Technical Report
Spring 2011 Operational Test
Administration

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

June 30, 2011

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 the spring of 2011 to assess students' skills in Grades 3–8 and 10 in Reading and Mathematics; in Grades 5 and 8 in Science; in high school Biology; and in 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, and Mathematics, 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 of 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, 8, and 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 2011 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 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 2011 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, 2009; see http://www.ed.gov/policy/elsec/guid/saaprguidance.pdf).

Quality Assurance for DC CAS Psychometric Analyses: CTB's Research Process Upgrade (RPU) System

CTB first implemented its Research Process Upgrade (RPU) system for DC CAS in 2009. RPU provides standardized, automated procedures for data acquisition, cleaning, and confirmation; item analysis, calibration, and DIF analysis; and scaling, equating, and production of scoring tables. The standardization provides efficiencies that are necessary to meet the aggressive DC CAS schedules and quality assurance that CTB and the Office of the State Superintendent of Education (OSSE) continually strive for. RPU for the DC CAS program provides the following features.

Decision-Based Process Flow

Each step of the process of analyzing DC CAS data, from specifying data formats and psychometric analyses through delivering scoring tables to enable score reporting, is organized by specifying the operational objectives, research and psychometric questions, and decisions that must be made in each process. For example, RPU requires specification of item flagging criteria so that items that do not meet specifications can be flagged automatically for review by content specialists.

Standardized and Certified Software and Validation of Analyses

RPU standardizes all analyses and reports and implements formal validation steps for every analysis using accuracy-certified software. RPU runs validations automatically, which increases consistency and efficiency. The validation analyses produce diagnostics when concerns are identified so that DC CAS Research Scientists can determine appropriate courses of action.

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 2011 DC CAS. It also is organized to facilitate finding information easily. The report can be used in several ways:

- Read the report from front to back.
- Scan section headers and subheaders and read selected subsections.
- Locate specific topics in the table of contents (e.g., “Internal Consistency Reliability” in “Section 5. Evidence for Reliability and Validity”; “Section 7. IRT Analyses”).
- 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, 2009; 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 2011 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 are also reporting categories, applicable to all grades in Reading, Mathematics, Science/Biology, and Composition. These reporting categories have been re-articulated in Science and Biology for the 2011 administration, which means that while the text in the standards has been maintained, it has been reorganized under new content strand headings. The total numbers of operational items included in the Science/Biology test design remained the same, although they were distributed differently under the new strand headings. The numbers of items in the test design and blueprints for 2011 are presented in the next section.

Table 1. DC CAS 2011 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	Science and Technology	Items of this category measure students’ knowledge of scientific inquiry and the impacts of technology on society.
2	Earth and Space Science	Items of this category measure students’ knowledge of the solar system and the physical characteristics of Earth.
3	Physical Science	Items of this category measure students’ ability to describe how objects are affected by force, motion, and changes in temperature.
4	Life Science	Items of this category measure students’ basic understanding of the periodic table, force and motion, and heat transfer.

Science Grade 8		
Content Strand		**Description of Items**
1	Scientific Thinking and Inquiry	Items of this category measure students’ understanding and application of scientific design.
2	Matter and Reactions	Items of this category measure students’ basic understanding of the properties and structures of elements and chemical reactions.
3	Forces	Items of this category measure students’ understanding of the concepts of force and motion.
4	Energy and Waves	Items of this category measure students’ basic knowledge of energy and how it is transferred.

High School Biology		
Content Strand		Description of Items
1	Cell Biology and Biochemistry	Items of this category measure students' basic understanding of the chemistry of living things and knowledge of cell structures and functions.
2	Genetics and Evolution	Items of this category measure students' knowledge of genes, biodiversity, and the theory of evolution.
3	Multicellular Organisms— Plants and Animals	Items of this category measure students' knowledge of plant and animal biology.
4	Ecosystems	Items of this category measure students' knowledge of biotic and abiotic factors in 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.

2011 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 Office of Assessment, continued in 2011 the design established by tests of the DC CAS administered in 2006, 2007, 2008, 2009, and 2010. That design is represented in Table 2. Tables 3–5 display the blueprints for the operational item sets for the 2011 DC CAS in Reading, Mathematics, and Science/Biology.

Table 2. DC CAS 2011 Test Design: Reading, Mathematics, and Science/Biology

Content	Operational Items	Anchor Items (Operational Subset)	Embedded Field Test Items, Total across Two Forms
Reading	45 MC, 3 CR = 54 points	Gr. 3–8, 10: 23 MC, 1 CR	Gr. 3–7: 34 MC, 4 CR Gr. 8, 10: 36 MC, 4 CR
Math-ematics	51 MC, 3 CR = 60 points	Gr. 3–5, 7, 8: 25 MC Gr. 6, 10: 24 MC	Gr. 3–8, 10: 28 MC, 4 CR
Science/ Biology	47 MC, 3 CR = 53 points	Gr. 5, 8, Bio: 23 MC, 1 CR	Gr. 5, 8, Bio: 24 MC, 4 CR

Note. MC is Multiple-Choice and CR is Constructed-Response

2011 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 primarily selected operational items from the pool of operational and field test items from the 2010 administration of DC CAS. A small number of items also appeared in the 2006, 2007, and 2008 operational DC CAS.[1] A forms equating anchor set was selected for each grade/content area test using CTB's ItemWin (Burket, 2000) 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.

All proposed selections for operational forms were pre-equated in ItemWin to ensure that 2011 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 2011 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 2011 operational test forms. The 2011 test forms included two sets of field test items to expand the pool of items available for operational use in 2012 and beyond.

All operational items were reviewed for content standards alignment and appropriateness in previous years by CTB test developers and the DC Office of Assessment in order to be eligible for inclusion in 2011 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 2011 forms. DC Office of Assessment invited educators and community

[1] Items from the 2009 operational and field test administrations are no longer included in the DC CAS item pool because of concerns about exposure.

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 2011 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 Section 2, “Composition Test,” and Table 6. DC CAS 2011 Operational Test Form Scoring Rubrics: Composition.

Organization of Test Booklets and Other Test Materials

All students in public and charter schools in the District of Columbia took one of the two 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 two 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 and Grade 10 students were given mathematics reference cards. Students in Grades 7, 8, and 10 were allowed to use calculators in Mathematics 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 two 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 Composition form each was provided to each student in 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.

Table 3. DC CAS 2011 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	10	10	0	0	10	18.52%
	3	Informational Text	15	15	1	3	18	33.33%
	4	Literary Text	20	20	2	6	26	48.15%
		Total	45	45	3	9	54	100%
5	1	Language Development	10	10	0	0	10	18.52%
	3	Informational Text	15	15	1	3	18	33.33%
	4	Literary Text	20	20	2	6	26	48.15%
		Total	45	45	3	9	54	100%
6	1	Language Development	10	10	0	0	10	18.52%
	3	Informational Text	14	14	1	3	17	31.48%
	4	Literary Text	21	21	2	6	27	50.00%
		Total	45	45	3	9	54	100%
7	1	Language Development	10	10	0	0	10	18.52%
	3	Informational Text	16	16	1	3	19	35.19%
	4	Literary Text	19	19	2	6	25	46.30%
		Total	45	45	3	9	54	100%
8	1	Language Development	10	10	0	0	10	18.52%
	3	Informational Text	17	17	1	3	20	37.04%
	4	Literary Text	18	18	2	6	24	44.44%
		Total	45	45	3	9	54	100%
10	1	Language Development	9	9	0	0	9	16.67%
	3	Informational Text	17	17	1	3	20	37.04%
	4	Literary Text	19	19	2	6	25	46.30%
		Total	45	45	3	9	54	100%

Note. MC is Multiple-Choice and CR is Constructed-Response

Table 4. DC CAS 2011 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	12	12	0	0	12	20.00%
	3	Geometry	5	5	1	3	8	13.33%
	4	Measurement	8	8	0	0	8	13.33%
	5	Data Analysis, Statistics & Probability	10	10	1	3	13	21.67%
		Total	51	51	3	9	60	100%
4	1	Number Sense & Operations	19	19	0	0	19	31.67%
	2	Patterns, Relations & Algebra	11	11	0	0	11	18.33%
	3	Geometry	6	6	1	3	9	15.00%
	4	Measurement	4	4	1	3	7	11.67%
	5	Data Analysis, Statistics & Probability	11	11	1	3	14	23.33%
		Total	51	51	3	9	60	100%
5	1	Number Sense & Operations	19	19	0	0	19	31.67%
	2	Patterns, Relations & Algebra	11	11	1	3	14	23.33%
	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	15	15	1	3	18	30.00%
	2	Patterns, Relations & Algebra	13	13	1	3	16	26.67%
	3	Geometry	8	8	0	0	8	13.33%
	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	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%

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	7	7	1	3	10	16.67%
	4	Measurement	8	8	0	0	8	13.33%
	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 2011 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	Science and Technology	14	14	1	2	16	30.19%
	2	Earth and Space Science	12	12	1	2	14	26.42%
	3	Physical Science	10	10	0	0	10	18.87%
	4	Life Science	11	11	1	2	13	24.53%
		Total	47	47	3	6	53	100%
8	1	Scientific Thinking and Inquiry	8	8	1	2	10	18.87%
	2	Matter and Reactions	20	20	1	2	22	41.51%
	3	Forces	8	8	1	2	10	18.87%
	4	Energy and Waves	11	11	0	0	11	20.75%
		Total	47	47	3	6	53	100%
High School	1	Cell Biology & Biochemistry	13	13	1	2	15	28.30%
	2	Genetics and Evolution	15	15	1	2	17	32.08%
	3	Multicellular Organisms	11	11	0	0	11	20.75%
	4	Ecosystems	8	8	1	2	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 Development using a six-point rubric and once for use of English Language Conventions using a four-point rubric. The rubrics used to score these essays can be found in Appendix B. Scorable essays are assigned scores 1-6 for Topic/Idea Development and 1-4 for Standard English Conventions. Non-scorable responses are assigned a condition code (e.g., off topic, response not in English) and are not assigned a numerical score. Student scores on the two rubrics are summed so that total Composition scores range from 2 to 10.

Table 6. DC CAS 2011 Operational Test Form Scoring Rubrics: Composition

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

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 4 and April 14, 2011.

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 high school 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 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 language learners (ELLs) who participate in DC CAS grade level administrations may be provided approved test administration accommodations that are specified by special education IEP teams, Section 504 teams, or ELL teams.

Participation in the 2011 DC CAS Test Administrations and Use of Data for Analysis and Score Reporting

Approximately 4,500 students were assessed in Reading and Mathematics at each tested grade, with slightly fewer in each tested grade of Composition, and Science/Biology. We report information below about participating students at each grade, numbers of examinees in special programs, and numbers of examinees in special education and ELL programs who received test administration accommodations. Only students with a valid test administration as required by the type of analysis, as defined below, are included in the reports.

Definition of *Valid Test Administration*

In this technical report, two sets of rules are used to define a valid test administration. The first set of rules is for psychometric analyses included in this report (e.g., reliability, DIF, item parameter calibration, and equating), answer documents are excluded when any of the following conditions are observed:

- 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.

The second set of valid test administration rules are for analyses summarizing test performance (e.g., overall numbers of examinees, descriptive statistics, and correlations of test scores). All students who have a valid test score, as defined in the DC CAS Spring 2011 Business Requirements, are included in these analyses. For the Reading, Mathematics, Science and Biology assessments, the requirements document outlines a valid attempt on the test as:

- At least one item marked with a correct response

OR

- At least 5 items validly marked in the content area

And for Composition, a valid attempt is defined as:

- A score of non-zero on both parts of the item

Note: To maintain confidentiality of individual student results, this report does not show subgroup results for fewer than 25 students. The race/ethnicity subgroups Native Hawaiian/Pacific Islander and American Indian/Alaska Native contain fewer than 25 students per grade and are not shown in the following tables.

Table 7. Numbers of Examinees with Valid Test Administrations in 2011: Reading

Grade	Students with Test Scores	Males	Females	Asian	African American	Hispanic	White
3	4,796	2,458	2,320	103	3,523	673	470
4	4,841	2,427	2,389	74	3,748	605	388
5	4,797	2,417	2,366	76	3,764	607	334
6	4,403	2,228	2,162	40	3,582	505	254
7	4,456	2,220	2,212	57	3,670	464	245
8	4,327	2,156	2,152	58	3,616	417	213
10	4,491	2,111	2,307	55	3,743	467	163

Table 8. Numbers of Examinees with Valid Test Administrations in 2011: Mathematics

Grade	Students with Test Scores	Males	Females	Asian	African American	Hispanic	White
3	4,823	2,470	2,335	108	3,524	690	474
4	4,873	2,442	2,405	84	3,752	616	394
5	4,817	2,430	2,373	78	3,764	621	337
6	4,433	2,244	2,176	46	3,591	517	255
7	4,485	2,236	2,225	60	3,673	485	247
8	4,370	2,181	2,170	66	3,619	449	213
10	4,464	2,098	2,297	54	3,722	464	161

Table 9. Numbers of Examinees with Valid Test Administrations in 2011: Science/Biology

Grade	Students with Test Scores	Males	Females	Asian	African American	Hispanic	White
5	4,765	2,401	2,348	78	3,729	609	332
8	4,223	2,081	2,100	66	3,475	441	203
High School	3,790	1,757	1,952	49	3,166	436	100

Table 10. Numbers of Examinees with Valid Test Administrations in 2011: Composition

Grade	Students with Test Scores	Males	Females	Asian	African American	Hispanic	White
4	4,755	2,373	2,356	75	3,672	595	386
7	4,301	2,126	2,149	54	3,528	453	241
10	3,761	1,723	1,978	52	3,110	396	152

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 ELL 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 2011 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	4,826	446 (9%)	426 (9%)	16 (0%)	315 (7%)	3 (0%)
4	4,879	536 (11%)	318 (7%)	28 (1%)	297 (6%)	1 (0%)
5	4,819	589 (12%)	298 (6%)	29 (1%)	279 (6%)	0 (0%)
6	4,435	609 (14%)	214 (5%)	25 (1%)	121 (3%)	0 (0%)
7	4,487	657 (15%)	247 (6%)	26 (1%)	149 (3%)	0 (0%)
8	4,376	696 (16%)	229 (5%)	29 (1%)	160 (4%)	1 (0%)
10	4,508	621 (14%)	184 (4%)	19 (0%)	107 (2%)	0 (0%)
Science/Biology						
5	4,765	562 (12%)	282 (6%)	27 (1%)	301 (6%)	0 (0%)
8	4,223	600 (14%)	205 (5%)	21 (0%)	154 (4%)	0 (0%)
High School	3,790	458 (12%)	150 (4%)	11 (0%)	106 (3%)	1 (0%)
Composition						
4	4,755	492 (10%)	263 (6%)	25 (1%)	289 (6%)	0 (0%)
7	4,301	531 (12%)	194 (5%)	19 (0%)	143 (3%)	0 (0%)
10	3,761	420 (11%)	134 (4%)	12 (0%)	101 (3%)	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 Code
Reading and/or Mathematics							
3	4,826	481 (10%)	520 (11%)	471 (10%)	271 (6%)	22 (0%)	446 (9%)
4	4,879	633 (13%)	627 (13%)	613 (13%)	375 (8%)	10 (0%)	536 (11%)
5	4,819	654 (14%)	663 (14%)	623 (13%)	389 (8%)	16 (0%)	589 (12%)
6	4,435	684 (15%)	671 (15%)	643 (14%)	505 (11%)	14 (0%)	609 (14%)
7	4,487	634 (14%)	624 (14%)	616 (14%)	542 (12%)	11 (0%)	657 (15%)
8	4,376	701 (16%)	678 (15%)	620 (14%)	573 (13%)	11 (0%)	696 (16%)
10	4,508	604 (13%)	573 (13%)	455 (10%)	540 (12%)	19 (0%)	621 (14%)
Science/Biology							
5	4,765	632 (13%)	634 (13%)	594 (12%)	351 (7%)	16 (0%)	562 (12%)
8	4,223	567 (13%)	562 (13%)	508 (12%)	410 (10%)	6 (0%)	600 (14%)
High School	3,790	444 (12%)	419 (11%)	338 (9%)	313 (8%)	29 (1%)	458 (12%)
Composition							
4	4,755	554 (12%)	550 (12%)	544 (11%)	271 (6%)	6 (0%)	492 (10%)
7	4,301	544 (13%)	543 (13%)	521 (12%)	380 (9%)	16 (0%)	531 (12%)
10	3,761	437 (12%)	423 (11%)	349 (9%)	296 (8%)	10 (0%)	420 (11%)

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 to students 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 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	4,826	428 (9%)	502 (10%)	343 (7%)	95 (2%)
4	4,879	541 (11%)	607 (12%)	473 (10%)	79 (2%)
5	4,819	561 (12%)	649 (13%)	459 (10%)	82 (2%)
6	4,435	558 (13%)	654 (15%)	464 (10%)	87 (2%)
7	4,487	518 (12%)	617 (14%)	420 (9%)	51 (1%)
8	4,376	568 (13%)	658 (15%)	389 (9%)	40 (1%)
10	4,508	471 (10%)	540 (12%)	191 (4%)	53 (1%)
Science/Biology					
5	4,765	539 (11%)	622 (13%)	423 (9%)	86 (2%)
8	4,223	469 (11%)	544 (13%)	355 (8%)	33 (1%)
High School	3,790	344 (9%)	398 (11%)	184 (5%)	45 (1%)
Composition					
4	4,755	464 (10%)	534 (11%)	401 (8%)	66 (1%)
7	4,301	445 (10%)	535 (12%)	344 (8%)	46 (1%)
10	3,761	332 (9%)	397 (11%)	156 (4%)	45 (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. Tables 14-16 display 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	4,826	390 (8%)	108 (2%)	408 (8%)	7 (0%)
4	4,879	305 (6%)	113 (2%)	323 (7%)	6 (0%)
5	4,819	258 (5%)	135 (3%)	270 (6%)	2 (0%)
6	4,435	199 (4%)	93 (2%)	209 (5%)	1 (0%)
7	4,487	219 (5%)	151 (3%)	233 (5%)	2 (0%)
8	4,376	200 (5%)	140 (3%)	214 (5%)	2 (0%)
10	4,508	176 (4%)	161 (4%)	181 (4%)	3 (0%)
Science/Biology					
5	4,765	238 (5%)	114 (2%)	262 (5%)	1 (0%)
8	4,223	189 (4%)	131 (3%)	197 (5%)	2 (0%)
High School	3,790	123 (3%)	112 (3%)	137 (4%)	11 (0%)
Composition					
4	4,755	272 (6%)	120 (3%)	283 (6%)	6 (0%)
7	4,301	183 (4%)	126 (3%)	196 (5%)	1 (0%)
10	3,761	103 (3%)	126 (3%)	108 (3%)	5 (0%)

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 to students 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 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	4,826	45 (1%)	51 (1%)	188 (4%)	181 (4%)
4	4,879	43 (1%)	38 (1%)	106 (2%)	168 (3%)
5	4,819	33 (1%)	42 (1%)	100 (2%)	121 (3%)
6	4,435	37 (1%)	31 (1%)	67 (2%)	82 (2%)
7	4,487	48 (1%)	38 (1%)	83 (2%)	91 (2%)
8	4,376	79 (2%)	41 (1%)	74 (2%)	56 (1%)
10	4,508	9 (0%)	26 (1%)	82 (2%)	79 (2%)
Science/Biology					
5	4,765	29 (1%)	37 (1%)	93 (2%)	115 (2%)
8	4,223	60 (1%)	35 (1%)	57 (1%)	46 (1%)
High School	3,790	11 (0%)	44 (1%)	37 (1%)	46 (1%)
Composition					
4	4,755	9 (0%)	24 (1%)	99 (2%)	148 (3%)
7	4,301	19 (0%)	35 (1%)	75 (2%)	66 (2%)
10	3,761	2 (0%)	12 (0%)	52 (1%)	58 (2%)

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	4,826	48 (1%)	96 (2%)	390 (8%)
4	4,879	38 (1%)	82 (2%)	308 (6%)
5	4,819	35 (1%)	108 (2%)	243 (5%)
6	4,435	36 (1%)	84 (2%)	191 (4%)
7	4,487	45 (1%)	111 (2%)	219 (5%)
8	4,376	57 (1%)	115 (3%)	202 (5%)
10	4,508	5 (0%)	148 (3%)	171 (4%)
Science/Biology				
5	4,765	43 (1%)	90 (2%)	244 (5%)
8	4,223	54 (1%)	105 (2%)	190 (4%)
High School	3,790	11 (0%)	89 (2%)	134 (4%)
Composition				
4	4,755	27 (1%)	48 (1%)	229 (5%)
7	4,301	25 (1%)	36 (1%)	184 (4%)
10	3,761	8 (0%)	112 (3%)	105 (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 2011 *Test Chairperson's Manual*. They, in turn, trained all test administrators and proctors. Test administrators administered all DC CAS assessments according to requirements and steps in the 2011 *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 Test 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 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 [only] Mathematics 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 next in this section). Corresponding security checklists were also produced. Daily tracking reports were provided to the 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 the 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 for DC CAS, 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 members 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 nonscorable 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 shipped security barcodes 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 27, 2011, approximately 99.86% of secure materials were accounted for; only 103 secure test booklets were missing for the 2011 administration, compared to 106 test booklets missing in 2010.

