Technical Report
Spring 2012 Test Administration

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

December 31, 2012

CTB/McGraw-Hill
Monterey, California 93940

Table of Contents

List of Tables

Section 1. Overview

The primary purpose of the DC CAS is to measure the mastery of Reading, Mathematics, Science, Biology, and Composition content standards of all District of Columbia (DC) public school students annually. The assessments provide the foundation for an accountability system that enables the District 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 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.

This document describes the operational District of Columbia Comprehensive Assessment System (DC CAS) that was administered to students in the spring of 2012 to assess students' skills in Grades 2–10 Reading; Grades 2–8 and 10 Mathematics; Grades 5 and 8 Science and high school Biology; and Grades 4, 7, and 10 Composition. The DC CAS consists of multiple choice (MC) and constructed response (CR) items in Reading, Mathematics, and Science/Biology, and writing prompts for Composition. All items are administered under standardized conditions, where students are allowed accommodations when eligible.

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. Furthermore, the evidence provided in this report is directly relevant to the *Standards and Assessments Peer Review Guidance*, Critical Elements (January 12, 2009; see http://www.ed.gov/policy/elsec/guid/saaprguidance.pdf).

This technical report is written to document procedures and results from developing, analyzing, and validating the 2012 DC CAS. It provides information relevant to an evaluation of the validity of intended interpretations and uses of results from the 2012 DC CAS tests. The design of the test administration, content development and forms construction, classical item analysis, differential item functioning (DIF), item response theory analyses (IRT), and proficiency level data are provided.

Section 2. Item and Test Development

This section contains information relevant to the *Standards and Assessments Peer Review Guidance*, Critical Elements 4.1 and 5.1:

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)?

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?

Overview

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 item and test specifications. Test development involves creating a design framework from the statement of the achievement construct to be measured. Design elements include numbers and types of items and score points allocated to each content strand in each content area test.

Content Standards

The DC CAS tests are aligned to either DC Content Standards, Common Core State Standards (CCSS), or to both for content areas transitioning to the common core. The standards serve as reporting categories. Reading content is fully aligned to the CCSS in Reading. Mathematics content has been transitioning to the CCSS and remains partially aligned to the DC Content Standards, which serve as the reporting categories. In 2013, the Mathematics content will be fully aligned to the CCSS. in Mathematics. In 2012, the total numbers of operational items included in the Science/Biology test design remained the same, although they were distributed differently under the new standard headings.

Item Development

Each year, newly developed items are field tested in DC CAS in all grades and content areas. These items are developed by CTB and, prior to field testing, go through a rigorous content and psychometric review and approval process. CTB content and style editors, supervisors, and managers review all items and then provide items to participants in Content and Bias/Sensitivity Review workshops conducted in DC. OSSE invited educators, members of the DC Public Schools

(DCPS) administration, and community representatives to participate in workshops to review the items. A training session was provided by CTB, after which the participants reviewed all items for content and grade appropriateness, as well as for alignment to the content standards, then rated each item for acceptance, revision, or rejection. The reviewers used the criteria in the checklist in Appendix A to guide their rating decisions.

Test Development

CTB's Research and Development teams, with the approval of the OSSE, developed test forms designed to measure student performance through both multiple choice (MC) and constructed response (CR) item types. The total number of items and score points included n each reporting category serves as the test blueprint, details of which are provided in Tables 1–4.

The items that are available for selection in the DC CAS 2012 assessments originated from a pool of operational and formerly field tested items from the 2011 DC CAS administration, with a small number of items selected from older pools (excluding 2009). The Grade 2 Reading and Mathematics items and the Grade 9 Reading items originated from CTB-owned items in the *TerraNova*™ item pool.

DC CAS assessments are equated each year so that, from one form and year to the next, student scores remain comparable. The equating process requires a set of items to link or anchor one year to the next. These anchor items are a small subset of items that, proportionally, reflect the overall test blueprints. They are typically selected first, around which the remaining operational items are selected for each form.

The forms were assembled by CTB Development staff who attended to the blueprint requirements, as well as various other content and psychometric requirements, such as test length, score points, item types, statistical comparability, and quality. For example, all proposed selections for operational forms were compared to previous DC CAS test forms to ensure they remain parallel and comparable in terms of test difficulty and coverage of the DC CAS content standards, as specified in the 2012 test blueprints.

Once the forms were assembled, they went through an iterative review and approval process where they were reviewed and approved by CTB Research, and then by OSSE. The items were reviewed for content standards alignment and appropriateness by CTB test developers and by OSSE.

Test Design

The DC CAS tests are designed as operational tests with embedded field test items. In this way, newly developed items can be field tested in and amongst operational items. This is an advantage over separate field test designs that highlight the items that do not "count" towards students' scores and can decrease the motivation of their serious effort and response.

To maximize the number of items field tested while minimally impacting the testing time required, two forms were developed for all grades and content areasexcept Composition. Each form included the same core set of operational items (which comprised the 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 across all students in the District (i.e., no sample was drawn).

The Composition tests were designed as operational field tests. This design captures student performance and score on field tested items. Four writing prompts in each of Grades 4, 7, and 10 were administered in spiral fashion within the classroom. Therefore, each student only needed to respond to one prompt. The prompts were developed to reflect Common Core State Standards and scored three times based on the following traits/rubrics: Writing Topic Development, Writing Language Conventions, and Understanding Literary Text or Understanding Informational Text (depending upon the type of passage associated with each prompt: literary or informational). The rubrics used to score these items can be found in Appendix B. Note that student reports reflected the scores from all rubrics; however, only the two Writing traits contributed to the overall scale score and proficiency level designations this year.

Test Administration Design

For Grades 4–8 and 10, both Reading and Mathematics items were included in the same test books. Reading items were in a stand-alone test book at Grade 9 since Mathematics was not tested at that grade level. For all grades, test books and answer booklets were color-coded. Students in Grades 2 and 3 used scannable test books in which they recorded their answers.

For Grades 3–10, each Reading and Mathematics test was divided into four sessions, for a total of eight sessions per grade level test. For Grade 2, each Reading and Mathematics test was divided into three sessions, for a total of six sessions. For all grades, 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 provided for each of the four prompts. The test books were scannable documents that included the following: directions to students, evaluation criteria, a writing prompt, three lined pages, and a biogrid. The prompts were administered within the established two-week testing window. Students were also issued two sheets of double-sided, lined draft paper, specially developed for the Composition test, for planning their writing.

Additional information regarding administration manuals and procedures is provided in Section 3.

Table 1. DC CAS 2012 Operational Test Form Blueprints: Reading

Grade	Content Strand		Operational						Anchor		Field Test
			Number of MC Items/ Points	% of MC Points	Number of CR Items	Number of CR Points	% of CR Points	Total Number of Points	Number of Items	% of Points	Number of Items
2	1	Vocabulary Acquisition & Use	7	100.00%	0	0	0.00%	7	--	--	4
	3	Reading Informational Text	13	81.25%	1	3	18.75%	16	--	--	19
	4	Reading Literary Text	13	81.25%	1	3	18.75%	16	--	--	19
		Total	33	84.62%	2	6	15.38%	39	--	--	42
3	1	Vocabulary Acquisition & Use	8	100.00%	0	0	0.00%	8	5	62.50%	4
	3	Reading Informational Text	18	85.71%	1	3	14.29%	21	9	42.86%	20
	4	Reading Literary Text	19	76.00%	2	6	24.00%	25	9	36.00%	14
		Total	45	83.33%	3	9	16.67%	54	23	42.59%	38
4	1	Vocabulary Acquisition & Use	8	100.00%	0	0	0.00%	8	4	50.00%	4
	3	Reading Informational Text	17	85.00%	1	3	15.00%	20	8	40.00%	20
	4	Reading Literary Text	20	76.92%	2	6	23.08%	26	11	42.31%	14
		Total	45	83.33%	3	9	16.67%	54	23	42.59%	38
5	1	Vocabulary Acquisition & Use	8	100.00%	0	0	0.00%	8	4	50.00%	4
	3	Reading Informational Text	16	72.73%	2	6	27.27%	22	8	36.36%	14
	4	Reading Literary Text	21	87.50%	1	3	12.50%	24	11	45.83%	20
		Total	45	83.33%	3	9	16.67%	54	23	42.59%	38
6	1	Vocabulary Acquisition & Use	9	100.00%	0	0	0.00%	9	4	44.44%	4
	3	Reading Informational Text	15	71.43%	2	6	28.57%	21	8	38.10%	14
	4	Reading Literary Text	21	87.50%	1	3	12.50%	24	10	41.67%	20
		Total	45	83.33%	3	9	16.67%	54	22	40.74%	38

Table 1. DC CAS 2012 Operational Test Form Blueprints: Reading *(continued)*

Grade	Content Strand		Operational						Anchor		Field Test
			Number of MC Items/ Points	% of MC Points	Number of CR Items	Number of CR Points	% of CR Points	Total Number of Points	Number of Items	% of Points	Number of Items
7	1	Vocabulary Acquisition & Use	8	100.00%	0	0	0.00%	8	5	62.50%	4
	3	Reading Informational Text	18	75.00%	2	6	25.00%	24	9	37.50%	14
	4	Reading Literary Text	19	86.36%	1	3	13.64%	22	9	40.91%	20
		Total	45	83.33%	3	9	16.67%	54	23	42.59%	38
8	1	Vocabulary Acquisition & Use	7	100.00%	0	0	0.00%	7	5	71.43%	4
	3	Reading Informational Text	20	76.92%	2	6	23.08%	26	10	38.46%	18
	4	Reading Literary Text	18	85.71%	1	3	14.29%	21	8	38.10%	18
		Total	45	83.33%	3	9	16.67%	54	23	42.59%	40
9	1	Vocabulary Acquisition & Use	8	100.00%	0	0	0.00%	8	4	50.00%	4
	3	Reading Informational Text	21	80.77%	2	5	19.23%	26	11	42.31%	18
	4	Reading Literary Text	16	84.21%	1	3	15.79%	19	8	42.11%	18
		Total	45	84.91%	3	8	15.09%	53	23	43.40%	40
10	1	Vocabulary Acquisition & Use	9	100.00%	0	0	0.00%	9	5	55.56%	4
	3	Reading Informational Text	18	75.00%	2	6	25.00%	24	8	33.33%	18
	4	Reading Literary Text	18	85.71%	1	3	14.29%	21	10	47.62%	18
		Total	45	83.33%	3	9	16.67%	54	23	42.59%	40

Table 2. DC CAS 2012 Operational Test Form Blueprints: Mathematics

Grade		Content Strand	Operational						Anchor		Field Test
			Number of MC Items/ Points	% of MC Points	Number of CR Items	Number of CR Points	% of CR Points	Total Number of Points	Number of Items	% of Points	Number of Items
2	1	Operations & Algebraic Thinking	8	80.00%	1	2	20.00%	10	--	--	7
	2	Numbers & Operations Base Ten	11	100.00%	0	0	0.00%	11	--	--	12
	3	Geometry	7	100.00%	0	0	0.00%	7	--	--	3
	4	Measurement and Data	12	85.71%	1	2	14.29%	14	--	--	14
		Total	38	90.48%	2	4	9.52%	42	--	--	29
3	1	Number Sense & Operations	16	84.21%	1	3	15.79%	19	9	47.37%	9
	2	Patterns, Relations & Algebra	9	100.00%	0	0	0.00%	9	4	44.44%	6
	3	Geometry	4	57.14%	1	3	42.86%	7	2	28.57%	7
	4	Measurement	12	100.00%	0	0	0.00%	12	4	33.33%	6
	5	Data Analysis, Statistics & Probability	10	76.92%	1	3	23.08%	13	6	46.15%	4
		Total	51	85.00%	3	9	15.00%	60	25	41.67%	32
4	1	Number Sense & Operations	23	100.00%	0	0	0.00%	23	11	47.83%	10
	2	Patterns, Relations & Algebra	7	70.00%	1	3	30.00%	10	6	60.00%	6
	3	Geometry	4	57.14%	1	3	42.86%	7	2	28.57%	4
	4	Measurement	7	100.00%	0	0	0.00%	7	2	28.57%	9
	5	Data Analysis, Statistics & Probability	10	76.92%	1	3	23.08%	13	4	30.77%	3
		Total	51	85.00%	3	9	15.00%	60	25	41.67%	32
5	1	Number Sense & Operations	20	100.00%	0	0	0.00%	20	10	50.00%	10
	2	Patterns, Relations & Algebra	10	76.92%	1	3	23.08%	13	6	46.15%	6
	3	Geometry	6	66.67%	1	3	33.33%	9	3	33.33%	4
	4	Measurement	9	100.00%	0	0	0.00%	9	2	22.22%	8
	5	Data Analysis, Statistics & Probability	6	66.67%	1	3	33.33%	9	3	33.33%	4
		Total	51	85.00%	3	9	15.00%	60	24	40.00%	32

Table 2. DC CAS 2012 Operational Test Form Blueprints: Mathematics *(continued)*

Grade	Content Strand		Operational						Anchor		Field Test
			Number of MC Items/ Points	% of MC Points	Number of CR Items	Number of CR Points	% of CR Points	Total Number of Points	Number of Items	% of Points	Number of Items
6	1	Number Sense & Operations	15	83.33%	1	3	16.67%	18	8	44.44%	10
	2	Patterns, Relations & Algebra	13	81.25%	1	3	18.75%	16	5	31.25%	7
	3	Geometry	8	100.00%	0	0	0.00%	8	5	62.50%	6
	4	Measurement	5	62.50%	1	3	37.50%	8	2	25.00%	5
	5	Data Analysis, Statistics & Probability	10	100.00%	0	0	0.00%	10	4	40.00%	4
		Total	51	85.00%	3	9	15.00%	60	24	40.00%	32
7	1	Number Sense & Operations	16	84.21%	1	3	15.79%	19	8	42.11%	10
	2	Patterns, Relations & Algebra	12	80.00%	1	3	20.00%	15	6	40.00%	7
	3	Geometry	8	100.00%	0	0	0.00%	8	2	25.00%	9
	4	Measurement	7	70.00%	1	3	30.00%	10	2	20.00%	3
	5	Data Analysis, Statistics & Probability	8	100.00%	0	0	0.00%	8	5	62.50%	3
		Total	51	85.00%	3	9	15.00%	60	23	38.33%	32
8	1	Number Sense & Operations	16	100.00%	0	0	0.00%	16	8	50.00%	6
	2	Patterns, Relations & Algebra	18	85.71%	1	3	14.29%	21	8	38.10%	12
	3	Geometry	5	62.50%	1	3	37.50%	8	4	50.00%	4
	4	Measurement	3	50.00%	1	3	50.00%	6	1	16.67%	5
	5	Data Analysis, Statistics & Probability	9	100.00%	0	0	0.00%	9	4	44.44%	5
		Total	51	85.00%	3	9	15.00%	60	25	41.67%	32
10	1	Number Sense & Operations	11	100.00%	0	0	0.00%	11	3	27.27%	5
	2	Patterns, Relations & Algebra	18	85.71%	1	3	14.29%	21	11	52.38%	10
	3	Geometry	6	66.67%	1	3	33.33%	9	2	22.22%	9
	4	Measurement	7	100.00%	0	0	0.00%	7	4	57.14%	4
	5	Data Analysis, Statistics & Probability	9	75.00%	1	3	25.00%	12	5	41.67%	3
		Total	51	85.00%	3	9	15.00%	60	25	41.67%	31

Table 3. DC CAS 2012 Operational Test Form Blueprints: Science/Biology

Grade		Content Strand	Operational						Anchor		Field Test
			Number of MC Items/ Points	% of MC Points	Number of CR Items	Number of CR Points	% of CR Points	Total Number of Points	Number of Items	% of Points	Number of Items
5	1	Science and Technology	14	87.50%	1	2	12.50%	16	--	--	7
	2	Earth and Space Science	12	85.71%	1	2	14.29%	14	--	--	12
	3	Physical Science	10	100.00%	0	0	0.00%	10	--	--	3
	4	Life Science	11	84.62%	1	2	15.38%	13	--	--	14
		Total	47	88.68%	3	6	11.32%	53	--	--	29
8	1	Scientific Thinking and Inquiry	6	75.00%	1	2	25.00%	8	9	47.37%	9
	2	Matter and Reactions	21	91.30%	1	2	8.70%	23	4	44.44%	6
	3	Forces	8	80.00%	1	2	20.00%	10	2	28.57%	7
	4	Energy and Waves	12	100.00%	0	0	0.00%	12	4	33.33%	6
		Total	47	88.68%	3	6	11.32%	53	6	46.15%	4
High School	1	Cell Biology & Biochemistry	13	86.67%	1	2	13.33%	15	25	41.67%	32
	2	Genetics and Evolution	15	100.00%	0	0	0.00%	15	11	47.83%	10
	3	Multicellular Organisms	10	83.33%	1	2	16.67%	12	6	60.00%	6
	4	Ecosystems	9	81.82%	1	2	18.18%	11	2	28.57%	4
		Total	47	88.68%	3	6	11.32%	53	2	28.57%	9

Table 4. DC CAS 2012 Operational Test Form Blueprints: Composition

Grade	Scoring Rubric	Number of CR Items	Number of CR Points	Contribution to Overall Scale Score	
				Number of Points	% of Points
4, 7, 10	Writing Topic Development	4	6	6	60.00%
	Writing Language Conventions	4	4	4	40.00%
	Understanding Literary Text*	2	4	--	--
	Understanding Informational Text*	2	4	--	--
	Total Possible Points	--	14	10	100.00%

* Understanding Literary or Informational Text Rubric was considered as a field test rubric and did not contribute to students' overall scores.

