Technical Report
Spring 2010 Operational Test
Administration

Washington, D.C. Comprehensive Assessment System (DC CAS)

September 9, 2010

CTB
McGraw-Hill

CTB/McGraw-Hill
Monterey, California 93940

Developed and published by CTB/McGraw-Hill LLC, a subsidiary of The McGraw-Hill Companies, Inc., 20 Ryan Ranch Road, Monterey, California 93940-5703.

Table of Contents

List of Tables

Section 1. Overview

This document describes the operational District of Columbia Comprehensive Assessment System (DC CAS) that was administered to students in Grades 3–8 and 10 in the spring of 2010 to assess students' skills in Reading and Mathematics; Grades 5 and 8 in Science; Grade 10 in Biology; and Grades 4, 7, and 10 in Composition. The DC CAS in Reading, Mathematics, and Science/Biology contains multiple–choice and constructed–response items that are administered under standardized conditions. The suggested time allotment for each section is approximately 30 to 40 minutes. The tests have suggested time limits instead of fixed time limits because the DC CAS tests are designed to measure proficiency in Reading, Mathematics, and Science/Biology with the goal of measuring Adequate Yearly Progress (AYP) as the program continues from year to year. The Composition assessment is a single essay prompt that is scored twice using two different rubrics. Composition and Science/Biology are not included in AYP calculations.

Purpose of the DC CAS Assessments in Reading, Mathematics, Science/Biology, and Composition

The primary purpose for the DC CAS is to measure the mastery of content standards of all District of Columbia (DC) public school students annually at the elementary and secondary levels in Reading, Mathematics, Science, and Composition in selected grades. These high quality, standards–based assessments are administered in Reading in Grades 3–8 and 10, Mathematics in Grades 3–8 and 10, Science in Grades 5 and 8, high school Biology, and Composition in Grades 4, 7, and 10. In summary, the assessments provide the foundation for an accountability system which enables the State to determine whether students and schools are making adequate yearly progress on DC content standards as required by the No Child Left Behind (NCLB) Act.

In addition, the assessments are used by district and school-based instructional staff to focus their lessons on state content standards and evaluate whether students and schools are achieving those standards. Parents use the results to monitor their children's educational progress and the effectiveness of their school and school district.

Highlights of This Technical Report

This technical report provides information, discussion, and assertions relevant to an evaluation of the validity of intended interpretations and uses of results from the 2010 DC CAS tests. The design of the test administration, content development and forms construction, statistical item review, classical item analysis, and item response theory analyses are covered. Following are some highlights of this report:

- Throughout, the report provides evidence, discussion, and assertions about the reliability of DC CAS scores and the validity of inferences about what students in the District of Columbia schools know and can do in relation to (a) DC content standards in Reading, Mathematics, Science/Biology, and Composition; and (b) the performance level descriptors that define levels of performance on DC CAS assessments in Grades 3–8 and high school.
- The report includes evidence about the 2010 DC CAS in sections on student participation, test content and design, reliability and validity, reliability and

accuracy of hand-scoring, DC CAS Percent Index scores, standard setting, Item Response Theory (IRT) and other analyses, student performance, and analyses of field test items.

- Throughout the report, shaded text indicates sections of the report that provide evidence that is directly relevant to the S*tandards and Assessment Peer Review Guidance*, Critical Elements (January 12, 2010; see http://www.ed.gov/policy/elsec/guid/saaprguidance.pdf).

Suggestions for How to Use This Technical Report

Technical reports for assessment programs are the primary means for test developers and assessment program managers to communicate with test users (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2009, p. 67). The standards require technical reports to document, for example, rationales and recommended uses for tests (Standard 6.3) and technical characteristics such as score reliability and validity of score interpretations (Standard 6.5). Because of the technical nature of developing, implementing, and validating achievement tests like the DC CAS, technical reports target audiences with some level of technical training and understanding.

This technical report is written to document procedures and results from developing, analyzing, and validating the 2010 DC CAS. It also is organized to facilitate finding information easily. Users of the report can use the report in several ways:

- Read the report from front to back.
- Scan section headers and sub-headers and read selected sub-sections.
- Locate specific topics in the table of contents (e.g., Section 7. IRT Analyses; *Internal Consistency Reliability* in Section 5. Evidence for Reliability and Validity).
- Review specific tables.
- Locate data and text that provide evidence relevant to one or more critical elements in the S*tandards and Assessment Peer Review Guidance*, Critical Elements (January 12, 2010; see http://www.ed.gov/policy/elsec/guid/saaprguidance.pdf).

Section 2. Test Content, Design, and Development for Reading, Mathematics, Science/Biology, and Composition

This section contains information relevant to the *Standards and Assessment Peer Review Guidance*, Critical Element 5.1:

Has the State outlined a coherent approach to ensuring alignment between each of its assessments…based on grade–level achievement standards, and the academic content standards and academic achievement standards the assessment is designed to measure?

A key piece of validity evidence is provided by the procedures used to develop the test's content and the alignment of items with the test blueprint and specifications. By setting forth a description of the events that took place in a test's development, we establish evidence of validity for the DC CAS based on test development procedures and test content.

Evidence of validity based on test content includes information about the test and item specifications. Test development involves creating a design framework from the statement of the achievement construct to be measured. The design for the 2010 test is based on test specifications that were developed in 2006 and 2007. Design elements include numbers and types of items and score points allocated to each content strand in each content area test.

Table 1 includes the description of DC CAS content strands, which also are reporting categories, applicable for all grades in Reading, Mathematics, Science/Biology, and Composition.

Table 1. DC CAS 2010 Test Strand Descriptions: Reading, Mathematics, Science/Biology, and Composition

Reading		
Content Strand		**Description of Items**
1	Language Development	Items of this category measure students' ability to identify meanings of words using prior knowledge, word structure, and/or context.
3	Informational Text	Items of this category measure students' ability to read, comprehend, and respond to informational passages.
4	Literary Text	Items of this category measure students' ability to read, comprehend, and respond to literary passages.

Mathematics		
Content Strand		**Description of Items**
1	Numbers & Operations	Items of this category measure students' ability to use numbers and number relationships.
2	Algebra	Items of this category measure students' ability to use algebraic methods to describe patterns and functions.
3	Geometry	Items of this category measure students' ability to use geometric concepts, properties, and relationships.
4	Measurement	Items of this category measure students' ability to use tools and techniques to measure.
5	Data Analysis	Items of this category measure students' ability to use data analysis, statistics, and probability.
Science Grade 5		
Content Strand		**Description of Items**
1	Scientific Inquiry	Items of this category measure students' ability to create and analyze scientific investigations.
2	Science and Technology	Items of this category measure students' ability to identify examples of technology and advances in science and technology that have impacted society.
3	Earth Science	Items of this category measure students' ability to describe Earth's position in the solar system and its weather and climate.
5	Physical Science	Items of this category measure students' ability to describe how objects are affected by force, motion, and changes in temperature.
7	Life Science	Items of this category measure students' ability to describe characteristics of organisms and how the environment influences their survival.

Science Grade 8		
Content Strand		**Description of Items**
1	Scientific Thinking and Inquiry	Items of this category measure students' ability to use graphics and models to describe scientific phenomena.
2	Structure of Matter	Items of this category measure students' ability to describe the characteristics and behavior of atoms.
3	Reactions	Items of this category measure students' ability to explain chemical reactions.
4	Forces / Density and Buoyancy	Items of this category measure students' ability to describe how density, buoyancy, and force affect objects.
5	Conservation of Energy	Items of this category measure students' ability to describe different types of energy and energy transformation.
High School Biology		
Content Strand		**Description of Items**
1	Scientific Inquiry	Items of this category measure students' ability to apply scientific information to different situations.
2	Biochemistry (Chemistry of Living Things)	Items of this category measure students' ability to describe molecules that are important to living things.
3	Cell Biology	Items of this category measure students' ability to describe important structure and functions of a cell.
4	Genetics	Items of this category measure students' ability to describe the structure and function of hereditary molecules.
5	Evolution	Items of this category measure students' ability to explain evolution from modern and historical perspectives.
6	Plants / Mammalian Body	Items of this category measure students' ability to explain the function of major systems and processes that occur in plants and animals.
8	Ecology	Items of this category measure students' ability to describe ecosystems and factors that affect ecosystems.

For the Composition tests, prompts guide examinees to write coherent essays that require narration (Grade 4), explanation (Grade 7), and persuasion (Grade 10). Each student responds to one prompt.

2010 Test Design and Coverage of the Content Strands

This section contains information relevant to the *Standards and Assessment Peer Review Guidance*, Critical Element 4.1:

For each assessment, including all alternate assessments, has the State documented the issue of **validity** (in addition to the alignment of the assessment with the content standards), as described in the *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 1999), with respect to *all* of the following categories:

(d) Has the State ascertained that the scoring and reporting structures are consistent with the sub-domain structures of its academic content standards (i.e., are item interrelationships consistent with the framework from which the test arises)?

CTB's Research and Development teams, with the approval of the District of Columbia's Office of Assessment, continued in 2010 the design established by tests of the DC CAS administered in 2006, 2007, 2008, and 2009. That design is represented in Table 2. Tables 3-5 display the blueprints for the operational item sets for the 2010 DC CAS in Reading, Mathematics, and Science/Biology.

Table 2. DC CAS 2010 Test Design: Reading, Mathematics, and Science/Biology

Content	Operational Items	Anchor Items (Operational Subset)	Embedded Field Test Items, Total across Four Forms
Reading	45 MC, 3 CR = 54 points	Gr: 3, 5, 6, 7: 45 MC, 3 CR Gr. 4, 8: 44 MC, 3 CR Gr. 10: 38 MC, 3 CR	Gr. 3–6: 68 MC, 8 CR Gr. 7: 58 MC, 7 CR [1] Gr. 8, 10: 72 MC, 8 CR
Math-ematics	51 MC, 3 CR = 60 points	Gr. 3: 45 MC, 3 CR Gr. 4, 8: 46 MC, 3 CR Gr 5: 51 MC, 3 CR Gr. 6: 49 MC, 3 CR Gr. 7, 10: 47 MC, 3 CR	Gr. 3–8, 10: 56 MC, 8 CR
Science/ Biology	47 MC, 3 CR = 53 points	Gr 5: 23 MC, 1 CR Gr. 8, Bio.: 22 MC, 1 CR	Gr. 5, 8, Bio: 48 MC, 8 CR

***Note*.** MC= multiple choice item, CR = constructed response item.

[1] Field test item sets 2 and 3 share 11 common items.

Table 3. DC CAS 2010 Operational Test Form Blueprints: Reading

Grade	Content Strand		Number of MC Items	Number of MC Points	Number of CR Items	Number of CR Points	Number of Points	% of Points
3	1	Language Development	11	11	0	0	11	20.37%
	3	Informational Text	16	16	1	3	19	35.19%
	4	Literary Text	18	18	2	6	24	44.44%
		Total	45	45	3	9	54	100%
4	1	Language Development	11	11	0	0	11	20.37%
	3	Informational Text	15	15	1	3	18	33.33%
	4	Literary Text	19	19	2	6	25	46.30%
		Total	45	45	3	9	54	100%
5	1	Language Development	12	12	0	0	12	22.22%
	3	Informational Text	15	15	1	3	18	33.33%
	4	Literary Text	18	18	2	6	24	44.44%
		Total	45	45	3	9	54	100%
6	1	Language Development	9	9	0	0	9	16.67%
	3	Informational Text	16	16	1	3	19	35.19%
	4	Literary Text	20	20	2	6	26	48.15%
		Total	45	45	3	9	54	100%
7	1	Language Development	10	10	0	0	10	18.52%
	3	Informational Text	13	13	1	3	16	29.63%
	4	Literary Text	22	22	2	6	28	51.85%
		Total	45	45	3	9	54	100%
8	1	Language Development	9	9	0	0	9	16.67%
	3	Informational Text	12	12	1	3	15	27.78%
	4	Literary Text	24	24	2	6	30	55.56%
		Total	45	45	3	9	54	100%
10	1	Language Development	11	11	0	0	11	20.37%
	3	Informational Text	16	16	1	3	19	35.19%
	4	Literary Text	18	18	2	6	24	44.44%
		Total	45	45	3	9	54	100%

Note. MC is Multiple Choice and CR is Constructed Response

Table 4. DC CAS 2010 Operational Test Form Blueprints: Mathematics

Grade	Content Standard		Number of MC Items	Number of MC Points	Number of CR Items	Number of CR Points	Number of Points	% of Points
3	1	Number Sense & Operations	16	16	1	3	19	31.67%
	2	Patterns, Relations, & Algebra	11	11	0	0	11	18.33%
	3	Geometry	5	5	1	3	8	13.33%
	4	Measurement	8	8	0	0	8	13.33%
	5	Data Analysis, Statistics, & Probability	11	11	1	3	14	23.33%
		Total	51	51	3	9	60	100%
4	1	Number Sense & Operations	19	19	0	0	19	31.67%
	2	Patterns, Relations, & Algebra	9	9	1	3	12	20.00%
	3	Geometry	5	5	1	3	8	13.33%
	4	Measurement	9	9	0	0	9	15.00%
	5	Data Analysis, Statistics, & Probability	9	9	1	3	12	20.00%
		Total	51	51	3	9	60	100%
5	1	Number Sense & Operations	18	18	0	0	18	30.00%
	2	Patterns, Relations, & Algebra	12	12	1	3	15	25.00%
	3	Geometry	9	9	0	0	9	15.00%
	4	Measurement	6	6	1	3	9	15.00%
	5	Data Analysis, Statistics, & Probability	6	6	1	3	9	15.00%
		Total	51	51	3	9	60	100%
6	1	Number Sense & Operations	16	16	1	3	19	31.67%
	2	Patterns, Relations, & Algebra	13	13	1	3	16	26.67%
	3	Geometry	7	7	0	0	7	11.67%
	4	Measurement	8	8	0	0	8	13.33%
	5	Data Analysis, Statistics, & Probability	7	7	1	3	10	16.67%
		Total	51	51	3	9	60	100%
7	1	Number Sense & Operations	17	17	0	0	17	28.33%
	2	Patterns, Relations, & Algebra	13	13	1	3	16	26.67%
	3	Geometry	6	6	1	3	9	15.00%
	4	Measurement	5	5	1	3	8	13.33%
	5	Data Analysis, Statistics, & Probability	10	10	0	0	10	16.67%
		Total	51	51	3	9	60	100%

Grade	Content Standard		Number of MC Items	Number of MC Points	Number of CR Items	Number of CR Points	Number of Points	% of Points
8	1	Number Sense & Operations	17	17	0	0	17	28.33%
	2	Patterns, Relations, & Algebra	13	13	1	3	16	26.67%
	3	Geometry	9	9	0	0	9	15.00%
	4	Measurement	5	5	1	3	8	13.33%
	5	Data Analysis, Statistics, & Probability	7	7	1	3	10	16.67%
		Total	51	51	3	9	60	100%
10	1	Number Sense & Operations	9	9	1	3	12	20.00%
	2	Patterns, Relations, & Algebra	14	14	1	3	17	28.33%
	3	Geometry	8	8	1	3	11	18.33%
	4	Measurement	7	7	0	0	7	11.67%
	5	Data Analysis, Statistics, & Probability	13	13	0	0	13	21.67%
		Total	51	51	3	9	60	100%

Note. MC is Multiple Choice and CR is Constructed Response

Table 5. DC CAS 2010 Operational Test Form Blueprints: Science/Biology

Grade	Content Standard		Number of MC Items	Number of MC Points	Number of CR Items	Number of CR Points	Number of Points	% of Total Points
5	1	Scientific Inquiry	10	10	1	2	12	22.64%
	2	Science & Technology	7	7	0	0	7	13.21%
	3	Earth Science	10	10	1	2	12	22.64%
	5	Physical Science	10	10	0	0	10	18.87%
	7	Life Science	10	10	1	2	12	22.64%
		Total	47	47	3	6	53	100%
8	1	Scientific Thinking and Inquiry	10	10	0	0	10	18.87%
	2	Structure of Matter	9	9	2	4	13	24.53%
	3	Reactions	7	7	0	0	7	13.21%
	4	Forces/Density and Buoyancy	10	10	1	2	12	22.64%
	5	Conservation of Energy	11	11	0	0	11	20.75%
		Total	47	47	3	6	53	100%
High School	1	Scientific Inquiry	8	8	0	0	8	15.09%
	2	Biochemistry	5	5	0	0	5	9.43%
	3	Cell Biology	6	6	1	2	8	15.09%
	4	Genetics	8	8	0	0	8	15.09%
	5	Evolution	5	5	0	0	5	9.43%
	6	Plants/Mammalian Body	9	9	0	0	9	16.98%
	8	Ecology	6	6	2	4	10	18.87%
		Total	47	47	3	6	53	100%

Note. MC is Multiple Choice and CR is Constructed Response

Composition Test

The Composition test includes one prompt per grade. In Grade 4, students write a personal narrative; in Grade 7, a descriptive-expository essay; and in Grade 10, a persuasive-argumentative essay. Each essay is scored once for Topic/Idea Development using a six-point rubric and once for use of English Language Conventions using a four-point rubric. Student scores on the two rubrics are summed so that total Composition scores range from 2 to 10.

Table 6. DC CAS 2010 Operational Test Form Scoring Rubrics: Composition

Grade	Scoring Rubric	Number of Points	% of Points
4, 7, 10	Topic/Idea Development	6	60%
	Language Conventions	4	40%
	Total possible points	10	100%

2010 Test Development Procedures

Test developers followed the test blueprints and DC CAS psychometric specifications to select the anchor item subsets and the operational items for the Reading, Mathematics, and Science/Biology tests. CTB test developers selected operational items from the pool of operational and field test items from the 2006, 2007, and 2008 administrations of DC CAS. A forms equating anchor set was selected for each grade/content area test using CTB's ItemWin software program and DC CAS content and statistical specifications. Each anchor set is a mini-blueprint of the full operational blueprint (see Tables 3-5 below). Test developers also used ItemWin to select the remainder of the operational items according to content and measurement constraints for measurement of reporting categories, as well as psychometric requirements, to the extent possible. The Reading and Mathematics items were selected primarily from among the 2006 operational test forms. The Science/Biology items were selected primarily from among the 2007 and 2008 operational test forms and 2007 field test items.

All proposed selections for operational forms were pre-equated in ItemWin to ensure that 2010 test forms are parallel to previous DC CAS test forms in terms of test difficulty and coverage of the DC CAS content standards, as specified in the 2010 test blueprints. CTB's Research team reviewed and approved the pre-equating results and requested revisions as necessary. Upon approval, page production of the forms began.

CTB test developers also developed items for the field test item sets that were embedded in the 2010 operational test forms. The 2010 test forms included four sets of field test items, rather than the usual two sets. OSSE chose to expand the field test item sets items to expand the pool of items available for operational use in 2011 and beyond.

All operational items were reviewed in previous years by CTB test developers and the DC Office of Assessment for content standards alignment and appropriateness in order to be eligible for inclusion in 2010 test forms. Similarly, all items newly written for DC

CAS by CTB test developers were evaluated by CTB content and style editors, supervisors, and managers prior to being taken to Content and Bias/Sensitivity Reviews in the District of Columbia. Content and Bias/Sensitivity Reviews were conducted to review items to be field tested in 2010 forms. DC Office of Assessment invited educators and community representatives to participate in the reviews. Following a training session conducted by CTB, the participants reviewed all items for content and grade appropriateness, and all items were accepted, revised, or rejected. The reviews were conducted during a workshop, and the reviewers used the criteria in the checklist in Appendix A to guide their decisions.

Analysis of the 2010 field test items will be completed subsequent to release of this operational technical report. Results from the field test analyses will be documented in a separate technical memo.

The Composition tests include one essay prompt per grade. Student essays are scored twice; see the section *Composition Test* and Table 6. *DC CAS 2010 Operational Test Form Scoring Rubrics: Composition*, below.

Organization of Test Booklets and Other Test Materials

All students in public and charter schools in the District of Columbia took one of the four DC CAS test forms. Each form included the same core set of operational items (with a forms equating anchor subset) and a set of unique embedded field test items. The four forms were spiraled together and packaged to ensure near equal distribution of the forms in classrooms and so that field test data were based on randomly equivalent groups.

Both Reading and Mathematics items were included in the same test books. Test books and answer booklets for Grades 4–8 and 10 were color-coded. Students in Grade 3 used scannable test books in which they recorded their answers. Students were also given calibrated card-stock rulers; Grade 10 students were given Mathematics reference cards. Students in grades 7, 8, and 10 were allowed to use calculators in session 1 only.

Each Reading and Mathematics test was divided into four sessions, for a total of eight sessions per grade level test. Each session included both multiple choice and constructed response items.

A similar configuration was used for the Science/Biology tests. Students responded to the test items in one of four test books. They recorded their answers in scannable answer documents. No manipulatives were provided. The Science/Biology tests were divided into three sessions, each with both multiple choice and constructed response items.

Composition test books were produced by CTB. The test books were scannable documents that included the following: directions to students, evaluation criteria, a writing prompt, three lined pages, and a biogrid. One form each was provided to each student for Grades 4, 7, and 10. The selected prompts were administered within the established two-week testing window. Each student responded to one prompt. Students were also issued two sheets of double-sided, lined, draft paper, specially developed for the Composition test, for planning their writing. Test administration instructions were included in the *Test Directions* for grades 4–8 and 10.

All student responses were scored with both a six-point holistic rubric for Topic Development and a four-point holistic rubric for Language Conventions. The rubrics used to score these items can be found in Appendix B.

Section 3. Student Participation

This section contains information relevant to *Standards and Assessment Peer Review Guidance*, Critical Elements 6.1 and 6.2:

6.1

1. Do the State's participation data indicate that all students in the tested grade levels or grade ranges are included in the assessment system (e.g., students with disabilities, students with limited English proficiency, economically disadvantaged students, race/ethnicity, migrant students, homeless students, etc.)?

2. Does the State report separately the number and percent of students with disabilities assessed on the regular assessment without accommodations, on the regular assessment with accommodations, on an alternate assessment against grade level standards, and, if applicable, on an alternate assessment against alternate achievement standards and/or on an alternate assessment against modified academic achievement standards?

6.2.

1. What guidelines does the State have in place for including all students with disabilities in the assessment system?

(a) Has the State developed, disseminated information on, and promoted use of appropriate accommodations to increase the number of students with disabilities who are tested against academic achievement standards for the grade in which they are enrolled?

Tests Administered

All DC schools administered the DC CAS tests between April 19 and April 30, 2010.

The tests administered were:

- Reading and Mathematics, Grades 3–8, and 10
- Composition, Grades 4, 7, and 10
- Science, Grades 5 and 8
- Biology, for those students in Grades 8–12 who were enrolled in a Biology course

Eligibility for Participation in DC CAS

The DC CAS *Test Chairperson's Manual* states that all students enrolled in District of Columbia schools must participate in DC CAS grade level test administrations, with one exception: A student with significant cognitive disabilities and whose Individualized Education Program (IEP) indicates that the student meets OSSE's established criteria may participate in the DC CAS alternate assessment portfolio. Students with disabilities and English learners who participate in DC CAS grade level administrations may be provided approved test administration accommodations that are specified by special education IEP teams, 504 teams, or ELL teams.

Participation in the 2010 DC CAS Test Administrations and Use of Data for Analysis and Score Reporting

Approximately 4,500 students were assessed in Reading, Mathematics, Science/Biology, and Composition at each tested grade. We report information below about participating students at each grade, numbers of examinees in special programs, and numbers of examinees in special education and English language learner (ELL) programs who received test administration accommodations.

Definition of *Valid Test Administration*

Only those students whose answer documents provided evidence of a valid test administration were included in the psychometric analyses described below. For the 2010 DC CAS, a test administration is valid if the answer document includes validly marked item responses, except under the following conditions.

- Three or more of the first five items are invalidly marked or omitted.
- The operational test total raw score equals zero and the sum of the operational and field test item valid responses is less than 5.
- All operational and field test items are omitted.

In addition, any answer sheets on which the marked grade level is inconsistent with the test grade level are excluded from psychometric analyses. (This rule is not applied for the Biology test because students from grades 8-12 are eligible.)

When we use the term “valid test administration” in subsequent sections of this report, we refer to any students who received a test score and was included in subsequent analyses because they met the requirements for a valid test administration.

Table 7. Numbers of Examinees with Valid Test Administrations in 2010: Reading

Grade	Students with Test Scores	Males	Females	Asian/ Pacific Islander	African American	Hispanic	White
3	4,932	2,471	2,440	86	3,822	594	404
4	4,841	2,430	2,393	75	3,797	586	362
5	4,518	2,297	2,208	61	3,636	505	301
6	4,537	2,286	2,230	54	3,732	457	267
7	4,389	2,180	2,186	48	3,721	394	207
8	4,542	2,237	2,271	62	3,844	426	176
10	4,416	2,078	2,264	58	3,742	380	162

Table 8. Numbers of Examinees with Valid Test Administrations in 2010: Mathematics

Grade	Students with Test Scores	Males	Females	Asian/ Pacific Islander	African American	Hispanic	White
3	4,956	2,487	2,447	90	3,830	604	405
4	4,868	2,448	2,399	81	3,800	600	365
5	4,536	2,310	2,213	64	3,639	514	304
6	4,561	2,293	2,245	57	3,739	469	269
7	4,403	2,189	2,189	60	3,713	403	207
8	4,547	2,236	2,277	65	3,822	450	176
10	4,388	2,058	2,255	58	3,717	377	161

Table 9. Numbers of Examinees with Valid Test Administrations in 2010: Science/Biology

Grade	Students with Test Scores	Males	Females	Asian/ Pacific Islander	African American	Hispanic	White
5	4,463	2,261	2,181	64	3,559	510	303
8	4,407	2,140	2,217	64	3,689	441	170
High School	4,113	1,925	2,060	52	3,441	378	141

Table 10. Numbers of Examinees with Valid Test Administrations in 2010: Composition

Grade	Students with Test Scores	Males	Females	Asian/ Pacific Islander	African American	Hispanic	White
4	4,555	2,266	2,263	73	3,551	556	348
7	4,229	2,080	2,110	50	3,559	388	202
10	3,837	1,711	2,039	51	3,235	331	149

When appropriate, students with disabilities who receive educational services under special education or Section 504 received test administration accommodations in one or more of four categories: timing/scheduling, setting, presentation, and response. For a student to receive an accommodation, the accommodation had to be in place during the school year and specified in the student's IEP or 504 plan. Students in English language learner programs received test administration accommodations in one or more of three categories: direct linguistic support oral, direct linguistic support written, and indirect linguistic support.

For more information on these accommodations, please refer to the DC CAS *Test Chairperson's Manual*.

Table 11. Number (and Percentage) of Students in Special Programs with Test Scores on the 2010 DC CAS in Reading, Mathematics, Science/Biology, or Composition

Grade	Students with Test Scores	Special Education	English Language Learner	Section 504	Title I Targeted	Home Schooling
Reading and/or Mathematics						
3	4,958	510 (10%)	406 (8%)	9 (0%)	332 (7%)	2 (0%)
4	4,872	574 (12%)	329 (7%)	14 (0%)	303 (6%)	1 (0%)
5	4,539	613 (14%)	246 (5%)	24 (1%)	265 (6%)	1 (0%)
6	4,568	626 (14%)	256 (6%)	14 (0%)	128 (3%)	.
7	4,422	676 (15%)	216 (5%)	13 (0%)	124 (3%)	1 (0%)
8	4,578	755 (16%)	177 (4%)	14 (0%)	124 (3%)	0
10	4,429	721 (16%)	191 (4%)	18 (0%)	376 (8%)	0
Science/Biology						
5	4,463	552 (12%)	223 (5%)	23 (1%)	275 (6%)	0
8	4,407	660 (15%)	204 (5%)	8 (0%)	125 (3%)	0
High School	4,113	577 (14%)	195 (5%)	4 (0%)	344 (8%)	0
Composition						
4	4,555	479 (11%)	280 (6%)	8 (0%)	302 (7%)	1 (0%)
7	4,229	577 (14%)	150 (4%)	12 (0%)	123 (3%)	0
10	3,837	550 (14%)	181 (5%)	17 (0%)	302 (8%)	1 (0%)

Note. Students who participated in more than one test administration are counted only once. Student subgroups are indicated in the Program Participation section on the biogrid on each student's answer document.

Table 12. Number (and Percentage) of Students Receiving One or More Special Education Test Administration Accommodations in Reading, Mathematics, Science/Biology, or Composition

Grade	Students with Test Scores	Timing/ Scheduling	Setting	Presentation	Response	Other	Students with Special Education Codes
Reading and/or Mathematics							
3	4,958	521 (11%)	538 (11%)	497 (10%)	260 (5%)	29 (1%)	510 (10%)
4	4,872	608 (12%)	607 (12%)	546 (11%)	306 (6%)	37 (1%)	574 (12%)
5	4,539	609 (13%)	583 (13%)	539 (12%)	338 (7%)	30 (1%)	613 (14%)
6	4,568	631 (14%)	596 (13%)	557 (12%)	432 (9%)	32 (1%)	626 (14%)
7	4,422	668 (15%)	618 (14%)	522 (12%)	484 (11%)	26 (1%)	676 (15%)
8	4,578	722 (16%)	663 (14%)	534 (12%)	533 (12%)	15 (0%)	755 (16%)
10	4,429	561 (13%)	538 (12%)	405 (9%)	393 (9%)	46 (1%)	721 (16%)
Science/Biology							
5	4,463	541 (12%)	520 (12%)	487 (11%)	288 (6%)	28 (1%)	552 (12%)
8	4,407	580 (13%)	552 (13%)	448 (10%)	401 (9%)	12 (0%)	660 (15%)
High School	4,113	452 (11%)	412 (10%)	355 (9%)	246 (6%)	42 (1%)	577 (14%)
Composition							
4	4,555	471 (10%)	488 (11%)	440 (10%)	232 (5%)	31 (1%)	479 (11%)
7	4,229	556 (13%)	520 (12%)	461 (11%)	353 (8%)	28 (1%)	577 (14%)
10	3,837	404 (11%)	395 (10%)	306 (8%)	223 (6%)	21 (1%)	550 (14%)

Note. Students who received more than one accommodation in a single content area test can be counted in multiple columns. Students who received accommodations in more than one content area test administration are counted only once. The Special Education code is recorded by test administrators on the biogrid section of each student's answer document. Accommodations provided are recorded by test administrators in the Accommodations section on the biogrid. Students for whom the Special Education bubble was not completed and who did receive test administration accommodations are not counted here.

Table 13. Number (and Percentage) of Students Receiving One or More Selected Special Education Test Administration Accommodations in Reading, Mathematics, Science/Biology, or Composition

Grade	Students with Test Scores	Breaks	Small Group and Individual Administrations	Read or Translate Test Questions (MA, SC and WR only)	Responses Dictated
Reading and/or Mathematics					
3	4,958	408 (8%)	522 (11%)	349 (7%)	69 (1%)
4	4,872	453 (9%)	588 (12%)	383 (8%)	79 (2%)
5	4,539	459 (10%)	562 (12%)	370 (8%)	74 (2%)
6	4,568	502 (11%)	579 (13%)	354 (8%)	48 (1%)
7	4,422	509 (12%)	589 (13%)	307 (7%)	26 (1%)
8	4,578	575 (13%)	649 (14%)	283 (6%)	37 (1%)
10	4,429	415 (9%)	515 (12%)	125 (3%)	55 (1%)
Science/Biology					
5	4,463	416 (9%)	503 (11%)	335 (8%)	72 (2%)
8	4,407	440 (10%)	537 (12%)	236 (5%)	31 (1%)
High School	4,113	329 (8%)	386 (9%)	107 (3%)	27 (1%)
Composition					
4	4,555	372 (8%)	473 (10%)	300 (7%)	68 (1%)
7	4,229	401 (9%)	501 (12%)	245 (6%)	27 (1%)
10	3,837	284 (7%)	378 (10%)	86 (2%)	37 (1%)

Note. Students who received more than one accommodation in a single content area test can be counted in multiple columns. Students who received accommodations in more than one content area test administration are counted only once. Accommodations are recorded by test administrators in the Accommodations section on the biogrid on each student's answer document.

Definitions:

Breaks: Timing/Scheduling codes 2, 3, and 5
Small Group and Individual Administrations: Setting codes 1, 3, and 4
Read or Translate Test Questions (Math, Science, or Composition only): Presentation codes 3 and 5
Responses Dictated: Response codes 3, 4, 6, and 7

ELLs were classified by their schools into one of four language proficiency levels. These levels are related to levels of language instruction services and participation in school instruction. In addition, students classified as ELL were eligible to receive test administration accommodations in one or more of three categories: direct linguistic support oral, direct linguistic support written, and indirect linguistic support. Table 14 displays information on ELL students. Details on accommodations are available in the DC CAS *Test Chairperson's Manual*.

