	.70
31	S4	4	1	.41	.3	.93	112	S6	4	1	.45	.6	.74
34	S4	4	2	.54	.1	.88	113	S6	4	1	.54	.7	.87
35	S4	1	2	.48	1.1	.78	114	S6	6	1	.43	.7	.92
36	S4	5		.14	.3	.46	115	S6	4	1	.44	.8	.80
37	S4	1	3	.17	.3	.41	116	S6	6	1	.49	.8	.64
38	S4	1	3	.45	.4	.63	117	S6	1	3	.04	.8	.59
39	S4	5		.24	.6	.15							

Mean R_ITT (item 117 excluded): .37
Mean p-value (item 117 excluded): .66

Table 53. Constructed-Response Item Analyses, Grade 8 Reading

Item #	Session	CS	SA	R_ITT	Percent of Students Obtaining Score Level					Omit %	p-Value
					0	1	2	3	4		
15	S3	1	2	.57	14.0	18.6	65.8			1.7	.75
16	S3	1	2	.56	23.6	30.0	42.6			3.7	.58
17	S3	1	2	.49	13.6	18.7	66.2			1.6	.76
18	S3	1	2	.52	10.8	25.3	61.7			2.2	.74
26	S4	5		.49	19.2	29.7	48.8			2.3	.64
32	S4	6	1	.34	29.8	39.3	12.8	10.4		7.7	.32
33	S4	1	1	.48	10.1	.3	29.2	58.2		2.2	.78
40	S4	5		.39	10.5	9.3	79.3			.9	.84
45	S4	6	4	.63	11.4	33.7	35.9	14.2		4.9	.49
90	S6	4	2	.50	7.4	52.1	38.2			2.2	.64
97	S6	1	4	.50	11.6	40.9	18.4	25.4		3.7	.51
98	S6	4	4	.59	8.7	16.5	25.3	32.2	11.7	5.5	.53
104	S6	6	4	.56	42.3	32.2	18.3			7.3	.34
110	S6	5		.56	46.8	23.5	26.9			2.8	.39
111	S6	6	1	.37	62.8	19.8	11.0			6.4	.21

Mean R_ITT: .51
Mean p-value: .57

Table 54. Multiple-Choice Item Analyses, Grade 8 Writing

Item #	Session	CS	SA	R_ITT	Omit %	p-Value	Item #	Session	CS	SA	R_ITT	Omit %	p-Value
47	S5	3	8	.31	.2	.81	69	S5	3	7	.31	.7	.60
48	S5	3	8	.30	.3	.44	70	S5	3	8	.37	.4	.30
49	S5	2	7	.42	.2	.86	71	S5	3	8	.44	.4	.53
50	S5	2	7	.34	3.6	.82	72	S5	2	5	.24	.3	.50
51	S5	2		.39	1.1	.69	73	S5	2	5	.45	.3	.69
52	S5	3	7	.41	.2	.82	74	S5	2	5	.26	.2	.56
53	S5	3	7	.34	.4	.63	75	S5	2	5	.38	.2	.60
54	S5	3	7	.53	.1	.74	76	S5	2	5	.51	.2	.63
55	S5	2	8	.47	.2	.85	77	S5	2	5	.49	.3	.75
56	S5	3	7	.46	.6	.79	78	S5	3	8	.21	.2	.46
57	S5	2	5	.32	.3	.70	79	S5	2	8	.53	.3	.81
59	S5	2	8	.48	.3	.66	80	S5	2	8	.41	.4	.56
60	S5	3	7	.55	.1	.85	81	S5	2	8	.58	.3	.83
61	S5	3	7	.48	.8	.88	82	S5	2	7	.37	.3	.51
62	S5	2	5	.21	2.0	.42	83	S5	3	7	.53	.4	.83
63	S5	3	7	.52	.2	.78	84	S5	3	7	.57	.3	.83
64	S5	3	7	.51	.3	.76	85	S5	3	7	.51	1.2	.82
65	S5	3	7	.46	.3	.79	86	S5	2	5	.35	.5	.67
66	S5	2	5	.44	.3	.90	87	S5	2	5	.47	.3	.78
67	S5	3	7	.03	.5	.08	88	S5	2	5	.42	.3	.67
68	S5	3	7	.45	.6	.64							

Mean R_ITT: .42
Mean p-value: .68

Table 55. Constructed-Response Item Analyses, Grade 8 Writing

Item #	Part	Session	CS	SA	R_ITT	Percent of Students Obtaining Score Level					Omit %	p-Value
						0	1	2	3	4		
1		S1	2	6	.15	1.2	98.8				.0	.99
2	A	S2	3	7	.02	52.5	47.4				.1	.47
2	B	S2	3	8	.28	86.0	14.0				.1	.14
2	C	S2	3	7	.43	46.7	53.2				.1	.53
2	D	S2	3	8	.41	38.0	61.9				.1	.62
2	E	S2	3	8	.44	22.6	77.3				.1	.77
2	F	S2	3	8	.40	30.9	69.0				.1	.69
3	A	S2	2	6	.61	1.5	3.6	24.5	52.9	17.1	.3	.70
3	B	S2	2	6	.65	.2	4.2	28.0	53.5	13.8	.3	.69
3	C	S2	3	6	.43	.3	6.8	92.5			.3	.96
25		S3	2	5	.61	.5	9.4	47.1	35.7	5.3	2.0	.58
58		S5	2	5	.55	.3	8.1	52.2	32.4	5.7	1.3	.58
89		S5	2	5	.57	.3	10.5	57.1	27.1	3.4	1.6	.55

Mean R_ITT (item 2A excluded): .47
Mean p-value (item 2A excluded): .65

Table 56. Multiple-Choice Item Analyses, Grade 8 Mathematics

Item #	Session	CS	SA	R_ITT	Omit %	p-Value	Item #	Session	CS	SA	R_ITT	Omit %	p-Value
1	S1	1/6		.19	.3	.10	32	S2	1/6	2	.27	.4	.30
2	S1	1/6		.53	.2	.52	33	S2	4/5	2	.32	.4	.37
4	S1	4/5	2	.33	.3	.20	35	S2	1/6		.50	.3	.38
5	S1	3		.19	2.2	.88	36	S2	4/5	3	.17	.6	.27
6	S1	3	2	.41	.1	.58	37	S2	1/6		.43	.8	.55
7	S1	3		.43	.8	.50	38	S2	1/6		.55	1.4	.53
8	S1	2		.15	.7	.57	39	S2	2	1	.42	1.1	.42
10	S1	2		.41	.5	.47	41	S3	2	2	.31	.3	.28
11	S1	2	2	.37	.2	.56	42	S3	4/5	3	.31	.3	.66
13	S1	4/5		.56	.6	.38	44	S3	4/5		.39	.2	.70
15	S1	1/6		.03	.8	.15	45	S3	1/6		.41	.4	.46
16	S1	4/5	3	.44	.8	.37	46	S3	3		.36	.2	.81
17	S1	4/5		.33	.3	.62	47	S3	1/6		.55	.3	.73
18	S1	1/6		.49	.7	.63	49	S3	3		.31	.6	.65
19	S1	4/5	2	.39	.4	.53	51	S3	4/5		.50	.2	.27
21	S2	3		.33	.2	.26	52	S3	2		.47	.3	.73
22	S2	2	1	.21	.2	.35	54	S3	1/6		.40	.5	.66
23	S2	2	1	.51	.4	.50	55	S3	4/5	2	.44	.2	.57
24	S2	2		.25	.2	.73	56	S3	3		.24	.2	.55
26	S2	3	2	.48	.1	.65	57	S3	2	1	.53	.6	.49
27	S2	3		.26	.3	.16	58	S3	4/5	3	.19	.2	.81
29	S2	4/5	3	.13	.5	.25	59	S3	2	1	.19	.3	.38
31	S2	4/5		.47	.9	.50							

Mean R_ITT (item 15 excluded): .37
Mean p-value (item 15 excluded): .50

Table 57. Constructed-Response Item Analyses, Grade 8 Mathematics

Item #	Session	CS	SA	R_ITT	Percent of Students Obtaining Score Level					Omit %	p-Value
					0	1	2	3	4		
3	S1	2		.59	42.6	16.3	37.6			3.5	.46
9	S1	3		.67	19.7	26.2	14.0	38.6		1.6	.57
12	S1	1/6	2	.59	70.7	10.2	16.2			2.9	.21
14	S1	1/6	2	.44	30.4	54.8	8.9	3.3		2.5	.28
20	S1	1/6	2	.69	64.3	11.8	9.7	4.5	7.8	1.9	.19
25	S2	2	1	.44	81.7	5.2	4.8			8.2	.07
28	S2	3		.65	16.8	21.1	24.6	35.5		2.1	.59
30	S2	4/5	2	.53	64.1	4.2	27.9			3.8	.30
34	S2	4/5	3	.64	43.1	38.3	10.7	5.4		2.4	.25
40	S2	4/5	3	.71	11.5	31.7	21.7	15.5	17.8	1.8	.48
43	S2	4/5	2	.53	67.5	23.5	6.9			2.2	.19
48	S3	4/5		.63	19.6	14.6	29.5	33.9		2.4	.58
50	S3	1/6	2	.49	61.0	5.7	30.2			3.1	.33
53	S3	2	1	.69	40.2	23.0	18.7	14.4		3.8	.35
60	S3	2	1	.65	3.6	25.0	48.0	6.0	16.6	.8	.51

Mean R_ITT: .60
Mean p-value: .36

Table 58. Multiple-Choice Item Analyses, Grade 8 Science

Item #	Session	CS	SA	R_ITT	Omit %	p-Value	Item #	Session	CS	SA	R_ITT	Omit %	p-Value
1	S1	3		.06	.1	.36	45	S2	5/6		.45	.2	.60
3	S1	3		.20	.3	.32	46	S2	4	5	.39	.3	.60
4	S1	3		.19	.2	.44	47	S2	1	1	.30	.7	.35
5	S1	3		.34	.1	.55	48	S2	1	2	.19	.1	.67
6	S1	3		.37	.1	.69	49	S2	4	5	.43	.3	.54
8	S1	2	3	.26	.1	.36	50	S2	5/6		.40	.3	.88
9	S1	2	3	.44	.2	.81	52	S2	4	5	.14	.3	.38
10	S1	4		.36	.4	.47	53	S2	3		.10	.2	.39
11	S1	2	3	.44	.4	.71	54	S2	1	2	.54	.2	.79
12	S1	4	5	.21	.2	.48	55	S2	1	1	.35	3.7	.84
13	S1	4		.18	.2	.41	56	S3	3		.36	.1	.67
14	S1	4	5	.35	.2	.55	57	S3	3		.27	.1	.49
15	S1	4	5	.36	.1	.89	58	S3	2	4	.00	.2	.14
16	S1	4	5	.36	.2	.86	59	S3	4		.32	.1	.75
17	S1	2	4	.26	.2	.51	60	S3	1	2	.35	.2	.83
18	S1	5/6		.32	.2	.55	61	S3	4		.27	.7	.45
25	S2	3		.12	.2	.52	63	S3	2	3	.41	.6	.47
26	S2	1	1	.26	.2	.64	64	S3	1	2	.31	.2	.94
27	S2	1	2	.39	.1	.87	65	S3	1	2	.37	.4	.88
29	S2	4	5	.23	.2	.23	67	S3	5/6		.50	.5	.86
31	S2	1	1	.49	.2	.55	68	S3	4		.25	.2	.79
32	S2	4	5	.41	.4	.44	69	S3	2	3	.35	.2	.44
34	S2	2	4	.30	.4	.47	70	S3	4	5	.33	.3	.57
35	S2	3		.24	.2	.76	71	S3	4	5	.27	.3	.24
36	S2	5/6		.17	.4	.49	74	S3	5/6		.13	.3	.31
37	S2	2	3	.26	.4	.49	76	S3	2	4	.45	.2	.75
38	S2	2	4	.33	.8	.51	78	S3	1	1	.43	.2	.73
39	S2	2	4	.29	.6	.42	79	S3	2	4	.23	.2	.24
41	S2	2	4	.14	.3	.30	80	S3	2	3	.40	.1	.79
42	S2	2	3	.10	.3	.41	81	S3	1	1	.40	.3	.73
43	S2	5/6		.36	.2	.84	82	S3	5/6		.39	.2	.60
44	S2	3		.33	.2	.74	83	S3	1	2	.41	.1	.83

Mean R_ITT: .31
Mean p-value: .58

Table 59. Constructed-Response Item Analyses, Grade 8 Science

Item #	Session	CS	SA	R_ITT	Percent of Students Obtaining Score Level					Omit %	p-Value
					0	1	2	3	4		
2	S1	1		.55	42.8	41.3	6.2			9.6	.27
7	S1	2	4	.49	36.7	30.1	23.9			9.3	.39
19	S1	1	1	.46	49.6	49.5				.9	.50
20	S1	1	2	.55	20.4	11.1	14.6	28.7	23.2	1.9	.55
21	S1	1	2	.52	34.0	62.7				3.4	.63
22	S1	1	2	.30	39.6	58.3				2.1	.58
23	S1	1	1	.39	29.2	67.8				3.0	.68
24	S1	1	1	.59	21.1	20.3	22.4	31.7		4.6	.53
28	S2	2	3	.56	48.9	31.4	18.0			1.7	.34
30	S2	5/6		.35	77.6	21.2				1.2	.21
33	S2	4	5	.55	37.0	52.5	5.8			4.7	.32
40	S2	3		.62	22.6	27.1	48.5			1.8	.62
51	S2	3		.30	75.7	13.5	1.4			9.4	.08
62	S3	4	5	.52	43.8	40.4	6.4			9.4	.27
66	S3	3		.39	5.9	14.3	78.0			1.8	.85
72	S3	3		.53	56.2	30.6	7.4			5.8	.23
73	S3	4	5	.30	71.2	21.8	1.1			5.8	.12
75	S3	1	2	.41	18.1	79.0				2.9	.79
77	S3	3		.38	49.4	48.0				2.7	.48
84	S3	5/6		.35	53.6	43.6				2.9	.44

Mean R_ITT: .46
Mean p-value: .44

Table 60. Multiple-Choice Item Analyses, Grade 9 Reading

Item #	Session	CS	SA	R_ITT	Omit %	p-Value	Item #	Session	CS	SA	R_ITT	Omit %	p-Value
4	S3	6	1	.23	.4	.55	41	S4	1	2	.41	.6	.72
5	S3	6	1	.37	1.7	.70	43	S4	1	3	.53	.5	.82
6	S3	1	3	.41	.1	.93	44	S4	1	3	.40	1.9	.68
7	S3	4	1	.43	.2	.70	46	S4	5		.43	.3	.69
8	S3	6	1	.47	.2	.69	47	S4	5		.22	.5	.28
11	S3	6	1	.31	.4	.75	91	S6	5		.37	.3	.75
12	S3	6	1	.35	.5	.64	92	S6	6	2	.29	.2	.54
13	S3	1	3	.40	.2	.48	93	S6	4	2	.53	.5	.84
14	S3	1	3	.49	.3	.76	94	S6	4	2	.53	.3	.82
15	S3	1	3	.22	.4	.36	95	S6	5		.14	1.6	.47
16	S3	4	1	.45	.4	.84	96	S6	6	1	.20	.3	.60
17	S3	1	3	.23	1.0	.36	97	S6	6	1	.40	.2	.75
18	S3	5		.43	.3	.78	98	S6	1	3	.50	.2	.80
19	S3	5		.35	.9	.61	99	S6	5		.29	.4	.58
20	S3	5		.27	.3	.56	101	S6	5		.28	.2	.42
21	S3	5		.48	.3	.67	102	S6	1	3	.15	.4	.34
26	S3	4	2	.46	1.7	.86	104	S6	1	2	.46	.3	.73
27	S3	4	2	.53	1.8	.85	105	S6	1	2	.47	.3	.84
28	S3	4	2	.52	1.9	.89	106	S6	4	2	.48	.2	.85
29	S3	4	2	.33	2.0	.84	107	S6	4	2	.54	.3	.83
30	S4	6	1	.34	.1	.90	108	S6	5		.39	.3	.75
31	S4	1	3	.24	.1	.37	109	S6	5		.39	.3	.60
32	S4	4	1	.37	.1	.77	110	S6	1	3	.42	.2	.63
33	S4	6	1	.41	.4	.68	111	S6	5		.49	.4	.66
34	S4	5		.41	.3	.83	114	S6	5		.20	1.0	.31
35	S4	6	4	.43	.1	.70	115	S6	4	4	.29	.4	.54
36	S4	6	4	.44	.2	.72	116	S6	6	4	.25	.4	.44
37	S4	1	3	.25	.2	.71	117	S6	6	4	.24	.3	.38
38	S4	4	4	.45	.2	.68	118	S6	5		.22	.5	.37