Section 3. Test Administration Guidelines and Requirements

This section contains information relevant to the *Standards and Assessments 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 the State Superintendent of Education (OSSE), 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 2012 *Test Chairperson's Manual*. Test Chairpersons, in turn, trained all Test Administrators and proctors. Test Administrators administered all DC CAS assessments according to requirements and steps in the *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 for 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 for English language learners. They specify approved accommodations that maintain standard testing conditions (e.g., reading only Mathematics, Science, or Health questions or Composition writing prompts to examinees) and identify accommodations that are considered modifications to the test that 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, Science, or Health test questions or Composition writing prompts), and response accommodations (e.g., dictating responses). The *Test Chairperson's Manual* and *Test Directions* specify accommodations approved for English language learners; they are in the following areas: direct linguistic support―oral, direct linguistic support—written, and indirect linguisticsupport. 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 and OSSE by completing the online Security Exceptions Survey.

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 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 September 20, 2012, approximately 99.68% of secure materials were accounted for; 212 secure test books were missing for the 2012 administration, compared with 103 test books missing in 2011.

Section 4. Student Participation

This section contains information relevant to S*tandards and Assessments 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 public schools in the District of Columbia administered the DC CAS tests between April 17 and April 27, 2012.

The tests administered were:

- Reading, Grades 2–10
- Mathematics, Grades 2–8 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

Participation in DC CAS

The DC CAS *Test Chairperson's Manual* states that all students enrolled in all public schools in the District of Columbia 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.

Approximately 4,500 students were assessed in Reading and Mathematics at each tested grade, with slightly fewer in each tested grade in Composition and Science/Biology. 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 2012 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.

The total number and percent of students with valid tests are provided at the total and subgroups of gender and race/ethnicity in Table 5. Participation rates for students in special populations, such as special education, Title 1, English Language Learners, and students with 504 plans is provided in Table 6. ELL students who participate in the DC CAS 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, the number and percent of which are provided in Table 7.

Special Accommodation

Students with disabilities and 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. Test administration accommodations are

categorized into 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. Within prescribed parameters, 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. The number and percent of the various accommodations documented are provided in Tables 6–10. For more information on these accommodations, please refer to the DC CAS *Test Chairperson's Manual*.

Table 5. Number and Percent of Examinees with Valid Test Administrations on the 2012 DC CAS in Reading, Mathematics, Science/Biology, or Composition

Grade	Students with Test Scores	Males		Females		Asian		African American		Hispanic		White	
		N	%	N	%	N	%	N	%	N	%	N	%
Reading													
2	4,491	2,274	50.63%	2,194	48.85%	95	2.12%	3,216	71.61%	626	13.94%	521	11.60%
3	4,754	2,402	50.53%	2,334	49.10%	94	1.98%	3,475	73.10%	665	13.99%	479	10.08%
4	4,589	2,317	50.49%	2,253	49.10%	102	2.22%	3,357	73.15%	632	13.77%	461	10.05%
5	4,744	2,402	50.63%	2,326	49.03%	78	1.64%	3,694	77.87%	578	12.18%	366	7.72%
6	4,545	2,297	50.54%	2,222	48.89%	68	1.50%	3,596	79.12%	566	12.45%	268	5.90%
7	4,301	2,160	50.22%	2,126	49.43%	55	1.28%	3,458	80.40%	508	11.81%	240	5.58%
8	4,359	2,172	49.83%	2,163	49.62%	55	1.26%	3,545	81.33%	476	10.92%	236	5.41%
9	4,164	2,031	48.78%	2,061	49.50%	77	1.85%	3,296	79.15%	489	11.74%	197	4.73%
10	4,272	2,039	47.73%	2,186	51.17%	64	1.50%	3,559	83.31%	445	10.42%	153	3.58%
Mathematics													
2	4,514	2,284	50.60%	2,205	48.85%	100	2.22%	3,224	71.42%	632	14.00%	523	11.59%
3	4,781	2,418	50.58%	2,344	49.03%	97	2.03%	3,486	72.91%	675	14.12%	482	10.08%
4	4,603	2,320	50.40%	2,264	49.19%	104	2.26%	3,360	73.00%	638	13.86%	464	10.08%
5	4,759	2,415	50.75%	2,328	48.92%	81	1.70%	3,692	77.58%	587	12.33%	368	7.73%
6	4,567	2,304	50.45%	2,236	48.96%	69	1.51%	3,608	79.00%	573	12.55%	269	5.89%
7	4,325	2,161	49.97%	2,148	49.66%	55	1.27%	3,463	80.07%	527	12.18%	240	5.55%
8	4,381	2,179	49.74%	2,178	49.71%	58	1.32%	3,541	80.83%	497	11.34%	236	5.39%
10	4,245	2,027	47.75%	2,173	51.19%	64	1.51%	3,533	83.23%	445	10.48%	152	3.58%
Science/Biology													
5	4,707	2,381	50.58%	2,299	48.84%	79	1.68%	3,641	77.35%	588	12.49%	366	7.78%
8	4,263	2,096	49.17%	2,122	49.78%	57	1.34%	3426	80.37%	493	11.56%	233	5.47%
High School	3,715	1,744	46.94%	1,890	50.87%	69	1.86%	2,968	79.89%	396	10.66%	197	5.30%
Composition													
4	4,470	2,236	50.02%	2,206	49.35%	103	2.30%	3,244	72.57%	622	13.91%	456	10.20%
7	4,146	2,049	49.42%	2,062	49.73%	55	1.33%	3,304	79.69%	498	12.01%	230	5.55%
10	3,511	1,638	46.65%	1,830	52.12%	58	1.65%	2,886	82.20%	388	11.05%	140	3.99%

Table 6. Number and Percent of Students in Special Programs with Test Scores on the 2012 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	
		N	%	N	%	N	%	N	%	N	%
Reading and/or Mathematics											
2	4,518	304	7%	432	10%	27	1%	309	7%	1	0%
3	4,783	458	10%	392	8%	33	1%	268	6%	1	0%
4	4,603	514	11%	259	6%	32	1%	265	6%	2	0%
5	4,763	669	14%	233	5%	35	1%	238	5%	0	0%
6	4,572	675	15%	244	5%	36	1%	51	1%	1	0%
7	4,332	662	15%	237	5%	40	1%	136	3%	2	0%
8	4,394	636	14%	250	6%	27	1%	141	3%	2	0%
9	4,164	568	14%	242	6%	10	0%	170	4%	0	0%
10	4,282	699	16%	182	4%	7	0%	444	10%	0	0%
Science/Biology											
5	4,707	570	12%	208	4%	33	1%	237	5%	0	0%
8	4,263	475	11%	241	6%	25	1%	134	3%	2	0%
High School	3,715	404	11%	140	4%	7	0%	149	4%	0	0%
Composition											
4	4,470	422	9%	221	5%	29	1%	248	6%	1	0%
7	4,146	512	12%	186	4%	31	1%	124	3%	2	0%
10	3,511	459	13%	151	4%	7	0%	189	5%	0	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 7. Number and Percent of Students Coded for ELL Access for Proficiency Levels 1–4 in Reading, Mathematics, Science/Biology, or Composition

Grade	Students with Test Scores	Level 1		Level 2		Level 3		Level 4	
		N	%	N	%	N	%	N	%
Reading and/or Mathematics									
2	4,518	48	1%	93	2%	189	4%	153	3%
3	4,783	28	1%	49	1%	178	4%	202	4%
4	4,603	20	0%	23	0%	74	2%	183	4%
5	4,763	23	0%	30	1%	65	1%	130	3%
6	4,572	33	1%	48	1%	83	2%	99	2%
7	4,332	42	1%	32	1%	82	2%	75	2%
8	4,394	43	1%	47	1%	78	2%	82	2%
9	4,164	62	1%	67	2%	57	1%	40	1%
10	4,282	5	0%	21	0%	78	2%	83	2%
Science/Biology									
5	4,707	19	0%	25	1%	59	1%	121	3%
8	4,263	39	1%	45	1%	77	2%	80	2%
High School	3,715	21	1%	50	1%	46	1%	38	1%
Composition									
4	4,470	10	0%	20	0%	60	1%	149	3%
7	4,146	16	0%	20	0%	74	2%	76	2%
10	3,511	6	0%	24	1%	65	2%	67	2%

Table 8. Number and Percent 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[1]		Direct Linguistic Support—Written		Indirect Linguistic Support		Other	
		N	%	N	%	N	%	N	%
Reading and/or Mathematics									
2	4,518	374	8%	200	4%	392	9%	11	0%
3	4,783	375	8%	192	4%	389	8%	3	0%
4	4,603	214	5%	120	3%	226	5%	2	0%
5	4,763	193	4%	144	3%	203	4%	1	0%
6	4,572	195	4%	135	3%	199	4%	1	0%
7	4,332	149	3%	127	3%	171	4%	3	0%
8	4,394	195	4%	146	3%	210	5%	1	0%
9	4,164	208	5%	94	2%	204	5%	1	0%
10	4,282	166	4%	141	3%	171	4%	10	0%
Science/Biology									
5	4,707	181	4%	144	3%	187	4%	0	0%
8	4,263	189	4%	144	3%	204	5%	0	0%
High School	3,715	127	3%	130	3%	130	3%	10	0%
Composition									
4	4,470	199	4%	105	2%	202	5%	1	0%
7	4,146	109	3%	35	1%	134	3%	0	0%
10	3,511	146	4%	50	1%	145	4%	7	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. Students for whom the ELL bubble was not completed but who did receive these ELL test administration accommodations are counted here.

[1] The "Oral Reading of Test in English" accommodation is typically not permitted for the Reading test.

Table 9. Number and Percent 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[1]		Response		Other	
		N	%	N	%	N	%	N	%	N	%
Reading/Mathematics											
2	4,518	352	8%	371	8%	352	8%	191	4%	21	0%
3	4,783	506	11%	487	10%	476	10%	332	7%	10	0%
4	4,603	594	13%	598	13%	561	12%	393	9%	25	1%
5	4,763	718	15%	714	15%	669	14%	434	9%	21	0%
6	4,572	714	16%	717	16%	677	15%	508	11%	9	0%
7	4,332	701	16%	699	16%	658	15%	556	13%	13	0%
8	4,394	669	15%	673	15%	648	15%	591	13%	6	0%
9	4,164	539	13%	543	13%	420	10%	276	7%	11	0%
10	4,282	646	15%	660	15%	503	12%	603	14%	15	0%
Science/Biology											
5	4,707	648	14%	649	14%	608	13%	354	8%	23	0%
8	4,263	613	14%	612	14%	577	14%	468	11%	5	0%
High School	3,715	388	10%	414	11%	297	8%	220	6%	5	0%
Composition											
4	4,470	484	11%	497	11%	456	10%	256	6%	24	1%
7	4,146	557	13%	562	14%	529	13%	378	9%	13	0%
10	3,511	389	11%	430	12%	300	9%	243	7%	6	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. Students for whom the Special Education bubble was not completed and who did receive these Special Education test administration accommodations are counted here.

[1] The "Presentation" column contains 10 accommodations, two of which are not typically permitted for Reading assessments: "Reading Test Questions" and "Translation of Words or Phrases" is available for Mathematics, Science/Biology, and Composition only.

Table 10. Number and Percent 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)[1]		Responses Dictated	
		N	%	N	%	N	%	N	%
Reading and/or Mathematics									
2	4,518	303	7%	361	8%	267	6%	70	2%
3	4,783	447	9%	478	10%	374	8%	103	2%
4	4,603	508	11%	584	13%	460	10%	130	3%
5	4,763	623	13%	702	15%	560	12%	104	2%
6	4,572	618	14%	704	15%	535	12%	70	2%
7	4,332	584	13%	690	16%	502	12%	77	2%
8	4,394	563	13%	663	15%	486	11%	47	1%
9	4,164	449	11%	518	12%	111	3%	36	1%
10	4,282	539	13%	637	15%	211	5%	46	1%
Science/Biology									
5	4,707	569	12%	637	14%	500	11%	88	2%
8	4,263	503	12%	605	14%	439	10%	52	1%
High School	3,715	310	8%	392	11%	152	4%	29	1%
Composition									
4	4,470	411	9%	484	11%	368	8%	97	2%
7	4,146	442	11%	553	13%	416	10%	58	1%
10	3,511	324	9%	417	12%	149	4%	31	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.

[1] The "Reading Test Questions" and "Translation of Words or Phrases" accommodations are typically not permitted for the Reading test.

Section 5. Scoring

This section contains information relevant to S*tandards and Assessments 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?

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 2012 constructed response items were written by CTB's Development teams in conjunction with OSSE and, for Reading Grade 9 DCPS. Composition's Understanding

Literary Text and Understanding Informational Text rubric was added this year to the scoring guides, and also underwent a rangefinding process in DC to identify anchor papers, which represent the exemplars at each score point.

Prior to actual scoring, CTB supervisors studied and internalized the scoring guides along with existing materials that were then used in training raters to hand-score the constructed response items for all content areas. This ensured consistency in scoring the same items across administrations (such as field test to operational), with the same anchor papers and training philosophy.

Preparation and Meeting Logistics for Rangefinding

Rangefinding is the process of reviewing student responses to newly tested (field tested) items to identify anchor or exemplar papers at each score point. The anchor papers are concrete examples of particular score points and are delineated in the scoring guides used during training and scoring. All DC CAS constructed response items go through this process prior to operational scoring. For example, for the newly field tested Composition prompts (four in each of Grades 4, 7, and 10), an extensive rangefinding workshop was held in DC, from June 18, 2012, through June 22, 2012, with discussion groups of three or four DC teachers per grade. These groups of teachers chose the anchor papers to be used during subsequent rubric training.

In preparation for rangefinding, CTB content supervisors reviewed hundreds of student responses to identify a variety of papers for the reviews. These potential anchors were then assembled for review at rangefinding. During 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

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), each of which consisted of a single item to be scored. Raters and team leaders were trained in the following steps:

- Reviewing the student answer booklet
- Reviewing rubrics
- Reviewing anchor papers and training papers and answering questions arising from the established scores

- Explaining scoring strategies, followed by a question-and-answer period
- Administering Qualifying Round 1
- Reviewing Qualifying Round 1 established scores, and answering questions arising from the scores
- 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 previously. The CTB content experts who supervised the training of the team leaders also supervised the training of the raters.

Breakdown of Scoring Teams

Groups of CTB content experts oversaw the training and scoring of the constructed response items for 2012 in Reading, Mathematics, Science/Biology, and Composition. Each of the content experts was responsible for training and scoring all of the items in his or her content area. Teams of between raters (numbers of which depend 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.