Section 5. Evidence for Reliability and Validity

The *Standards and Assessment Peer Review Guidance* (dated January 12, 2009) 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 2011 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 2011 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 multiple-choice and constructed-response 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 0.80 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 2011 DC CAS Operational Tests

Content	Grade	Students with Test Scores	Number of Items	Alpha	Stratified Alpha	Feldt-Raju
Reading	3	4,773	48	0.927	0.933	0.932
	4	4,817	48	0.923	0.927	0.927
	5	4,791	48	0.928	0.931	0.931
	6	4,393	48	0.920	0.922	0.922
	7	4,440	48	0.912	0.917	0.915
	8	4,310	48	0.910	0.918	0.916
	10	4,442	48	0.924	0.931	0.930
Mathematics	3	4,805	54	0.934	0.940	0.941
	4	4,858	54	0.923	0.929	0.930
	5	4,812	54	0.929	0.932	0.934
	6	4,423	54	0.929	0.936	0.936
	7	4,458	54	0.920	0.924	0.925
	8	4,354	54	0.916	0.922	0.921
	10	4,415	54	0.912	0.914	0.915
Science	5	4,764	50	0.889	0.890	0.891
	8	4,213	50	0.876	0.880	0.880
Biology	High School	3,760	50	0.856	0.858	0.858

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 are, on average, 0.92. This is strong evidence for the reliability of scores for the Reading, Mathematics, Science, and Biology tests. The lowest reliability was in Biology (0.86).

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 2011 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 23, range between 0.51 and 0.87. 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	0.7426
	3	Informational Text	17	0.8416
	4	Literary Text	20	0.8318
	Total Number of Items on DC CAS		48	--
4	1	Language Development	10	0.7163
	3	Informational Text	16	0.8065
	4	Literary Text	22	0.8427
	Total Number of Items on DC CAS		48	--
5	1	Language Development	10	0.6970
	3	Informational Text	16	0.8118
	4	Literary Text	22	0.8654
	Total Number of Items on DC CAS		48	--
6	1	Language Development	10	0.7573
	3	Informational Text	15	0.7570
	4	Literary Text	23	0.8517
	Total Number of Items on DC CAS		48	--
7	1	Language Development	10	0.6940
	3	Informational Text	17	0.7906
	4	Literary Text	21	0.8184
	Total Number of Items on DC CAS		48	--
8	1	Language Development	10	0.6338
	3	Informational Text	18	0.8116
	4	Literary Text	20	0.8103
	Total Number of Items on DC CAS		48	--
10	1	Language Development	9	0.6968
	3	Informational Text	18	0.8404
	4	Literary Text	21	0.8296
	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	0.8087
	2	Patterns, Relations & Algebra	12	0.7925
	3	Geometry	6	0.5110
	4	Measurement	8	0.6929
	5	Data Analysis, Statistics & Probability	11	0.8002
	Total	Number of Items on DC CAS	54	--
4	1	Number Sense & Operations	19	0.8107
	2	Patterns, Relations & Algebra	11	0.7528
	3	Geometry	7	0.5724
	4	Measurement	5	0.5378
	5	Data Analysis, Statistics & Probability	12	0.6902
	Total	Number of Items on DC CAS	54	--
5	1	Number Sense & Operations	19	0.8232
	2	Patterns, Relations & Algebra	12	0.7288
	3	Geometry	9	0.6470
	4	Measurement	7	0.6700
	5	Data Analysis, Statistics & Probability	7	0.6966
	Total	Number of Items on DC CAS	54	--
6	1	Number Sense & Operations	16	0.8141
	2	Patterns, Relations & Algebra	14	0.7867
	3	Geometry	8	0.5799
	4	Measurement	8	0.6219
	5	Data Analysis, Statistics & Probability	8	0.7146
	Total	Number of Items on DC CAS	54	--
7	1	Number Sense & Operations	17	0.8137
	2	Patterns, Relations & Algebra	14	0.7161
	3	Geometry	9	0.6748
	4	Measurement	6	0.6179
	5	Data Analysis, Statistics & Probability	8	0.6222
	Total	Number of Items on DC CAS	54	--
8	1	Number Sense & Operations	17	0.7549
	2	Patterns, Relations & Algebra	14	0.7504
	3	Geometry	9	0.5749
	4	Measurement	6	0.5977
	5	Data Analysis, Statistics & Probability	8	0.6507
	Total	Number of Items on DC CAS	54	--
10	1	Number Sense & Operations	10	0.6619
	2	Patterns, Relations & Algebra	15	0.8056
	3	Geometry	8	0.5909
	4	Measurement	8	0.5429
	5	Data Analysis, Statistics & Probability	13	0.6914
	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	Science and Technology	15	0.6695
	2	Earth and Space Science	13	0.7180
	3	Physical Science	10	0.6559
	4	Life Science	12	0.6554
	Total	Number of Items on DC CAS	50	--
8	1	Scientific Thinking and Inquiry	9	0.6398
	2	Matter and Reactions	21	0.7273
	3	Forces	9	0.6257
	4	Energy and Waves	11	0.5505
	Total	Number of Items on DC CAS	50	--
High School	1	Cell Biology & Biochemistry	14	0.5883
	2	Genetics and Evolution	16	0.6939
	3	Multicellular Organisms	11	0.5844
	4	Ecosystems	9	0.5410
	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 2011 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	35	500	35	600	37	700	35	800	38	900	35
1	300	30	400	35	500	35	600	37	700	35	800	38	900	35
2	300	30	400	35	500	35	600	37	700	35	800	38	900	35
3	300	30	400	35	500	35	600	37	700	35	800	38	900	35
4	300	30	400	35	500	35	600	37	700	35	800	38	900	35
5	300	30	400	35	500	35	600	37	700	35	800	38	900	35
6	300	30	400	35	500	35	600	37	700	35	800	38	900	35
7	300	30	400	35	500	35	600	37	700	35	800	38	900	35
8	300	30	400	35	500	35	600	37	700	35	800	38	900	35
9	300	30	400	35	500	35	600	37	700	35	800	38	900	35
10	300	30	412	23	507	28	600	37	703	32	812	26	906	29
11	306	24	420	15	519	17	613	24	715	21	822	17	917	18
12	314	16	424	11	524	12	620	16	721	15	826	12	922	13
13	319	11	427	8	527	8	625	12	725	11	830	9	926	9
14	322	9	429	7	529	7	628	9	728	9	832	7	929	8
15	325	7	432	6	531	6	631	7	731	7	835	6	931	7
16	327	6	433	5	533	5	633	6	733	6	836	6	933	6
17	329	6	435	5	534	4	635	5	735	6	838	5	935	5
18	330	5	436	5	536	4	636	5	736	5	840	5	936	5
19	332	5	438	4	537	4	638	4	738	5	841	5	937	5
20	333	4	439	4	538	4	639	4	739	5	842	4	939	4
21	334	4	440	4	539	3	640	4	740	4	843	4	940	4
22	336	4	441	4	540	3	641	4	742	4	845	4	941	4
23	337	4	442	4	541	3	642	3	743	4	846	4	942	4
24	338	4	443	4	542	3	643	3	744	4	847	4	943	4
25	339	4	444	4	543	3	644	3	745	4	848	4	944	4
26	340	3	445	3	544	3	645	3	746	4	849	4	945	4
27	341	3	446	3	545	3	646	3	747	3	850	3	946	3
28	342	3	447	3	546	3	647	3	748	3	851	3	947	3
29	343	3	448	3	546	3	647	3	749	3	852	3	948	3
30	344	3	449	3	547	3	648	3	750	3	853	3	949	3
31	345	3	450	3	548	3	649	3	751	3	854	3	950	3
32	345	3	451	3	549	3	650	3	752	3	855	3	951	3
33	346	3	452	3	550	3	651	3	753	3	856	3	952	3
34	347	3	453	3	551	3	651	3	754	3	857	3	953	3
35	348	3	454	3	552	3	652	3	755	3	858	3	954	3
36	349	3	455	3	553	3	653	3	755	3	859	3	955	3
37	350	3	456	3	554	3	654	3	756	3	860	3	956	3
38	351	3	457	3	555	3	655	3	757	3	861	3	957	3
39	352	3	458	3	556	3	656	3	758	3	862	3	958	3
40	354	4	459	3	557	3	657	3	759	3	863	3	959	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
41	355	4	460	3	558	3	658	3	760	3	865	3	960	3
42	356	4	462	4	559	4	659	3	761	3	866	4	961	3
43	357	4	463	4	561	4	661	3	763	3	867	4	963	4
44	359	4	464	4	562	4	662	4	764	3	869	4	964	4
45	360	4	466	4	564	4	663	4	765	4	870	4	965	4
46	362	5	468	4	566	4	665	4	767	4	872	4	967	4
47	364	5	470	5	568	5	667	5	768	4	873	4	969	4
48	366	5	472	5	571	5	669	5	770	5	875	4	970	4
49	369	6	475	6	573	5	672	6	773	5	877	5	973	5
50	372	6	478	7	576	6	676	7	776	6	880	5	975	5
51	375	7	482	8	580	7	681	9	780	7	883	6	978	6
52	381	9	489	10	586	9	688	11	785	10	888	7	983	8
53	390	14	499	15	597	15	699	16	796	14	895	11	991	12
54	399	19	499	15	599	16	699	16	799	16	899	13	999	17

Table 22. DC CAS 2011 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	21	400	36	500	31	600	35	700	33	800	42	900	35
1	300	21	400	36	500	31	600	35	700	33	800	42	900	35
2	300	21	400	36	500	31	600	35	700	33	800	42	900	35
3	300	21	400	36	500	31	600	35	700	33	800	42	900	35
4	300	21	400	36	500	31	600	35	700	33	800	42	900	35
5	300	21	400	36	500	31	600	35	700	33	800	42	900	35
6	300	21	400	36	500	31	600	35	700	33	800	42	900	35
7	300	21	400	36	500	31	600	35	700	33	800	42	900	35
8	300	21	400	36	500	31	600	35	700	33	800	42	900	35
9	300	21	400	36	500	31	600	35	700	33	800	42	900	35
10	300	21	400	36	500	31	600	35	700	33	800	42	900	35
11	300	21	400	36	501	30	612	23	700	33	800	42	905	30
12	300	21	404	31	511	20	620	16	708	25	812	30	914	21
13	304	17	415	21	516	15	624	11	716	17	823	19	919	15
14	309	13	421	15	520	11	627	9	721	13	829	13	924	12
15	313	11	425	12	524	10	630	8	724	10	832	10	927	10
16	316	9	428	10	526	8	632	7	727	8	835	8	930	9
17	318	8	431	8	529	7	634	6	729	7	837	7	932	8
18	321	8	433	7	531	7	636	5	731	7	839	6	934	7
19	323	7	435	7	533	6	637	5	733	6	840	5	936	7
20	325	6	437	6	534	6	638	5	735	6	842	5	938	6
21	327	6	439	6	536	5	640	4	737	6	843	5	940	6
22	329	6	440	5	538	5	641	4	738	5	844	4	941	5
23	330	5	442	5	539	5	642	4	740	5	846	4	943	5
24	332	5	443	5	541	5	643	4	741	5	847	4	944	5
25	333	5	444	4	542	5	644	4	742	5	848	4	945	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
26	335	5	446	4	543	4	645	4	744	4	849	4	947	5
27	336	5	447	4	544	4	646	4	745	4	850	3	948	4
28	337	4	448	4	546	4	647	3	746	4	851	3	949	4
29	339	4	449	4	547	4	648	3	747	4	852	3	950	4
30	340	4	450	4	548	4	649	3	748	4	852	3	951	4
31	341	4	451	4	549	4	650	3	749	4	853	3	952	4
32	342	4	452	3	550	4	651	3	751	4	854	3	953	4
33	343	4	453	3	551	4	652	3	752	4	855	3	955	4
34	344	4	454	3	552	4	652	3	753	4	856	3	956	4
35	346	4	455	3	553	3	653	3	754	4	857	3	957	4
36	347	4	456	3	554	3	654	3	755	4	857	3	958	4
37	348	4	456	3	555	3	655	3	756	4	858	3	959	4
38	349	4	457	3	556	3	656	3	757	3	859	3	960	4
39	350	4	458	3	557	3	657	3	758	3	860	3	961	4
40	351	4	459	3	558	3	658	3	759	3	861	3	962	4
41	352	4	460	3	559	3	659	3	760	3	861	3	963	4
42	353	4	461	3	560	3	659	3	761	4	862	3	964	4
43	354	4	462	3	561	3	660	3	762	4	863	3	965	4
44	356	4	463	3	562	3	661	3	763	4	864	3	966	4
45	357	4	464	3	563	3	662	3	764	4	865	3	967	4
46	358	4	465	3	564	3	663	3	765	4	866	3	969	4
47	359	4	466	3	565	3	664	3	767	4	867	3	970	4
48	361	4	467	3	566	3	666	3	768	4	868	3	971	4
49	362	4	468	4	568	4	667	3	769	4	869	3	973	4
50	364	4	469	4	569	4	668	4	771	4	870	3	974	4
51	365	5	471	4	570	4	669	4	772	4	871	3	976	4
52	367	5	472	4	572	4	671	4	774	4	873	4	977	5
53	369	5	474	4	573	4	672	4	776	4	874	4	979	5
54	371	5	475	4	575	5	674	4	778	5	876	4	982	5
55	373	6	477	5	577	5	676	5	780	5	878	5	984	6
56	376	6	480	5	580	6	678	5	783	6	881	5	988	7
57	380	7	483	6	583	6	682	6	787	7	884	6	993	9
58	385	9	487	7	588	8	686	8	793	10	888	7	999	12
59	394	14	494	11	596	12	695	12	799	13	896	11	999	12
60	399	17	499	14	599	15	699	15	799	13	899	12	999	12

Table 23. DC CAS 2011 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	44	800	51	900	49
1	500	44	800	51	900	49
2	500	44	800	51	900	49
3	500	44	800	51	900	49
4	500	44	800	51	900	49
5	500	44	800	51	900	49
6	500	44	800	51	900	49
7	500	44	800	51	900	49
8	500	44	800	51	900	49
9	500	44	800	51	900	49
10	515	29	800	51	928	22
11	527	17	830	21	936	13
12	532	12	837	14	940	9
13	535	9	841	10	943	7
14	538	7	844	8	945	6
15	540	6	846	6	947	5
16	541	5	848	5	948	4
17	543	4	849	5	949	4
18	544	4	851	4	950	4
19	545	4	852	4	952	3
20	546	4	853	3	952	3
21	547	3	854	3	953	3
22	548	3	855	3	954	3
23	549	3	855	3	955	2
24	550	3	856	3	956	2
25	551	3	857	3	956	2
26	552	3	858	3	957	2
27	553	3	859	2	958	2
28	554	3	859	2	958	2
29	554	3	860	2	959	2
30	555	2	861	2	959	2
31	556	2	861	2	960	2
32	557	2	862	2	960	2
33	557	2	862	2	961	2
34	558	2	863	2	961	2
35	559	2	864	2	962	2
36	560	2	864	2	963	2
37	560	2	865	2	963	2
38	561	2	866	2	964	2
39	562	2	866	2	964	2
40	563	2	867	2	965	2
41	563	2	868	2	966	2
42	564	2	869	2	966	2
43	565	2	870	2	967	2
44	566	2	870	2	968	2
45	567	3	871	3	969	2

Raw Score	Grade 5		Grade 8		High School	
	Scale Score	SEM	Scale Score	SEM	Scale Score	SEM
46	568	3	873	3	970	2
47	570	3	874	3	971	3
48	571	3	875	3	972	3
49	573	4	877	4	974	4
50	576	5	880	5	977	5
51	580	6	884	7	981	7
52	586	9	892	12	990	14
53	599	19	899	18	999	23

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.

CTB used the KKCLASS program (Kim, 2007) to calculate these statistics on the 2011 DC CAS results. The KKCLASS program implements an IRT-based procedure that is consistent with DC CAS IRT scaling and scoring. The procedure is described below.

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 2011 DC CAS in Reading, Mathematics, and Science/Biology. As can be seen in these tables below, the classification consistency results range from 0.66 to 0.78 in all content areas and grades. The results are comparable to those in 2010, which ranged between 0.66 and 0.77. Kappa coefficients range between 0.48 and 0.68, which is comparable to the 2010 results (0.49 to 0.68). 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 2011 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 2010, which also ranged between 0.73 and 0.84. These results suggest that the 2011 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 2011 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

	Classification Consistency		Classification Accuracy		
Grade	Consistency	Kappa	Accuracy	False Positive Errors	False Negative Errors
3	0.7787	0.6764	0.8408	0.0533	0.1059
4	0.7591	0.6524	0.8271	0.0623	0.1106
5	0.7671	0.6584	0.8298	0.0569	0.1133
6	0.7702	0.6560	0.8327	0.0548	0.1125
7	0.7414	0.6248	0.8170	0.0817	0.1013
8	0.7433	0.6314	0.8164	0.0705	0.1131
10	0.7502	0.6471	0.8204	0.0654	0.1142

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

	Classification Consistency		Classification Accuracy		
Grade	Consistency	Kappa	Accuracy	False Positive Errors	False Negative Errors
3	0.7669	0.6705	0.8303	0.0637	0.1060
4	0.7532	0.6535	0.8235	0.0568	0.1197
5	0.7718	0.6800	0.8297	0.0570	0.1133
6	0.7684	0.6714	0.8292	0.0590	0.1119
7	0.7579	0.6520	0.8263	0.0808	0.0928
8	0.7379	0.6146	0.8123	0.0798	0.1079
10	0.7425	0.6327	0.8097	0.0960	0.0943

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

	Classification Consistency		Classification Accuracy		
Grade	Consistency	Kappa	Accuracy	False Positive Errors	False Negative Errors
5	0.7140	0.5824	0.7940	0.0967	0.1093
8	0.6815	0.5419	0.7722	0.0860	0.1418
High School	0.6578	0.4838	0.7282	0.1088	0.1630

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 the comparison 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 2011 administration were female and male students and African American, Asian, Hispanic, and White students. Comparing these subgroups in DIF analyses is conventional practice in the United States. Male and African American students were the reference groups. Selecting males as the reference group is conventional practice in the United States. African American students are selected as the reference group for DC CAS because they are the largest subgroup enrolled in District of Columbia schools.

Items flagged for DIF may or may not provide an unfair advantage or disadvantage for one examinee subgroup compared to 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 obviously sensitive terms, phrases, scenarios, or illustrations that could influence examinee performance appear in DC CAS items prior to field testing and selection for operational test forms. OSSE and CTB then screen all field tested items that are flagged for DIF after each administration to identify items that may favor or disadvantage examinee subgroups. Items that are flagged are rarely disqualified from entering the item pool for operational use, 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 also are conducted on operational items. Results for the Reading, Mathematics, and Science/Biology assessments are reported below. 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 multiple-choice 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 $-1.5 \leq \Delta_{MH} \leq -1$ or $1 \leq \Delta_{MH} \leq 1.5$.

- 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 gender and race/ethnicity. 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 2011 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 2010. 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 of 126 Reading item flags for DIF[2] represents 13.2% of the 957 flagging opportunities[3] in Reading; the total of 108 item flags in Mathematics for DIF represents 10% of the 1,080 flagging opportunities in Mathematics; and the total of 18 item flags in Science/Biology for DIF represent 5.1% of the 356 flagging opportunities.

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

[2] In the following tables, the “total items flagged for DIF” is the total number of items with DIF statistics shown in columns B, B-, C, and C- and summed across all DIF comparisons and grades listed within a content area. Items listed in column A are excluded as, statistically, a flag of “A” indicates no DIF.

[3] ”Flagging opportunities” are the total number of items with DIF statistics that were examined for DIF in a content area. In the following tables, this is the total of all items shown in columns A, B, B-, C, and C- and summed across all DIF comparisons and grades listed within a content area.

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	N/A	N/A	N/A	N/A	N/A
	Hispanic	48	0	0	0	0
	White	32	6	0	9	1
Grade 4 (total 48 items)						
Male	Female	47	1	0	0	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	46	0	2	0	0
	White	36	7	0	5	0
Grade 5 (total 48 items)						
Male	Female	47	0	1	0	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	41	2	4	0	1
	White [1]	36	5	0	6	0
Grade 6 (total 48 items)						
Male	Female	45	2	1	0	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	44	2	1	0	1
	White	34	5	1	8	0
Grade 7 (total 48 items)						
Male	Female	46	0	1	0	1
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	46	1	0	0	1
	White	31	3	2	12	0
Grade 8 (total 48 items)						
Male	Female	45	1	1	1	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	46	1	1	0	0
	White[1]	31	5	1	9	0
Grade 10 (total 48 items)						
Male	Female	43	2	2	1	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	39	3	2	2	2
	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 grades 5 & 8, no matching pairs of reference and focal group examinees were found for some total test scores for item 14 in grade 5, and items 6 and 27 in grade 8. As a result, DIF statistics were not calculated for these items.