Table 14. Number (and Percentage) of Students Receiving One or More English Language Learner Test Administration Accommodations in Reading, Mathematics, Science/Biology, or Composition

Grade	Students with Test Scores	Direct Linguistic Support - Oral	Direct Linguistic Support - Written	Indirect Linguistic Support	Other
Reading and/or Mathematics					
3	4,958	394 (8%)	141 (3%)	372 (8%)	5 (0%)
4	4,872	339 (7%)	152 (3%)	308 (6%)	1 (0%)
5	4,539	236 (5%)	120 (3%)	228 (5%)	2 (0%)
6	4,568	176 (4%)	112 (2%)	189 (4%)	1 (0%)
7	4,422	188 (4%)	108 (2%)	189 (4%)	3 (0%)
8	4,578	149 (3%)	67 (1%)	155 (3%)	2 (0%)
10	4,429	111 (3%)	98 (2%)	124 (3%)	.
Science/Biology					
5	4,463	202 (5%)	105 (2%)	186 (4%)	1 (0%)
8	4,407	168 (4%)	119 (3%)	193 (4%)	.
High School	4,113	146 (4%)	137 (3%)	164 (4%)	1 (0%)
Composition					
4	4,555	271 (6%)	175 (4%)	245 (5%)	1 (0%)
7	4,229	139 (3%)	85 (2%)	151 (4%)	1 (0%)
10	3,837	128 (3%)	99 (3%)	132 (3%)	.

Note. Students who received more than one accommodation in a single content area test can be counted in multiple columns. Students who received accommodations in more than one content area test administration are counted only once. The English Language Learner code is recorded by test administrators on the biogrid section of each student’s answer document. Accommodations provided are recorded by test administrators in the Accommodations section on the biogrid. Students for whom the English Language Learner bubble was not completed and who did receive test administration accommodations are not counted here.

Table 15. Number (and Percentage) of Students Coded for ELL Proficiency Levels 1–4 in Reading, Mathematics, Science/Biology, or Composition

Grade	Students with Test Scores	ELL: Access for ELL Proficiency Level 1	ELL: Access for ELL Proficiency Level 2	ELL: Access for ELL Proficiency Level 3	ELL: Access for ELL Proficiency Level 4
Reading and/or Mathematics					
3	4,958	40 (1%)	62 (1%)	162 (3%)	178 (4%)
4	4,872	36 (1%)	41 (1%)	133 (3%)	153 (3%)
5	4,539	28 (1%)	32 (1%)	68 (1%)	103 (2%)
6	4,568	35 (1%)	40 (1%)	84 (2%)	91 (2%)
7	4,422	41 (1%)	27 (1%)	71 (2%)	89 (2%)
8	4,578	44 (1%)	47 (1%)	88 (2%)	68 (1%)
10	4,429	12 (0%)	48 (1%)	87 (2%)	64 (1%)
Science/Biology					
5	4,463	30 (1%)	24 (1%)	51 (1%)	104 (2%)
8	4,407	42 (1%)	41 (1%)	66 (1%)	59 (1%)
High School	4,113	33 (1%)	38 (1%)	69 (2%)	66 (2%)
Composition					
4	4,555	13 (0%)	31 (1%)	120 (3%)	141 (3%)
7	4,229	11 (0%)	21 (0%)	56 (1%)	64 (2%)
10	3,837	3 (0%)	34 (1%)	79 (2%)	55 (1%)

Table 16. Number (and Percentage) of Students Receiving One or More Selected English Language Learner Test Administration Accommodations in Reading, Mathematics, Science/Biology, or Composition

Grade	Students with Test Scores	Direct Linguistic Support - Oral: Oral Reading of Test in English[1]	Direct Linguistic Support - Written: Bilingual Word to Word Dictionary	Indirect Linguistic Support: Extended Time
Reading and/or Mathematics				
3	4,958	45 (1%)	75 (2%)	365 (7%)
4	4,872	78 (2%)	105 (2%)	297 (6%)
5	4,539	50 (1%)	93 (2%)	214 (5%)
6	4,568	50 (1%)	74 (2%)	177 (4%)
7	4,422	31 (1%)	95 (2%)	184 (4%)
8	4,578	30 (1%)	47 (1%)	146 (3%)
10	4,429	18 (0%)	88 (2%)	124 (3%)
Science/Biology				
5	4,463	32 (1%)	85 (2%)	176 (4%)
8	4,407	44 (1%)	100 (2%)	187 (4%)
High School	4,113	47 (1%)	121 (3%)	162 (4%)
Composition				
4	4,555	17 (0%)	104 (2%)	236 (5%)
7	4,229	15 (0%)	68 (2%)	143 (3%)
10	3,837	8 (0%)	79 (2%)	130 (3%)

Note. Students who received more than one accommodation in a single content area test can be counted in multiple columns. Students who received accommodations in more than one content area test administration are counted only once. Accommodations are recorded by test administrators in the Accommodations section on the biogrid on each student's answer document.

Definitions:

Direct Linguistic Support—Oral: Oral Reading of Test in English (code 5)
Direct Linguistic Support—Written: Bilingual Word to Word Dictionary (code 2)
Indirect Linguistic Support: Extended time codes 1, 2, 3 and 5

[1] Oral reading of the Reading test is not allowed.

Section 4. Test Administration Guidelines and Requirements

This section contains information relevant to the *Standards and Assessment Peer Review Guidance*, Critical Elements 4.3, 4.5, and 6.2:

4.3
Has the State ensured that its assessment system is fair and accessible to all students, including students with disabilities and students with limited English proficiency, with respect to each of the following issues:

(a) Has the State ensured that the assessments provide an appropriate variety of accommodations for students with disabilities? *and*

(b) Has the State ensured that the assessments provide an appropriate variety of linguistic accommodations for students with limited English proficiency?

4.5
Has the State established clear criteria for the administration, scoring, analysis, and reporting components of its assessment system, including all alternate assessments, and does the State have a system for monitoring and improving the on-going quality of its assessment system?

6.2
1. What guidelines does the State have in place for including all students with disabilities in the assessment system?

(a) Has the State developed, disseminated information on, and promoted use of appropriate accommodations to increase the number of students with disabilities who are tested against academic achievement standards for the grade in which they are enrolled?

(b) Has the State ensured that general and special education teachers and other appropriate staff know how to administer assessments, including making use of accommodations, for students with disabilities and students covered under Section 504?

Overview

Administration of the DC CAS assessments each spring is managed by the Office of Assessment and Accountability, coordinated in each school by a Test Chairperson, and conducted by classroom teachers. Assessment office staff trained school Test Chairpersons on test administration guidelines and requirements using the 2010 *Test Chairperson's Manual*. They, in turn, trained all test administrators and proctors. Test administrators administer all DC CAS assessments according to requirements and steps in the 2010 *Test Directions*.

The *Test Chairperson's Manual* directs Test Chairpersons to follow the procedures for training test administrators and proctors on required procedures for administering each test and maintaining test security before, during, and after test administrations. It also provides information on available accommodations for students with disabilities and English language learners.

The *Test Directions* covers similar topics and requirements. In addition, it provides instructions on scheduling test administrations, preparing students for the test administration, using standardized testing procedures, and verbatim instructions for administering each test to students. It also provides information on available accommodations for students with disabilities and English language learners.

Guidelines and Requirements for Administering DC CAS

The *Test Chairperson's Manual* indicates that DC CAS administrations should be scheduled to ensure that all students have adequate time to respond to all test items under unhurried conditions. It also describes testing condition requirements to ensure that students can feel as comfortable as possible and are not distracted during administration. The manual requires each Test Chairperson to complete a testing Site Observation Report to ensure that adequate testing conditions can be provided. It also contains instructions on distributing test materials to test administrators, retrieving the materials, accounting for 100% of all secure materials, shipping the materials to CTB for processing, and maintaining security of the materials at all times and throughout the entire process.

The *Test Chairperson's Manual* and *Test Directions* provide detailed information on available test administration accommodations for students with disabilities and English language learners. It specifies approved accommodations that maintain standard testing conditions (e.g., reading only Mathematics test questions to examinees) and identifies accommodations that are considered modifications to the test which will result in invalidated test scores (e.g., assisted reading of Reading passages).

The *Test Chairperson's Manual* and *Test Directions* specify accommodations approved for students with disabilities in the following areas: timing/scheduling (e.g., providing breaks between prescribed timing sections of the tests), setting (e.g., individual and small group administrations), presentation (e.g., reading of Mathematics [only] test questions), and response accommodations (e.g., dictating responses). The *Test Chairperson's Manual* and *Test Directions* specify accommodations approved for English language learners in the following areas: direct linguistic support oral, direct linguistic support written, and indirect linguistic support. Both manuals indicate that test administrators must record on the student's answer document all test administration accommodations that are provided.

CTB provides test administration sessions for school Test Chairpersons in the month prior to test administration. School Test Chairpersons are required to conduct training sessions, and all school staff who will handle test materials must attend these sessions. School Test Chairpersons are explicitly required in the *Test Chairperson's Manual* to oversee the test administrations in their schools. They are required to ensure that test materials are available in adequate numbers and that school staff adhere to test security requirements, track materials by using security checklists, report breaches if they occur, document disruptions during testing, sign test materials in and out each day, account for 100% of secure test materials, and report missing or damaged materials immediately to CTB Customer Service.

Materials Orders, Delivery, and Retrieval

Customer orders were managed in CTB's Online Enrollment System. Schools updated and validated their enrollments or indicated non-participation. CTB used the results for order fulfillment.

Prior to shipment of materials, barcodes were applied to the secure materials for the purpose of secure inventory tracking (a description of the Secure Inventory process is provided later in this section). Corresponding security checklists were also produced. Daily tracking reports were provided to the Office of the State Superintendent of Education (OSSE) for the purpose of monitoring the deliveries.

The appropriate district and school staff were previously trained to maintain security and monitor quantities of materials. Shortly after delivery, they unpacked and reviewed materials to ensure readiness for administration, as described in a previous section of this report (Guidelines and Requirements for Administering DC CAS). In the event that the materials received were not sufficient for administration, a short/add window functioned to permit CTB customer service to process requests for additional materials while maintaining a secure inventory.

After the test administration was complete, the materials were packaged for retrieval and picked up according to a verified schedule. Daily tracking reports also served for OSSE to monitor retrievals. When the materials were back in CTB's custody, all books with security barcodes were accounted for as described in the following section of this report, Secure Inventory.

Secure Inventory

To further support the full range of test security requirements that are the reality for today's State assessments, CTB has instituted a comprehensive Test Security/Test Inventory System. This system was created using industry best practices. Upon request, CTB further customized a security model to precisely match the needs of DC CAS security requirements. This security model for the DC CAS assessment maintains its own list of material deliverables and services from assessment barcoding to inventory checking and shipment tracking, as described in the steps below.

1. Secure materials are barcoded at the printer, vertically banded, and inventoried. Barcode files are sent to CTB. Packing lists and test materials are sent to the schools.

2. Materials are distributed into the schools.

3. Following the test administration, school staff separate secure and non-secure materials and package them for return to CTB following *Test Chairperson's Manual* instructions.

4. The dedicated/secure carrier contacts the schools to schedule retrieval of their materials on a specified date.

5. Scorable secure documents are accounted for during answer document scanning, and non-scorable secure documents are scanned into an inventory

return system. Materials sent to the wrong CTB facility are forwarded to the appropriate site, as needed.

6. Missing Materials Reports are sent to OSSE for resolution once scanning is completed. Given a list of security barcodes shipped, minus the barcode numbers already received, the remaining list is considered to be Missing Inventory.

7. OSSE contacts schools and reports back to CTB on findings, including additional books that have been located, contaminated books that could not be returned to CTB, and damaged or destroyed books where no barcode was available for scanning.

8. CTB processes additional, received inventory and approved exceptions, and produces a final missing inventory report.

As of June 22, 2010, approximately 99.86% of secure materials were accounted for; only 106 secure test booklets were missing for 2010 administration, compared to 1,184 test booklets unaccounted for in 2009.

Section 5. Evidence for Reliability and Validity

The *Standards and Assessment Peer Review Guidance* (dated January 12, 2010) requires states to develop evidence in five categories to support the validity of interpretations of state assessment results that are consistent with intended purposes: evidence based on (a) test content, (b) the test's relationships with other variables, (c) examinee response processes, (d) the test's internal structure, and (e) positive and negative consequences of interpreting and using test scores. In addition, the guidance requires states to provide evidence on (a) score reliability and sources of error, including traditional score reliability estimates (e.g., internal consistency coefficients), classical standard errors of measurement, and item response theory (IRT) conditional standard errors; (b) examinee proficiency level classification accuracy and consistency estimates, and error estimates for aggregates (e.g., percentages of examinees in each proficiency level); and (c) estimates of the accuracy of year-to-year changes in scores. Finally, the guidance identifies other characteristics of state assessments that support valid interpretations of test scores, including fairness and accessibility; comparability of results; procedures for testing administration, scoring, analysis, and reporting; and efforts to ensure valid interpretations and warranted uses of results.

This technical report focuses specifically on 2010 DC CAS test development procedures and psychometric evaluation procedures and results. Section 5 of the report provides evidence relevant to the critical elements identified in each subsection. Other reliability and validity evidence is available in sources beyond this technical report.

Construct, Purpose, and Interpretation of Scores

This section contains information relevant to *Standards and Assessment Peer Review Guidance*, Critical Element 4.1:

For each assessment, including all alternate assessments, has the State documented the issue of **validity** (in addition to the alignment of the assessment with the content standards), as described in the *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 1999), with respect to *all* of the following categories:

(a) Has the State specified the purposes of the assessments, delineating the types of uses and decisions most appropriate to each?

As stated in Section 1, the primary purpose for the DC CAS is to measure the progress of all District of Columbia public school students annually at the elementary and secondary levels in Reading, Mathematics, Science/Biology and Composition in selected grades. These high quality, standards-based assessments are administered in Reading in Grades 3–8 and 10, Mathematics in Grades 3–8 and 10, Science in Grades 5 and 8, high school Biology, and Composition in Grades 4, 7, and 10. In summary, the assessments provide the foundation for an accountability system which enables the State to determine whether students and schools are making adequate yearly progress on DC content standards as required by the NCLB Act.

In addition, the assessments are used by district and school-based instructional staff to focus their lessons on state content standards and evaluate whether students and schools are achieving those standards. Parents use the results to monitor their children's educational progress and the effectiveness of their school and school district.

The evidence and arguments in Section 5 are relevant to and support the validity of these intended interpretations and uses of DC CAS test scores

Internal Consistency Reliability

This section contains information relevant to the *Standards and Assessment Peer Review Guidance*, Critical Element 4.2:

For each assessment, including all alternate assessments, has the State considered the issue of **reliability**, as described in the *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 1999), with respect to *all* of the following categories:

(a) Has the State determined the reliability of the scores it reports, based on data for its own student population and each reported subpopulation?

The degree of score reliability that is required for an interpretation of an individual student's test score must be carefully considered. Individual score reliability is estimated using internal consistency coefficients that are computed on all student responses in each grade and content area of the DC CAS. They are computed using the operational items administered to all students in a grade and content area. Generally, the number of students who took a DC CAS 2010 operational form and were included in the calibration sample was approximately 4,500 for each grade and content area. Unless otherwise noted, all data reported are from the operational calibration data.

The various reliability coefficients, reported in Table 17, differ slightly in their assumptions. The preferred coefficient for these tests is the stratified alpha coefficient. This coefficient is most appropriate for tests comprised of a combination of multiple–choice (MC) and constructed–response (CR) items, as in the DC CAS tests. Table 17 also contains Cronbach's alpha and Feldt-Raju score reliability estimates, which we discuss below.

Cronbach's alpha reliability coefficient is frequently used to assess internal consistency. This measure is used when both MC and CR items are in a test. The alpha reliability is based on a single test administration and provides reliability estimates that equal the average of all split-half reliability coefficients that would have been obtained on all possible divisions of the test into halves. This measure of reliability is the lower bound of a test's score reliability.

The stratified coefficient alpha is another internal consistency score reliability index. It measures the internal consistency of a test that contains both multiple–choice and constructed-response items. The stratified alpha treats the multiple-choice and constructed-response sections as separate subtests, estimates the reliability of the two subtests, and combines those estimates to estimate total test score internal consistency.

The Feldt-Raju index is a third index of internal consistency. It is also designed for mixed-format tests. Unlike the stratified alpha that stratifies the items based on the number of score points, the Feldt-Raju corrects the underestimation of Cronbach's alpha which assumes that tests are parallel in classical test theory terms; mixed format tests are more appropriately assumed to be congeneric.

As a rule of thumb, reliability coefficients for test scores that are equal to or greater than 0.8 are considered acceptable for tests of moderate lengths. All of the reliability indices calculated provide evidence that these tests are performing as expected and that they support inferences about what students know and can do in relation to the content knowledge and skills that the tests target.

Table 17. Internal Consistency Reliability Coefficients for the 2010 DC CAS Operational Tests

Content	Grade	Students with Test Scores	Number of Items	Alpha	Stratified Alpha	Feldt-Raju
Reading	3	4,928	48	0.930	0.936	0.935
	4	4,829	48	0.922	0.926	0.925
	5	4,510	48	0.930	0.933	0.933
	6	4,520	48	0.915	0.917	0.918
	7	4,382	48	0.917	0.922	0.922
	8	4,526	48	0.909	0.914	0.914
	10	4,394	48	0.924	0.931	0.929
Mathematics	3	4,944	54	0.931	0.935	0.936
	4	4,865	54	0.921	0.925	0.925
	5	4,533	54	0.917	0.922	0.922
	6	4,548	54	0.932	0.936	0.937
	7	4,390	54	0.925	0.932	0.932
	8	4,527	54	0.902	0.908	0.908
	10	4,359	54	0.920	0.925	0.926
Science	5	4,458	50	0.883	0.884	0.885
	8	4,393	50	0.867	0.868	0.870
Biology	High School	4,097	50	0.831	0.833	0.835

Note. Case counts (i.e., numbers of students with test scores) in this and all other tables may differ slightly. Rules for counting cases and including and excluding them from counts and statistics are different for classical item analyses, IRT calibrations and equatings, and total test summaries.

The stratified alpha reliabilities for all content areas and grades is, on average, 0.91. This is strong evidence for the reliability of scores for Reading, Mathematics, Science, and Biology tests. The lowest reliability was in Biology (0.83).

Internal consistency reliability estimates for examinee subgroups appear in Appendix C.

Reliabilities of Content Strand Scores

The alpha reliability coefficients of each strand score reported for the 2010 DC CAS are presented in Tables 18–20. The degree of reliability that is required to interpret these strand scores, as for any test score, must be carefully considered. These coefficients are computed on all student responses in each grade and content area for each content strand. The internal reliability estimates for these strand scores, which include as few as

5 items and as many as 26, range between 0.28 and 0.86, though many are 0.60 and higher. Interpretation of strand score internal consistency reliabilities requires caution because of the small numbers of items that make up each strand score. Likewise, interpretation of strand scores for individual examinees requires caution.

Table 18. Coefficient Alpha Reliability for Reading Strand Scores

Grade	Content Strand		Number of Items	Reliability
3	1	Language Development	11	0.7797
	3	Informational Text	17	0.8268
	4	Literary Text	20	0.8403
	Total Number of Items on DC CAS		48	--
4	1	Language Development	11	0.7466
	3	Informational Text	16	0.8137
	4	Literary Text	21	0.8257
	Total Number of Items on DC CAS		48	--
5	1	Language Development	12	0.7648
	3	Informational Text	16	0.8183
	4	Literary Text	20	0.8600
	Total Number of Items on DC CAS		48	--
6	1	Language Development	9	0.7071
	3	Informational Text	17	0.7823
	4	Literary Text	22	0.8373
	Total Number of Items on DC CAS		48	--
7	1	Language Development	10	0.7157
	3	Informational Text	14	0.7863
	4	Literary Text	24	0.8405
	Total Number of Items on DC CAS		48	--
8	1	Language Development	9	0.6826
	3	Informational Text	13	0.6966
	4	Literary Text	26	0.8508
	Total Number of Items on DC CAS		48	--
10	1	Language Development	11	0.6826
	3	Informational Text	17	0.8372
	4	Literary Text	20	0.8374
	Total Number of Items on DC CAS		48	--

Table 19. Coefficient Alpha Reliability for Mathematics Strand Scores

Grade	Content Strand		Number of Items	Reliability
3	1	Number Sense & Operations	17	0.7976
	2	Patterns, Relations, & Algebra	11	0.7762
	3	Geometry	6	0.5647
	4	Measurement	8	0.6400
	5	Data Analysis, Statistics, & Probability	12	0.7777
	Total	Number of Items on DC CAS	54	--
4	1	Number Sense & Operations	19	0.8270
	2	Patterns, Relations, & Algebra	10	0.7348
	3	Geometry	6	0.5134
	4	Measurement	9	0.6033
	5	Data Analysis, Statistics, & Probability	10	0.6287
	Total	Number of Items on DC CAS	54	--
5	1	Number Sense & Operations	18	0.8133
	2	Patterns, Relations, & Algebra	13	0.7343
	3	Geometry	9	0.5974
	4	Measurement	7	0.6162
	5	Data Analysis, Statistics, & Probability	7	0.5803
	Total	Number of Items on DC CAS	54	--
6	1	Number Sense & Operations	17	0.8272
	2	Patterns, Relations, & Algebra	14	0.8145
	3	Geometry	7	0.4898
	4	Measurement	8	0.6512
	5	Data Analysis, Statistics, & Probability	8	0.7016
	Total	Number of Items on DC CAS	54	--
7	1	Number Sense & Operations	17	0.8036
	2	Patterns, Relations, & Algebra	14	0.7512
	3	Geometry	7	0.6512
	4	Measurement	6	0.5290
	5	Data Analysis, Statistics, & Probability	10	0.7274
	Total	Number of Items on DC CAS	54	--
8	1	Number Sense & Operations	17	0.7621
	2	Patterns, Relations, & Algebra	14	0.7412
	3	Geometry	9	0.4797
	4	Measurement	6	0.4500
	5	Data Analysis, Statistics, & Probability	8	0.6271
	Total	Number of Items on DC CAS	54	--
10	1	Number Sense & Operations	10	0.6779
	2	Patterns, Relations, & Algebra	15	0.7730
	3	Geometry	9	0.6848
	4	Measurement	7	0.5719
	5	Data Analysis, Statistics, & Probability	13	0.7453
	Total	Number of Items on DC CAS	54	--

Table 20. Coefficient Alpha Reliability for Science/Biology Strand Scores

Grade	Content Strand		Number of Items	Reliability
5	1	Scientific Inquiry	11	0.6318
	2	Science & Technology	7	0.6775
	3	Earth Science	11	0.6480
	5	Physical Science	10	0.4756
	7	Life Science	11	0.6128
	Total Number of Items on DC CAS		50	--
8	1	Scientific Thinking and Inquiry	10	0.7052
	2	Structure of Matter	11	0.5843
	3	Reactions	7	0.3339
	4	Forces/Density and Buoyancy	11	0.5723
	5	Conservation of Energy	11	0.6003
	Total Number of Items on DC CAS		50	--
High School	1	Scientific Inquiry	8	0.3459
	2	Biochemistry	5	0.2784
	3	Cell Biology	7	0.5323
	4	Genetics	8	0.4409
	5	Evolution	5	0.3373
	6	Plants/Mammalian Body	9	0.5004
	8	Ecology	8	0.5210
	Total Number of Items on DC CAS		50	--

Conditional Standard Error of Measurement

This section contains information relevant to the *Standards and Assessment Peer Review Guidance*, Critical Element 4.2:

For each assessment, including all alternate assessments, has the State considered the issue of **reliability**, as described in the *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 1999), with respect to *all* of the following categories:

(a) Has the State quantified and reported within the technical documentation for its assessments the conditional standard error of measurement and student classification that are consistent at each cut score specified in its academic achievement standards?

Whereas reliability coefficients indicate the degree of consistency in test scores, the standard error of measurement (SEM) indicates the degree of unreliability in test scores. The standard error is an estimate of the standard deviation of observed scores to expect if an examinee were retested under unchanged conditions. Conditional standard deviations of observed scores can be found for each score level. The conditional estimate of measurement error increases as the number of items that coincide with examinees' levels of performance decreases. Generally, there are few students with extreme scores; these score levels are measured less accurately than moderate scores. If all of the items are very difficult or very easy for examinees, the error of measurement will be larger than when the items' difficulties are distributed across the ability levels of the students being tested.

In addition to classic internal consistency reliability coefficients, the SEM based on IRT is also provided as reliability evidence for DC CAS scores. The IRT SEM provides conditional standard errors that are specific to each scale score. These standard errors were estimated as a function of the scale scores, using IRT. Accuracy of measurement is especially important when applied to individual scores. The IRT-based SEM indicates the expected standard deviation of observed scores if an examinee at a specific level of ability were tested repeatedly under unchanged conditions. Tables 21–23 list the number correct to scale score values, along with their associated SEM values, for Reading, Mathematics, and Science/Biology.

Table 21. DC CAS 2010 Number Correct to Scale Score Conversions with Associated Standard Errors of Measurement (SEM): Reading

Raw Score	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 10	
	Scale Score	SEM	Scale Score	SEM	Scale Score	SEM	Scale Score	SEM	Scale Score	SEM	Scale Score	SEM	Scale Score	SEM
0	300	30	400	36	500	36	600	38	700	36	800	38	900	33
1	300	30	400	36	500	36	600	38	700	36	800	38	900	33
2	300	30	400	36	500	36	600	38	700	36	800	38	900	33
3	300	30	400	36	500	36	600	38	700	36	800	38	900	33
4	300	30	400	36	500	36	600	38	700	36	800	38	900	33
5	300	30	400	36	500	36	600	38	700	36	800	38	900	33
6	300	30	400	36	500	36	600	38	700	36	800	38	900	33
7	300	30	400	36	500	36	600	38	700	36	800	38	900	33
8	300	30	400	36	500	36	600	38	700	36	800	38	900	33
9	300	30	400	36	500	36	600	38	700	36	800	38	911	22
10	300	30	411	25	500	36	600	38	712	24	814	24	919	14
11	313	17	419	16	517	19	617	21	720	16	823	15	923	10
12	318	12	424	11	523	13	624	14	725	11	827	11	926	8
13	322	9	427	9	527	9	628	10	728	9	830	8	929	7
14	324	7	430	7	529	7	631	8	731	7	833	7	931	6
15	326	6	432	6	531	6	633	7	733	6	835	6	932	5
16	328	5	434	6	533	5	635	6	735	6	837	5	934	5
17	330	5	435	5	534	5	637	5	736	5	838	5	935	4
18	331	5	437	5	536	4	638	5	738	5	840	5	936	4
19	332	4	438	4	537	4	640	4	739	4	841	4	938	4
20	334	4	439	4	538	4	641	4	741	4	842	4	939	4
21	335	4	441	4	539	3	642	4	742	4	844	4	940	4
22	336	4	442	4	540	3	643	3	743	4	845	4	941	3
23	337	4	443	4	541	3	644	3	744	4	846	4	942	3
24	338	4	444	4	542	3	645	3	745	4	847	4	943	3
25	339	3	445	3	543	3	646	3	746	3	848	4	944	3
26	340	3	446	3	544	3	647	3	747	3	849	3	945	3
27	341	3	447	3	545	3	647	3	748	3	850	3	946	3
28	342	3	448	3	546	3	648	3	749	3	851	3	946	3
29	343	3	449	3	546	3	649	3	750	3	852	3	947	3
30	344	3	450	3	547	3	650	3	751	3	853	3	948	3
31	345	3	451	3	548	3	651	3	752	3	854	3	949	3
32	346	3	452	3	549	3	652	3	753	3	855	3	950	3
33	347	3	453	3	550	3	652	3	754	3	856	3	951	3
34	348	3	454	3	551	3	653	3	755	3	857	3	952	3
35	349	3	455	3	552	3	654	3	755	3	858	3	953	3
36	350	3	456	3	553	3	655	3	756	3	859	3	953	3

Raw Score	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 10	
	Scale Score	SEM	Scale Score	SEM	Scale Score	SEM	Scale Score	SEM	Scale Score	SEM	Scale Score	SEM	Scale Score	SEM
37	351	3	457	3	554	3	656	3	757	3	860	3	954	3
38	352	3	458	3	555	3	657	3	758	3	861	3	955	3
39	353	4	459	3	556	3	658	3	759	3	862	3	956	3
40	355	4	460	3	557	3	659	3	760	3	863	4	957	3
41	356	4	461	4	558	3	660	3	761	3	864	4	959	3
42	357	4	463	4	560	3	662	3	762	3	865	4	960	3
43	358	4	464	4	561	4	663	4	763	3	867	4	961	3
44	360	4	466	4	563	4	665	4	764	3	868	4	962	3
45	362	4	467	4	564	4	666	4	765	3	870	4	964	4
46	363	5	469	4	566	4	668	4	767	3	871	4	965	4
47	365	5	471	5	568	4	670	5	768	4	873	5	967	4
48	367	5	473	5	571	5	672	5	770	4	875	5	969	4
49	370	6	476	6	573	5	675	6	772	4	877	6	971	5
50	373	6	480	7	577	6	679	8	774	5	880	6	973	5
51	377	7	484	8	581	7	684	9	777	6	884	7	977	6
52	382	9	490	10	588	9	691	12	781	7	889	9	981	8
53	391	13	499	14	599	14	699	16	789	11	897	13	990	12
54	399	19	499	14	599	14	699	16	799	19	899	14	999	17

Table 22. DC CAS 2010 Number Correct to Scale Score Conversions with Associated Standard Errors of Measurement (SEM): Mathematics

Raw Score	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 10	
	Scale Score	SEM	Scale Score	SEM	Scale Score	SEM	Scale Score	SEM	Scale Score	SEM	Scale Score	SEM	Scale Score	SEM
0	300	20	400	35	500	31	600	35	700	34	800	40	900	33
1	300	20	400	35	500	31	600	35	700	34	800	40	900	33
2	300	20	400	35	500	31	600	35	700	34	800	40	900	33
3	300	20	400	35	500	31	600	35	700	34	800	40	900	33
4	300	20	400	35	500	31	600	35	700	34	800	40	900	33
5	300	20	400	35	500	31	600	35	700	34	800	40	900	33
6	300	20	400	35	500	31	600	35	700	34	800	40	900	33
7	300	20	400	35	500	31	600	35	700	34	800	40	900	33
8	300	20	400	35	500	31	600	35	700	34	800	40	900	33
9	300	20	400	35	500	31	600	35	700	34	800	40	900	33
10	300	20	400	35	500	31	600	35	704	30	800	40	900	33
11	300	20	400	35	500	31	608	27	714	19	800	40	906	27
12	302	18	408	27	509	22	617	18	720	14	816	24	914	19
13	307	13	417	19	515	16	622	13	724	11	823	16	919	14
14	311	11	422	14	519	12	626	10	727	9	828	12	923	11
15	314	9	425	11	523	10	629	8	729	7	831	9	926	10
16	317	8	428	9	526	8	631	7	732	7	834	8	929	8
17	319	8	431	8	528	8	633	6	733	6	836	7	932	8
18	322	7	433	7	530	7	635	6	735	6	838	6	934	7
19	324	7	435	6	532	6	636	5	737	5	839	6	936	6
20	326	6	437	6	534	6	638	5	738	5	841	5	938	6
21	327	6	438	5	536	6	639	4	740	5	842	5	939	5

Raw Score	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 10	
	Scale Score	SEM	Scale Score	SEM	Scale Score	SEM	Scale Score	SEM	Scale Score	SEM	Scale Score	SEM	Scale Score	SEM
22	329	6	440	5	537	5	640	4	741	4	844	5	941	5
23	331	5	441	5	539	5	641	4	742	4	845	4	942	5
24	332	5	442	5	540	5	643	4	743	4	846	4	943	5
25	334	5	443	4	542	5	644	4	745	4	848	4	945	4
26	335	5	445	4	543	5	645	4	746	4	849	4	946	4
27	336	5	446	4	544	4	646	3	747	4	850	4	947	4
28	338	5	447	4	545	4	646	3	748	4	851	4	948	4
29	339	5	448	4	547	4	647	3	749	4	852	4	949	4
30	340	4	449	4	548	4	648	3	750	4	853	4	950	4
31	342	4	450	4	549	4	649	3	751	4	854	4	951	4
32	343	4	451	4	550	4	650	3	752	3	855	4	952	4
33	344	4	452	3	551	4	651	3	753	3	856	3	953	3
34	345	4	453	3	552	4	652	3	754	3	857	3	954	3
35	347	4	454	3	553	4	653	3	755	3	858	3	955	3
36	348	4	455	3	554	4	654	3	756	3	859	3	956	3
37	349	4	456	3	556	4	654	3	757	3	860	3	957	3
38	350	4	457	3	557	4	655	3	758	3	861	3	958	3
39	351	4	458	3	558	4	656	3	759	3	862	3	959	3
40	352	4	459	3	559	4	657	3	760	3	863	3	960	3
41	354	4	460	3	560	4	658	3	761	3	864	3	961	3
42	355	4	461	3	561	4	659	3	762	3	865	3	962	3
43	356	4	462	3	562	4	660	3	763	3	866	3	963	3
44	357	4	463	3	563	4	661	3	764	3	867	3	964	3
45	358	4	464	3	564	4	662	3	765	3	868	3	965	3
46	360	4	465	3	566	4	663	3	766	3	869	3	966	3
47	361	4	467	4	567	4	665	4	767	4	870	3	967	3
48	362	4	468	4	568	4	666	4	768	4	871	3	968	3
49	364	4	469	4	569	4	667	4	769	4	872	3	969	4
50	365	4	471	4	571	4	668	4	771	4	874	3	971	4
51	367	4	472	4	573	4	670	4	772	4	875	3	972	4
52	368	4	474	4	574	5	671	4	774	4	876	3	974	4
53	370	5	476	4	576	5	673	4	776	4	878	4	975	4
54	372	5	478	5	578	5	675	5	778	5	879	4	977	4
55	375	5	480	5	581	5	677	5	780	5	881	4	979	5
56	377	6	483	5	583	6	680	6	783	5	884	5	982	5
57	381	7	487	6	587	7	683	6	786	6	887	6	985	6
58	386	9	492	8	591	8	687	8	791	7	892	8	990	8
59	395	13	499	12	599	12	695	12	798	11	899	12	998	12
60	399	15	499	12	599	12	699	14	799	12	899	12	999	13

Table 23. DC CAS 2010 Number Correct to Scale Score Conversions with Associated Standard Errors of Measurement (SEM): Science/Biology

Raw Score	Grade 5		Grade 8		High School	
	Scale Score	SEM	Scale Score	SEM	Scale Score	SEM
0	500	42	800	50	900	48
1	500	42	800	50	900	48
2	500	42	800	50	900	48
3	500	42	800	50	900	48
4	500	42	800	50	900	48
5	500	42	800	50	900	48
6	500	42	800	50	900	48
7	500	42	800	50	900	48
8	500	42	800	50	900	48
9	500	42	800	50	900	48
10	517	26	800	50	932	16
11	526	16	831	19	938	11
12	531	12	837	12	941	8
13	534	9	841	9	943	6
14	536	7	843	7	945	5
15	538	6	845	6	946	5
16	540	5	847	5	948	4
17	542	4	848	4	949	4
18	543	4	850	4	950	4
19	544	4	851	4	951	3
20	545	3	852	3	952	3
21	546	3	853	3	953	3
22	547	3	854	3	954	3
23	548	3	855	3	955	3
24	549	3	856	3	956	3
25	550	3	857	3	956	2
26	551	3	858	3	957	2
27	552	3	858	3	958	2
28	553	3	859	3	958	2
29	553	3	860	2	959	2
30	554	3	861	2	960	2
31	555	3	861	2	960	2
32	556	3	862	2	961	2
33	557	3	863	2	962	2
34	558	3	864	2	962	2
35	558	3	864	2	963	2
36	559	2	865	2	963	2
37	560	2	866	2	964	2
38	561	2	867	2	965	2
39	562	2	868	2	965	2
40	563	2	868	2	966	2
41	563	2	869	2	966	2
42	564	2	870	2	967	2
43	565	2	871	2	968	2
44	566	3	872	2	969	2
45	567	3	873	2	969	2

Raw Score	Grade 5		Grade 8		High School	
	Scale Score	SEM	Scale Score	SEM	Scale Score	SEM
46	568	3	874	3	970	2
47	570	3	875	3	971	2
48	571	3	876	3	973	3
49	573	4	878	3	974	3
50	575	4	880	4	976	4
51	578	5	884	6	979	5
52	584	8	890	9	984	8
53	599	22	899	17	999	21

Classification Consistency and Accuracy

This section contains information relevant to the *Standards and Assessment Peer Review Guidance*, Critical Element 4.2:

For each assessment, including all alternate assessments, has the State considered the issue of **reliability**, as described in the *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 1999), with respect to *all* of the following categories:

(b) Has the State quantified and reported within the technical documentation for its assessments the conditional standard error of measurement and student classification that are consistent at each cut score specified in its academic achievement standards?