Mean R_ITT: .38
Mean p-value: .65

Table 61. Constructed-Response Item Analyses, Grade 9 Reading

Item #	Session	CS	SA	R_ITT	Percent of Students Obtaining Score Level					Omit %	p-Value
					0	1	2	3	4		
9	S3	6	1	.38	2.9	16.5	43.1	35.8		1.7	.70
10	S3	6	1	.54	8.6	11.4	23.8	51.2		5.0	.71
22	S3	1	2	.62	12.1	16.2	67.4			4.2	.76
23	S3	1	2	.57	25.5	28.8	37.5			8.3	.52
24	S3	1	2	.55	14.2	17.5	63.4			5.0	.72
25	S3	1	2	.57	14.7	29.3	50.2			5.7	.65
39	S4	6	4	.50	22.1	43.1	28.1			6.7	.50
40	S4	4	4	.56	31.8	31.3	26.8			10.1	.42
42	S4	4	2	.50	16.6	36.2	24.4	17.7		5.0	.46
45	S4	4	2	.62	15.3	42.0	33.4			9.2	.54
100	S6	1	1	.49	46.0	14.5	21.6	10.0		7.9	.29
103	S6	5		.62	8.8	13.4	24.5	48.3		5.1	.69
112	S6	1	4	.55	7.5	35.6	19.0	30.9		7.0	.55
113	S6	4	4	.61	5.7	12.0	20.2	36.8	15.8	9.6	.57
119	S6	6	4	.58	37.3	32.4	21.2			9.1	.37

Mean R_ITT: .55
Mean p-value: .56

Table 62. Multiple-Choice Item Analyses, Grade 9 Writing

Item #	Session	CS	SA	R_ITT	Omit %	p-Value	Item #	Session	CS	SA	R_ITT	Omit %	p-Value
49	S5	2		.39	.1	.57	70	S5	2	8	.47	.3	.75
50	S5	3	8	.25	.1	.28	71	S5	3	7	.38	.3	.54
51	S5	3	8	.37	.3	.57	72	S5	2	7	.38	.3	.66
52	S5	3	8	.36	.1	.78	73	S5	2	5	.47	.5	.76
53	S5	3	8	.33	.2	.48	74	S5	2		.35	.3	.50
54	S5	3	8	.45	.2	.70	75	S5	2		.44	.3	.91
55	S5	3	8	.34	4.4	.52	76	S5	2		.22	.4	.31
56	S5	3	7	.31	.1	.61	77	S5	2		.48	.4	.67
57	S5	3	7	.46	.2	.82	78	S5	2		.44	.3	.68
58	S5	3	7	.47	.3	.79	79	S5	2	5	.49	.3	.79
59	S5	3	7	.38	.5	.61	80	S5	2	5	.41	.3	.60
60	S5	3	7	.41	.3	.69	81	S5	2	5	.43	.4	.56
61	S5	3	7	.31	1.2	.62	82	S5	2	5	.53	.4	.78
62	S5	3	7	.47	.3	.72	83	S5	3	8	.22	.3	.48
63	S5	3	7	.39	.4	.82	84	S5	2	8	.55	.4	.85
64	S5	2	5	.53	.1	.80	85	S5	3	7	.48	.3	.87
65	S5	2	5	.23	.2	.54	86	S5	2	7	.51	.4	.79
66	S5	2	5	.49	.3	.71	87	S5	3	7	.33	.5	.58
67	S5	2	5	.49	.3	.86	88	S5	3	7	.38	.6	.44
68	S5	3	8	.41	.3	.38	89	S5	2	5	.37	.4	.70

Mean R_ITT: .41
Mean p-value: .65

Table 63. Constructed-Response Item Analyses, Grade 9 Writing

Item #	Part	Session	CS	SA	R_ITT	Percent of Students Obtaining Score Level					Omit %	p-Value
						0	1	2	3	4		
1		S1	2	6	.16	1.2	98.8				.0	.99
2	A	S2	3	7	.35	14.9	83.7				1.4	.84
2	B	S2	3	8	.39	65.6	33.0				1.4	.33
2	C	S2	3	7	.31	73.3	25.3				1.4	.25
2	D	S2	3	8	.17	23.7	74.9				1.4	.75
2	E	S2	3	8	.39	26.9	71.7				1.4	.72
2	F	S2	3	8	.24	39.5	59.1				1.4	.59
3	A	S2	2	6	.61	2.2	3.0	24.3	49.3	20.8	.5	.71
3	B	S2	2	6	.65	.3	3.8	27.2	49.4	18.9	.5	.71
3	C	S2	3	6	.47	.4	9.2	90.0			.4	.95
48		S4	2	5	.58	.4	3.2	34.3	48.1	10.8	3.1	.65
69		S5	2	5	.55	.3	2.5	36.9	46.4	11.7	2.1	.66
90		S5	2	5	.60	.4	4.0	32.3	45.9	13.8	3.6	.65

Mean R_ITT: .44
Mean p-value: .68

Table 64. Multiple-Choice Item Analyses, Grade 9 Mathematics

Item #	Session	CS	SA	R_ITT	Omit %	p-Value	Item #	Session	CS	SA	R_ITT	Omit %	p-Value
1	S1	3	2	.38	.1	.60	31	S2	1/6		.43	4.6	.68
2	S1	4/5		.28	.2	.62	33	S2	2		.35	1.0	.24
4	S1	1/6		.24	.3	.10	35	S2	4/5		.32	.6	.33
5	S1	1/6		.37	.3	.64	36	S2	2	1	.52	.5	.58
6	S1	4/5		.19	.3	.44	37	S2	4/5		.33	.5	.51
7	S1	4/5	2	.38	.2	.24	38	S2	4/5		.18	.9	.43
8	S1	4/5		.23	.2	.27	39	S2	2		.37	1.5	.55
9	S1	3		.37	.3	.53	41	S3	1/6		.38	.4	.37
11	S1	2	1	.50	.3	.64	42	S3	2		.39	.2	.64
12	S1	2		.24	.6	.53	43	S3	1/6		.45	.2	.70
13	S1	3		.38	.3	.61	44	S3	3		.21	1.8	.87
15	S1	3		.35	.6	.81	46	S3	4/5		.28	.4	.55
17	S1	2	1	.44	1.0	.53	48	S3	4/5		.40	.5	.64
18	S1	4/5	2	.48	.7	.46	50	S3	2		.44	.3	.50
19	S1	3		.24	.8	.69	51	S3	2		.06	.5	.33
21	S2	1/6		.29	.4	.33	52	S3	1/6		.41	.9	.27
22	S2	3		.37	.2	.58	54	S3	4/5		.34	.3	.63
24	S2	1/6		.02	.4	.16	55	S3	2		.22	.4	.15
25	S2	3		.16	.2	.24	56	S3	1/6		.51	.7	.58
26	S2	3		.43	.2	.71	57	S3	2		.47	.6	.37
27	S2	4/5		.13	.5	.28	58	S3	2		.46	.5	.60
29	S2	3		.33	1.3	.20	59	S3	1/6		.34	.8	.26
30	S2	3		.25	.8	.28							

Mean R_ITT (item 51 excluded): .34
Mean p-value (item 51 excluded): .48

Table 65. Constructed-Response Item Analyses, Grade 9 Mathematics

Item #	Session	CS	SA	R_ITT	Percent of Students Obtaining Score Level					Omit %	p-Value
					0	1	2	3	4		
3	S1	3		.51	57.0	35.8	5.4			1.8	.23
10	S1	1/6	2	.57	60.4	19.6	8.3	9.2		2.5	.21
14	S1	1/6		.56	70.9	6.0	19.1			4.0	.22
16	S1	2		.63	72.3	18.7	4.5	1.1		3.3	.10
20	S1	4/5	2	.71	31.2	36.4	18.9	8.1	2.5	2.9	.27
23	S2	2	1	.56	68.9	14.3	9.3			7.4	.16
28	S2	2	1	.68	34.9	22.1	20.2	18.4		4.4	.39
32	S2	2		.49	76.3	5.8	11.8			6.2	.15
34	S2	4/5		.64	43.6	34.7	11.3	6.5		3.9	.26
40	S2	2	1	.62	9.5	24.4	33.0	19.6	10.6	3.0	.48
45	S3	1/6		.49	80.6	.7	10.8			8.0	.11
47	S3	3		.42	90.1	5.5	1.5	.7		2.1	.04
49	S3	3		.48	22.8	55.7	18.5			3.0	.46
53	S3	4/5	2	.64	54.1	32.4	6.0	6.8		.8	.22
60	S3	3	1	.62	23.1	46.9	19.9	4.4	2.9	2.9	.28

Mean R_ITT: .58
Mean p-value: .24

Table 66. Multiple-Choice Item Analyses, Grade 10 Reading

Item #	Session	CS	SA	R_ITT	Omit %	p-Value	Item #	Session	CS	SA	R_ITT	Omit %	p-Value
4	S3	1	1	.33	.1	.62	43	S4	6	4	.42	.7	.71
5	S3	6	1	.43	.1	.70	44	S4	6	4	.27	1.2	.54
6	S3	1	1	.39	.1	.82	45	S4	6	4	.14	.4	.32
7	S3	6	1	.30	.2	.75	47	S4	5		.13	.2	.35
8	S3	6	1	.29	.6	.30	48	S4	5		.11	.2	.38
9	S3	1	3	.27	.9	.73	50	S4	4	2	.45	.4	.90
10	S3	1	3	.41	.1	.73	51	S4	1	2	.23	.6	.46
13	S3	5		.24	.2	.77	54	S4	5		.34	2.4	.65
15	S3	6	1	.31	.4	.81	55	S4	4	2	.40	2.4	.89
16	S3	6	1	.34	.5	.60	56	S4	1	2	.45	2.5	.90
17	S3	1	3	.44	.2	.57	99	S6	5		.30	.4	.61
18	S3	1	3	.49	.2	.82	100	S6	1	2	.35	.1	.43
19	S3	1	3	.24	.2	.39	101	S6	4	2	.40	.1	.62
20	S3	4	1	.46	.4	.86	102	S6	1	2	.49	.2	.75
22	S3	1	2	.27	.5	.53	103	S6	5		.34	.3	.71
23	S3	1	2	.43	.5	.73	105	S6	5		.20	.2	.38
24	S3	1	2	.26	.4	.64	106	S6	5		.43	.2	.62
25	S3	1	3	.48	.4	.90	108	S6	1	2	.48	.2	.79
26	S3	5		.40	1.2	.62	109	S6	1	2	.49	.2	.88
27	S3	5		.40	1.2	.69	110	S6	4	2	.54	.3	.88
28	S3	4	2	.24	1.3	.64	111	S6	4	2	.57	.1	.86
30	S4	6	1	.42	.2	.82	112	S6	5		.39	.2	.80
31	S4	4	1	.53	.1	.77	113	S6	5		.41	.2	.65
32	S4	6	1	.20	.3	.41	114	S6	1	3	.46	.2	.70
33	S4	6	1	.44	.2	.69	115	S6	4	1	.58	.2	.80
36	S4	6	4	.14	.2	.35	116	S6	6	1	.32	.2	.68
37	S4	6	4	.42	.1	.78	117	S6	6	1	.54	1.0	.73
39	S4	6	4	.18	.2	.66	118	S6	1	3	.55	1.2	.72
40	S4	6	4	.45	.2	.84	119	S6	4	1	.53	.2	.82
41	S4	6	4	.28	.3	.66	120	S6	1	3	.23	.3	.55
42	S4	1	3	.36	.6	.83							

Mean R_ITT: .37
Mean p-value: .70

Table 67. Constructed-Response Item Analyses, Grade 10 Reading

Item #	Session	CS	SA	R_ITT	Percent of Students Obtaining Score Level					Omit %	p-Value
					0	1	2	3	4		
11	S3	6	1	.53	14.0	36.9	46.4			2.7	.65
12	S3	6	1	.43	3.7	16.9	43.0	34.1		2.3	.68
14	S3	6	1	.54	7.0	9.7	23.4	55.4		4.6	.74
21	S3	5		.58	29.6	20.7	11.1	31.8		6.9	.46
34	S4	1	1	.57	24.1	4.6	19.1	48.1		4.0	.62
35	S4	4	1	.59	9.9	1.1	27.2	57.7		4.2	.76
38	S4	6	4	.51	13.5	57.3	23.6			5.6	.52
46	S4	6	4	.49	27.8	16.6	22.8	11.8	6.3	14.8	.31
49	S4	4	2	.44	40.7	35.8	18.2			5.3	.36
52	S4	4	2	.60	8.2	7.7	23.9	54.0		6.2	.73
53	S4	4	2	.55	11.3	28.5	20.6	31.8		7.8	.55
104	S6	5		.50	17.9	43.7	23.0	9.2		6.2	.39
107	S6	5		.61	8.0	11.2	22.6	53.0		5.2	.72
121	S6	4	1	.63	6.7	6.4	19.8	60.6		6.4	.76

Mean R_ITT: .54
Mean p-value: .59

Table 68. Multiple-Choice Item Analyses, Grade 10 Writing

Item #	Session	CS	SA	R_ITT	Omit %	p-Value	Item #	Session	CS	SA	R_ITT	Omit %	p-Value
57	S5	3	7	.28	.1	.57	77	S5	2	8	.34	.9	.75
58	S5	2		.33	.4	.85	78	S5	2	5	.34	.2	.61
59	S5	3	8	.44	.1	.76	79	S5	3	7	.39	.2	.75
60	S5	3	8	.37	.4	.59	80	S5	3	7	.32	.2	.46
61	S5	3	7	.37	.2	.77	81	S5	3	7	.56	.2	.87
62	S5	3	7	.45	.2	.90	82	S5	2	8	.52	.5	.79
63	S5	3	7	.51	.2	.88	83	S5	3	8	.20	.7	.56
64	S5	3	7	.33	.2	.70	85	S5	2		.40	.3	.85
65	S5	3	7	.43	.2	.89	86	S5	2		.49	.3	.74
66	S5	3	7	.49	.9	.89	87	S5	2		.37	.4	.89
67	S5	3	7	.47	.2	.75	88	S5	2		.43	.5	.56
68	S5	3	7	.37	.5	.85	89	S5	2	7	.50	.3	.85
69	S5	2	5	.21	.1	.54	90	S5	3	7	.34	.4	.59
70	S5	2	5	.51	.1	.88	91	S5	3	7	.41	.5	.48
71	S5	2	5	.37	.1	.73	92	S5	3	7	.50	.6	.78
72	S5	2	5	.49	.2	.72	93	S5	2	5	.57	.2	.78
73	S5	2	5	.49	.3	.90	94	S5	2	5	.52	.2	.87
74	S5	3	7	.26	.2	.65	95	S5	2	5	.52	.3	.73
75	S5	3	7	.29	.2	.46	96	S5	2	8	.49	.3	.71
76	S5	2	8	.38	.7	.69	97	S5	2	8	.35	.4	.47