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	53	1	0	0	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	53	0	0	1	0
	White	41	5	1	5	2
Grade 4 (total 54 items)						
Male	Female	53	1	0	0	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	53	1	0	0	0
	White	41	2	2	7	2
Grade 5 (total 54 items)						
Male	Female	53	1	0	0	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	54	0	0	0	0
	White	38	2	4	9	1
Grade 6 (total 54 items)						
Male	Female	50	1	1	1	1
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	52	1	1	0	0
	White	35	9	2	6	2
Grade 7 (total 54 items)						
Male	Female	53	1	0	0	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	52	2	0	0	0
	White	37	7	2	5	3
Grade 8 (total 54 items)						
Male	Female	53	0	1	0	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	52	0	2	0	0
	White	43	4	2	4	1
Grade 10 (total 54 items)						
Male	Female	53	0	1	0	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	53	1	0	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	50	0	0	0	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	49	1	0	0	0
	White	38	6	0	6	0
Grade 8 (total 50 items)						
Male	Female	50	0	0	0	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	50	0	0	0	0
	White[1]	4	1	1	0	0
High School (total 50 items)						
Male	Female	48	1	0	1	0
African American	Asian	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

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.

In Science grade 8 forty-four items did not meet the focal group N count criterion, therefore DIF statistics were not calculated for these items.

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

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 2011 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 2011 as had been used when they were scored as field test items.

Preparation and Meeting Logistics for Rangefinding Prior to 2011 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 participants became familiar with the items and the criteria that raters 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 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 a DC participant provided sign-off that consensus on the scoring of the items was achieved.

Training and Qualifying Procedures in 2011

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 CTB content experts who supervised the training of the team leaders also supervised the training of the raters.

Breakdown of Scoring Teams

Four CTB content experts oversaw the training and scoring of the constructed-response items for 2011 in Reading, Mathematics, Science/Biology, and Writing: one expert for Reading, one for Mathematics, and one each for Science/Biology and Writing. Each of these four content experts was responsible for training and scoring all of the items in his or her content area.

Teams of between 8 and 13 raters (depending on the content and grade) trained on and scored all the operational items at their respective grades, 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 April 26, 2011, through May 6, 2011.

Across all grades, 21 unique constructed-response items were scored 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-response items in each test, for a total of 9 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 anchor papers for each item. Raters then took qualifying rounds, 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 6% of books that the raters received were “checkset” papers rather than live books. 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. The CTB 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 involved a one-on-one discussion between the supervisor (or a team leader) and the rater, who discussed 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 2011, 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 the ideas rather than the specific written expressions 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 content area composed of discrete knowledge. This rate of perfect agreement drops rapidly, however, for a content 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 2011 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 content areas.

In Reading, the average perfect agreement was 65%, with a high of 82% and a low of 56%. For perfect and adjacent agreement, the average was 95%, with a high of 99% and a low of 87%. In Mathematics, the average perfect agreement was 84%, with a high of 96% and a low of 64%. For perfect and adjacent agreement, the average was 98%, with a high of 100% and a low of 92%. In Science/Biology, the average perfect agreement rate was 85%, with a high of 95% and a low of 73%. For perfect and adjacent agreement, the average was 98%, with a high of 99% and a low of 97%. In Composition, the average perfect agreement was 58%, with a high of 64% and a low of 51% agreement. For perfect and adjacent agreement, the average was 96%, with a high of 99% and a low of 95%. 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 2011 Writing Prompts

The 2011 Writing prompts for Grades 4, 7, and 10, were all from the 2006 test administration.

Prior to their initial scoring as field test items, the prompts underwent extensive rangefinding 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 during the operational test administration in 2011 as had been used in 2006 when these prompts were originally scored as field test items.

Table 31. DC CAS 2011 Operational Inter-Rater Agreement for Constructed-Response Items: Reading

				% of Agreement			
Grade	Form	Item No.	Score Points	Perfect	Adjacent	Perfect + Adjacent	Checkset Average Agreement Percentages
3	1-2	9	0-3	73	23	96	84
		20	0-3	71	27	97	65
		45	0-3	60	34	94	79
4	1-2	9	0-3	63	29	92	75
		18	0-3	66	32	98	83
		45	0-3	65	34	99	72
5	1-2	9	0-3	58	29	87	83
		19	0-3	73	22	95	77
		45	0-3	68	30	98	71
6	1-2	8	0-3	67	31	98	84
		18	0-3	82	16	98	90
		28	0-3	56	40	96	65
7	1-2	7	0-3	60	35	95	66
		19	0-3	72	25	96	72
		45	0-3	60	33	92	66
8	1-2	8	0-3	66	31	97	67
		19	0-3	64	33	97	75
		52	0-3	56	39	94	70
10	1-2	7	0-3	60	33	93	60
		19	0-3	69	27	97	74
		52	0-3	60	36	95	72

***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 2011 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	1-2	27	0-3	96	4	100	99
		45	0-3	70	28	99	78
		61	0-3	86	13	100	78
4	1-2	28	0-3	94	6	100	95
		42	0-3	92	8	100	91
		61	0-3	93	6	99	96
5	1-2	23	0-3	89	10	99	89
		45	0-3	77	15	92	81
		58	0-3	81	18	99	83
6	1-2	31	0-3	90	10	99	93
		39	0-3	82	17	99	65
		64	0-3	90	9	99	82
7	1-2	25	0-3	86	14	100	84
		41	0-3	80	19	99	81
		60	0-3	64	31	95	76
8	1-2	23	0-3	77	19	96	90
		41	0-3	79	17	96	85
		60	0-3	78	20	98	87
10	1-2	21	0-3	85	13	98	86
		39	0-3	89	9	99	84
		59	0-3	84	14	98	81
	1-2	21	0-3	81	15	95	91
		39	0-3	92	6	98	92
		59	0-3	92	7	99	69

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 2011 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	1-2	13	0-2	87	12	99	90
		26	0-2	73	25	98	89
		50	0-2	92	7	99	92
8	1-2	13	0-2	86	12	98	86
		27	0-2	89	9	99	94
		51	0-2	80	17	97	86
High School	1-2	26	0-2	81	17	99	85
		45	0-2	95	3	97	88
		54	0-2	79	18	97	89

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 2011 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	1-6	51	44	95	65
	1B	1-4	56	39	95	70
7	1A	1-6	59	37	96	81
	1B	1-4	64	35	99	83
10	1A	1-6	56	39	95	77
	1B	1-4	62	36	98	73

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 2011 DC CAS assessments in Reading, Mathematics, and Science/Biology underwent both classical test theory and IRT 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 the Stocking and Lord (1983) procedure was used 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 3PL model (used for multiple-choice items) and the 2PPC model (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 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 multiple-choice and constructed-response 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 10% 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 multiple-choice and constructed-response items and among constructed-response 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 (*Q*1) statistic that takes into account differing numbers of score levels as well as sample size:

$$Z_j = \frac{(Q_{1j} - DF)}{\sqrt{2DF}}$$, where *j* = 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 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 2011 Numbers of Operational Items Flagged for Poor Fit During Calibration

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

Item Calibration

The 2011 items were calibrated using approximately 99% of all Reading, Mathematics, and Science/Biology student data. (Approximately 1% of student records were removed from the calibration sample for students with multiple records and exclusion rules described on page 23.) The number of students within each calibration dataset is presented in Table 36. All DC CAS grades and content areas converged successfully during calibration.

Table 36. Numbers of Students in 2011 Calibration Datasets

Content	Grade	Number of Students
Reading	3	4,773
	4	4,817
	5	4,791
	6	4,393
	7	4,440
	8	4,310
	10	4,442
Mathematics	3	4,805
	4	4,858
	5	4,812
	6	4,423
	7	4,458
	8	4,354
	10	4,415
Science/Biology	5	4,764
	8	4,213
	High School	3,760

Establishing Upper and Lower Bounds for the Grade Level Scales for the Base Years: 2006 for Reading and Mathematics and 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, LOSS and HOSS were set to be equal at the same grade for each content area. For example, the Grade 3 LOSS and HOSS are 300 and 399, (respectively) and the Grade 5 LOSS and HOSS are 500 and 599, respectively, 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 LOSS and HOSS for all grades are provided in Table 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/Biology	900	999

***Note**. In Science, the LOSS and HOSS apply only to the tested Grades 5, 8, and Biology (Grades 8–12). 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, IRT 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 2011, anchor item sets that link the current test forms to the 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; 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 location as the original administration 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 test form and the DC CAS scale score scales.

The equating design and test form equating procedures for the 2011 DC CAS followed the standard design and procedures used in 2007–2009.[4] Anchor items in the 2011 test forms were 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 (1983) equating procedure was 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:

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

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

The TCC method determines the scaling constants (multiplicative -- M1 and additive -- M2) by minimizing the following quadratic loss function (F):

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

where *N* is the number of examinees in the arbitrary group.

Anchor 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. Correlation coefficients for the reference and equated IRT item parameters should be very high (0.90–1.00).

2. Reference and equated anchor item TCCs should be closely aligned.

[4] 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 subsequently. 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 a small number of items from the 2007 statewide field test.

3. Stocking-Lord linear transformation parameters (i.e., scaling transformation constants) should be fairly stable across administrations.

4. *P* values of the anchor items for the estimated new form and the reference form should be similar and aligned on a regression line.

5. *P* values of the anchor items should show the same direction and magnitude of change as do the scale scores.

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

7. Changes in the percentages of examinees in each proficiency level should be reasonable across administrations (e.g. no more than 3 to 5 percent in either direction).

These standard CTB Research team quality checks were followed during calibration and equating analyses for all grades and content areas. Additional anchor item checks were conducted for items that were flagged for review as per the business rules. After reviewing all flagged anchor items, all anchor items were retained.

Anchor Item Parameter Comparisons

Differential anchor item performance between the 2010 and 2011 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. IRT guessing (*c*) parameters typically fluctuate considerably, are held to fixed values during 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, as expected. This is an indication that the anchor items performed similarly in the examinee populations in 2010 and 2011.

The correlations in Table 38 for the discrimination (*a*) and difficulty (*b*) parameters are moderate to high, ranging from 0.84 to 0.98 for *a* parameters (0.52–0.94 in 2010) and from 0.94 to 1.00 for *b* parameters (0.95–0.99 in 2010). 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 2011 DC CAS Operational Test Form

Content	Grade	Discrimination (a)	Difficulty (b)	*P* Value
Reading	3	0.94	0.99	0.99
	4	0.85	0.99	0.99
	5	0.94	0.99	0.99
	6	0.97	0.99	0.99
	7	0.95	0.99	0.99
	8	0.92	0.99	0.99
	10	0.98	1.00	0.98
Mathematics	3	0.93	0.97	0.96
	4	0.92	0.99	0.99
	5	0.85	0.99	0.99
	6	0.84	0.98	0.99
	7	0.88	0.98	0.98
	8	0.90	0.98	0.99
	10	0.93	0.99	0.99
Science/Biology	5	0.97	0.99	0.99
	8	0.92	0.94	0.98
	High School	0.91	0.98	0.99

Scaling Constants

The scaling constants, or Stocking-Lord 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 2010–2011 differences in these scaling constants indicates the degree of differences in average difficulty of the reference and new test form administrations.

The scaling constants for the DC CAS grades and content areas are displayed in Table 39 for the 2007–2011 administrations. Table 39 indicates that both scaling constants are reasonably similar across the 2010 and 2011 administrations. To aid comparison, differences between scaling constants since 2007 are provided in Table 40.

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

Content	Grade	2007		2008		2009		2010		2011	
		Multiplicative	Additive	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	13.6	350.4
	4	11.8	451.2	11.7	453.3	12.4	453.4	13.4	451.6	13.5	451.1
	5	11.4	552.2	11.3	554.9	11.4	553.7	12.4	553.2	12.2	554.2
	6	10.8	652.1	10.4	652.9	10.4	653.0	11.4	651.6	11.2	652.7
	7	10.4	751.3	10.4	752.7	10.2	754.7	11.5	754.3	11.6	754.3
	8	11.1	851.8	10.4	853.8	11.1	853.5	12.3	854.6	12.0	856.9
	10	11.3	954.5	10.9	953.4	10.7	954.1	12.1	952.1	13.0	955.6
Math-ematics	3	14.5	353.9	16.2	354.0	17.3	357.0	16.7	352.4	17.3	353.8
	4	14.1	452.1	13.2	456.4	14.1	457.9	13.8	455.1	14.1	455.1
	5	14.1	552.2	14.8	555.9	15.1	556.4	14.2	556.8	14.7	557.0
	6	13.4	647.3	14.5	649.8	14.3	650.1	14.3	650.3	14.2	652.4
	7	13.7	746.9	13.4	750.0	14.6	751.0	15.1	751.2	14.5	753.6
	8	13.0	844.8	12.5	847.4	12.9	848.5	13.5	848.6	13.0	851.6
	10	15.5	945.2	16.9	945.4	16.7	947.3	16.5	944.3	16.2	948.0
Science	5	N/A		8	550	8.7	549.9	9.1	549.2	9.2	549.4
	8			8	850	8.9	851	9.4	851.9	9.4	852.9
Biology	High School	N/A		8	950	7.7	946.6	7.5	949.5	7.8	949.9

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

Content	Grade	Difference 2008-2007		Difference 2009-2008		Difference 2010-2009		Difference 2011-2010	
		Multiplicative	Additive	Multiplicative	Additive	Multiplicative	Additive	Multiplicative	Additive
Reading	3	0.3	1.4	0.0	-0.9	3.6	-3.5	-0.7	0.8
	4	-0.1	2.1	0.7	0.1	1.0	-1.8	0.1	-0.5
	5	-0.1	2.7	0.1	-1.2	1.0	-0.5	-0.2	1.0
	6	-0.4	0.8	0.0	0.1	1.0	-1.4	-0.2	1.1
	7	0.0	1.4	-0.2	2.0	1.3	-0.4	0.1	0.0
	8	-0.7	2.0	0.7	-0.3	1.2	1.1	-0.3	2.3
	10	-0.4	-1.1	-0.2	0.7	1.4	-2.0	0.9	3.5
Mathematics	3	1.7	0.1	1.1	3.0	-0.6	-4.6	0.6	1.4
	4	-0.9	4.3	0.9	1.5	-0.3	-2.8	0.3	0.0
	5	0.7	3.7	0.3	0.5	-0.9	0.4	0.5	0.2
	6	1.1	2.5	-0.2	0.3	0.0	0.2	-0.1	2.1
	7	-0.3	3.1	1.2	1.0	0.5	0.2	-0.6	2.4
	8	-0.5	2.6	0.4	1.1	0.6	0.1	-0.5	3.0
	10	1.4	0.2	-0.2	1.9	-0.2	-3.0	-0.3	3.7
Science	5	N/A		0.7	-0.1	0.4	-0.7	0.1	0.2
	8			0.9	1.0	0.5	0.9	0.0	1.0
Biology	High School	N/A		-0.3	-3.4	-0.2	2.9	0.3	0.4

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 were drafted for Reading and Mathematics 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 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, DC 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 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 Reading and Mathematics in 2006 and Science, Biology, and Composition in 2008. 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 be expected to 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 content area 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 (EPM) score. 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	36	67
	5	556	39	72
	6	655	38	70
	7	756	37	69
	8	856	33	61
	10	956	37	69
Mathematics	3	360	48	80
	4	458	39	65
	5	560	42	70
	6	654	36	60
	7	752	33	55
	8	850	27	45
	10	951	30	50
Science	5	553	27	51
	8	856	24	45
Biology	High School	952	19	36

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, raw score standard deviations, 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 collection of items on a test ranged from moderately difficult (mean *p* value of 0.38 for Biology) to moderately easy (mean *p* value of 0.70 for Grade 3 Mathematics).

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 2.78% in Mathematics Grade 10. Overall, these omit rates are low. CTB flags items when more than 5% 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%, except for in Reading Grade 10 (1.16%), Mathematics Grade 10 (1.61%), and high school Biology (1.13%) indicating that the DC CAS tests, which generally are 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 2011 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	348.21 (16.39)	34.48 (11.75)	0.67	0.46	1.08	0.31
	4	450.83 (15.83)	32.02 (11.57)	0.61	0.44	0.57	0.16
	5	552.86 (14.66)	35.01 (11.08)	0.68	0.46	0.42	0.12
	6	650.88 (14.34)	33.36 (10.94)	0.64	0.43	0.49	0.17
	7	753.84 (13.94)	34.45 (10.62)	0.66	0.41	0.65	0.27
	8	853.50 (14.93)	31.22 (10.88)	0.59	0.40	0.75	0.30
	10	951.53 (16.49)	32.90 (11.90)	0.63	0.44	2.04	1.16
Mathematics	3	352.54 (19.28)	40.25 (12.77)	0.70	0.46	0.74	0.10
	4	455.21 (16.08)	36.37 (12.50)	0.63	0.42	0.43	0.12
	5	556.14 (16.69)	38.26 (12.45)	0.65	0.44	0.33	0.07
	6	650.44 (16.93)	33.20 (13.37)	0.57	0.44	0.55	0.24
	7	752.39 (17.27)	34.63 (12.19)	0.59	0.41	0.83	0.33
	8	850.50 (15.93)	30.55 (12.31)	0.52	0.39	0.95	0.43
	10	944.77 (19.10)	27.44 (11.96)	0.47	0.39	2.78	1.61
Science/ Biology	5	547.92 (12.53)	24.35 (9.65)	0.47	0.35	0.89	0.49
	8	849.05 (16.76)	21.75 (9.45)	0.42	0.33	1.60	0.76
	High School	946.87 (14.83)	19.43 (8.47)	0.38	0.30	2.33	1.13
Composition	4	N/A	5.75 (1.84)	0.58	0.92	N/A	N/A
	7	N/A	5.80 (1.64)	0.58	0.92	N/A	N/A
	10	N/A	5.58 (1.92)	0.56	0.93	N/A	N/A

Note. Omit and not reached rates are percentages.

DC CAS Performance Level Distributions

The 2011 student performance results using all valid test scores are presented in Table 47.

Table 47. DC CAS 2011 Percentages of Students at Each Performance Level

Content	Performance Level	Grade						
		3	4	5	6	7	8	10[1]
Reading	N	4,796	4,841	4,797	4,403	4,456	4,327	4,491
	Below Basic	22.29	18.65	15.13	15.47	11.11	13.73	18.77
	Basic	36.74	37.57	38.65	42.31	40.82	37.60	37.16
	Proficient	37.82	35.98	39.13	37.52	35.19	36.54	33.22
	Advanced	3.15	7.79	7.09	4.70	12.88	12.13	10.84
Mathematics	N	4,823	4,873	4,817	4,433	4,485	4,370	4,464
	Below Basic	22.79	18.67	18.37	15.11	14.47	13.57	23.97
	Basic	41.84	35.50	37.22	39.66	29.39	28.60	35.44
	Proficient	24.26	34.95	32.61	31.81	43.41	46.59	34.68
	Advanced	11.11	10.88	11.79	13.42	12.73	11.24	5.91
Science/ Biology	N	--	--	4,765	--	--	4,223	3,790
	Below Basic	--	--	20.29	--	--	34.48	32.22
	Basic	--	--	42.25	--	--	29.24	23.17
	Proficient	--	--	31.21	--	--	32.02	42.14
	Advanced	--	--	6.25	--	--	4.26	2.48
Composition	N	--	4,755	--	--	4,301	--	3,761
	Below Basic	--	10.91	--	--	5.98	--	12.28
	Basic	--	54.97	--	--	60.71	--	56.71
	Proficient	--	25.95	--	--	27.44	--	22.63
	Advanced	--	8.16	--	--	5.88	--	8.38

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 Asians. There are similar numbers of males and females in each grade. Mean performance by race/ethnicity generally shows that White students achieve the highest mean scores, followed by Asian students, Hispanic students, and African American students.

Table 48. 2011 DC CAS Subgroup Scale Score Means and Standard Deviations: Reading

Grade	Subgroup	N	Mean	SD
3	All Examinees	4,796	348.21	16.39
	Male	2,458	346.06	17.33
	Female	2,320	350.50	15.02
	Asian	103	355.47	14.04
	African American	3,523	345.81	16.09
	Hispanic	673	348.34	14.69
	White	470	364.12	11.14
4	All Examinees	4,841	450.83	15.83
	Male	2,427	448.71	16.37
	Female	2,389	453.04	14.94
	Asian	74	463.49	14.11
	African American	3,748	448.63	14.99
	Hispanic	605	450.69	14.48
	White	388	469.69	12.00
5	All Examinees	4,797	552.86	14.66
	Male	2,417	551.22	15.05
	Female	2,366	554.58	14.04
	Asian	76	563.64	12.71
	African American	3,764	551.13	13.97
	Hispanic	607	553.10	14.12
	White	334	569.50	11.99
6	All Examinees	4,403	650.88	14.34
	Male	2,228	648.40	15.23
	Female	2,162	653.48	12.88
	Asian	40	664.73	14.13
	African American	3,582	649.44	13.61
	Hispanic	505	652.34	14.77
	White	254	665.98	13.38
7	All Examinees	4,456	753.84	13.94
	Male	2,220	751.45	14.62
	Female	2,212	756.33	12.71
	Asian	57	761.86	12.44
	African American	3,670	752.68	13.60
	Hispanic	464	754.31	12.98
	White	245	768.76	11.92

Grade	Subgroup	N	Mean	SD
8	All Examinees	4,327	853.50	14.93
	Male	2,156	851.15	15.52
	Female	2,152	855.92	13.90
	Asian	58	860.81	13.20
	African American	3,616	852.41	14.46
	Hispanic	417	853.69	15.30
	White	213	869.74	12.58
10	All Examinees	4,491	951.53	16.49
	Male	2,111	949.08	17.23
	Female	2,307	954.12	15.32
	Asian	55	959.96	13.86
	African American	3,743	950.29	16.15
	Hispanic	467	955.05	14.20
	White	163	968.85	18.00

Note. Results are based on students with valid test scores. See the section *Participation in the 2011 DC CAS Test Administrations and Use of Data for Analysis and Score Reporting* for an explanation.