Training and scoring of the operational constructed response items occurred May 8–18, 2012, for Reading, Mathematics, and Science/Biology, and July 9–20, 2012, for Composition. Training and scoring of the field test constructed response items occurred July 11–18, 2012, Training consisted of a review of the rubrics, followed by analysis of the anchor papers for each item. Raters then participated in qualifying rounds, which consisted of ten books of sample papers for the item in a given RIB. Raters were given two opportunities to achieve acceptable qualification ratings; those not meeting the minimum were dismissed.

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 remained consistent in scoring accuracy.

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, where the checksets were "blind" or unknown to the rater. Raters whose checkset accuracy repeatedly dipped below the quality standards were flagged and retrained. 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. 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. If the rater's accuracy on

checkset scores did not meet the quality standards after this retraining, the rater was 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, results of which are provided in Section 8. This procedure is called a "double-blind read" because the second rater does not know the first rater's score.

Scoring Security

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 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.

Section 6. Methods

This section contains information relevant to *Standards and Assessments Peer Review Guidance*, Critical Elements 4.4, 4.5, 4.6, and 5.6.

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?

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?

5.6
Assessment results must be expressed in terms of the achievement standards, not just scale scores or percentiles.

This section describes the methods used to analyze the item and test level data for the DC CAS. Results of the item and test level analyses described here are provided as evidence for reliability and validity in Section 8.

Classical Item Level Analyses

Each operational test item was first reviewed in terms of classical raw score statistics. Each item's frequency distribution (number of students responding for each answer choice or score level) as well as each item's overall *p* value (proportion of students choosing the correct answer) and point biserial or item-test correlation (how correlated each individual item is with the test as a whole based on the correct response) were reviewed. Typically, *p* values should range between 0.30 and 0.90. Items with a *p* value less than 0.30 are considered more difficult since less than 30% of the students are getting the correct answer. Values greater than 0.90 indicate a fairly easy item, with more than 90% of students getting the correct answer. With newly tested content, the *p* value may dip lower than 0.30, at which point the item should be evaluated in light of the newness of content or students' opportunity to learn the content. Point biserials or item-test correlations are usually in the range of 0.30 and above, although some items can be acceptable when as low as 0.15. The point biserials of each item's distractors, or incorrect responses, were also analyzed, as well as any distractor with a positive point biserial, either of which was reviewed for the possibility of an additional correct response or no correct response.

It is also important to track the rate at which students do not respond to, or omit, items. Omitted items receive a zero score. The rate of omission often provides some information about test times, or speededness, particularly if there is a high rate of items omitted at the end of a test session. It also provides an indication of items that may simply be unclear or illogically presented. When more than 5% of students omit an item, the item is reviewed by both CTB Research and Development and shared with OSSE.

Item Bias Analyses

Differential item functioning (DIF) statistics provide a measure of the systematic errors by

subgroups that are specifically attributed to some bias or systematic over- or under-representation of subgroup performance when compared with total group performance. To evaluate the potential bias, items are first reviewed from content perspectives. 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 the DC CAS items prior to field testing and selection for operational test forms

For the DC CAS program, CTB uses Mantel-Haenszel statistics (Mantel & Haenszel, 1959) to evaluate DIF for both operational and field test items. The subgroups compared in the DIF analyses for the 2012 administration reflect conventional subgroupings, and were based on gender (male—reference and female—focal) and race/ethnicity (African American—reference, and Asian, Hispanic, and White—focal). As with all statistical tests, Mantel-Haenszel DIF statistics are subject to Type I and II errors. An item flagged for DIF may or may not provide an unfair advantage or disadvantage for one examinee subgroup compared with another. However, the flag does show when an item is more difficult for a particular focal subgroup of students than would be expected based on their total test scores, when compared with the difficulty of the item for the comparison or reference subgroup with equivalent total test scores. OSSE and CTB screen all items that are flagged for DIF after each administration to identify items that may favor or disadvantage examinee subgroups.

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

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

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

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.

Calibration and Equating

Scaling and linking was accomplished using the PARDUX and SAS 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:

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

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

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

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

Goodness-of-fit statistics were computed for each item to examine how closely the item's data conform to the item response models. This provides a measure of validity. 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:

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

The *Z*-statistic is an index of the degree to which obtained proportions of students with each item score are close to the proportions that would be predicted by the estimated thetas and item parameters. These values are computed for ten intervals corresponding to deciles of the theta distribution (Burket, 1995). The *Z*-statistic is used to characterize item fit. The critical value of *Z* is different for each grade because it is dependent on sample size.

Evidence of the validity of the scalings is provided by model fit. If the IRT model fits the empirical item response distributions for the population we want to generalize to (i.e., District of Columbia students), then the claim that the scores are valid indicators of an underlying proficiency is strengthened. Fit statistics 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 11 indicates that only small numbers of operational items were flagged for poor fit to the IRT model. No items were removed from operational scaling and scoring due to poor fit.

Year-to-Year Equating Procedures

Once the IRT scaling is accomplished, equating the scale across years enables comparability of scores from one year to the next and across all test forms in the same content area and grade. In 2007 through 2012, anchor item sets that equate the current test forms to the previous year's scale were used in a Stocking and Lord (1983) equating methodology.

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 items consist of multiple choice and/or constructed response items. The Reading and Science/Biology equating anchor items for 2012 sets included multiple choice items and one constructed response item; in Mathematics, all of the anchor items were multiple choice items. Anchor items are rotated in and out of use each year, to the degree possible, to minimize item over-exposure. Anchor items are placed in approximately the same location or same third of the location as the original administration. Anchor item *a* and *b* parameters are calibrated freely (i.e., not fixed during calibration). The number and representativeness of the anchor items relative to the overall test and blueprints is provided in Tables 1–4. The blueprint should be proportionally represented in the anchor sets.

Because Composition prompts are so few, the "items" or scores from each of the Writing rubrics were linked to the Reading scale by first matching students' Reading and Composition item-level scored responses. The Reading operational items were treated as anchor items and the Stocking and Lord common-item equating procedure was conducted.

Once calibrated, the anchor item set and equating results are carefully reviewed to ensure that it is performing very similarly in both current and reference (just prior) year. These standard CTB Research team quality checks are followed during calibration and equating analyses for all grades and content areas. Additional anchor item checks were conducted for items flagged in any

of the following verifications, which 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). Specifically, differential anchor item performance between the 2011 and 2012 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 correlations are shown in Table 12 for the discrimination (*a*) and difficulty (*b*) parameters and are moderate to high, ranging from 0.85 to 0.97 for *a* parameters (0.84–0.98 in 2011) and from 0.96 to .099 for *b* parameters (0.94–1.00 in 2011). These correlations indicate that the items performed similarly in the two administrations and provide evidence that the equating results are reasonable and accurate.

2. Reference and equated anchor item parameters and TCCs should be closely aligned. The TCCs are reviewed after each equating cycle for each grade and content area. Further, statistical differences were evaluated with four difference statistics: root mean squared difference, mean absolute difference, maximum absolute difference, and the absolute value of the mean signed difference.

3. The scaling constants, or Stocking-Lord linear transformation parameters, should be fairly stable across administrations. 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 2011–2012 differences in these scaling constants indicates the degree of differences in average difficulty of the reference and new test form administrations. The scaling constants from the 2012 administration along with constants across the 2007–current years of the DC CAS administration and scales are provided in Table 13.

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, show the same direction and magnitude of change as do the scale scores. The correlations of the anchor item *p* values in Table 12 are highly correlated, ranging from 0.96 to 0.99 for all grades and content areas. This is an indication that the anchor items performed similarly in the examinee populations in 2011 and 2012.

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.

Establishing Upper and Lower Bounds for the Grade Level Scales

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, the 2008 base scale for Science/Biology, the 2011 base scale for Reading Grade 9, and the 2012 Reading scale for Composition. These values remain constant from year to year. The LOSS and HOSS for all grades are provided in Table 14.

Reliability Coefficients

Total test reliability statistics (alpha and CSEMs) measure the level of consistency (reliability) of performance over all test questions in a given form, the results of which imply how well the questions measure the content domain and could continue to do so over repeated administrations. Total test reliability coefficients (in this case measured by Cronbach's alpha [α; 1951]) may range from 0.00 to 1.00, where 1.00 refers to a perfectly reliable test. The DC CAS reliability data are based on DC students in the calibration sample of approximately 4,500 students per grade/content.

The total test reliabilities of the operational forms were evaluated first by Cronbach's α index of internal consistency. The specific calculation for Cronbach's α is calculated as

$$\hat{\alpha} = \frac{k}{k-1}\left(1 - \frac{\sum \hat{\sigma}_i^2}{\hat{\sigma}_X^2}\right), \tag{8.1}$$

where k is the number of items on the test form, $\hat{\sigma}_i^2$ is the variance of item i, and $\hat{\sigma}_X^2$ is the total test variance.

The stratified coefficient alpha is an 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.

Standard Errors of Measurement

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 the 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.

Proficiency Level Analyses

One of the cornerstones of the NCLB Act (US DOE, 2002) is the measurement of Adequate Yearly Progress (AYP) for states with respect to the percentage of students at or above the academic performance standards established by states. Because of a heavy emphasis on moving all students to or above the "Proficient" category by year 2014, the consistency and accuracy of the classification of students into these proficiency categories is of particular interest.
The statistical quality of cut scores that define the proficiency levels in which students are placed per their performance serves as additional validity evidence. Details about the standard setting workshops and Bookmark Standard Setting Procedure used to set the cut scores are given in the DC CAS Cut Score Setting Technical Report (CTB/McGraw-Hill, 2012). It may be useful to note that the Bookmark procedure (Mitzel, Lewis, Patz, & Green, 2001) is a well-documented and highly regarded procedure that has been demonstrated by independent research to produce reasonable cut scores on tests across the country.

It is also important to review the specific scale score SEM for each cut score. Comparison of these SEMs to the SEMs associated with other DC CAS scale scores for each test should almost always be among the lowest, meaning that the DC CAS tests tend to measure most accurately near the cut score. This is a desirable quality when cut scores are used to classify examinees.

Classification Consistency

Not only is it important that the amount of measurement error around the cut score be minimal; also important is the expected consistency with which students would be classified into performance levels if given the test over repeated occasions. 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, the table below 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.

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	$P_{1\cdot}$	$P_{2\cdot}$	$P_{3\cdot}$	*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:

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

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 2012 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 the contingency table 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).

Table 11. DC CAS 2012 Numbers of Operational Items Flagged for Poor Fit During Calibration

Content	Grade	Flagged for Poor Fit
Reading	2	3
	3	2
	4	0
	5	2
	6	0
	7	3
	8	0
	9	0
	10	1
Mathematics	2	2
	3	2
	4	2
	5	1
	6	0
	7	0
	8	0
	10	3
Science/Biology	5	0
	8	3
	High School	1
Composition	4	0
	7	3
	10	1

Table 12. Correlations Between the Item Parameters for the Reference Form and 2012 DC CAS Operational Test Form

Grade	Discrimination (a)	Difficulty (b)	*P* Value Correlation
Reading			
2	N/A	N/A	N/A
3	0.95	0.99	0.99
4	0.97	0.99	0.99
5	0.97	0.99	0.99
6	0.96	0.98	0.98
7	0.95	0.97	0.98
8	0.97	0.99	0.99
9	0.85	0.98	0.98
10	0.95	0.97	0.97
Mathematics			
2	N/A	N/A	N/A
3	0.96	0.98	0.98
4	0.93	0.99	0.99
5	0.97	0.97	0.98
6	0.93	0.98	0.99
7	0.90	0.98	0.98
8	0.94	0.98	0.98
10	0.88	0.97	0.97
Science/Biology			
5	0.92	0.99	0.99
8	0.89	0.97	0.96
High School	0.96	0.96	0.98

Table 13. Scaling Constants Across Administrations, All Grades and Content Areas

Grade	2007		2008		2009		2010		2011		2012	
	Mult	Add	Mult	Add	Mult	Add	Mult	Add	Mult	Add	Mult	Add
Reading												
2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
3	10.40	352.60	10.70	354.00	10.70	353.10	14.30	349.60	13.60	350.40	13.09	349.66
4	11.80	451.20	11.70	453.30	12.40	453.40	13.40	451.60	13.50	451.10	12.22	453.49
5	11.40	552.20	11.30	554.90	11.40	553.70	12.40	553.20	12.20	554.20	12.23	555.64
6	10.80	652.10	10.40	652.90	10.40	653.00	11.40	651.60	11.20	652.70	11.39	652.04
7	10.40	751.30	10.40	752.70	10.20	754.70	11.50	754.30	11.60	754.30	11.55	755.52
8	11.10	851.80	10.40	853.80	11.10	853.50	12.30	854.60	12.00	856.90	11.75	855.29
9	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	13.50	950.00	13.85	948.32
10	11.30	954.50	10.90	953.40	10.70	954.10	12.10	952.10	13.00	955.60	12.60	954.87
Mathematics												
2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
3	14.50	353.90	16.20	354.00	17.30	357.00	16.70	352.40	17.30	353.80	15.87	353.15
4	14.10	452.10	13.20	456.40	14.10	457.90	13.80	455.10	14.10	455.10	13.68	457.78
5	14.10	552.20	14.80	555.90	15.10	556.40	14.20	556.80	14.70	557.00	14.62	558.84
6	13.40	647.30	14.50	649.80	14.30	650.10	14.30	650.30	14.20	652.40	14.36	652.35
7	13.70	746.90	13.40	750.00	14.60	751.00	15.10	751.20	14.50	753.60	15.02	754.39
8	13.00	844.80	12.50	847.40	12.90	848.50	13.50	848.60	13.00	851.60	13.08	851.59
10	15.50	945.20	16.90	945.40	16.70	947.30	16.50	944.30	16.20	948.00	16.25	948.32
Science/Biology												
5	N/A	N/A	8.00	550.00	8.70	549.90	9.10	549.20	9.20	549.40	8.91	550.22
8	N/A	N/A	8.00	850.00	8.90	851.00	9.40	851.90	9.40	852.90	9.52	852.86
High School	N/A	N/A	8.00	950.00	7.70	946.60	7.50	949.50	7.80	949.90	8.36	950.88
Composition												
4	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	12.78	452.38
7	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	11.34	754.68
10	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	13.11	952.72

Table 14. LOSS and HOSS for Relevant Grades in Reading, Mathematics, Science/Biology and Composition

Grade	LOSS	HOSS
2	200	300
3	300	399
4	400	499
5	500	599
6	600	699
7	700	799
8	800	899
9	900	999
10	900	999

Section 7. Standard Setting

This section contains information relevant to the *Standards and Assessments 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?

The DC CAS cut scores associated with each of the four proficiency levels (*Below Basic, Basic, Proficient, Advanced*) for each grade and content area were all set through a content, statistical, and policy-based process. The content and related statistics were reviewed through standard setting workshops conducted with DC teachers in Washington, DC, and the resulting cut score recommendations were provided to the DC Technical Advisory Council and OSSE approvals. 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. DC staff reviewed, refined, and approved the descriptions prior to each workshop.

In 2012, standard setting workshops were conducted to review and recommend cut scores for Reading Grades 2–10, Mathematics Grade 2, and Composition Grades 4, 7, and 10. Previous standard settings were conducted to determine cut scores in 2006 for Mathematics Grades 3–8 and 10, in 2008 for Science Grades 5 and 8 and Biology High School, and in 2011 for Reading Grade 9.

The Bookmark Standard Setting Procedure (BSSP; Lewis, Mitzel, & Green, 1996; Lewis, Mitzel, Mercado, & Schultz, 2012) was implemented to establish performance standards for the assessments of Reading, Mathematics, and Science/Biology. Cut scores for Grade 2 Reading and Mathematics were established in July 2012. Cut scores for Grades 3–10 Reading were reviewed in July 2012, extending work from the original standard setting committee in July 2006 and July

2011. The Grades 3–8 and 10 Mathematics and Science/Biology cut scores were established in July 2008. The Judgmental Policy Capturing procedure (Jaeger, 1995) was used to set standards for the Composition assessments in July 2012. District of Columbia educators who participated in standard setting workshops recommended cut scores for each test and grade level.