Mean R_ITT: .41
Mean p-value: .73

Table 69. Constructed-Response Item Analyses, Grade 10 Writing

Item #	Part	Session	CS	SA	R_ITT	Percent of Students Obtaining Score Level					Omit %	p-Value
						0	1	2	3	4		
1		S1	2	6	.09	1.6	98.4				.0	.98
2	A	S2	3	8	.39	66.3	32.4				1.3	.32
2	B	S2	3	7	.32	37.7	61.0				1.3	.61
2	C	S2	3	7	.39	52.7	46.1				1.3	.46
2	D	S2	3	7	.23	14.1	84.6				1.3	.85
2	E	S2	3	7	.46	29.9	68.8				1.3	.69
2	F	S2	3	8	.28	33.0	65.7				1.3	.66
3	A	S2	2	6	.60	1.3	3.8	25.9	40.5	28.0	.5	.72
3	B	S2	2	6	.66	.2	5.2	26.6	40.3	27.3	.5	.72
3	C	S2	3	6	.50	.3	13.9	85.3			.5	.92
29		S3	2	5	.60	.3	9.7	36.9	38.4	11.5	3.3	.61
84		S5	2	5	.62	.3	7.3	31.0	43.2	15.4	2.9	.65
98		S5	2	5	.61	.3	7.1	28.9	42.7	17.3	3.7	.66

Mean R_ITT: .49
Mean p-value: .68

Table 70. Multiple-Choice Item Analyses, Grade 10 Mathematics

Item #	Session	CS	SA	R_ITT	Omit %	p-Value	Item #	Session	CS	SA	R_ITT	Omit %	p-Value
1	S1	1/6		.40	.4	.57	30	S2	3	2	.14	.3	.25
2	S1	4/5		.35	.2	.27	31	S2	1/6		.48	.4	.45
3	S1	2		.31	.4	.36	32	S2	1/6		.34	.3	.56
4	S1	4/5		.23	.2	.47	34	S2	1/6		.38	.3	.41
5	S1	3		.38	.2	.43	36	S2	4/5		.17	.4	.49
7	S1	1/6		.36	.3	.63	37	S2	1/6		.35	.5	.64
8	S1	3		.37	.1	.47	38	S2	2		.36	1.3	.50
9	S1	3		.21	.1	.75	41	S3	1/6		.39	.3	.61
11	S1	3		.40	.2	.61	42	S3	4/5		.49	.4	.31
13	S1	3	2	.34	.2	.83	44	S3	4/5		.17	.8	.47
14	S1	1/6		.22	.3	.54	45	S3	2		.34	.2	.46
16	S1	1/6		.37	.6	.66	46	S3	2	1	.32	.2	.19
17	S1	4/5		.50	.5	.29	47	S3	2		.50	.5	.45
18	S1	4/5		.29	.4	.76	48	S3	1/6		.51	.4	.34
19	S1	2	1	.30	.6	.66	50	S3	2		.22	.6	.34
21	S2	4/5		.36	.2	.34	51	S3	3	1	.29	.3	.75
22	S2	1/6		.05	.7	.34	53	S3	3		.02	.5	.25
23	S2	2		.41	.4	.27	54	S3	2		.44	.6	.65
24	S2	4/5		.30	.2	.50	55	S3	1/6		.43	.3	.50
26	S2	2		.39	.5	.65	56	S3	3	2	.36	.3	.23
27	S2	3	2	.22	.5	.41	57	S3	2		-.01	.6	.54
28	S2	3	2	.38	.3	.61	59	S3	4/5		.35	.6	.40
29	S2	2	1	.46	.2	.73							

Mean R_ITT (items 53 and 57 excluded): .34
Mean p-value (items 53 and 57 excluded): .49

Table 71. Constructed-Response Item Analyses, Grade 10 Mathematics

Item #	Session	CS	SA	R_ITT	Percent of Students Obtaining Score Level					Omit %	p-Value
					0	1	2	3	4		
6	S1	3		.55	40.1	51.3	7.1			1.6	.33
10	S1	2		.70	62.9	22.9	9.0	3.0		2.2	.17
12	S1	1/6		.50	81.6	1.1	11.7			5.6	.12
15	S1	1/6		.69	54.5	32.9	9.6	1.2		1.8	.19
20	S1	2		.75	42.8	20.5	11.6	14.0	7.3	3.8	.29
25	S2	3	2	.46	18.3	59.2	20.3			2.3	.50
33	S2	2	1	.53	7.1	20.2	34.2	38.0		.4	.68
35	S2	2		.58	69.4	7.7	17.8			5.0	.22
39	S2	4/5		.66	46.5	41.7	6.2	4.8		.8	.23
40	S2	3	1	.64	4.0	37.2	29.7	20.1	6.9	2.7	.46
43	S3	2	1	.54	33.4	46.0	16.4			4.2	.39
49	S3	4/5		.55	44.5	45.8	3.5	3.3		3.0	.21
52	S3	4/5		.66	63.2	20.5	10.4			5.9	.21
58	S3	3	2	.55	82.6	11.5	2.9	.9		2.2	.07
60	S3	4/5		.64	52.1	36.3	5.0	1.7	.9	3.9	.14

Mean R_ITT: .61
Mean p-value: .28

**Table 72. Item Parameter and Fit Summary, Multiple-Choice Items,
Grade 3 Reading (N=49,522)**

Item	a	b	c	Fit-Z	Item	a	b	c	Fit-Z
1	1.41	-2.57	.20	15.7	22	.83	-.02	.12	10.8
2	.53	-3.18	.20	33.1	23	.96	-.90	.22	2.8
3	1.30	-1.50	.20	5.1	24	.92	-.16	.09	21.7
6	.65	-.97	.24	4.9	25	1.65	-.94	.16	17.7
7	1.40	-2.19	.14	7.8	26	.96	-.61	.19	6.3
8	1.39	-2.36	.16	7.6	27	.94	1.43	.15	50.4
9	.62	-2.27	.13	11.0	28	1.42	-.72	.31	15.4
10	1.16	-2.34	.28	46.4	29	1.12	-.87	.27	11.4
11	1.18	-2.41	.13	8.9	30	.94	-.61	.20	7.6
12	.81	-1.65	.16	1.8	31	1.06	-.58	.13	10.0
13	.87	-.97	.23	7.5	35	1.12	-1.99	.04	2.0
14	1.25	-.50	.27	14.4	36	.80	-1.48	.20	13.8
17	.86	-.89	.17	8.7	37	1.38	-.83	.18	13.6
18	1.12	-.27	.17	19.5	38	.70	-2.02	.13	7.9
19	1.18	-1.37	.11	4.7	39	.86	-.20	.16	7.3
20	.82	-1.85	.06	4.7	40	1.11	-1.01	.16	5.3
21	.70	-1.23	.13	15.4	41	.82	-1.46	.03	47.7

**Table 73. Item Parameter and Fit Summary,
Constructed-Response Items, Grade 3
Reading (N=49,522)**

Item	f	g1	g2	g3	g4	Fit-Z
4	.33	-.07	.64			25.2
5	.76	-2.05	-2.16	-1.35		66.6
15	.46	-.64	-.86	-0.14	-0.24	69.7
16	.51	.87	.10			213.7
32	1.08	-2.32	-1.07			23.2
33	.75	-1.08	-.49	-1.31		22.1
34	1.12	-1.77	-1.58			8.5

**Table 74. Item Parameter and Fit Summary, Multiple-Choice Items,
Grade 3 Spanish Reading (N=1659)**

Item	a	b	c	Fit-Z	Item	a	b	c	Fit-Z
1	.53	-3.09	.20	.4	21	.86	-1.43	.20	4.6
2	.95	1.36	.19	1.8	22	1.34	-2.00	.20	2.0
5	.63	-1.31	.20	.1	23	.65	-.60	.20	-.3
6	.45	-1.16	.28	1.8	24	1.06	-1.55	.20	-.9
7	.59	-.80	.16	-.2	25	.83	-1.72	.20	2.8
8	.90	-1.75	.28	-.9	28	.61	-1.95	.20	2.0
9	.74	.30	.21	-.6	30	.67	-.89	.20	.0
10	.63	-.18	.28	-.3	31	.44	-1.27	.20	-.1
11	.64	-2.11	.28	1.1	33	.99	-1.24	.16	-.8
12	.64	2.00	.15	.3	34	.70	-.48	.17	.4
14	.61	.81	.20	-.6	35	.88	-.42	.20	1.4
15	.62	-.56	.16	.4	36	.51	.37	.20	.9
16	.76	-1.63	.20	.6	37	.94	-.88	.20	-1.0
18	.63	-1.20	.20	2.4	38	.70	1.06	.19	.9
19	.80	-.89	.20	1.0	39	.68	-.75	.16	-.4
20	.84	-.95	.16	.0	40	.56	.97	.18	-.4

**Table 75. Item Parameter and Fit Summary,
Constructed-Response Items, Grade 3
Spanish Reading (N=1659)**

Item	f	g1	g2	g3	g4	Fit-Z
3	.54	-.13	-.93	-1.18		-.3
4	.46	.04	.78			-.3
13	.87	-.74	-.44			.0
17	.86	-1.00	-2.42			2.9
26	.78	-1.82	-.80			-.8
27	.58	.54	-.12			5.8
29	.48	1.12	.12			.5
32	.94	-1.23	-.22			-.4

Table 76. Item Parameter and Fit Summary, Multiple-Choice Items, Grade 3 Writing (N=9194)

Item	a	b	c	Fit-Z	Item	a	b	c	Fit-Z
1	1.25	-2.45	.28	.1	21	.95	-2.15	.20	.2
2	1.18	-1.67	.22	.5	22	1.29	-1.45	.31	2.1
3	.97	.58	.42	7.1	23	1.07	-1.97	.20	.7
4	.81	-1.60	.28	1.7	24	.91	-1.65	.15	2.5
6	1.40	-2.64	.20	.2	25	1.08	-2.34	.20	.7
7	1.31	-2.30	.20	-.4	26	.75	-2.17	.20	3.3
8	.50	-2.00	.20	2.1	27	.83	-1.96	.20	1.5
9	1.08	-1.52	.15	.9	28	.90	-2.32	.20	3.8
10	.48	-1.22	.20	5.7	30	.72	.05	.24	7.9
11	.84	-1.46	.20	1.1	31	.74	-1.89	.20	2.0
12	.67	-1.58	.20	1.2	32	.98	-1.70	.20	-1.1
14	.78	-2.53	.20	.2	33	1.28	-1.79	.17	1.5
15	.93	-3.33	.20	.6	34	.70	-1.32	.19	17.7
16	.81	-2.21	.20	.4	35	.72	-2.42	.20	-1.1
17	.48	-1.09	.20	6.1	36	.88	-1.62	.20	1.9
18	1.00	-.90	.17	3.1	37	.76	-1.01	.15	.7
19	1.03	-.09	.31	8.3	38	.96	-2.73	.20	-1.2
20	.81	-1.67	.20	2.3					

Table 77. Item Parameter and Fit Summary, Constructed-Response Items, Grade 3 Writing (N=9194)

Item	Part	f	g1	g2	g3	g4	Fit-Z
5	A	1.70	-2.81				-.1
5	B	1.49	-1.95				.5
5	C	1.56	-2.58				1.9
5	D	.77	-2.85				-.1
5	E	1.32	-1.43				.7
5	F	1.60	-.99				8.3
13	A	1.02	-4.36	2.12			14.6
13	B	1.45	-4.46				-.3
13	C	1.42	-2.37				2.6
13	D	1.88	-5.15				.3
29	A	1.10	-4.15	2.07			15.9
29	B	1.73	-5.45				.2
29	C	1.59	-2.89				4.1
29	D	2.09	-6.21				-1.1
39	A	1.14	-4.88	2.49			11.1
39	B	1.65	-5.13				.2
39	C	1.53	-2.67				4.1
39	D	2.02	-6.33				-.8

**Table 78. Item Parameter and Fit Summary, Multiple-Choice Items,
Grade 3 Spanish Writing (N=1474)**

Item	a	b	c	Fit-Z	Item	a	b	c	Fit-Z
1	.77	-2.25	.20	.2	21	.54	-.75	.20	.7
2	.88	-1.12	.20	-.4	22	.82	.29	.18	-.1
3	.65	-2.94	.20	-.4	23	.75	-1.14	.20	-1.0
4	.59	-.60	.20	.1	24	.62	-.70	.20	-1.1
5	.88	-1.67	.20	-1.3	25	.88	-1.91	.20	-.4
6	.49	-1.57	.20	-.7	26	.74	-1.60	.20	.1
8	.37	-1.40	.20	6.5	27	.90	-.68	.20	-.2
9	.44	.20	.20	1.0	28	.55	-1.84	.20	-.7
10	.60	-.52	.20	1.2	30	.56	-.88	.20	.5
11	.48	-1.12	.20	1.2	31	.45	3.58	.19	.9
12	.58	.19	.20	-.1	32	.74	-1.23	.20	.3
13	.95	.68	.32	.9	33	.78	-1.05	.20	.1
14	.46	-.84	.20	.3	34	.47	.17	.20	-.1
15	.66	-2.12	.20	.0	35	.50	.33	.20	1.5
16	.76	-2.25	.20	.6	36	.69	-.59	.20	-1.0
17	.82	-2.17	.20	-1.1	37	.41	.75	.20	2.9
18	.58	-1.53	.20	1.5	38	.37	-1.76	.20	-1.1
20	.91	.25	.27	-.5					

**Table 79. Item Parameter and Fit Summary,
Constructed-Response Items, Grade 3
Spanish Writing (N=1474)**

Item	Part	f	g1	g2	g3	g4	Fit-Z
7	A	.95	-1.93				1.3
7	B	1.52	-2.60				-1.1
7	C	.98	.79				-.6
7	D	.45	-.92				1.1
7	E	1.11	-.43				5.5
7	F	.59	-.23				.6
19	A	1.10	-2.63	.91			.0
19	B	1.38	-.64				1.1
19	C	1.80	-.79				.6
19	D	2.77	-5.70				.7
29	A	1.25	-3.16	.96			.7
29	B	1.64	-1.30				.4
29	C	1.99	-.70				.8
29	D	2.33	-4.87				-.7
39	A	.92	-.94	1.40			4.2
39	B	1.47	-.64				.6
39	C	1.86	-.21				.1
39	D	2.12	-3.92				-.6