Table 49. DC CAS 2011 Subgroup Scale Score Means and Standard Deviations: Mathematics

Grade	Subgroup	N	Mean	SD
3	All Examinees	4,823	352.54	19.28
	Male	2,470	351.44	20.38
	Female	2,335	353.74	17.97
	Asian	108	367.76	17.08
	African American	3,524	348.62	18.11
	Hispanic	690	355.70	17.12
	White	474	373.20	14.82
4	All Examinees	4,873	455.21	16.08
	Male	2,442	454.31	16.84
	Female	2,405	456.17	15.21
	Asian	84	472.51	13.08
	African American	3,752	452.75	15.16
	Hispanic	616	456.25	15.18
	White	394	473.13	12.06
5	All Examinees	4,817	556.14	16.69
	Male	2,430	555.62	17.11
	Female	2,373	556.73	16.22
	Asian	78	572.33	12.94
	African American	3,764	554.04	15.96
	Hispanic	621	557.79	15.57
	White	337	572.87	15.65
6	All Examinees	4,433	650.44	16.93
	Male	2,244	649.25	17.32
	Female	2,176	651.73	16.42
	Asian	46	666.35	13.66
	African American	3,591	648.45	16.03
	Hispanic	517	654.20	16.75
	White	255	667.87	17.35

Grade	Subgroup	N	Mean	SD
7	All Examinees	4,485	752.39	17.27
	Male	2,236	750.37	17.98
	Female	2,225	754.49	16.22
	Asian	60	768.15	17.15
	African American	3,673	750.57	16.45
	Hispanic	485	753.80	16.03
	White	247	772.78	16.51
8	All Examinees	4,370	850.50	15.93
	Male	2,181	849.55	16.39
	Female	2,170	851.54	15.34
	Asian	66	863.74	14.74
	African American	3,619	849.22	15.55
	Hispanic	449	851.65	14.77
	White	213	866.21	14.30
10	All Examinees	4,464	944.77	19.10
	Male	2,098	943.55	19.70
	Female	2,297	946.26	18.34
	Asian	54	964.98	18.41
	African American	3,722	943.17	18.55
	Hispanic	464	949.08	16.90
	White	161	964.77	19.40

Note. Results are based on students with valid test scores. See the section *Participation in the 2011 DC CAS Test Administrations and Use of Data for Analysis and Score Reporting* for an explanation.

Table 50. DC CAS 2011 Subgroup Scale Score Means and Standard Deviations: Science/Biology

Grade	Subgroup	N	Mean	SD
5	All Examinees	4,765	547.92	12.53
	Male	2,401	547.43	12.98
	Female	2,348	548.45	12.03
	Asian	78	559.14	8.18
	African American	3,729	546.20	11.96
	Hispanic	609	548.81	11.88
	White	332	562.84	8.67
8	All Examinees	4,223	849.05	16.76
	Male	2,081	848.45	17.59
	Female	2,100	849.94	15.60
	Asian	66	858.23	13.37
	African American	3,475	847.81	16.67
	Hispanic	441	850.76	15.79
	White	203	864.94	9.69
High School	All Examinees	3,790	946.87	14.83
	Male	1,757	946.05	15.64
	Female	1,952	947.95	13.77
	Asian	49	950.65	17.27
	African American	3,166	946.36	14.80
	Hispanic	436	948.12	13.87
	White	100	959.14	9.87

Note. Results are based on students with valid test scores. See the section *Participation in the 2011 DC CAS Test Administrations and Use of Data for Analysis and Score Reporting* for an explanation.

Table 51. DC CAS 2011 Subgroup Raw Score Means and Standard Deviations: Composition

Grade	Subgroup	N	Mean	SD
4	All Examinees	4,755	5.75	1.84
	Male	2,373	5.40	1.82
	Female	2,356	6.11	1.78
	Asian	75	7.39	1.70
	African American	3,672	5.49	1.72
	Hispanic	595	5.76	1.64
	White	386	7.92	1.65
7	All Examinees	4,301	5.80	1.64
	Male	2,126	5.46	1.66
	Female	2,149	6.15	1.54
	Asian	54	6.78	1.91
	African American	3,528	5.65	1.54
	Hispanic	453	5.87	1.64
	White	241	7.75	1.70
10	All Examinees	3,761	5.58	1.92
	Male	1,723	5.27	1.84
	Female	1,978	5.86	1.95
	Asian	52	6.77	1.79
	African American	3,110	5.42	1.85
	Hispanic	396	6.00	1.92
	White	152	7.47	1.99

Note. Results are based on students with valid test scores. See the section *Participation in the 2011 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.

Correlations are somewhat higher in the elementary grades than in the middle and high school grades. Correlations between Reading and Mathematics are 0.74 and higher; correlations of Reading and Mathematics scores with Science/Biology scores are 0.61 and higher; correlations with the Composition total scores are in the range of 0.53 to 0.69. Composition correlations are relatively lower because Composition scores range from 2 to 10, which restricts variability and covariance. These results are consistent with typical content area correlations for educational achievement tests in these content areas.

These correlations are moderately high. They indicate that approximately 25%–50% 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 approximately 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 (i.e., disregard measurement error).

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

Grade	Mathematics	Science/ Biology*	Composition
Reading			
Grade 3	0.79	--	--
Grade 4	0.79	--	0.67
Grade 5	0.76	0.75	--
Grade 6	0.77	--	--
Grade 7	0.77	--	0.66
Grade 8	0.77	0.75	--
Grade 10	0.74	0.64	0.69
Mathematics			
Grade 4	--	--	0.63
Grade 5	--	0.73	--
Grade 7	--	--	0.62
Grade 8	--	0.78	--
Grade 10	--	0.61	0.64
Science/Biology			
Grade 10	--	--	0.53

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 area 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 2011 Reading strand and total test correlations for all grades are presented. The Reading strand correlations are moderate to high for all grades.

Table 54 displays the correlations for the DC CAS 2011 Mathematics strand and total test correlations by grade. The correlations are mostly moderate to high. The correlations between Geometry and the other Mathematics strands tend to be lower than for the other strands. Geometry and Measurement also tend to have the lowest correlations 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 2011 Science/Biology strand and total test correlations for all grades are presented. The correlations are moderate to high, although somewhat lower in general than the correlations in Reading and Mathematics.

Table 53. DC CAS 2011 Reading Strand Correlations by Grade

Grade	Content Strand	Language Development	Informational Text	Literary Text	Total Reading
3	Language Development	--	0.78	0.76	0.88
	Informational Text	0.78	--	0.82	0.94
	Literary Text	0.76	0.82	--	0.95
	Total Raw Score	0.88	0.94	0.95	--
4	Language Development	--	0.76	0.77	0.87
	Informational Text	0.76	--	0.82	0.93
	Literary Text	0.77	0.82	--	0.96
	Total Raw Score	0.87	0.93	0.96	--
5	Language Development	--	0.75	0.77	0.87
	Informational Text	0.75	--	0.82	0.93
	Literary Text	0.77	0.82	--	0.96
	Total Raw Score	0.87	0.93	0.96	--
6	Language Development	--	0.73	0.77	0.88
	Informational Text	0.73	--	0.78	0.90
	Literary Text	0.77	0.78	--	0.96
	Total Raw Score	0.88	0.90	0.96	--
7	Language Development	--	0.72	0.74	0.86
	Informational Text	0.72	--	0.80	0.92
	Literary Text	0.74	0.80	--	0.95
	Total Raw Score	0.86	0.92	0.95	--
8	Language Development	--	0.69	0.72	0.83
	Informational Text	0.69	--	0.80	0.93
	Literary Text	0.72	0.80	--	0.95
	Total Raw Score	0.83	0.93	0.95	--
10	Language Development	--	0.77	0.75	0.86
	Informational Text	0.77	--	0.82	0.94
	Literary Text	0.75	0.82	--	0.95
	Total Raw Score	0.86	0.94	0.95	--

Table 54. DC CAS 2011 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	0.69	0.74	0.75	0.93
	Patterns, Relations & Algebra	0.80	--	0.64	0.69	0.70	0.88
	Geometry	0.69	0.64	--	0.64	0.64	0.80
	Measurement	0.74	0.69	0.64	--	0.67	0.84
	Data Analysis, Statistics & Probability	0.75	0.70	0.64	0.67	--	0.88
	Total Raw Score	0.93	0.88	0.80	0.84	0.88	--
4	Number Sense & Operations	--	0.77	0.68	0.70	0.76	0.92
	Patterns, Relations & Algebra	0.77	--	0.63	0.71	0.73	0.89
	Geometry	0.68	0.63	--	0.61	0.65	0.80
	Measurement	0.70	0.71	0.61	--	0.68	0.83
	Data Analysis, Statistics & Probability	0.76	0.73	0.65	0.68	--	0.89
	Total Raw Score	0.92	0.89	0.80	0.83	0.89	--
5	Number Sense & Operations	--	0.78	0.71	0.72	0.77	0.93
	Patterns, Relations & Algebra	0.78	--	0.68	0.68	0.74	0.90
	Geometry	0.71	0.68	--	0.64	0.68	0.82
	Measurement	0.72	0.68	0.64	--	0.66	0.84
	Data Analysis, Statistics & Probability	0.77	0.74	0.68	0.66	--	0.87
	Total Raw Score	0.93	0.90	0.82	0.84	0.87	--
6	Number Sense & Operations	--	0.80	0.66	0.67	0.75	0.92
	Patterns, Relations & Algebra	0.80	--	0.65	0.70	0.75	0.92
	Geometry	0.66	0.65	--	0.57	0.60	0.77
	Measurement	0.67	0.70	0.57	--	0.64	0.80
	Data Analysis, Statistics & Probability	0.75	0.75	0.60	0.64	--	0.87
	Total Raw Score	0.92	0.92	0.77	0.80	0.87	--

Grade	Content Strand	Number Sense & Operations	Patterns, Relations & Algebra	Geometry	Measurement	Data Analysis, Statistics & Probability	Total Mathematics
7	Number Sense & Operations	--	0.76	0.63	0.73	0.69	0.92
	Patterns, Relations & Algebra	0.76	--	0.61	0.70	0.68	0.89
	Geometry	0.63	0.61	--	0.60	0.56	0.78
	Measurement	0.73	0.70	0.60	--	0.67	0.85
	Data Analysis, Statistics & Probability	0.69	0.68	0.56	0.67	--	0.83
	Total Raw Score	0.92	0.89	0.78	0.85	0.83	--
8	Number Sense & Operations	--	0.75	0.64	0.68	0.72	0.90
	Patterns, Relations & Algebra	0.75	--	0.65	0.69	0.72	0.91
	Geometry	0.64	0.65	--	0.61	0.62	0.79
	Measurement	0.68	0.69	0.61	--	0.67	0.83
	Data Analysis, Statistics & Probability	0.72	0.72	0.62	0.67	--	0.86
	Total Raw Score	0.90	0.91	0.79	0.83	0.86	--
10	Number Sense & Operations	--	0.72	0.59	0.52	0.62	0.84
	Patterns, Relations & Algebra	0.72	--	0.70	0.60	0.72	0.92
	Geometry	0.59	0.70	--	0.58	0.65	0.82
	Measurement	0.52	0.60	0.58	--	0.59	0.74
	Data Analysis, Statistics & Probability	0.62	0.72	0.65	0.59	--	0.86
	Total Raw Score	0.84	0.92	0.82	0.74	0.86	--

Table 55. DC CAS 2011 Science/Biology Strand Correlations by Grade

Grade	Content Strand	Science and Technology	Earth and Space Science	Physical Science	Life Science	Total Science
5	Science and Technology	--	0.63	0.65	0.65	0.86
	Earth and Space Science	0.63	--	0.64	0.67	0.86
	Physical Science	0.65	0.64	--	0.66	0.84
	Life Science	0.65	0.67	0.66	--	0.87
	Total Raw Score	0.86	0.86	0.84	0.87	--

Grade	Content Strand	Scientific Thinking and Inquiry	Matter and Reactions	Forces	Energy and Waves	Total Science
8	Scientific Thinking and Inquiry	--	0.65	0.63	0.54	0.82
	Matter and Reactions	0.65	--	0.69	0.61	0.91
	Forces	0.63	0.69	--	0.58	0.85
	Energy and Waves	0.54	0.61	0.58	--	0.79
	Total Raw Score	0.82	0.91	0.85	0.79	--

Grade	Content Strand	Cell Biology and Biochemistry	Genetics and Evolution	Multicellular Organisms	Ecosystems	Total Biology
High School	Cell Biology and Biochemistry	--	0.62	0.55	0.55	0.83
	Genetics and Evolution	0.62	--	0.56	0.58	0.87
	Multicellular Organisms	0.55	0.56	--	0.57	0.80
	Ecosystems	0.55	0.58	0.57	--	0.79
	Total Raw Score	0.83	0.87	0.80	0.79	--

The DC CAS 2011 rubric score and total Composition test correlations for all grades are presented in Table 56. The correlations between the Topic Development and Language Conventions scores are moderate, 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 2011 Composition Rubric Score Correlations by Grade

Grade	Content Strand	Topic Development	Language Conventions	Total Composition
4	Topic Development	--	0.68	0.93
	Language Conventions	0.68	--	0.90
7	Topic Development	--	0.69	0.94
	Language Conventions	0.69	--	0.89
10	Topic Development	--	0.74	0.96
	Language Conventions	0.74	--	0.90

Section 11. DC CAS 2011 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 2011, two sets of field test items were embedded in the operational test forms for Reading, Mathematics, and Science/Biology to expand the bank of items available for use in 2012 operational test forms and to address specific item bank needs (e.g., relatively easy items). Analysis of the 2011 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 section headers on previous years' field test results contained in the technical reports for those years.

References

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.

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. (2000). ITEMWIN [Computer program]. Unpublished.

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, 2009). Retrieved December 7, 2010 from http://www.ed.gov/policy/elsec/guid/saaprguidance.pdf.

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

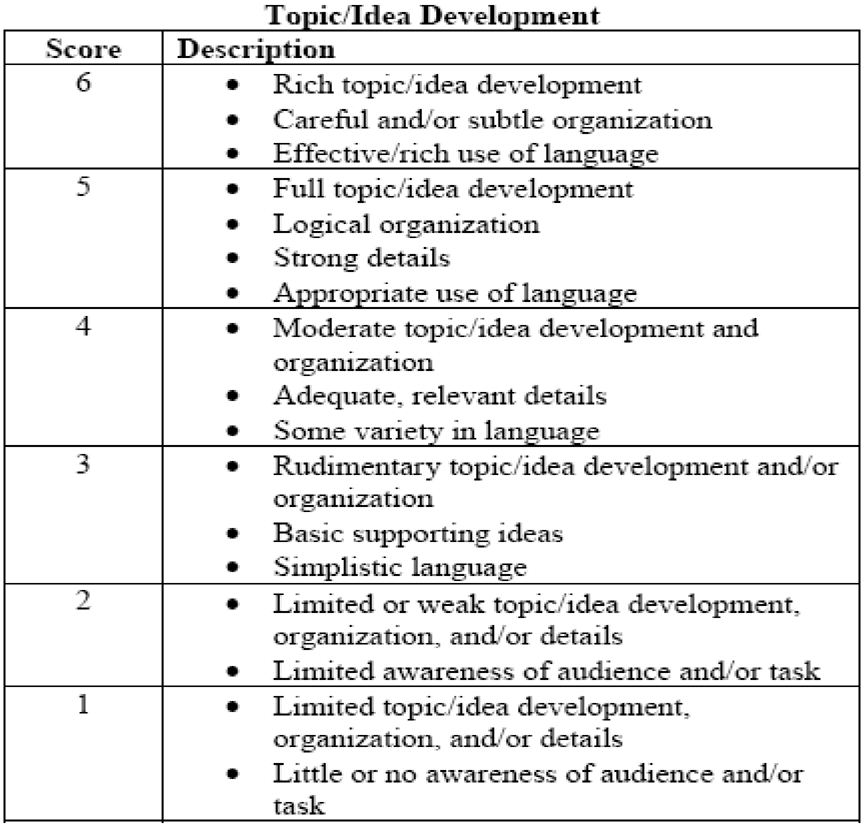
Check to ensure that if any religious, political, social, or philosophical issues are addressed:

- ❑ 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:

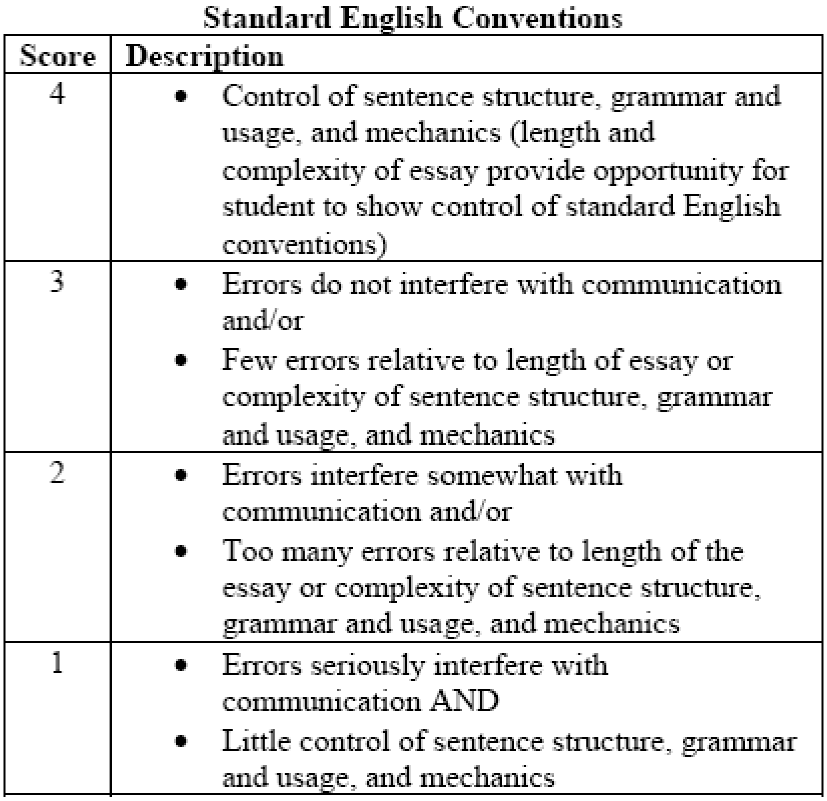
- ❑ be free of offensive, disturbing, or inappropriate language or content
- ❑ be free of stereotyping based on:
  - gender
  - race
  - ethnicity
  - religion
  - socioeconomic status
  - age
  - regional or geographic area
  - disability
  - occupation
- ❑ demonstrate sensitivity to historical representation 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*, 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	2,440	48	0.932	0.937	0.936
Females	2,315		0.919	0.925	0.924
Asian	103		0.912	0.917	0.918
African American	3,501		0.923	0.928	0.928
Hispanic	673		0.916	0.920	0.921
White	469		0.839	0.851	0.849
4					
Males	2,414	48	0.926	0.930	0.930
Females	2,378		0.916	0.920	0.920
Asian	74		0.915	0.921	0.922
African American	3,726		0.913	0.918	0.917
Hispanic	603		0.913	0.917	0.917
White	388		0.880	0.885	0.887
5					
Males	2,413	48	0.930	0.933	0.933
Females	2,364		0.923	0.926	0.927
Asian	76		0.897	0.904	0.905
African American	3,760		0.923	0.925	0.926
Hispanic	606		0.924	0.926	0.927
White	333		0.892	0.899	0.901
6					
Males	2,222	48	0.925	0.927	0.927
Females	2,158		0.908	0.910	0.910
Asian	40		0.864	0.873	0.874
African American	3,573		0.914	0.916	0.916
Hispanic	504		0.917	0.920	0.920
White	254		0.911	0.918	0.916
7					
Males	2,213	48	0.916	0.921	0.919
Females	2,203		0.901	0.906	0.904
Asian	57		0.915	0.920	0.921
African American	3,657		0.907	0.912	0.911
Hispanic	462		0.902	0.908	0.906
White	244		0.883	0.893	0.891
8					
Males	2,148	48	0.911	0.918	0.917
Females	2,143		0.907	0.914	0.912
Asian	58		0.917	0.926	0.924
African American	3,601		0.902	0.910	0.908
Hispanic	416		0.914	0.920	0.919
White	212		0.903	0.911	0.911

Grade/Subgroup	Students with Valid Test Scores	Number of Items	Alpha	Stratified Alpha	Feldt-Raju
10					
Males	2,085	48	0.927	0.934	0.933
Females	2,284		0.917	0.925	0.924
Asian	55		0.913	0.918	0.919
African American	3,700		0.921	0.928	0.927
Hispanic	462		0.911	0.917	0.917
White	163		0.942	0.949	0.949