Classification Consistency

Classification consistency, or decision consistency, is defined as the extent to which the classifications of examinees agree on the basis of two independent administrations of a test, or administration of two parallel test forms. However, it is practically infeasible to obtain data from repeated administrations of a test because of cost, time, and students' recall of the first administration. Therefore, a common practice is to estimate decision consistency from one administration of a test.

Classification Accuracy

Classification accuracy, or decision accuracy, is defined as the extent to which the actual classifications of test-takers based on observed test scores agree with classifications that would be made on the basis of their true scores (Livingston & Lewis, 1995). It is common practice to estimate decision accuracy using a psychometric model to estimate true scores that correspond to observed scores as the basis for estimating classification accuracy.

In other words, classification *consistency* refers to the agreement between two observed scores, while classification *accuracy* refers to the agreement between the observed score and the estimated true score.

A straightforward classification consistency estimation can be expressed in terms of a contingency table representing the probability of a particular classification outcome under specific scenarios. For example, Table 24 is a contingency table of (H+1) rows × (H+1) columns, where H is the number of cut scores, such that two cut scores yield a 3×3 contingency table.

Table 24. Example of Contingency Table with Two Cut Scores

	Level 1	Level 2	Level 3	Sum
Level 1	P_{11}	P_{21}	P_{31}	$P_{\cdot 1}$
Level 2	P_{12}	P_{22}	P_{32}	$P_{\cdot 2}$
Level 3	P_{13}	P_{23}	P_{33}	$P_{\cdot 3}$
Sum	$P_{1.}$	$P_{2.}$	$P_{3.}$	1.0

Hambleton and Novick (1973) proposed P as a measure of classification consistency, where P is defined as the sum of the diagonal values of the contingency table (shaded above):

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

To account for statistical chance agreement, Swaminathan, Hambleton, & Algina (1974) suggested using Cohen’s kappa (1960):

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

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

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

Kolen and Kim (2005) suggested a method for estimating consistency and accuracy that involves the generation of item responses using item parameters based on the IRT model (see also Kim, Choi, Um, & Kim, 2006, as well as Kim, Barton, & Kim, 2008). Two sets of item responses are generated using a set of item parameters and an examinee’s ability distribution from a single test administration.

In 2010, CTB used the KKCLASS (Kim, 2007) program to calculate these statistics on the 2010 DC CAS results. The KKCLASS procedure is an IRT-based procedure and is more consistent with the scaling and scoring of student results. These two sets of item responses are considered as an examinee’s responses on two administrations of the same form. The procedure is described below and is implemented with KKCLASS software (Kim, 2007):

Step 1: Obtain item parameters (**I**) and ability distribution weight ($\hat{g}(\theta)$) at each quadrature point from a single test.

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

Step 3: Construct a classification matrix at each quadrature point. Determine the joint event for the cells in Table 24 using the raw scores obtained from Step 2.

Step 4: Repeat Steps 2 and 3 *R* times and get average values over *R* replications.

Step 5: Multiply distribution weight ($\hat{g}(\theta)$) by average values in Step 4 for each quadrature point, and sum across all quadrature points. From this final contingency table, classification consistency indices, such as consistency agreement and kappa, can be computed.

Step 6: Because examinees' abilities are estimated at each quadrature point, this quadrature point can be considered the true score. Therefore, classification accuracy is computed using both examinees' estimated abilities (observed scores) and quadrature point (true score).

Tables 25–27 display the classification consistency and accuracy results for the 2010 DC CAS in Reading, Mathematics, and Science/Biology. As can be seen in the tables below, the classification consistency results range from 0.66 to 0.77 in all content areas and grades. The results are comparable to those in 2009, which ranged between 0.72 and 0.77. Kappa coefficients range between 0.49 and 0.68, which is comparable to the 2009 results (0.57 to 0.67). The kappa values, which indicate classification consistency beyond chance consistency, represent moderate to substantial consistency levels (Landis & Koch, 1997). The classification consistency results suggest that the 2010 DC CAS assessments in Reading, Mathematics, and Science/Biology would classify examinees into the same DC CAS proficiency levels across multiple test administrations with reasonably strong consistency.

The classification accuracy results range from 0.73 to 0.84 in all content areas and grades. The results are comparable to those in 2009, which also ranged between 0.78 and 0.84. These results suggest that the 2010 DC CAS assessments in Reading, Mathematics, and Science/Biology classify examinees into DC CAS proficiency levels based on observed test scores with reasonably strong accuracy.

The false positive rates are estimates of the percentages of examinees that are classified into a proficiency level higher than their true proficiency level. The false negative rates are estimates of the percentages of examinees that are classified into a proficiency level lower than their true proficiency level. These are reasonably low false positive and negative rates in absolute terms. It is a policy question as to how much higher or lower false positive rates should be relative to false negative rates.

The magnitude of classification consistency and accuracy measures is influenced by key features of the test design, including the number of items and number of cut scores, score reliability and associated standard errors of measurement, and the locations of the cut scores in relation to the examinee proficiency frequency distributions. The classification consistency and accuracy results observed for 2010 suggest that consistent and accurate performance level classifications are being made for students based on the DC CAS assessments.

Table 25. Classification Consistency and Accuracy Rates for All Cut Scores: Reading

Grade	Classification Consistency		Classification Accuracy		
	Consistency	Kappa	Accuracy	False Positive Errors	False Negative Errors
3	0.7740	0.6767	0.8300	0.0527	0.1172
4	0.7652	0.6570	0.8359	0.0645	0.0996
5	0.7684	0.6561	0.8295	0.0528	0.1177
6	0.7628	0.6410	0.8252	0.0585	0.1163
7	0.7487	0.6342	0.8233	0.0841	0.0926
8	0.7373	0.6226	0.8028	0.0511	0.1460
10	0.7636	0.6510	0.8288	0.0595	0.1117

Table 26. Classification Consistency and Accuracy Rates for All Cut Scores: Mathematics

Grade	Classification Consistency		Classification Accuracy		
	Consistency	Kappa	Accuracy	False Positive Errors	False Negative Errors
3	0.7670	0.6689	0.8316	0.0609	0.1074
4	0.7576	0.6574	0.8207	0.0586	0.1207
5	0.7515	0.6455	0.8196	0.0699	0.1105
6	0.7684	0.6725	0.8278	0.0599	0.1123
7	0.7696	0.6707	0.8369	0.0796	0.0835
8	0.7131	0.5862	0.7934	0.0700	0.1366
10	0.7560	0.6504	0.8180	0.0663	0.1157

Table 27. Classification Consistency and Accuracy Rates for All Cut Scores: Science/Biology

Grade	Classification Consistency		Classification Accuracy		
	Consistency	Kappa	Accuracy	False Positive Errors	False Negative Errors
5	0.7208	0.5924	0.7916	0.0782	0.1302
8	0.6854	0.5409	0.7655	0.0920	0.1425
High School	0.6555	0.4869	0.7323	0.1025	0.1652

Classification consistency and accuracy estimates for the Basic, Proficient, and Advanced cut scores appear in Appendix D. Classification consistency and accuracy estimates for all cut scores for examinee subgroups appear in Appendix E.

Differential Item Functioning

This section contains information relevant to the *Standards and Assessment Peer Review Guidance*, Critical Element 4.3:

Has the State ensured that its assessment system is fair and accessible to all students, including students with disabilities and students with limited English proficiency, with respect to each of the following issues:

(c) Has the State taken steps to ensure fairness in the development of the assessments?

An item flagged for differential item functioning (DIF) is more difficult for a particular group of students than would be expected based on their total test scores, compared to the difficulty of the item for a focal group with equivalent total test scores. For the DC CAS program, CTB uses Mantel-Haenszel statistics (Mantel & Haenszel, 1959) to evaluate DIF for both operational and field test items. The groups compared in the DIF analyses for the 2010 administration were female and male students, and African American, Asian/Pacific Islander, Hispanic, and White students. Comparing these subgroups in DIF analyses is conventional practice in the US. Male and African American students were the reference groups. Selecting males as the reference group is conventional practice in the US. African American students are selected as the reference group for DIF analyses because they are the largest subgroup enrolled in DC schools. DIF is examined to identify operational items that may favor one group over another.

Items flagged for DIF may or may not favor one examinee subgroup over another. As with all statistical tests, Mantel-Haenszel DIF statistics are subject to Type I and II errors. All items are screened in Content and Bias Review meetings comprised of DC educators to ensure that no obvious sensitive and unfair terms, phrases, scenarios, or illustrations that could influence examinee performance on items appear in DC CAS items prior to field testing and selection for operational test forms. OSSE and CTB screen items that are flagged for DIF after each operational administration to identify items that may favor or disadvantage examinee subgroups. Items that are flagged are rarely disqualified from operational scoring, typically because no plausible explanations for the flags are apparent in the item content and response requirements. In these cases, statistical flagging is attributed to statistical error. Statistical DIF analyses are not conducted for the Composition test. Composition prompts are subjected to standard Content and Bias Reviews.

The statistical procedures and flagging criteria used by CTB to identify items that exhibit DIF are those used by the Educational Testing Service (ETS) for the National Assessment of Educational Progress (NAEP). For MC items, the Mantel-Haenszel (χ^2_{MH}) statistic (Mantel & Haenszel, 1959) was used to evaluate potential DIF in items. In this procedure, items with A, B, and C level DIF are flagged.

For multiple-choice items, the Mantel-Haenszel (χ^2_{MH}) statistic flags items for potential DIF using the following criteria:

- B level DIF, where a "B" indicates DIF and has an absolute value of the Mantel-Haenszel (Δ_{MH}) that is significantly greater than zero (at the 0.05 level) and $-1.5 \le \Delta_{MH} \le -1$ or $1 \le \Delta_{MH} \le 1.5$.

- C level DIF, where a "C" indicates DIF and has an absolute value of the Mantel-Haenszel (Δ_{MH}) that is significantly greater than zero (at the 0.05 level) and $|\Delta_{MH}|$ exceeds 1.5.

For constructed-response items, an effect size (ES) statistic based on the Mantel χ^2 is used to flag items for potential DIF. ES is obtained by dividing the standardized mean difference (SMD) statistics by the standard deviation of the item. Items are flagged using the same rules that are used in NAEP:

- BB level, where the Mantel statistic is significant ($p < 0.05$) and |ES| is between 0.17 and 0.25.

- CC level, where the Mantel statistic is significant ($p < 0.05$) and |ES| ≥ 0.25

C and CC level flags indicate moderate to severe DIF. B and BB level flags indicate moderate DIF. A-level flags indicate negligible DIF. (A detailed description of these procedures can be found in Zwick, Donoghue, & Grima, 1993.)

Positive DIF values indicate items that favor the focal group, while negative values indicate items that disadvantage the focal group.

Results of the Differential Item Functioning Analyses

The DIF analyses were conducted for all grades and content areas for race/ethnicity and gender. DIF analyses were conducted with at least 400 cases for reference groups and 200 cases for focal groups to provide data adequate for Mantel-Haenszel DIF analysis procedures, which require subdividing each comparison group based on total test raw scores.

Tables 28–30 summarize the 2010 DIF analysis results for operational items. Modest numbers of multiple choice and constructed response items were flagged for DIF at levels B and C. This is similar to the results in 2009. The majority of items flagged for DIF were in race/ethnicity comparisons; many of those were positive values that indicated DIF that favored the focal group (e.g., Hispanic and White students).

Overall, the number of items flagged for DIF was moderate. For example, the total 106 Reading items flagged for DIF represent 11.6 percent of the 911 flagging opportunities in Reading. (Flagging opportunities in Reading were calculated as [48 items per grade x 5 grades x 3 DIF comparisons – 1] + [48 items x 2 grades x 2 comparisons]; the Asian/Pacific Islander "NA" comparisons are not counted, nor are the one Reading item in footnote 1 and White comparisons in grades 8 and 10.) The percentages of flagged items in Mathematics (116 flags, 1,026 flagging opportunities) and Science/Biology (34 flags, 350 flagging opportunities, taken together), calculated in similar fashion, are 11.3 and 9.7 percent, respectively.

Appendix F lists all flagged items and their respective Mantel-Haenszel DIF output, including the focal subgroups for which each item was flagged.

Table 28. Numbers of Items Flagged for DIF Using the Mantel-Haenszel Procedure: Reading

Reference Group	Focal Group	A	B	B-	C	C-
Grade 3 (total 48 items)						
Male	Female	48	0	0	0	0
African American	Asian/Pacific Islander	N/A	N/A	N/A	N/A	N/A
	Hispanic	44	2	2	0	0
	White [1]	31	6	1	7	2
Grade 4 (total 48 items)						
Male	Female	46	1	1	0	0
African American	Asian/Pacific Islander	N/A	N/A	N/A	N/A	N/A
	Hispanic	47	0	1	0	0
	White	35	7	0	6	0
Grade 5 (total 48 items)						
Male	Female	45	2	1	0	0
African American	Asian/Pacific Islander	N/A	N/A	N/A	N/A	N/A
	Hispanic	46	1	1	0	0
	White	42	0	0	6	0
Grade 6 (total 48 items)						
Male	Female	47	1	0	0	0
African American	Asian/Pacific Islander	N/A	N/A	N/A	N/A	N/A
	Hispanic	43	2	1	0	2
	White	33	2	1	12	0
Grade 7 (total 48 items)						
Male	Female	48	0	0	0	0
African American	Asian/Pacific Islander	N/A	N/A	N/A	N/A	N/A
	Hispanic	45	0	2	1	0
	White	35	2	1	10	0

Grade 8 (total 48 items)						
Male	Female	44	3	1	0	0
African American	Asian/Pacific Islander	N/A	N/A	N/A	N/A	N/A
	Hispanic	42	2	4	0	0
	White	N/A	N/A	N/A	N/A	N/A
Grade 10 (total 48 items)						
Male	Female	43	2	3	0	0
African American	Asian/Pacific Islander	N/A	N/A	N/A	N/A	N/A
	Hispanic	41	2	4	1	0
	White	N/A	N/A	N/A	N/A	N/A

Note. Positive flags indicate DIF that favors the focal group. Statistics with fewer than 200 focal group examinees and 400 reference group examinees are not calculated for these analyses to provide appropriate subgroup comparisons. A=no DIF; B=moderate DIF; C=considerable DIF.

N/A= not applicable because case count requirements for the reference (400) and focal (200) groups were not met. See Table 7 for the numbers of examinees in each grade and subgroup.

[1] Although the minimum case counts for the reference (400) and focal (200) groups were available in grade 3 Reading, no matching pairs of reference and focal group examinees were found for some total test scores for item 16. As a result, DIF statistics were not calculated for this item.

Table 29. Numbers of Items Flagged for DIF Using the Mantel-Haenszel Procedure: Mathematics

Reference Group	Focal Group	A	B	B-	C	C-
Grade 3 (total 54 items)						
Male	Female	52	2	0	0	0
African American	Asian/Pacific Islander	N/A	N/A	N/A	N/A	N/A
	Hispanic	50	2	2	0	0
	White	39	4	4	6	1
Grade 4 (total 54 items)						
Male	Female	54	0	0	0	0
African American	Asian/Pacific Islander	N/A	N/A	N/A	N/A	N/A
	Hispanic	51	2	1	0	0
	White	36	9	2	3	4
Grade 5 (total 54 items)						
Male	Female	52	0	0	2	0
African American	Asian/Pacific Islander	N/A	N/A	N/A	N/A	N/A
	Hispanic	50	4	0	0	0
	White	37	7	4	6	0
Grade 6 (total 54 items)						
Male	Female	54	0	0	0	0
African American	Asian/Pacific Islander	N/A	N/A	N/A	N/A	N/A
	Hispanic	51	2	0	1	0
	White	37	5	2	7	3

Grade 7 (total 54 items)						
Male	Female	52	2	0	0	0
African American	Asian/Pacific Islander	N/A	N/A	N/A	N/A	N/A
	Hispanic	51	2	0	0	1
	White	39	1	2	8	4
Grade 8 (total 54 items)						
Male	Female	52	1	1	0	0
African American	Asian/Pacific Islander	N/A	N/A	N/A	N/A	N/A
	Hispanic	50	2	2	0	0
	White	N/A	N/A	N/A	N/A	N/A
Grade 10 (total 54 items)						
Male	Female	50	1	3	0	0
African American	Asian/Pacific Islander	N/A	N/A	N/A	N/A	N/A
	Hispanic	53	0	1	0	0
	White	N/A	N/A	N/A	N/A	N/A

Note. Positive flags indicate DIF that favors the focal group. Statistics with fewer than 200 focal group examinees and 400 reference group examinees are not calculated for these analyses to provide appropriate subgroup comparisons. A=no DIF; B=moderate DIF; C=considerable DIF.

N/A= not applicable because case count requirements for the reference (400) and focal (200) groups were not met. See Table 8 for the numbers of examinees in each grade and subgroup.

Table 30. Numbers of Items Flagged for DIF Using the Mantel-Haenszel Procedure: Science/Biology

Reference Group	Focal Group	A	B	B-	C	C-
Grade 5 (total 50 items)						
Male	Female	48	1	1	0	0
African American	Asian/Pacific Islander	N/A	N/A	N/A	N/A	N/A
	Hispanic	48	2	0	0	0
	White	29	13	1	5	2
Grade 8 (total 50 items)						
Male	Female	46	2	2	0	0
African American	Asian/Pacific Islander	N/A	N/A	N/A	N/A	N/A
	Hispanic	49	0	1	0	0
	White	N/A	N/A	N/A	N/A	N/A
High School (total 50 items)						
Male	Female	49	1	0	0	0
African American	Asian/Pacific Islander	N/A	N/A	N/A	N/A	N/A
	Hispanic	47	2	1	0	0
	White	N/A	N/A	N/A	N/A	N/A

Note. Positive flags indicate DIF that favors the focal group. Statistics with fewer than 200 focal group examinees and 400 reference group examinees are not calculated for these analyses to provide appropriate subgroup comparisons. A=no DIF; B=moderate DIF; C=considerable DIF.

N/A= not applicable because case count requirements for the reference (400) and focal (200) groups were not met. See Table 9 for the numbers of examinees in each grade and subgroup.

Section 6. Reliability and Validity of Hand-Scoring

This section contains information relevant to the *Standards and Assessment Peer Review Guidance*, Critical Element 4.2:

For each assessment, including all alternate assessments, has the State considered the issue of **reliability**, as described in the *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 1999), with respect to *all* of the following categories:

(c) Has the State reported evidence of generalizability for all relevant sources, such as variability of groups, internal consistency of item responses, variability among schools, consistency from form to form of the test, and inter-rater consistency in scoring?

In this section, we first describe the scoring process used for DC CAS. In particular, we focus on the hand-scoring process. At the end of this section, we describe and report the results of the inter-rater reliability study conducted on the hand-scoring of the constructed-response items. Inter-rater reliability assesses the consistency of how the rating system is implemented.

DC CAS Scoring Process

Multiple-choice items were scored by CTB using electronic scanning equipment. Constructed-response items were scored by human raters who were trained by CTB. Evidence of validity is provided by the procedures for hand-scoring described below.

Selection of Scoring Raters

CTB/McGraw-Hill and Kelly Services Inc. strive to develop a highly qualified, experienced core of raters so that the integrity of all projects is appropriately maintained.

Recruitment

The DC CAS 2010 project was staffed with a large number of returning raters and team leaders who had previous experience with hand-scoring projects. Kelly Services Inc. also recruited new team leaders and raters for employment.

CTB requires that all team leaders and raters possess a bachelor’s degree or higher. Kelly Services Inc. carefully screened all new applicants and required them to produce either a transcript or a copy of the degree. Kelly Services Inc. also required a one- to two-hour interview/screening process. Individuals who did not present proper documentation or had less than desirable work records were eliminated during this process. Kelly Services Inc. verified that 100% of all potential raters met the degree requirement. All experienced raters and team leaders had already successfully completed the screening process.

The Interview Process

All potential raters completed a pre-interview activity. For some parts of the pre-interview activity, applicants were shown examples of test responses and were supplied with a scoring guide. In a brief introduction, they became acquainted with the application of a rubric. After the introduction, applicants applied the scoring guide to score the sample responses.

Each applicant's scores were used for discussion during the interview process to determine the applicant's trainability, as well as his or her ability to understand and implement the standards set forth in the sample scoring guide.

Kelly Services Inc. interviewed each applicant and determined the applicant's suitability for a specific content area and grade level. Applicants with strong leadership skills were questioned further to determine whether they were qualified to be team leaders.

When Kelly Services Inc. felt applicants were qualified, the applicants were recommended for employment. All assignments were made according to availability and suitability. Before being hired, all employees were required to read, agree to, and sign a nondisclosure agreement outlining the CTB/McGraw-Hill business ethics and security procedures.

Training Material Development

Scoring guides for the 2010 constructed response items in Reading, Mathematics, Writing, and Science/Biology were developed by CTB's Content and Development teams in conjunction with DC Public Schools (DCPS). Prior to actual scoring, CTB supervisors studied and internalized these guides along with existing materials that were then used in training raters to hand score the constructed response items for all four content areas. This ensured that the same Anchor papers and training philosophy were used while scoring the items operationally in 2010 as had been used when they were scored as field test items.

Due to budget and schedule constraints in 2006, Rangefinding for Reading and Mathematics was not conducted in DC. Instead, Anchor papers for Reading and Mathematics were chosen by CTB staff and approved by OSSE via email that year. Rangefinding for Writing only was conducted in DC in 2006. Rangefinding for Reading, Mathematics, and Science/Biology began in 2007, the first year that Reading and Mathematics tests were operational and that Science items were included in the DC-CAS in astatewide field test. That Rangefinding process is described in the next section.

Preparation and Meeting Logistics for Rangefinding Prior to 2010 Operational Scoring

Prior to Rangefinding in DC, CTB content supervisors looked at hundreds of student responses to identify a variety of papers for the reviews. These potential Anchors were then assembled for review at Rangefinding. (An Anchor paper is a concrete example of a particular score point, as delineated in the scoring guides, that is used during training and scoring by the CTB raters.)

Rangefinding participants were placed in groups of three or more (plus the CTB content supervisor/facilitator) to discuss a particular grade and content area, and were involved in discussion of all field test items for that grade. Rubrics were passed out and discussed so that all raters became familiar with the item and the criteria that they would use to score the student responses after Rangefinding.

DC participants, along with their CTB facilitator, then reviewed packets containing approximately 35 to 50 responses per item and applied the rubrics and scoring criteria in order to choose appropriate Anchor papers. This process effectively set the range of responses for each score point for each item. At least one Anchor paper for each score point was chosen for every item, and discussion within each group included insights, suggestions, and summary statements for future training on the item. These were recorded by the CTB facilitator. The chosen Anchor papers and their final scores were also recorded by the CTB representative, and raters provided sign-off that consensus on the scoring of the items was achieved.

Training and Qualifying Procedures in 2010

Hand-scoring involves training and qualifying team leaders and raters, monitoring scoring accuracy and production, and ensuring the security of both the test materials and the scoring facilities. An explanation of the training and qualification procedures follows.

All raters were trained and qualified in specific rater item blocks (RIBs), which consisted of a group of items to be scored. Raters and team leaders were trained using the following steps:

- Reviewing the student answer booklet
- Reviewing rubrics
- Reviewing anchor papers
- Explaining scoring strategies, followed by a question-and-answer period
- Scoring a training set, followed by sharing established scores, discussing responses, and answering questions arising from scores
- Scoring and discussing additional training sets
- Administering Qualifying Round 1
- Administering Qualifying Round 2 (if necessary)
- Explaining condition codes and sensitive paper procedures
- Explaining nonstandard response or computer-generated response (nsr/cgr) procedures
- Explaining unscannable image procedures

All raters were trained and qualified using the same procedures and criteria used for the team leaders, who had been trained prior to the training of the raters. The same CTB content experts who supervised the training of the team leaders also supervised the training of the raters.

Breakdown of Scoring Teams

Six CTB content experts oversaw the training and scoring of the constructed–response items for 2010 in Reading, Writing, Mathematics and Science/Biology: two experts for Reading, two for Mathematics, and one each for Writing and Science/Biology. Each of these six content experts was responsible for training and scoring items across several grades of a content area.

Teams of between 8 and 17 raters (depending on the content and grade) trained on and scored all three operational items at their respective grade, and some cross-training was done across grades to ensure on-time completion.

The window for training and scoring the constructed-response items for Reading, Mathematics, Science/Biology, and Writing was from May 10, 2010, through May 21, 2010.

Twenty-one unique constructed–response items were scored (across all grades) for Reading and 21 for Mathematics. Each of the seven grades contained 3 operational items for Reading and 3 for Mathematics (which were common to all four test forms).

For Writing, one operational prompt was administered and scored at each of the three grades (4, 7, and 10). The same rubrics were used to score all three grades of Writing, and each Writing response was scored twice, once for Content and once for Conventions.

Science testing and scoring included the Science tests at grades 5 and 8 and high school Biology and consisted of three operational constructed responses items in each test, for a total of nine constructed response items across the three levels of science.

Reading utilized 66 raters across all grades and 4 team leaders (more experienced raters) over the course of the scoring window. Mathematics utilized 56 raters and 4 team leaders.

Writing utilized 39 raters and 3 team leaders across the three grades, and Science/Biology utilized 24 raters and 2 team leaders across the three levels of science.

Training consisted of a review of the rubrics, followed by analysis of the anchors for each item. Raters then took a training round in preparation for qualifying, which consisted of ten books of sample papers for the items in that RIB. Raters were given two chances to pass and were dismissed if a rating of at least 80% (overall and by item) was not achieved after the second, unique qualification round.

Monitoring the Scoring Process

After training was completed and live scoring began, a number of quality control measures were put in place to ensure that books were scored accurately and that raters did not drift.

Throughout the course of hand-scoring, calibration sets of pre-scored papers (checksets/validity sets) were administered daily to each rater to monitor scoring accuracy and to maintain a consistent focus on the established rubrics and guidelines. Approximately nine percent of books that the raters received were “checkset” papers rather than a live book. Checksets were executed via imaging software that provided images in such a way that the rater did not know when a checkset was being administered.

Raters whose checkset accuracy repeatedly dipped below the quality standards were flagged and retrained. Our Data Monitoring staff also ran inter-rater reliability reports throughout live scoring to look for any raters who were struggling and in need of retraining.

Retraining involves a one-on-one discussion between the supervisor (or a team leader) and the rater, who discuss the problem item(s) as well as the scoring guides and, if necessary, training papers.

In addition to the checkset process, CTB's hand-scoring protocol included the use of read-behinds (spot-checks during live scoring). The read-behind was another valuable rater-reliability monitoring technique that allowed a team leader to review a rater's scored documents, providing feedback and counseling as appropriate.

In 2010, team leaders again conducted read-behinds on raters who had been retrained, as soon as they returned to scoring. If the rater's accuracy on read-behinds and their checkset scores both did not improve after this retraining, they were dismissed from the project immediately.

Approximately 10% of all DC CAS tests were scored by a second rater to establish inter-rater reliability statistics for all constructed-response items. This procedure is called a "double-blind read" because the second rater does not know the first rater's score.

All raters had to sign nondisclosure forms indicating that they were not to disclose the items they were scoring.

Security guards were on site whenever employees were present in the building. All employees were issued photo identification badges and were required to wear them in plain view at all times. Visitors and employees who forgot their badges were issued visitors' badges and were required to wear them in plain view. All employees and visitors were subject to inspection of their personal effects.

Hand-Scoring Agreement

The DC CAS constructed-response questions require a response composed by the examinee, usually in the form of one or more sentences, where the ideas expressed are scored as correct, partially correct, or incorrect. Since it is the ideas rather than the specific written expressions that are scored, the response cannot be scored by applying a clerical key. Raters use judgment to determine whether the ideas expressed match those described in a scoring guide. In other words, raters interpret what the student has written. In order to minimize the difference in interpretations that raters make, raters are required to have certain hiring qualifications and on-site training using examples of responses that match and do not match the desired answers. Even so, the match between a student's response and the scoring guide description of a correct response is a matter of degree. As a result, perfect agreement between different raters of the same student response is not expected in order for the test to be valid. High perfect agreement between raters (70%–80% agreement and above) can be obtained when the ideas being expressed and scored are rather narrowly defined instances of principles or algorithms within a subject area composed of discrete knowledge. This rate of perfect agreement drops rapidly, however, for a subject area such as Reading, where the ideas being expressed are not highly constrained by content; instead, the form and coherence of the expression of the ideas is the target of the testing and scoring.

Nevertheless, relatively high adjacent agreement (scores only one point different) can be obtained. This adjacent agreement still varies with known characteristics of the question and scoring guides. Adjacent agreement of 95% or more is desirable when analytic rubrics are used. When holistic rubrics are used and scoring is deliberately impressionistic, adjacent agreement may drop below 90%.

The inter-rater agreement for DC CAS 2010 operational tests is reported in Tables 31–34 as the percentage value of the difference between the first and second score assigned to a student response on each constructed-response item. Inter-rater reliability assesses the consistency of how the rating system is implemented. The inter-rater reliability analyses show that the DC CAS hand-scoring results have an acceptable perfect-agreement rate and a high adjacent-agreement rate across grades and subject areas.

In Reading, the average perfect agreement was 71%, with a high of 81% and a low of 55%. For perfect and adjacent agreement, the average was 96%, with a high of 100% and a low of 89%. In Mathematics, the average perfect agreement was 88%, with a high of 99% and a low of 76% agreement. For perfect and adjacent agreement, the average was 98%, with a high of 100% and a low of 95%. In Science/Biology, the average perfect agreement rate was 86%, with a high of 98% and a low of 72%. For perfect and adjacent, the average was 98%, with a high of 99% and a low of 97%. In Composition, the average perfect agreement was 68%, with a high of 70% and a low of 64% agreement. For perfect and adjacent agreement, the average was 98%, with a high of 99% and a low of 97%. These rater agreement rates are consistent with industry standards for Reading, Mathematics, and Science short constructed-response items and for essay prompts scored with 4- and 6-point rubrics.

Selection of the 2010 Writing Prompts

The 2010 Writing prompts for Grades 4 and 10 were from the 2007 test administration, and the prompt for Grade 7 was from 2006.

Prior to their initial scoring as field test items, the prompts underwent extensive range-finding in DC with discussion groups of 4–6 teachers per grade, who chose the anchor papers to be used during subsequent training and scoring and who also helped define the parameters in the Writing rubrics.

Thus, the same anchor papers, training materials, and scoring criteria were utilized in 2010 during the operational test administration as had been used in 2006 and 2007 when these prompts were originally scored as field test items.