The July 2006 standard setting workshops for 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 Judgmental Policy Capturing method in 2012 was implemented to set standards for the Composition test in Grades 4, 7, and 10. Judgmental Policy Capturing is a rubric-centered, content-based method that has been used in recent years to establish performance standards on unscaled assessments, (see Jaeger, 1995; Perie, 2007; Roeber, 2002). During the one-and-a-half-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. A third rubric for Composition, Literary Analysis (0–4 points), was also used to score students' responses; however, scores from this rubric did not contribute to students' total scores in 2012.) 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 for each writing prompt.

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 2012 and the OSSE in 2008 and 2012. Certain cut scores were adjusted each time to achieve articulated standards and impact data. The DC Board of Education approved these cut scores in 2006 and 2008, and the OSSE approved the cut scores for Composition, Grade 2 Reading and Mathematics, and Grades 3–10 Reading in 2012.

Grades 3–10 Reading Cut Score Review

In recognition that there may have been subtle shifts in the DC CAS Reading tests and expectations of student performance since the original standard setting in 2006—except for Grade 9, where cut scores were set in 2011—OSSE decided to conduct a review of the cut scores for Grades 3–10 Reading in 2012.

The Bookmark Standard Setting Procedure (BSSP; Lewis, Mitzel, & Green, 1996; Lewis, Miztel, Mercado, & Schulz, 2012) was implemented to review the cut scores for Grades 3–10 Reading. The workshop lasted one-and-a-half days. Participants reviewed cut scores at the Basic, Proficient, and Advanced levels, which separate students into four performance levels: Below Basic, Basic, Proficient, and Advanced.

Participants engaged in training, discussion, and two rounds of bookmark placements. Participants reviewed the existing cut scores and recommended adjustments for content-based reasons. Participants also adjusted the performance level descriptions (PLDs) to improve their clarity and alignment with the tested content.

NOTE: The Reading cut scores reviewed and approved during the 2012 standard setting for Grades 3–10 were NOT used for scoring, reporting, or accountability purposes in 2012. These will be applied in 2013.

Grade 2 Reading and Mathematics Standard Setting

The foundation of the standard setting for the Grade 2 Reading and Mathematics assessments were based on the guidance and procedures used for DC CAS in Reading and Mathematics for Grades 3–8 and 10. Prior to setting performance standards for the Grade 2 assessments, CTB Test Development staff drafted performance level descriptors, which summarize the knowledge, skills, and abilities expected of students in each performance level on the tests.

The Bookmark Standard Setting Procedure was implemented to set standards for the Grade 2 Reading and Mathematics assessments. The standard setting workshop lasted one-and-a-half days. Participants recommended 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 two rounds of bookmark placements. To help participants recommend cut scores that were well articulated with Grades 3–10 Reading and Grades 3–8 and 10 Mathematics, participants were shown *target cut scores* that were calculated statistically from the Grades 3–10 cut scores. Participants were free to recommend any set of cut scores that were consistent with the tested content and the expectations of students in each performance level.

Grades 4, 7, and 10 Composition Standard Setting

The pool of writing prompts for the Composition assessment was refreshed in 2012 with new prompts. Simultaneously, a test scale (based on DC CAS Reading) was implemented on the Composition test for the first time. In recognition of these changes in the Composition assessment, the OSSE decided to hold a standard setting for Grades 4, 7, and 10 Composition.

The Judgmental Policy Capturing procedure (Jaeger, 1995) was implemented to set standards for the Composition assessments. The standard setting workshop lasted one-and-a-half days. Participants recommended cut scores at the Basic, Proficient, and Advanced levels, which separate students into four performance levels: Below Basic, Basic, Proficient, and Advanced.

Participants engaged in training and in three rounds of discussion and decisions. For each writing prompt, participants recommended cut scores in terms of raw score. These raw scores were later transformed onto the test scale. Participants considered students' performance on the two scored rubrics, Topic Development and Standard English Conventions, and on one unscored rubric, Literary Analysis. Only scores from the first two rubrics contribute to students' scores in 2012.

Final, Approved DC CAS Cut Scores

The cut score recommendations from the 2012 committees were reviewed by staff from the OSSE. In addition, the OSSE reviewed the impact data associated with the recommended cut scores: impact data are the percentages of students classified in each performance level, based on the cut

scores. The cut scores, as recommended by District of Columbia educators, were reviewed by the Technical Advisory Committee and approved by the OSSE in August 2012. The standard setting technical report summarizes procedures and results of the 2012 standard settings and cut score review.

The report includes round-by-round synopses, agendas, training materials, and the recommended cut scores. See District of Columbia Comprehensive Assessment System (DC CAS) Standard Setting Technical Report 2012 (CTB/McGraw-Hill, 2012).

Table 15 shows the final, approved cut scores. Note that the 2012 Reading cut scores for all grades except Grade 2 were applied to final scores reported in 2012. The new Reading cut scores will be applied to all grades in 2013.

Table 15. Final Cut Score Ranges

Reading: Applied in 2012				
Grade	**Below Basic**	**Basic**	**Proficient**	**Advanced**
2	200 – 231	232 – 245	246 – 263	264 – 299
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
9	900 – 930	931 – 949	950 – 959	960 – 999
10	900 – 939	940 – 955	956 – 969	970 – 999
Reading: To Be Applied in 2013				
Grade	**Below Basic**	**Basic**	**Proficient**	**Advanced**
2	200 – 231	232 – 245	246 – 263	264 – 299
3	300 – 339	340 – 351	352 – 366	367 – 399
4	400 – 443	444 – 455	456 – 469	470 – 499
5	500 – 544	545 – 554	555 – 568	569 – 599
6	600 – 639	640 – 651	652 – 665	666 – 699
7	700 – 743	744 – 755	756 – 766	767 – 799
8	800 – 841	842 – 855	856 – 867	868 – 899
9	900 – 938	939 – 950	951 – 964	965 – 999
10	900 – 942	943 – 954	955 – 966	967 – 999
Mathematics				
Grade	**Below Basic**	**Basic**	**Proficient**	**Advanced**
2	200 – 243	244 – 254	255 – 267	268 – 299
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
Science/Biology				
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
Composition				
Grade	**Below Basic**	**Basic**	**Proficient**	**Advanced**
4	400 – 443	444 – 455	456 – 469	470 – 499
7	700 – 743	744 – 755	756 – 766	767 – 799
10	900 – 942	943 – 954	955 – 966	967 – 999

Section 8. Evidence for Reliability and Validity

This section contains information relevant to *Standards and Assessments Peer Review Guidance*, Critical Elements 4.1, 4.2, and 4.3:

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?
(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)?
(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)?

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? *and*
(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? *and*
(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?

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?

Reliability

Reliability refers to the degree to which students' scores are free from such effects and provides a measure of consistency. In other words, reliability helps to describe how consistent students' performances would be if given the assessment over multiple occasions. 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.

Validity

The collection of reliability evidences is a necessary precursor to establishing evidence of validity. How the scores are ultimately used is a key component to validity evidence, such that the trustworthiness of the scores is well established. As noted in the introduction, test validation is an ongoing process of gathering evidence from many sources to evaluate the trustworthiness of the desired score interpretation or use. This evidence is provided throughout this technical report specific to procedures and processes that support the integrity of the content of the test, test development, blueprints, alignment, scoring and rater reliability, psychometric analyses (item analyses, scaling, equating, and comparative analyses across administrations), and student-level performance results.

Item Level Evidence

Classical Item Statistics

DC CAS operational and field test items are all reviewed for statistical accuracy and quality. Table 16 summarizes item level classical statistics for operational and field test items. 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 operational items on a test ranged from moderately difficult (mean *p* value of 0.41 for Science Grade 8 and Biology) to moderately easy (mean *p* value of 0.74 for Grade 2 Mathematics). Tables in Appendix C display the item difficulty for each item at each grade. With respect to field test items, a test ranged from moderately difficult (mean *p* value of 0.35 for Science Grade 8 and Biology) to moderately easy (mean *p* value of 0.69 for Grade 2 Mathematics).

The point biserial, or item-test correlation, a type of internal consistency measure, is one measure of the correlation between each item and the overall test. The item-test correlations for each content area and grade for operational and field test items are shown in Table 16. The operational test form correlations range from 0.38 to 0.45 (Reading); from 0.38 to 0.45 (Mathematics); from 0.29 to 0.35 (Science/Biology); and from 0.60 to 0.68 (Composition). Field test form correlations range from 0.27 to 0.37 (Reading); from 0.33 to 0.45 (Mathematics); and from 0.25 to 0.34 (Science/Biology).

Table 16 also displays the mean item omit rates calculated across students for each grade and content area. 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 to ensure the administration conditions, such as testing time and accurate printing and scanning. Overall, the omit rates are low. The largest mean percentage omit rate is 5.80% in Composition Grade 10. All of the not reached rates are less than 1% except for in Reading Grade 9 (1.92%), Reading Grade 10 (1.15%), Composition Grade 4 (1.18%), Composition Grade 7 (1.75%), and Composition Grade 10 (5.80%), indicating that the students were provided with ample time to complete the DC CAS tests.

Inter-Rater Reliability

The inter-rater reliabilities of constructed response items rely heavily on the solid and consistent training of the hand-scorers, as was described in Section 4. 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 differing by only one point) 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%.

Statistical agreement data are presented in terms of the percentage of perfect, adjacent, and discrepant agreement. Adjacent agreement occurs when two raters differ by one point, and discrepant agreement is when two raters differ by more than one point. Tables 17–20 provide the inter-rater agreement statistics for operational constructed response items. In general, the values are within acceptable limits. For operational items, in Reading, the average perfect agreement was 72%, with a high of 86% and a low of 56%. For perfect and adjacent agreement, the average was 96%, with a high of 99% and a low of 91%. In Mathematics, the average perfect agreement was 90%, with a high of 97% and a low of 79%. For perfect and adjacent agreement, the average was 99%, with a high of 100% and a low of 98%. In Science/Biology, the average perfect agreement rate was 87%, with a high of 94% and a low of 80%. For perfect and adjacent agreement, the average was 99%, with a high of 100% and a low of 97%. In Composition, the average perfect agreement was 59%, with a high of 87% and a low of 42%. For perfect and adjacent agreement, the average was 94%, with a high of 100% and a low of 83%.

Field test items with perfect plus adjacent inter-rater agreement rates below 90% or lower checkset agreement rates will be avoided as much as possible during the process of selecting items for operational use. These items and their rubric can be investigated to determine whether the rubric may be difficult to apply in live scoring, and such items can be revised and re-field tested. Tables 21–23 provide the agreement rates for field test constructed response items. For field test items, in Reading, the average perfect agreement was 68%, with a high of 84% and a low of 57%. For perfect and adjacent agreement, the average was 96%, with a high of 99% and a low of 90%. In Mathematics, the average perfect agreement was 89%, with a high of 97% and a low of 77%. For

perfect and adjacent agreement, the average was 99%, with a high of 100% and a low of 95%. In Science/Biology, the average perfect agreement rate was 87%, with a high of 97% and a low of 73%. For perfect and adjacent agreement, the average was 99%, with a high of 100% and a low of 97%.

Differential Item Function

Differential item function (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 24–27 summarize the 2012 DIF analysis results for operational items, and Tables 28–30 for pilot items. 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. Recall that A corresponds to no DIF, B to moderate DIF, and C to considerable DIF. Modest numbers of multiple choice and constructed response items were flagged for DIF at levels B and C. 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 operational items flagged for DIF was moderate. For example, the total of 126 Reading item flags for DIF represents 13.2% of the 957 flagging opportunities 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.

The number of field test items flagged for DIF was moderate. For example, the total of 126 Reading item flags for DIF represents 13.2% of the 957 flagging opportunities 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 represents 5.1% of the 356 flagging opportunities.

Test and Strand Level Evidence

Operational Test Scores

Operational test level raw score and scale score means and standard deviations for the District are provided in Table 31, along with the test level reliability coefficients, including Cronbach alpha, stratified coefficient alpha, and Feldt-Raju. The scale score and raw score means and standard deviations are consistent across grades within content area. The reliabilities all show high levels of internal consistency, with reliabilities all greater than 0.85. Subgroup performance and total test reliabilities are provided in Appendix D.

Similarly, the content strand means, standard deviations, average*p* values, and reliabilities are provided for each grade and content area in Tables 32–35. 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).

Strand Level Scores

The raw score means and standard deviations highlight strands in which students show better or lesser mean performance, and the variability of that performance given the spread represented by the standard deviations. The average *p* values are a better indicator of the strand level difficulty, however, given they are not swayed by the number of items in a given strand, as the mean raw score is. Therefore, a review of the average *p* values in each strand highlights the strands that tend to be the more or less difficult for students. Specifically, the strands that tend to be the most difficult in each content area are Reading Informational Text in Reading, Measurement in Mathematics, and in Science—Science and Technology (Grade 5), Energy and Waves (Grade 8), and Cell Biology and Biochemistry (HS). In Composition, we look to the mean raw scores, noting that each strand represents a single rubric of 4 to 6 points. The mean raw scores are very similar across strands, where the Writing Language Conventions rubric or strand was slightly more difficult in Grades 4 and 7, while the Writing Topic Development was slightly more difficult in Grade 10.

In strands where there are very few items, reliabilities are lower, as would be expected. The degree of reliability that is required to interpret these strand scores, as for any test score, must therefore be carefully considered. These coefficients are computed on all valid 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 4 items and as many as 23, range between 0.40 and 0.88.

As an additional measure of internal consistency, correlations have been produced between strands within each grade and content area. These are provided in Tables 36–39. A review of the correlations shows fairly strong relationships amongst strands within content area. Specifically, in Table 36, the DC CAS 2012 Reading strand and total test correlations for all grades are presented. The Reading strand correlations are moderate to high for all grades.

Table 37 displays the correlations for the DC CAS 2012 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 38, the DC CAS 2012 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.

The DC CAS 2012 rubric score and total Composition test correlations for all grades are presented in Table 39. 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.

Standard Errors of Measurement

Standard errors of measurement (SEMs) indicate the degree of unreliability in the test scores, and conditional SEMs specific to each scale score provide further evidence. Tables 40–43 list the number correct to scale score values, along with their associated IRT SEM values. It is most

important to review these at the cut scores that differentiated students by proficient level. The cut score SEMs range from 3 to 8 (Reading), 3 to 9 (Mathematics), 2 to 6 (Science/Biology) and 5 to 13 (Composition). The lowest SEMs are typically at the "Proficient" cut in all grades and content areas.

Proficiency Level Evidence

Student performance relative to their score is classified into one of four proficiency levels: Below Basic, Basic, Proficient, and Advanced. The categorizations are important for accountability purposes, as well as for teacher, students, and parents to understand the content meaning of the associated scale scores. The percentage of students in each category, referred to as "impact data," is provided in Table 44. The "overall pass rate" represents the combined impact data of the two upper levels, Proficient and Advanced, and is often the sum percentage referenced in accountability measures.

Tables 45-48 display the classification consistency and accuracy results for each cut score and across all cut scores for the 2012 DC CAS in Reading, Mathematics, Science/Biology, and Composition. (The same information is provided for each subgroup in tables in Appendix D.) These statistics provide indication of the reliability of the proficiency cut scores, which designate the categories within which student performance would be classified over multiple administrations of the same assessment. The classification consistency statistics can be interpreted like the correlations, where the closer to 1.00 the statistics, the stronger the reliability. As with other measures of reliability, the statistics are impacted by the number of data points or, in this case, items and score points. Step 2 of the classification consistency calculations rests on the total raw scores. For that reason, the reliabilities for Composition are likely to be lower than the assessments in other content areas with higher possible total raw scores/points. What can be seen from the results described, however, is that Composition remains comparable to the other content areas, even with fewer points.

The classification consistency in all grades in Reading, Mathematics, and Science/Biology range from 0.65 to 0.82, and are comparable to those in 2011, which ranged between 0.66 and 0.78. The classification consistency ranged from 0.52 to 0.86 in Composition. The kappa values, which indicate classification consistency beyond chance consistency, represent moderate to substantial consistency levels (Landis & Koch, 1997). The kappa coefficients in Reading, Mathematics, and Science/Biology coefficients range between 0.48 and 0.77, which is comparable with the 2011 results (0.48 to 0.68). Kappa coefficients in Composition this year range from 0.34 to 0.62.