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	2,459	54	0.939	0.945	0.945
Females	2,328		0.928	0.934	0.935
Asian	108		0.914	0.924	0.929
African American	3,508		0.928	0.932	0.933
Hispanic	690		0.924	0.930	0.931
White	472		0.884	0.893	0.897
4					
Males	2,431	54	0.925	0.932	0.932
Females	2,401		0.920	0.927	0.928
Asian	84		0.892	0.895	0.903
African American	3,739		0.913	0.919	0.920
Hispanic	614		0.914	0.921	0.922
White	394		0.879	0.885	0.889
5					
Males	2,428	54	0.932	0.935	0.936
Females	2,370		0.926	0.929	0.931
Asian	78		0.903	0.905	0.913
African American	3,759		0.923	0.926	0.927
Hispanic	621		0.924	0.928	0.929
White	337		0.919	0.923	0.926
6					
Males	2,240	54	0.931	0.937	0.938
Females	2,170		0.927	0.933	0.934
Asian	46		0.903	0.906	0.913
African American	3,582		0.921	0.927	0.928
Hispanic	517		0.929	0.936	0.937
White	254		0.942	0.948	0.949
7					
Males	2,222	54	0.922	0.926	0.927
Females	2,214		0.916	0.920	0.921
Asian	60		0.932	0.938	0.939
African American	3,650		0.911	0.915	0.916
Hispanic	483		0.913	0.917	0.918
White	245		0.927	0.930	0.932

Grade/Subgroup	Students with Valid Test Scores	Number of Items	Alpha	Stratified Alpha	Feldt-Raju
8					
Males	2,174	54	0.918	0.924	0.923
Females	2,161		0.913	0.919	0.919
Asian	66		0.923	0.928	0.930
African American	3,605		0.907	0.913	0.913
Hispanic	447		0.912	0.919	0.918
White	213		0.925	0.930	0.931
10					
Males	2,073	54	0.915	0.917	0.919
Females	2,276		0.908	0.911	0.912
Asian	54		0.933	0.934	0.937
African American	3,685		0.901	0.904	0.906
Hispanic	458		0.909	0.911	0.912
White	159		0.933	0.936	0.938

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	2,400	50	0.893	0.894	0.894
Females	2,348		0.886	0.886	0.887
Asian	78		0.895	0.896	0.899
African American	3,728		0.852	0.853	0.855
Hispanic	609		0.875	0.876	0.877
White	332		0.890	0.891	0.892
8					
Males	2,076	50	0.886	0.889	0.890
Females	2,095		0.866	0.869	0.870
Asian	66		0.902	0.906	0.906
African American	3,468		0.851	0.855	0.856
Hispanic	440		0.869	0.874	0.874
White	201		0.890	0.892	0.894
High School					
Males	1,741	50	0.861	0.862	0.862
Females	1,941		0.853	0.854	0.855
Asian	48		0.927	0.929	0.931
African American	3,143		0.838	0.839	0.840
Hispanic	431		0.861	0.863	0.864
White	100		0.910	0.910	0.913

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	0.9371	0.8935	0.9482	0.7787
		Kappa	0.8175	0.7813	0.4674	0.6764
	Classification Accuracy	Accuracy	0.9539	0.9233	0.9636	0.8408
		False Positive Errors	0.0189	0.0254	0.0089	0.0533
		False Negative Errors	0.0272	0.0512	0.0275	0.1059
4	Classification Consistency	Consistency	0.9282	0.8906	0.9401	0.7591
		Kappa	0.7579	0.7784	0.6339	0.6524
	Classification Accuracy	Accuracy	0.9468	0.9241	0.9562	0.8271
		False Positive Errors	0.0167	0.0365	0.0091	0.0623
		False Negative Errors	0.0365	0.0394	0.0347	0.1106
5	Classification Consistency	Consistency	0.9490	0.8857	0.9323	0.7671
		Kappa	0.8031	0.7701	0.5863	0.6584
	Classification Accuracy	Accuracy	0.9637	0.9155	0.9506	0.8298
		False Positive Errors	0.0174	0.0294	0.0101	0.0569
		False Negative Errors	0.0190	0.0551	0.0393	0.1133
6	Classification Consistency	Consistency	0.9326	0.8881	0.9494	0.7702
		Kappa	0.7517	0.7714	0.5668	0.6560
	Classification Accuracy	Accuracy	0.9526	0.9169	0.9632	0.8327
		False Positive Errors	0.0254	0.0233	0.0062	0.0548
		False Negative Errors	0.0220	0.0598	0.0307	0.1125
7	Classification Consistency	Consistency	0.9457	0.8811	0.9137	0.7414
		Kappa	0.7313	0.7618	0.6516	0.6248
	Classification Accuracy	Accuracy	0.9630	0.9152	0.9388	0.8170
		False Positive Errors	0.0190	0.0406	0.0221	0.0817
		False Negative Errors	0.0180	0.0442	0.0391	0.1013
8	Classification Consistency	Consistency	0.9304	0.8849	0.9278	0.7433
		Kappa	0.7052	0.7695	0.6857	0.6314
	Classification Accuracy	Accuracy	0.9514	0.9147	0.9504	0.8164
		False Positive Errors	0.0228	0.0286	0.0190	0.0705
		False Negative Errors	0.0258	0.0567	0.0306	0.1131
10	Classification Consistency	Consistency	0.9273	0.8918	0.9304	0.7502
		Kappa	0.7552	0.7809	0.6745	0.6471
	Classification Accuracy	Accuracy	0.9492	0.9226	0.9485	0.8204
		False Positive Errors	0.0286	0.0264	0.0104	0.0654
		False Negative Errors	0.0222	0.0510	0.0411	0.1142

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

Grade	Mathematics Classification Consistency and Accuracy		Basic	Proficient	Ad-vanced	All Cuts
3	Classification Consistency	Consistency	0.9263	0.9119	0.9266	0.7669
		Kappa	0.7910	0.8089	0.6633	0.6705
	Classification Accuracy	Accuracy	0.9463	0.9364	0.9475	0.8303
		False Positive Errors	0.0156	0.0318	0.0164	0.0637
		False Negative Errors	0.0381	0.0318	0.0361	0.1060
4	Classification Consistency	Consistency	0.9130	0.9037	0.9362	0.7532
		Kappa	0.7161	0.8063	0.7032	0.6535
	Classification Accuracy	Accuracy	0.9382	0.9314	0.9539	0.8235
		False Positive Errors	0.0256	0.0217	0.0095	0.0568
		False Negative Errors	0.0361	0.0469	0.0366	0.1197
5	Classification Consistency	Consistency	0.9285	0.9094	0.9338	0.7718
		Kappa	0.7622	0.8169	0.7078	0.6800
	Classification Accuracy	Accuracy	0.9492	0.9301	0.9504	0.8297
		False Positive Errors	0.0240	0.0203	0.0127	0.0570
		False Negative Errors	0.0268	0.0496	0.0369	0.1133
6	Classification Consistency	Consistency	0.9213	0.9104	0.9365	0.7684
		Kappa	0.6953	0.8191	0.7423	0.6714
	Classification Accuracy	Accuracy	0.9413	0.9314	0.9565	0.8292
		False Positive Errors	0.0201	0.0220	0.0168	0.0590
		False Negative Errors	0.0386	0.0466	0.0267	0.1119
7	Classification Consistency	Consistency	0.9211	0.8976	0.9378	0.7579
		Kappa	0.6909	0.7916	0.7411	0.6520
	Classification Accuracy	Accuracy	0.9432	0.9283	0.9548	0.8263
		False Positive Errors	0.0369	0.0295	0.0145	0.0808
		False Negative Errors	0.0199	0.0422	0.0308	0.0928
8	Classification Consistency	Consistency	0.8985	0.8893	0.9472	0.7379
		Kappa	0.5959	0.7734	0.7519	0.6146
	Classification Accuracy	Accuracy	0.9284	0.9212	0.9625	0.8123
		False Positive Errors	0.0347	0.0344	0.0109	0.0798
		False Negative Errors	0.0370	0.0444	0.0266	0.1079
10	Classification Consistency	Consistency	0.8842	0.8942	0.9633	0.7425
		Kappa	0.6853	0.7820	0.7246	0.6327
	Classification Accuracy	Accuracy	0.9144	0.9227	0.9725	0.8097
		False Positive Errors	0.0506	0.0393	0.0061	0.0960
		False Negative Errors	0.0350	0.0380	0.0214	0.0943

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	0.8661	0.8822	0.9643	0.7140
		Kappa	0.5908	0.7502	0.7214	0.5824
	Classification Accuracy	Accuracy	0.9037	0.9161	0.9742	0.7940
		False Positive Errors	0.0521	0.0337	0.0109	0.0967
		False Negative Errors	0.0441	0.0502	0.0149	0.1093
8	Classification Consistency	Consistency	0.8238	0.8740	0.9698	0.6815
		Kappa	0.6141	0.7307	0.7000	0.5419
	Classification Accuracy	Accuracy	0.8799	0.9125	0.9782	0.7722
		False Positive Errors	0.0538	0.0294	0.0033	0.0860
		False Negative Errors	0.0663	0.0581	0.0185	0.1418
High School	Classification Consistency	Consistency	0.8019	0.8282	0.9808	0.6578
		Kappa	0.5469	0.6541	0.6653	0.4838
	Classification Accuracy	Accuracy	0.8521	0.8799	0.9841	0.7282
		False Positive Errors	0.0761	0.0341	0.0021	0.1088
		False Negative Errors	0.0718	0.0860	0.0138	0.1630

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	0.7896	0.6957	0.8559	0.0639	0.0802
Females	0.7755	0.6646	0.8460	0.0626	0.0914
Asian	0.7400	0.5748	0.8160	0.0573	0.1267
African American	0.7896	0.6895	0.8568	0.0626	0.0806
Hispanic	0.7803	0.6711	0.8496	0.0642	0.0862
White	0.7458	0.4963	0.8198	0.0682	0.1121
Grade 4					
Males	0.7643	0.6636	0.8327	0.0783	0.0890
Females	0.7579	0.6462	0.8295	0.0772	0.0932
Asian	0.7294	0.5858	0.8002	0.0877	0.1121
African American	0.7640	0.6515	0.8336	0.0763	0.0902
Hispanic	0.7508	0.6322	0.8221	0.0830	0.0949
White	0.7592	0.5861	0.8304	0.0818	0.0877
Grade 5					
Males	0.7593	0.6483	0.8328	0.0743	0.0929
Females	0.7491	0.6258	0.8257	0.0772	0.0971
Asian	0.7349	0.5846	0.8156	0.0989	0.0854
African American	0.7578	0.6364	0.8319	0.0741	0.0940
Hispanic	0.7539	0.6282	0.8293	0.0711	0.0996
White	0.7189	0.5239	0.8026	0.0974	0.1000
Grade 6					
Males	0.7776	0.6701	0.8459	0.0742	0.0800
Females	0.7611	0.6324	0.8335	0.0801	0.0864
Asian	0.7377	0.5792	0.8166	0.1172	0.0662
African American	0.7700	0.6490	0.8403	0.0767	0.0830
Hispanic	0.7763	0.6590	0.8468	0.0761	0.0771
White	0.7518	0.6035	0.8204	0.0832	0.0964
Grade 7					
Males	0.7351	0.6153	0.8066	0.0921	0.1013
Females	0.7256	0.5965	0.8008	0.1015	0.0977
Asian	0.7372	0.6213	0.8096	0.0889	0.1014
African American	0.7288	0.5992	0.8023	0.0965	0.1011
Hispanic	0.7264	0.5949	0.8007	0.0981	0.1012
White	0.7663	0.5873	0.8334	0.0972	0.0694
Grade 8					
Males	0.7472	0.6384	0.8170	0.0849	0.0981
Females	0.7541	0.6459	0.8232	0.0842	0.0926
Asian	0.7570	0.6399	0.8299	0.0938	0.0763
African American	0.7478	0.6337	0.8181	0.0850	0.0969
Hispanic	0.7575	0.6537	0.8237	0.0831	0.0933

Grade/Subgroup	Classification Consistency		Classification Accuracy		
	Consistency	Kappa	Accuracy	False Positive Errors	False Negative Errors
White	0.7819	0.6116	0.8430	0.0754	0.0816
Grade 10					
Males	0.7550	0.6545	0.8261	0.0839	0.0899
Females	0.7475	0.6420	0.8219	0.0859	0.0922
Asian	0.7213	0.6043	0.8029	0.1101	0.0869
African American	0.7521	0.6472	0.8251	0.0843	0.0906
Hispanic	0.7313	0.6140	0.8073	0.0967	0.0960
White	0.7996	0.6675	0.8572	0.0624	0.0804

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	0.7677	0.6735	0.8372	0.0708	0.0920
Females	0.7535	0.6486	0.8277	0.0780	0.0943
Asian	0.7564	0.6496	0.8280	0.0897	0.0823
African American	0.7705	0.6591	0.8396	0.0678	0.0926
Hispanic	0.7450	0.6360	0.8222	0.0832	0.0945
White	0.7122	0.5376	0.7958	0.1075	0.0966
Grade 4					
Males	0.7613	0.6672	0.8333	0.0773	0.0894
Females	0.7500	0.6465	0.8258	0.0846	0.0896
Asian	0.7947	0.6608	0.8588	0.0620	0.0793
African American	0.7546	0.6475	0.8290	0.0803	0.0907
Hispanic	0.7477	0.6365	0.8229	0.0881	0.0891
White	0.7698	0.5981	0.8392	0.0813	0.0795
Grade 5					
Males	0.7720	0.6820	0.8326	0.0796	0.0878
Females	0.7716	0.6768	0.8327	0.0819	0.0855
Asian	0.7797	0.6354	0.8386	0.0947	0.0667
African American	0.7731	0.6744	0.8337	0.0794	0.0870
Hispanic	0.7619	0.6601	0.8245	0.0821	0.0934
White	0.7790	0.6375	0.8388	0.0870	0.0742
Grade 6					
Males	0.7566	0.6564	0.8306	0.0832	0.0861
Females	0.7668	0.6679	0.8388	0.0812	0.0800
Asian	0.7851	0.6568	0.8480	0.0603	0.0917
African American	0.7552	0.6442	0.8303	0.0844	0.0853
Hispanic	0.7712	0.6762	0.8408	0.0801	0.0791
White	0.8227	0.6983	0.8754	0.0639	0.0608
Grade 7					
Males	0.7672	0.6709	0.8374	0.0798	0.0828
Females	0.7635	0.6491	0.8327	0.0751	0.0923

Grade/Subgroup	Classification Consistency		Classification Accuracy		
	Consistency	Kappa	Accuracy	False Positive Errors	False Negative Errors
Asian	0.8397	0.7434	0.8905	0.0436	0.0659
African American	0.7577	0.6458	0.8297	0.0806	0.0897
Hispanic	0.7731	0.6575	0.8405	0.0680	0.0915
White	0.8498	0.7132	0.8926	0.0579	0.0495
Grade 8					
Males	0.7317	0.6094	0.8069	0.0942	0.0990
Females	0.7457	0.6150	0.8175	0.0954	0.0872
Asian	0.7620	0.6112	0.8252	0.0900	0.0848
African American	0.7338	0.6021	0.8086	0.0963	0.0952
Hispanic	0.7332	0.5961	0.8077	0.0954	0.0968
White	0.8245	0.6954	0.8750	0.0681	0.0569
Grade 10					
Males	0.7425	0.6345	0.8118	0.0851	0.1031
Females	0.7458	0.6337	0.8144	0.0874	0.0982
Asian	0.8004	0.6933	0.8576	0.0683	0.0741
African American	0.7437	0.6288	0.8131	0.0855	0.1015
Hispanic	0.7194	0.5956	0.7938	0.1017	0.1044
White	0.8047	0.7006	0.8525	0.0680	0.0795

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	0.7148	0.5877	0.7946	0.1004	0.1050
Females	0.7187	0.5888	0.7995	0.0978	0.1027
Asian	0.7279	0.5715	0.8073	0.0890	0.1037
African American	0.7093	0.5623	0.7908	0.1022	0.1071
Hispanic	0.7236	0.5870	0.8049	0.0935	0.1016
White	0.7821	0.6223	0.8482	0.0757	0.0761
Grade 8					
Males	0.7023	0.5725	0.7819	0.0909	0.1272
Females	0.6863	0.5465	0.7673	0.1050	0.1277
Asian	0.7211	0.5806	0.8026	0.1038	0.0936
African American	0.6881	0.5426	0.7691	0.1000	0.1310
Hispanic	0.6939	0.5559	0.7769	0.0982	0.1249
White	0.8014	0.6592	0.8602	0.0585	0.0813
High School					
Males	0.6771	0.5122	0.7496	0.1168	0.1336
Females	0.6678	0.4910	0.7408	0.1267	0.1325
Asian	0.6688	0.5228	0.7402	0.1443	0.1155
African American	0.6682	0.4940	0.7417	0.1236	0.1347
Hispanic	0.6763	0.5026	0.7478	0.1171	0.1350
White	0.7874	0.6186	0.8467	0.0739	0.0794

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	0.9318	0.8281	0.9531	0.0240	0.0229
Females	0.9402	0.7893	0.9580	0.0210	0.0210
Asian	0.9473	0.7303	0.9610	0.0112	0.0278
African American	0.9286	0.8138	0.9506	0.0254	0.0239
Hispanic	0.9305	0.7835	0.9508	0.0241	0.0251
White	0.9958	0.8995	0.9976	0.0013	0.0011
Grade 4					
Males	0.9180	0.7739	0.9428	0.0272	0.0300
Females	0.9395	0.7416	0.9587	0.0227	0.0186
Asian	0.9736	0.6622	0.9803	0.0099	0.0098
African American	0.9207	0.7609	0.9452	0.0282	0.0266
Hispanic	0.9296	0.7597	0.9514	0.0218	0.0269
White	0.9945	0.7978	0.9966	0.0022	0.0012
Grade 5					
Males	0.9379	0.7937	0.9574	0.0228	0.0197
Females	0.9557	0.7919	0.9700	0.0159	0.0141
Asian	0.9949	0.8374	0.9974	0.0026	0.0000
African American	0.9413	0.7919	0.9602	0.0211	0.0187
Hispanic	0.9497	0.7923	0.9644	0.0191	0.0164
White	0.9901	0.8016	0.9930	0.0043	0.0027
Grade 6					
Males	0.9218	0.7701	0.9441	0.0317	0.0241
Females	0.9486	0.7345	0.9633	0.0241	0.0126
Asian	0.9827	0.1714	0.9895	0.0106	0.0000
African American	0.9280	0.7570	0.9486	0.0307	0.0208
Hispanic	0.9525	0.7822	0.9663	0.0218	0.0119
White	0.9866	0.7800	0.9904	0.0067	0.0029
Grade 7					
Males	0.9331	0.7499	0.9520	0.0265	0.0215
Females	0.9611	0.7146	0.9727	0.0170	0.0103
Asian	0.9676	0.6974	0.9765	0.0050	0.0185
African American	0.9433	0.7451	0.9595	0.0234	0.0170
Hispanic	0.9527	0.7292	0.9673	0.0194	0.0133
White	0.9923	0.7572	0.9947	0.0045	0.0008
Grade 8					
Males	0.9211	0.7353	0.9442	0.0300	0.0258
Females	0.9436	0.6921	0.9605	0.0229	0.0166
Asian	0.9679	0.7822	0.9772	0.0170	0.0058
African American	0.9273	0.7161	0.9489	0.0281	0.0230
Hispanic	0.9411	0.7565	0.9577	0.0244	0.0179
White	0.9905	0.7707	0.9932	0.0046	0.0022
Grade 10					
Males	0.9157	0.7693	0.9417	0.0314	0.0269
Females	0.9353	0.7347	0.9565	0.0243	0.0192

Grade/Subgroup	Classification Consistency		Classification Accuracy		
	Consistency	Kappa	Accuracy	False Positive Errors	False Negative Errors
Asian	0.9353	0.4991	0.9572	0.0288	0.0140
African American	0.9216	0.7611	0.9467	0.0294	0.0239
Hispanic	0.9359	0.6934	0.9550	0.0261	0.0189
White	0.9821	0.8551	0.9877	0.0024	0.0099

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	0.9238	0.7987	0.9458	0.0267	0.0275
Females	0.9255	0.7690	0.9474	0.0276	0.0249
Asian	0.9717	0.7500	0.9783	0.0039	0.0178
African American	0.9122	0.7800	0.9380	0.0317	0.0303
Hispanic	0.9333	0.7545	0.9519	0.0247	0.0234
White	0.9898	0.7857	0.9930	0.0031	0.0040
Grade 4					
Males	0.9203	0.7616	0.9417	0.0289	0.0294
Females	0.9232	0.7275	0.9429	0.0336	0.0235
Asian	0.9909	0.4499	0.9927	0.0019	0.0054
African American	0.9107	0.7413	0.9343	0.0352	0.0305
Hispanic	0.9317	0.7335	0.9494	0.0293	0.0213
White	0.9944	0.7057	0.9957	0.0032	0.0012
Grade 5					
Males	0.9251	0.7688	0.9454	0.0284	0.0263
Females	0.9311	0.7547	0.9504	0.0273	0.0223
Asian	0.9993	0.9867	0.9997	0.0002	0.0001
African American	0.9200	0.7590	0.9420	0.0315	0.0264
Hispanic	0.9374	0.7545	0.9545	0.0220	0.0235
White	0.9866	0.8022	0.9901	0.0033	0.0066
Grade 6					
Males	0.9091	0.6894	0.9360	0.0361	0.0279
Females	0.9309	0.7028	0.9523	0.0255	0.0221
Asian	0.9855	0.6949	0.9910	0.0058	0.0032
African American	0.9100	0.6850	0.9370	0.0350	0.0280
Hispanic	0.9505	0.7577	0.9663	0.0175	0.0162
White	0.9779	0.7469	0.9849	0.0069	0.0082
Grade 7					
Males	0.9136	0.7097	0.9394	0.0358	0.0247
Females	0.9370	0.7011	0.9558	0.0232	0.0209
Asian	0.9761	0.6984	0.9861	0.0109	0.0030
African American	0.9186	0.7051	0.9427	0.0324	0.0249
Hispanic	0.9392	0.7207	0.9587	0.0240	0.0173
White	0.9839	0.7132	0.9888	0.0059	0.0053