Table 31. DC CAS 2010 Operational Inter-Rater Agreement for Constructed-Response Items: Reading

Grade	Form	Item No.	Score Points	% of Agreement			Checkset Average Agreement Percentages
				Perfect	Adjacent	Perfect + Adjacent	
3	1-4	9	3	71	22	93	81
		20	3	71	25	96	81
		45	3	73	23	96	82
4	1-4	9	3	55	39	94	69
		19	3	71	27	98	86
		45	3	62	36	98	74
5	1-4	7	3	81	19	100	90
		17	3	72	26	98	78
		44	3	78	20	98	80
6	1-4	6	3	76	22	98	87
		17	3	79	19	98	88
		44	3	81	10	91	77
7	1-4	8	3	67	29	96	88
		18	3	79	16	95	90
		46	3	61	34	95	79
8	1-4	9	3	67	29	96	71
		19	3	76	21	97	87
		52	3	60	29	89	82
10	1-4	7	3	76	19	95	88
		19	3	67	31	98	81
		47	3	66	28	94	78

Note**.** Perfect + Adjacent agreement percentages may not equal the sum of Perfect and Adjacent percentages due to rounding. Checkset average agreement percentages are calculated across all checksets and raters.

Table 32. DC CAS 2010 Operational Inter-Rater Agreement for Constructed-Response Items: Mathematics

Grade	Form	Item No.	Score Points	% of Agreement			Checkset Average Agreement Percentages
				Perfect	Adjacent	Perfect + Adjacent	
3	1-4	10	3	85	14	99	95
		27	3	94	5	99	98
		56	3	80	19	99	88
4	1-4	10	3	95	4	99	98
		23	3	84	14	98	92
		59	3	76	20	96	71
5	1-4	6	3	87	12	99	96
		25	3	87	11	98	92
		60	3	84	14	98	90
6	1-4	6	3	88	10	98	90
		25	3	82	17	99	87
		60	3	87	8	95	90
7	1-4	6	3	86	12	98	89
		25	3	83	16	99	96
		60	3	92	7	99	90
8	1-4	6	3	92	8	100	98
		25	3	86	9	95	96
		60	3	83	15	98	95
10	1-3	6	3	96	3	99	97
		25	3	98	2	100	99
		60	3	87	12	99	96
	4	6	3	99	1	100	100
		25	3	96	4	100	98
		60	3	88	11	99	100

Note. Perfect + Adjacent agreement percentages may not equal the sum of Perfect and Adjacent percentages due to rounding. Checkset average agreement percentages are calculated across all checksets and raters.

Table 33. DC CAS 2010 Operational Inter-Rater Agreement for Constructed-Response Items: Science/Biology

Grade	Form	Item No.	Score Points	% of Agreement			Checkset Average Agreement Percentages
				Perfect	Adjacent	Perfect + Adjacent	
5	1-4	13	2	87	13	100	97
		27	2	72	27	99	86
		51	2	83	15	98	91
8	1-4	13	2	84	13	97	87
		26	2	86	10	96	91
		51	2	94	4	98	98
High School	1-4	13	2	81	17	98	92
		27	2	86	13	99	89
		51	2	98	1	99	98

Note. Perfect + Adjacent agreement percentages may not equal the sum of Perfect and Adjacent percentages due to rounding. Checkset average agreement percentages are calculated across all checksets and raters.

Table 34. DC CAS 2010 Operational Inter-Rater Agreement for Constructed-Response Items: Composition

Grade	Item No.	Score Points	% of Agreement			Checkset Average Agreement Percentages
			Perfect	Adjacent	Perfect + Adjacent	
4	1A	6	69	29	98	82
	1B	4	68	30	98	81
7	1A	6	68	30	98	78
	1B	4	70	29	99	80
10	1A	6	64	33	97	73
	1B	4	66	32	98	67

Note. Perfect + Adjacent agreement percentages may not equal the sum of Perfect and Adjacent percentages due to rounding. Checkset average agreement percentages are calculated across all checksets and raters.

Section 7. IRT Analyses

This section contains information relevant to the *Standards and Assessment Peer Review Guidance*, Critical Element 4.4:

When different test forms or formats are used, the State must ensure that the meaning and interpretation of results are consistent.

(a) Has the State taken steps to ensure consistency of test forms over time?

The 2010 DC CAS assessments in Reading, Mathematics, and Science/Biology underwent both classical test theory and item response theory analyses. In other sections, we describe results from classical analyses (e.g., score reliability, Mantel-Haenszel DIF) for the tests in these four content areas and for the Composition test. In this section, we describe procedures and results from IRT analyses for Reading, Mathematics, and Science/Biology.

Calibration and Equating Models

Scaling and linking was accomplished using the PARDUX and FLUX computer programs to implement the three-parameter logistic model (3PL) and the two-parameter partial-credit (2PPC) IRT models for item calibration and scaling and Stocking and Lord (1983) procedure for equating. These software programs were developed at CTB/McGraw-Hill to enable scaling and linking of complex assessment data.

In PARDUX, a marginal maximum likelihood procedure was used to simultaneously estimate the item parameters under the three-parameter logistic model (3PL, used for multiple-choice items) and the two-parameter partial-credit model (2PPC, used for constructed-response items) (Bock & Aitkin, 1981; Thissen, 1982). These models were implemented using the microcomputer program PARDUX (Burket, 1995). For setting the 2006 base scales for Reading and Mathematics, all scales were also calibrated in PARSCALE (Muraki & Bock, 1991) as verification of the PARDUX results.

Under the 3PL model, the probability that a student with trait or scale score θ responds correctly to multiple-choice item *j* is as follows:

$$P_j(\theta) = c_j + (1 - c_j)/[1 + \exp(-1.7a_j(\theta - b_j))]. \qquad (1)$$

In equation (1), a_j is the item discrimination, b_j is the item difficulty, and c_j is the probability of a correct response by a very low-scoring student. The 2PPC model holds that the probability that a student with trait or scale score θ will respond in category *k* to partial-credit item *j* is given by

$$P_{jk}(\theta) = \exp(z_{jk}) / \sum_{i=1}^{m_j} \exp(z_{ji}), \qquad (2)$$

where $z_{jk} = (k-1) f_j - \sum_{i=0}^{k-1} g_{ji}$, and $g_{j0} = 0$ for all *j*.

The summary output of the above equations is in two different metrics, corresponding to the two item response models (3PL and 2PPC). The location and discrimination parameters for the multiple-choice items are in the traditional 3PL metric (labeled *b* and *a*, respectively). In the 2PPC model, *f* (alpha) and *g* (gamma) are analogous to *b* and a, where alpha is the discrimination parameter and gamma over alpha (*g*/*f*) is the location where adjacent trace lines cross on the ability scale. Because of the different metrics used, the 3PL (multiple-choice) parameters *b* and *a* are not directly comparable to the 2PPC 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 (Burket, 1995). Application of this procedure locates both the MC and CR items on the same scale. Note that for the 2PPC model there are m_j - 1 (where m_j is a score level *j*), independent *g*'s, and one *f*, for a total of m_j independent parameters estimated for each item, while there is one *a* and one *b* per item in the 3PL model.

Goodness of Fit to the IRT Models

Goodness-of-fit statistics were computed for each item to examine how closely the item's data conform to the item response models. A procedure described by Yen (1981) was used to measure fit. In this procedure, students are rank ordered on the basis of their $\hat{\theta}$ values and sorted into ten cells with ten percent of the sample in each cell. Each item *j* in each decile *I* has a response from N_{ij} examinees. The fitted IRT models are used to calculate an expected proportion E_{ijk} of examinees who respond to item *j* in category *k*. The observed proportion O_{ijk} is also tabulated for each decile, and the approximate chi-square statistic

$$Q_{1j} = \sum_{i=1}^{10} \sum_{k=1}^{m_j} \frac{N_{ij}(O_{ijk} - E_{ijk})^2}{E_{ijk}},$$

Q_{1j} should be approximately chi-square distributed with degrees of freedom (*DF*) equal to the number of "independent" cells, 10(m_j-1), minus the number of estimated parameters. For the 3PL model, m_j = 2, so $DF = 10(2 - 1) - 3 = 7$. For the 2PPC model, $DF = 10(m_j - 1) - m_j = 9m_j - 1$. Since *DF* differs between MC and CR items and among CR items with different score levels m_j, Q_{1j} is transformed, yielding the test statistic

$$Z_j = \frac{Q_{1j} - DF}{\sqrt{2DF}}.$$

This statistic is useful for flagging items that fit relatively poorly. Z_j is sensitive to sample size, and cut-off values for flagging an item based on Z_j have been developed and were used to identify items for the item review. The cut-off value is (N/1500 x 4) for a given test, where N is the sample size.

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:

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

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 are computed for ten intervals corresponding to deciles of the theta distribution (Burket, 1995). The Z-statistic is used to characterize item fit. The critical value of Z is different for each grade because it is dependent on sample size.

Evidence of the validity of the scalings is provided by model fit. If the IRT model fits the empirical item response distributions for the population we want to generalize to (i.e., District of Columbia students), then the claim that the scores are valid indicators of an underlying proficiency is strengthened. Fit statistics that indicate the degree of difference between (a) expected probabilities of correct responses at each proficiency level, and (b) observed probabilities examined when items are field tested and when they are used operationally. Table 35 indicates that only small numbers of items were flagged for poor fit to the IRT model. No items were removed from operational scaling and scoring due to poor fit.

Table 35. DC CAS 2010 Numbers of Operational Items Flagged for Poor Fit During Calibration

Content	Grade	Flagged for Poor Fit
Reading	3	2
	4	0
	5	0
	6	0
	7	1
	8	0
	10	0
Mathematics	3	1
	4	0
	5	1
	6	0
	7	3
	8	1
	10	2
Science/Biology	5	0
	8	2
	High School	1

Item Calibration

The 2010 items were calibrated using approximately 99 percent of all Reading and Mathematics student data and 98 percent of all Science/Biology student data. (One to two percent of student records were removed from the calibration sample for students with multiple records, fewer than five valid attempts, and according to other calibration data cleaning rules.) The number of students within each calibration dataset is presented in Table 36. All DC CAS grades and content areas converged successfully during calibration. For quality assurance, all analyses were run twice, independently, by different analysts and reviewed by senior staff.

Table 36. Numbers of Students in 2010 Calibration Datasets

Content	Grade	Number of Students
Reading	3	4,928
	4	4,829
	5	4,510
	6	4,520
	7	4,382
	8	4,526
	10	4,394
Mathematics	3	4,944
	4	4,865
	5	4,533
	6	4,548
	7	4,390
	8	4,527
	10	4,359
Science/Biology	5	4,458
	8	4,393
	High School	4,097

Establishing Upper and Lower Bounds for the Grade Level Scales for the Base Years: 2006 for Reading and Mathematics, 2008 for Science/Biology

Upper and lower bound scale scores are called the lowest obtainable scale score (LOSS) and highest obtainable scale score (HOSS). A maximum likelihood procedure cannot produce scale score estimates for students with perfect scores or scores below the level expected from guessing. Also, while maximum likelihood estimates are available for students with extreme scores other than zero or perfect scores, occasionally these estimates have standard errors of measurement that are very large, and differences between these extreme values have very little meaning. Therefore, scores are established for these students based on a rational but necessarily non-maximum likelihood procedure.

For the DC CAS, LOSSes and HOSSes were set to be equal at the same grade for each content area. For example, the the Grade 3 LOSS and HOSS are 300 and 399 (respectively) and the Grade 5 LOSS is 500 and HOSS is 599 for Reading, Mathematics, and Science. These values were established on the 2006 base scale for Reading and Mathematics and the 2008 base scale for Science/Biology. These values remain constant from year to year. The LOSSes and HOSSes for all grades are provided in Tables 37.

Table 37. LOSS and HOSS for Reading, Mathematics, and Science/Biology Grades 3–8 and 10

Grade	LOSS	HOSS
3	300	399
4	400	499
5	500	599
6	600	699
7	700	799
8	800	899
10/High School	900	999

Note. The LOSS and HOSS apply for Science only in grades 5 and 8 and for Biology at all high school grades. Students may take a Biology course and the required DC CAS Biology test in any grade, 8-12.

Year to Year Equating Procedures

As previously discussed, item response theory models were used to calibrate DC CAS Reading, Mathematics, and Science/Biology items and create new test scale score scales in 2006 and 2008. These scale score scales enable comparability of scores from one year to the next and across all test forms in the same content area and grade. In 2007 through 2010, anchor item sets that link the current test forms to a previous year's scale were used in a Stocking and Lord (1983) equating to maintain the equivalence of DC CAS test forms and interpretation of scale score scales.

Through a common item equating design, the scaled item parameters for each grade level/content area test were placed onto grade and test year specific scales. Using the data from the calibration sample, Stocking and Lord (1983) equating produced parameters expressed on the scales for each content area that are constant across all test years.

Ordinarily, the Reading, and Science/Biology equating anchor item sets include multiple-choice items and one constructed-response item; and in Mathematics, all of the anchor items are multiple-choice items. Anchor items are rotated in and out of use each year, to the degree possible given the limitations of the small DC CAS item pools, to minimize exposure. Anchor items are placed in approximately the same location or same third of the original administration location each year. Anchor item *a* and *b* parameters are calibrated freely (i.e., not fixed during calibration) and used in equating procedures defined by Stocking and Lord (1983). New operational items that were field tested in the previous year's administration are calibrated in conjunction with the anchor items, and the anchor items are used to equate the current year's operational test form to the previous year's operational form and the DC CAS scale score scales.

Procedures for selecting operational anchor items and equating the 2010 test forms did not follow standard DC CAS procedures. No 2009 operational and field test items were available for use in 2010 test forms (and will not be available for use in the future). The 2010 Reading and Mathematics operational test forms consisted of operational items from 2006, 2007, and 2008 operational test forms. The 2010 Science/Biology operational forms consisted of operational items from 2008 and small numbers of items from the 2007 statewide field test. Most operational items in the 2010 Reading and Mathematics tests were used as anchor items to equate the 2010 operational test forms to the 2008 operational forms and DC CAS scale score scale. The numbers of anchor items for the Science/Biology tests are consistent with standard DC CAS equating procedures. The numbers of equating anchor items are documented in Table 2.

Beginning in 2011 and beyond, the equating design and test forms equating procedures will follow those used during 2007-2009. Using 2011 as the example, anchor items will be selected from the 2010 operational items. (Required numbers of anchor items for Reading, Mathematics, and Science/Biology are specified in Table 2.) And, as described earlier in this section, the Stocking and Lord equating procedure will be applied to transform 2011 calibrated item parameters and equate the 2011 test forms to the DC CAS test scales.

The Stocking and Lord (1983) procedure, also called test characteristic curve (TCC) method, was used to place each grade on the vertical scale that had been developed for each content area. It minimizes the mean squared difference between the two characteristic curves, one based on estimates from the previous calibration and the other on transformed estimates from the current calibration. Let $\hat{\psi}_j$ be the test characteristic curve based on estimates from the previous calibration and $\hat{\psi}_j^*$ be the test characteristic curve based on transformed estimates from the current calibration.

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

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

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

following quadratic loss function (F):

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

Anchor Set Review Process

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

1. *P* values of the anchor items are compared. The anchor items should be similar in difficulty in both new and reference administrations. The estimated new form and the reference form *p* values should be aligned on the regression line.

2. IRT item parameters are compared. The correlation coefficients between the reference and equated item parameters should be very high (0.90-1.00).

3. The test characteristic curves for the anchor items are compared before and after the equating transformation is applied. The reference and equated anchor item TCCs should be closely overlapping.

4. The linear transformation parameters (i.e., scaling transformation constants) should be fairly stable across administrations.

Additional analyses of the equating include:

5. The *p* values of the common anchor items between the two administrations show the same direction and magnitude of change as do the scale scores.

6. The full distribution of scale scores is reasonably comparable across administrations and reflects any differences in ability that are indicated by the anchor items.

7. Changes in the percentages of examinees in each proficiency level are reasonable across administrations.

These routine CTB Research team quality check steps were followed during equating for all grades and content areas. No anchor items were flagged for performing differentially from the Stocking and Lord (1983) procedures outlined above.

Anchor Item Parameter Comparisons

Differential anchor item functioning between the two administrations was evaluated by comparing the correlations between the reference and new form item difficulty (*b* parameter), discrimination (*a* parameter), and proportion correct (*p* value) values after equating. The guessing (*c*) parameters typically fluctuate considerably, are held to fixed values in equating, and were not considered in this evaluation.

The anchor item *p* values in Table 38 are highly correlated, ranging from 0.96 to 0.99 for all grades and content areas (see Table 38), as expected. This indicates that the anchor items performed similarly in the examinee populations in 2008 and 2010.

The correlations in Table 38 for the discrimination (*a*) and difficulty (*b*) parameters are moderate to high, ranging from 0.52 to 0.94 for *a* parameters (0.80-0.97 in 2009) and 0.95 to 0.99 for *b* parameters (0.92-0.99 in 2009). These correlations indicate that the items performed similarly in the two administrations and provide evidence that the equating results are reasonable and accurate.

Table 38. Correlations Between the Item Parameters for the Reference Form and 2010 DC CAS Operational Test Form

Content	Grade	Discrimination (a)	Difficulty (b)	*P* Value
Reading	3	0.92	0.97	0.97
	4	0.83	0.97	0.97
	5	0.85	0.98	0.98
	6	0.90	0.98	0.96
	7	0.94	0.99	0.98
	8	0.88	0.98	0.99
	10	0.92	0.98	0.98
Mathematics	3	0.82	0.98	0.98
	4	0.89	0.96	0.96
	5	0.85	0.98	0.98
	6	0.88	0.95	0.97
	7	0.77	0.97	0.96
	8	0.83	0.98	0.97
	10	0.88	0.99	0.98
Science/Biology	5	0.92	0.96	0.97
	8	0.52	0.95	0.96
	High School	0.80	0.95	0.96

Scaling Constants

The scaling constants, or linear transformation parameters, were examined to determine whether performance differences on anchor items are similar across years. There are two constants, a multiplicative constant (M1) and an additive constant (M2). Because PARDUX calibrations center the IRT scale close to the average proficiency of the test-takers, the magnitude of the 2009–2010 differences in these scaling constants indicate the degree of differences in average proficiency between the reference test form and new test form administrations. The scaling constants for the DC CAS grades and content areas are displayed in Table 39 for the 2007–2010 administrations. Table 39 indicates that the additive scaling constants are reasonably similar across all administrations. The multiplicative scaling constants for Reading are higher than in past years, indicating that the slopes of the 2010 test characteristic curves had to be adjusted more than usual to align 2010 to 2006. This larger than usual multiplicative adjustment to the 2010 test characteristic curves for Reading indicates that the 2010 variance of student scores on the anchor set and the resulting untransformed anchor item parameters differed from previous years more than usual. Although these differences are unusual for DC CAS, they are not outside the range of differences found in other state assessment programs. To aid comparison, the differences between scaling constants since 2007-2008 are provided in Table 40.

Table 39. Scaling Constants Across Administrations, All Grades and Content Areas

Content	Grade	2007		2008		2009		2010	
		Multiplicative	Additive	Multiplicative	Additive	Multiplicative	Additive	Multiplicative	Additive
Reading	3	10.4	352.6	10.7	354.0	10.7	353.1	14.3	349.6
	4	11.8	451.2	11.7	453.3	12.4	453.4	13.4	451.6
	5	11.4	552.2	11.3	554.9	11.4	553.7	12.4	553.2
	6	10.8	652.1	10.4	652.9	10.4	653.0	11.4	651.6
	7	10.4	751.3	10.4	752.7	10.2	754.7	11.5	754.3
	8	11.1	851.8	10.4	853.8	11.1	853.5	12.3	854.6
	10	11.3	954.5	10.9	953.4	10.7	954.1	12.1	952.1
Mathematics	3	14.5	353.9	16.2	354.0	17.3	357.0	16.7	352.4
	4	14.1	452.1	13.2	456.4	14.1	457.9	13.8	455.1
	5	14.1	552.2	14.8	555.9	15.1	556.4	14.2	556.8
	6	13.4	647.3	14.5	649.8	14.3	650.1	14.3	650.3
	7	13.7	746.9	13.4	750.0	14.6	751.0	15.1	751.2
	8	13.0	844.8	12.5	847.4	12.9	848.5	13.5	848.6
	10	15.5	945.2	16.9	945.4	16.7	947.3	16.5	944.3
Science	5	N/A		8.0	550.0	8.7	549.9	9.1	549.2
	8			8.0	850.0	8.9	851.0	9.4	851.9
Biology	High School	N/A		8.0	950.0	7.7	946.6	7.5	949.5

Table 40. Differences Between Scaling Constants Across Administrations, All Grades and Content Areas

Content	Grade	Difference 2008-2007		Difference 2009-2008		Difference 2010-2009	
		Multiplicative	Additive	Multiplicative	Additive	Multiplicative	Additive
Reading	3	0.3	1.4	0.0	-0.9	3.6	-3.5
	4	-0.1	2.1	0.7	0.1	1.0	-1.8
	5	-0.1	2.7	0.1	-1.2	1.0	-0.5
	6	-0.4	0.8	0.0	0.1	1.0	-1.4
	7	0.0	1.4	-0.2	2.0	1.3	-0.4
	8	-0.7	2.0	0.7	-0.3	1.2	1.1
	10	-0.4	-1.1	-0.2	0.7	1.4	-2.0
Mathematics	3	1.7	0.1	1.1	3.0	-0.6	-4.6
	4	-0.9	4.3	0.9	1.5	-0.3	-2.8
	5	0.7	3.7	0.3	0.5	-0.9	0.4
	6	1.1	2.5	-0.2	0.3	0.0	0.2
	7	-0.3	3.1	1.2	1.0	0.5	0.2
	8	-0.5	2.6	0.4	1.1	0.6	0.1
	10	1.4	0.2	-0.2	1.9	-0.2	-3.0
Science	5	N/A		0.7	-0.1	0.4	-0.7
	8			0.9	1.0	0.5	0.9
Biology	High School	N/A		-0.3	-3.4	-0.2	2.9

Once the tests are equated, final parameter tables are developed into scoring tables, from which each student's scale score is derived. Examinee scale scores are estimated for DC CAS using number correct scoring.

Section 8. Standard Setting

This section contains information relevant to the *Standards and Assessment Peer Review Guidance*, Critical Elements 2.1, 2.2, and 2.3:

2.1
Has the State formally approved/adopted challenging academic achievement standards in Reading/Language Arts and Mathematics for each of Grades 3 through 8 and for the 10-12 grade range? These standards were to be completed by school year 2005-2006.

2.2
Has the State formally approved/adopted academic achievement descriptors in Science for each of the grade spans 3-5, 6-9, and 10-12 as required by school year 2005-06?

2.3
1. Do these academic achievement standards (including modified and alternate academic achievement standards, if applicable) include for each content area--

(a) At least three levels of achievement, including two levels of high achievement (proficient and advanced) that determine how well students are mastering a State's academic content standards and a third level of achievement (basic) to provide information about the progress of lower-achieving students toward mastering the proficient and advanced levels of achievement; *and*

(b) Descriptions of the competencies associated with each achievement level; *and*

(c) Assessment scores ("cut scores") that differentiate among the achievement levels and a rationale and procedure used to determine each achievement level?

Prior to setting performance standards for the DC CAS Reading, Mathematics, Science/Biology, and Composition tests, CTB test development staff drafted performance level descriptions for each grade and content area. Performance level descriptors for Reading and Mathematics were drafted in 2006, and for Science/Biology and Composition in 2008. DCPS staff reviewed, refined, and approved the descriptions prior to each workshop.

A modification of the Bookmark Standard Setting Procedure (BSSP; Lewis, Mitzel, & Green, 1996) was implemented to set standards for the Reading and Mathematics assessments in July 2006 and for the Science/Biology assessments in July 2008. The Reasoned Judgment method (Perie, 2007; Roeber, 2002) was used to set standards for the Composition assessments in August 2008. DCPS staff who participated in standard setting workshops recommended cut scores for each test and grade level.

The standard setting workshops for Reading, Mathematics, and Science/Biology lasted four-and-a-half days, with the morning of the first day devoted to orientation and bookmark training, two and a half days to standard setting, and one and a half days to description writing. Participants recommended three cut scores at the Basic, Proficient, and Advanced levels, which would separate students into four performance levels: Below Basic, Basic, Proficient, and Advanced. Participants engaged in training, discussion, and three rounds of bookmark placements. The table leaders reviewed the participant-recommended cut scores and associated impact data and suggested changes to promote cross-grade articulation. Impact data are the percentages of students who are classified in each performance level based on the recommended cut scores.

The Reasoned Judgment (RJ) method was implemented to set standards for the Composition test in Grades 4, 7, and 10. The Reasoned Judgment procedure is a rubric-centered, content-based method that has been used in recent years to establish performance standards on unscaled assessments, such as many alternate assessments (Perie, 2007; Roeber, 2002). During the three-day procedure, D.C. educators were trained to examine the DC CAS scoring rubrics and to consider the knowledge and skills associated with the attainment of each successive score level. Two separate rubrics were used to score the Composition tests: students received 0–6 points for Topic/Idea Development, and 0–4 points for Standard English Conventions. (Total Composition scores range from 2 to 10.) Participants studied these scoring rubrics, the DC CAS content standards, and performance level descriptions and discussed their expectations of the knowledge and skills students must have in order to associate a score level with a performance level.

The cut score recommendations from the committees for all content areas and grades were reviewed by the DC CAS Technical Advisory Committee and DCPS (in 2006) and the OSSE in 2008. Small numbers of cut scores were adjusted both times to achieve articulated standards and impact data. The DC Board of Education approved these cut scores.

Tables 41–44 show the final, approved cut scores. Complete Standard Setting Technical Reports summarize procedures and results of the DC CAS standard settings for all content area assessments. The reports include a round-by-round synopsis, agendas, all training materials, recommended cut scores, and reference papers. (See *Bookmark Standard Setting Technical Report 2008 for Grades 5 and 8 Science and High School Biology* and *Reasoned Judgment Standard Setting Technical Report 2008 for Grades 4, 7, and 10 Composition*.)

Table 41. Final Reading Cut Score Ranges

Grade	Below Basic	Basic	Proficient	Advanced
3	300 – 338	339 – 353	354 – 372	373 – 399
4	400 – 438	439 – 454	455 – 471	472 – 499
5	500 – 539	540 – 555	556 – 572	573 – 599
6	600 – 639	640 – 654	655 – 671	672 – 699
7	700 – 738	739 – 755	756 – 767	768 – 799
8	800 – 839	840 – 855	856 – 869	870 – 899
10	900 – 939	940 – 955	956 – 969	970 – 999

Table 42. Final Mathematics Cut Score Ranges

Grade	Below Basic	Basic	Proficient	Advanced
3	300 – 339	340 – 359	360 – 375	376 – 399
4	400 – 442	443 – 457	458 – 473	474 – 499
5	500 – 542	543 – 559	560 – 574	575 – 599
6	600 – 635	636 – 653	654 – 667	668 – 699
7	700 – 735	736 – 751	752 – 769	770 – 799
8	800 – 835	836 – 849	850 – 867	868 – 899
10	900 – 932	933 – 950	951 – 970	971 – 999

Table 43. Final Science/Biology Cut Score Ranges

Grade	Below Basic	Basic	Proficient	Advanced
5	500 – 540	541 – 552	553 – 563	564 – 599
8	800 – 848	849 – 855	856 – 867	868 – 899
High School	900 – 945	946 – 951	952 – 965	966 – 999

Table 44. Final Composition Cut Score Ranges

Grade	Below Basic	Basic	Proficient	Advanced
4	0 – 3	4 – 6	7 – 8	9 – 10
7	0 – 3	4 – 6	7 – 8	9 – 10
10	0 – 3	4 – 6	7 – 8	9 – 10

Section 9. Percent Indices for the State and for Content Areas and Content Strands

State Percent Index for Content Areas

The DC CAS assessments provide a State Percent Index for Reading, Mathematics, and Science/Biology. The Percent Index is determined by using all of the test information to provide additional indication about examinee performance within content areas. For each content area scale score, the corresponding IRT-based expected percent of maximum (EPM) score is identified through the test characteristic curve. The state Performance Indexes are these EPM scores. State Performance Indexes range from 0 to 100 and are interpreted similarly to, but not the same as, a percent correct score.

Percent Index Score for Content Strands

Teachers and educational decision-makers frequently want diagnostic information that can be used to inform instructional strategies within a content area and to help identify student strengths and weaknesses. This information can be derived from student scores on subsets of test questions called content strands (e.g., Informational Text, Number Sense). Results from the DC CAS can be used to calculate a Percent Index for each content strand in Reading, Mathematics, Science, and Biology. The Percent Index represents the score a student or class would have achieved had they taken every item in the DC CAS item pool. This estimate can be calculated based on every item in the pool for each content strand.

Percent Index results for every student and class can be found on score reports provided to schools. The results are scaled so that the numbers range from 0 to 100. They can be interpreted similarly to, but not the same as, a percent correct score. Student performance in a content strand can also be identified as at or above a proficient Percent Index cut score. (The proficient cut scores for each subject and grade level can be found in Table 45.)

Strand Percent Indexes should be interpreted with caution. In designing tests, some compromises must be made regarding the specificity of strands, test length, student guessing behavior, and breadth of content coverage. When used with due caution, DC CAS information on the performance at or above Proficient in the content strands can be useful in augmenting information from other sources, such as teacher observations and classroom assessments.

Calculating Proficient Percent Index using Expected Percent of Maximum Score

The proficient Percent Index cut scores are determined by using overall test characteristics and converting the information to an expected percent of maximum score (EPM). The resulting EPM score is the cut score for mastery for each strand in that grade and content area combination.

More specifically, the Strand PI is an estimate of the true score for the strand (i.e., the estimated proportion of maximum points possible within a strand) based on the performance of a student in the total content area. Because most strands are measured by a relatively small number of items, a Bayesian procedure that takes into account the overall test performance is used to improve the reliability of the strand scores. Given a student's scale score in the content area, the 3PL IRT model for multiple-choice items and the 2PPC model for constructed-response items are used to compute the estimated proportion of the maximum points obtained for that strand.

The estimated proportion of the maximum points obtained for the strand provides the initial (Bayesian prior) estimate of the student's score. If this initial estimate is consistent with the student's observed proportion, as indicated by a chi-square test, the two scores are combined as a weighted average to obtain the Strand PI score (i.e., the estimated true score). The appropriate weight for the Bayesian prior estimate is computed as a function of the standard error of the scale score on which it is based; the smaller the standard error, the larger the weight. If the prior estimate and the observed proportion differ significantly, the observed proportion of the maximum score is used without the prior estimate to compute the student's PI score on that strand.

Performance At or Above Proficient

Student performance in a content strand also can be characterized as at or above the Proficient cut score. The Performance Index cut scores are the Proficient cut score on the total test scale determined via standard setting. This cut score, in scale score units, was transformed to the State Performance Index value using the overall test characteristic curve and converting the information to the expected percent of maximum score. The resulting EPM score became the cut score for mastery for each strand in that grade and content area combination.

A student's PI score and performance level are affected by the difficulty of the items in a given test form and level; the more difficult the items, the lower the PI will tend to be, and this will be reflected in the strand performance level.

Strand performance designations should be interpreted with caution. In designing tests that are both usable and useful, some compromises must be made regarding the specificity of strands, test length, and breadth of content coverage. Some strands are clearly broader than others, and it cannot be assumed that the items measuring various strands are equally representative samples of their respective skill domains. Moreover, the scale score performance cut scores used to separate PI ranges are based on the Proficient level cut scores for the total test, not for the individual content strands. Other reasons for caution in interpreting this information include students' guessing behavior, the limited generalizability of strand scores, which are based on relatively few items and score points, and variations in the difficulty of strands. When used with due caution, DC CAS information on the performance at or above Proficient in the content strands is useful in augmenting information from other sources, such as teacher observations and classroom assessments.

Cut Scores for Performance At or Above Proficient for Percent Index Scores

Each year, the raw score cut score that corresponds to the expected percent of maximum that relates to the Proficient scale score cut score for the content area can change. The Proficient cut scores for the content strand PIs are provided in Table 45.

Table 45. Content Strand Percent Index Cut Scores

Subtest	Grade	Proficient Scale Score Cut Score	Total Test Number Correct Score	EPM Score
Reading	3	354	40	74
	4	455	35	65
	5	556	39	72
	6	655	36	67
	7	756	36	67
	8	856	33	61
	10	956	39	72
Mathematics	3	360	46	77
	4	458	39	65
	5	560	41	68
	6	654	36	60
	7	752	32	53
	8	850	27	45
	10	951	31	52
Science	5	553	28	53
	8	856	24	45
Biology	High School	952	20	38

Section 10. Results

Test and Item Characteristics

Table 46 summarizes the DC CAS Reading, Mathematics, and Science/Biology results for the total population of students at each grade. The table displays mean scale scores, scale score standard deviations, raw score means, and raw score standard deviations and mean *p* values and item-total correlations. For multiple-choice items, percent correct (*p* values) is reported. For constructed-response items, the *p* value is calculated as the mean score across all students divided by the maximum number of score points possible. On average, the items are of moderate difficulty or easy.

Table 46 also displays the mean item omit rates calculated across students for each grade and content area. The largest mean percentage omit rate is 2.22 in Mathematics grade 10. Overall, these omit rates are low. CTB flags items when more than 5 percent of students omit an item. Flagged items are reviewed to ensure that they are appropriate for examinees in the tested grade. In addition, omitted items near the end of the test are reviewed as not reached items. All of the not reached rates are less than 1 percent, except for in Mathematics grade 10 (1.17 percent), indicating that the DC CAS tests, while somewhat difficult for students, are not speeded.