The classification accuracy results range from 0.73 to 0.85 in Reading, Mathematics, and Science/Biology. The results are comparable with those in 2011, which also ranged between 0.73 and 0.84. In Composition, classification accuracies range from 0.62 to 0.91. These results suggest that the 2012 DC CAS assessments in all content areas 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. A review of the tables, though shows these rates quite low, ranging from 0.03 to 0.30 in Composition and from 0.00 to 0.17 in Reading, Mathematics, and Science/Biology.

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 2012 suggest that consistent and accurate performance level classifications are being made for students based on the DC CAS assessments.

Correlational Evidence across Content Areas

Using all scored data, the correlations across 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 between Reading, Mathematics, Science/Biology, and Composition total raw scores appear in Table 49.

Correlations are somewhat higher in the elementary grades than in the middle and high school grades. Correlations between Reading and Mathematics are 0.72 and higher; correlations of Reading and Mathematics scores with Science/Biology scores are 0.56 and higher; correlations with the Composition total scores are in the range of 0.46 to 0.64. 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 16. DC CAS 2012 Classical Item Level Statistics

Grade	Operational					Field Test				
	Number of Items	Mean *p* value	Mean Item-Total Correlation	Mean Omit Rate	Mean Not Reached Rate	Number of Items	Mean *p* value	Mean Item-Total Correlation	Mean Omit Rate	Mean Not Reached Rate
Reading										
2	35	0.64	0.40	2.33	0.72	42	0.45	0.36	1.59	0.17
3	48	0.65	0.45	1.07	0.26	38	0.50	0.37	1.05	0.13
4	48	0.62	0.42	0.55	0.14	38	0.45	0.34	0.59	0.08
5	48	0.65	0.42	0.66	0.35	38	0.48	0.37	0.66	0.13
6	48	0.64	0.41	0.57	0.23	38	0.52	0.35	0.82	0.28
7	48	0.63	0.38	0.64	0.23	38	0.43	0.27	0.82	0.33
8	48	0.59	0.39	0.70	0.33	40	0.48	0.34	1.41	0.52
9	48	0.58	0.42	3.18	1.92	40	0.48	0.34	3.60	2.55
10	48	0.61	0.41	1.97	1.15	40	0.50	0.36	2.66	1.50
Mathematics										
2	32	0.74	0.43	0.62	0.14	36	0.69	0.43	0.74	0.17
3	53	0.66	0.45	0.92	0.09	32	0.63	0.45	1.17	0.20
4	54	0.63	0.43	0.51	0.12	32	0.53	0.43	1.27	0.23
5	53	0.67	0.43	0.42	0.13	32	0.44	0.37	0.86	0.18
6	54	0.58	0.43	0.46	0.10	32	0.45	0.37	0.73	0.14
7	52	0.58	0.41	0.70	0.24	32	0.41	0.33	1.26	0.29
8	54	0.51	0.39	0.81	0.34	32	0.38	0.33	1.35	0.37
10	54	0.48	0.38	2.10	0.93	32	0.37	0.33	3.50	1.06
Science/Biology										
5	50	0.47	0.35	0.74	0.38	28	0.47	0.34	0.90	0.27
8	50	0.41	0.33	1.27	0.43	28	0.35	0.26	2.05	0.29
High School	50	0.41	0.29	1.76	0.92	28	0.35	0.25	2.70	0.62
Composition										
4	8	0.48	0.60*	1.18	1.18	N/A	N/A	N/A	N/A	N/A
7	8	0.56	0.68*	1.75	1.75	N/A	N/A	N/A	N/A	N/A
10	8	0.53	0.61*	5.80	5.80	N/A	N/A	N/A	N/A	N/A

*Item-total correlations for Composition include the Reading items along with which the Composition prompts were scaled.

Table 17. DC CAS 2012 Operational Inter-Rater Agreement for Constructed Response Items: Reading

Grade	Form	Item No.	Score Points	% of Agreement			Checkset Average Agreement Percentages
				Perfect	Adjacent	Perfect + Adjacent	
2	1-2	8	0-3	68	27	94	93
		33	0-3	86	5	91	98
3	1-2	12	0-3	70	25	95	91
		18	0-3	69	28	97	93
		38	0-3	73	23	95	95
4	1-2	5	0-3	78	20	98	91
		18	0-3	77	21	98	84
		67	0-3	68	27	95	93
5	1-2	19	0-3	73	25	98	80
		23	0-3	56	37	93	66
		67	0-3	69	22	91	64
6	1-2	14	0-3	63	34	97	66
		28	0-3	64	32	96	57
		66	0-3	75	23	98	77
7	1-2	13	0-3	68	28	97	80
		17	0-3	77	20	97	85
		44	0-3	73	20	93	83
8	1-2	17	0-3	74	20	94	84
		28	0-3	64	32	96	87
		68	0-3	80	19	98	81
9	1-2	9	0-2	75	24	99	89
		18	0-3	72	24	96	79
		54	0-3	77	22	98	77
10	1-2	6	0-3	69	26	95	75
		16	0-3	74	23	97	85
		38	0-3	78	21	99	92

***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 18. DC CAS 2012 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	
2	1-2	6	0-2	88	12	100	96
		26	0-2	97	3	100	97
3	1-2	6	0-3	81	18	99	93
		25	0-3	91	7	98	98
		60	0-3	89	11	100	96
4	1-2	6	0-3	94	6	100	96
		25	0-3	94	6	100	92
		60	0-3	97	3	100	98
5	1-2	6	0-3	89	9	98	97
		25	0-3	97	3	100	100
		60	0-3	89	10	99	98
6	1-2	6	0-3	90	9	100	99
		25	0-3	95	5	100	98
		60	0-3	95	5	99	95
7	1-2	6	0-3	87	12	99	90
		25	0-3	79	20	99	94
		60	0-3	94	4	99	95
8	1-2	6	0-3	90	10	99	100
		25	0-3	94	5	99	96
		60	0-3	83	15	98	88
10	1-2	6	0-3	79	19	98	84
		25	0-3	95	4	99	96
		60	0-3	87	12	99	93

***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 19. DC CAS 2012 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	92	8	100	95
		27	0-2	80	19	99	89
		51	0-2	94	4	99	94
8	1-2	13	0-2	91	6	97	85
		27	0-2	89	10	99	94
		51	0-2	81	17	98	83
High School	1-2	13	0-2	85	14	98	86
		27	0-2	86	13	99	93
		51	0-2	81	18	100	86

***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 20. DC CAS 2012 Operational Inter-Rater Agreement for Constructed Response Items: Composition

Grade	Form	Item No.	Score Points	% of Agreement			Checkset Average Agreement Percentages
				Perfect	Adjacent	Perfect + Adjacent	
4	1	1A	1-6	50	47	97	60
		1B	1-4	58	41	99	77
		1C	1-4	57	37	94	72
4	2	1A	1-6	48	38	86	73
		1B	1-4	58	38	96	76
		1C	1-4	58	33	92	76
4	3	1A	1-6	57	38	95	65
		1B	1-4	67	31	98	70
		1C	1-4	68	30	97	71
4	4	1A	1-6	55	40	95	78
		1B	1-4	61	39	100	82
		1C	1-4	58	40	98	76
7	1	1A	1-6	57	39	96	74
		1B	1-4	62	38	100	79
		1C	1-4	68	32	100	75
7	2	1A	1-6	51	39	90	67
		1B	1-4	64	33	97	76
		1C	1-4	61	34	95	78
7	3	1A	1-6	55	42	97	69
		1B	1-4	59	41	99	77
		1C	1-4	57	37	94	74
7	4	1A	1-6	62	36	98	79
		1B	1-4	61	38	99	79
		1C	1-4	64	33	97	83
10	1	1A	1-6	49	36	85	76
		1B	1-4	42	53	95	79
		1C	1-4	54	29	83	79
10	2	1A	1-6	51	38	90	75
		1B	1-4	65	31	95	79
		1C	1-4	62	26	88	79
10	3	1A	1-6	52	35	87	78
		1B	1-4	74	25	98	86
		1C	1-4	87	8	95	93
10	4	1A	1-6	49	40	89	82
		1B	1-4	61	34	95	84
		1C	1-4	52	38	90	86

***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 21. DC CAS 2012 Field Test Inter-Rater Agreement for Constructed Response Items: Reading

Grade	Form	Item No.	Score Points	% of Agreement			Checkset Average Agreement Percentages
				Perfect	Adjacent	Perfect + Adjacent	
2	1	49	0-3	70	25	95	85
	2	49	0-3	82	14	97	85
3	1	33	0-3	66	31	97	95
		59	0-3	74	22	95	82
	2	33	0-3	73	25	98	77
		59	0-3	76	23	98	86
4	1	39	0-3	74	25	99	86
		62	0-3	65	31	96	70
	2	39	0-3	72	24	96	69
		62	0-3	71	26	97	89
5	1	36	0-3	68	29	97	64
		60	0-3	57	36	94	72
	2	36	0-3	77	20	98	71
		60	0-3	69	24	93	64
6	1	36	0-3	65	31	96	75
		61	0-3	74	21	95	77
	2	36	0-3	67	30	97	78
		61	0-3	81	15	97	83
7	1	33	0-3	60	34	94	76
		60	0-3	71	25	96	72
	2	33	0-3	57	33	90	72
		60	0-3	60	34	95	72
8	1	38	0-3	58	39	97	82
		62	0-3	57	38	95	83
	2	38	0-3	63	33	96	77
		62	0-3	84	14	98	89
9	1	37	0-3	71	28	99	89
	2	64	0-3	78	19	97	80
10	1	32	0-3	59	36	95	82
		60	0-3	72	20	92	87
	2	32	0-3	60	36	96	86
		60	0-3	57	38	95	82

***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 22. DC CAS 2012 Field Test Inter-Rater Agreement for Constructed Response Items: Mathematics

Grade	Form	Item No.	Score Points	% of Agreement			Checkset Average Agreement Percentages
				Perfect	Adjacent	Perfect + Adjacent	
2	1	53	0-3	89	10	99	91
	2	53	0-3	87	12	99	96
3	1	32	0-3	91	8	99	96
		49	0-3	84	15	99	93
	2	32	0-3	92	6	98	98
		49	0-3	91	7	98	98
4	1	32	0-3	92	7	99	92
		49	0-3	93	7	99	96
	2	32	0-3	91	9	100	96
		49	0-3	96	4	100	97
5	1	32	0-3	88	11	99	89
		49	0-3	97	2	99	96
	2	32	0-3	82	17	99	95
		49	0-3	93	8	100	98
6	1	32	0-3	91	9	100	91
		49	0-3	77	19	96	88
	2	32	0-3	90	8	98	93
		49	0-3	89	11	100	84
7	1	32	0-3	96	2	98	95
		49	0-3	83	15	98	87
	2	32	0-3	95	5	99	93
		49	0-3	79	18	98	94
8	1	32	0-3	88	12	100	97
		49	0-3	86	12	98	91
	2	32	0-3	91	8	98	92
		49	0-3	83	16	99	92
10	1	32	0-3	89	6	95	93
		49	0-3	93	4	97	90
	2	32	0-3	92	6	98	85
		49	0-3	95	3	98	98

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

Table 23. DC CAS 2012 Field Test 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	17	0-2	93	6	99	98
		41	0-2	94	5	99	92
	2	17	0-2	73	24	97	82
		41	0-2	83	16	100	87
8	1	17	0-2	92	6	98	85
		41	0-2	87	12	99	92
	2	17	0-2	95	4	99	99
		41	0-2	97	2	99	99
High School	1	17	0-2	93	6	99	87
		41	0-2	76	22	98	78
	2	17	0-2	81	18	99	84
		41	0-2	83	17	99	88

***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 24. Numbers of Operational Items Flagged for DIF Using the Mantel-Haenszel Procedure: Reading

Reference Group	Focal Group	A	B	B-	C	C-
Grade 2 (total 35 items)						
Male	Female	35	0	0	0	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	35	0	0	0	0
	White	23	8	1	3	0
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	47	0	1	0	0
	White	35	7	1	5	0
Grade 4 (total 48 items)						
Male	Female	47	0	1	0	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	46	2	0	0	0
	White	32	9	1	6	0
Grade 5 (total 48 items)						
Male	Female	46	2	0	0	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	47	1	0	0	0
	White	37	5	0	6	0
Grade 6 (total 48 items)						
Male	Female	44	2	2	0	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	47	1	0	0	0
	White	38	1	1	8	0
Grade 7 (total 48 items)						
Male	Female	44	2	1	0	1
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	45	0	2	1	0
	White	29	9	2	8	0
Grade 8 (total 48 items)						
Male	Female	46	2	0	0	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	45	1	2	0	0
	White[1]	31	6	0	10	0

N/A= not applicable because case count requirements for the reference (400) and focal (200) groups were not met. See Table 5 for the numbers of examinees in each grade and subgroup.

Table 24. Numbers of Operational Items Flagged for DIF Using the Mantel-Haenszel Procedure: Reading *(continued)*

Grade 9 (total 48 items)						
Male	Female	45	1	0	1	1
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	45	0	3	0	0
	White	N/A	N/A	N/A	N/A	N/A
Grade 10 (total 48 items)						
Male	Female	47	0	1	0	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	45	2	1	0	0
	White	N/A	N/A	N/A	N/A	N/A

N/A= not applicable because case count requirements for the reference (400) and focal (200) groups were not met. See Table 5 for the numbers of examinees in each grade and subgroup.

Table 25. Numbers of Operational Items Flagged for DIF Using the Mantel-Haenszel Procedure: Mathematics

Reference Group	Focal Group	A	B	B-	C	C-
Grade 2 (total 32 items)						
Male	Female	31	0	1	0	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	30	1	1	0	0
	White	23	4	0	5	0
Grade 3 (total 53 items)						
Male	Female	50	1	2	0	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	53	0	0	0	0
	White	37	6	5	5	0
Grade 4 (total 54 items)						
Male	Female	51	2	1	0	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	53	0	0	1	0
	White	39	3	2	10	0
Grade 5 (total 53 items)						
Male	Female	49	4	0	0	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	52	1	0	0	0
	White	41	2	3	6	1
Grade 6 (total 54 items)						
Male	Female	53	0	0	0	1
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	51	2	0	1	0
	White	37	3	3	8	3
Grade 7 (total 52 items)						
Male	Female	50	1	1	0	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	51	0	1	0	0
	White	41	5	0	4	2
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	50	2	2	0	0
	White	35	7	4	5	3
Grade 10 (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	N/A	N/A	N/A	N/A	N/A

N/A= not applicable because case count requirements for the reference (400) and focal (200) groups were not met. See Table 5 for the numbers of examinees in each grade and subgroup.

Table 26. Numbers of Operational 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	50	0	0	0	0
	White	40	8	0	2	0
Grade 8 (total 50 items)						
Male	Female	49	1	0	0	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	49	1	0	0	0
	White	39	6	2	3	0
High School (total 50 items)						
Male	Female	49	1	0	0	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	48	0	2	0	0
	White	N/A	N/A	N/A	N/A	N/A

N/A= not applicable because case count requirements for the reference (400) and focal (200) groups were not met. See Table 5 for the numbers of examinees in each grade and subgroup.

Table 27. Numbers of Operational/Field Test Items Flagged for DIF Using the Mantel-Haenszel Procedure: Composition

Reference Group	Focal Group	A	B	B-	C	C-
Grade 4 (total 8 items)						
Male	Female	4	3	0	1	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	6	0	0	2	0
	White	4	0	1	0	0
Grade 7 (total 8 items)						
Male	Female	2	5	0	1	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	3	2	0	3	0
	White	N/A	N/A	N/A	N/A	N/A
Grade 10 (total 8 items)						
Male	Female	1	0	0	1	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	N/A	N/A	N/A	N/A	N/A
	White	N/A	N/A	N/A	N/A	N/A

N/A= not applicable because case count requirements for the reference (400) and focal (200) groups were not met. See Table 5 for the numbers of examinees in each grade and subgroup.