Grade/Subgroup	Classification Consistency		Classification Accuracy		
	Consistency	Kappa	Accuracy	False Positive Errors	False Negative Errors
Grade 8					
Males	0.8931	0.6109	0.9222	0.0434	0.0345
Females	0.9116	0.6046	0.9358	0.0389	0.0254
Asian	0.9806	0.6567	0.9852	0.0068	0.0080
African American	0.8954	0.6058	0.9241	0.0447	0.0312
Hispanic	0.9098	0.5933	0.9329	0.0347	0.0324
White	0.9777	0.7123	0.9835	0.0049	0.0115
Grade 10					
Males	0.8765	0.6889	0.9109	0.0463	0.0429
Females	0.8927	0.6820	0.9236	0.0431	0.0332
Asian	0.9642	0.6643	0.9772	0.0119	0.0109
African American	0.8781	0.6868	0.9125	0.0467	0.0407
Hispanic	0.8958	0.6289	0.9259	0.0439	0.0302
White	0.9662	0.7405	0.9760	0.0160	0.0081

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	0.8616	0.6078	0.9006	0.0546	0.0448
Females	0.8703	0.5923	0.9079	0.0528	0.0392
Asian	0.9727	0.4337	0.9813	0.0089	0.0098
African American	0.8495	0.5890	0.8923	0.0600	0.0477
Hispanic	0.8853	0.6078	0.9195	0.0480	0.0326
White	0.9860	0.6941	0.9903	0.0045	0.0052
Grade 8					
Males	0.8383	0.6485	0.8829	0.0552	0.0619
Females	0.8381	0.6285	0.8820	0.0604	0.0576
Asian	0.9273	0.7080	0.9472	0.0310	0.0218
African American	0.8267	0.6282	0.8738	0.0609	0.0653
Hispanic	0.8522	0.6386	0.8951	0.0597	0.0452
White	0.9735	0.7158	0.9794	0.0077	0.0129
High School					
Males	0.8033	0.5679	0.8610	0.0711	0.0679
Females	0.8069	0.5443	0.8638	0.0760	0.0602
Asian	0.8451	0.5669	0.8923	0.0684	0.0392
African American	0.7997	0.5544	0.8586	0.0754	0.0660
Hispanic	0.8051	0.5428	0.8619	0.0728	0.0652
White	0.9441	0.5628	0.9607	0.0239	0.0153

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	0.9040	0.7936	0.9365	0.0317	0.0318
Females	0.8910	0.7805	0.9277	0.0339	0.0384
Asian	0.8858	0.7511	0.9207	0.0431	0.0362
African American	0.8952	0.7687	0.9307	0.0329	0.0364
Hispanic	0.8875	0.7638	0.9248	0.0342	0.0410
White	0.9322	0.6966	0.9561	0.0275	0.0164
Grade 4					
Males	0.8962	0.7835	0.9251	0.0402	0.0347
Females	0.8863	0.7722	0.9182	0.0377	0.0441
Asian	0.9041	0.7241	0.9284	0.0329	0.0387
African American	0.8869	0.7605	0.9187	0.0398	0.0415
Hispanic	0.8723	0.7412	0.9059	0.0504	0.0437
White	0.9625	0.7796	0.9755	0.0141	0.0104
Grade 5					
Males	0.8776	0.7487	0.9129	0.0389	0.0482
Females	0.8729	0.7455	0.9093	0.0441	0.0466
Asian	0.9154	0.7616	0.9426	0.0388	0.0186
African American	0.8689	0.7310	0.9064	0.0437	0.0500
Hispanic	0.8645	0.7288	0.9036	0.0423	0.0541
White	0.9578	0.7697	0.9714	0.0174	0.0112
Grade 6					
Males	0.8967	0.7772	0.9297	0.0341	0.0362
Females	0.8686	0.7369	0.9091	0.0444	0.0465
Asian	0.8806	0.6600	0.9175	0.0548	0.0277
African American	0.8782	0.7426	0.9162	0.0404	0.0434
Hispanic	0.8784	0.7559	0.9171	0.0439	0.0390
White	0.9571	0.8460	0.9716	0.0133	0.0152
Grade 7					
Males	0.8717	0.7360	0.9059	0.0463	0.0478
Females	0.8616	0.7214	0.8994	0.0540	0.0466
Asian	0.8809	0.7209	0.9137	0.0482	0.0380
African American	0.8613	0.7195	0.8986	0.0520	0.0494
Hispanic	0.8602	0.7203	0.8975	0.0528	0.0497
White	0.9555	0.7868	0.9700	0.0171	0.0129
Grade 8					
Males	0.8924	0.7812	0.9191	0.0393	0.0416
Females	0.8942	0.7867	0.9205	0.0381	0.0414
Asian	0.9426	0.8487	0.9633	0.0250	0.0117
African American	0.8875	0.7735	0.9157	0.0412	0.0431
Hispanic	0.9013	0.8024	0.9243	0.0348	0.0409
White	0.9601	0.8084	0.9666	0.0077	0.0257
Grade 10					
Males	0.8943	0.7779	0.9239	0.0377	0.0384
Females	0.8917	0.7833	0.9234	0.0391	0.0376

Grade/Subgroup	Classification Consistency		Classification Accuracy		
	Consistency	Kappa	Accuracy	False Positive Errors	False Negative Errors
Asian	0.8758	0.7442	0.9089	0.0350	0.0560
African American	0.8906	0.7741	0.9220	0.0389	0.0391
Hispanic	0.8895	0.7785	0.9209	0.0430	0.0361
White	0.9655	0.8583	0.9768	0.0119	0.0113

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	0.9146	0.8116	0.9406	0.0258	0.0337
Females	0.9042	0.7955	0.9334	0.0310	0.0355
Asian	0.9048	0.7763	0.9331	0.0482	0.0187
African American	0.9101	0.7725	0.9373	0.0259	0.0369
Hispanic	0.9000	0.7950	0.9316	0.0364	0.0319
White	0.9216	0.7018	0.9447	0.0306	0.0247
Grade 4					
Males	0.9014	0.7999	0.9334	0.0343	0.0323
Females	0.8941	0.7877	0.9290	0.0350	0.0359
Asian	0.9424	0.7467	0.9636	0.0181	0.0183
African American	0.8932	0.7765	0.9284	0.0352	0.0365
Hispanic	0.8854	0.7707	0.9212	0.0434	0.0354
White	0.9516	0.6990	0.9675	0.0210	0.0115
Grade 5					
Males	0.9135	0.8240	0.9364	0.0305	0.0331
Females	0.9070	0.8124	0.9319	0.0351	0.0330
Asian	0.9542	0.7960	0.9666	0.0106	0.0228
African American	0.9072	0.8046	0.9319	0.0339	0.0342
Hispanic	0.8999	0.7998	0.9265	0.0373	0.0362
White	0.9553	0.8063	0.9678	0.0173	0.0149
Grade 6					
Males	0.9055	0.8064	0.9356	0.0318	0.0325
Females	0.9000	0.7996	0.9316	0.0365	0.0319
Asian	0.9312	0.7453	0.9518	0.0176	0.0306
African American	0.8978	0.7868	0.9304	0.0350	0.0346
Hispanic	0.9029	0.8012	0.9322	0.0400	0.0277
White	0.9657	0.8683	0.9769	0.0159	0.0072
Grade 7					
Males	0.9044	0.8085	0.9343	0.0290	0.0367
Females	0.8958	0.7782	0.9266	0.0339	0.0395
Asian	0.9644	0.8610	0.9784	0.0112	0.0104
African American	0.8936	0.7862	0.9260	0.0337	0.0404
Hispanic	0.9054	0.7985	0.9338	0.0277	0.0386

Grade/Subgroup	Classification Consistency		Classification Accuracy		
	Consistency	Kappa	Accuracy	False Positive Errors	False Negative Errors
White	0.9746	0.8590	0.9821	0.0084	0.0095
Grade 8					
Males	0.8839	0.7654	0.9179	0.0364	0.0457
Females	0.8870	0.7613	0.9207	0.0399	0.0394
Asian	0.9325	0.7160	0.9499	0.0334	0.0167
African American	0.8787	0.7546	0.9144	0.0397	0.0459
Hispanic	0.8873	0.7613	0.9210	0.0391	0.0400
White	0.9798	0.8694	0.9863	0.0115	0.0022
Grade 10					
Males	0.9009	0.7924	0.9280	0.0302	0.0418
Females	0.8884	0.7739	0.9177	0.0366	0.0456
Asian	0.9272	0.7677	0.9454	0.0320	0.0226
African American	0.8938	0.7762	0.9225	0.0326	0.0449
Hispanic	0.8774	0.7545	0.9078	0.0438	0.0484
White	0.9499	0.8393	0.9639	0.0216	0.0145

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	0.8876	0.7593	0.9194	0.0350	0.0456
Females	0.8831	0.7569	0.9163	0.0359	0.0478
Asian	0.8831	0.6533	0.9165	0.0530	0.0306
African American	0.8795	0.7242	0.9133	0.0367	0.0500
Hispanic	0.8783	0.7520	0.9146	0.0366	0.0489
White	0.9645	0.7780	0.9749	0.0143	0.0108
Grade 8					
Males	0.8838	0.7518	0.9174	0.0295	0.0531
Females	0.8671	0.7219	0.9039	0.0401	0.0559
Asian	0.9142	0.7942	0.9385	0.0273	0.0341
African American	0.8713	0.7136	0.9075	0.0361	0.0564
Hispanic	0.8648	0.7249	0.9038	0.0350	0.0612
White	0.9646	0.8116	0.9760	0.0151	0.0088
High School					
Males	0.8438	0.6839	0.8872	0.0471	0.0657
Females	0.8311	0.6620	0.8771	0.0536	0.0693
Asian	0.8446	0.6787	0.8819	0.0498	0.0684
African American	0.8340	0.6645	0.8797	0.0517	0.0686
Hispanic	0.8397	0.6793	0.8834	0.0485	0.0680
White	0.9293	0.7051	0.9515	0.0239	0.0246

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	0.9536	0.4865	0.9663	0.0082	0.0255
Females	0.9441	0.4845	0.9603	0.0077	0.0320
Asian	0.9068	0.4243	0.9343	0.0030	0.0627
African American	0.9656	0.4533	0.9755	0.0042	0.0203
Hispanic	0.9622	0.5001	0.9740	0.0059	0.0201
White	0.8178	0.4480	0.8660	0.0393	0.0947
Grade 4					
Males	0.9499	0.6357	0.9648	0.0109	0.0243
Females	0.9320	0.6500	0.9526	0.0168	0.0305
Asian	0.8515	0.6565	0.8915	0.0449	0.0636
African American	0.9562	0.5460	0.9697	0.0083	0.0221
Hispanic	0.9487	0.5761	0.9649	0.0109	0.0243
White	0.8020	0.6031	0.8583	0.0656	0.0761
Grade 5					
Males	0.9437	0.5879	0.9625	0.0125	0.0250
Females	0.9202	0.5588	0.9464	0.0171	0.0365
Asian	0.8244	0.5697	0.8757	0.0575	0.0668
African American	0.9474	0.4843	0.9654	0.0094	0.0253
Hispanic	0.9396	0.5288	0.9613	0.0096	0.0291
White	0.7709	0.5362	0.8383	0.0757	0.0860
Grade 6					
Males	0.9589	0.5307	0.9721	0.0083	0.0196
Females	0.9438	0.5985	0.9611	0.0116	0.0273
Asian	0.8742	0.6745	0.9097	0.0518	0.0386
African American	0.9636	0.4716	0.9756	0.0057	0.0188
Hispanic	0.9452	0.5862	0.9634	0.0104	0.0262
White	0.8080	0.5800	0.8585	0.0632	0.0783
Grade 7					
Males	0.9300	0.6442	0.9487	0.0192	0.0320
Females	0.9026	0.6569	0.9288	0.0304	0.0408
Asian	0.8882	0.7574	0.9194	0.0357	0.0449
African American	0.9239	0.6087	0.9442	0.0211	0.0347
Hispanic	0.9132	0.6288	0.9358	0.0259	0.0383
White	0.8183	0.6289	0.8686	0.0756	0.0558
Grade 8					
Males	0.9333	0.6371	0.9537	0.0156	0.0308
Females	0.9158	0.6980	0.9422	0.0233	0.0346
Asian	0.8459	0.6333	0.8894	0.0518	0.0589
African American	0.9326	0.6413	0.9535	0.0157	0.0308
Hispanic	0.9147	0.6410	0.9417	0.0238	0.0345
White	0.8311	0.6529	0.8832	0.0631	0.0537
Grade 10					
Males	0.9448	0.6764	0.9605	0.0148	0.0247

Grade/Subgroup	Classification Consistency		Classification Accuracy		
	Consistency	Kappa	Accuracy	False Positive Errors	False Negative Errors
Females	0.9204	0.6816	0.9420	0.0225	0.0355
Asian	0.9101	0.7769	0.9368	0.0463	0.0169
African American	0.9397	0.6537	0.9565	0.0160	0.0276
Hispanic	0.9057	0.6315	0.9314	0.0275	0.0411
White	0.8519	0.7030	0.8926	0.0481	0.0593

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	0.9279	0.6721	0.9508	0.0184	0.0308
Females	0.9223	0.6403	0.9467	0.0194	0.0339
Asian	0.8776	0.7242	0.9166	0.0376	0.0459
African American	0.9470	0.5922	0.9643	0.0102	0.0256
Hispanic	0.9100	0.6171	0.9386	0.0221	0.0393
White	0.7983	0.5944	0.8581	0.0739	0.0680
Grade 4					
Males	0.9394	0.7132	0.9582	0.0141	0.0277
Females	0.9325	0.6935	0.9539	0.0160	0.0301
Asian	0.8611	0.7209	0.9024	0.0420	0.0556
African American	0.9505	0.6457	0.9664	0.0099	0.0237
Hispanic	0.9304	0.6567	0.9523	0.0154	0.0324
White	0.8236	0.6471	0.8760	0.0572	0.0668
Grade 5					
Males	0.9332	0.6990	0.9509	0.0206	0.0285
Females	0.9333	0.7007	0.9504	0.0195	0.0301
Asian	0.8258	0.6514	0.8723	0.0839	0.0438
African American	0.9456	0.6624	0.9597	0.0140	0.0263
Hispanic	0.9244	0.6624	0.9435	0.0228	0.0337
White	0.8368	0.6729	0.8809	0.0665	0.0527
Grade 6					
Males	0.9418	0.7519	0.9590	0.0153	0.0257
Females	0.9358	0.7498	0.9548	0.0191	0.0260
Asian	0.8678	0.7336	0.9052	0.0369	0.0579
African American	0.9472	0.7122	0.9629	0.0144	0.0227
Hispanic	0.9175	0.7328	0.9422	0.0225	0.0353
White	0.8788	0.7534	0.9136	0.0410	0.0454
Grade 7					
Males	0.9481	0.7621	0.9637	0.0149	0.0215
Females	0.9297	0.7259	0.9502	0.0180	0.0319
Asian	0.8987	0.7952	0.9260	0.0215	0.0524

Grade/Subgroup	Classification Consistency		Classification Accuracy		
	Consistency	Kappa	Accuracy	False Positive Errors	False Negative Errors
African American	0.9444	0.6911	0.9611	0.0145	0.0245
Hispanic	0.9274	0.6847	0.9481	0.0163	0.0356
White	0.8911	0.7661	0.9216	0.0436	0.0348
Grade 8					
Males	0.9520	0.7598	0.9668	0.0145	0.0188
Females	0.9447	0.7443	0.9610	0.0166	0.0224
Asian	0.8482	0.6806	0.8900	0.0499	0.0600
African American	0.9569	0.7388	0.9700	0.0119	0.0181
Hispanic	0.9339	0.6984	0.9538	0.0217	0.0245
White	0.8666	0.7328	0.9052	0.0516	0.0432
Grade 10					
Males	0.9637	0.7209	0.9730	0.0086	0.0184
Females	0.9632	0.7242	0.9730	0.0076	0.0194
Asian	0.9084	0.8032	0.9350	0.0244	0.0406
African American	0.9702	0.6625	0.9780	0.0061	0.0159
Hispanic	0.9446	0.6930	0.9602	0.0140	0.0258
White	0.8881	0.7726	0.9126	0.0304	0.0569

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	0.9638	0.7191	0.9746	0.0108	0.0147
Females	0.9635	0.7068	0.9753	0.0091	0.0157
Asian	0.8714	0.6893	0.9095	0.0272	0.0633
African American	0.9784	0.6035	0.9852	0.0054	0.0093
Hispanic	0.9583	0.5784	0.9709	0.0090	0.0202
White	0.8314	0.6626	0.8830	0.0569	0.0600
Grade 8					
Males	0.9701	0.7247	0.9802	0.0066	0.0132
Females	0.9698	0.6486	0.9796	0.0053	0.0151
Asian	0.8738	0.6120	0.9161	0.0458	0.0381
African American	0.9788	0.6322	0.9860	0.0037	0.0103
Hispanic	0.9656	0.6320	0.9768	0.0040	0.0192
White	0.8615	0.7042	0.9044	0.0358	0.0598
High School					
Males	0.9823	0.7002	0.9880	0.0047	0.0073
Females	0.9792	0.6251	0.9860	0.0042	0.0097
Asian	0.9320	0.7224	0.9537	0.0333	0.0130
African American	0.9842	0.6036	0.9893	0.0034	0.0073
Hispanic	0.9828	0.6743	0.9890	0.0023	0.0088
White	0.8962	0.7259	0.9302	0.0281	0.0417

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, Hispanic, White

Note. See the subsection *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**
8	MC	ETHNIC	White	0.085	2,737	469	B
22	MC	ETHNIC	White	0.122	2,765	469	B
34	MC	ETHNIC	White	0.041	2,650	469	B
35	MC	ETHNIC	White	0.075	2,723	469	B
37	CR	ETHNIC	White	0.227	2,805	469	B
41	MC	ETHNIC	White	0.062	2,776	469	B
4	MC	ETHNIC	White	0.023	2,783	469	C
10	MC	ETHNIC	White	0.050	2,852	469	C
14	MC	ETHNIC	White	0.078	2,775	469	C
19	MC	ETHNIC	White	0.061	2,608	469	C
29	MC	ETHNIC	White	0.073	2,857	469	C
36	MC	ETHNIC	White	0.119	2,767	469	C
38	MC	ETHNIC	White	0.070	2,807	469	C
40	MC	ETHNIC	White	0.057	2,809	469	C
45	MC	ETHNIC	White	0.076	2,716	469	C
6	MC	ETHNIC	White	-0.071	2,677	469	-C

Note. *DIF is Differential Item Functioning and SMD is Standardized Mean Difference. DIF levels A, B, and C are explained in the subsection 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**
1	MC	GENDER	Female	0.064	2,404	2,376	B
10	MC	ETHNIC	White	0.059	2,669	388	B
16	MC	ETHNIC	White	0.067	2,590	388	B
18	CR	ETHNIC	White	0.228	2,802	388	B
26	MC	ETHNIC	White	0.097	2,805	388	B
32	MC	ETHNIC	White	0.106	2,618	388	B
36	MC	ETHNIC	White	0.064	2,656	388	B
44	MC	ETHNIC	White	0.072	2,716	388	B
33	MC	ETHNIC	Hispanic	-0.066	3,715	603	-B
41	MC	ETHNIC	Hispanic	-0.059	3,689	603	-B
7	MC	ETHNIC	White	0.055	2,716	388	C
11	MC	ETHNIC	White	0.035	2,796	388	C
12	MC	ETHNIC	White	0.175	2,750	388	C
17	MC	ETHNIC	White	0.016	2,538	388	C
45	MC	ETHNIC	White	0.056	2,835	388	C

Note. *DIF is Differential Item Functioning and SMD is Standardized Mean Difference. DIF levels A, B, and C are explained in the subsection 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**
40	MC	ETHNIC	Hispanic	0.030	3,753	604	B
47	MC	ETHNIC	Hispanic	0.055	3,757	604	B
17	MC	ETHNIC	White	0.040	2,959	326	B
22	MC	ETHNIC	White	0.109	2,864	333	B
39	MC	ETHNIC	White	0.056	3,075	326	B
41	MC	ETHNIC	White	0.032	3,068	325	B
48	MC	ETHNIC	White	0.059	2,943	326	B
4	MC	GENDER	Female	-0.069	2,411	2,364	-B
1	MC	ETHNIC	Hispanic	-0.057	3,726	604	-B
20	MC	ETHNIC	Hispanic	-0.058	3,755	604	-B
29	MC	ETHNIC	Hispanic	-0.056	3,755	606	-B
35	MC	ETHNIC	Hispanic	-0.092	3,702	604	-B
2	MC	ETHNIC	White	0.037	3,016	326	C
4	MC	ETHNIC	White	0.047	2,809	326	C
11	MC	ETHNIC	White	0.176	2,716	333	C
13	MC	ETHNIC	White	0.082	2,859	326	C
19	CR	ETHNIC	White	0.452	2,983	333	C
36	MC	ETHNIC	White	0.112	2,853	326	C
6	MC	ETHNIC	Hispanic	-0.064	3,755	604	-C