Tables in Appendix G display the item difficulty for each item at each grade.

Table 46. DC CAS 2010 Operational Test Scale Score and Raw Score Descriptive Statistics

Content	Grade	Mean Scale Score (SD)	Mean Raw Score (SD)	Mean *p* value	Mean Item-Total Correlation	Mean Omit Rate	Mean Not Reached Rate
Reading	3	349.09 (16.36)	34.47 (11.72)	0.67	0.47	0.78	0.15
	4	451.07 (15.78)	31.65 (11.33)	0.61	0.43	0.55	0.15
	5	552.62 (15.01)	34.94 (10.94)	0.68	0.46	0.40	0.09
	6	650.89 (14.45)	31.63 (10.84)	0.62	0.42	0.50	0.24
	7	753.72 (13.73)	33.55 (10.96)	0.64	0.43	0.55	0.20
	8	853.50 (15.06)	31.13 (10.97)	0.60	0.41	0.84	0.51
	10	951.18 (14.27)	33.09 (11.51)	0.64	0.44	1.39	0.72
Mathematics	3	352.60 (17.96)	39.69 (12.15)	0.68	0.44	0.84	0.13
	4	454.66 (15.71)	36.03 (11.82)	0.61	0.41	0.61	0.27
	5	556.54 (15.99)	37.88 (11.49)	0.66	0.40	0.38	0.12
	6	649.21 (16.88)	32.54 (13.10)	0.57	0.44	0.57	0.21
	7	750.72 (17.00)	32.14 (12.83)	0.55	0.42	0.75	0.28
	8	847.48 (16.66)	27.61 (11.14)	0.48	0.36	1.09	0.60
	10	943.60 (18.80)	27.51 (12.45)	0.47	0.40	2.22	1.17
Science/ Biology	5	548.34 (11.47)	25.16 (9.37)	0.48	0.34	0.77	0.37
	8	848.58 (15.67)	21.04 (8.75)	0.41	0.31	1.92	0.76
	High School	947.18 (14.27)	19.19 (7.85)	0.37	0.27	1.95	0.74
Composition	4	N/A	5.73 (1.76)	0.57	0.91	N/A	N/A
	7	N/A	6.23 (1.86)	0.62	0.92	N/A	N/A
	10	N/A	5.35 (2.06)	0.54	0.94	N/A	N/A

Note. Omit and not reached rates are percentages.

DC CAS Performance Level Distributions

Using scores for all students with valid test administrations, the 2010 results are presented in Table 47.

Table 47. DC CAS 2010 Percentages of Students at Each Performance Level

Content	Performance Level	Grade						
		3	4	5	6	7	8	10[1]
Reading	N	4,932	4,841	4,518	4,537	4,389	4,542	4,416
	Below Basic	22.34	18.30	14.98	15.41	11.30	13.45	17.48
	Basic	36.19	36.58	39.18	43.16	40.90	38.68	44.18
	Proficient	35.67	39.10	39.66	36.85	35.86	36.33	30.82
	Advanced	5.80	6.01	6.18	4.58	11.94	11.54	7.52
Mathematics	N	4,956	4,868	4,536	4,561	4,403	4,547	4,388
	Below Basic	22.70	19.54	16.56	18.50	18.40	17.81	25.30
	Basic	40.03	35.29	38.67	38.24	29.73	32.46	37.60
	Proficient	29.38	35.50	34.19	31.68	41.40	40.77	30.97
	Advanced	7.89	9.68	10.58	11.58	10.47	8.95	6.13
Science/ Biology	N	--	--	4,463	--	--	4,407	4,113
	Below Basic	--	--	19.83	--	--	41.48	30.37
	Basic	--	--	42.19	--	--	24.78	29.44
	Proficient	--	--	31.46	--	--	28.82	38.17
	Advanced	--	--	6.52	--	--	4.92	2.02
Composition	N	--	4,555	--	--	4,229	--	3,837
	Below Basic	--	11.33	--	--	6.83	--	17.64
	Basic	--	56.47	--	--	47.79	--	53.56
	Proficient	--	28.01	--	--	33.84	--	22.41
	Advanced	--	4.19	--	--	11.54	--	6.39

Note. Total percentages for a grade may not sum to 100 due to rounding.

[1] Biology is administered to students in grades 8-12, the grade in which they elect to take the Biology course.

Means and Standard Deviations by Race/Ethnicity and Gender

Means and standard deviations for subgroups of the examinee population are presented in Tables 48–51 for Reading, Mathematics, Science/Biology, and Composition, respectively. African Americans make up the largest subgroup of students at each grade, followed by Hispanics, Whites, and Asian/Pacific Islanders. There are similar numbers of males and females, especially at the elementary grades. Mean performance by race/ethnicity generally shows that White students achieve the highest mean scores, followed by Asian/Pacific Islander students, Hispanic students, and African American students.

Table 48. 2010 Subgroup Scale Score Means and Standard Deviations: Reading

Grade	Subgroup	N	Mean	SD
3	All Examinees	4,932	349.09	16.36
	Male	2,471	346.38	16.90
	Female	2,440	351.84	15.31
	Asian/Pacific Islander	86	359.93	12.11
	African American	3,822	347.19	15.81
	Hispanic	594	347.87	15.76
	White	404	365.61	11.88
4	All Examinees	4,841	451.07	15.78
	Male	2,430	449.14	16.52
	Female	2,393	453.05	14.77
	Asian/Pacific Islander	75	461.79	15.63
	African American	3,797	449.24	15.07
	Hispanic	586	451.21	15.14
	White	362	467.53	13.41
5	All Examinees	4,518	552.62	15.01
	Male	2,297	550.58	15.87
	Female	2,208	554.76	13.75
	Asian/Pacific Islander	61	565.98	11.41
	African American	3,636	550.90	14.65
	Hispanic	505	554.21	13.44
	White	301	567.50	12.27
6	All Examinees	4,537	650.89	14.45
	Male	2,286	649.14	15.28
	Female	2,230	652.73	13.31
	Asian/Pacific Islander	54	660.39	14.42
	African American	3,732	649.51	13.81
	Hispanic	457	650.82	13.35
	White	267	668.15	12.45
7	All Examinees	4,389	753.72	13.73
	Male	2,180	751.60	14.36
	Female	2,186	755.89	12.68
	Asian/Pacific Islander	48	763.60	9.10
	African American	3,721	752.69	13.06
	Hispanic	394	753.98	13.78
	White	207	769.96	14.17
8	All Examinees	4,542	853.50	15.06
	Male	2,237	851.50	15.97
	Female	2,271	855.62	13.73
	Asian/Pacific Islander	62	865.11	15.03
	African American	3,844	852.30	14.45
	Hispanic	426	854.97	14.29
	White	176	873.28	13.58

10	All Examinees	4,416	951.18	14.27
	Male	2,078	949.72	15.09
	Female	2,264	952.86	13.09
	Asian/Pacific Islander	58	958.50	14.56
	African American	3,742	950.49	13.71
	Hispanic	380	951.49	13.55
	White	162	968.09	14.34

Note. Results are based on students with valid test scores. See the section *Participation in the 2010 DC CAS Test Administrations and Use of Data for Analysis and Score Reporting* for an explanation.

Table 49. 2010 Subgroup Scale Score Means and Standard Deviations: Mathematics

Grade	Subgroup	N	Mean	SD
3	All Examinees	4,956	352.60	17.96
	Male	2,487	351.32	18.47
	Female	2,447	353.96	17.25
	Asian/Pacific Islander	90	367.93	13.50
	African American	3,830	350.20	17.38
	Hispanic	604	352.48	15.97
	White	405	371.33	13.42
4	All Examinees	4,868	454.66	15.71
	Male	2,448	454.24	16.24
	Female	2,399	455.23	15.05
	Asian/Pacific Islander	81	470.12	13.15
	African American	3,800	452.52	14.84
	Hispanic	600	456.30	14.66
	White	365	470.94	14.78
5	All Examinees	4,536	556.54	15.99
	Male	2,310	555.71	16.81
	Female	2,213	557.44	15.03
	Asian/Pacific Islander	64	571.16	12.84
	African American	3,639	554.43	15.18
	Hispanic	514	559.04	14.87
	White	304	573.88	14.55
6	All Examinees	4,561	649.21	16.88
	Male	2,293	647.94	17.51
	Female	2,245	650.63	16.03
	Asian/Pacific Islander	57	664.49	19.31
	African American	3,739	647.54	16.08
	Hispanic	469	650.20	15.67
	White	269	667.19	16.23
7	All Examinees	4,403	750.72	17.00
	Male	2,189	749.67	17.72
	Female	2,189	751.91	16.10
	Asian/Pacific Islander	60	766.97	14.89
	African American	3,713	749.23	16.29
	Hispanic	403	751.90	16.00
	White	207	771.20	14.68

8	All Examinees	4,547	847.48	16.66
	Male	2,236	846.63	17.24
	Female	2,277	848.44	16.02
	Asian/Pacific Islander	65	864.15	14.07
	African American	3,822	846.28	16.16
	Hispanic	450	848.96	15.71
	White	176	865.81	15.12
10	All Examinees	4,388	943.60	18.80
	Male	2,058	942.61	19.92
	Female	2,255	944.82	17.57
	Asian/Pacific Islander	58	960.45	15.81
	African American	3,717	942.33	18.20
	Hispanic	377	945.92	16.92
	White	161	965.94	18.84

Note. Results are based on students with valid test scores. See the section *Participation in the 2010 DC CAS Test Administrations and Use of Data for Analysis and Score Reporting* for an explanation.

Table 50. DC CAS 2010 Subgroup Scale Score Means and Standard Deviations: Science/Biology

Grade	Subgroup	N	Mean	SD
5	All Examinees	4,463	548.34	11.47
	Male	2,261	547.91	12.05
	Female	2,181	548.80	10.77
	Asian/Pacific Islander	64	559.02	8.85
	African American	3,559	546.75	11.01
	Hispanic	510	549.77	9.46
	White	303	562.36	8.61
8	All Examinees	4,407	848.58	15.67
	Male	2,140	847.95	16.47
	Female	2,217	849.38	14.75
	Asian/Pacific Islander	64	857.50	14.29
	African American	3,689	847.66	15.62
	Hispanic	441	849.81	14.23
	White	170	864.16	8.42
High School	All Examinees	4,113	947.18	14.27
	Male	1,925	946.49	14.94
	Female	2,060	948.27	13.18
	Asian/Pacific Islander	52	957.38	7.02
	African American	3,441	946.52	14.07
	Hispanic	378	948.88	13.48
	White	141	961.25	7.01

Note. Results are based on students with valid test scores. See the section *Participation in the 2010 DC CAS Test Administrations and Use of Data for Analysis and Score Reporting* for an explanation.

Table 51. DC CAS 2010 Subgroup Scale Score Means and Standard Deviations: Composition

Grade	Subgroup	N	Mean	SD
4	All Examinees	4,555	5.73	1.76
	Male	2,266	5.42	1.77
	Female	2,263	6.05	1.68
	Asian/Pacific Islander	73	6.84	1.65
	African American	3,551	5.54	1.70
	Hispanic	556	5.83	1.66
	White	348	7.27	1.61
7	All Examinees	4,229	6.23	1.86
	Male	2,080	5.83	1.84
	Female	2,110	6.65	1.78
	Asian/Pacific Islander	50	7.28	1.52
	African American	3,559	6.10	1.80
	Hispanic	388	6.33	1.85
	White	202	8.30	1.50
10	All Examinees	3,837	5.35	2.06
	Male	1,711	5.03	2.06
	Female	2,039	5.66	2.01
	Asian/Pacific Islander	51	6.71	1.83
	African American	3,235	5.24	2.02
	Hispanic	331	5.58	1.93
	White	149	7.25	2.01

Note. Results are based on students with valid test scores. See the section *Participation in the 2010 DC CAS Test Administrations and Use of Data for Analysis and Score Reporting* for an explanation.

Correlations

This section contains information relevant to the *Standards and Assessment Peer Review Guidance*, Critical Element 4.1:

For each assessment, including all alternate assessments, has the State documented the issue of **validity** (in addition to the alignment of the assessment with the content standards), as described in the *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 1999), with respect to *all* of the following categories:

(e) Has the State ascertained that test and item scores are related to outside variables as intended (e.g., scores are correlated strongly with relevant measures of academic achievement and are weakly correlated, if at all, with irrelevant characteristics, such as demographics)?

Using all scored data, the correlations among the Reading, Mathematics, Science/Biology, and Composition raw scores were calculated as a way of examining evidence of the validity of inferences about student achievement based on relationships between content area tests. This evidence is referred to as evidence of convergent and discriminant validity. The correlations among Reading, Mathematics, Science/Biology, and Composition total raw scores appear in Table 52.

These results are consistent with typical content area correlations for educational achievement tests in these content areas. The correlations are somewhat higher in the elementary grades than in the middle and high school grades. Correlations between Reading and Mathematics are 0.73 and higher; correlations of Reading and Mathematics scores with Science/Biology scores are 0.64 and higher; correlations with the Composition total scores are in the range of 0.45 to 0.64. Composition correlations are relatively lower because Composition scores range from 2 to 10, which restricts variability and covariance. Correlations in all content areas are based on 3,670 to 4,899 cases, except in Biology. The differences between numbers of test takers in each grade (more than 4,400 per grade; see Tables 7–10) and the numbers in these correlations is due to loss of cases when files were merged to calculate the correlations.

These correlations are moderately high. They indicate that approximately 25–50 percent of the variability in performance on these separate content area tests can be accounted for by skills and proficiency shared across the content areas (i.e., disregarding measurement error). This observation suggests that one half to three quarters of the performance on each content area assessment can be explained by knowledge, skills, and proficiency that are unique to each content area.

Table 52. Correlations Among Reading, Mathematics, Science/Biology, and Composition Total Test Raw Scores, by Grade

Grade	Mathematics	Science/ Biology*	Composition
Reading			
Grade 3	0.80	--	--
Grade 4	0.77	--	0.60
Grade 5	0.75	0.74	--
Grade 6	0.75	--	--
Grade 7	0.76	--	0.64
Grade 8	0.73	0.72	--
Grade 10	0.73	0.64	0.61
Mathematics			
Grade 4	--	--	0.55
Grade 5	--	0.72	--
Grade 7	--	--	0.58
Grade 8	--	0.74	--
Grade 10	--	0.64	0.58
Science/Biology			
Grade 10	--	--	0.45

Note. “--“ = not applicable.
*In Biology all grades were used in the analyses but only Grade 10 can be used for the correlations since the other grades are not in common.

Correlations of Strand Scores and Total Content Area Scores

This section contains information relevant to the *Standards and Assessment Peer Review Guidance,* Critical Element 4.1:

For each assessment, including all alternate assessments, has the State documented the issue of **validity** (in addition to the alignment of the assessment with the content standards), as described in the *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 1999), with respect to *all* of the following categories:

(c) Has the State ascertained that the scoring and reporting structures are consistent with the sub-domain structures of its academic content standards (i.e., are item interrelationships consistent with the framework from which the test arises)?

Correlations among strand and total content areas raw scores also provide evidence to support the validity of interpretations of test scores. Correlations among strand scores within a content area test indicate the degree to which strand scores provide unique evidence about student proficiency. In Table 53, the DC CAS 2010 Reading strand and total test correlations for all grades are presented. The Reading correlations are moderate to high among each other and the total Reading test for all grades.

Table 54 displays the correlations for the DC CAS 2010 Mathematics strand and total test scores by grade. The correlations are mostly moderate to high. The correlations between Geometry and the other Mathematics strands are lower than for the other strands. Geometry and Measurement also tend to have the lowest correlation with the Mathematics total raw score at each grade. This is due in part to the smaller number of items used to measure Geometry and Measurement in relation to the rest of the content strands.

In Table 55, the DC CAS 2010 Science/Biology strand and total test correlations for all grades are presented. The correlations are moderate to high and somewhat lower in general than the correlations in Reading and Mathematics.

Table 53. DC CAS 2010 Reading Strand Correlations by Grade

Grade	Content Strand	Language Development	Informational Text	Literary Text	Total Reading
3	Language Development	--	0.79	0.79	0.89
	Informational Text	0.79	--	0.83	0.95
	Literary Text	0.79	0.83	--	0.95
	Total Raw Score	0.89	0.95	0.95	--
4	Language Development	--	0.77	0.78	0.89
	Informational Text	0.77	--	0.80	0.93
	Literary Text	0.78	0.80	--	0.95
	Total Raw Score	0.89	0.93	0.95	--
5	Language Development	--	0.76	0.75	0.88
	Informational Text	0.76	--	0.82	0.93
	Literary Text	0.75	0.82	--	0.95
	Total Raw Score	0.88	0.93	0.95	--
6	Language Development	--	0.74	0.75	0.86
	Informational Text	0.74	--	0.78	0.92
	Literary Text	0.75	0.78	--	0.95
	Total Raw Score	0.86	0.92	0.95	--
7	Language Development	--	0.73	0.75	0.86
	Informational Text	0.73	--	0.78	0.91
	Literary Text	0.75	0.78	--	0.95
	Total Raw Score	0.86	0.91	0.95	--
8	Language Development	--	0.68	0.76	0.85
	Informational Text	0.68	--	0.77	0.88
	Literary Text	0.76	0.77	--	0.97
	Total Raw Score	0.85	0.88	0.97	--
10	Language Development	--	0.75	0.76	0.87
	Informational Text	0.75	--	0.82	0.94
	Literary Text	0.76	0.82	--	0.95
	Total Raw Score	0.87	0.94	0.95	--

Table 54. DC CAS 2010 Mathematics Strand Correlations by Grade

Grade	Content Strand	Number Sense & Operations	Patterns, Relations, & Algebra	Geometry	Measurement	Data Analysis, Statistics, & Probability	Total Mathematics
3	Number Sense & Operations	--	0.80	0.68	0.69	0.78	0.93
	Patterns, Relations, & Algebra	0.80	--	0.65	0.66	0.75	0.90
	Geometry	0.68	0.65	--	0.59	0.68	0.80
	Measurement	0.69	0.66	0.59	--	0.66	0.80
	Data Analysis, Statistics, & Probability	0.78	0.75	0.68	0.66	--	0.90
	Total Raw Score	0.93	0.90	0.80	0.80	0.90	--
4	Number Sense & Operations	--	0.80	0.66	0.71	0.71	0.93
	Patterns, Relations, & Algebra	0.80	--	0.64	0.69	0.68	0.90
	Geometry	0.66	0.64	--	0.60	0.62	0.78
	Measurement	0.71	0.69	0.60	--	0.63	0.82
	Data Analysis, Statistics, & Probability	0.71	0.68	0.62	0.63	--	0.84
	Total Raw Score	0.93	0.90	0.78	0.82	0.84	--
5	Number Sense & Operations	--	0.76	0.65	0.69	0.69	0.92
	Patterns, Relations, & Algebra	0.76	--	0.64	0.65	0.67	0.89
	Geometry	0.65	0.64	--	0.59	0.58	0.79
	Measurement	0.69	0.65	0.59	--	0.57	0.82
	Data Analysis, Statistics, & Probability	0.69	0.67	0.58	0.57	--	0.81
	Total Raw Score	0.92	0.89	0.79	0.82	0.81	--
6	Number Sense & Operations	--	0.82	0.58	0.71	0.76	0.93
	Patterns, Relations, & Algebra	0.82	--	0.60	0.71	0.77	0.93
	Geometry	0.58	0.60	--	0.56	0.56	0.71
	Measurement	0.71	0.71	0.56	--	0.68	0.82
	Data Analysis, Statistics, & Probability	0.76	0.77	0.56	0.68	--	0.88
	Total Raw Score	0.93	0.93	0.71	0.82	0.88	--

7	Number Sense & Operations	--	0.79	0.66	0.66	0.74	0.91
	Patterns, Relations, & Algebra	0.79	--	0.69	0.67	0.74	0.92
	Geometry	0.66	0.69	--	0.61	0.66	0.82
	Measurement	0.66	0.67	0.61	--	0.70	0.80
	Data Analysis, Statistics, & Probability	0.74	0.74	0.66	0.70	--	0.87
	Total Raw Score	0.91	0.92	0.82	0.80	0.87	--
8	Number Sense & Operations	--	0.72	0.54	0.63	0.67	0.89
	Patterns, Relations, & Algebra	0.72	--	0.55	0.67	0.71	0.90
	Geometry	0.54	0.55	--	0.49	0.52	0.71
	Measurement	0.63	0.67	0.49	--	0.61	0.79
	Data Analysis, Statistics, & Probability	0.67	0.71	0.52	0.61	--	0.84
	Total Raw Score	0.89	0.90	0.71	0.79	0.84	--
10	Number Sense & Operations	--	0.74	0.70	0.59	0.65	0.86
	Patterns, Relations, & Algebra	0.74	--	0.75	0.59	0.74	0.92
	Geometry	0.70	0.75	--	0.58	0.71	0.88
	Measurement	0.59	0.59	0.58	--	0.56	0.73
	Data Analysis, Statistics, & Probability	0.65	0.74	0.71	0.56	--	0.87
	Total Raw Score	0.86	0.92	0.88	0.73	0.87	--

Table 55. DC CAS 2010 Science/Biology Strand Correlations by Grade

Grade	Content Strand	Scientific Inquiry	Science & Technology	Earth Science	Physical Science	Life Science	Total Science
5	Scientific Inquiry	--	0.56	0.58	0.56	0.59	0.81
	Science & Technology	0.56	--	0.65	0.53	0.64	0.81
	Earth Science	0.58	0.65	--	0.57	0.64	0.85
	Physical Science	0.56	0.53	0.57	--	0.55	0.77
	Life Science	0.59	0.64	0.64	0.55	--	0.85
	Total Raw Score	0.81	0.81	0.85	0.77	0.85	--

Grade	Content Strand	Structure of Matter	Reactions	Conservation of Energy	Forces/ Density & Buoyancy	Scientific Thinking	Total Science
8	Structure of Matter	--	0.55	0.41	0.56	0.64	0.82
	Reactions	0.55	--	0.44	0.58	0.58	0.81
	Conservation of Energy	0.41	0.44	--	0.42	0.44	0.63
	Forces/Density & Buoyancy	0.56	0.58	0.42	--	0.59	0.81
	Scientific Thinking	0.64	0.58	0.44	0.59	--	0.84
	Total Raw Score	0.82	0.81	0.63	0.81	0.84	--

Grade	Content Strand	Scientific Inquiry	Biochemistry	Cell Biology	Genetics	Evolution	Plants/ Mammalian Body	Ecology	Total Biology
High School	Scientific Inquiry	--	0.20	0.40	0.42	0.36	0.40	0.44	0.68
	Biochemistry	0.20	--	0.28	0.24	0.21	0.23	0.23	0.44
	Cell Biology	0.40	0.28	--	0.47	0.39	0.45	0.44	0.72
	Genetics	0.42	0.24	0.47	--	0.39	0.48	0.51	0.75
	Evolution	0.36	0.21	0.39	0.39	--	0.37	0.40	0.62
	Plants/Mammalian Body	0.40	0.23	0.45	0.48	0.37	--	0.54	0.76
	Ecology	0.44	0.23	0.44	0.51	0.40	0.54	--	0.78
	Total Raw Score	0.68	0.44	0.72	0.75	0.62	0.76	0.78	--

The DC CAS 2010 rubric score and total Composition test correlations for all grades are presented in Table 56. The correlations between the Topic/Idea Development and Language Conventions scores are moderately high, suggesting that each rubric assesses somewhat different composing skills, as intended. The correlations between the rubric scores and total Composition scores are high, as expected.

Table 56. DC CAS 2010 Composition Rubric Score Correlations by Grade

Grade	Content Strand	Topic/Idea Development	Language Conventions	Composition Total
4	Topic/Idea Development	--	0.65	0.91
	Language Conventions	0.65	--	0.90
7	Topic/Idea Development	--	0.72	0.95
	Language Conventions	0.72	--	0.90
10	Topic/Idea Development	--	0.78	0.95
	Language Conventions	0.78	--	0.93

Note. Correlations based on 4,555 cases in grade 4; 4,229 in grade 7; and 3,837 in grade 10.

Section 11. DC CAS 2010 Field Test

This section contains information relevant to the *Standards and Assessment Peer Review Guidance*, Critical Element 4.5:

Has the State established clear criteria for the administration, scoring, analysis, and reporting components of its assessment system, including all alternate assessments, and does the State have a system for monitoring and improving the on-going quality of its assessment system?

In spring 2010, four sets of field test items were embedded in the operational test forms for Reading, Mathematics, and Science/Biology. In past years, only two sets of field test items were embedded in operational forms. The increase in 2010 was intended to expand the bank of items available for use in 2011 and subsequent and operational test forms and to address specific item bank needs (e.g., relatively easy items). Analysis of the 2010 field test items will be completed subsequent to release of this operational technical report. Results from the field test analyses will be documented in a technical memo that will cover the following topics:

- Field test item development
- Hand-scoring of field test items
- Calibrating and scaling the field test items
- Field testing of Composition prompts in 2006 and selection for operational use

These topics are consistent with the the section headers on previous years' field test results, contained in the technical reports for those years.

References

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

Burket, G. R. (1995). PARDUX (Version 1.7) [Computer program]. Unpublished.

Hambleton, R. K. & Novick, M. R. (1973). Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement*, *10*, 159–170.

Kim, D. (2007). KKCLASS [Computer program]. Unpublished.

Kim, D., Barton, K, & Kim, X. (2008). *Estimating Classification Consistency and Classification Accuracy With Pattern Scoring.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Kim, D., Choi, S., Um, K., & Kim, J. (2006). *A Comparison of Methods for Estimating Classification Consistency*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Kolen, M. J. & Kim, D. (2005). Personal correspondence.

Landis, J. R. & Koch, G. G. (1997). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33,* 159–174.

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 presented at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix, AZ.

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

Mantel, N. & Haenszel, W. (1959) Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*, 719–748.

Muraki, E. & Bock, R. D. (1991). *PARSCALE*: Parameter Scaling of Rating Data [Computer program]. Chicago, IL: Scientific Software, Inc.

Perie, M. (2007, June). *Setting alternate achievement standards.* Dover, NH: National Center for the Improvement of Educational Assessment. Retrieved January 11, 2008 from http://www.nciea.org/publications/CCSSO_MAP07.pdf.

Roeber, E. (2002). *Setting standards on alternate assessments (Synthesis Report 42).* Minneapolis, MN: National Center on Educational Outcomes. Retrieved January 11, 2008 from http://cehd.umn.edu/NCEO/OnlinePubs/Synthesis42.html

Standards and Assessment Peer Review Guidance. (January 12, 2010). Retrieved December 7, 2010 from http://www.ed.gov/policy/elsec/guid/saaprguidance.pdf.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2009). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

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

Swaminathan, H., Hambleton, R. K., & Algina, J. (1974). Reliability of Criterion-Referenced Tests: A Decision-Theoretic Formulation, *Journal of Educational Measurement*, Vol. *11*, No. 4 (Winter, 1974), pp. 263–267.

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

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

Zwick, R., Donoghue, J.R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*, 233–251.

Appendix A: Checklist for DC Educator Review of DC CAS Items

A. Checklist for the Content Reviewer

For All Items:

Check to ensure that the content of each item:

- □ is targeted to assess only one strand or skill
- □ deals with material that is important in testing the targeted strand or skill
- □ uses grade-appropriate content and thinking skills
- □ is presented at a reading level suitable for the grade level being tested
- □ is accurate and documented against reliable, up-to-date sources

For Multiple-Choice Items:

Check to ensure that the content of each item:

- □ has a stem that facilitates answering the question or completing the statement without looking at the answer choices
- □ has a stem that does not present clues to the correct answer choice
- □ has answer choices that are plausible and attractive to the student who has not mastered the Strand or skill
- □ is conceptually, grammatically, and syntactically consistent—between the stem and answer choices, and among the answer choices
- □ has mutually exclusive distractors
- □ has one and only one correct answer choice

For Constructed-Response Items:

Check to ensure that the content of each item:

- □ is written so that a student possessing the knowledge or skill being tested can construct a response that is scorable with the specified rubric or scoring tool; that is, the range of possible correct responses must be wide enough to allow for diversity of responses, but narrow enough so that students who do not clearly show their grasp of the Strand or skill being assessed cannot obtain the maximum score
- □ is presented without clues to the correct response
- □ has precise and unambiguous directions for the desired response
- □ is free of extraneous words or expressions
- □ is appropriate for the question being asked and the intended response (For example, the item does not ask students to draw pictures of abstract ideas.)
- □ is conceptually, grammatically, and syntactically consistent

B. Checklist for the Sensitivity Reviewer

To have confidence in test results, it is important to ensure that students are given a reasonable chance to do their best on the test. Test items must be accessible to a diverse student population with respect to gender, race, ethnicity, geographic region, socioeconomic status, and other factors.