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

Reference Group	Focal Group	A	B	B-	C	C-
Grade 2 (total 42 items)						
Male	Female	42	0	0	0	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	40	1	0	0	1
	White	17	9	0	16	0
Grade 3 (total 38 items)						
Male	Female	37	1	0	0	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	35	1	2	0	0
	White	31	4	0	2	1
Grade 4 (total 38 items)						
Male	Female	37	0	0	1	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	37	0	1	0	0
	White	27	4	0	7	0
Grade 5 (total 38 items)						
Male	Female	36	2	0	0	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	38	0	0	0	0
	White	29	6	0	3	0
Grade 6 (total 38 items)						
Male	Female	38	0	0	0	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	37	0	1	0	0
	White[1]	26	4	0	7	0
Grade 7 (total 38 items)						
Male	Female	38	0	0	0	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	38	0	0	0	0
	White	26	6	0	6	0
Grade 8 (total 40 items)						
Male	Female	36	3	1	0	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	38	0	1	1	0
	White	16	1	1	1	1

N/A= not applicable because case count requirements for the reference (400) and focal (200) groups were not met.
See Table 5 for the numbers of examinees in each grade and subgroup.

Table 28. Numbers of Field Test Items Flagged for DIF Using the Mantel-Haenszel Procedure: Reading *(continued)*

Grade 9 (total 40 items)						
Male	Female	38	2	0	0	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	35	4	0	0	1
	White	N/A	N/A	N/A	N/A	N/A
Grade 10 (total 40 items)						
Male	Female	35	4	1	0	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	36	2	1	0	1
	White	N/A	N/A	N/A	N/A	N/A

N/A= not applicable because case count requirements for the reference (400) and focal (200) groups were not met.
See Table 5 for the numbers of examinees in each grade and subgroup.

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

Reference Group	Focal Group	A	B	B-	C	C-
Grade 2 (total 36 items)						
Male	Female	35	0	1	0	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	35	1	0	0	0
	White	25	3	2	5	1
Grade 3 (total 32 items)						
Male	Female	31	0	0	1	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	32	0	0	0	0
	White	25	1	0	3	2
Grade 4 (total 32 items)						
Male	Female	30	0	2	0	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	32	0	0	0	0
	White	19	3	2	7	1
Grade 5 (total 32 items)						
Male	Female	31	0	1	0	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	28	3	1	0	0
	White	15	2	0	15	0
Grade 6 (total 32 items)						
Male	Female	29	1	1	0	1
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	28	2	2	0	0
	White	26	0	2	3	1
Grade 7 (total 32 items)						
Male	Female	32	0	0	0	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	27	5	0	0	0
	White	23	3	0	6	0
Grade 8 (total 32 items)						
Male	Female	32	0	0	0	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	32	0	0	0	0
	White	21	5	1	4	1
Grade 10 (total 32 items)						
Male	Female	30	1	1	0	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	30	1	1	0	0
	White	N/A	N/A	N/A	N/A	N/A

N/A= not applicable because case count requirements for the reference (400) and focal (200) groups were not met. See Table 5 for the numbers of examinees in each grade and subgroup.

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

Reference Group	Focal Group	A	B	B-	C	C-
Grade 5 (total 28 items)						
Male	Female	28	0	0	0	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	27	1	0	0	0
	White	20	2	0	6	0
Grade 8 (total 28 items)						
Male	Female	26	0	2	0	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	28	0	0	0	0
	White	22	3	0	3	0
High School (total 28 items)						
Male	Female	28	0	0	0	0
African American	Asian	N/A	N/A	N/A	N/A	N/A
	Hispanic	27	1	0	0	0
	White	N/A	N/A	N/A	N/A	N/A

N/A= not applicable because case count requirements for the reference (400) and focal (200) groups were not met. See Table 5 for the numbers of examinees in each grade and subgroup.

Table 31. Total Test Scale and Raw Score Means and Reliability Statistics

Grade	Students with Test Scores	Number of Items	Alpha	Stratified Alpha	Feldt-Raju	Scale Score		Raw Score	
						Mean	SD	Mean	SD
Reading									
2	4,469	35	0.88	0.88	0.88	241.97	15.78	23.25	7.82
3	4,737	48	0.93	0.94	0.94	348.65	15.37	33.48	11.60
4	4,559	48	0.92	0.92	0.92	452.42	15.09	31.82	11.15
5	4,734	48	0.92	0.92	0.92	553.75	15.09	32.95	10.73
6	4,539	48	0.91	0.92	0.92	650.16	14.20	33.13	10.96
7	4,283	48	0.90	0.91	0.90	754.13	14.25	33.03	10.33
8	4,337	48	0.90	0.91	0.91	853.86	14.32	30.26	10.42
9	3,534	48	0.92	0.93	0.93	947.17	16.94	28.27	11.34
10	4,230	48	0.92	0.92	0.92	951.32	15.45	30.94	11.23
Mathematics									
2	4,499	32	0.89	0.89	0.90	253.94	15.27	37.03	10.68
3	4,771	53	0.93	0.94	0.94	352.25	17.70	37.04	12.53
4	4,590	54	0.93	0.93	0.93	456.65	15.75	36.73	12.29
5	4,747	53	0.93	0.93	0.93	557.66	16.67	39.20	12.30
6	4,551	54	0.93	0.94	0.94	651.21	17.11	33.02	13.01
7	4,297	52	0.92	0.92	0.93	753.33	17.49	31.49	11.84
8	4,341	54	0.92	0.92	0.92	850.23	16.59	29.39	11.75
10	3,466	54	0.91	0.92	0.92	946.80	18.80	27.28	11.83
Science/Biology									
5	4,697	50	0.89	0.89	0.89	548.40	13.28	24.55	9.69
8	4,253	50	0.88	0.88	0.88	848.66	17.76	21.00	9.49
HS	3,693	50	0.85	0.86	0.86	947.91	14.76	21.34	8.62
Composition*									
4	4,508	8	0.92	0.92	0.93	451.73	18.87	4.51	1.91
7	4,176	8	0.91	0.90	0.92	754.33	15.76	5.27	1.98
10	3,429	8	0.92	0.92	0.93	952.18	20.11	4.76	2.17

*8 items = 4 prompts scored twice with two Writing rubrics

Table 31. Coefficient Alpha Reliability for Reading Strand Scores

Grade	Content Strand		Number of Items	Mean *p* value	Standard Deviation	Reliability
2	1	Vocabulary Acquisition & Use	7	0.70	0.15	0.58
	3	Reading Informational Text	14	0.54	0.15	0.76
	4	Reading Literary Text	14	0.71	0.18	0.74
	Total Number of Items on DC CAS		35			
3	1	Vocabulary Acquisition & Use	8	0.72	0.13	0.71
	3	Reading Informational Text	19	0.56	0.14	0.83
	4	Reading Literary Text	21	0.70	0.14	0.87
	Total Number of Items on DC CAS		48			
4	1	Vocabulary Acquisition & Use	8	0.68	0.13	0.64
	3	Reading Informational Text	18	0.58	0.13	0.82
	4	Reading Literary Text	22	0.62	0.12	0.83
	Total Number of Items on DC CAS		48			
5	1	Vocabulary Acquisition & Use	8	0.67	0.12	0.63
	3	Reading Informational Text	18	0.59	0.17	0.78
	4	Reading Literary Text	22	0.70	0.12	0.86
	Total Number of Items on DC CAS		48			
6	1	Vocabulary Acquisition & Use	9	0.63	0.15	0.69
	3	Reading Informational Text	17	0.60	0.15	0.80
	4	Reading Literary Text	22	0.68	0.14	0.82
	Total Number of Items on DC CAS		48			
7	1	Vocabulary Acquisition & Use	8	0.62	0.12	0.67
	3	Reading Informational Text	20	0.62	0.14	0.78
	4	Reading Literary Text	20	0.64	0.10	0.79
	Total Number of Items on DC CAS		48			
8	1	Vocabulary Acquisition & Use	7	0.65	0.09	0.58
	3	Reading Informational Text	22	0.58	0.11	0.83
	4	Reading Literary Text	19	0.58	0.22	0.76
	Total Number of Items on DC CAS		48			
9	1	Vocabulary Acquisition & Use	8	0.58	0.20	0.58
	3	Reading Informational Text	23	0.58	0.14	0.87
	4	Reading Literary Text	17	0.58	0.14	0.82
	Total Number of Items on DC CAS		48			
10	1	Vocabulary Acquisition & Use	9	0.69	0.09	0.69
	3	Reading Informational Text	20	0.61	0.11	0.84
	4	Reading Literary Text	19	0.58	0.15	0.80
	Total Number of Items on DC CAS		48			

Table 33. Coefficient Alpha Reliability for Mathematics Strand Scores

Grade	Content Strand		Number of Items	Mean *p* value	Standard Deviation	Reliability
2	1	Operations & Algebraic Thinking	8	0.67	0.16	0.74
	2	Numbers & Operations Base Ten	9	0.80	0.10	0.70
	3	Geometry	4	0.83	0.03	0.41
	4	Measurement and Data	11	0.70	0.14	0.76
	Total Number of Items on DC CAS		32			
3	1	Number Sense & Operations	17	0.68	0.17	0.82
	2	Patterns, Relations & Algebra	9	0.72	0.16	0.75
	3	Geometry	5	0.65	0.25	0.46
	4	Measurement	11	0.51	0.13	0.80
	5	Data Analysis, Statistics & Probability	11	0.71	0.12	0.76
	Total Number of Items on DC CAS		53			
4	1	Number Sense & Operations	23	0.68	0.14	0.88
	2	Patterns, Relations & Algebra	8	0.65	0.16	0.66
	3	Geometry	5	0.64	0.27	0.49
	4	Measurement	7	0.45	0.12	0.54
	5	Data Analysis, Statistics & Probability	11	0.65	0.17	0.71
	Total Number of Items on DC CAS		54			
5	1	Number Sense & Operations	20	0.70	0.14	0.82
	2	Patterns, Relations & Algebra	11	0.72	0.12	0.75
	3	Geometry	6	0.64	0.20	0.53
	4	Measurement	9	0.57	0.11	0.69
	5	Data Analysis, Statistics & Probability	7	0.68	0.11	0.67
	Total Number of Items on DC CAS		53			
6	1	Number Sense & Operations	16	0.65	0.14	0.80
	2	Patterns, Relations & Algebra	14	0.49	0.13	0.79
	3	Geometry	8	0.63	0.11	0.53
	4	Measurement	6	0.53	0.15	0.66
	5	Data Analysis, Statistics & Probability	10	0.57	0.12	0.74
	Total Number of Items on DC CAS		54			
7	1	Number Sense & Operations	17	0.57	0.14	0.82
	2	Patterns, Relations & Algebra	13	0.59	0.12	0.74
	3	Geometry	7	0.55	0.20	0.50
	4	Measurement	7	0.53	0.11	0.70
	5	Data Analysis, Statistics & Probability	8	0.67	0.14	0.62
	Total Number of Items on DC CAS		52			
8	1	Number Sense & Operations	16	0.51	0.16	0.76
	2	Patterns, Relations & Algebra	19	0.53	0.08	0.81
	3	Geometry	6	0.48	0.09	0.48
	4	Measurement	4	0.36	0.15	0.41
	5	Data Analysis, Statistics & Probability	9	0.55	0.16	0.67
	Total Number of Items on DC CAS		54			
10	1	Number Sense & Operations	11	0.55	0.18	0.69
	2	Patterns, Relations & Algebra	19	0.45	0.11	0.84
	3	Geometry	7	0.53	0.10	0.58
	4	Measurement	7	0.40	0.11	0.40
	5	Data Analysis, Statistics & Probability	10	0.50	0.17	0.62
	Total Number of Items on DC CAS		54			

Table 34. Coefficient Alpha Reliability for Science/Biology Strand Scores

Grade	Content Strand		Number of Items	Mean *p* value	Standard Deviation	Reliability
5	1	Science and Technology	15	0.43	0.15	0.71
	2	Earth and Space Science	13	0.51	0.19	0.66
	3	Physical Science	10	0.49	0.10	0.66
	4	Life Science	12	0.47	0.13	0.65
	Total Number of Items on DC CAS		50			
8	1	Scientific Thinking and Inquiry	7	0.42	0.11	0.60
	2	Matter and Reactions	22	0.40	0.11	0.75
	3	Forces	9	0.45	0.09	0.63
	4	Energy and Waves	12	0.39	0.10	0.56
	Total Number of Items on DC CAS		50			
High School	1	Cell Biology & Biochemistry	14	0.38	0.13	0.58
	2	Genetics and Evolution	15	0.40	0.10	0.65
	3	Multicellular Organisms	11	0.47	0.13	0.61
	4	Ecosystems	10	0.40	0.12	0.54
	Total Number of Items on DC CAS		50			

Table 35. Coefficient Alpha Reliability for Composition Strand Scores

Grade	Content Strand		Number of Items Across Four Forms	Students with Test Scores Across the Four Forms	Mean Raw Score	STD of Raw Score	Correlation Between the Two Strands
4		Writing Topic Development	4	4,508	2.33	1.16	0.80
		Writing Language Conventions	4	4,508	2.18	0.85	--
7		Writing Topic Development	4	4,176	2.75	1.18	0.84
		Writing Language Conventions	4	4,176	2.52	0.89	--
10		Writing Topic Development	4	3,429	2.34	1.29	0.77
		Writing Language Conventions	4	3,429	2.42	1.02	--

Table 36. DC CAS 2012 Reading Strand Correlations by Grade

Grade	Content Strand	Acquisition & Use	Informational Text	Literary Text	Total Reading
2	Acquisition & Use	--	0.61	0.59	0.77
	Informational Text	0.61	--	0.73	0.93
	Literary Text	0.59	0.73	--	0.90
	Total Raw Score	0.77	0.93	0.90	--
3	Acquisition & Use	--	0.75	0.78	0.86
	Informational Text	0.75	--	0.81	0.94
	Literary Text	0.78	0.81	--	0.96
	Total Raw Score	0.86	0.94	0.96	--
4	Acquisition & Use	--	0.71	0.73	0.82
	Informational Text	0.71	--	0.82	0.94
	Literary Text	0.73	0.82	--	0.96
	Total Raw Score	0.82	0.94	0.96	--
5	Acquisition & Use	--	0.69	0.71	0.82
	Informational Text	0.69	--	0.79	0.93
	Literary Text	0.71	0.79	--	0.95
	Total Raw Score	0.82	0.93	0.95	--
6	Acquisition & Use	--	0.73	0.72	0.84
	Informational Text	0.73	--	0.80	0.94
	Literary Text	0.72	0.80	--	0.94
	Total Raw Score	0.84	0.94	0.94	--
7	Acquisition & Use	--	0.68	0.72	0.83
	Informational Text	0.68	--	0.77	0.93
	Literary Text	0.72	0.77	--	0.93
	Total Raw Score	0.83	0.93	0.93	--
8	Acquisition & Use	--	0.68	0.68	0.79
	Informational Text	0.68	--	0.77	0.95
	Literary Text	0.68	0.77	--	0.92
	Total Raw Score	0.79	0.95	0.92	--
9	Acquisition & Use	--	0.69	0.69	0.80
	Informational Text	0.69	--	0.82	0.96
	Literary Text	0.69	0.82	--	0.93
	Total Raw Score	0.80	0.96	0.93	--
10	Acquisition & Use	--	0.76	0.73	0.86
	Informational Text	0.76	--	0.79	0.95
	Literary Text	0.73	0.79	--	0.93
	Total Raw Score	0.86	0.95	0.93	--

Table 37. DC CAS 2012 Mathematics Strand Correlations by Grade

Grade	Content Strand	Operations & Algebraic Thinking	Numbers & Operations Base Ten	Geometry	Measurement & Data	Total Mathematics
2	Operations & Algebraic Thinking	--	0.70	N/A	0.72	0.9
	Numbers & Operations Base Ten	0.70	--	N/A	0.68	0.86
	Geometry	N/A	N/A	N/A	N/A	N/A
	Measurement & Data	0.72	0.68	N/A	--	0.90
	Total Raw Score	0.90	0.86	N/A	0.90	--