**Table 80. Item Parameter and Fit Summary, Multiple-Choice Items,
Grade 4 Reading (N=10,328)**

Item	a	b	c	Fit-Z	Item	a	b	c	Fit-Z
4	.94	-2.25	.20	2.4	41	1.09	-1.78	.20	1.3
5	.93	-1.98	.20	1.1	42	1.18	-1.79	.11	.3
6	.86	-.99	.09	-.3	43	1.37	-.69	.22	3.5
7	.78	-1.77	.20	2.2	44	1.28	-1.72	.23	8.0
10	.96	-.26	.32	2.5	45	.69	-1.22	.16	2.3
11	.67	-.63	.14	1.4	46	.76	-.67	.16	1.1
12	.24	1.85	.20	10.2	47	1.24	-.48	.18	2.0
13	1.10	-.28	.26	2.3	91	1.00	-1.32	.21	.8
14	.74	.37	.14	2.5	92	.89	-1.37	.12	1.6
15	.78	-1.72	.20	8.1	95	.91	-2.50	.20	3.6
16	.73	.45	.26	.8	96	.74	-2.25	.20	3.1
21	1.04	-.27	.21	2.1	97	1.17	-1.61	.14	1.0
22	1.18	-1.85	.16	1.1	98	.55	-2.33	.20	2.4
23	.91	-.93	.17	.1	99	.80	-1.03	.20	-.2
24	.80	-.40	.21	-.1	100	1.10	-1.75	.20	-.3
25	1.00	-.48	.32	-.1	101	.91	-1.01	.15	-.4
26	.66	-1.47	.20	1.9	102	.69	-2.06	.20	29.8
27	.92	.67	.17	.6	103	.64	-.20	.10	-.2
28	.89	-.72	.22	4.3	104	1.12	-.61	.21	-.6
29	.93	-1.13	.20	-.2	106	.86	.36	.18	-.2
30	.76	-.70	.30	3.7	107	.85	-.32	.21	.7
31	.59	.18	.14	1.5	108	.80	-.24	.18	2.1
32	.94	-1.04	.13	-.3	109	.64	-.73	.18	9.9
33	.85	-.35	.23	1.2	110	.84	-.25	.18	-.9
34	.69	-1.12	.13	-1.1	111	.86	-1.09	.20	1.2
37	.56	.37	.20	5.0	112	1.30	-.69	.27	.5
39	.62	-1.24	.20	12.8	113	.86	-.27	.14	3.6
40	.94	-.83	.14	-.4	115	1.41	-.21	.22	4.4

**Table 81. Item Parameter and Fit Summary,
Constructed-Response Items, Grade 4
Reading (N=10,328)**

Item	f	g1	g2	g3	g4	Fit-Z
8	.60	-.15	-.75			16.8
9	.41	.42	.34			10.3
17	.88	-.38	.81			5.5
18	.67	-.01	2.33			.9
19	.43	4.46	-4.22	.59		16.2
20	.56	1.72	.15	.25		12.2
35	.51	.58	-2.48			14.5
36	.36	-.23	-1.11	-.31	-.57	9.5
38	.65	.65	-.32			39.2
93	.81	-2.39	-1.54			3.4
94	.62	-1.05	-1.04	-1.49		.2
105	.77	-.44	-1.24	-1.80		17.5
114	.73	-1.37	.90	.14		18.2
116	.73	-.52	-.02			19.2

Table 82. Item Parameter and Fit Summary, Multiple-Choice Items, Grade 4 Spanish Reading (N=702)

Item	a	b	c	Fit-Z	Item	a	b	c	Fit-Z
4	1.06	.65	.17	.7	38	.73	.42	.18	-.8
5	.14	2.15	.20	-.9	41	.54	1.59	.20	1.2
6	.47	-.60	.20	-.9	42	.60	-1.37	.20	.6
7	.79	1.04	.17	.4	43	.54	.82	.19	1.8
8	.68	.41	.18	.7	44	.75	2.69	.15	-.1
10	.78	.19	.26	.0	45	.81	-1.42	.20	-.1
12	.37	-.31	.20	.4	46	.99	-.63	.20	-.7
15	.89	.40	.18	.5	47	1.20	-.23	.18	.1
17	.34	-.80	.20	.5	48	.92	.65	.13	1.2
18	.78	-.56	.20	.0	92	.70	-1.66	.20	-.1
19	.38	-1.58	.20	-.2	93	.63	-.57	.20	3.3
20	.76	-.64	.20	-.4	94	1.18	.59	.28	-1.0
21	.61	-2.60	.20	.9	95	.37	-1.56	.20	2.0
22	.25	1.13	.20	.3	96	.82	-1.30	.20	.1
23	.80	.05	.16	-.5	97	.70	.27	.20	-1.1
24	.78	.06	.16	.0	98	.75	-1.30	.20	1.7
25	.43	.30	.19	-.9	99	1.21	-1.76	.20	-.5
26	.37	.51	.20	2.7	100	.42	1.05	.18	-1.1
27	.82	-1.02	.20	-.9	102	1.20	.02	.26	-.2
28	1.08	.03	.34	-.8	104	1.05	1.27	.18	-1.4
30	.62	.95	.17	.2	105	.81	.54	.24	-.1
31	.69	-.28	.20	2.3	106	.73	1.33	.15	-.4
33	.74	2.09	.23	.1	108	1.66	2.02	.11	.8
34	.81	-1.28	.20	.9	109	.61	.14	.18	.6
35	.75	-.62	.20	.8	110	.21	.20	.20	.6
36	.53	.67	.24	-.5	111	1.15	1.82	.20	.8

Table 83. Item Parameter and Fit Summary, Constructed-Response Items, Grade 4 Spanish Reading (N=702)

Item	f	g1	g2	g3	g4	Fit-Z
9	.82	-.40	-1.40			2.5
11	.56	1.62	-.91	-0.24	-1.96	-1.2
13	.33	1.60	.66	0.70		.7
14	.66	1.37	.52			-.6
16	.86	.14	-.01			-1.4
29	.98	-1.34	-.77	-1.90		.5
32	.94	1.18	.31			1.3
37	.64	.34	.54	1.02		-1.2
39	1.14	-.10				-.1
40	1.21	.00				-.9
101	.70	-2.01	-.40			.7
103	.65	1.01	-1.69			.2
112	.28	1.53	-.47			1.6
113	.23	1.83	-.04			.5

Table 84. Item Parameter and Fit Summary, Multiple-Choice Items, Grade 4 Writing (N=9970)

Item	a	b	c	Fit-Z	Item	a	b	c	Fit-Z
49	1.21	-3.43	.20	1.4	69	.62	-.62	.14	-.6
50	1.08	-3.10	.20	.1	70	.75	-1.07	.20	.4
51	1.02	-1.76	.20	-.1	72	.54	-1.28	.19	2.5
52	.56	-1.79	.20	3.2	73	.59	.28	.22	2.1
53	.44	-2.37	.20	-.6	74	.87	-.15	.17	.0
54	.81	-2.84	.20	8.5	75	.78	-1.57	.20	.5
55	1.19	-1.95	.14	1.8	76	.92	-.20	.19	.4
56	.55	-1.47	.20	9.3	77	1.22	.05	.21	3.1
57	.78	-2.64	.20	.2	78	1.06	-1.76	.16	.4
58	.51	-1.61	.19	1.3	79	1.18	-2.50	.18	31.2
59	.64	-.86	.19	-1.0	80	.92	-2.11	.20	.2
60	.50	-1.65	.20	-.5	81	.77	.54	.22	1.4
61	.64	-2.32	.20	1.2	82	.87	.04	.21	2.1
62	1.00	.47	.13	5.4	83	1.13	-1.00	.26	-.3
63	.53	-.81	.20	3.4	84	1.05	-1.44	.22	1.0
64	.75	-2.84	.20	.7	85	.61	-1.10	.20	.7
65	.80	-.76	.10	1.1	86	.75	-1.17	.20	4.0
66	.87	-.26	.23	2.7	87	.86	.46	.28	2.7
67	.83	-.24	.22	.9	88	1.09	-.22	.15	2.8
68	.66	-.26	.26	1.9	89	.86	-1.38	.20	1.8

Table 85. Item Parameter and Fit Summary, Constructed-Response Items Grade 4 Writing (N=9970)

Item	Part	f	g1	g2	g3	g4	Fit-Z
1		.84	-5.87				-1.5
2	A	1.09	-.43				2.3
2	B	1.13	1.00				-.6
2	C	.72	.13				.2
2	D	1.26	-3.11				3.3
2	E	.29	-2.41				.3
2	F	1.17	.34				1.3
3	A	.81	-1.26	-2.47	-.08	1.64	190.3
3	B	1.20	-5.14	-2.75	-.08	2.00	6.0
3	C	1.61	-6.26	-1.28			.3
48		1.02	-3.08	-2.37	.07	1.78	21.4
71		1.01	-3.97	-2.79	-.16	1.65	12.5
90		1.00	-3.74	-2.42	.32	2.05	20.6

Table 86. Item Parameter and Fit Summary, Multiple-Choice Items, Grade 4 Spanish Writing (N=703)

Item	a	b	c	Fit-Z	Item	a	b	c	Fit-Z
50	.65	-.02	.20	.0	71	1.20	.25	.32	.6
51	.77	-1.47	.20	.1	72	1.19	.67	.24	-.3
52	.74	-1.07	.20	-.2	73	.91	.52	.28	-1.2
53	.69	-1.59	.20	1.3	75	.36	.05	.20	.0
55	.58	.95	.22	1.5	76	.55	.14	.20	3.7
56	.73	-2.28	.20	.8	77	.41	1.68	.20	-1.1
57	.66	-.17	.20	.4	78	.92	1.26	.21	2.7
58	.59	-1.08	.20	-.1	79	.69	.52	.23	2.0
59	.85	-2.26	.20	-1.1	80	.53	1.83	.25	-1.0
60	.69	-1.57	.20	.7	81	.82	1.63	.16	.8
61	.78	-2.95	.20	-1.4	82	.85	1.74	.20	-1.0
62	.89	-1.74	.20	2.1	83	.74	.73	.25	-.4
63	1.37	.73	.18	1.2	84	1.76	1.46	.28	-.3
64	1.18	-.46	.32	-.3	85	.51	.29	.20	1.7
65	1.21	.39	.34	-.6	86	.56	.75	.21	2.4
66	1.55	.10	.30	-.7	87	.70	.15	.20	1.6
67	1.40	-.08	.38	-.5	88	.46	3.40	.20	1.7
68	1.42	.83	.28	.1	89	.66	.58	.26	.4
69	1.64	-.12	.29	1.9	90	1.19	1.00	.28	-.2
70	1.39	.76	.22	.6					

Table 87. Item Parameter and Fit Summary, Constructed-Response Items, Grade 4 Spanish Writing (N=703)

Item	Part	f	g1	g2	g3	g4	Fit-Z
1		1.02	-4.77				-.8
2	A	.89	-.93				.4
2	B	.82	.81				.3
2	C	.79	.70				2.4
2	D	1.20	4.33				-.8
2	E	.78	.50				.2
2	F	.56	-1.02				1.1
3	A	.61	.41	-1.49	-.20	3.12	2.2
3	B	1.24	-3.46	-2.26	.00	4.12	-.8
3	C	1.80	-4.23	4.71			-2.0
49		.68	-.66	-.84	.44	2.79	-1.0
74		.30	2.75	-1.35	.09	2.12	1.9
91		.61	-.96	-1.54	.62	2.22	-.2

**Table 88. Item Parameter and Fit Summary, Multiple-Choice Items,
Grade 5 Reading (N=10,415)**

Item	a	b	c	Fit-Z	Item	a	b	c	Fit-Z
4	.98	-.60	.25	3.2	43	.83	-1.04	.18	.7
5	.48	-.87	.20	3.0	44	.82	-.85	.23	1.3
6	.49	-.07	.12	-1.1	45	1.02	-1.68	.20	2.6
9	1.01	-1.25	.31	-.6	46	1.00	1.74	.11	8.2
10	.99	.17	.41	2.1	47	.57	-.20	.20	3.3
11	.95	-1.14	.12	.3	48	.64	-.91	.12	.3
12	.93	-1.24	.19	.1	49	1.14	.24	.19	2.6
15	1.28	-2.31	.20	1.8	50	1.04	-1.36	.18	-.3
16	.93	-1.30	.20	3.2	93	.54	-1.78	.20	.2
17	1.22	-2.09	.11	.1	94	1.38	-1.63	.21	1.0
18	1.04	-2.55	.20	.9	95	.69	-1.92	.20	3.5
21	.32	-.62	.20	14.2	96	.41	-1.68	.20	7.0
22	.32	.64	.18	5.7	97	.53	-1.02	.19	1.3
23	.29	.40	.20	-.6	100	.94	.16	.18	2.7
24	.80	-1.19	.20	1.3	101	.73	-1.78	.19	.9
25	.74	-1.23	.20	4.7	102	.64	.45	.19	1.6
26	1.05	-.72	.28	3.5	103	1.00	-.93	.24	1.5
28	.73	1.04	.24	-.3	104	.51	.74	.20	-.8
29	.83	-.26	.27	-.1	105	.68	-.72	.09	4.0
30	.84	-1.90	.20	.1	106	1.18	-.81	.23	2.4
31	.44	-1.25	.20	7.6	107	.93	.02	.17	3.4
34	1.07	-.93	.27	1.7	108	.96	-.40	.22	1.6
35	.66	-1.93	.20	3.0	109	.74	-.52	.30	-1.6
36	.93	.13	.17	.4	110	.76	-.55	.14	-.1
37	.96	-.81	.20	1.7	111	.55	1.48	.18	2.8
38	1.08	-.88	.23	-1.2	112	1.22	-1.03	.29	1.4
41	.85	-1.77	.13	.5	113	.82	-.74	.10	3.0
42	.99	-1.29	.20	-.3	115	1.32	-.42	.24	.2

**Table 89. Item Parameter and Fit Summary,
Constructed-Response Items, Grade 5
Reading (N=10,415)**

Item	f	g1	g2	g3	g4	Fit-Z
7	.92	-1.00	.16	.97		9.2
8	.79	-1.68	-1.70			6.2
13	.40	4.30	-4.38	.27		19.9
14	.57	1.02	.19	-.07		18.0
19	.65	-.65	-.49	.21		7.9
20	.81	-.59	1.00			12.0
32	.45	-.51	-.27			2.1
33	.67	-.85	-.60			9.5
39	.66	.00	-.24			3.3
40	.82	-.55	-.01	.28		-.1
98	.62	-1.41	.22			3.9
99	.79	-.65	.95			11.9
114	.70	-1.42	.47	-.05		12.3
116	.67	-.69	-.15			13.7

Table 90. Item Parameter and Fit Summary, Multiple-Choice Items, Grade 5 Writing (N=10,005)

Item	a	b	c	Fit-Z	Item	a	b	c	Fit-Z
51	.81	-.50	.21	-.4	72	1.01	-1.12	.10	-.2
52	.61	-1.55	.18	-.3	73	.95	-1.35	.15	1.4
53	.63	-1.99	.20	-.5	74	.98	-.19	.28	.6
54	.66	1.04	.16	.5	75	.99	-.84	.15	.9
55	.32	-1.92	.23	-.4	76	.77	-1.95	.20	2.4
56	.62	-2.64	.20	.0	77	.66	-.93	.14	-.2
57	.50	-2.01	.20	-.7	78	.84	-.92	.15	.9
58	.64	1.07	.16	.8	79	.71	-1.27	.17	-.7
59	.83	-1.43	.20	-.8	80	.78	-.41	.25	-.7
61	.43	-.82	.20	3.1	81	.87	-.19	.19	3.9
62	.46	-1.91	.20	3.3	82	1.25	-.30	.22	2.0
63	.69	2.11	.19	1.7	83	.51	-.21	.18	4.8
64	.18	2.62	.20	.5	84	.57	.64	.20	1.0
65	.98	.93	.21	.8	85	.83	-.45	.21	.4
66	.48	-.41	.20	5.4	86	1.00	-.78	.19	1.1
67	.47	-2.24	.20	19.5	87	.67	-.80	.16	.1
68	.96	-1.45	.13	2.1	88	.61	-.93	.14	.7
69	.95	-1.72	.20	9.4	89	.72	-.24	.21	3.2
70	.56	-1.24	.20	.6	90	1.37	-.90	.20	2.9
71	.73	-3.07	.20	1.1	91	1.25	-.69	.15	4.5

Table 91. Item Parameter and Fit Summary, Constructed-Response Items, Grade 5 Writing (N=10,005)