Check to ensure that the content of each item is free of explicit references to or descriptions of:

- ❑ events involving extreme sadness or adversity
- ❑ acts of physical or psychological violence
- ❑ alcohol or drug abuse
- ❑ vulgar language
- ❑ sex

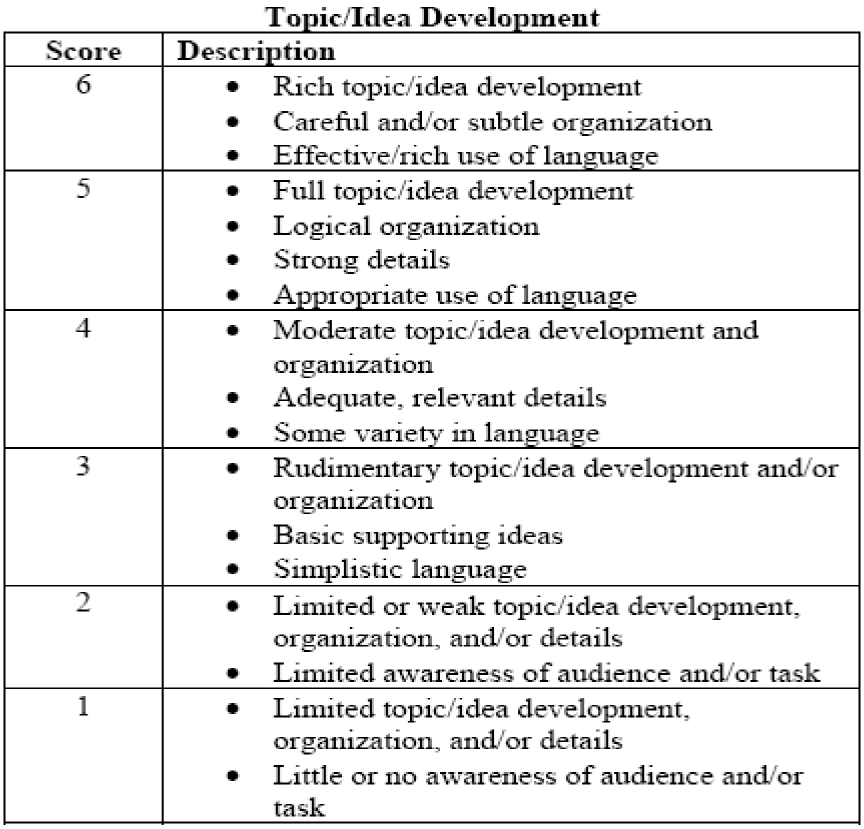
Check to ensure that if any religious, political, social, or philosophical issues are addressed:

- ❑ more than one point of view is expressed
- ❑ beliefs or biases do not interfere with factual accuracy
- ❑ contemporary issues that have already been proven to be controversial are absent
- ❑ stereotypic descriptions of beliefs or customs are absent

Test items must:

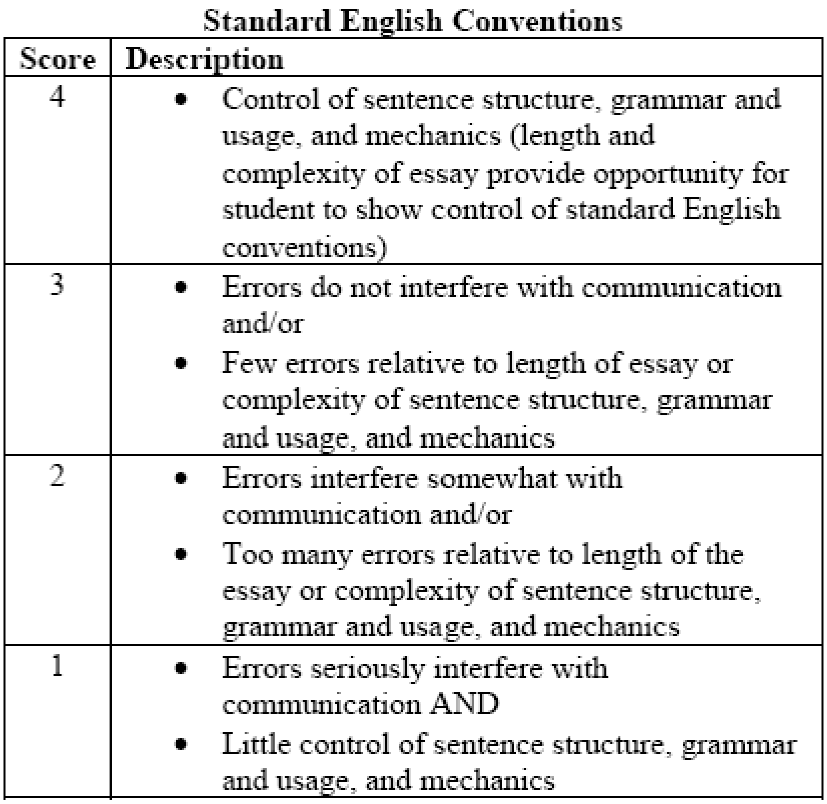
- ❑ be free of offensive, disturbing, or inappropriate language or content
- ❑ be free of stereotyping based on:
  - gender
  - race
  - ethnicity
  - religion
  - socioeconomic status
  - age
  - regional or geographic area
  - disability
  - occupation
- ❑ demonstrate sensitivity to historical representation of groups
- ❑ be free of differential familiarity for any group based on:
  - language
  - socioeconomic status
  - regional or geographic area
  - prior knowledge or experiences unrelated to the subject matter being tested

Appendix B: DC CAS Composition Scoring Rubrics



Topic/Idea Development

Score	Description
6	• Rich topic/idea development • Careful and/or subtle organization • Effective/rich use of language
5	• Full topic/idea development • Logical organization • Strong details • Appropriate use of language
4	• Moderate topic/idea development and organization • Adequate, relevant details • Some variety in language
3	• Rudimentary topic/idea development and/or organization • Basic supporting ideas • Simplistic language
2	• Limited or weak topic/idea development, organization, and/or details • Limited awareness of audience and/or task
1	• Limited topic/idea development, organization, and/or details • Little or no awareness of audience and/or task



Standard English Conventions

Score	Description
4	• Control of sentence structure, grammar and usage, and mechanics (length and complexity of essay provide opportunity for student to show control of standard English conventions)
3	• Errors do not interfere with communication and/or • Few errors relative to length of essay or complexity of sentence structure, grammar and usage, and mechanics
2	• Errors interfere somewhat with communication and/or • Too many errors relative to length of the essay or complexity of sentence structure, grammar and usage, and mechanics
1	• Errors seriously interfere with communication AND • Little control of sentence structure, grammar and usage, and mechanics

Appendix C: Internal Consistency Reliability Coefficients for Examinee Subgroups

(See Section 5. Evidence for Reliability and Validity, *Internal Consistency Reliability*, at Table 17)

Table C1. Internal Consistency Reliability Coefficients for Examinee Subgroups: Reading

Grade/Subgroup	Students with Valid Test Scores	Number of Items	Alpha	Stratified Alpha	Feldt-Raju
3					
Males	2,470	48	0.933	0.938	0.937
Females	2,437		0.922	0.929	0.928
Asian/Pacific Islander	86		0.882	0.894	0.891
African American	3,820		0.927	0.932	0.931
Hispanic	592		0.923	0.929	0.928
White	404		0.860	0.867	0.868
4					
Males	2,421	48	0.925	0.928	0.928
Females	2,390		0.917	0.921	0.921
Asian/Pacific Islander	75		0.915	0.917	0.918
African American	3,791		0.913	0.917	0.916
Hispanic	584		0.918	0.922	0.922
White	360		0.907	0.914	0.913
5					
Males	2,293	48	0.935	0.938	0.937
Females	2,204		0.921	0.925	0.924
Asian/Pacific Islander	61		0.862	0.872	0.873
African American	3,630		0.927	0.930	0.929
Hispanic	503		0.920	0.922	0.922
White	301		0.905	0.913	0.910
6					
Males	2,275	48	0.920	0.921	0.923
Females	2,224		0.907	0.908	0.909
Asian/Pacific Islander	54		0.920	0.923	0.925
African American	3,720		0.908	0.909	0.911
Hispanic	454		0.905	0.906	0.908
White	266		0.895	0.899	0.902
7					
Males	2,175	48	0.920	0.925	0.924
Females	2,184		0.910	0.915	0.915
Asian/Pacific Islander	48		0.865	0.866	0.876
African American	3,715		0.911	0.916	0.915
Hispanic	394		0.921	0.926	0.925
White	207		0.906	0.919	0.916
8					
Males	2,229	48	0.914	0.919	0.919
Females	2,265		0.900	0.905	0.905
Asian/Pacific Islander	62		0.911	0.917	0.919
African American	3,830		0.901	0.907	0.906
Hispanic	426		0.907	0.912	0.912

Grade/Subgroup	Students with Valid Test Scores	Number of Items	Alpha	Stratified Alpha	Feldt-Raju
White	175		0.890	0.903	0.903
10					
Males	2,068	48	0.930	0.936	0.934
Females	2,253		0.915	0.922	0.920
Asian/Pacific Islander	58		0.922	0.932	0.928
African American	3,723		0.920	0.927	0.925
Hispanic	380		0.918	0.925	0.923
White	161		0.932	0.940	0.939

Table C2. Internal Consistency Reliability Coefficients for Examinee Subgroups: Mathematics

Grade/Subgroup	Students with Valid Test Scores	Number of Items	Alpha	Stratified Alpha	Feldt-Raju
3					
Males	2,480	54	0.933	0.937	0.938
Females	2,442		0.927	0.931	0.932
Asian/Pacific Islander	90		0.873	0.878	0.884
African American	3,820		0.926	0.930	0.931
Hispanic	603		0.921	0.924	0.926
White	404		0.872	0.876	0.880
4					
Males	2,445	54	0.923	0.927	0.927
Females	2,399		0.919	0.922	0.923
Asian/Pacific Islander	81		0.907	0.912	0.914
African American	3,800		0.911	0.914	0.915
Hispanic	600		0.918	0.921	0.921
White	364		0.922	0.925	0.926
5					
Males	2,308	54	0.923	0.928	0.928
Females	2,212		0.910	0.915	0.916
Asian/Pacific Islander	64		0.880	0.886	0.892
African American	3,637		0.909	0.913	0.914
Hispanic	513		0.910	0.915	0.916
White	304		0.904	0.913	0.912
6					
Males	2,285	54	0.935	0.938	0.939
Females	2,241		0.929	0.933	0.933
Asian/Pacific Islander	57		0.952	0.958	0.959
African American	3,729		0.925	0.928	0.929
Hispanic	468		0.927	0.930	0.931
White	267		0.938	0.943	0.945
7					
Males	2,184	54	0.929	0.936	0.936
Females	2,182		0.919	0.927	0.927
Asian/Pacific Islander	60		0.926	0.933	0.934
African American	3,702		0.918	0.924	0.925

Grade/Subgroup	Students with Valid Test Scores	Number of Items	Alpha	Stratified Alpha	Feldt-Raju
Hispanic	403		0.923	0.930	0.931
White	207		0.922	0.928	0.930
8					
Males	2,225	54	0.906	0.911	0.912
Females	2,269		0.899	0.904	0.905
Asian/Pacific Islander	65		0.921	0.922	0.925
African American	3,808		0.890	0.895	0.896
Hispanic	448		0.898	0.905	0.905
White	174		0.916	0.920	0.921
10					
Males	2,041	54	0.926	0.931	0.932
Females	2,244		0.914	0.919	0.921
Asian/Pacific Islander	58		0.926	0.929	0.933
African American	3,692		0.911	0.917	0.918
Hispanic	375		0.910	0.915	0.917
White	160		0.945	0.949	0.951

Table C3. Internal Consistency Reliability Coefficients for Examinee Subgroups: Science/Biology

Grade/Subgroup	Students with Valid Test Scores	Number of Items	Alpha	Stratified Alpha	Feldt-Raju
5					
Males	2,259	50	0.891	0.892	0.893
Females	2,178		0.874	0.875	0.876
Asian/Pacific Islander	64		0.898	0.900	0.902
African American	3,554		0.850	0.852	0.853
Hispanic	510		0.865	0.866	0.867
White	303		0.885	0.886	0.889
8					
Males	2,135	50	0.878	0.879	0.880
Females	2,208		0.856	0.857	0.859
Asian/Pacific Islander	64		0.918	0.920	0.922
African American	3,676		0.847	0.847	0.849
Hispanic	440		0.852	0.853	0.855
White	170		0.892	0.897	0.897
High School					
Males	1,917	50	0.838	0.840	0.841
Females	2,052		0.827	0.828	0.830
Asian/Pacific Islander	52		0.877	0.877	0.880
African American	3,427		0.795	0.797	0.800
Hispanic	377		0.834	0.836	0.837
White	141		0.891	0.892	0.894

Appendix D: Classification Consistency and Accuracy Results for Each Proficiency Level in Each Grade and Content Area Assessment

Table D1. Classification Consistency and Accuracy Rates by Grade and Cut Score: Reading

Grade	Reading Classification Consistency and Accuracy		Basic	Proficient	Advanced	All Cuts
3	Classification Consistency	Consistency	0.9431	0.8930	0.9379	0.7740
		Kappa	0.8356	0.7807	0.5602	0.6767
	Classification Accuracy	Accuracy	0.9593	0.9186	0.9522	0.8300
		False Positive Errors	0.0196	0.0267	0.0064	0.0527
		False Negative Errors	0.0211	0.0547	0.0414	0.1172
4	Classification Consistency	Consistency	0.9335	0.8874	0.9442	0.7652
		Kappa	0.7771	0.7728	0.5931	0.6570
	Classification Accuracy	Accuracy	0.9527	0.9215	0.9616	0.8359
		False Positive Errors	0.0243	0.0267	0.0135	0.0645
		False Negative Errors	0.0229	0.0518	0.0248	0.0996
5	Classification Consistency	Consistency	0.9467	0.8785	0.9430	0.7684
		Kappa	0.7893	0.7552	0.6096	0.6561
	Classification Accuracy	Accuracy	0.9605	0.9105	0.9585	0.8295
		False Positive Errors	0.0164	0.0296	0.0068	0.0528
		False Negative Errors	0.0231	0.0599	0.0348	0.1177
6	Classification Consistency	Consistency	0.9302	0.8794	0.9532	0.7628
		Kappa	0.7298	0.7527	0.5825	0.6410
	Classification Accuracy	Accuracy	0.9486	0.9102	0.9663	0.8252
		False Positive Errors	0.0214	0.0278	0.0094	0.0585
		False Negative Errors	0.0300	0.0620	0.0243	0.1163
7	Classification Consistency	Consistency	0.9388	0.8850	0.9248	0.7487
		Kappa	0.7049	0.7696	0.6781	0.6342
	Classification Accuracy	Accuracy	0.9566	0.9204	0.9464	0.8233
		False Positive Errors	0.0257	0.0386	0.0198	0.0841
		False Negative Errors	0.0177	0.0411	0.0338	0.0926
8	Classification Consistency	Consistency	0.9273	0.8786	0.9307	0.7373
		Kappa	0.6942	0.7570	0.6933	0.6226
	Classification Accuracy	Accuracy	0.9476	0.9074	0.9479	0.8028
		False Positive Errors	0.0186	0.0214	0.0111	0.0511
		False Negative Errors	0.0338	0.0712	0.0410	0.1460
10	Classification Consistency	Consistency	0.9329	0.8881	0.9421	0.7636
		Kappa	0.7651	0.7630	0.6371	0.6510
	Classification Accuracy	Accuracy	0.9480	0.9210	0.9598	0.8288
		False Positive Errors	0.0153	0.0313	0.0128	0.0595
		False Negative Errors	0.0367	0.0477	0.0274	0.1117

Table D2. Classification Consistency and Accuracy Rates by Grade and Cut Score: Mathematics

Grade	Mathematics Classification Consistency and Accuracy		Basic	Proficient	Advanced	All Cuts
3	Classification Consistency	Consistency	0.9228	0.9018	0.9421	0.7670
		Kappa	0.7795	0.7922	0.6586	0.6689
	Classification Accuracy	Accuracy	0.9485	0.9263	0.9567	0.8316
		False Positive Errors	0.0279	0.0222	0.0109	0.0609
		False Negative Errors	0.0236	0.0515	0.0324	0.1074
4	Classification Consistency	Consistency	0.9143	0.8999	0.9435	0.7576
		Kappa	0.7293	0.7983	0.7025	0.6574
	Classification Accuracy	Accuracy	0.9393	0.9232	0.9582	0.8207
		False Positive Errors	0.0276	0.0198	0.0112	0.0586
		False Negative Errors	0.0332	0.0570	0.0306	0.1207
5	Classification Consistency	Consistency	0.9233	0.8916	0.9366	0.7515
		Kappa	0.7241	0.7814	0.6921	0.6455
	Classification Accuracy	Accuracy	0.9462	0.9179	0.9555	0.8196
		False Positive Errors	0.0250	0.0245	0.0204	0.0699
		False Negative Errors	0.0288	0.0576	0.0241	0.1105
6	Classification Consistency	Consistency	0.9167	0.9089	0.9426	0.7684
		Kappa	0.7180	0.8145	0.7355	0.6725
	Classification Accuracy	Accuracy	0.9382	0.9309	0.9586	0.8278
		False Positive Errors	0.0226	0.0238	0.0136	0.0599
		False Negative Errors	0.0392	0.0453	0.0278	0.1123
7	Classification Consistency	Consistency	0.9152	0.9069	0.9471	0.7696
		Kappa	0.7140	0.8136	0.7344	0.6707
	Classification Accuracy	Accuracy	0.9409	0.9331	0.9630	0.8369
		False Positive Errors	0.0352	0.0294	0.0150	0.0796
		False Negative Errors	0.0239	0.0375	0.0220	0.0835
8	Classification Consistency	Consistency	0.8790	0.8764	0.9551	0.7131
		Kappa	0.5963	0.7527	0.7423	0.5862
	Classification Accuracy	Accuracy	0.9146	0.9106	0.9683	0.7934
		False Positive Errors	0.0352	0.0259	0.0089	0.0700
		False Negative Errors	0.0502	0.0635	0.0229	0.1366
10	Classification Consistency	Consistency	0.8815	0.9082	0.9659	0.7560
		Kappa	0.6824	0.8046	0.7332	0.6504
	Classification Accuracy	Accuracy	0.9137	0.9321	0.9722	0.8180
		False Positive Errors	0.0391	0.0232	0.0040	0.0663
		False Negative Errors	0.0472	0.0447	0.0239	0.1157

Table D3. Classification Consistency and Accuracy Rates by Grade and Cut Score: Science/Biology

Grade	Science/Biology Classification Consistency and Accuracy		Basic	Proficient	Ad-vanced	All Cuts
5	Classification Consistency	Consistency	0.8774	0.8778	0.9644	0.7208
		Kappa	0.6209	0.7422	0.7258	0.5924
	Classification Accuracy	Accuracy	0.9115	0.9052	0.9749	0.7916
		False Positive Errors	0.0491	0.0214	0.0077	0.0782
		False Negative Errors	0.0394	0.0735	0.0174	0.1302
8	Classification Consistency	Consistency	0.8254	0.8711	0.9717	0.6854
		Kappa	0.6379	0.7152	0.7180	0.5409
	Classification Accuracy	Accuracy	0.8752	0.9103	0.9778	0.7655
		False Positive Errors	0.0635	0.0268	0.0024	0.0920
		False Negative Errors	0.0613	0.0629	0.0199	0.1425
High School	Classification Consistency	Consistency	0.8104	0.8260	0.9839	0.6555
		Kappa	0.5614	0.6441	0.6505	0.4869
	Classification Accuracy	Accuracy	0.8590	0.8758	0.9885	0.7323
		False Positive Errors	0.0580	0.0433	0.0028	0.1025
		False Negative Errors	0.0830	0.0809	0.0087	0.1652

Appendix E: Classification Consistency and Accuracy Estimates for All Proficiency Levels for Examinee Subgroups

Table E1. Classification Consistency and Accuracy Rates for All Cut Scores and Examinee Subgroups: Reading

Grade/Subgroup	Classification Consistency		Classification Accuracy		
	Consistency	Kappa	Accuracy	False Positive Errors	False Negative Errors
Grade 3					
Males	0.7856	0.6925	0.8517	0.0651	0.0832
Females	0.7613	0.6547	0.8358	0.0725	0.0917
Asian/Pacific Islander	0.7379	0.5736	0.8161	0.0863	0.0976
African American	0.7780	0.6774	0.8478	0.0681	0.0841
Hispanic	0.7733	0.6704	0.8425	0.0644	0.0931
White	0.7382	0.5490	0.8127	0.0777	0.1096
Grade 4					
Males	0.7701	0.6666	0.8366	0.0681	0.0953
Females	0.7676	0.6563	0.8346	0.0710	0.0944
Asian/Pacific Islander	0.8119	0.6958	0.8746	0.0590	0.0665
African American	0.7700	0.6582	0.8370	0.0682	0.0948
Hispanic	0.7712	0.6610	0.8376	0.0691	0.0933
White	0.7407	0.5677	0.8081	0.0878	0.1041
Grade 5					
Males	0.7646	0.6546	0.8320	0.0753	0.0927
Females	0.7596	0.6347	0.8308	0.0784	0.0908
Asian/Pacific Islander	0.7644	0.6038	0.8274	0.0782	0.0944
African American	0.7643	0.6435	0.8331	0.0755	0.0914
Hispanic	0.7539	0.6265	0.8275	0.0781	0.0944
White	0.7519	0.5832	0.8204	0.0900	0.0896
Grade 6					
Males	0.7680	0.6541	0.8405	0.0723	0.0871
Females	0.7522	0.6169	0.8283	0.0836	0.0881
Asian/Pacific Islander	0.7366	0.5937	0.8087	0.0838	0.1075
African American	0.7632	0.6336	0.8374	0.0769	0.0857
Hispanic	0.7574	0.6231	0.8300	0.0771	0.0929
White	0.7307	0.5514	0.8092	0.0919	0.0989
Grade 7					
Males	0.7533	0.6414	0.8231	0.0848	0.0921
Females	0.7496	0.6310	0.8201	0.0889	0.0910
Asian/Pacific Islander	0.7114	0.5463	0.7903	0.0933	0.1164
African American	0.7494	0.6275	0.8202	0.0887	0.0911
Hispanic	0.7523	0.6406	0.8234	0.0743	0.1023
White	0.7929	0.6204	0.8482	0.0789	0.0729

Grade 8					
Males	0.7493	0.6438	0.8236	0.0786	0.0978
Females	0.7500	0.6360	0.8253	0.0769	0.0978
Asian/Pacific Islander	0.7781	0.6720	0.8450	0.0850	0.0700
African American	0.7469	0.6312	0.8224	0.0780	0.0996
Hispanic	0.7434	0.6310	0.8194	0.0779	0.1027
White	0.8183	0.6166	0.8759	0.0725	0.0516
Grade 10					
Males	0.7746	0.6696	0.8418	0.0764	0.0818
Females	0.7589	0.6374	0.8287	0.0855	0.0857
Asian/Pacific Islander	0.7360	0.6197	0.8022	0.1033	0.0945
African American	0.7668	0.6494	0.8358	0.0807	0.0834
Hispanic	0.7623	0.6420	0.8313	0.0786	0.0901
White	0.7697	0.6300	0.8291	0.0907	0.0803

Table E2. Classification Consistency and Accuracy Rates for All Cut Scores and Examinee Subgroups: Mathematics

Grade/Subgroup	Classification Consistency		Classification Accuracy		
	Consistency	Kappa	Accuracy	False Positive Errors	False Negative Errors
Grade 3					
Males	0.7715	0.6758	0.8410	0.0740	0.0850
Females	0.7508	0.6450	0.8256	0.0734	0.1011
Asian/Pacific Islander	0.7193	0.5753	0.7962	0.0814	0.1224
African American	0.7681	0.6622	0.8390	0.0715	0.0896
Hispanic	0.7416	0.6243	0.8180	0.0741	0.1080
White	0.7351	0.5716	0.8110	0.0943	0.0947
Grade 4					
Males	0.7637	0.6667	0.8294	0.0794	0.0912
Females	0.7597	0.6600	0.8277	0.0796	0.0927
Asian/Pacific Islander	0.7655	0.6191	0.8276	0.0750	0.0974
African American	0.7604	0.6540	0.8279	0.0796	0.0926
Hispanic	0.7500	0.6429	0.8185	0.0809	0.1005
White	0.7958	0.6688	0.8543	0.0768	0.0688
Grade 5					
Males	0.7512	0.6493	0.8221	0.0831	0.0948
Females	0.7465	0.6347	0.8197	0.0891	0.0912
Asian/Pacific Islander	0.7730	0.6363	0.8434	0.0880	0.0686
African American	0.7483	0.6337	0.8208	0.0838	0.0954
Hispanic	0.7490	0.6323	0.8197	0.0894	0.0909
White	0.7524	0.5854	0.8194	0.1082	0.0723
Grade 6					
Males	0.7651	0.6703	0.8300	0.0843	0.0857
Females	0.7580	0.6558	0.8248	0.0859	0.0893
Asian/Pacific Islander	0.8319	0.7434	0.8884	0.0725	0.0392
African American	0.7583	0.6515	0.8249	0.0854	0.0897
Hispanic	0.7533	0.6473	0.8196	0.0900	0.0904
White	0.8056	0.6779	0.8624	0.0741	0.0635

Grade/Subgroup	Classification Consistency		Classification Accuracy		
	Consistency	Kappa	Accuracy	False Positive Errors	False Negative Errors
Grade 7					
Males	0.7627	0.6662	0.8359	0.0774	0.0868
Females	0.7609	0.6507	0.8367	0.0780	0.0853
Asian/Pacific Islander	0.8187	0.6979	0.8776	0.0643	0.0581
African American	0.7570	0.6484	0.8333	0.0787	0.0879
Hispanic	0.7676	0.6598	0.8400	0.0765	0.0835
White	0.8178	0.6755	0.8701	0.0642	0.0657
Grade 8					
Males	0.7111	0.5870	0.7923	0.1048	0.1029
Females	0.7141	0.5805	0.7957	0.1011	0.1032
Asian/Pacific Islander	0.8227	0.7056	0.8789	0.0753	0.0459
African American	0.7064	0.5702	0.7892	0.1040	0.1068
Hispanic	0.7229	0.5904	0.8057	0.1035	0.0908
White	0.7813	0.6300	0.8384	0.0876	0.0741
Grade 10					
Males	0.7592	0.6580	0.8294	0.0791	0.0914
Females	0.7492	0.6357	0.8229	0.0932	0.0839
Asian/Pacific Islander	0.8330	0.7458	0.8888	0.0421	0.0691
African American	0.7515	0.6382	0.8245	0.0872	0.0883
Hispanic	0.7416	0.6205	0.8145	0.0918	0.0937
White	0.8056	0.6934	0.8601	0.0887	0.0512

Table E3. Classification Consistency and Accuracy Rates for All Cut Scores and Examinee Subgroups: Science/Biology

Grade/Subgroup	Classification Consistency		Classification Accuracy		
	Consistency	Kappa	Accuracy	False Positive Errors	False Negative Errors
Grade 5					
Males	0.7127	0.5875	0.7957	0.1047	0.0996
Females	0.7090	0.5711	0.7930	0.1051	0.1019
Asian/Pacific Islander	0.7906	0.6747	0.8500	0.0843	0.0657
African American	0.7061	0.5588	0.7915	0.1049	0.1035
Hispanic	0.7013	0.5536	0.7862	0.1073	0.1064
White	0.7681	0.6072	0.8322	0.1017	0.0661
Grade 8					
Males	0.7131	0.5768	0.7903	0.0944	0.1154
Females	0.6867	0.5438	0.7686	0.1095	0.1219
Asian/Pacific Islander	0.7264	0.6171	0.7935	0.1177	0.0888
African American	0.6971	0.5467	0.7771	0.1015	0.1213
Hispanic	0.6867	0.5447	0.7675	0.1098	0.1228
White	0.7842	0.6434	0.8564	0.0811	0.0624
High School					
Males	0.6691	0.5091	0.7474	0.1159	0.1367
Females	0.6698	0.5040	0.7473	0.1213	0.1314
Asian/Pacific Islander	0.7619	0.5486	0.8204	0.0824	0.0972

Grade/Subgroup	Classification Consistency		Classification Accuracy		
	Consistency	Kappa	Accuracy	False Positive Errors	False Negative Errors
African American	0.6590	0.4897	0.7388	0.1222	0.1390
Hispanic	0.7024	0.5373	0.7749	0.1132	0.1120
White	0.7870	0.6067	0.8431	0.0883	0.0686

Table E4. Classification Consistency and Accuracy Rates for Basic Scores and Examinee Subgroups: Reading

Grade/Subgroup	Classification Consistency		Classification Accuracy		
	Consistency	Kappa	Accuracy	False Positive Errors	False Negative Errors
Grade 3					
Males	0.9334	0.8360	0.9532	0.0245	0.0222
Females	0.9434	0.7984	0.9606	0.0220	0.0174
Asian/Pacific Islander	0.9791	0.7841	0.9851	0.0097	0.0052
African American	0.9319	0.8195	0.9525	0.0258	0.0216
Hispanic	0.9340	0.8160	0.9532	0.0235	0.0233
White	0.9957	0.8928	0.9971	0.0019	0.0010
Grade 4					
Males	0.9289	0.7938	0.9505	0.0244	0.0252
Females	0.9393	0.7586	0.9576	0.0235	0.0189
Asian/Pacific Islander	0.9828	0.8020	0.9892	0.0082	0.0026
African American	0.9264	0.7737	0.9486	0.0266	0.0249
Hispanic	0.9396	0.7879	0.9580	0.0225	0.0195
White	0.9918	0.8697	0.9947	0.0036	0.0017
Grade 5					
Males	0.9414	0.8100	0.9577	0.0214	0.0209
Females	0.9545	0.7649	0.9672	0.0168	0.0160
Asian/Pacific Islander	0.9982	-0.0008	0.9991	0.0009	0.0000
African American	0.9412	0.7906	0.9575	0.0217	0.0208
Hispanic	0.9609	0.8054	0.9721	0.0132	0.0147
White	0.9939	0.8859	0.9959	0.0026	0.0016
Grade 6					
Males	0.9258	0.7650	0.9501	0.0239	0.0260
Females	0.9406	0.7067	0.9587	0.0241	0.0173
Asian/Pacific Islander	0.9869	0.8536	0.9894	0.0103	0.0003
African American	0.9274	0.7410	0.9505	0.0259	0.0236
Hispanic	0.9423	0.7665	0.9598	0.0234	0.0168
White	0.9809	0.6419	0.9867	0.0054	0.0079
Grade 7					
Males	0.9272	0.7266	0.9491	0.0297	0.0212
Females	0.9556	0.7022	0.9697	0.0177	0.0126
Asian/Pacific Islander	0.9928	0.8116	0.9961	0.0013	0.0025
African American	0.9364	0.7114	0.9557	0.0265	0.0178
Hispanic	0.9506	0.7672	0.9669	0.0147	0.0184
White	0.9978	0.9453	0.9985	0.0014	0.0000

Grade/Subgroup	Classification Consistency		Classification Accuracy		
	Consistency	Kappa	Accuracy	False Positive Errors	False Negative Errors
Grade 8					
Males	0.9169	0.7209	0.9428	0.0315	0.0257
Females	0.9434	0.6845	0.9618	0.0238	0.0144
Asian/Pacific Islander	0.9766	0.6681	0.9843	0.0100	0.0057
African American	0.9263	0.7119	0.9497	0.0288	0.0215
Hispanic	0.9335	0.6841	0.9540	0.0300	0.0160
White	0.9880	0.6808	0.9920	0.0051	0.0030
Grade 10					
Males	0.9273	0.7835	0.9507	0.0244	0.0250
Females	0.9420	0.7479	0.9605	0.0209	0.0186
Asian/Pacific Islander	0.9736	0.7921	0.9836	0.0078	0.0086
African American	0.9316	0.7689	0.9533	0.0240	0.0227
Hispanic	0.9362	0.7598	0.9570	0.0227	0.0203
White	0.9901	0.8793	0.9930	0.0002	0.0067

Table E5. Classification Consistency and Accuracy Rates for Basic Scores and Examinee Subgroups: Mathematics

Grade/Subgroup	Classification Consistency		Classification Accuracy		
	Consistency	Kappa	Accuracy	False Positive Errors	False Negative Errors
Grade 3					
Males	0.9204	0.7908	0.9438	0.0310	0.0252
Females	0.9228	0.7618	0.9453	0.0277	0.0270
Asian/Pacific Islander	0.9806	0.6002	0.9851	0.0088	0.0061
African American	0.9128	0.7728	0.9383	0.0326	0.0291
Hispanic	0.9205	0.7691	0.9440	0.0310	0.0250
White	0.9917	0.8123	0.9940	0.0019	0.0040
Grade 4					
Males	0.9224	0.7656	0.9442	0.0305	0.0253
Females	0.9157	0.7259	0.9393	0.0333	0.0274
Asian/Pacific Islander	0.9896	0.8587	0.9929	0.0042	0.0029
African American	0.9090	0.7387	0.9348	0.0357	0.0296
Hispanic	0.9329	0.7577	0.9503	0.0279	0.0218
White	0.9842	0.8055	0.9892	0.0061	0.0048
Grade 5					
Males	0.9227	0.7527	0.9464	0.0287	0.0249
Females	0.9270	0.7120	0.9497	0.0298	0.0205
Asian/Pacific Islander	0.9886	0.6621	0.9937	0.0040	0.0024
African American	0.9160	0.7321	0.9420	0.0318	0.0262
Hispanic	0.9412	0.7177	0.9586	0.0281	0.0133
White	0.9889	0.8159	0.9925	0.0045	0.0029
Grade 6					
Males	0.9108	0.7361	0.9362	0.0340	0.0298
Females	0.9194	0.6971	0.9412	0.0304	0.0284

Grade/Subgroup	Classification Consistency		Classification Accuracy		
	Consistency	Kappa	Accuracy	False Positive Errors	False Negative Errors
Asian/Pacific Islander	0.9666	0.7831	0.9804	0.0080	0.0116
African American	0.9088	0.7163	0.9342	0.0346	0.0312
Hispanic	0.9229	0.7260	0.9427	0.0303	0.0271
White	0.9768	0.7545	0.9843	0.0085	0.0072
Grade 7					
Males	0.9081	0.7270	0.9363	0.0318	0.0318
Females	0.9230	0.7059	0.9476	0.0283	0.0240
Asian/Pacific Islander	0.9809	0.7756	0.9858	0.0044	0.0098
African American	0.9090	0.7129	0.9375	0.0325	0.0301
Hispanic	0.9226	0.7297	0.9471	0.0261	0.0268
White	0.9918	0.8580	0.9952	0.0027	0.0021
Grade 8					
Males	0.8732	0.6097	0.9077	0.0490	0.0433
Females	0.8904	0.6030	0.9219	0.0438	0.0343
Asian/Pacific Islander	0.9761	0.4656	0.9851	0.0112	0.0037
African American	0.8752	0.6040	0.9099	0.0487	0.0413
Hispanic	0.8878	0.5866	0.9202	0.0465	0.0334
White	0.9721	0.6516	0.9795	0.0113	0.0093
Grade 10					
Males	0.8799	0.7052	0.9171	0.0386	0.0443
Females	0.8868	0.6718	0.9226	0.0430	0.0344
Asian/Pacific Islander	0.9851	0.8038	0.9918	0.0066	0.0016
African American	0.8767	0.6856	0.9154	0.0431	0.0416
Hispanic	0.8911	0.6628	0.9241	0.0437	0.0322
White	0.9830	0.8233	0.9883	0.0064	0.0053

Table E6. Classification Consistency and Accuracy Rates for Basic Scores and Examinee Subgroups: Science/Biology

Grade/Subgroup	Classification Consistency		Classification Accuracy		
	Consistency	Kappa	Accuracy	False Positive Errors	False Negative Errors
Grade 5					
Males	0.8666	0.6259	0.9064	0.0525	0.0410
Females	0.8744	0.5958	0.9124	0.0528	0.0348
Asian/Pacific Islander	0.9800	0.6671	0.9829	0.0163	0.0009
African American	0.8564	0.6052	0.8996	0.0579	0.0425
Hispanic	0.8880	0.5988	0.9223	0.0452	0.0325
White	0.9841	0.5424	0.9881	0.0098	0.0021
Grade 8					
Males	0.8420	0.6780	0.8876	0.0528	0.0596
Females	0.8342	0.6512	0.8818	0.0609	0.0573
Asian/Pacific Islander	0.9128	0.6886	0.9347	0.0387	0.0266
African American	0.8306	0.6554	0.8795	0.0590	0.0615
Hispanic	0.8342	0.6478	0.8804	0.0612	0.0585
White	0.9729	0.7773	0.9826	0.0059	0.0115

Grade/Subgroup	Classification Consistency		Classification Accuracy		
	Consistency	Kappa	Accuracy	False Positive Errors	False Negative Errors
High School					
Males	0.8116	0.5759	0.8645	0.0719	0.0636
Females	0.8245	0.5705	0.8742	0.0689	0.0569
Asian/Pacific Islander	0.9426	0.4932	0.9642	0.0288	0.0070
African American	0.8074	0.5637	0.8616	0.0745	0.0639
Hispanic	0.8395	0.5893	0.8847	0.0651	0.0502
White	0.9665	0.4839	0.9761	0.0181	0.0058

Table E7. Classification Consistency and Accuracy Rates for Proficient Scores and Examinee Subgroups: Reading