Grade	Content Strand	Number Sense & Operations	Patterns, Relations & Algebra	Geometry	Measurement	Data Analysis, Statistics & Probability	Total Mathematics
3	Number Sense & Operations	--	0.79	0.62	0.78	0.75	0.94
	Patterns, Relations & Algebra	0.79	--	0.54	0.69	0.71	0.86
	Geometry	0.62	0.54	--	0.58	0.58	0.72
	Measurement	0.78	0.69	0.58	--	0.68	0.88
	Data Analysis, Statistics & Probability	0.75	0.71	0.58	0.68	--	0.87
	Total Raw Score	0.94	0.86	0.72	0.88	0.87	--
4	Number Sense & Operations	--	0.79	0.65	0.61	0.75	0.95
	Patterns, Relations & Algebra	0.79	--	0.60	0.56	0.69	0.87
	Geometry	0.65	0.60	--	0.50	0.61	0.75
	Measurement	0.61	0.56	0.50	--	0.53	0.72
	Data Analysis, Statistics & Probability	0.75	0.69	0.61	0.53	--	0.86
	Total Raw Score	0.95	0.87	0.75	0.72	0.86	--
5	Number Sense & Operations	--	0.79	0.67	0.74	0.74	0.93
	Patterns, Relations & Algebra	0.79	--	0.66	0.70	0.73	0.90
	Geometry	0.67	0.66	--	0.64	0.63	0.80
	Measurement	0.74	0.70	0.64	--	0.67	0.85
	Data Analysis, Statistics & Probability	0.74	0.73	0.63	0.67	--	0.86
	Total Raw Score	0.93	0.90	0.80	0.85	0.86	--

Table 37. DC CAS 2012 Mathematics Strand Correlations by Grade *(continued)*

Grade	Content Strand	Number Sense & Operations	Patterns, Relations & Algebra	Geometry	Measurement	Data Analysis, Statistics & Probability	Total Mathematics
6	Number Sense & Operations	--	0.79	0.64	0.74	0.77	0.93
	Patterns, Relations & Algebra	0.79	--	0.60	0.73	0.74	0.91
	Geometry	0.64	0.60	--	0.57	0.58	0.74
	Measurement	0.74	0.73	0.57	--	0.69	0.85
	Data Analysis, Statistics & Probability	0.77	0.74	0.58	0.69	--	0.87
	Total Raw Score	0.93	0.91	0.74	0.85	0.87	--
7	Number Sense & Operations	--	0.8	0.62	0.74	0.7	0.94
	Patterns, Relations & Algebra	0.80	--	0.60	0.72	0.67	0.92
	Geometry	0.62	0.60	--	0.54	0.52	0.73
	Measurement	0.74	0.72	0.54	--	0.62	0.84
	Data Analysis, Statistics & Probability	0.70	0.67	0.52	0.62	--	0.80
	Total Raw Score	0.94	0.92	0.73	0.84	0.80	--
8	Number Sense & Operations	--	0.77	0.58	0.56	0.69	0.90
	Patterns, Relations & Algebra	0.77	--	0.62	0.57	0.72	0.94
	Geometry	0.58	0.62	--	0.46	0.57	0.75
	Measurement	0.56	0.57	0.46	--	0.51	0.68
	Data Analysis, Statistics & Probability	0.69	0.72	0.57	0.51	--	0.84
	Total Raw Score	0.90	0.94	0.75	0.68	0.84	--
10	Number Sense & Operations	--	0.73	0.66	0.52	0.65	0.86
	Patterns, Relations & Algebra	0.73	--	0.71	0.57	0.69	0.93
	Geometry	0.66	0.71	--	0.51	0.61	0.83
	Measurement	0.52	0.57	0.51	--	0.51	0.69
	Data Analysis, Statistics & Probability	0.65	0.69	0.61	0.51	--	0.82
	Total Raw Score	0.86	0.93	0.83	0.69	0.82	--

Table 38. DC CAS 2012 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	--	0.66	0.68	0.68	0.89
	Earth and Space Science	0.66	--	0.65	0.65	0.86
	Physical Science	0.68	0.65	--	0.65	0.85
	Life Science	0.68	0.65	0.65	--	0.86
	Total Raw Score	0.89	0.86	0.85	0.86	--
Grade	**Content Strand**	**Scientific Thinking and Inquiry**	**Matter and Reactions**	**Forces**	**Energy and Waves**	**Total Science**
8	Scientific Thinking and Inquiry	--	0.64	0.63	0.54	0.80
	Matter and Reactions	0.64	--	0.67	0.61	0.92
	Forces	0.63	0.67	--	0.59	0.84
	Energy and Waves	0.54	0.61	0.59	--	0.79
	Total Raw Score	0.80	0.92	0.84	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	0.57	0.55	0.83
	Genetics and Evolution	0.62	--	0.59	0.59	0.86
	Multicellular Organisms	0.57	0.59	--	0.58	0.82
	Ecosystems	0.55	0.59	0.58	--	0.80
	Total Raw Score	0.83	0.86	0.82	0.80	--

Table 39. DC CAS 2012 Composition Rubric Score Correlations by Grade

Grade	Content Strand	Topic Development	Language Conventions	Total Composition
4	Topic Development	--	0.78	0.96
	Language Conventions	0.78	--	0.92
	Total Raw Score	0.96	0.92	--
7	Topic Development	--	0.81	0.97
	Language Conventions	0.81	--	0.94
	Total Raw Score	0.97	0.94	--
10	Topic Development	--	0.69	0.95
	Language Conventions	0.69	--	0.89
	Total Raw Score	0.95	0.89	--

Table 40. DC CAS 2012 Number Correct to Scale Score Conversions with Associated Standard Errors of Measurement (SEM): Reading

Raw Score	Grade 2		Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 9		Grade 10	
	Scale Score	SEM	Scale Score	SEM	Scale Score	SEM	Scale Score	SEM	Scale Score	SEM	Scale Score	SEM	Scale Score	SEM	Scale Score	SEM	Scale Score	SEM
0	200	25	300	31	400	37	500	38	600	35	700	38	800	39	900	38	900	37
1	200	25	300	31	400	37	500	38	600	35	700	38	800	39	900	38	900	37
2	200	25	300	31	400	37	500	38	600	35	700	38	800	39	900	38	900	37
3	200	25	300	31	400	37	500	38	600	35	700	38	800	39	900	38	900	37
4	200	25	300	31	400	37	500	38	600	35	700	38	800	39	900	38	900	37
5	200	25	300	31	400	37	500	38	600	35	700	38	800	39	900	38	900	37
6	200	25	300	31	400	37	500	38	600	35	700	38	800	39	900	38	900	37
7	200	25	300	31	400	37	500	38	600	35	700	38	800	39	900	38	900	37
8	206	19	300	31	400	37	500	38	600	35	700	38	800	39	900	38	900	37
9	213	12	300	31	400	37	500	38	600	35	700	38	800	39	900	38	900	37
10	217	9	302	29	412	25	500	38	600	35	700	38	816	23	900	38	909	28
11	220	7	313	18	421	16	518	20	616	19	714	24	824	15	910	28	919	18
12	222	6	319	12	425	12	524	14	622	13	721	17	829	10	920	18	924	13
13	225	6	322	9	429	9	528	10	626	9	726	12	832	8	925	13	928	10
14	227	5	325	8	431	7	531	8	628	7	729	10	834	7	929	10	931	8
15	228	5	327	6	433	6	533	6	630	6	732	8	836	6	932*	8	933	7
16	230	5	329	6	435	6	535	6	632	5	734	7	838	6	934	7	935	6
17	232*	5	331	5	437	5	537	5	634	5	736	6	839	5	936	6	937	5
18	233	5	332	5	438	5	538	5	635	4	738	5	841*	5	938	5	938	5
19	235	4	334	4	439*	4	539	4	636	4	739*	5	842	5	939	5	940*	5
20	236	4	335	4	441	4	541*	4	638	4	741	5	844	4	941	4	941	4
21	238	4	336	4	442	4	542	4	639	4	742	4	845	4	942	4	942	4
22	239	4	337	4	443	4	543	4	640*	4	743	4	846	4	943	4	943	4
23	241	4	339*	4	444	4	544	4	641	3	744	4	847	4	944	4	944	4
24	243	4	340	3	445	4	545	3	642	3	745	4	848	4	945	4	945	4
25	244	4	341	3	446	3	546	3	643	3	747	4	849	4	946	3	946	4
26	246*	5	342	3	447	3	547	3	644	3	748	4	851	4	947	3	947	3
27	248	5	343	3	448	3	548	3	645	3	749	4	852	3	948	3	948	3

*Proficiency Level Scale Score cuts (Basic, Proficient, Advanced)

Table 40. DC CAS 2012 Number Correct to Scale Score Conversions with Associated Standard Errors of Measurement (SEM): Reading *(continued)*

Raw Score	Grade 2		Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 9		Grade 10	
	Scale Score	SEM	Scale Score	SEM	Scale Score	SEM	Scale Score	SEM	Scale Score	SEM	Scale Score	SEM	Scale Score	SEM	Scale Score	SEM	Scale Score	SEM
28	249	5	343	3	449	3	549	3	646	3	750	3	853	3	949	3	949	3
29	251	5	344	3	450	3	550	3	646	3	751	3	854	3	950*	3	950	3
30	253	5	345	3	451	3	551	3	647	3	752	3	854	3	951	3	951	3
31	255	5	346	3	452	3	552	3	648	3	753	3	855	3	952	3	952	3
32	258	6	347	3	453	3	552	3	649	3	753	3	856*	3	953	3	953	3
33	261	6	348	3	454	3	553	3	650	3	754	3	857	3	954	3	954	3
34	264*	7	349	3	455*	3	554	3	651	3	755	3	858	3	955	3	955	3
35	268	8	350	3	456	3	555	3	652	3	756*	3	859	3	956	3	956*	3
36	273	9	351	3	457	3	556*	3	653	3	757	3	860	3	957	3	957	3
37	280	10	352	3	458	3	558	3	654	3	758	3	862	3	958	3	958	3
38	290	16	353	3	459	3	559	3	655*	3	759	3	863	3	959	3	959	3
39	299	23	354*	3	460	3	560	4	656	3	761	3	864	4	960*	3	960	3
40	.	.	355	3	461	3	561	4	657	3	762	3	865	4	961	3	961	3
41	.	.	356	3	462	3	562	4	658	3	763	4	866	4	962	3	963	4
42	.	.	357	3	463	4	564	4	659	3	764	4	867	4	964	3	964	4
43	.	.	358	3	464	4	565	4	660	3	765	4	869	4	965	3	965	4
44	.	.	360	4	466	4	567	4	662	4	767	4	870*	4	966	4	967	4
45	.	.	361	4	467	4	569	5	663	4	768*	4	872	4	968	4	968	4
46	.	.	363	4	469	5	571	5	665	4	770	4	873	4	969	4	970*	5
47	.	.	365	4	471	5	573*	5	666	4	772	5	875	5	971	4	972	5
48	.	.	367	4	474*	6	576	6	668	4	774	5	877	5	974	5	974	5
49	.	.	369	4	477	7	579	6	670	4	776	6	880	5	976	5	977	6
50	.	.	371	5	481	8	582	7	672*	5	779	6	882	6	980	6	981	7
51	.	.	375*	6	486	10	587	8	676	6	783	8	886	7	985	8	985	8
52	.	.	379	8	494	13	594	11	680	8	789	10	892	10	993	12	992	11
53	.	.	388	12	499	15	599	13	689	13	799	16	899	14	999	16	999	13
54	.	.	399	19	499	15	599	13	699	21	799	16	899	14	.	.	999	13

*Proficiency Level Scale Score cuts (Basic, Proficient, Advanced)

Table 41. DC CAS 2012 Number Correct to Scale Score Conversions with Associated Standard Errors of Measurement (SEM): Mathematics

Raw Score	Grade 2		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	Scale Score	SEM
0	200	35	300	26	400	38	500	32	600	37	700	39	800	43	900	35
1	200	35	300	26	400	38	500	32	600	37	700	39	800	43	900	35
2	200	35	300	26	400	38	500	32	600	37	700	39	800	43	900	35
3	200	35	300	26	400	38	500	32	600	37	700	39	800	43	900	35
4	200	35	300	26	400	38	500	32	600	37	700	39	800	43	900	35
5	200	35	300	26	400	38	500	32	600	37	700	39	800	43	900	35
6	200	35	300	26	400	38	500	32	600	37	700	39	800	43	900	35
7	209	26	300	26	400	38	500	32	600	37	700	39	800	43	900	35
8	220	15	300	26	400	38	500	32	600	37	700	39	800	43	900	35
9	225	10	300	26	400	38	500	32	600	37	700	39	800	43	900	35
10	228	8	300	26	400	38	500	32	600	37	700	39	800	43	900	35
11	231	6	302	24	400	38	500	32	600	37	706	33	800	43	907	28
12	233	6	309	17	409	28	507	25	614	22	717	22	806	37	915	20
13	236	5	314	13	419	19	514	18	621	16	723	15	820	23	921	15
14	237	5	317	11	424	14	519	13	625	12	728	11	827	16	925	12
15	239	4	320	9	428	10	523	11	628	9	731	9	831	12	929	10
16	241	4	323	8	431	9	526	9	631	8	734	8	834	9	931	9
17	242	4	325	7	433	8	529	8	633	7	736*	7	837*	8	934*	8
18	244*	4	327	7	435	7	531	7	635	6	738	6	839	7	936	7
19	245	4	329	6	437	6	533	7	637*	6	740	6	841	6	938	7
20	247	3	331	6	439	6	535	6	639	5	741	5	842	5	940	6
21	248	3	333	6	440	5	537	6	640	5	743	5	844	5	942	6
22	249	3	334	5	442	5	538	5	641	5	744	5	845	5	943	5
23	251	3	336	5	443*	5	540	5	643	4	745	4	846	4	945	5
24	252	3	337	5	444	4	541	5	644	4	747	4	848	4	946	5
25	253	3	339	5	445	4	542	5	645	4	748	4	849	4	948	5
26	255*	3	340*	4	446	4	544*	4	646	4	749	4	850*	4	949	4
27	257	4	341	4	448	4	545	4	647	4	750	4	851	4	950	4

*Proficiency Level Scale Score cuts (Basic, Proficient, Advanced)

Table 41. DC CAS 2012 Number Correct to Scale Score Conversions with Associated Standard Errors of Measurement (SEM): Mathematics *(continued)*

Raw Score	Grade 2		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	Scale Score	SEM
28	258	4	342	4	449	4	546	4	648	4	751	4	852	4	951*	4
29	261	4	344	4	450	4	547	4	649	3	752*	4	853	3	952	4
30	263	5	345	4	451	3	548	4	650	3	753	4	854	3	954	4
31	267	6	346	4	452	3	549	4	651	3	754	4	855	3	955	4
32	273*	9	347	4	453	3	550	4	652	3	755	3	855	3	956	4
33	299	35	348	4	454	3	551	4	653	3	756	3	856	3	957	4
34	.	.	349	4	454	3	552	4	654*	3	757	3	857	3	958	4
35	.	.	350	3	455	3	553	3	655	3	758	3	858	3	959	4
36	.	.	351	3	456	3	554	3	656	3	759	3	859	3	960	3
37	.	.	352	3	457	3	555	3	657	3	760	3	860	3	961	3
38	.	.	353	3	458*	3	556	3	657	3	761	3	861	3	962	3
39	.	.	354	3	459	3	557	3	658	3	762	3	861	3	963	3
40	.	.	355	3	460	3	558	3	659	3	763	3	862	3	964	3
41	.	.	356	3	461	3	559	3	660	3	764	3	863	3	965	3
42	.	.	357	3	462	3	560*	3	661	3	765	3	864	3	966	3
43	.	.	359	3	463	3	561	3	662	3	767	4	865	3	967	4
44	.	.	360*	3	464	3	562	3	663	3	768	4	866	3	968	4
45	.	.	361	4	465	3	563	3	664	3	769	4	867	3	969	4
46	.	.	362	4	466	3	564	3	665	3	771*	4	868*	3	970	4
47	.	.	363	4	467	3	565	3	666	3	772	4	869	3	972*	4
48	.	.	365	4	468	3	567	4	667	3	774	4	870	3	973	4
49	.	.	366	4	470	4	568	4	668*	3	776	5	871	3	974	4
50	.	.	368	4	471	4	569	4	669	3	778	5	873	4	976	4
51	.	.	369	4	472	4	571	4	671	4	781	6	874	4	978	5
52	.	.	371	5	474*	4	572	4	672	4	785	7	876	4	980	5
53	.	.	373	5	476	4	574	5	673	4	790	9	878	4	982	5
54	.	.	376*	5	477	4	576*	5	675	4	798	12	880	5	985	6
55	.	.	378	6	480	5	579	6	677	5	799	13	883	5	988	7