Item	Part	f	g1	g2	g3	g4	Fit-Z
1		.54	-5.45				-.9
2	A	1.16	-1.77				.5
2	B	1.41	-1.99				.4
2	C	.84	-2.73				2.6
2	D	1.28	-2.35				1.3
2	E	1.04	-.59				.5
2	F	.99	-.41				6.2
3	A	.96	-2.39	-2.46	-.17	.96	170.0
3	B	1.26	-4.10	-2.79	-.20	1.21	27.1
3	C	1.76	-5.33	-3.12			68.3
27		.73	-1.61	-1.93	.07	1.84	54.7
60		.88	-3.53	-2.62	-.27	1.07	7.8
92		1.07	-4.19	-1.48	1.16	2.68	9.7

**Table 92. Item Parameter and Fit Summary, Multiple-Choice Items,
Grade 5 Mathematics (N=10,030)**

Item	a	b	c	Fit-Z	Item	a	b	c	Fit-Z
1	1.17	1.20	.22	1.1	35	1.09	-.01	.09	-.3
2	.63	-.09	.17	.2	36	.86	-.88	.24	.4
3	.38	-2.91	.20	-.2	38	.43	-.32	.13	-1.2
4	.63	-1.76	.20	2.3	39	.60	-2.56	.20	5.8
6	.73	-3.61	.20	2.8	40	.89	-1.13	.14	.3
7	.87	-1.03	.28	-.5	41	1.05	-.52	.20	1.2
8	.85	-1.02	.37	-.9	43	1.30	-.91	.11	4.3
10	.94	-1.41	.22	-.4	44	.74	-1.41	.19	17.5
11	.87	-1.11	.27	-.5	45	.47	-.39	.12	.6
12	.76	-.04	.35	1.6	47	.48	-2.01	.20	6.0
13	.75	-1.13	.19	.0	48	.59	-1.52	.16	1.6
14	.78	-.77	.23	3.1	49	.92	-2.84	.20	4.9
15	.78	-1.25	.10	1.8	50	.62	-1.96	.19	2.4
16	.96	-.59	.24	.1	52	.60	-.01	.18	.3
17	1.00	.30	.32	2.5	53	1.34	-1.71	.15	3.8
19	.76	1.34	.09	4.3	55	.89	.80	.32	3.4
20	.78	-1.41	.10	1.9	56	.41	-.89	.20	4.9
22	.45	-2.69	.20	2.2	57	1.21	-.02	.25	2.8
24	.57	-2.74	.20	-.1	59	.84	-.81	.15	.5
25	.60	.10	.17	-.5	60	.64	-.05	.25	.8
26	.78	-.54	.18	2.6	62	.67	-.06	.24	-.3
27	.94	-.45	.47	-.3	63	.92	-.84	.22	2.1
28	1.07	-.86	.12	-.8	64	.74	-1.12	.19	21.8
30	.82	-1.41	.20	.6	65	1.09	-1.57	.27	1.9
31	.81	-.84	.21	-.5	66	1.30	-.13	.17	.0
32	.32	-1.74	.20	2.5	67	.60	-1.64	.19	4.6
34	.79	1.48	.23	1.1	68	.68	-.97	.20	3.4

**Table 93. Item Parameter and Fit Summary,
Constructed-Response Items, Grade 5
Mathematics (N=10,030)**

Item	f	g1	g2	g3	g4	Fit-Z
5	.69	-.86	-.20			7.6
9	1.28	-4.54	-.63	.01		2.0
18	1.20	-.71	.05	.78		2.3
21	.68	-.89	-1.86			3.2
23	.73	-1.35	-2.06	-.30	-1.34	25.0
29	1.27	-1.57	-.63			20.7
33	.80	-2.02	-.53	-1.12		13.8
37	.99	-1.19	.06			26.1
42	.66	-1.20	-1.00	-.94		7.1
46	.70	-3.97	.90	-.69	-.20	8.0
51	.68	-1.21	-.99			14.1
54	.79	-2.09	-1.51	-.85		1.2
58	.67	-1.12	.53	1.17		8.4
61	1.21	-1.51	.23			10.3
69	.96	-2.78	.63	-1.11	1.07	13.3

**Table 94. Item Parameter and Fit Summary, Multiple-Choice Items,
Grade 6 Reading (N=9498)**

Item	a	b	c	Fit-Z	Item	a	b	c	Fit-Z
5	.90	-1.90	.20	3.4	37	.99	-.76	.30	-.3
6	.84	-1.68	.20	3.9	38	.83	-.29	.26	.2
7	.93	-1.88	.15	.3	39	1.14	-1.26	.19	.4
8	1.09	-.43	.27	.6	42	1.01	-1.51	.29	1.1
9	.55	1.30	.13	1.0	43	.36	-.11	.20	9.5
10	1.15	1.06	.26	2.0	47	.93	-.89	.36	-.8
13	.54	-1.31	.20	2.9	48	.52	-.51	.20	.7
14	.78	.05	.23	3.3	49	.91	-.70	.24	-.4
15	.83	.03	.23	-.7	95	.54	-.73	.20	6.8
16	1.46	.80	.24	12.5	96	.70	.67	.25	.0
17	1.04	-.21	.21	3.2	97	.72	-1.39	.20	7.3
18	.86	-.06	.16	.6	98	1.04	.63	.26	2.5
20	.52	-.65	.20	.4	99	.58	-1.33	.20	4.0
21	1.01	-.28	.20	.3	100	1.72	-1.00	.21	.1
22	.78	-1.23	.11	4.1	101	1.04	-.18	.18	3.0
23	.64	-1.28	.20	5.4	102	.78	-2.05	.17	-.1
24	.58	-.76	.20	23.3	103	.73	.34	.20	-.7
25	.33	-.46	.20	11.9	104	.51	-.27	.16	2.3
26	1.17	.79	.12	7.9	105	.65	-.85	.20	39.4
29	.34	-.01	.20	10.2	106	.61	.28	.20	4.8
30	.44	.02	.16	1.0	107	1.11	-.86	.14	.6
31	.27	.85	.20	.5	108	1.10	-.38	.19	3.9
32	.79	-1.30	.15	-1.1	110	.78	-.85	.20	1.0
33	.77	-1.33	.20	3.0	111	1.20	1.56	.22	30.4
34	1.20	-.94	.25	.6	112	1.15	-.17	.27	.4
35	.81	-1.80	.15	2.4	113	1.08	-.29	.20	1.1
36	.73	-.47	.27	4.8	114	1.15	.01	.23	1.6

**Table 95. Item Parameter and Fit Summary,
Constructed-Response Items, Grade 6
Reading (N=9498)**

Item	f	g1	g2	g3	g4	Fit-Z
4	.79	-2.20	-2.20			5.0
11	.59	.43	-1.52	-.18	-1.07	13.5
12	.80	.13	-.65			2.7
19	.98	.24	-1.44			5.0
27	.73	-.65	-.90	-.07		3.8
28	.92	-.97	1.35			16.3
40	.78	-2.38	2.38			24.4
41	.59	1.42	-1.20	.17	.34	9.5
44	1.00	-1.72	-.10			1.6
45	.68	-1.31	-1.08			42.6
46	1.04	-.85	-2.46			4.0
93	.63	-1.14	-.35	-.97		7.1
94	1.01	-.05	-.14			2.9
109	.60	-.40	.80			13.6
115	.48	.65	-.06	.28		14.4

Table 96. Item Parameter and Fit Summary, Multiple-Choice Items, Grade 6 Writing (N=9427)

Item	c	b	c	Fit-Z	Item	c	b	c	Fit-Z
51	.51	-1.51	.20	3.7	72	1.40	-1.77	.18	1.7
52	.94	-.70	.28	.5	73	.96	-.71	.12	1.7
53	.86	-1.08	.23	.3	74	.60	-1.27	.20	13.6
54	.84	-1.92	.16	-.8	75	.43	.33	.13	-0.6
55	.82	-.54	.28	.5	76	.58	.65	.19	9.7
56	.61	.44	.17	.9	77	.68	-.97	.14	12.6
57	.72	.88	.25	1.3	78	.93	-.93	.19	-0.1
58	1.17	-.12	.32	1.3	79	.89	-.51	.26	0.5
59	1.07	-.51	.26	1.1	80	.68	1.33	.18	0.6
60	.79	-.23	.22	1.5	81	.69	-.87	.20	5.8
61	.97	-1.42	.23	2.2	82	.74	-.60	.20	10.7
62	.67	-2.24	.20	.0	83	1.15	-1.02	.14	3.5
63	.66	1.66	.16	4.0	84	.89	-.52	.15	-0.3
64	.27	.43	.20	3.0	85	1.56	-1.18	.16	5.8
65	.88	-.44	.35	3.8	86	1.43	-.96	.16	8.0
66	.84	-.05	.38	2.7	87	1.56	-.88	.18	2.3
67	.50	-1.66	.20	2.1	88	1.16	-.91	.17	2.5
68	.80	.84	.21	.5	89	.46	-.40	.20	8.2
69	.84	-1.65	.13	-.4	90	.77	-1.00	.21	2.1
71	.88	-.97	.19	1.1	91	.80	-.42	.21	0.8

Table 97. Item Parameter and Fit Summary, Constructed-Response Items, Grade 6 Writing (N=9427)

Item	Part	f	g1	g2	g3	g4	Fit-Z
1		.78	-5.48				-1.4
2	A	.85	2.20				2.0
2	B	.79	-.46				2.0
2	C	.99	-.45				.8
2	D	.57	-.95				-.3
2	E	.47	.45				4.9
2	F	.46	-1.62				-1.0
3	A	.95	-2.65	-2.34	-.03	1.47	104.8
3	B	1.26	-5.56	-2.75	.19	1.71	12.7
3	C	1.87	-7.18	-2.70			16.4
50		1.15	-2.88	-2.35	-.04	2.73	36.8
70		1.30	-4.01	-2.50	.24	2.71	36.8
92		1.33	-4.58	-1.95	.63	2.87	12.6

**Table 98. Item Parameter and Fit Summary, Multiple-Choice Items,
Grade 6 Mathematics (N=9921)**

Item	a	b	c	Fit-Z	Item	a	b	c	Fit-Z
1	1.23	.72	.22	1.9	30	.73	-1.02	.18	4.0
2	.78	-.12	.27	1.9	32	1.18	-.07	.12	2.1
3	.58	-1.45	.18	1.1	33	.83	.61	.19	-.1
4	.61	-2.21	.20	.8	34	.77	.09	.16	7.0
6	1.58	.69	.31	2.0	36	.52	-.69	.17	1.9
8	.37	-3.33	.20	-.7	37	.43	-.82	.18	-.6
9	1.07	-.94	.35	.7	38	.61	-2.74	.20	.7
10	.87	1.22	.09	-.2	41	1.17	-.09	.19	-.6
11	.55	-.94	.15	.7	42	1.12	.47	.12	1.7
13	.68	.64	.07	4.4	43	.99	.72	.18	2.1
14	.83	-1.46	.20	-.9	45	.61	-.85	.20	2.3
16	.97	.62	.07	6.2	46	.60	-1.91	.20	-.4
17	.80	-.40	.37	5.3	48	.49	-.91	.20	1.3
18	.86	-1.13	.28	2.4	50	.52	-.96	.20	3.1
19	.98	-.82	.49	1.6	51	.69	-.17	.24	3.0
21	.78	-.56	.27	.5	52	.56	-2.42	.20	.2
22	.67	-.76	.07	4.5	53	.74	-.92	.12	.5
23	.64	-.48	.20	3.7	55	1.70	.95	.34	21.9
25	1.06	.72	.10	2.1	56	.94	1.01	.18	6.7
26	.50	-.87	.20	1.9	57	.93	-1.68	.18	4.6
27	.76	1.15	.24	-.1	58	.54	-1.21	.13	1.0
28	1.09	-.06	.34	-.6	59	.58	-2.06	.20	7.7
29	.61	.35	.23	.6					

**Table 99. Item Parameter and Fit Summary,
Constructed-Response Items, Grade 6
Mathematics (N=9921)**

Item	f	g1	g2	g3	g4	Fit-Z
5	1.19	-4.60	-1.15	-.49		.7
7	.96	-2.07	-.71			12.5
12	.89	-.72	-.18			30.4
15	1.23	-1.20	-.48	.26		5.4
20	.97	-2.55	-2.17	-1.05	-.30	16.4
24	.73	1.78	-.06			9.1
31	.73	-1.24	.29	.98		15.1
35	.60	-.91	-.94	-1.47		9.5
39	.99	1.57	2.76			10.8
40	1.24	-1.78	.87	.83	1.42	15.9
44	1.50	-3.05	-.32			20.1
47	.88	-1.88	-.59	2.26		21.2
49	.92	.40	.25			157.2
54	.84	-.32	-.43	1.76		3.3
60	.99	-3.32	.06	-.99	.55	21.3

Table 100. Item Parameter and Fit Summary, Multiple-Choice Items, Grade 7 Reading (N=9323)

Item	a	b	c	Fit-Z	Item	a	b	c	Fit-Z
4	1.00	-2.77	.20	1.1	43	.98	.08	.20	.6
5	.72	-1.54	.20	3.3	44	1.16	1.87	.05	35.6
6	.53	-1.17	.20	6.3	46	.52	-1.02	.20	6.1
7	.73	-2.19	.20	2.3	47	.65	-1.75	.20	5.5
10	.58	-1.56	.20	3.1	48	.82	-.26	.27	3.5
11	.65	-.24	.21	7.9	49	.65	-1.22	.17	2.6
12	.82	-.53	.23	.1	50	.90	.14	.25	7.7
13	1.33	.38	.24	7.1	51	1.20	-1.21	.17	.5
14	1.33	-.36	.32	1.3	52	1.21	.61	.20	3.9
15	.72	-.48	.15	6.5	53	1.08	-1.40	.26	-.2
16	.82	.80	.16	2.4	54	1.04	-.28	.23	.3
17	1.01	-1.59	.10	.6	55	.29	-.12	.20	7.3
18	1.29	.96	.26	10.0	100	.54	-1.17	.19	2.4
19	.98	-.02	.24	.5	101	.76	-.15	.21	-.3
20	.88	-.92	.15	2.8	102	.73	-1.71	.20	5.7
21	.60	-1.61	.17	1.3	103	.93	.43	.28	1.7
24	.80	-1.22	.11	-1.6	104	.61	-1.64	.20	1.0
26	1.03	-.51	.19	-.6	105	1.69	-1.23	.22	1.6
28	.62	.70	.20	46.7	106	1.44	-1.90	.21	.0
29	.62	.56	.14	.0	107	1.24	-1.05	.17	12.8
32	.75	-.55	.26	7.8	108	.81	-.65	.25	2.6
33	.71	-1.23	.12	.1	109	.69	-1.74	.20	.1
34	1.33	.52	.20	4.4	112	1.24	-.64	.29	-1.1
35	.96	-2.20	.20	2.9	113	1.17	-.04	.23	1.7
36	.98	-2.22	.16	10.8	114	.23	-1.15	.20	3.5
39	1.53	-1.13	.25	-.1	115	1.46	.13	.29	1.3
40	.90	-.91	.15	.1	116	.48	.77	.21	5.0
41	.44	2.46	.32	-.4	117	.69	.35	.19	.7
42	.55	2.43	.32	1.5					

Table 101. Item Parameter and Fit Summary, Constructed-Response Items, Grade 7 Reading (N=9323)

Item	f	g1	g2	g3	g4	Fit-Z
8	.55	.07	-1.40	-.24	-1.20	6.7
9	.85	-.21	-.84			.9
22	.53	-.16	-.24	.07		5.8
23	.70	-.53	-.66			2.1
25	.62	.64	.79			8.8
27	.79	-1.41	-.84			8.4
31	.62	-.42	-.46			3.8
37	.38	-.12	1.02	.29		33.0
38	.54	2.01	-3.64	-.71		7.5
45	.55	-.25	-2.04			8.1
98	.60	-1.03	-.42	-.96		7.1
99	.95	-.22	-.40			3.2
110	.98	-1.63	-1.90			5.8
111	.57	-.99	1.49			20.9