Grade/Subgroup	Classification Consistency		Classification Accuracy		
	Consistency	Kappa	Accuracy	False Positive Errors	False Negative Errors
Grade 3					
Males	0.8975	0.7774	0.9304	0.0312	0.0384
Females	0.8878	0.7753	0.9248	0.0370	0.0382
Asian/Pacific Islander	0.8845	0.7107	0.9195	0.0443	0.0361
African American	0.8895	0.7629	0.9257	0.0352	0.0391
Hispanic	0.8891	0.7613	0.9241	0.0313	0.0446
White	0.9260	0.6682	0.9497	0.0260	0.0243
Grade 4					
Males	0.8912	0.7775	0.9217	0.0323	0.0460
Females	0.8906	0.7812	0.9218	0.0329	0.0454
Asian/Pacific Islander	0.9214	0.7778	0.9481	0.0273	0.0246
African American	0.8839	0.7614	0.9165	0.0343	0.0492
Hispanic	0.8901	0.7790	0.9211	0.0323	0.0466
White	0.9564	0.7960	0.9697	0.0167	0.0136
Grade 5					
Males	0.8724	0.7363	0.9100	0.0438	0.0462
Females	0.8666	0.7332	0.9079	0.0474	0.0448
Asian/Pacific Islander	0.9166	0.7148	0.9437	0.0351	0.0212
African American	0.8654	0.7225	0.9057	0.0463	0.0480
Hispanic	0.8537	0.7070	0.8997	0.0510	0.0493
White	0.9360	0.7017	0.9568	0.0312	0.0120
Grade 6					
Males	0.8846	0.7532	0.9197	0.0405	0.0398
Females	0.8661	0.7300	0.9075	0.0481	0.0444
Asian/Pacific Islander	0.8972	0.7561	0.9257	0.0399	0.0344
African American	0.8724	0.7278	0.9118	0.0449	0.0433
Hispanic	0.8532	0.6956	0.8961	0.0505	0.0534
White	0.9518	0.7713	0.9671	0.0239	0.0090
Grade 7					
Males	0.8900	0.7743	0.9196	0.0384	0.0420
Females	0.8748	0.7484	0.9079	0.0491	0.0430
Asian/Pacific Islander	0.8870	0.6290	0.9146	0.0427	0.0428

Grade/Subgroup	Classification Consistency		Classification Accuracy		
	Consistency	Kappa	Accuracy	False Positive Errors	False Negative Errors
African American	0.8767	0.7505	0.9094	0.0459	0.0447
Hispanic	0.8952	0.7902	0.9239	0.0378	0.0383
White	0.9596	0.7503	0.9721	0.0173	0.0106
Grade 8					
Males	0.8951	0.7880	0.9232	0.0343	0.0425
Females	0.8872	0.7731	0.9175	0.0332	0.0493
Asian/Pacific Islander	0.9195	0.7956	0.9420	0.0346	0.0234
African American	0.8860	0.7708	0.9165	0.0354	0.0482
Hispanic	0.8953	0.7899	0.9234	0.0311	0.0455
White	0.9847	0.8994	0.9900	0.0039	0.0061
Grade 10					
Males	0.8983	0.7737	0.9288	0.0392	0.0320
Females	0.8783	0.7484	0.9138	0.0491	0.0372
Asian/Pacific Islander	0.8706	0.7347	0.9054	0.0684	0.0262
African American	0.8851	0.7488	0.9191	0.0450	0.0359
Hispanic	0.8887	0.7618	0.9209	0.0432	0.0359
White	0.9522	0.8239	0.9679	0.0229	0.0092

Table E8. Classification Consistency and Accuracy Rates for Proficient Scores and Examinee Subgroups: Mathematics

Grade/Subgroup	Classification Consistency		Classification Accuracy		
	Consistency	Kappa	Accuracy	False Positive Errors	False Negative Errors
Grade 3					
Males	0.9067	0.7972	0.9371	0.0286	0.0343
Females	0.8914	0.7741	0.9266	0.0322	0.0413
Asian/Pacific Islander	0.8956	0.7369	0.9285	0.0390	0.0326
African American	0.8989	0.7691	0.9320	0.0302	0.0377
Hispanic	0.8825	0.7448	0.9193	0.0312	0.0495
White	0.9218	0.6989	0.9470	0.0304	0.0227
Grade 4					
Males	0.8978	0.7943	0.9264	0.0332	0.0404
Females	0.8974	0.7937	0.9264	0.0318	0.0418
Asian/Pacific Islander	0.9295	0.7026	0.9492	0.0200	0.0308
African American	0.8934	0.7791	0.9233	0.0339	0.0428
Hispanic	0.8840	0.7681	0.9167	0.0349	0.0484
White	0.9565	0.8180	0.9697	0.0174	0.0129
Grade 5					
Males	0.8935	0.7840	0.9223	0.0360	0.0417
Females	0.8815	0.7616	0.9140	0.0422	0.0438
Asian/Pacific Islander	0.9246	0.7217	0.9487	0.0362	0.0151
African American	0.8829	0.7554	0.9148	0.0392	0.0459
Hispanic	0.8812	0.7615	0.9127	0.0447	0.0426
White	0.9447	0.7334	0.9606	0.0292	0.0102

Grade/Subgroup	Classification Consistency		Classification Accuracy		
	Consistency	Kappa	Accuracy	False Positive Errors	False Negative Errors
Grade 6					
Males	0.9074	0.8076	0.9309	0.0345	0.0347
Females	0.8996	0.7982	0.9267	0.0375	0.0358
Asian/Pacific Islander	0.9678	0.9148	0.9771	0.0153	0.0076
African American	0.8995	0.7890	0.9258	0.0368	0.0374
Hispanic	0.8962	0.7920	0.9233	0.0405	0.0363
White	0.9577	0.8408	0.9682	0.0205	0.0113
Grade 7					
Males	0.9088	0.8175	0.9372	0.0299	0.0329
Females	0.8936	0.7850	0.9270	0.0350	0.0380
Asian/Pacific Islander	0.9705	0.8532	0.9814	0.0160	0.0026
African American	0.8957	0.7912	0.9284	0.0337	0.0379
Hispanic	0.9067	0.8097	0.9356	0.0351	0.0293
White	0.9743	0.8240	0.9811	0.0105	0.0084
Grade 8					
Males	0.8767	0.7529	0.9150	0.0415	0.0436
Females	0.8689	0.7370	0.9096	0.0411	0.0493
Asian/Pacific Islander	0.9364	0.7145	0.9587	0.0229	0.0184
African American	0.8657	0.7307	0.9070	0.0428	0.0502
Hispanic	0.8854	0.7688	0.9236	0.0425	0.0338
White	0.9726	0.8648	0.9822	0.0093	0.0085
Grade 10					
Males	0.9097	0.8035	0.9341	0.0330	0.0329
Females	0.8940	0.7764	0.9233	0.0404	0.0363
Asian/Pacific Islander	0.9197	0.7907	0.9462	0.0185	0.0353
African American	0.9010	0.7801	0.9280	0.0375	0.0344
Hispanic	0.8847	0.7636	0.9167	0.0378	0.0455
White	0.9450	0.8096	0.9591	0.0272	0.0138

Table E9. Classification Consistency and Accuracy Rates for Proficient Scores and Examinee Subgroups: Science/Biology

Grade/Subgroup	Classification Consistency		Classification Accuracy		
	Consistency	Kappa	Accuracy	False Positive Errors	False Negative Errors
Grade 5					
Males	0.8834	0.7523	0.9160	0.0399	0.0441
Females	0.8698	0.7277	0.9062	0.0417	0.0520
Asian/Pacific Islander	0.9015	0.7047	0.9311	0.0387	0.0303
African American	0.8732	0.7129	0.9089	0.0411	0.0500
Hispanic	0.8528	0.7007	0.8914	0.0504	0.0582
White	0.9519	0.7468	0.9660	0.0223	0.0118
Grade 8					

Grade/Subgroup	Classification Consistency		Classification Accuracy		
	Consistency	Kappa	Accuracy	False Positive Errors	False Negative Errors
Males	0.8853	0.7462	0.9182	0.0352	0.0466
Females	0.8640	0.7048	0.9023	0.0422	0.0555
Asian/Pacific Islander	0.8709	0.7145	0.9031	0.0493	0.0476
African American	0.8725	0.7071	0.9088	0.0388	0.0524
Hispanic	0.8659	0.7131	0.9029	0.0425	0.0546
White	0.9488	0.7667	0.9664	0.0208	0.0129
High School					
Males	0.8417	0.6722	0.8881	0.0427	0.0691
Females	0.8289	0.6550	0.8778	0.0512	0.0710
Asian/Pacific Islander	0.8548	0.5675	0.8886	0.0470	0.0643
African American	0.8289	0.6432	0.8783	0.0484	0.0733
Hispanic	0.8473	0.6946	0.8933	0.0481	0.0586
White	0.9426	0.6859	0.9600	0.0212	0.0188

Table E10. Classification Consistency and Accuracy Rates for Advanced Scores and Examinee Subgroups: Reading

Grade/Subgroup	Classification Consistency		Classification Accuracy		
	Consistency	Kappa	Accuracy	False Positive Errors	False Negative Errors
Grade 3					
Males	0.9546	0.5735	0.9681	0.0094	0.0225
Females	0.9299	0.5945	0.9504	0.0135	0.0361
Asian/Pacific Islander	0.8741	0.5646	0.9115	0.0322	0.0563
African American	0.9564	0.5298	0.9696	0.0071	0.0233
Hispanic	0.9500	0.5514	0.9652	0.0097	0.0251
White	0.8165	0.5777	0.8659	0.0498	0.0843
Grade 4					
Males	0.9498	0.5608	0.9645	0.0114	0.0241
Females	0.9375	0.6116	0.9552	0.0146	0.0302
Asian/Pacific Islander	0.9075	0.7382	0.9372	0.0235	0.0393
African American	0.9595	0.5014	0.9719	0.0073	0.0207
Hispanic	0.9412	0.5220	0.9585	0.0143	0.0272
White	0.7925	0.5622	0.8437	0.0675	0.0889
Grade 5					
Males	0.9508	0.6014	0.9643	0.0101	0.0256
Females	0.9385	0.6230	0.9557	0.0142	0.0300
Asian/Pacific Islander	0.8496	0.6245	0.8846	0.0422	0.0732
African American	0.9577	0.5358	0.9699	0.0075	0.0226
Hispanic	0.9392	0.6130	0.9557	0.0139	0.0305
White	0.8220	0.6177	0.8677	0.0562	0.0761
Grade 6					
Males	0.9577	0.5765	0.9708	0.0079	0.0213
Females	0.9456	0.5667	0.9621	0.0115	0.0264

Grade/Subgroup	Classification Consistency		Classification Accuracy		
	Consistency	Kappa	Accuracy	False Positive Errors	False Negative Errors
Asian/Pacific Islander	0.8525	0.5492	0.8936	0.0336	0.0728
African American	0.9635	0.4496	0.9751	0.0061	0.0188
Hispanic	0.9619	0.5413	0.9740	0.0033	0.0227
White	0.7980	0.5798	0.8553	0.0627	0.0820
Grade 7					
Males	0.9358	0.6655	0.9545	0.0167	0.0288
Females	0.9188	0.6940	0.9425	0.0221	0.0354
Asian/Pacific Islander	0.8315	0.6321	0.8796	0.0493	0.0711
African American	0.9359	0.6548	0.9551	0.0164	0.0285
Hispanic	0.9061	0.5929	0.9326	0.0218	0.0456
White	0.8353	0.6604	0.8776	0.0602	0.0622
Grade 8					
Males	0.9370	0.6797	0.9577	0.0128	0.0295
Females	0.9193	0.6869	0.9461	0.0199	0.0340
Asian/Pacific Islander	0.8819	0.7512	0.9188	0.0404	0.0408
African American	0.9344	0.6352	0.9563	0.0138	0.0299
Hispanic	0.9143	0.6619	0.9420	0.0168	0.0412
White	0.8455	0.6429	0.8940	0.0636	0.0425
Grade 10					
Males	0.9489	0.6615	0.9623	0.0128	0.0249
Females	0.9384	0.6423	0.9545	0.0156	0.0299
Asian/Pacific Islander	0.8915	0.6863	0.9132	0.0271	0.0597
African American	0.9499	0.6150	0.9634	0.0118	0.0248
Hispanic	0.9373	0.5940	0.9534	0.0127	0.0339
White	0.8272	0.6544	0.8681	0.0676	0.0643

Table E11. Classification Consistency and Accuracy Rates for Advanced Scores and Examinee Subgroups: Mathematics

Grade/Subgroup	Classification Consistency		Classification Accuracy		
	Consistency	Kappa	Accuracy	False Positive Errors	False Negative Errors
Grade 3					
Males	0.9439	0.6503	0.9601	0.0144	0.0255
Females	0.9362	0.6461	0.9537	0.0135	0.0328
Asian/Pacific Islander	0.8429	0.6133	0.8826	0.0337	0.0838
African American	0.9559	0.6224	0.9686	0.0086	0.0228
Hispanic	0.9382	0.5425	0.9547	0.0119	0.0334
White	0.8211	0.6160	0.8700	0.0620	0.0680
Grade 4					
Males	0.9427	0.6912	0.9588	0.0157	0.0256
Females	0.9459	0.7237	0.9620	0.0144	0.0235
Asian/Pacific Islander	0.8463	0.6819	0.8855	0.0507	0.0637
African American	0.9571	0.6455	0.9698	0.0100	0.0202

Grade/Subgroup	Classification Consistency		Classification Accuracy		
	Consistency	Kappa	Accuracy	False Positive Errors	False Negative Errors
Hispanic	0.9325	0.6694	0.9516	0.0182	0.0303
White	0.8549	0.7077	0.8954	0.0534	0.0512
Grade 5					
Males	0.9350	0.6787	0.9534	0.0183	0.0283
Females	0.9379	0.7003	0.9560	0.0171	0.0269
Asian/Pacific Islander	0.8598	0.7018	0.9010	0.0479	0.0511
African American	0.9493	0.6433	0.9640	0.0127	0.0232
Hispanic	0.9265	0.6649	0.9484	0.0166	0.0350
White	0.8189	0.6377	0.8663	0.0745	0.0592
Grade 6					
Males	0.9469	0.7419	0.9630	0.0158	0.0213
Females	0.9389	0.7284	0.9568	0.0180	0.0251
Asian/Pacific Islander	0.8975	0.7951	0.9309	0.0492	0.0199
African American	0.9501	0.6867	0.9649	0.0140	0.0210
Hispanic	0.9342	0.6644	0.9537	0.0193	0.0270
White	0.8711	0.7407	0.9099	0.0451	0.0450
Grade 7					
Males	0.9456	0.7302	0.9623	0.0156	0.0221
Females	0.9441	0.7187	0.9621	0.0146	0.0233
Asian/Pacific Islander	0.8672	0.7299	0.9103	0.0439	0.0458
African American	0.9521	0.6851	0.9674	0.0126	0.0200
Hispanic	0.9380	0.6931	0.9573	0.0152	0.0275
White	0.8516	0.7017	0.8938	0.0510	0.0552
Grade 8					
Males	0.9578	0.7466	0.9695	0.0143	0.0161
Females	0.9510	0.7124	0.9641	0.0162	0.0197
Asian/Pacific Islander	0.9087	0.8067	0.9351	0.0411	0.0238
African American	0.9616	0.6977	0.9721	0.0125	0.0154
Hispanic	0.9466	0.6923	0.9618	0.0144	0.0237
White	0.8355	0.6709	0.8767	0.0670	0.0563
Grade 10					
Males	0.9692	0.7672	0.9782	0.0076	0.0142
Females	0.9677	0.7248	0.9770	0.0098	0.0131
Asian/Pacific Islander	0.9277	0.8182	0.9508	0.0170	0.0322
African American	0.9732	0.6983	0.9811	0.0066	0.0123
Hispanic	0.9653	0.6717	0.9737	0.0103	0.0160
White	0.8775	0.7543	0.9128	0.0551	0.0321

Table E12. Classification Consistency and Accuracy Rates for Advanced Scores and Examinee Subgroups: Science/Biology

Grade/Subgroup	Classification Consistency		Classification Accuracy		
	Consistency	Kappa	Accuracy	False Positive Errors	False Negative Errors
Grade 5					
Males	0.9612	0.7183	0.9733	0.0122	0.0145
Females	0.9633	0.6879	0.9744	0.0105	0.0150
Asian/Pacific Islander	0.9088	0.7897	0.9360	0.0294	0.0346
African American	0.9749	0.5728	0.9830	0.0060	0.0110
Hispanic	0.9591	0.6314	0.9725	0.0117	0.0158
White	0.8317	0.6627	0.8781	0.0697	0.0522
Grade 8					
Males	0.9710	0.7338	0.9809	0.0075	0.0116
Females	0.9716	0.6985	0.9809	0.0078	0.0114
Asian/Pacific Islander	0.9306	0.8154	0.9528	0.0310	0.0162
African American	0.9775	0.6665	0.9850	0.0050	0.0099
Hispanic	0.9701	0.6853	0.9804	0.0075	0.0122
White	0.8586	0.7026	0.9066	0.0549	0.0385
High School					
Males	0.9826	0.6298	0.9881	0.0038	0.0081
Females	0.9835	0.6616	0.9887	0.0041	0.0072
Asian/Pacific Islander	0.9462	0.7659	0.9646	0.0092	0.0262
African American	0.9880	0.5088	0.9919	0.0021	0.0060
Hispanic	0.9860	0.6339	0.9910	0.0027	0.0064
White	0.8698	0.6854	0.9054	0.0501	0.0445

Appendix F: Items Flagged for DIF Using Mantel-Haenszel Procedures

Table F1. Focal and Reference Groups for All Tables

Comparison	Reference	Focal
Gender	Male	Female
Race/Ethnicity	African American	Asian/Pacific Islander, Hispanic, White

Note. See the sub-section *Differential Item Functioning* for the rationales for including these subgroups.

Table F2. Items Flagged for DIF Using Mantel-Haenszel: Reading

Reading Grade 3							
Operational Item Sequence Number	**Item Type**	**DIF Comparison**	**Focal Group**	**SMD**	**Reference Group N**	**Focal Group N**	**DIF Level**
16	MC	ETHNIC	Hispanic	0.034	3,801	592	B
20	CR	ETHNIC	Hispanic	0.189	3,817	592	B
12	MC	ETHNIC	White	0.033	2,779	404	B
17	MC	ETHNIC	White	0.040	2,967	404	B
18	MC	ETHNIC	White	0.039	3,001	404	B
24	MC	ETHNIC	White	0.048	2,986	404	B
34	MC	ETHNIC	White	0.053	2,925	404	B
36	MC	ETHNIC	White	0.061	2,975	404	B
28	MC	ETHNIC	Hispanic	-0.099	3,819	592	-B
35	MC	ETHNIC	Hispanic	-0.116	3,813	591	-B
40	MC	ETHNIC	White	-0.065	2,947	404	-B
2	MC	ETHNIC	White	0.157	2,882	404	C
3	MC	ETHNIC	White	0.042	2,888	404	C
6	MC	ETHNIC	White	0.066	2,922	404	C
7	MC	ETHNIC	White	0.028	2,848	404	C
20	CR	ETHNIC	White	0.277	2,896	404	C
22	MC	ETHNIC	White	0.152	2,961	404	C
45	MC	ETHNIC	White	0.110	2,961	404	C
42	MC	ETHNIC	White	-0.033	2,908	404	-C
48	MC	ETHNIC	White	-0.037	3,009	404	-C

Note. DIF is Differential Item Functioning and SMD is Standardized Mean Difference. DIF levels A, B, and C are explained in the sub-section *Differential Item Functioning*.

Reading Grade 4							
Operational Item Sequence Number	**Item Type**	**DIF Comparison**	**Focal Group**	**SMD**	**Reference Group N**	**Focal Group N**	**DIF Level**
10	MC	ETHNIC	White	0.031	3,253	357	B
11	MC	ETHNIC	White	0.116	3,382	360	B
16	MC	ETHNIC	White	0.024	3,338	357	B
26	MC	ETHNIC	White	0.080	3,338	357	B
28	MC	ETHNIC	White	0.078	3,313	357	B
36	MC	ETHNIC	White	0.075	3,157	357	B
44	MC	ETHNIC	White	0.103	3,166	357	B
9	CR	GENDER	Female	0.178	2,413	2,382	B
41	MC	ETHNIC	Hispanic	-0.059	3,756	584	-B
14	MC	GENDER	Female	-0.083	2,411	2,384	-B
7	MC	ETHNIC	White	0.077	3,200	357	C
14	MC	ETHNIC	White	0.122	3,272	356	C
17	MC	ETHNIC	White	0.059	3,260	356	C
30	MC	ETHNIC	White	0.171	3,185	357	C
33	MC	ETHNIC	White	0.074	3,222	357	C
45	MC	ETHNIC	White	0.050	3,312	357	C

Note. DIF is Differential Item Functioning and SMD is Standardized Mean Difference. DIF levels A, B, and C are explained in the sub-section Differential Item Functioning.

Reading Grade 5							
Operational Item Sequence Number	**Item Type**	**DIF Comparison**	**Focal Group**	**SMD**	**Reference Group N**	**Focal Group N**	**DIF Level**
46	MC	ETHNIC	Hispanic	0.039	3,613	502	B
7	CR	GENDER	Female	0.189	2,291	2,202	B
10	MC	GENDER	Female	0.058	2,291	2,196	B
19	MC	ETHNIC	Hispanic	-0.052	3,618	502	-B
27	MC	GENDER	Female	-0.041	2,291	2,194	-B
9	MC	ETHNIC	White	0.126	3,141	301	C
14	MC	ETHNIC	White	0.065	2,975	301	C
15	MC	ETHNIC	White	0.069	3,054	301	C
25	MC	ETHNIC	White	0.094	3,060	301	C
28	MC	ETHNIC	White	0.059	3,016	301	C
48	MC	ETHNIC	White	0.133	2,873	301	C

Note. DIF is Differential Item Functioning and SMD is Standardized Mean Difference.

Reading Grade 6							
Operational Item Sequence Number	Item Type	DIF Comparison	Focal Group	SMD	Reference Group N	Focal Group N	DIF Level
7	MC	ETHNIC	Hispanic	0.061	3,696	454	B
10	MC	ETHNIC	Hispanic	0.071	3,705	454	B
13	MC	ETHNIC	White	0.049	2,897	265	B
24	MC	ETHNIC	White	0.090	2,898	265	B
17	CR	GENDER	Female	0.209	2,275	2,224	B
28	MC	ETHNIC	Hispanic	-0.062	3,709	454	-B
18	MC	ETHNIC	White	-0.035	2,912	266	-B
2	MC	ETHNIC	White	0.056	2,951	265	C
5	MC	ETHNIC	White	0.147	3,030	265	C
9	MC	ETHNIC	White	0.083	2,864	265	C
12	MC	ETHNIC	White	0.072	2,739	265	C
16	MC	ETHNIC	White	0.073	2,777	265	C
20	MC	ETHNIC	White	0.049	3,163	265	C
22	MC	ETHNIC	White	0.114	2,967	265	C
32	MC	ETHNIC	White	0.043	2,842	265	C
36	CR	ETHNIC	White	0.344	3,076	266	C
37	MC	ETHNIC	White	0.102	3,032	265	C
38	MC	ETHNIC	White	0.111	2,976	265	C
41	MC	ETHNIC	White	0.158	2,907	265	C
5	MC	ETHNIC	Hispanic	-0.158	3,714	454	-C
35	MC	ETHNIC	Hispanic	-0.133	3,710	454	-C

Note. DIF is Differential Item Functioning and SMD is Standardized Mean Difference. DIF levels A, B, and C are explained in the sub-section *Differential Item Functioning*.

Reading Grade 7							
Operational Item Sequence Number	**Item Type**	**DIF Comparison**	**Focal Group**	**SMD**	**Reference Group N**	**Focal Group N**	**DIF Level**
25	MC	ETHNIC	White	0.050	2,517	206	B
28	MC	ETHNIC	White	0.067	2,513	207	B
4	MC	ETHNIC	Hispanic	-0.040	3,684	394	-B
23	MC	ETHNIC	Hispanic	-0.109	3,679	394	-B
29	MC	ETHNIC	White	-0.121	2,542	207	-B
12	MC	ETHNIC	Hispanic	0.057	3,656	394	C
17	MC	ETHNIC	White	0.070	2,615	207	C
19	MC	ETHNIC	White	0.081	2,658	207	C
22	MC	ETHNIC	White	0.146	2,662	207	C
30	MC	ETHNIC	White	0.124	2,676	207	C
36	MC	ETHNIC	White	0.080	2,587	207	C
37	MC	ETHNIC	White	0.111	2,442	207	C
40	MC	ETHNIC	White	0.163	2,633	207	C
42	MC	ETHNIC	White	0.049	2,248	207	C
44	MC	ETHNIC	White	0.046	2,589	207	C
48	MC	ETHNIC	White	0.072	2,568	207	C

Note. DIF is Differential Item Functioning and SMD is Standardized Mean Difference. DIF levels A, B, and C are explained in the sub-section *Differential Item Functioning*.

Reading Grade 8							
Operational Item Sequence Number	**Item Type**	**DIF Comparison**	**Focal Group**	**SMD**	**Reference Group N**	**Focal Group N**	**DIF Level**
29	MC	ETHNIC	Hispanic	0.107	3,810	426	B
35	MC	ETHNIC	Hispanic	0.120	3,795	426	B
9	CR	GENDER	Female	0.225	2,222	2,265	B
15	MC	GENDER	Female	0.056	2,221	2,265	B
16	MC	GENDER	Female	0.069	2,223	2,265	B
3	MC	ETHNIC	Hispanic	-0.110	3,794	426	-B
6	MC	ETHNIC	Hispanic	-0.105	3,763	426	-B
7	MC	ETHNIC	Hispanic	-0.048	3,792	426	-B
48	MC	ETHNIC	Hispanic	-0.058	3,797	426	-B
34	MC	GENDER	Female	-0.098	2,223	2,265	-B

Note. DIF is Differential Item Functioning and SMD is Standardized Mean Difference. DIF levels A, B, and C are explained in the sub-section *Differential Item Functioning*.

Reading Grade 10							
Operational Item Sequence Number	**Item Type**	**DIF Comparison**	**Focal Group**	**SMD**	**Reference Group N**	**Focal Group N**	**DIF Level**
19	CR	ETHNIC	Hispanic	0.220	3,712	379	B
37	CR	ETHNIC	Hispanic	0.192	3,688	379	B
2	MC	GENDER	Female	0.029	2,064	2,253	B
7	CR	GENDER	Female	0.217	2,062	2,253	B
9	MC	ETHNIC	Hispanic	-0.056	3,705	379	-B
17	MC	ETHNIC	Hispanic	-0.104	3,644	380	-B
24	MC	ETHNIC	Hispanic	-0.071	3,715	380	-B
42	MC	ETHNIC	Hispanic	-0.114	3,695	379	-B
17	MC	GENDER	Female	-0.099	2,065	2,253	-B
31	MC	GENDER	Female	-0.121	2,061	2,253	-B
33	MC	GENDER	Female	-0.090	2,063	2,253	-B
22	MC	ETHNIC	Hispanic	0.225	3,686	379	C

Note. DIF is Differential Item Functioning and SMD is Standardized Mean Difference. DIF levels A, B, and C are explained in the sub-section *Differential Item Functioning*.

Table F3. Items Flagged for DIF Using Mantel-Haenszel: Mathematics

Mathematics Grade 3							
Operational Item Sequence Number	**Item Type**	**DIF Comparison**	**Focal Group**	**SMD**	**Reference Group N**	**Focal Group N**	**DIF Level**
1	MC	ETHNIC	Hispanic	0.037	3,775	603	B
5	MC	ETHNIC	Hispanic	0.037	3,746	603	B
18	MC	ETHNIC	White	0.110	2,998	404	B
33	MC	ETHNIC	White	0.069	2,881	404	B
35	MC	ETHNIC	White	0.036	3,000	404	B
40	MC	ETHNIC	White	0.065	3,025	404	B
19	MC	GENDER	Female	0.017	2,477	2,441	B
24	MC	GENDER	Female	0.065	2,472	2,442	B
34	MC	ETHNIC	Hispanic	-0.102	3,798	603	-B
38	MC	ETHNIC	Hispanic	-0.102	3,797	603	-B
17	MC	ETHNIC	White	-0.040	2,950	404	-B
27	MC	ETHNIC	White	-0.057	2,910	404	-B
43	MC	ETHNIC	White	-0.055	2,995	404	-B
50	MC	ETHNIC	White	-0.020	2,648	404	-B
7	MC	ETHNIC	White	0.095	3,018	404	C
8	MC	ETHNIC	White	0.183	2,808	404	C
32	MC	ETHNIC	White	0.138	2,885	404	C
34	MC	ETHNIC	White	0.167	2,847	404	C
38	MC	ETHNIC	White	0.085	2,793	404	C
51	MC	ETHNIC	White	0.214	2,734	404	C
6	MC	ETHNIC	White	-0.116	2,846	404	-C

Note. DIF is Differential Item Functioning and SMD is Standardized Mean Difference. DIF levels A, B, and C are explained in the sub-section *Differential Item Functioning*.

Mathematics Grade 4							
Operational Item Sequence Number	**Item Type**	**DIF Comparison**	**Focal Group**	**SMD**	**Reference Group N**	**Focal Group N**	**DIF Level**
47	CR	ETHNIC	Hispanic	0.225	3,784	597	B
54	MC	ETHNIC	Hispanic	0.032	3,786	599	B
22	MC	ETHNIC	White	0.082	3,163	358	B
26	MC	ETHNIC	White	0.029	3,615	359	B
31	MC	ETHNIC	White	0.056	3,252	359	B
41	MC	ETHNIC	White	0.058	3,510	359	B
43	MC	ETHNIC	White	0.080	3,511	359	B
46	MC	ETHNIC	White	0.147	3,411	359	B
47	CR	ETHNIC	White	0.197	3,227	358	B
48	MC	ETHNIC	White	0.071	3,413	359	B
50	MC	ETHNIC	White	0.135	3,563	364	B
38	MC	ETHNIC	Hispanic	-0.073	3,775	599	-B
16	MC	ETHNIC	White	-0.035	3,462	364	-B
39	MC	ETHNIC	White	-0.037	3,292	359	-B
11	MC	ETHNIC	White	0.184	3,398	359	C
44	MC	ETHNIC	White	0.073	3,291	359	C
51	MC	ETHNIC	White	0.156	3,494	359	C
5	MC	ETHNIC	White	-0.101	3,486	359	-C
17	MC	ETHNIC	White	-0.054	3,398	359	-C
25	MC	ETHNIC	White	-0.039	3,418	359	-C
35	MC	ETHNIC	White	-0.149	3,492	359	-C

Note. DIF is Differential Item Functioning and SMD is Standardized Mean Difference. DIF levels A, B, and C are explained in the sub-section *Differential Item Functioning*.

Mathematics Grade 5							
Operational Item Sequence Number	**Item Type**	**DIF Comparison**	**Focal Group**	**SMD**	**Reference Group N**	**Focal Group N**	**DIF Level**
11	MC	ETHNIC	Hispanic	0.025	3,632	513	B
13	MC	ETHNIC	Hispanic	0.116	3,590	513	B
45	MC	ETHNIC	Hispanic	0.099	3,578	513	B
49	MC	ETHNIC	Hispanic	0.044	3,626	513	B
5	MC	ETHNIC	White	0.080	2,965	304	B
16	MC	ETHNIC	White	0.039	2,986	304	B
18	MC	ETHNIC	White	0.045	2,848	304	B
21	CR	ETHNIC	White	0.222	2,789	304	B
23	MC	ETHNIC	White	0.048	2,767	304	B
40	MC	ETHNIC	White	0.050	2,639	304	B
45	MC	ETHNIC	White	0.073	2,711	304	B
6	CR	ETHNIC	White	-0.208	2,872	304	-B
27	MC	ETHNIC	White	-0.053	2,723	304	-B
41	MC	ETHNIC	White	-0.060	2,902	304	-B
53	MC	ETHNIC	White	-0.048	2,861	304	-B
13	MC	ETHNIC	White	0.147	2,841	304	C
14	MC	ETHNIC	White	0.092	2,693	304	C
15	MC	ETHNIC	White	0.091	2,893	304	C
17	MC	ETHNIC	White	0.116	2,997	304	C
33	MC	ETHNIC	White	0.096	2,823	304	C
54	MC	ETHNIC	White	0.039	2,874	304	C
35	MC	GENDER	Female	0.022	2,306	2,212	C
37	MC	GENDER	Female	0.038	2,306	2,211	C

Note. DIF is Differential Item Functioning and SMD is Standardized Mean Difference. DIF levels A, B, and C are explained in the sub-section *Differential Item Functioning*.

Mathematics Grade 6							
Operational Item Sequence Number	**Item Type**	**DIF Comparison**	**Focal Group**	**SMD**	**Reference Group N**	**Focal Group N**	**DIF Level**
15	MC	ETHNIC	Hispanic	0.111	3,707	468	B
41	MC	ETHNIC	Hispanic	0.087	3,709	468	B
7	MC	ETHNIC	White	0.039	3,229	267	B
25	MC	ETHNIC	White	0.099	3,153	267	B
39	MC	ETHNIC	White	0.078	3,184	267	B
42	MC	ETHNIC	White	0.111	3,322	267	B
48	CR	ETHNIC	White	0.189	3,277	267	B
11	MC	ETHNIC	White	-0.089	3,159	267	-B
14	MC	ETHNIC	White	-0.076	3,099	267	-B
39	MC	ETHNIC	Hispanic	0.133	3,709	468	C
2	MC	ETHNIC	White	0.077	3,048	267	C
12	MC	ETHNIC	White	0.063	3,154	267	C
29	MC	ETHNIC	White	0.062	3,159	267	C
30	MC	ETHNIC	White	0.045	2,991	267	C
32	MC	ETHNIC	White	0.128	3,179	267	C
41	MC	ETHNIC	White	0.248	3,175	267	C
53	MC	ETHNIC	White	0.045	3,226	267	C
17	MC	ETHNIC	White	-0.119	3,055	267	-C
44	MC	ETHNIC	White	-0.054	2,907	267	-C
49	MC	ETHNIC	White	-0.048	3,349	267	-C

Note. DIF is Differential Item Functioning and SMD is Standardized Mean Difference. DIF levels A, B, and C are explained in the sub-section *Differential Item Functioning*.