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

Reading Grade 6							
Operational Item Sequence Number	**Item Type**	**DIF Comparison**	**Focal Group**	**SMD**	**Reference Group N**	**Focal Group N**	**DIF Level**
18	CR	GENDER	Female	0.178	2,222	2,158	B
21	MC	GENDER	Female	0.076	2,222	2,156	B
18	CR	ETHNIC	Hispanic	0.177	3,573	503	B
34	MC	ETHNIC	Hispanic	0.055	3,573	503	B
7	MC	ETHNIC	White	0.065	3,127	254	B
18	CR	ETHNIC	White	0.185	3,157	248	B
20	MC	ETHNIC	White	0.070	3,095	254	B
21	MC	ETHNIC	White	0.050	3,191	254	B
43	MC	ETHNIC	White	0.051	2,906	254	B
27	MC	GENDER	Female	-0.128	2,222	2,156	-B
7	MC	ETHNIC	Hispanic	-0.126	3,573	503	-B
4	MC	ETHNIC	White	-0.037	2,968	254	-B
5	MC	ETHNIC	White	0.083	3,136	254	C
6	MC	ETHNIC	White	0.046	3,016	254	C
14	MC	ETHNIC	White	0.050	3,214	254	C
16	MC	ETHNIC	White	0.081	3,058	254	C
23	MC	ETHNIC	White	0.180	3,022	254	C
27	MC	ETHNIC	White	0.145	3,025	254	C
31	MC	ETHNIC	White	0.115	3,204	254	C
48	MC	ETHNIC	White	0.059	3,083	254	C
37	MC	ETHNIC	Hispanic	-0.101	3,573	503	-C

Note. *DIF is Differential Item Functioning and SMD is Standardized Mean Difference. DIF levels A, B, and C are explained in the subsection 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
48	MC	ETHNIC	Hispanic	0.086	3,634	462	B
22	MC	ETHNIC	White	0.074	2,793	242	B
32	MC	ETHNIC	White	0.077	3,064	242	B
39	MC	ETHNIC	White	0.090	2,783	242	B
36	MC	GENDER	Female	-0.092	2,211	2,201	-B
19	CR	ETHNIC	White	-0.192	2,954	244	-B
40	MC	ETHNIC	White	-0.070	2,994	242	-B
2	MC	ETHNIC	White	0.045	2,847	242	C
5	MC	ETHNIC	White	0.095	2,994	242	C
10	MC	ETHNIC	White	0.041	3,039	242	C
14	MC	ETHNIC	White	0.065	2,890	242	C
18	MC	ETHNIC	White	0.089	3,130	242	C
27	MC	ETHNIC	White	0.063	2,925	242	C
30	MC	ETHNIC	White	0.105	3,023	242	C
35	MC	ETHNIC	White	0.081	2,994	242	C
36	MC	ETHNIC	White	0.117	2,775	242	C
41	MC	ETHNIC	White	0.055	3,017	242	C
43	MC	ETHNIC	White	0.116	3,006	242	C
45	MC	ETHNIC	White	0.078	2,900	242	C
43	MC	GENDER	Female	-0.169	2,207	2,201	-C
44	MC	ETHNIC	Hispanic	-0.106	3,636	462	-C

Note. *DIF is Differential Item Functioning and SMD is Standardized Mean Difference. DIF levels A, B, and C are explained in the subsection 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**
8	CR	GENDER	Female	0.230	2,145	2,136	B
29	MC	ETHNIC	Hispanic	0.117	3,579	415	B
1	MC	ETHNIC	White	0.069	2,654	206	B
5	MC	ETHNIC	White	0.129	2,681	212	B
13	MC	ETHNIC	White	0.074	2,848	212	B
33	MC	ETHNIC	White	0.071	2,972	212	B
45	MC	ETHNIC	White	0.073	2,872	212	B
34	MC	GENDER	Female	-0.114	2,141	2,142	-B
48	MC	ETHNIC	Hispanic	-0.075	3,573	415	-B
42	CR	ETHNIC	White	-0.186	2,867	212	-B
19	CR	GENDER	Female	0.257	2,144	2,142	C
7	MC	ETHNIC	White	0.034	2,684	212	C
12	MC	ETHNIC	White	0.084	2,830	212	C
14	MC	ETHNIC	White	0.039	2,694	212	C
23	MC	ETHNIC	White	0.082	2,889	212	C
28	MC	ETHNIC	White	0.075	2,805	212	C
29	MC	ETHNIC	White	0.163	2,842	212	C
30	MC	ETHNIC	White	0.184	2,843	212	C
34	MC	ETHNIC	White	0.135	2,628	212	C
44	MC	ETHNIC	White	0.107	2,527	212	C

Note. *DIF is Differential Item Functioning and SMD is Standardized Mean Difference. DIF levels A, B, and C are explained in the subsection 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**
2	MC	GENDER	Female	0.026	2,085	2,284	B
42	CR	GENDER	Female	0.194	2,083	2,283	B
7	CR	ETHNIC	Hispanic	0.204	3,670	462	B
17	MC	ETHNIC	Hispanic	0.044	3,668	462	B
42	CR	ETHNIC	Hispanic	0.220	3,671	462	B
28	MC	GENDER	Female	-0.072	2,083	2,283	-B
32	MC	GENDER	Female	-0.073	2,083	2,283	-B
2	MC	ETHNIC	Hispanic	-0.040	3,652	462	-B
6	MC	ETHNIC	Hispanic	-0.047	3,667	462	-B
7	CR	GENDER	Female	0.292	2,083	2,283	C
19	CR	ETHNIC	Hispanic	0.299	3,671	462	C
37	MC	ETHNIC	Hispanic	0.145	3,638	462	C
9	MC	ETHNIC	Hispanic	-0.132	3,655	462	-C
18	MC	ETHNIC	Hispanic	-0.104	3,668	462	-C

Note. *DIF is Differential Item Functioning and SMD is Standardized Mean Difference. DIF levels A, B, and C are explained in the subsection 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**
50	MC	GENDER	Female	0.027	2,459	2,325	B
38	MC	ETHNIC	White	0.036	2,880	472	B
42	MC	ETHNIC	White	0.035	2,749	472	B
46	MC	ETHNIC	White	0.097	2,915	472	B
53	MC	ETHNIC	White	0.071	2,732	472	B
54	MC	ETHNIC	White	0.064	2,823	472	B
24	MC	ETHNIC	White	-0.064	2,718	472	-B
37	CR	ETHNIC	Hispanic	0.271	3,456	690	C
7	MC	ETHNIC	White	0.033	2,755	472	C
37	CR	ETHNIC	White	0.506	2,485	472	C
41	MC	ETHNIC	White	0.050	2,617	472	C
51	MC	ETHNIC	White	0.118	2,738	472	C
52	MC	ETHNIC	White	0.032	2,674	472	C
3	MC	ETHNIC	White	-0.023	2,730	472	-C
44	MC	ETHNIC	White	-0.020	2,762	472	-C

Note. *DIF is Differential Item Functioning and SMD is Standardized Mean Difference. DIF levels A, B, and C are explained in the subsection 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**
17	MC	GENDER	Female	0.032	2,428	2,401	B
13	MC	ETHNIC	Hispanic	0.054	3,725	614	B
20	MC	ETHNIC	White	0.105	2,823	394	B
43	MC	ETHNIC	White	0.047	2,768	394	B
22	MC	ETHNIC	White	-0.049	2,924	394	-B
46	MC	ETHNIC	White	-0.043	2,815	394	-B
11	MC	ETHNIC	White	0.131	2,642	394	C
15	MC	ETHNIC	White	0.044	2,758	394	C
24	CR	ETHNIC	White	0.312	2,728	394	C
35	MC	ETHNIC	White	0.090	2,660	394	C
42	MC	ETHNIC	White	0.080	2,934	394	C
51	MC	ETHNIC	White	0.109	2,840	394	C
54	MC	ETHNIC	White	0.091	2,855	394	C
16	MC	ETHNIC	White	-0.066	2,719	394	-C
28	MC	ETHNIC	White	-0.079	2,912	394	-C

Note. *DIF is Differential Item Functioning and SMD is Standardized Mean Difference. DIF levels A, B, and C are explained in the subsection 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**
52	MC	GENDER	Female	0.042	2,427	2,370	B
23	MC	ETHNIC	White	0.056	3,063	337	B
38	MC	ETHNIC	White	0.071	3,087	337	B
4	MC	ETHNIC	White	-0.046	2,897	337	-B
29	MC	ETHNIC	White	-0.055	3,011	337	-B
30	MC	ETHNIC	White	-0.039	3,126	337	-B
43	MC	ETHNIC	White	-0.058	3,015	337	-B
3	MC	ETHNIC	White	0.205	3,139	337	C
17	MC	ETHNIC	White	0.082	3,137	337	C
20	MC	ETHNIC	White	0.151	2,970	337	C
21	MC	ETHNIC	White	0.084	3,127	337	C
22	MC	ETHNIC	White	0.052	3,030	337	C
24	MC	ETHNIC	White	0.041	3,034	337	C
36	MC	ETHNIC	White	0.068	3,168	337	C
37	CR	ETHNIC	White	0.336	3,070	337	C
44	MC	ETHNIC	White	0.154	3,142	337	C
52	MC	ETHNIC	White	-0.037	2,908	337	-C

Note. *DIF is Differential Item Functioning and SMD is Standardized Mean Difference. DIF levels A, B, and C are explained in the subsection 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**
4	MC	GENDER	Female	0.065	2,240	2,170	B
52	CR	ETHNIC	Hispanic	0.245	3,575	517	B
15	MC	ETHNIC	White	0.051	2,617	254	B
21	MC	ETHNIC	White	0.083	3,051	254	B
22	MC	ETHNIC	White	0.070	2,887	254	B
26	MC	ETHNIC	White	0.101	2,997	254	B
27	CR	ETHNIC	White	0.181	2,986	254	B
38	MC	ETHNIC	White	0.067	2,904	254	B
42	MC	ETHNIC	White	0.130	3,004	254	B
49	MC	ETHNIC	White	0.128	3,013	254	B
50	MC	ETHNIC	White	0.068	2,811	254	B
50	MC	GENDER	Female	-0.091	2,240	2,170	-B
7	MC	ETHNIC	Hispanic	-0.046	3,569	517	-B
4	MC	ETHNIC	White	-0.052	2,996	254	-B
31	CR	ETHNIC	White	-0.179	2,971	254	-B
52	CR	GENDER	Female	0.328	2,240	2,170	C
6	MC	ETHNIC	White	0.059	3,010	254	C
10	MC	ETHNIC	White	0.058	2,862	254	C
35	MC	ETHNIC	White	0.072	3,005	254	C
40	MC	ETHNIC	White	0.165	2,908	254	C
46	MC	ETHNIC	White	0.047	3,122	254	C
54	MC	ETHNIC	White	0.152	3,059	254	C
6	MC	GENDER	Female	-0.112	2,240	2,170	-C
7	MC	ETHNIC	White	-0.025	3,046	254	-C
44	MC	ETHNIC	White	-0.156	2,979	254	-C

Note. *DIF is Differential Item Functioning and SMD is Standardized Mean Difference. DIF levels A, B, and C are explained in the subsection 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**
2	MC	GENDER	Female	0.062	2,216	2,211	B
33	CR	ETHNIC	Hispanic	0.193	3,632	483	B
49	MC	ETHNIC	Hispanic	0.106	3,627	483	B
7	MC	ETHNIC	White	0.055	3,007	245	B
10	MC	ETHNIC	White	0.041	3,072	245	B
19	MC	ETHNIC	White	0.049	3,122	245	B
25	MC	ETHNIC	White	0.051	3,204	245	B
30	MC	ETHNIC	White	0.065	3,105	245	B
33	CR	ETHNIC	White	0.192	2,893	245	B
49	MC	ETHNIC	White	0.070	2,920	245	B
28	MC	ETHNIC	White	-0.049	3,295	245	-B
42	MC	ETHNIC	White	-0.036	2,752	245	-B
4	MC	ETHNIC	White	0.170	2,994	245	C
8	MC	ETHNIC	White	0.152	3,122	245	C
12	MC	ETHNIC	White	0.078	2,919	245	C
31	MC	ETHNIC	White	0.148	3,185	245	C
53	MC	ETHNIC	White	0.117	3,210	245	C
9	MC	ETHNIC	White	-0.019	3,174	245	-C
44	MC	ETHNIC	White	-0.073	3,090	245	-C
52	MC	ETHNIC	White	-0.031	3,111	245	-C

Note. *DIF is Differential Item Functioning and SMD is Standardized Mean Difference. DIF levels A, B, and C are explained in the subsection 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
5	MC	ETHNIC	White	0.067	2,775	213	B
42	MC	ETHNIC	White	0.081	2,928	213	B
53	MC	ETHNIC	White	0.095	2,630	213	B
54	MC	ETHNIC	White	0.113	2,710	213	B
41	MC	GENDER	Female	-0.096	2,173	2,159	-B
26	MC	ETHNIC	Hispanic	-0.054	3,568	447	-B
37	MC	ETHNIC	Hispanic	-0.106	3,525	447	-B
11	MC	ETHNIC	White	-0.103	3,059	213	-B
30	MC	ETHNIC	White	-0.098	2,921	213	-B
1	MC	ETHNIC	White	0.042	2,672	213	C
2	MC	ETHNIC	White	0.059	2,927	213	C
7	MC	ETHNIC	White	0.281	3,076	213	C
39	MC	ETHNIC	White	0.165	2,844	213	C
22	MC	ETHNIC	White	-0.124	2,866	213	-C

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

Mathematics Grade 10							
Operational Item Sequence Number	Item Type	DIF Comparison	Focal Group	SMD	Reference Group N	Focal Group N	DIF Level
14	MC	ETHNIC	Hispanic	0.076	3,616	458	B
16	MC	GENDER	Female	-0.102	2,072	2,274	-B

Note. *DIF is Differential Item Functioning and SMD is Standardized Mean Difference. DIF levels A, B, and C are explained in the subsection 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
25	MC	ETHNIC	Hispanic	0.094	3,717	609	B
4	MC	ETHNIC	White	0.045	3,125	330	B
10	CR	ETHNIC	White	0.190	3,127	331	B
13	MC	ETHNIC	White	0.091	3,146	330	B
40	MC	ETHNIC	White	0.075	3,468	330	B
44	MC	ETHNIC	White	0.140	3,315	319	B
49	MC	ETHNIC	White	0.081	3,377	330	B
11	MC	ETHNIC	White	0.078	3,458	330	C
16	MC	ETHNIC	White	0.082	3,304	330	C
17	MC	ETHNIC	White	0.055	3,106	330	C
18	MC	ETHNIC	White	0.194	3,110	318	C
35	MC	ETHNIC	White	0.162	3,325	330	C
43	MC	ETHNIC	White	0.100	3,155	330	C

Note. *DIF is Differential Item Functioning and SMD is Standardized Mean Difference. DIF levels A, B, and C are explained in the subsection 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
8	MC	ETHNIC	White	0.142	2,778	200	B
15	MC	ETHNIC	White	-0.134	2,971	200	-B

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

High School Biology							
Operational Item Sequence Number	Item Type	DIF Comparison	Focal Group	SMD	Reference Group N	Focal Group N	DIF Level
46	MC	GENDER	Female	0.112	1,736	1,941	B
10	MC	ETHNIC	Hispanic	-0.094	3,111	430	-B
44	MC	GENDER	Female	0.103	1,736	1,941	C

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

Appendix G: Operational Item Adjusted *P* Values

Table G1. DC CAS 2011 Operational Form Item Characteristics: Reading

Reading Grade 3								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	4,771	1	0.88		25	4,749	1	0.76
2	4,762	1	0.81		26	4,742	1	0.78
3	4,765	1	0.67		27	4,740	1	0.71
4	4,768	1	0.83		28	4,741	1	0.88
5	4,757	1	0.71		29	4,758	1	0.61
6	4,758	1	0.78		30	4,749	1	0.86
7	4,757	1	0.79		31	4,624	1	0.82
8	4,742	1	0.43		32	4,753	1	0.84
9	4,631	3	0.37		33	4,744	1	0.79
10	4,763	1	0.70		34	4,730	1	0.75
11	4,747	1	0.73		35	4,751	1	0.62
12	4,722	1	0.82		36	4,739	1	0.64
13	4,663	1	0.55		37	4,684	3	0.42
14	4,627	1	0.62		38	4,732	1	0.64
15	4,736	1	0.51		39	4,728	1	0.63
16	4,681	1	0.70		40	4,740	1	0.72
17	4,598	1	0.59		41	4,734	1	0.57
18	4,743	1	0.74		42	4,732	1	0.71
19	4,751	1	0.70		43	4,622	1	0.68
20	4,635	3	0.51		44	4,669	1	0.61
21	4,755	1	0.44		45	4,664	1	0.63
22	4,753	1	0.33		46	4,619	1	0.67
23	4,753	1	0.82		47	4,662	1	0.70
24	4,735	1	0.62		48	4,658	1	0.73

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

Reading Grade 4								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	4,816	1	0.74		25	4,793	1	0.64
2	4,815	1	0.84		26	4,797	1	0.55
3	4,813	1	0.69		27	4,798	1	0.60
4	4,813	1	0.76		28	4,812	1	0.47
5	4,804	1	0.61		29	4,809	1	0.71
6	4,801	1	0.66		30	4,809	1	0.64
7	4,805	1	0.64		31	4,808	1	0.50
8	4,783	1	0.54		32	4,807	1	0.45
9	4,730	3	0.42		33	4,806	1	0.73
10	4,808	1	0.48		34	4,806	1	0.69
11	4,812	1	0.68		35	4,807	1	0.55
12	4,800	1	0.36		36	4,789	1	0.58
13	4,803	1	0.74		37	4,756	3	0.50
14	4,797	1	0.66		38	4,755	1	0.78
15	4,797	1	0.71		39	4,752	1	0.50
16	4,783	1	0.50		40	4,754	1	0.81
17	4,757	1	0.89		41	4,737	1	0.81
18	4,747	3	0.46		42	4,771	1	0.76
19	4,807	1	0.80		43	4,770	1	0.65
20	4,804	1	0.54		44	4,766	1	0.48
21	4,801	1	0.63		45	4,753	1	0.65
22	4,805	1	0.53		46	4,761	1	0.26
23	4,801	1	0.60		47	4,759	1	0.45
24	4,801	1	0.71		48	4,754	1	0.62

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

Reading Grade 5

Operational Item Sequence Number	N	Max Points	Adjusted *P* Value		Operational Item Sequence Number	N	Max Points	Adjusted *P* Value
1	4,788	1	0.82		25	4,781	1	0.75
2	4,788	1	0.70		26	4,782	1	0.70
3	4,788	1	0.83		27	4,774	1	0.65
4	4,786	1	0.70		28	4,781	1	0.53
5	4,786	1	0.72		29	4,781	1	0.84
6	4,785	1	0.86		30	4,781	1	0.87
7	4,782	1	0.51		31	4,779	1	0.84
8	4,765	1	0.76		32	4,776	1	0.84
9	4,716	3	0.44		33	4,780	1	0.84
10	4,787	1	0.79		34	4,778	1	0.84
11	4,784	1	0.35		35	4,777	1	0.58
12	4,781	1	0.71		36	4,753	1	0.57
13	4,784	1	0.53		37	4,737	3	0.43
14	4,784	1	0.82		38	4,762	1	0.75
15	4,775	1	0.44		39	4,763	1	0.54
16	4,784	1	0.71		40	4,761	1	0.87
17	4,770	1	0.73		41	4,756	1	0.69
18	4,741	1	0.69		42	4,751	1	0.61
19	4,668	3	0.28		43	4,762	1	0.79
20	4,782	1	0.80		44	4,758	1	0.73
21	4,782	1	0.79		45	4,759	1	0.82
22	4,778	1	0.34		46	4,757	1	0.63
23	4,781	1	0.77		47	4,754	1	0.85
24	4,778	1	0.67		48	4,740	1	0.58

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

Reading Grade 6

Operational Item Sequence Number	N	Max Points	Adjusted *P* Value		Operational Item Sequence Number	N	Max Points	Adjusted *P* Value
1	4,390	1	0.58		25	4,384	1	0.78
2	4,389	1	0.68		26	4,386	1	0.53
3	4,391	1	0.54		27	4,366	1	0.48
4	4,387	1	0.71		28	4,317	3	0.54
5	4,378	1	0.68		29	4,384	1	0.72
6	4,388	1	0.75		30	4,383	1	0.77
7	4,350	1	0.51		31	4,371	1	0.46
8	4,334	3	0.52		32	4,382	1	0.90
9	4,386	1	0.73		33	4,381	1	0.72
10	4,387	1	0.45		34	4,383	1	0.74
11	4,383	1	0.53		35	4,375	1	0.58
12	4,383	1	0.56		36	4,377	1	0.68
13	4,381	1	0.77		37	4,382	1	0.73
14	4,359	1	0.72		38	4,382	1	0.87
15	4,371	1	0.66		39	4,383	1	0.67
16	4,367	1	0.63		40	4,382	1	0.76
17	4,350	1	0.73		41	4,376	1	0.60
18	4,275	3	0.28		42	4,382	1	0.66
19	4,386	1	0.82		43	4,376	1	0.68
20	4,387	1	0.55		44	4,339	1	0.52
21	4,385	1	0.64		45	4,335	1	0.55
22	4,382	1	0.73		46	4,340	1	0.70
23	4,383	1	0.50		47	4,339	1	0.66
24	4,383	1	0.69		48	4,334	1	0.66