Table 102. Item Parameter and Fit Summary, Multiple-Choice Items, Grade 7 Writing (N=9536)

Item	a	b	c	Fit-Z	Item	a	b	c	Fit-Z
56	.23	1.77	.20	3.3	77	.60	-1.54	.20	2.6
57	.93	-1.71	.20	.9	78	.52	.35	.43	1.5
58	.75	-2.03	.20	-.1	79	.73	-2.91	.20	6.2
59	.52	-1.75	.13	.8	80	.65	-2.12	.20	2.4
60	.85	-.83	.28	.2	81	.92	.69	.25	3.1
61	.31	1.44	.20	1.0	82	.97	.25	.22	3.2
62	1.08	3.20	.07	16.1	83	.86	-1.31	.09	10.5
63	.85	-.55	.16	-.2	84	.78	-.82	.24	-.6
64	.47	-1.09	.11	-.6	85	.41	-.01	.12	1.9
65	.64	2.60	.25	.9	86	.89	-.51	.12	4.6
66	.95	-1.50	.19	-.3	87	.38	-.37	.20	8.8
67	.93	-1.75	.20	-.6	88	.54	.26	.15	10.8
68	.80	-2.35	.15	.2	89	.65	-1.33	.09	6.5
69	.82	-1.42	.20	3.7	90	.62	-1.18	.15	6.2
71	.55	-1.15	.20	3.2	91	1.17	-1.45	.20	-.2
72	.48	.11	.07	4.3	92	.75	-.22	.12	2.9
73	.39	-.92	.13	.4	93	1.09	-1.59	.30	0.0
74	.67	-1.60	.10	1.6	94	.85	-.75	.16	2.0
75	.58	-1.66	.15	10.9	95	.70	-.27	.12	4.0
76	.46	-.17	.09	1.5	96	.74	1.25	.15	7.8

Table 103. Item Parameter and Fit Summary, Constructed-Response Items, Grade 7 Writing (N=9536)

Item	Part	f	g1	g2	g3	g4	Fit-Z
1		1.23	-5.85				1.4
2	A	.94	.82				.6
2	B	.61	.05				2.2
2	C	1.05	.11				.2
2	D	.95	-.64				5.3
2	E	1.29	-.30				1.6
2	F	.69	.58				5.2
3	A	1.17	-3.88	-2.37	-.38	1.35	64.4
3	B	1.31	-5.99	-2.55	-.26	1.46	4.9
3	C	1.72	-7.04	-1.84			1.1
30		1.08	-2.86	-2.02	.15	2.21	73.6
70		1.25	-3.52	-2.76	-.57	1.73	19.5
97		1.33	-4.54	-2.57	.08	2.51	22.5

Table 104. Item Parameter and Fit Summary, Multiple-Choice Items, Grade 7 Mathematics (N=10,737)

Item	a	b	c	Fit-Z	Item	a	b	c	Fit-Z
1	.74	.21	.23	-0.4	32	.66	-.80	.20	0.2
2	.76	.31	.36	6.1	33	.82	.55	.20	0.3
4	.59	-.68	.11	0.1	35	.61	-.73	.12	1.8
5	.54	.08	.20	4.9	36	.81	.96	.19	0.0
6	1.44	.61	.34	2.9	37	.88	-.59	.29	0.9
8	1.04	.81	.27	4.4	38	.52	-.30	.15	0.6
9	.87	.89	.20	3.3	39	.85	.95	.08	2.0
10	.44	-3.07	.20	1.9	41	.93	-1.42	.20	2.3
12	1.45	1.07	.26	6.8	42	.74	-.48	.11	2.3
13	.47	.21	.16	5.0	44	1.28	.78	.37	6.4
14	1.28	.21	.23	3.3	45	.92	.29	.34	-1.3
16	1.07	-.26	.09	1.0	46	1.54	.60	.33	6.2
17	.58	.76	.27	1.4	47	.98	.33	.11	0.3
18	.85	.82	.16	1.6	48	.90	-.12	.25	0.5
19	.90	-.70	.49	3.6	50	1.17	.73	.22	0.7
21	.88	-.35	.33	0.9	51	1.11	-.69	.24	3.0
22	.74	-.19	.25	5.1	52	1.13	.57	.11	-0.6
23	.77	.13	.28	0.1	54	.32	-2.35	.20	2.5
25	1.76	1.63	.20	53.7	55	.88	.92	.18	1.1
27	.74	1.53	.06	0.2	56	.95	.54	.45	-0.7
28	.74	.24	.18	1.4	58	1.61	.26	.33	0.7
29	.66	-3.36	.20	6.3	59	1.10	-.16	.05	5.8
30	1.04	-.35	.11	2.6					

Table 105. Item Parameter and Fit Summary, Constructed-Response Items, Grade 7 Mathematics (N=10,737)

Item	f	g1	g2	g3	g4	Fit-Z
3	.75	1.57	-.25			21.0
7	.87	1.42	2.09	.30		17.2
11	1.06	1.82	1.44			11.6
15	.91	-.56	-.69	1.63		21.8
20	.91	-2.39	1.29	2.13	1.2	4.7
24	.98	-.40	-.26	.75		8.4
26	1.07	1.24	2.98			26.5
31	.99	1.90	-.07			1.5
34	1.13	-2.14	-1.27	.46		11.0
40	1.12	-1.83	.29	.87	1.4	9.4
43	.74	2.24	-1.35			37.1
49	.90	.23	-.25	.39		12.0
53	.94	.49	-.06			135.8
57	1.02	-2.11	-.65	2.05		26.6
60	.83	-.20	.32	.64	2.9	15.3

Table 106. Item Parameter and Fit Summary, Multiple-Choice Items, Grade 8 Reading (N=9188)

Item	a	b	c	Fit-Z	Item	a	b	c	Fit-Z
4	.28	-.95	.20	16.7	39	1.07	1.60	.06	19.8
5	.69	.58	.30	4.8	41	.51	-1.34	.20	3.3
6	.79	-1.59	.20	3.3	42	.59	-1.99	.20	5.9
7	.72	-1.21	.16	1.4	43	1.03	.06	.31	4.2
8	.65	-1.17	.20	1.7	44	.75	-.53	.15	.1
9	1.06	.81	.31	4.9	46	1.00	.73	.21	3.2
10	.80	.37	.21	1.4	91	.48	-.73	.20	16.5
11	.86	-2.28	.20	3.0	92	.91	-1.27	.25	2.2
12	1.16	-1.94	.20	-1.2	93	1.00	-.55	.26	.7
13	1.14	-2.02	.16	3.1	94	.91	.64	.20	6.3
14	.88	.33	.19	2.9	95	.79	2.06	.15	.4
19	.82	-2.01	.20	5.0	96	1.08	-1.59	.26	-.3
20	1.33	-1.35	.23	-.6	99	.90	1.83	.21	7.7
21	1.28	-1.90	.20	7.7	100	.67	.46	.24	2.4
22	.46	-2.36	.20	5.4	101	.50	.99	.19	3.6
23	.95	-.58	.25	-.4	102	.60	1.44	.19	-.3
24	.81	-.72	.19	1.4	103	.90	1.41	.21	3.5
27	.76	-.86	.20	4.6	105	1.16	.17	.21	5.5
28	.84	-1.17	.20	1.8	106	1.17	.49	.30	.7
29	1.25	.32	.23	3.0	107	1.07	1.34	.27	.7
30	1.08	-2.13	.20	5.7	108	1.16	-.77	.38	-.9
31	1.02	-2.17	.20	35.1	109	.65	-.70	.22	8.0
34	1.71	-1.30	.26	-.5	112	.98	-.64	.24	5.3
35	.97	-.95	.20	.3	113	1.43	-1.31	.21	-.3
36	.24	1.65	.20	3.1	114	1.40	-1.60	.42	-.6
37	.60	1.64	.27	1.4	115	.96	-1.01	.26	-.1
38	.94	-.21	.20	.4	116	1.16	-.24	.20	4.0

Table 107. Item Parameter and Fit Summary, Constructed-Response Items, Grade 8 Reading (N=9188)

Item	f	g1	g2	g3	g4	Fit-Z
15	1.02	-1.09	-1.32			5.7
16	.97	-.61	-.23			19.7
17	.74	-.76	-1.23			6.7
18	.91	-1.33	-.89			2.6
26	.75	-.81	-.38			4.8
32	.41	-.22	1.23	.44		23.2
33	.57	2.95	-4.76	-.73		6.0
40	.58	-.35	-2.17			8.5
45	1.18	-1.56	-.13	1.54		9.1
90	1.02	-2.23	.40			3.1
97	.71	-1.34	.64	.06		4.2
98	.75	-.84	-.69	-.16	1.42	4.2
104	1.29	.19	1.24			10.0
110	1.14	.40	.40			.8
111	.79	1.26	1.00			2.0

**Table 108. Item Parameter and Fit Summary, Multiple-Choice
Items, Grade 8 Writing (N=9537)**

Item	a	b	c	Fit-Z	Item	a	b	c	Fit-Z
47	.47	-1.90	.20	2.6	69	.43	-.41	.14	0.1
48	.59	.68	.15	1.2	70	.81	.94	.06	4.5
49	.74	-1.96	.15	.4	71	.67	-.11	.08	1.2
50	.56	-1.94	.15	3.0	72	.37	.66	.20	4.9
51	.86	-.46	.33	-.2	73	.76	-.68	.20	16.0
52	.96	-1.01	.38	1.8	74	.34	.10	.20	27.4
53	.73	.00	.32	1.4	75	.58	-.44	.15	13.4
54	1.25	-.78	.22	1.0	76	.89	-.60	.08	.4
55	.98	-1.71	.09	2.9	77	.94	-.97	.19	2.1
56	.81	-1.43	.08	.9	78	.33	1.29	.20	10.2
57	.50	-.88	.20	1.2	79	1.04	-1.46	.07	7.8
59	.84	-.68	.11	2.7	80	.86	.00	.19	.6
60	1.46	-1.45	.12	1.3	81	1.51	-1.38	.11	-.2
61	1.27	-1.56	.24	1.5	82	.61	.04	.14	.9
62	.34	.95	.10	.4	83	1.28	-1.30	.21	4.5
63	1.07	-1.13	.11	3.7	84	1.39	-1.36	.12	.2
64	.90	-1.24	.07	20.6	85	1.08	-1.43	.13	-1.3
65	.92	-1.23	.20	1.0	86	.49	-.95	.11	2.0
66	1.05	-1.88	.20	6.4	87	.82	-1.23	.20	19.3
67	.96	3.08	.07	5.0	88	.75	-.63	.19	.0
68	.83	-.41	.20	1.1					

**Table 109. Item Parameter and Fit Summary,
Constructed-Response Items, Grade 8 Writing
(N=9537)**

Item	Part	f	g1	g2	g3	g4	Fit-Z
1		1.33	-5.72				-1.0
2	B	.93	1.89				.9
2	C	.97	-.40				6.6
2	D	.87	-.78				-.5
2	E	1.18	-1.85				4.3
2	F	.94	-1.12				2.0
3	A	1.20	-2.99	-3.31	-1.01	1.86	70.8
3	B	1.47	-5.56	-3.76	-.96	2.24	8.2
3	C	1.63	-5.92	-3.78			2.3
25		1.40	-3.90	-2.79	.32	3.24	31.7
58		1.21	-3.62	-2.90	.44	2.62	36.2
89		1.36	-4.18	-2.74	.97	3.46	44.4

**Table 110. Item Parameter and Fit Summary, Multiple-Choice
Items, Grade 8 Mathematics (N=10,164)**

Item	a	b	c	Fit-Z	Item	a	b	c	Fit-Z
1	1.47	2.10	.05	13.6	32	.84	1.52	.18	1.0
2	1.30	.20	.17	1.3	33	1.77	1.11	.20	37.1
4	1.46	1.51	.11	2.0	35	1.26	.62	.12	20.9
5	.44	-2.85	.20	2.0	36	1.59	1.78	.22	.8
6	.73	.03	.20	.8	37	1.07	.37	.27	4.1
7	.92	.37	.18	4.1	38	1.30	.06	.18	2.0
8	.21	.30	.20	5.7	39	1.32	.80	.22	9.8
10	1.32	.67	.25	8.4	41	.70	1.46	.10	-1.0
11	.78	.46	.30	.0	42	.97	.59	.49	-.1
13	1.13	.50	.04	4.7	44	.87	-.24	.35	-.6
16	1.01	.83	.13	3.3	45	.64	.42	.12	3.2
17	.73	.29	.34	-1.0	46	.70	-1.45	.19	2.6
18	.92	-.42	.12	.8	47	1.23	-.95	.04	2.6
19	.74	.33	.21	.8	49	.90	.44	.44	2.6
21	.58	1.51	.05	2.2	51	1.11	.98	.04	.7
22	1.18	1.49	.25	8.3	52	1.12	-.64	.23	3.1
23	1.37	.44	.21	2.9	54	.89	-.12	.32	3.5
24	.43	-1.24	.20	2.5	55	1.05	.25	.26	1.0
26	.89	-.53	.10	1.3	56	.37	.21	.17	2.2
27	1.04	1.91	.07	.3	57	1.31	.33	.17	7.4
29	.92	2.26	.20	.1	58	.32	-2.42	.20	1.1
31	1.44	.54	.25	.0	59	1.45	1.56	.31	2.7

**Table 111. Item Parameter and Fit Summary,
Constructed-Response Items, Grade 8
Mathematics (N=10,164)**

Item	f	g1	g2	g3	g4	Fit-Z
3	0.94	0.71	-0.56			7.7
9	1.10	-1.32	0.37	-0.54		32.7
12	1.21	2.10	0.49			5.5
14	0.67	-0.72	2.13	1.62		56.7
20	1.36	1.51	1.04	2.10	1.17	7.6
25	1.51	3.44	1.93			5.3
28	1.14	-1.35	-0.42	0.14		12.3
30	0.78	2.83	-1.56			5.6
34	1.30	-0.26	2.07	2.20		5.4
40	1.13	-2.02	0.17	0.68	0.85	13.9
43	1.24	1.20	2.27			2.4
48	0.93	-0.35	-0.90	0.13		14.5
50	0.65	2.27	-1.34			38.4
53	1.22	0.10	0.63	1.18		28.7
60	1.05	-3.08	-1.15	2.22	0.02	16.1

**Table 112. Item Parameter and Fit Summary, Multiple-Choice
Items, Grade 8 Science (N=10,000)**

Item	a	b	c	Fit-Z	Item	a	b	c	Fit-Z
1	.91	2.45	.33	2.2	45	1.07	-.02	.24	1.7
3	.75	1.65	.21	.2	46	.88	.08	.27	.7
4	.89	1.29	.32	-.5	47	1.47	1.04	.20	16.4
5	.92	.48	.30	.2	48	.26	-.98	.20	16.9
6	.72	-.39	.26	1.3	49	.81	.02	.14	-.3
8	.94	1.25	.22	1.9	50	1.15	-1.22	.40	1.8
9	1.03	-.91	.32	.0	52	.91	1.78	.31	1.4
10	.94	.66	.25	5.5	53	.19	3.56	.20	9.5
11	.87	-.63	.20	5.5	54	1.33	-.99	.12	.6
12	.95	1.19	.37	.4	55	.68	-1.72	.20	1.5
13	.56	1.69	.28	.7	56	.86	-.07	.36	-.5
14	.87	.36	.28	-.9	57	.68	.95	.29	-.8
15	.78	-2.06	.20	2.1	58	1.65	2.41	.12	36.3
16	.76	-1.80	.20	2.1	59	.50	-1.15	.20	5.0
17	.69	.77	.30	1.0	60	.65	-1.59	.20	7.7
18	.63	.19	.23	.2	61	.47	.68	.14	1.2
25	.16	1.10	.20	1.1	63	.75	.25	.12	-.4
26	.97	.62	.47	1.3	64	.89	-2.41	.20	8.9
27	.79	-1.71	.20	2.1	65	.85	-1.81	.20	2.9
29	.62	1.90	.12	.3	67	1.42	-1.42	.12	.7
31	1.00	-.07	.11	-.4	68	.40	-1.81	.20	4.2
32	.84	.39	.13	.7	69	.75	.66	.16	1.1
34	.79	.77	.25	.4	70	.73	.16	.26	.2
35	.39	-1.55	.20	-.1	71	1.19	1.46	.13	-.2
36	.31	.98	.20	13.2	74	1.05	1.81	.24	8.4
37	.60	.79	.25	3.1	76	.88	-.90	.17	1.2
38	.96	.60	.30	.4	78	.76	-.93	.12	-.4
39	1.32	.99	.26	3.8	79	1.49	1.57	.16	-.2
41	1.43	1.87	.22	16.7	80	.73	-1.28	.14	-.4
42	.97	2.04	.38	-1.4	81	.79	-.57	.28	.3
43	.71	-1.67	.20	.5	82	.97	.09	.29	3.2
44	.54	-1.06	.20	.9	83	.93	-1.19	.30	-.4