Mathematics Grade 7							
Operational Item Sequence Number	**Item Type**	**DIF Comparison**	**Focal Group**	**SMD**	**Reference Group N**	**Focal Group N**	**DIF Level**
6	CR	ETHNIC	Hispanic	0.209	3,650	403	B
48	CR	ETHNIC	Hispanic	0.241	3,637	403	B
24	MC	ETHNIC	White	0.041	2,656	207	B
11	MC	GENDER	Female	0.044	2,183	2,181	B
36	MC	GENDER	Female	0.058	2,184	2,180	B
9	MC	ETHNIC	White	-0.070	2,565	207	-B
28	MC	ETHNIC	White	-0.061	2,841	207	-B
6	CR	ETHNIC	White	0.429	2,863	207	C
7	MC	ETHNIC	White	0.118	2,698	207	C
8	MC	ETHNIC	White	0.072	2,686	207	C
10	MC	ETHNIC	White	0.133	2,696	207	C
19	MC	ETHNIC	White	0.140	2,633	207	C
40	MC	ETHNIC	White	0.053	2,672	207	C
46	MC	ETHNIC	White	0.111	2,695	207	C
48	CR	ETHNIC	White	0.363	2,753	207	C
52	MC	ETHNIC	Hispanic	-0.119	3,648	403	-C
18	MC	ETHNIC	White	-0.087	2,766	207	-C
30	MC	ETHNIC	White	-0.043	2,659	207	-C
33	MC	ETHNIC	White	-0.067	2,800	207	-C
52	MC	ETHNIC	White	-0.173	2,811	207	-C

Note. DIF is Differential Item Functioning and SMD is Standardized Mean Difference. DIF levels A, B, and C are explained in the sub-section *Differential Item Functioning*.

Mathematics Grade 8							
Operational Item Sequence Number	**Item Type**	**DIF Comparison**	**Focal Group**	**SMD**	**Reference Group N**	**Focal Group N**	**DIF Level**
38	MC	ETHNIC	Hispanic	0.048	3,770	448	B
39	MC	ETHNIC	Hispanic	0.109	3,758	448	B
48	CR	GENDER	Female	0.201	2,222	2,266	B
7	MC	ETHNIC	Hispanic	-0.059	3,758	448	-B
15	MC	ETHNIC	Hispanic	-0.078	3,763	448	-B
34	MC	GENDER	Female	-0.094	2,220	2,269	-B

Note. DIF is Differential Item Functioning and SMD is Standardized Mean Difference.

Mathematics Grade 10							
Operational Item Sequence Number	**Item Type**	**DIF Comparison**	**Focal Group**	**SMD**	**Reference Group N**	**Focal Group N**	**DIF Level**
38	MC	GENDER	Female	0.076	2,041	2,244	B
45	MC	ETHNIC	Hispanic	-0.105	3,625	375	-B
12	MC	GENDER	Female	-0.070	2,041	2,244	-B
19	MC	GENDER	Female	-0.104	2,041	2,244	-B
37	MC	GENDER	Female	-0.093	2,041	2,244	-B

Note. DIF is Differential Item Functioning and SMD is Standardized Mean Difference. DIF levels A, B, and C are explained in the sub-section *Differential Item Functioning*.

Table F4. Items Flagged for DIF Using Mantel-Haenszel: Science/Biology

Science Grade 5							
Operational Item Sequence Number	**Item Type**	**DIF Comparison**	**Focal Group**	**SMD**	**Reference Group N**	**Focal Group N**	**DIF Level**
27	MC	ETHNIC	Hispanic	0.088	3,507	510	B
41	MC	ETHNIC	Hispanic	0.086	3,524	510	B
1	MC	ETHNIC	White	0.066	3,011	296	B
4	MC	ETHNIC	White	0.063	2,987	296	B
7	MC	ETHNIC	White	0.041	3,018	296	B
10	CR	ETHNIC	White	0.225	3,037	296	B
13	MC	ETHNIC	White	0.081	2,708	296	B
18	MC	ETHNIC	White	0.102	2,882	295	B
28	MC	ETHNIC	White	0.086	3,013	295	B
30	MC	ETHNIC	White	0.118	2,746	295	B
31	MC	ETHNIC	White	0.055	3,152	296	B
34	MC	ETHNIC	White	0.127	2,866	296	B
42	MC	ETHNIC	White	0.067	2,880	296	B
44	MC	ETHNIC	White	0.111	3,014	296	B
45	MC	ETHNIC	White	0.092	2,884	294	B
11	MC	GENDER	Female	0.087	2,253	2,178	B
38	MC	ETHNIC	White	-0.148	2,743	296	-B
42	MC	GENDER	Female	-0.112	2,255	2,178	-B
8	MC	ETHNIC	White	0.047	2,720	296	C
19	MC	ETHNIC	White	0.057	2,865	296	C
24	MC	ETHNIC	White	0.147	2,907	296	C
26	MC	ETHNIC	White	0.041	2,653	295	C
39	CR	ETHNIC	White	0.256	2,738	291	C
11	MC	ETHNIC	White	-0.065	2,828	296	-C
25	MC	ETHNIC	White	-0.052	2,868	296	-C

Note. DIF is Differential Item Functioning and SMD is Standardized Mean Difference. DIF levels A, B, and C are explained in the sub-section *Differential Item Functioning*.

Science Grade 8							
Operational Item Sequence Number	**Item Type**	**DIF Comparison**	**Focal Group**	**SMD**	**Reference Group N**	**Focal Group N**	**DIF Level**
32	MC	GENDER	Female	0.111	2,129	2,207	B
46	MC	GENDER	Female	0.096	2,129	2,207	B
4	MC	ETHNIC	Hispanic	-0.084	3,602	440	-B
9	MC	GENDER	Female	-0.117	2,128	2,207	-B
14	MC	GENDER	Female	-0.102	2,134	2,207	-B

Note. DIF is Differential Item Functioning and SMD is Standardized Mean Difference. DIF levels A, B, and C are explained in the sub-section *Differential Item Functioning*.

Biology							
Operational Item Sequence Number	**Item Type**	**DIF Comparison**	**Focal Group**	**SMD**	**Reference Group N**	**Focal Group N**	**DIF Level**
5	MC	ETHNIC	Hispanic	0.084	3,387	376	B
50	MC	ETHNIC	Hispanic	0.109	3,391	377	B
47	MC	GENDER	Female	0.101	1,910	2,050	B
44	MC	ETHNIC	Hispanic	-0.125	3,397	376	-B

Note. DIF is Differential Item Functioning and SMD is Standardized Mean Difference. DIF levels A, B, and C are explained in the sub-section *Differential Item Functioning*.

Appendix G: Operational Item Adjusted *P* Values

Table G1. DC CAS 2010 Operational Form Item Characteristics: Reading

Grade 3 Reading								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	4,918	1	0.91		25	4,861	1	0.81
2	4,870	1	0.43		26	4,826	1	0.80
3	4,919	1	0.67		27	4,897	1	0.41
4	4,909	1	0.91		28	4,901	1	0.62
5	4,920	1	0.75		29	4,917	1	0.65
6	4,907	1	0.64		30	4,914	1	0.86
7	4,908	1	0.71		31	4,816	1	0.82
8	4,874	1	0.65		32	4,907	1	0.85
9	4,811	3	0.36		33	4,907	1	0.80
10	4,916	1	0.73		34	4,887	1	0.74
11	4,894	1	0.74		35	4,910	1	0.61
12	4,816	1	0.82		36	4,902	1	0.63
13	4,897	1	0.57		37	4,862	3	0.42
14	4,860	1	0.62		38	4,900	1	0.75
15	4,908	1	0.86		39	4,898	1	0.43
16	4,901	1	0.89		40	4,895	1	0.69
17	4,898	1	0.77		41	4,896	1	0.69
18	4,905	1	0.71		42	4,900	1	0.76
19	4,912	1	0.72		43	4,818	1	0.67
20	4,818	3	0.38		44	4,878	1	0.61
21	4,919	1	0.47		45	4,857	1	0.62
22	4,911	1	0.35		46	4,828	1	0.68
23	4,913	1	0.82		47	4,860	1	0.71
24	4,910	1	0.61		48	4,863	1	0.77

Note. The adjusted *p* value for an item includes responses only for examinees with valid responses to that item.

Grade 4 Reading								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	4,828	1	0.76		25	4,821	1	0.56
2	4,827	1	0.85		26	4,823	1	0.69
3	4,824	1	0.72		27	4,816	1	0.77
4	4,815	1	0.78		28	4,821	1	0.58
5	4,825	1	0.61		29	4,822	1	0.66
6	4,811	1	0.65		30	4,815	1	0.37
7	4,813	1	0.62		31	4,815	1	0.54
8	4,801	1	0.55		32	4,819	1	0.40
9	4,731	3	0.44		33	4,819	1	0.72
10	4,815	1	0.69		34	4,817	1	0.57
11	4,810	1	0.47		35	4,817	1	0.77
12	4,811	1	0.62		36	4,790	1	0.54
13	4,813	1	0.61		37	4,770	3	0.29
14	4,803	1	0.50		38	4,780	1	0.78
15	4,802	1	0.72		39	4,772	1	0.55
16	4,795	1	0.72		40	4,770	1	0.81
17	4,787	1	0.60		41	4,761	1	0.79
18	4,756	1	0.81		42	4,791	1	0.76
19	4,731	3	0.50		43	4,792	1	0.66
20	4,824	1	0.79		44	4,788	1	0.48
21	4,821	1	0.54		45	4,776	1	0.63
22	4,819	1	0.63		46	4,785	1	0.24
23	4,819	1	0.53		47	4,781	1	0.45
24	4,822	1	0.56		48	4,780	1	0.61

Grade 5 Reading								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	4,509	1	0.59		25	4,494	1	0.58
2	4,510	1	0.74		26	4,498	1	0.74
3	4,508	1	0.68		27	4,505	1	0.85
4	4,505	1	0.81		28	4,504	1	0.65
5	4,509	1	0.34		29	4,502	1	0.86
6	4,487	1	0.55		30	4,504	1	0.84
7	4,453	3	0.47		31	4,503	1	0.85
8	4,497	1	0.84		32	4,501	1	0.84
9	4,498	1	0.51		33	4,502	1	0.84
10	4,495	1	0.75		34	4,502	1	0.58
11	4,497	1	0.78		35	4,478	1	0.57
12	4,489	1	0.78		36	4,465	3	0.43
13	4,484	1	0.33		37	4,486	1	0.78
14	4,487	1	0.68		38	4,487	1	0.84
15	4,477	1	0.56		39	4,485	1	0.67
16	4,461	1	0.82		40	4,481	1	0.61
17	4,423	3	0.33		41	4,481	1	0.86
18	4,504	1	0.68		42	4,486	1	0.82
19	4,503	1	0.79		43	4,489	1	0.74
20	4,502	1	0.76		44	4,483	1	0.82
21	4,506	1	0.66		45	4,479	1	0.65
22	4,494	1	0.81		46	4,485	1	0.87
23	4,497	1	0.76		47	4,485	1	0.60
24	4,500	1	0.34		48	4,484	1	0.42

Grade 6 Reading								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	4,519	1	0.84		25	4,512	1	0.65
2	4,517	1	0.68		26	4,511	1	0.68
3	4,515	1	0.81		27	4,511	1	0.72
4	4,514	1	0.37		28	4,508	1	0.77
5	4,484	1	0.46		29	4,507	1	0.47
6	4,448	3	0.65		30	4,511	1	0.90
7	4,514	1	0.71		31	4,514	1	0.71
8	4,517	1	0.71		32	4,510	1	0.75
9	4,514	1	0.48		33	4,501	1	0.58
10	4,511	1	0.69		34	4,501	1	0.69
11	4,517	1	0.76		35	4,485	1	0.75
12	4,492	1	0.61		36	4,447	3	0.13
13	4,500	1	0.67		37	4,500	1	0.46
14	4,503	1	0.65		38	4,497	1	0.48
15	4,500	1	0.63		39	4,499	1	0.65
16	4,492	1	0.72		40	4,497	1	0.69
17	4,398	3	0.32		41	4,490	1	0.34
18	4,515	1	0.82		42	4,462	1	0.51
19	4,516	1	0.56		43	4,459	1	0.53
20	4,514	1	0.65		44	4,460	1	0.68
21	4,514	1	0.70		45	4,458	1	0.48
22	4,513	1	0.50		46	4,458	1	0.52
23	4,512	1	0.79		47	4,456	1	0.65
24	4,511	1	0.44		48	4,455	1	0.69

Grade 7 Reading								
Operational Item Sequence Number	N	Max Points	Adjusted *P* Value		Operational Item Sequence Number	N	Max Points	Adjusted *P* Value
1	4,379	1	0.88		25	4,365	1	0.51
2	4,378	1	0.84		26	4,350	1	0.59
3	4,380	1	0.92		27	4,357	1	0.81
4	4,381	1	0.87		28	4,358	1	0.50
5	4,379	1	0.61		29	4,356	1	0.54
6	4,372	1	0.52		30	4,353	1	0.44
7	4,371	1	0.87		31	4,364	1	0.78
8	4,306	3	0.66		32	4,358	1	0.60
9	4,375	1	0.89		33	4,361	1	0.76
10	4,374	1	0.87		34	4,360	1	0.56
11	4,370	1	0.77		35	4,351	1	0.54
12	4,374	1	0.86		36	4,361	1	0.61
13	4,364	1	0.52		37	4,346	1	0.44
14	4,365	1	0.65		38	4,295	3	0.50
15	4,358	1	0.61		39	4,350	1	0.65
16	4,366	1	0.66		40	4,350	1	0.57
17	4,360	1	0.63		41	4,352	1	0.77
18	4,216	3	0.25		42	4,348	1	0.69
19	4,365	1	0.48		43	4,347	1	0.77
20	4,366	1	0.67		44	4,339	1	0.68
21	4,366	1	0.77		45	4,351	1	0.59
22	4,363	1	0.37		46	4,348	1	0.45
23	4,368	1	0.39		47	4,348	1	0.67
24	4,364	1	0.69		48	4,348	1	0.72

Grade 8 Reading								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	4,523	1	0.64		25	4,506	1	0.64
2	4,520	1	0.58		26	4,508	1	0.60
3	4,517	1	0.65		27	4,511	1	0.78
4	4,519	1	0.73		28	4,508	1	0.65
5	4,521	1	0.53		29	4,508	1	0.52
6	4,500	1	0.38		30	4,485	1	0.42
7	4,520	1	0.86		31	4,483	1	0.62
8	4,503	1	0.78		32	4,485	1	0.64
9	4,413	3	0.60		33	4,480	1	0.48
10	4,519	1	0.84		34	4,485	1	0.44
11	4,515	1	0.51		35	4,480	1	0.38
12	4,516	1	0.70		36	4,483	1	0.48
13	4,514	1	0.63		37	4,480	1	0.53
14	4,511	1	0.41		38	4,477	1	0.62
15	4,512	1	0.78		39	4,484	1	0.53
16	4,507	1	0.72		40	4,476	1	0.57
17	4,506	1	0.66		41	4,479	1	0.72
18	4,493	1	0.56		42	4,360	3	0.35
19	4,362	3	0.24		43	4,442	1	0.85
20	4,517	1	0.84		44	4,436	1	0.54
21	4,515	1	0.73		45	4,437	1	0.43
22	4,508	1	0.61		46	4,433	1	0.64
23	4,515	1	0.76		47	4,437	1	0.40
24	4,513	1	0.73		48	4,438	1	0.81

Grade 10 Reading								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	4,384	1	0.72		25	4,369	1	0.69
2	4,392	1	0.90		26	4,373	1	0.80
3	4,377	1	0.60		27	4,371	1	0.74
4	4,385	1	0.72		28	4,366	1	0.57
5	4,389	1	0.73		29	4,367	1	0.71
6	4,359	1	0.85		30	4,328	1	0.37
7	4,191	3	0.34		31	4,333	1	0.65
8	4,385	1	0.83		32	4,329	1	0.53
9	4,383	1	0.82		33	4,329	1	0.53
10	4,381	1	0.85		34	4,328	1	0.50
11	4,384	1	0.66		35	4,326	1	0.84
12	4,386	1	0.77		36	4,303	1	0.61
13	4,379	1	0.81		37	4,012	3	0.48
14	4,383	1	0.70		38	4,316	1	0.73
15	4,357	1	0.66		39	4,314	1	0.68
16	4,364	1	0.60		40	4,315	1	0.74
17	4,364	1	0.57		41	4,313	1	0.67
18	4,355	1	0.66		42	4,307	1	0.67
19	4,096	3	0.56		43	4,275	1	0.56
20	4,381	1	0.87		44	4,273	1	0.71
21	4,379	1	0.57		45	4,280	1	0.62
22	4,373	1	0.36		46	4,273	1	0.72
23	4,368	1	0.46		47	4,280	1	0.38
24	4,377	1	0.29		48	4,275	1	0.61

Table G2. DC CAS 2010 Operational Form Item Characteristics: Mathematics

Grade 3 Mathematics								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	4,937	1	0.87		28	4,927	1	0.66
2	4,935	1	0.92		29	4,932	1	0.81
3	4,929	1	0.91		30	4,920	1	0.73
4	4,934	1	0.84		31	4,928	1	0.90
5	4,930	1	0.89		32	4,925	1	0.53
6	4,930	1	0.64		33	4,900	1	0.65
7	4,928	1	0.59		34	4,903	1	0.39
8	4,923	1	0.36		35	4,865	1	0.73
9	4,901	1	0.56		36	4,891	1	0.58
10	4,880	3	0.40		37	4,879	1	0.63
11	4,914	1	0.51		38	4,904	1	0.64
12	4,891	1	0.78		39	4,918	1	0.80
13	4,652	1	0.70		40	4,907	1	0.52
14	4,912	1	0.38		41	4,925	1	0.81
15	4,932	1	0.80		42	4,920	1	0.96
16	4,933	1	0.79		43	4,895	1	0.61
17	4,914	1	0.72		44	4,859	3	0.58
18	4,914	1	0.51		45	4,895	1	0.59
19	4,924	1	0.95		46	4,883	1	0.57
20	4,906	1	0.87		47	4,884	1	0.93
21	4,872	1	0.64		48	4,908	1	0.88
22	4,857	1	0.60		49	4,887	1	0.47
23	4,874	3	0.70		50	4,913	1	0.88
24	4,906	1	0.82		51	4,892	1	0.30
25	4,886	1	0.68		52	4,905	1	0.50
26	4,923	1	0.66		53	4,897	1	0.68
27	4,830	1	0.70		54	4,902	1	0.61

Grade 4 Mathematics

Operational Item Sequence Number	N	Max Points	Adjusted *P* Value		Operational Item Sequence Number	N	Max Points	Adjusted *P* Value
1	4,861	1	0.72		28	4,836	1	0.55
2	4,849	1	0.69		29	4,838	1	0.75
3	4,860	1	0.84		30	4,839	1	0.64
4	4,856	1	0.69		31	4,836	1	0.61
5	4,854	1	0.63		32	4,832	1	0.65
6	4,845	1	0.49		33	4,835	1	0.56
7	4,861	1	0.72		34	4,837	1	0.51
8	4,845	1	0.76		35	4,831	1	0.60
9	4,823	1	0.53		36	4,831	1	0.63
10	4,790	3	0.41		37	4,830	1	0.56
11	4,853	1	0.28		38	4,834	1	0.67
12	4,856	1	0.25		39	4,829	1	0.79
13	4,853	1	0.69		40	4,784	1	0.75
14	4,853	1	0.75		41	4,818	1	0.54
15	4,857	1	0.94		42	4,820	1	0.57
16	4,853	1	0.68		43	4,819	1	0.57
17	4,849	1	0.84		44	4,820	1	0.55
18	4,822	1	0.44		45	4,827	1	0.83
19	4,803	3	0.54		46	4,810	1	0.31
20	4,849	1	0.63		47	4,703	3	0.64
21	4,853	1	0.67		48	4,832	1	0.62
22	4,854	1	0.62		49	4,829	1	0.49
23	4,848	1	0.45		50	4,824	1	0.18
24	4,854	1	0.59		51	4,828	1	0.32
25	4,847	1	0.85		52	4,827	1	0.29
26	4,837	1	0.81		53	4,821	1	0.80
27	4,820	1	0.75		54	4,832	1	0.88

Grade 5 Mathematics								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	4,529	1	0.86		28	4,519	1	0.76
2	4,524	1	0.85		29	4,515	1	0.67
3	4,531	1	0.87		30	4,522	1	0.91
4	4,525	1	0.86		31	4,518	1	0.60
5	4,505	1	0.49		32	4,519	1	0.56
6	4,490	3	0.34		33	4,521	1	0.51
7	4,526	1	0.61		34	4,523	1	0.90
8	4,530	1	0.66		35	4,522	1	0.96
9	4,527	1	0.69		36	4,517	1	0.60
10	4,530	1	0.79		37	4,522	1	0.92
11	4,528	1	0.89		38	4,506	1	0.44
12	4,531	1	0.84		39	4,518	1	0.77
13	4,526	1	0.45		40	4,494	1	0.75
14	4,527	1	0.61		41	4,514	1	0.64
15	4,524	1	0.74		42	4,514	1	0.71
16	4,519	1	0.60		43	4,507	1	0.64
17	4,522	1	0.53		44	4,503	1	0.57
18	4,521	1	0.59		45	4,507	1	0.66
19	4,513	1	0.62		46	4,508	1	0.62
20	4,500	1	0.75		47	4,486	1	0.25
21	4,490	3	0.42		48	4,455	3	0.49
22	4,519	1	0.48		49	4,511	1	0.83
23	4,523	1	0.75		50	4,513	1	0.82
24	4,516	1	0.47		51	4,509	1	0.27
25	4,519	1	0.64		52	4,508	1	0.54
26	4,522	1	0.75		53	4,508	1	0.76
27	4,503	1	0.58		54	4,510	1	0.70

Grade 6 Mathematics								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	4,547	1	0.86		28	4,534	1	0.86
2	4,540	1	0.57		29	4,530	1	0.50
3	4,535	1	0.50		30	4,531	1	0.77
4	4,524	1	0.79		31	4,533	1	0.46
5	4,518	1	0.70		32	4,532	1	0.43
6	4,375	3	0.16		33	4,532	1	0.76
7	4,546	1	0.70		34	4,530	1	0.58
8	4,545	1	0.71		35	4,532	1	0.75
9	4,535	1	0.68		36	4,529	1	0.63
10	4,544	1	0.66		37	4,526	1	0.58
11	4,539	1	0.45		38	4,526	1	0.67
12	4,541	1	0.61		39	4,526	1	0.55
13	4,542	1	0.35		40	4,501	1	0.40
14	4,539	1	0.31		41	4,518	1	0.29
15	4,538	1	0.67		42	4,513	1	0.30
16	4,523	1	0.54		43	4,517	1	0.65
17	4,531	1	0.62		44	4,515	1	0.66
18	4,534	1	0.64		45	4,517	1	0.78
19	4,538	1	0.52		46	4,513	1	0.63
20	4,493	1	0.73		47	4,499	1	0.32
21	4,468	3	0.44		48	4,440	3	0.45
22	4,536	1	0.75		49	4,516	1	0.62
23	4,536	1	0.56		50	4,516	1	0.36
24	4,537	1	0.62		51	4,516	1	0.44
25	4,534	1	0.53		52	4,514	1	0.67
26	4,530	1	0.51		53	4,513	1	0.62
27	4,523	1	0.48		54	4,512	1	0.29

Grade 7 Mathematics								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	4,373	1	0.71		28	4,372	1	0.77
2	4,383	1	0.75		29	4,348	1	0.37
3	4,383	1	0.78		30	4,365	1	0.75
4	4,343	1	0.44		31	4,356	1	0.61
5	4,350	1	0.71		32	4,362	1	0.30
6	4,301	3	0.37		33	4,358	1	0.75
7	4,374	1	0.36		34	4,367	1	0.40
8	4,375	1	0.55		35	4,359	1	0.73
9	4,378	1	0.63		36	4,367	1	0.81
10	4,373	1	0.25		37	4,368	1	0.72
11	4,381	1	0.84		38	4,368	1	0.62
12	4,370	1	0.37		39	4,367	1	0.53
13	4,376	1	0.69		40	4,363	1	0.64
14	4,376	1	0.83		41	4,347	1	0.65
15	4,376	1	0.13		42	4,346	1	0.45
16	4,371	1	0.75		43	4,354	1	0.62
17	4,354	1	0.49		44	4,348	1	0.78
18	4,378	1	0.68		45	4,336	1	0.21
19	4,367	1	0.37		46	4,347	1	0.50
20	4,311	1	0.62		47	4,313	1	0.47
21	4,234	3	0.47		48	4,266	3	0.33
22	4,379	1	0.79		49	4,333	1	0.34
23	4,370	1	0.77		50	4,353	1	0.53
24	4,372	1	0.64		51	4,347	1	0.56
25	4,370	1	0.38		52	4,355	1	0.27
26	4,371	1	0.49		53	4,352	1	0.40
27	4,364	1	0.49		54	4,352	1	0.69

Grade 8 Mathematics								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	4,518	1	0.78		28	4,474	1	0.35
2	4,516	1	0.13		29	4,480	1	0.67
3	4,513	1	0.30		30	4,479	1	0.44
4	4,512	1	0.44		31	4,487	1	0.61
5	4,461	1	0.23		32	4,481	1	0.75
6	4,426	3	0.39		33	4,485	1	0.37
7	4,522	1	0.14		34	4,485	1	0.42
8	4,504	1	0.39		35	4,475	1	0.52
9	4,512	1	0.33		36	4,484	1	0.58
10	4,512	1	0.62		37	4,483	1	0.52
11	4,509	1	0.63		38	4,485	1	0.88
12	4,512	1	0.20		39	4,479	1	0.28
13	4,508	1	0.36		40	4,463	1	0.40
14	4,518	1	0.56		41	4,457	1	0.55
15	4,505	1	0.71		42	4,460	1	0.17
16	4,511	1	0.68		43	4,462	1	0.63
17	4,502	1	0.60		44	4,459	1	0.34
18	4,484	1	0.38		45	4,448	1	0.48
19	4,506	1	0.37		46	4,457	1	0.77
20	4,479	1	0.47		47	4,441	1	0.61
21	4,312	3	0.26		48	4,235	3	0.40
22	4,506	1	0.76		49	4,451	1	0.50
23	4,506	1	0.48		50	4,456	1	0.21
24	4,500	1	0.34		51	4,453	1	0.50
25	4,503	1	0.81		52	4,457	1	0.42
26	4,504	1	0.74		53	4,460	1	0.41
27	4,505	1	0.67		54	4,454	1	0.36

Grade 10 Mathematics								
Operational Item Sequence Number	N	Max Points	Adjusted *P* Value		Operational Item Sequence Number	N	Max Points	Adjusted *P* Value
1	4,342	1	0.56		28	4,253	1	0.49
2	4,334	1	0.65		29	4,254	1	0.64
3	4,347	1	0.55		30	4,229	1	0.34
4	4,272	1	0.45		31	4,244	1	0.32
5	4,347	1	0.87		32	4,239	1	0.32
6	3,740	3	0.29		33	4,236	1	0.42
7	4,347	1	0.59		34	4,255	1	0.43
8	4,352	1	0.85		35	4,248	1	0.35
9	4,341	1	0.45		36	4,253	1	0.56
10	4,328	1	0.30		37	4,253	1	0.46
11	4,327	1	0.32		38	4,249	1	0.54
12	4,333	1	0.30		39	4,241	1	0.33
13	4,316	1	0.37		40	4,234	1	0.31
14	4,333	1	0.16		41	4,250	1	0.50
15	4,303	1	0.43		42	4,230	1	0.30
16	4,334	1	0.54		43	4,243	1	0.19
17	4,338	1	0.70		44	4,237	1	0.68
18	4,317	1	0.20		45	4,248	1	0.33
19	4,333	1	0.41		46	4,252	1	0.46
20	4,290	1	0.61		47	4,241	1	0.54
21	4,077	3	0.57		48	3,802	3	0.48
22	4,319	1	0.51		49	4,233	1	0.41
23	4,334	1	0.42		50	4,247	1	0.83
24	4,328	1	0.57		51	4,248	1	0.54
25	4,332	1	0.75		52	4,244	1	0.54
26	4,311	1	0.60		53	4,237	1	0.45
27	4,318	1	0.44		54	4,235	1	0.52

Table G3. DC CAS 2010 Operational Form Item Characteristics: Science/Biology

Grade 5 Science								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	4,455	1	0.49		26	4,443	1	0.53
2	4,456	1	0.44		27	4,430	1	0.61
3	4,444	1	0.32		28	4,440	1	0.62
4	4,454	1	0.67		29	4,436	1	0.70
5	4,447	1	0.65		30	4,427	1	0.26
6	4,451	1	0.31		31	4,429	1	0.68
7	4,450	1	0.58		32	4,424	1	0.31
8	4,445	1	0.62		33	4,424	1	0.59
9	4,390	1	0.91		34	4,414	1	0.43
10	4,367	2	0.28		35	4,408	1	0.33
11	4,446	1	0.65		36	4,406	1	0.46
12	4,445	1	0.64		37	4,398	1	0.31
13	4,431	1	0.43		38	4,377	1	0.35
14	4,444	1	0.51		39	4,271	2	0.22
15	4,437	1	0.35		40	4,415	1	0.26
16	4,439	1	0.23		41	4,413	1	0.52
17	4,424	1	0.33		42	4,412	1	0.46
18	4,432	1	0.31		43	4,408	1	0.55
19	4,438	1	0.65		44	4,410	1	0.40
20	4,428	1	0.31		45	4,407	1	0.49
21	4,363	2	0.68		46	4,413	1	0.84
22	4,450	1	0.79		47	4,412	1	0.65
23	4,444	1	0.56		48	4,404	1	0.26
24	4,446	1	0.41		49	4,405	1	0.34
25	4,441	1	0.67		50	4,402	1	0.27

Grade 8 Science								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	4,393	1	0.78		26	4,351	1	0.32
2	4,392	1	0.51		27	4,347	1	0.39
3	4,385	1	0.36		28	4,344	1	0.45
4	4,384	1	0.46		29	4,331	1	0.30
5	4,386	1	0.75		30	4,332	1	0.50
6	4,384	1	0.55		31	4,328	1	0.39
7	4,377	1	0.24		32	4,320	1	0.52
8	4,371	1	0.36		33	4,322	1	0.38
9	4,346	1	0.58		34	4,321	1	0.43
10	4,040	2	0.21		35	4,328	1	0.30
11	4,375	1	0.58		36	4,318	1	0.49
12	4,363	1	0.54		37	4,314	1	0.36
13	4,370	1	0.68		38	4,303	1	0.30
14	4,362	1	0.36		39	3,693	2	0.07
15	4,356	1	0.36		40	4,330	1	0.54
16	4,357	1	0.28		41	4,330	1	0.57
17	4,346	1	0.31		42	4,325	1	0.16
18	4,358	1	0.61		43	4,326	1	0.28
19	4,348	1	0.39		44	4,318	1	0.38
20	3,451	2	0.17		45	4,317	1	0.32
21	4,361	1	0.29		46	4,324	1	0.70
22	4,353	1	0.46		47	4,323	1	0.60
23	4,343	1	0.46		48	4,314	1	0.27
24	4,355	1	0.51		49	4,317	1	0.25
25	4,353	1	0.39		50	4,316	1	0.48

Biology								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	4,094	1	0.13		26	4,049	1	0.35
2	4,089	1	0.34		27	4,044	1	0.63
3	4,070	1	0.19		28	4,053	1	0.29
4	4,090	1	0.46		29	4,050	1	0.62
5	4,072	1	0.25		30	4,035	1	0.37
6	4,093	1	0.77		31	4,032	1	0.38
7	4,087	1	0.27		32	4,026	1	0.32
8	4,064	1	0.17		33	4,029	1	0.35
9	4,055	1	0.48		34	4,015	1	0.44
10	3,793	2	0.68		35	4,032	1	0.36
11	4,083	1	0.36		36	4,025	1	0.33
12	4,076	1	0.31		37	4,017	1	0.35
13	4,078	1	0.54		38	3,998	1	0.33
14	4,059	1	0.50		39	3,148	2	0.15
15	4,058	1	0.27		40	4,033	1	0.38
16	4,061	1	0.39		41	4,034	1	0.40
17	4,053	1	0.25		42	4,019	1	0.35
18	4,045	1	0.35		43	4,030	1	0.26
19	4,048	1	0.33		44	4,027	1	0.45
20	4,047	1	0.39		45	4,015	1	0.22
21	3,643	2	0.25		46	4,020	1	0.26
22	4,054	1	0.42		47	4,027	1	0.70
23	4,046	1	0.33		48	4,023	1	0.16
24	4,062	1	0.43		49	4,013	1	0.21
25	4,061	1	0.67		50	4,015	1	0.38