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

Reading Grade 7

Operational Item Sequence Number	N	Max Points	Adjusted *P* Value		Operational Item Sequence Number	N	Max Points	Adjusted *P* Value
1	4,436	1	0.81		25	4,427	1	0.74
2	4,439	1	0.74		26	4,420	1	0.70
3	4,429	1	0.63		27	4,421	1	0.53
4	4,431	1	0.64		28	4,417	1	0.60
5	4,436	1	0.79		29	4,413	1	0.84
6	4,421	1	0.72		30	4,414	1	0.49
7	4,395	3	0.47		31	4,410	1	0.57
8	4,430	1	0.89		32	4,417	1	0.61
9	4,426	1	0.87		33	4,412	1	0.59
10	4,427	1	0.78		34	4,407	1	0.57
11	4,427	1	0.87		35	4,413	1	0.63
12	4,424	1	0.51		36	4,398	1	0.44
13	4,416	1	0.64		37	4,371	3	0.47
14	4,417	1	0.62		38	4,410	1	0.66
15	4,415	1	0.64		39	4,409	1	0.59
16	4,404	1	0.83		40	4,407	1	0.80
17	4,402	1	0.88		41	4,399	1	0.75
18	4,373	1	0.60		42	4,408	1	0.61
19	4,343	3	0.54		43	4,409	1	0.58
20	4,425	1	0.66		44	4,408	1	0.79
21	4,420	1	0.58		45	4,372	1	0.61
22	4,416	1	0.60		46	4,374	1	0.46
23	4,421	1	0.71		47	4,373	1	0.68
24	4,425	1	0.79		48	4,368	1	0.71

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

Reading Grade 8

Operational Item Sequence Number	N	Max Points	Adjusted *P* Value		Operational Item Sequence Number	N	Max Points	Adjusted *P* Value
1	4,306	1	0.54		25	4,288	1	0.65
2	4,295	1	0.63		26	4,290	1	0.62
3	4,307	1	0.73		27	4,286	1	0.80
4	4,303	1	0.54		28	4,284	1	0.66
5	4,297	1	0.37		29	4,286	1	0.53
6	4,302	1	0.87		30	4,286	1	0.43
7	4,287	1	0.80		31	4,286	1	0.60
8	4,196	3	0.61		32	4,289	1	0.64
9	4,302	1	0.62		33	4,282	1	0.50
10	4,303	1	0.49		34	4,285	1	0.44
11	4,300	1	0.67		35	4,278	1	0.26
12	4,296	1	0.63		36	4,286	1	0.29
13	4,298	1	0.54		37	4,285	1	0.65
14	4,290	1	0.66		38	4,282	1	0.25
15	4,284	1	0.74		39	4,280	1	0.77
16	4,279	1	0.53		40	4,284	1	0.59
17	4,278	1	0.57		41	4,278	1	0.44
18	4,255	1	0.55		42	4,169	3	0.49
19	4,165	3	0.50		43	4,248	1	0.85
20	4,291	1	0.84		44	4,246	1	0.59
21	4,291	1	0.70		45	4,242	1	0.43
22	4,292	1	0.59		46	4,241	1	0.62
23	4,288	1	0.64		47	4,247	1	0.39
24	4,291	1	0.76		48	4,246	1	0.81

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

Reading Grade 10

Operational Item Sequence Number	N	Max Points	Adjusted *P* Value		Operational Item Sequence Number	N	Max Points	Adjusted *P* Value
1	4,433	1	0.72		25	4,382	1	0.61
2	4,441	1	0.89		26	4,366	1	0.45
3	4,424	1	0.60		27	4,374	1	0.61
4	4,437	1	0.72		28	4,373	1	0.72
5	4,438	1	0.71		29	4,368	1	0.48
6	4,419	1	0.84		30	4,351	1	0.46
7	4,153	3	0.51		31	4,355	1	0.72
8	4,424	1	0.84		32	4,353	1	0.60
9	4,424	1	0.81		33	4,347	1	0.69
10	4,419	1	0.84		34	4,351	1	0.56
11	4,420	1	0.66		35	4,337	1	0.49
12	4,421	1	0.76		36	4,347	1	0.60
13	4,417	1	0.81		37	4,342	1	0.26
14	4,418	1	0.69		38	4,347	1	0.50
15	4,407	1	0.66		39	4,344	1	0.57
16	4,401	1	0.68		40	4,344	1	0.74
17	4,396	1	0.81		41	4,328	1	0.59
18	4,393	1	0.74		42	3,949	3	0.52
19	4,055	3	0.66		43	4,271	1	0.55
20	4,383	1	0.87		44	4,273	1	0.72
21	4,382	1	0.57		45	4,273	1	0.63
22	4,358	1	0.46		46	4,270	1	0.72
23	4,380	1	0.31		47	4,268	1	0.40
24	4,374	1	0.70		48	4,267	1	0.63

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

Table G2. DC CAS 2011 Operational Form Item Characteristics: Mathematics

Mathematics Grade 3								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	4,788	1	0.58		28	4,788	1	0.77
2	4,797	1	0.56		29	4,777	1	0.81
3	4,777	1	0.92		30	4,714	1	0.88
4	4,773	1	0.61		31	4,780	1	0.78
5	4,796	1	0.61		32	4,779	1	0.66
6	4,793	1	0.80		33	4,755	1	0.61
7	4,780	1	0.69		34	4,778	1	0.67
8	4,775	1	0.77		35	4,774	1	0.74
9	4,775	1	0.48		36	4,767	1	0.63
10	4,787	1	0.79		37	4,739	3	0.37
11	4,785	1	0.94		38	4,654	1	0.68
12	4,767	1	0.48		39	4,764	1	0.52
13	4,717	1	0.78		40	4,777	1	0.94
14	4,702	1	0.75		41	4,780	1	0.85
15	4,790	1	0.84		42	4,787	1	0.78
16	4,787	1	0.72		43	4,788	1	0.68
17	4,789	1	0.78		44	4,779	1	0.86
18	4,763	1	0.92		45	4,765	1	0.68
19	4,769	1	0.80		46	4,769	1	0.46
20	4,770	1	0.88		47	4,752	1	0.85
21	4,791	1	0.86		48	4,762	1	0.63
22	4,739	1	0.65		49	4,733	3	0.30
23	4,754	3	0.56		50	4,742	1	0.94
24	4,785	1	0.52		51	4,769	1	0.48
25	4,763	1	0.85		52	4,775	1	0.81
26	4,771	1	0.64		53	4,771	1	0.63
27	4,779	1	0.75		54	4,736	1	0.64

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

Mathematics Grade 4

Operational Item Sequence Number	N	Max Points	Adjusted *P* Value		Operational Item Sequence Number	N	Max Points	Adjusted *P* Value
1	4,853	1	0.89		28	4,846	1	0.68
2	4,846	1	0.70		29	4,848	1	0.66
3	4,847	1	0.70		30	4,842	1	0.55
4	4,856	1	0.73		31	4,840	1	0.90
5	4,853	1	0.53		32	4,837	1	0.60
6	4,849	1	0.52		33	4,818	1	0.57
7	4,856	1	0.27		34	4,820	3	0.40
8	4,855	1	0.68		35	4,844	1	0.65
9	4,850	1	0.76		36	4,842	1	0.52
10	4,848	1	0.54		37	4,846	1	0.90
11	4,831	1	0.38		38	4,845	1	0.78
12	4,836	1	0.70		39	4,842	1	0.73
13	4,833	1	0.75		40	4,826	1	0.69
14	4,805	1	0.55		41	4,836	1	0.37
15	4,853	1	0.54		42	4,835	1	0.52
16	4,852	1	0.69		43	4,832	1	0.56
17	4,853	1	0.92		44	4,830	1	0.27
18	4,845	1	0.86		45	4,832	1	0.68
19	4,849	1	0.51		46	4,827	1	0.72
20	4,841	1	0.46		47	4,823	1	0.57
21	4,847	1	0.77		48	4,795	1	0.83
22	4,834	1	0.66		49	4,787	3	0.49
23	4,807	1	0.80		50	4,831	1	0.91
24	4,816	3	0.43		51	4,823	1	0.64
25	4,840	1	0.68		52	4,824	1	0.32
26	4,837	1	0.81		53	4,813	1	0.62
27	4,840	1	0.51		54	4,815	1	0.49

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

Mathematics Grade 5								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	4,800	1	0.61		28	4,799	1	0.80
2	4,811	1	0.82		29	4,798	1	0.57
3	4,804	1	0.42		30	4,799	1	0.75
4	4,809	1	0.67		31	4,799	1	0.79
5	4,803	1	0.71		32	4,797	1	0.59
6	4,799	1	0.55		33	4,797	1	0.38
7	4,806	1	0.65		34	4,789	1	0.61
8	4,800	1	0.62		35	4,798	1	0.89
9	4,804	1	0.78		36	4,783	1	0.59
10	4,790	1	0.29		37	4,771	3	0.39
11	4,799	1	0.67		38	4,798	1	0.47
12	4,803	1	0.88		39	4,798	1	0.74
13	4,803	1	0.58		40	4,783	1	0.87
14	4,784	1	0.89		41	4,789	1	0.60
15	4,808	1	0.94		42	4,789	1	0.70
16	4,805	1	0.84		43	4,782	1	0.64
17	4,795	1	0.55		44	4,786	1	0.45
18	4,800	1	0.60		45	4,778	1	0.64
19	4,733	3	0.38		46	4,777	3	0.81
20	4,802	1	0.46		47	4,794	1	0.59
21	4,801	1	0.78		48	4,796	1	0.61
22	4,799	1	0.64		49	4,795	1	0.85
23	4,805	1	0.62		50	4,796	1	0.63
24	4,806	1	0.75		51	4,797	1	0.81
25	4,801	1	0.58		52	4,795	1	0.89
26	4,806	1	0.45		53	4,797	1	0.30
27	4,808	1	0.84		54	4,798	1	0.72

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

Mathematics Grade 6								
Operational Item Sequence Number	N	Max Points	Adjusted *P* Value		Operational Item Sequence Number	N	Max Points	Adjusted *P* Value
1	4,416	1	0.42		28	4,401	1	0.50
2	4,420	1	0.79		29	4,394	1	0.34
3	4,420	1	0.88		30	4,388	1	0.51
4	4,413	1	0.72		31	4,345	3	0.38
5	4,416	1	0.70		32	4,396	1	0.66
6	4,418	1	0.68		33	4,405	1	0.72
7	4,420	1	0.87		34	4,403	1	0.77
8	4,413	1	0.69		35	4,401	1	0.59
9	4,408	1	0.44		36	4,403	1	0.72
10	4,417	1	0.54		37	4,404	1	0.71
11	4,419	1	0.63		38	4,401	1	0.28
12	4,414	1	0.38		39	4,401	1	0.43
13	4,410	1	0.32		40	4,401	1	0.47
14	4,401	1	0.67		41	4,387	1	0.39
15	4,417	1	0.62		42	4,379	1	0.30
16	4,412	1	0.62		43	4,386	1	0.78
17	4,411	1	0.61		44	4,388	1	0.56
18	4,419	1	0.65		45	4,390	1	0.34
19	4,415	1	0.79		46	4,385	1	0.63
20	4,411	1	0.52		47	4,378	1	0.34
21	4,416	1	0.45		48	4,382	1	0.69
22	4,413	1	0.58		49	4,380	1	0.35
23	4,418	1	0.69		50	4,368	1	0.46
24	4,415	1	0.73		51	4,354	1	0.70
25	4,401	1	0.54		52	4,335	3	0.55
26	4,367	1	0.52		53	4,388	1	0.65
27	4,347	3	0.43		54	4,387	1	0.38

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

Mathematics Grade 7								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	4,450	1	0.61		28	4,431	1	0.61
2	4,457	1	0.85		29	4,434	1	0.79
3	4,453	1	0.77		30	4,400	1	0.36
4	4,452	1	0.47		31	4,425	1	0.61
5	4,423	1	0.42		32	4,390	1	0.64
6	4,449	1	0.47		33	4,357	3	0.37
7	4,451	1	0.70		34	4,435	1	0.49
8	4,433	1	0.38		35	4,431	1	0.52
9	4,450	1	0.85		36	4,433	1	0.63
10	4,451	1	0.53		37	4,428	1	0.54
11	4,438	1	0.37		38	4,429	1	0.47
12	4,441	1	0.44		39	4,430	1	0.63
13	4,440	1	0.71		40	4,414	1	0.66
14	4,370	1	0.32		41	4,402	1	0.57
15	4,438	1	0.79		42	4,410	1	0.63
16	4,434	1	0.54		43	4,414	1	0.66
17	4,442	1	0.79		44	4,410	1	0.80
18	4,442	1	0.83		45	4,402	1	0.66
19	4,417	1	0.60		46	4,409	1	0.50
20	4,315	1	0.78		47	4,376	1	0.35
21	4,389	3	0.73		48	4,314	3	0.52
22	4,442	1	0.48		49	4,413	1	0.54
23	4,441	1	0.79		50	4,409	1	0.56
24	4,438	1	0.49		51	4,404	1	0.42
25	4,435	1	0.48		52	4,412	1	0.77
26	4,439	1	0.63		53	4,408	1	0.31
27	4,440	1	0.75		54	4,404	1	0.69

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

Mathematics Grade 8								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	4,352	1	0.68		28	4,308	1	0.49
2	4,338	1	0.59		29	4,324	1	0.78
3	4,351	1	0.32		30	4,317	1	0.38
4	4,338	1	0.48		31	4,313	1	0.77
5	4,346	1	0.55		32	4,277	1	0.55
6	4,325	1	0.41		33	4,239	3	0.42
7	4,340	1	0.33		34	4,328	1	0.54
8	4,338	1	0.50		35	4,316	1	0.33
9	4,337	1	0.41		36	4,327	1	0.88
10	4,343	1	0.60		37	4,297	1	0.43
11	4,326	1	0.37		38	4,320	1	0.54
12	4,303	1	0.49		39	4,326	1	0.51
13	4,341	1	0.33		40	4,307	1	0.46
14	4,334	1	0.57		41	4,300	1	0.65
15	4,339	1	0.67		42	4,286	1	0.44
16	4,330	1	0.44		43	4,293	1	0.41
17	4,314	1	0.33		44	4,288	1	0.50
18	4,315	1	0.52		45	4,289	1	0.61
19	4,263	3	0.58		46	4,297	1	0.66
20	4,335	1	0.51		47	4,252	1	0.53
21	4,345	1	0.53		48	4,156	3	0.40
22	4,332	1	0.37		49	4,285	1	0.48
23	4,339	1	0.61		50	4,288	1	0.53
24	4,339	1	0.81		51	4,293	1	0.55
25	4,335	1	0.71		52	4,288	1	0.43
26	4,343	1	0.86		53	4,292	1	0.43
27	4,313	1	0.40		54	4,284	1	0.53

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

Mathematics Grade 10								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	4,407	1	0.64		28	4,284	1	0.35
2	4,372	1	0.48		29	4,313	1	0.77
3	4,400	1	0.66		30	4,285	1	0.51
4	4,397	1	0.46		31	3,920	3	0.43
5	4,382	1	0.34		32	4,300	1	0.42
6	4,407	1	0.81		33	4,313	1	0.37
7	4,397	1	0.60		34	4,316	1	0.57
8	4,332	1	0.21		35	4,310	1	0.41
9	4,394	1	0.47		36	4,317	1	0.32
10	4,379	1	0.33		37	4,298	1	0.32
11	4,311	1	0.44		38	4,319	1	0.57
12	4,353	1	0.59		39	4,304	1	0.44
13	4,352	1	0.55		40	4,258	1	0.76
14	4,351	1	0.73		41	4,268	1	0.68
15	4,340	1	0.21		42	4,253	1	0.30
16	4,343	1	0.42		43	4,244	1	0.19
17	4,077	3	0.54		44	4,235	1	0.42
18	4,348	1	0.51		45	4,251	1	0.45
19	4,347	1	0.63		46	4,245	1	0.59
20	4,337	1	0.43		47	3,561	3	0.19
21	4,341	1	0.45		48	4,257	1	0.40
22	4,344	1	0.52		49	4,238	1	0.49
23	4,341	1	0.42		50	4,254	1	0.51
24	4,347	1	0.63		51	4,257	1	0.38
25	4,350	1	0.65		52	4,256	1	0.55
26	4,322	1	0.62		53	4,238	1	0.32
27	4,320	1	0.64		54	4,254	1	0.50

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

Table G3. DC CAS 2011 Operational Form Item Characteristics: Science/Biology

Science Grade 5								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	4,759	1	0.46		26	4,743	1	0.71
2	4,760	1	0.52		27	4,750	1	0.33
3	4,758	1	0.41		28	4,740	1	0.40
4	4,760	1	0.45		29	4,745	1	0.66
5	4,748	1	0.34		30	4,732	1	0.41
6	4,753	1	0.68		31	4,732	1	0.62
7	4,752	1	0.33		32	4,727	1	0.35
8	4,744	1	0.27		33	4,710	1	0.52
9	4,707	1	0.41		34	4,688	1	0.33
10	4,677	2	0.26		35	4,690	1	0.28
11	4,752	1	0.57		36	4,677	1	0.34
12	4,747	1	0.63		37	4,676	1	0.59
13	4,735	1	0.43		38	4,593	2	0.18
14	4,740	1	0.52		39	4,697	1	0.43
15	4,737	1	0.34		40	4,697	1	0.53
16	4,737	1	0.51		41	4,696	1	0.83
17	4,731	1	0.54		42	4,697	1	0.50
18	4,745	1	0.31		43	4,692	1	0.34
19	4,727	1	0.31		44	4,687	1	0.37
20	4,669	2	0.65		45	4,699	1	0.26
21	4,757	1	0.66		46	4,699	1	0.73
22	4,749	1	0.54		47	4,688	1	0.33
23	4,755	1	0.77		48	4,689	1	0.37
24	4,753	1	0.55		49	4,693	1	0.43
25	4,752	1	0.61		50	4,688	1	0.61

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

Science Grade 8

Operational Item Sequence Number	N	Max Points	Adjusted *P* Value		Operational Item Sequence Number	N	Max Points	Adjusted *P* Value
1	4,208	1	0.55		26	4,166	1	0.62
2	4,205	1	0.41		27	4,165	1	0.41
3	4,199	1	0.30		28	4,171	1	0.84
4	4,207	1	0.35		29	4,164	1	0.48
5	4,202	1	0.33		30	4,133	1	0.29
6	4,207	1	0.75		31	4,130	1	0.35
7	4,202	1	0.37		32	4,132	1	0.34
8	4,186	1	0.28		33	4,133	1	0.33
9	4,171	1	0.41		34	4,149	1	0.58
10	3,790	2	0.20		35	4,146	1	0.41
11	4,191	1	0.58		36	4,139	1	0.38
12	4,187	1	0.56		37	4,141	1	0.45
13	4,185	1	0.67		38	4,141	1	0.33
14	4,166	1	0.29		39	3,759	2	0.32
15	4,169	1	0.50		40	4,149	1	0.40
16	4,170	1	0.37		41	4,150	1	0.35
17	4,149	1	0.30		42	4,146	1	0.36
18	4,170	1	0.39		43	4,148	1	0.38
19	4,181	1	0.42		44	4,150	1	0.51
20	4,163	1	0.45		45	4,148	1	0.51
21	3,940	2	0.49		46	4,147	1	0.25
22	4,179	1	0.55		47	4,148	1	0.51
23	4,169	1	0.44		48	4,145	1	0.40
24	4,172	1	0.36		49	4,146	1	0.32
25	4,174	1	0.30		50	4,139	1	0.43

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

High School Biology

Operational Item Sequence Number	N	Max Points	Adjusted *P* Value		Operational Item Sequence Number	N	Max Points	Adjusted *P* Value
1	3,750	1	0.14		26	3,710	1	0.63
2	3,749	1	0.32		27	3,710	1	0.43
3	3,740	1	0.20		28	3,706	1	0.37
4	3,756	1	0.47		29	3,697	1	0.34
5	3,750	1	0.45		30	3,693	1	0.43
6	3,750	1	0.36		31	3,695	1	0.39
7	3,740	1	0.26		32	3,679	1	0.33
8	3,746	1	0.53		33	2,888	2	0.18
9	3,740	1	0.24		34	3,676	1	0.39
10	3,744	1	0.38		35	3,672	1	0.38
11	3,734	1	0.34		36	3,676	1	0.32
12	3,738	1	0.37		37	3,666	1	0.36
13	3,737	1	0.24		38	3,670	1	0.34
14	3,732	1	0.55		39	3,674	1	0.40
15	3,731	1	0.67		40	3,671	1	0.38
16	3,732	1	0.49		41	3,663	1	0.40
17	3,731	1	0.36		42	3,200	2	0.29
18	3,708	1	0.25		43	3,661	1	0.26
19	3,683	1	0.71		44	3,670	1	0.73
20	3,385	2	0.30		45	3,664	1	0.30
21	3,709	1	0.24		46	3,667	1	0.44
22	3,711	1	0.36		47	3,672	1	0.39
23	3,702	1	0.32		48	3,660	1	0.31
24	3,717	1	0.45		49	3,664	1	0.45
25	3,710	1	0.62		50	3,657	1	0.40

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