**Table 113. Item Parameter and Fit Summary,
Constructed-Response Items, Grade 8
Science (N=10,000)**

Item	f	g1	g2	g3	g4	Fit-Z
2	1.61	.26	4.10			18.0
7	.83	0.17	.59			6.2
19	1.16	-.08				4.5
20	.55	0.30	-.49	-.56	.52	12.2
21	1.44	-.87				4.8
22	.72	-0.48				4.4
23	.93	-0.97				8.1
24	.80	-0.34	-0.21	-.08		7.8
28	1.23	.30	1.29			3.2
30	1.10	1.43				6.1
33	1.37	-.55	2.98			1.1
40	1.28	-.99	-.43			18.2
51	.92	1.94	2.92			2.0
62	1.22	.10	2.76			5.0
66	.77	-1.40	-1.88			6.6
72	1.43	.84	2.89			8.0
73	.79	1.29	3.41			-.5
75	1.11	-1.83				2.3
77	.83	.04				5.2
84	.81	.23				2.4

Table 114. Item Parameter and Fit Summary, Multiple-Choice Items, Grade 9 Reading (N=9042)

Item	a	b	c	Fit-Z	Item	a	b	c	Fit-Z
4	.94	.84	.40	-.8	41	.63	-.94	.20	4.5
5	.53	-.89	.19	-.3	43	1.03	-1.36	.15	2.1
6	.97	-2.47	.20	2.1	44	.81	-.39	.31	.0
7	.89	-.45	.27	.8	46	.77	-.60	.23	1.5
8	.82	-.67	.15	-.1	47	1.10	1.37	.18	3.3
11	.48	-1.34	.20	3.7	91	.51	-1.31	.20	20.8
12	.69	-.24	.26	1.7	92	.74	.59	.31	2.7
13	.84	.29	.15	1.9	93	1.17	-1.33	.22	.8
14	1.31	-.60	.35	.9	94	1.42	-.98	.29	.4
15	.63	1.40	.18	4.2	95	.21	1.76	.20	1.6
16	.95	-1.31	.27	-1.0	96	.23	-.38	.20	2.5
17	1.34	1.16	.25	1.1	97	.69	-1.11	.20	1.4
18	.76	-1.20	.22	3.0	98	.97	-1.10	.24	.4
19	.74	.04	.28	3.7	99	.63	.39	.31	3.6
20	.89	.59	.35	.5	101	.75	.90	.21	11.7
21	1.16	-.34	.27	.0	102	.78	1.81	.24	-1.2
26	.86	-1.68	.23	4.6	104	1.12	-.56	.32	.3
27	1.23	-1.40	.19	1.8	105	.82	-1.69	.14	5.1
28	1.13	-1.82	.12	-.3	106	.97	-1.42	.25	-1.3
29	.51	-2.01	.20	4.8	107	1.43	-1.10	.27	-.2
30	.64	-2.41	.20	4.6	108	.69	-.98	.25	6.1
31	1.66	1.06	.26	3.5	109	.69	-.21	.19	1.8
32	.57	-1.29	.20	1.6	110	1.17	.02	.30	1.1
33	.81	-.54	.24	4.3	111	.87	-.56	.14	.1
34	.90	-1.02	.43	.6	114	.77	1.52	.18	-.4
35	.98	-.36	.34	3.3	115	.64	.43	.26	1.5
36	.73	-.84	.17	2.8	116	.53	.77	.18	5.1
37	.37	-1.12	.20	6.2	117	.69	1.24	.21	.1
38	.92	-.45	.22	-.3	118	1.12	1.16	.25	6.5

Table 115. Item Parameter and Fit Summary, Constructed-Response Items, Grade 9 Reading (N=9042)

Item	f	g1	g2	g3	g4	Fit-Z
9	.48	-1.84	-1.14	.28		3.4
10	.63	-.44	-1.06	-.72		7.8
22	1.13	-1.23	-1.62			6.5
23	.96	-.36	-.04			27.8
24	.84	-.64	-1.32			3.4
25	.99	-1.10	-.49			-.8
39	.83	-.89	.52			16.5
40	1.06	-.19	.47			4.4
42	.65	-.95	.29	.66		6.4
45	1.22	-1.42	.48			21.2
100	.72	1.07	-.24	1.20		12.8
103	.80	-1.01	-1.06	-.58		5.7
112	.67	-1.43	.41	-.19		9.8
113	.67	-.65	-.80	-.57	1.15	2.5
119	1.24	-.05	.97			4.6

Table 116. Item Parameter and Fit Summary, Multiple-Choice Items, Grade 9 Writing (N=9218)

Item	a	b	c	Fit-Z	Item	a	b	c	Fit-Z
49	.81	-.05	.23	-.8	70	.85	-1.19	.09	4.9
50	.96	1.25	.15	7.8	71	.76	.05	.19	2.7
51	.66	-.14	.16	.2	72	.62	-.68	.15	.7
52	.59	-1.37	.20	3.0	73	.82	-1.24	.10	7.5
53	.71	.44	.21	.5	74	.60	.35	.20	1.6
54	.91	-.68	.22	1.4	75	1.18	-1.92	.20	3.0
55	.80	.36	.25	1.5	76	.86	1.37	.18	10.1
56	.47	-.32	.18	2.1	77	1.19	-.40	.27	1.8
57	.95	-1.37	.19	.7	78	.76	-.66	.20	.7
58	1.10	-.96	.28	1.0	79	1.03	-1.20	.18	-.1
59	.94	-.05	.29	.3	80	.92	-.06	.24	1.0
60	.68	-.95	.11	.3	81	.77	-.17	.13	3.9
61	.48	-.67	.14	3.3	82	1.41	-.90	.24	.8
62	.83	-1.04	.11	1.5	83	.34	.87	.20	6.1
63	.69	-1.66	.15	2.5	84	1.45	-1.45	.14	4.0
64	1.14	-1.34	.09	3.9	85	1.29	-1.36	.32	7.5
65	.33	.44	.20	24.5	86	1.25	-1.06	.22	.3
66	1.38	-.51	.28	.2	87	.48	-.31	.12	3.5
67	1.34	-1.36	.30	.3	88	.83	.47	.14	4.5
68	.90	.54	.11	-.2	89	.56	-.96	.20	-.3

Table 117. Item Parameter and Fit Summary, Constructed-Response Items, Grade 9 Writing (N=9218)

Item	Part	f	g1	g2	g3	g4	Fit-Z
1		1.31	-5.64				.5
2	A	.92	-2.23				.8
2	B	1.04	.66				2.4
2	C	.86	1.00				5.3
2	D	.45	-1.22				4.0
2	E	.99	-1.44				1.3
2	F	.43	-.51				8.5
3	A	1.34	-3.06	-3.64	-1.21	1.45	141.9
3	B	1.64	-5.65	-4.14	-1.19	1.79	5.4
3	C	2.02	-6.99	-3.76			3.0
48		1.12	-2.11	-3.53	-.51	1.78	24.8
69		1.05	-2.20	-3.72	-.41	1.69	16.5
90		1.07	-1.91	-3.22	-.39	1.53	32.1

**Table 118. Item Parameter and Fit Summary, Multiple-Choice
Items, Grade 9 Mathematics (N=10,024)**

Item	a	b	c	Fit-Z	Item	a	b	c	Fit-Z
1	.68	-.19	.19	.2	30	1.29	1.49	.19	3.8
2	.44	-.40	.20	1.1	31	.74	-.89	.10	2.0
4	1.30	1.96	.05	9.0	33	1.07	1.36	.11	6.4
5	.88	-.01	.34	.1	35	1.21	1.24	.21	1.9
6	1.18	1.42	.36	-.7	36	1.33	-.14	.18	.3
7	1.54	1.23	.12	4.3	37	.76	.58	.27	-.5
8	1.62	1.53	.21	4.3	38	.56	1.57	.27	-.4
9	.61	.06	.17	-.1	39	.63	.02	.19	9.1
11	1.23	-.32	.21	4.8	41	.60	.64	.07	1.8
12	.45	.78	.27	5.9	42	.74	-.28	.22	1.8
13	.60	-.53	.10	2.0	43	.93	-.83	.12	2.2
15	.75	-1.35	.21	4.0	44	.45	-2.50	.20	7.4
17	1.34	.34	.27	1.1	46	.94	.79	.39	.5
18	1.21	.39	.16	3.9	48	.82	-.35	.22	5.8
19	.37	-1.01	.20	3.3	50	.89	.21	.15	.4
21	.82	1.28	.18	.2	52	1.18	1.06	.10	5.4
22	.66	.03	.21	-.5	54	.94	.29	.41	.0
24	1.90	2.33	.14	6.4	55	1.92	1.68	.10	13.9
25	.80	2.17	.17	1.0	56	1.37	.01	.22	-.2
26	.79	-.95	.11	4.6	57	1.30	.68	.14	7.2
27	.77	2.41	.23	.4	58	1.21	-.09	.25	2.5
29	.94	1.58	.08	1.7	59	1.57	1.27	.17	-.1

**Table 119. Item Parameter and Fit Summary,
Constructed-Response Items, Grade 9
Mathematics (N=10,024)**

Item	f	g1	g2	g3	g4	Fit-Z
3	1.12	.40	2.57			-.4
10	.92	.91	1.25	.66		3.0
14	1.11	2.40	-.45			7.8
16	1.98	1.61	3.42	4.55		2.0
20	1.45	-.89	1.06	1.97	3.21	3.4
23	1.35	1.74	1.52			13.5
28	1.26	-.22	.34	.94		31.4
32	1.02	2.68	.14			6.3
34	1.30	-.09	1.75	1.87		5.1
40	.96	-1.62	-.71	.78	1.26	21.7
45	1.13	4.99	-1.69			1.7
47	1.76	3.64	3.66	3.74		3.9
49	1.00	-1.33	1.39			9.9
53	1.36	.23	2.44	1.37		1.8
60	1.17	-1.26	1.04	2.29	1.94	10.5

Table 120. Item Parameter and Fit Summary, Multiple-Choice Items, Grade 10 Reading (N=9115)

Item	a	b	c	Fit-Z	Item	a	b	c	Fit-Z
4	.53	-.40	.16	-.7	43	.79	-.53	.25	1.5
5	1.54	-.09	.37	6.0	44	.44	.35	.20	-1.0
6	.64	-1.60	.20	3.7	45	.49	2.06	.18	-.3
7	.47	-1.35	.20	1.9	47	.59	2.09	.25	1.5
8	.68	1.12	.08	-.6	48	.40	2.43	.26	-.3
9	.39	-1.29	.20	1.4	50	.93	-1.91	.20	7.3
10	.81	-.56	.33	6.0	51	.73	.98	.27	.2
13	.38	-1.64	.20	5.8	54	.67	-.22	.25	.6
15	.42	-1.95	.20	12.2	55	.75	-2.07	.21	8.6
16	.95	.17	.27	2.1	56	.86	-2.03	.20	8.2
17	.94	-.05	.18	.0	99	.43	-.35	.20	4.2
18	1.14	-.99	.30	.6	100	.74	.49	.12	1.0
19	.56	1.13	.17	2.8	101	.87	-.08	.24	.7
20	.87	-1.72	.18	-.3	102	.96	-.84	.21	1.3
22	.40	.44	.20	6.6	103	.65	-.51	.31	.7
23	.97	-.51	.31	1.1	105	.45	1.36	.15	1.2
24	.37	-.52	.20	-.1	106	.91	-.19	.21	.5
25	1.38	-1.44	.41	.4	108	1.04	-.92	.26	.4
26	.78	-.28	.19	.2	109	.94	-1.80	.13	8.7
27	.66	-.72	.20	4.4	110	1.36	-1.48	.25	-.4
28	.46	.05	.32	2.5	111	1.51	-1.33	.22	-.5
30	.75	-1.43	.20	1.7	112	.71	-1.35	.20	2.0
31	.98	-.95	.15	1.5	113	.76	-.38	.20	3.7
32	.39	1.23	.15	.3	114	1.13	-.36	.28	3.2
33	1.38	-.16	.32	3.1	115	1.14	-1.21	.08	6.1
36	1.50	1.34	.26	3.0	116	.48	-.72	.20	7.0
37	.72	-1.17	.21	-.9	117	1.07	-.75	.13	9.2
39	.22	-1.00	.20	6.9	118	1.43	-.57	.23	1.3
40	.86	-1.46	.20	2.7	119	1.41	-.97	.27	-.1
41	.47	-.62	.20	5.5	120	.39	.24	.20	3.2
42	.54	-1.96	.20	2.6					

Table 121. Item Parameter and Fit Summary, Constructed-Response Items, Grade 10 Reading (N=9115)

Item	f	g1	g2	g3	g4	Fit-Z
11	.91	-1.54	-.29			4.1
12	.52	-1.66	-1.12	.24		3.5
14	.61	-.60	-1.15	-.85		10.9
21	.76	.12	.59	-.72		23.9
34	.59	1.35	-1.59	-.81		2.9
35	.70	1.73	-3.45	-.74		4.4
38	.99	-1.64	1.18			8.1
46	.63	.64	-.26	.94	1.10	24.2
49	.79	.02	.98			16.8
52	.72	-.28	-1.37	-0.75		6.1
53	.71	-.98	.43	-0.33		2.6
104	.76	-1.02	.75	1.41		12.9
107	.78	-.74	-1.10	-.83		3.9
121	.85	-.41	-1.60	-1.04		8.8

Table 122. Item Parameter and Fit Summary, Multiple-Choice Items, Grade 10 Writing (N=9028)

Item	a	b	c	Fit-Z	Item	a	b	c	Fit-Z
57	.38	-.23	.15	3.4	77	.55	-1.21	.20	6.3
58	.59	-1.98	.20	.9	78	.58	-.30	.20	2.2
59	.86	-.97	.24	.2	79	.65	-1.24	.14	-.2
60	.68	-.11	.22	-.5	80	.63	.40	.14	1.7
61	.58	-1.25	.20	9.8	81	1.61	-1.51	.16	.9
62	1.04	-1.91	.15	1.3	82	.91	-1.29	.09	12.8
63	1.36	-1.67	.16	.2	83	.31	.22	.20	2.4
64	.49	-.85	.20	5.5	85	.79	-1.64	.19	-.5
65	1.08	-1.65	.25	-.2	86	.92	-.97	.16	3.7
66	1.25	-1.75	.16	.9	87	.74	-2.14	.20	8.4
67	.80	-1.13	.14	2.1	88	.94	-.07	.19	1.5
68	.65	-1.85	.20	2.5	89	1.36	-1.20	.35	1.3
69	.30	.35	.20	14.8	90	.49	-.33	.13	1.0
70	1.20	-1.66	.15	1.1	91	.85	.23	.14	3.5
71	.56	-1.06	.20	17.8	92	1.09	-1.06	.17	-.2
72	1.05	-.72	.21	.6	93	1.14	-1.23	.07	3.9
73	1.42	-1.52	.30	.5	94	1.37	-1.55	.19	2.2
74	.36	-.73	.20	-.1	95	.93	-1.02	.10	7.2
75	.47	.71	.17	3.2	96	.95	-.78	.14	4.9
76	.57	-.76	.20	-.4	97	.89	.43	.19	2.1

Table 123. Item Parameter and Fit Summary, Constructed-Response Items, Grade 10 Writing (N=9028)

Item	Part	f	g1	g2	g3	g4	Fit-Z
1		1.33	-5.33				.3
2	A	1.04	.73				.9
2	B	.69	-.69				.9
2	C	.97	.05				.2
2	D	.70	-2.01				3.9
2	E	1.16	-1.22				3.9
2	F	.61	-.85				4.7
3	A	1.36	-3.51	-3.65	-1.15	.68	25.7
3	B	1.66	-5.69	-3.86	-1.29	.83	5.7
3	C	1.89	-6.65	-3.16			4.0
29		1.18	-3.01	-2.44	-.16	1.84	56.1
84		1.21	-2.94	-2.78	-.57	1.52	41.0
98		1.13	-2.41	-2.64	-.66	1.42	59.4

Table 124. Item Parameter and Fit Summary, Multiple-Choice Items, Grade 10 Mathematics (N=9919)

Item	a	b	c	Fit-Z	Item	a	b	c	Fit-Z
1	.69	-.17	.16	-1.3	29	1.14	-.74	.18	1.1
2	.87	1.28	.10	.0	30	.63	2.65	.19	6.0
3	1.23	1.11	.21	26.3	31	1.35	.47	.18	5.3
4	1.13	1.28	.38	2.5	32	.57	-.11	.12	.9
5	.67	.46	.08	4.7	34	.59	.51	.06	2.0
7	.70	-.25	.21	2.4	36	.83	1.66	.41	.6
8	.93	.69	.23	2.4	37	.62	-.40	.21	8.1
9	.33	-1.52	.20	7.8	38	1.15	.71	.29	3.0
11	.67	-.46	.12	.6	41	.74	-.17	.22	.7
13	.79	-1.44	.20	2.1	42	1.06	.81	.07	.3
14	1.61	1.12	.45	13.6	44	1.01	1.65	.40	-1.3
16	.88	-.05	.35	-1.1	45	1.31	.89	.30	-.7
17	1.18	.84	.07	18.7	46	1.37	1.55	.10	.3
18	.50	-1.35	.20	4.2	47	1.25	.45	.16	.4
19	1.02	.47	.48	9.1	48	1.18	.73	.09	2.1
21	1.02	1.12	.16	1.3	50	1.71	1.46	.27	2.3
22	1.71	2.14	.33	1.3	51	.55	-1.15	.20	7.8
23	1.09	1.13	.10	1.2	54	1.22	.00	.34	.4
24	.60	.60	.22	-.4	55	1.61	.59	.29	5.6
26	.65	-.60	.10	3.5	56	1.17	1.40	.10	23.5
27	.55	1.56	.24	6.1	59	1.33	.99	.24	4.9
28	1.06	.27	.35	-.6					

Table 125. Item Parameter and Fit Summary, Constructed-Response Items, Grade 10 Mathematics (N=9919)

Item	f	g1	g2	g3	g4	Fit-Z
6	1.34	-.54	2.85			-.4
10	1.97	1.04	2.54	3.79		3.9
12	1.10	4.57	-1.42			2.4
15	1.96	.32	2.84	4.99		7.0
20	1.42	.26	.87	.81	2.12	2.4
25	1.00	-1.74	1.40			12.0
33	.93	-2.18	-.94	.12		12.1
35	1.12	2.42	-.12			12.7
39	1.55	-.24	2.96	2.10		3.9
40	1.08	-2.95	-.11	.68	2.06	30.0
43	1.26	-.64	1.73			84.6
49	1.14	-.17	3.23	1.50		13.5
52	1.90	1.39	2.50			8.7
58	1.96	2.87	3.75	4.60		7.5
60	1.60	.19	3.23	3.26	3.47	3.2

**Table 126. Items Flagged for DIF by Focal Group,
in Favor of (+) and against (*)**

| Content | Grade | Item # | Item Type | Focal Group[1] | | | | Obs-Pred |
				F	M	B	H	
RD	3	15	CR		*			.118
RD	4	36	CR	+	*		+	.152, .141, .174
RD	4	105	CR				*	.144
RD	5	8	CR				+	.124
RD	6	11	CR				+	.120
RD	6	45	CR			*		.124
RD	7	22	CR	+	*			.101, .107
RD	7	23	CR			*		.124
RD	7	25	CR		*			.108
RD	7	37	CR			*		.137
RD	7	98	CR	+				.105
RD	8	17	CR			*		.142
RD	8	100	MC			*		.143
RD	9	-						-
RD	10	12	CR			+		.130
RD	10	21	CR	*				.110
RD	10	24	MC	*	+			.110, .109
RD	10	34	CR	+	*			.102, .105
RD	10	35	CR	+	*			.116, .103
RD	10	49	CR			+		.165
RD	10	52	CR			+	+	.203, .114
RD	10	53	CR			+		.204
RD	10	121	CR	+				.102
WR	3	-						-
WR	4	71	CR			+		.131
WR	4	90	CR			+		.124
WR	5	-						-
WR	6	3B	CR			+		.112
WR	6	70	CR	+	*			.106, .112
WR	6	92	CR		*			.101
WR	7	3A	CR	+	*			.108, .110
WR	7	3B	CR	+	*			.102, .110
WR	7	70	CR	+	*			.106, .106
WR	8	58	CR		*			.104
WR	9	69	CR	+	*			.144, .146
WR	9	74	MC	*	+			.127, .115
WR	9	76	MC	*	+			.125, .109
WR	10	84	CR	+	*	+		.133, .141, .146
MA	5	51	CR			+		.102
MA	6	54	CR			*		.121
MA	7	24	CR	+	*			.118, .117
MA	7	60	CR	+	*			.124, .132
MA	8	28	CR	+		+		.100, .217
MA	8	48	CR			*		.120
MA	9	40	CR	+	*			.154, .118
MA	10	-						-
SC	8	20	CR	+	*			.180, .181
Span RD	3	-						-
Span RD	4	-						-
Span WR	3	-						-
Span WR	4	-						-

[1]F = Female; M = Male; B = African American; H = Hispanic; + = favoring; * = disfavoring.

**Table 127. Overall Estimates of Standard
Error of Measurement**

Assessment	Standard Error Of Measurement
Grade 3 Reading	30.28
Grade 4 Reading	18.77
Grade 5 Reading	20.22
Grade 6 Reading	18.85
Grade 7 Reading	18.10
Grade 8 Reading	17.51
Grade 9 Reading	18.18
Grade 10 Reading	18.40
Grade 3 Spanish Reading	16.06
Grade 4 Spanish Reading	14.37
Grade 3 Writing	23.68
Grade 4 Writing	16.71
Grade 5 Writing	16.91
Grade 6 Writing	17.51
Grade 7 Writing	18.89
Grade 8 Writing	20.12
Grade 9 Writing	21.47
Grade 10 Writing	25.05
Grade 3 Spanish Writing	19.25
Grade 4 Spanish Writing	13.41
Grade 5 Mathematics	17.49
Grade 6 Mathematics	18.19
Grade 7 Mathematics	17.97
Grade 8 Mathematics	20.20
Grade 9 Mathematics	24.93
Grade 10 Mathematics	21.49
Grade 8 Science	16.01

Table 128. Reading Scale Scores and Associated Standard Errors

RD 3		RD 4		RD 5		RD 6		RD 7		RD 8		RD 9		RD 10	
SS	SE	SS	SE	SS	SE	SS	SE	SS	SE	SS	SE	SS	SE	SS	SE
150	200														
175	171	175	198												
200	145	200	170	200	185										
225	121	225	145	225	159										
250	101	250	124	250	137	250	162								
275	83	275	105	275	117	275	136								
300	68	300	88	300	99	300	114	300	123						
325	54	325	74	325	84	325	95	325	104	325	129				
350	41	350	60	350	70	350	79	350	87	350	106	350	124	350	144
375	30	375	48	375	58	375	65	375	72	375	87	375	104	375	120
400	23	400	38	400	46	400	53	400	58	400	71	400	87	400	99
425	19	425	29	425	36	425	43	425	45	425	57	425	71	425	81
450	18	450	22	450	27	450	34	450	35	450	45	450	58	450	66
475	17	475	17	475	21	475	27	475	27	475	34	475	46	475	53
500	16	500	14	500	17	500	21	500	21	500	26	500	35	500	41
525	16	525	13	525	15	525	17	525	18	525	20	525	26	525	32
550	18	550	12	550	14	550	14	550	15	550	16	550	20	550	24
575	21	575	12	575	13	575	13	575	14	575	14	575	15	575	18
600	25	600	13	600	13	600	12	600	13	600	13	600	12	600	14
625	31	625	15	625	14	625	12	625	13	625	13	625	11	625	12
650	37	650	18	650	16	650	13	650	14	650	14	650	11	650	11
675	45	675	22	675	19	675	15	675	15	675	14	675	12	675	11
700	54	700	27	700	22	700	17	700	17	700	15	700	12	700	12
725	65	725	33	725	25	725	21	725	20	725	16	725	13	725	14
750	79	750	40	750	30	750	26	750	24	750	18	750	16	750	16
775	96	775	49	775	36	775	33	775	29	775	21	775	20	775	19
800	116	800	58	800	44	800	41	800	36	800	26	800	26	800	23
		825	68	825	53	825	50	825	44	825	32	825	33	825	28
		850	80	850	63	850	60	850	53	850	39	850	43	850	33
		875	92	875	75	875	71	875	63	875	49	875	55	875	40
		900	105	900	89	900	84	900	74	900	61	900	69	900	48
		925	120	925	105	925	99	925	87	925	76	925	86	925	58
		950	136	950	123	950	116	950	101	950	93	950	105	950	70
						975	136	975	116	975	112	975	127	975	85
										1000	134	1000	153	1000	102

Table 129. Writing Scale Scores and Associated Standard Errors

WR 3		WR 4		WR 5		WR 6		WR 7		WR 8		WR 9		WR 10	
SS	SE	SS	SE	SS	SE	SS	SE	SS	SE	SS	SE	SS	SE	SS	SE
150	97														
175	75	175	68												
200	58	200	57	200	77										
225	45	225	46	225	61	225	56	225	58						
250	34	250	37	250	49	250	46	250	48	250	51	250	57	250	57
275	25	275	29	275	39	275	38	275	40	275	42	275	46	275	47
300	19	300	24	300	31	300	32	300	35	300	35	300	38	300	40
325	15	325	20	325	25	325	29	325	31	325	30	325	32	325	34
350	12	350	17	350	21	350	26	350	28	350	26	350	28	350	30
375	11	375	16	375	18	375	24	375	25	375	23	375	25	375	27
400	11	400	15	400	16	400	20	400	22	400	21	400	22	400	24
425	13	425	14	425	14	425	17	425	19	425	18	425	21	425	21
450	15	450	14	450	13	450	15	450	18	450	16	450	20	450	19
475	19	475	14	475	13	475	14	475	17	475	15	475	18	475	17
500	23	500	15	500	14	500	14	500	16	500	15	500	17	500	17
525	28	525	16	525	16	525	15	525	17	525	16	525	17	525	17
550	34	550	19	550	18	550	17	550	17	550	18	550	18	550	19
575	39	575	22	575	21	575	19	575	18	575	20	575	19	575	20
600	45	600	27	600	24	600	21	600	20	600	22	600	20	600	22
625	53	625	33	625	29	625	24	625	22	625	23	625	22	625	24
650	65	650	41	650	34	650	27	650	24	650	25	650	23	650	26
675	82	675	51	675	42	675	30	675	27	675	27	675	26	675	29
		700	65	700	52	700	36	700	30	700	29	700	29	700	32
		725	83	725	64	725	43	725	33	725	32	725	33	725	37
				750	80	750	52	750	37	750	35	750	39	750	42
				775	99	775	64	775	43	775	40	775	46	775	49
						800	79	800	52	800	46	800	55	800	57
						825	97	825	65	825	56	825	66	825	66
						850	118	850	81	850	68	850	79	850	78
								875	101	875	84	875	94	875	91
								900	125	900	105	900	113	900	106
												925	134	925	124
														950	145

Table 130. Spanish Reading and Writing Scale Scores and Associated Standard Errors

SP RD 3		SP RD 4			Sp WR 3		SP WR 4	
SS	SE	SS	SE		SS	SE	SS	SE
					250	79		
					275	59		
300	124	300	119		300	43	300	61
325	92	325	92		325	32	325	49
350	66	350	70		350	23	350	38
375	47	375	52		375	17	375	29
400	33	400	38		400	15	400	22
425	22	425	26		425	14	425	18
450	15	450	18		450	14	450	16
475	12	475	13		475	15	475	16
500	12	500	11		500	15	500	14
525	13	525	10		525	17	525	12
550	15	550	10		550	20	550	11
575	18	575	11		575	24	575	12
600	22	600	12		600	30	600	15
625	29	625	15		625	37	625	18
650	38	650	21		650	45	650	22
675	52	675	29		675	55	675	27
700	70	700	38		700	66	700	33
725	93	725	49		725	79	725	41
750	121	750	62		750	94	750	50
		775	76				775	62
		800	91				800	75

Table 131. Mathematics and Science Scale Scores and Associated Standard Errors

MA 5		MA 6		MA 7		MA 8		MA 9		MA 10			SC 8	
SS	SE	SS	SE	SS	SE	SS	SE	SS	SE	SS	SE		SS	SE
200	61													
225	51	225	64											
250	43	250	54											
275	36	275	46	275	75									
300	31	300	39	300	61	300	82						300	51
325	26	325	34	325	48	325	69	325	115				325	39
350	23	350	29	350	37	350	58	350	92	350	80		350	31
375	19	375	25	375	29	375	49	375	73	375	67		375	24
400	17	400	22	400	23	400	41	400	58	400	56		400	20
425	15	425	20	425	18	425	35	425	46	425	47		425	17
450	13	450	18	450	16	450	28	450	37	450	40		450	15
475	13	475	17	475	14	475	23	475	30	475	33		475	14
500	13	500	16	500	13	500	19	500	24	500	28		500	14
525	14	525	15	525	13	525	16	525	20	525	23		525	13
550	16	550	15	550	14	550	14	550	16	550	19		550	13
575	18	575	15	575	15	575	12	575	14	575	16		575	13
600	22	600	16	600	16	600	11	600	12	600	13		600	14
625	26	625	18	625	19	625	11	625	11	625	12		625	15
650	32	650	21	650	23	650	12	650	10	650	11		650	17
675	39	675	26	675	28	675	14	675	10	675	10		675	21
700	48	700	31	700	34	700	17	700	11	700	11		700	26
725	60	725	38	725	42	725	22	725	14	725	12		725	33
750	73	750	45	750	50	750	30	750	19	750	16		750	42
775	89	775	54	775	60	775	39	775	27	775	21		775	52
800	109	800	64	800	71	800	51	800	37	800	29		800	64
		825	77	825	83	825	66	825	49	825	39			
				850	96	850	84	850	66	850	53			
						875	105	875	86	875	70			
						900	130	900	111	900	93			
								925	143	925	120			
										950	155			