*Proficiency Level Scale Score cuts (Basic, Proficient, Advanced)

Table 41. DC CAS 2012 Number Correct to Scale Score Conversions with Associated Standard Errors of Measurement (SEM): Mathematics *(continued)*

Raw Score	Grade 2		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	Scale Score	SEM
56	.	.	382	7	482	5	583	7	679	5	799	13	886	6	991	8
57	.	.	387	9	485	6	587	9	682	6	.	.	890	7	996	9
58	.	.	397	14	490	8	596	14	687	8	.	.	894	8	999	10
59	.	.	399	15	498	11	599	15	695	13	.	.	899	10	999	10
60	.	.	.	.	499	12	.	.	699	16	.	.	899	10	999	10

*Proficiency Level Scale Score cuts (Basic, Proficient, Advanced)

Table 42: DC CAS 2012 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	45	800	53	900	50
1	500	45	800	53	900	50
2	500	45	800	53	900	50
3	500	45	800	53	900	50
4	500	45	800	53	900	50
5	500	45	800	53	900	50
6	500	45	800	53	900	50
7	500	45	800	53	900	50
8	500	45	800	53	900	50
9	500	45	800	53	900	50
10	508	36	800	53	900	50
11	526	18	830	23	925	25
12	532	13	839	14	934	16
13	536	9	843	10	938	12
14	538	7	846	7	942	9
15	540	6	848	6	944	7
16	542*	5	849*	5	946*	6
17	543	4	851	4	948	5
18	545	4	852	4	949	4
19	546	4	853	4	950	4
20	547	4	854	3	951	4
21	548	3	855	3	952*	3
22	549	3	856*	3	953	3
23	550	3	857	3	954	3
24	551	3	857	3	955	3
25	552	3	858	2	956	3
26	552	3	859	2	956	2
27	553*	3	859	2	957	2
28	554	3	860	2	958	2
29	555	2	861	2	959	2
30	556	2	861	2	959	2
31	556	2	862	2	960	2
32	557	2	863	2	960	2
33	558	2	863	2	961	2
34	559	2	864	2	962	2
35	559	2	864	2	962	2
36	560	2	865	2	963	2
37	561	2	866	2	964	2
38	561	2	866	2	964	2
39	562	2	867	2	965	2
40	563	2	868*	2	965	2
41	564*	2	869	2	966*	2
42	565	2	869	2	967	2
43	566	2	870	3	968	2
44	567	3	871	3	969	2

*Proficiency Level Scale Score cuts (Basic, Proficient, Advanced)

Table 42. DC CAS 2012 Number Correct to Scale Score Conversions with Associated Standard Errors of Measurement (SEM): Science/Biology *(continued)*

Raw Score	Grade 5		Grade 8		High School	
	Scale Score	SEM	Scale Score	SEM	Scale Score	SEM
45	568	3	872	3	970	2
46	569	3	874	3	971	3
47	570	3	875	3	972	3
48	572	3	877	4	973	3
49	574	4	879	4	975	4
50	576	4	881	5	977	4
51	579	6	885	6	981	6
52	585	8	891	9	987	9
53	599	21	899	15	999	20

Table 43. DC CAS 2012 Number Correct to Scale Score Conversions with Associated Standard Errors of Measurement (SEM): Composition

Raw Score	Prompt 1 Grade 4		Prompt 2 Grade 4		Prompt 3 Grade 4		Prompt 4 Grade 4	
	Scale Score	SEM	Scale Score	SEM	Scale Score	SEM	Scale Score	SEM
0	400	17	400	15	400	18	400	16
1	414	12	407	13	402	18	410	12
2	428	11	423	12	419	15	425	11
3	442	10	437	11	433	13	438	10
4	454*	10	449*	11	444*	13	447*	9
5	464*	9	459*	10	455	12	456*	9
6	472*	7	467	9	464*	12	464	9
7	477	7	474*	9	474*	12	472*	9
8	482	7	482	9	484	13	479	8
9	489	9	491	11	497	16	487	10
10	499	15	499	15	499	17	499	18
Raw Score	**Prompt 1 Grade 7**		**Prompt 2 Grade 7**		**Prompt 3 Grade 7**		**Prompt 4 Grade 7**	
	Scale Score	SEM	Scale Score	SEM	Scale Score	SEM	Scale Score	SEM
0	700	30	700	29	700	32	700	19
1	722	8	723	6	724	8	716	11
2	730	7	730	5	732	7	729	10
3	736	7	737	5	738	7	740	9
4	744*	7	743	6	746*	7	749*	8
5	752	7	750*	5	753	7	757*	8
6	760*	7	757*	6	760*	7	764	8
7	768*	7	763	5	767*	7	771*	8
8	775	7	769*	5	775	7	778	8
9	784	9	777	7	783	9	786	9
10	799	16	799	28	799	18	799	18
Raw Score	**Prompt 1 Grade 10**		**Prompt 2 Grade 10**		**Prompt 3 Grade 10**		**Prompt 4 Grade 10**	
	Scale Score	SEM	Scale Score	SEM	Scale Score	SEM	Scale Score	SEM
0	900	29	900	29	900	26	900	27
1	919	12	922	11	920	11	920	12
2	928	10	931	9	930	9	930	10
3	934	9	939	9	941	11	938	10
4	941	9	949*	9	953*	10	946*	9
5	948*	10	957*	9	962*	9	954	9
6	956*	10	964	8	968*	8	962*	10
7	964	11	971*	9	975	9	971*	11
8	974*	12	982	11	984	10	983	13
9	989	16	995	13	994	11	999	15
10	999	20	999	14	999	13	999	15

*Proficiency Level Scale Score cuts (Basic, Proficient, Advanced)

Table 44. DC CAS 2012 Percentages of Students at Each Performance Level

Content	Grade	Spring 2011 Impact Data					Spring 2012 Impact Data				
		N*	Percent of Students at Each Performance Level				N*	Percent of Students at Each Performance Level			
			Below Basic	Basic	Proficient	Advanced		Below Basic	Basic	Proficient	Advanced
Reading	2	--	--	--	--	--	4,491	22.24%	33.44%	36.12%	8.19%
	3	4796	22.29%	36.74%	37.82%	3.15%	4754	21.64%	38.16%	36.52%	3.68%
	4	4841	18.65%	37.57%	35.98%	7.79%	4589	15.86%	35.93%	41.73%	6.47%
	5	4797	15.13%	38.65%	39.13%	7.09%	4744	14.42%	38.26%	38.85%	8.47%
	6	4403	15.47%	42.31%	37.52%	4.70%	4545	17.43%	42.22%	36.13%	4.22%
	7	4456	11.11%	40.82%	35.19%	12.88%	4301	11.04%	39.64%	35.97%	13.35%
	8	4327	13.73%	37.60%	36.54%	12.13%	4359	14.18%	38.15%	37.62%	10.05%
	9	2891	18.13%	35.56%	22.73%	23.59%	4164	16.55%	41.50%	22.79%	19.16%
	10	4491	18.77%	37.16%	33.22%	10.84%	4272	18.38%	39.79%	32.44%	9.39%
Mathematics	2	--	--	--	--	--	4,514	20.36%	32.19%	36.89%	10.57%
	3	4823	22.79%	41.84%	24.26%	11.11%	4781	21.31%	42.25%	27.25%	9.18%
	4	4873	18.67%	35.50%	34.95%	10.88%	4603	15.75%	33.67%	37.98%	12.60%
	5	4817	18.37%	37.22%	32.61%	11.79%	4759	17.08%	34.33%	36.25%	12.33%
	6	4433	15.11%	39.66%	31.81%	13.42%	4567	16.33%	35.84%	32.95%	14.87%
	7	4485	14.47%	29.39%	43.41%	12.73%	4325	12.79%	29.36%	43.56%	14.29%
	8	4370	13.57%	28.60%	46.59%	11.24%	4381	14.13%	29.35%	45.04%	11.48%
	10	4464	23.97%	35.44%	34.68%	5.91%	4245	22.00%	36.35%	34.65%	7.00%
Science/ Biology	5	4765	20.29%	42.25%	31.21%	6.25%	4707	18.89%	42.91%	30.74%	7.46%
	8	4223	34.48%	29.24%	32.02%	4.26%	4263	35.05%	25.17%	34.37%	5.42%
	10	3790	32.22%	23.17%	42.14%	2.48%	3715	28.34%	26.59%	41.48%	3.58%
Composition	4	4755	10.91%	54.97%	25.95%	8.16%	4470	26.60%	31.99%	23.76%	17.65%
	7	4301	5.98%	60.71%	27.44%	5.88%	4146	18.07%	28.87%	32.34%	20.72%
	10	3761	12.28%	56.71%	22.63%	8.38%	3511	28.60%	25.43%	23.84%	22.13%

***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.

Table 45. Classification Consistency and Accuracy Rates by Grade and Cut Score: Reading

Grade	Reading Classification Consistency and Accuracy		Basic	Proficient	Advanced	All Cuts
2	Classification Consistency	Consistency	0.90	0.86	0.92	0.69
		Kappa	0.72	0.72	0.56	0.56
	Classification Accuracy	Accuracy	0.93	0.90	0.94	0.78
		False Positive Errors	0.03	0.04	0.01	0.08
		False Negative Errors	0.03	0.06	0.05	0.14
3	Classification Consistency	Consistency	0.93	0.90	0.96	0.79
		Kappa	0.78	0.79	0.58	0.68
	Classification Accuracy	Accuracy	0.95	0.92	0.97	0.85
		False Positive Errors	0.02	0.02	0.01	0.05
		False Negative Errors	0.03	0.05	0.02	0.10
4	Classification Consistency	Consistency	0.93	0.89	0.94	0.77
		Kappa	0.74	0.78	0.60	0.65
	Classification Accuracy	Accuracy	0.95	0.92	0.96	0.83
		False Positive Errors	0.02	0.02	0.01	0.06
		False Negative Errors	0.03	0.06	0.03	0.11
5	Classification Consistency	Consistency	0.94	0.88	0.93	0.75
		Kappa	0.75	0.77	0.61	0.64
	Classification Accuracy	Accuracy	0.96	0.92	0.95	0.82
		False Positive Errors	0.02	0.04	0.01	0.07
		False Negative Errors	0.02	0.05	0.04	0.11
6	Classification Consistency	Consistency	0.93	0.89	0.96	0.78
		Kappa	0.76	0.77	0.61	0.67
	Classification Accuracy	Accuracy	0.95	0.92	0.97	0.83
		False Positive Errors	0.02	0.03	0.01	0.06
		False Negative Errors	0.03	0.05	0.02	0.11
7	Classification Consistency	Consistency	0.94	0.86	0.91	0.71
		Kappa	0.70	0.72	0.65	0.59
	Classification Accuracy	Accuracy	0.96	0.90	0.94	0.80
		False Positive Errors	0.02	0.05	0.02	0.09
		False Negative Errors	0.02	0.05	0.04	0.11
8	Classification Consistency	Consistency	0.93	0.87	0.93	0.73
		Kappa	0.70	0.74	0.67	0.61
	Classification Accuracy	Accuracy	0.95	0.91	0.95	0.81
		False Positive Errors	0.03	0.05	0.02	0.10
		False Negative Errors	0.02	0.05	0.04	0.10
9	Classification Consistency	Consistency	0.91	0.90	0.92	0.73
		Kappa	0.64	0.80	0.77	0.63
	Classification Accuracy	Accuracy	0.94	0.93	0.94	0.81
		False Positive Errors	0.03	0.02	0.02	0.07
		False Negative Errors	0.03	0.04	0.04	0.11
10	Classification Consistency	Consistency	0.92	0.90	0.93	0.76
		Kappa	0.74	0.79	0.66	0.65
	Classification Accuracy	Accuracy	0.94	0.93	0.95	0.82
		False Positive Errors	0.02	0.03	0.01	0.06
		False Negative Errors	0.04	0.04	0.04	0.12

Table 46. Classification Consistency and Accuracy Rates by Grade and Cut Score: Mathematics

Grade	Mathematics Classification Consistency and Accuracy		Basic	Proficient	Advanced	All Cuts
2	Classification Consistency	Consistency	0.91	0.87	0.93	0.72
		Kappa	0.73	0.75	0.64	0.60
	Classification Accuracy	Accuracy	0.94	0.91	0.93	0.78
		False Positive Errors	0.03	0.04	0.04	0.10
		False Negative Errors	0.03	0.05	0.03	0.12
3	Classification Consistency	Consistency	0.93	0.91	0.94	0.78
		Kappa	0.79	0.81	0.68	0.69
	Classification Accuracy	Accuracy	0.95	0.93	0.95	0.83
		False Positive Errors	0.03	0.02	0.01	0.05
		False Negative Errors	0.03	0.05	0.04	0.11
4	Classification Consistency	Consistency	0.93	0.91	0.94	0.77
		Kappa	0.73	0.81	0.73	0.67
	Classification Accuracy	Accuracy	0.95	0.93	0.95	0.83
		False Positive Errors	0.02	0.03	0.02	0.06
		False Negative Errors	0.03	0.05	0.03	0.11
5	Classification Consistency	Consistency	0.94	0.91	0.92	0.77
		Kappa	0.78	0.82	0.68	0.68
	Classification Accuracy	Accuracy	0.95	0.93	0.95	0.83
		False Positive Errors	0.02	0.03	0.02	0.07
		False Negative Errors	0.03	0.03	0.04	0.10
6	Classification Consistency	Consistency	0.91	0.91	0.94	0.76
		Kappa	0.69	0.83	0.75	0.67
	Classification Accuracy	Accuracy	0.94	0.93	0.95	0.82
		False Positive Errors	0.03	0.02	0.02	0.07
		False Negative Errors	0.03	0.05	0.03	0.10
7	Classification Consistency	Consistency	0.92	0.95	0.96	0.82
		Kappa	0.76	0.90	0.89	0.77
	Classification Accuracy	Accuracy	0.90	0.96	0.89	0.75
		False Positive Errors	0.10	0.02	0.00	0.12
		False Negative Errors	0.01	0.01	0.11	0.13
8	Classification Consistency	Consistency	0.89	0.88	0.95	0.73
		Kappa	0.58	0.76	0.77	0.60
	Classification Accuracy	Accuracy	0.92	0.91	0.96	0.80
		False Positive Errors	0.04	0.04	0.01	0.09
		False Negative Errors	0.04	0.05	0.03	0.11
10	Classification Consistency	Consistency	0.88	0.89	0.96	0.74
		Kappa	0.65	0.78	0.75	0.62
	Classification Accuracy	Accuracy	0.91	0.92	0.97	0.81
		False Positive Errors	0.04	0.03	0.01	0.08
		False Negative Errors	0.05	0.05	0.02	0.12

Table 47. Classification Consistency and Accuracy Rates by Grade and Cut Score: Science/Biology

Grade	Science/Biology Classification Consistency and Accuracy		Basic	Proficient	Advanced	All Cuts
5	Classification Consistency	Consistency	0.87	0.88	0.96	0.71
		Kappa	0.60	0.76	0.70	0.58
	Classification Accuracy	Accuracy	0.91	0.92	0.96	0.79
		False Positive Errors	0.06	0.03	0.01	0.09
		False Negative Errors	0.04	0.05	0.03	0.12
8	Classification Consistency	Consistency	0.82	0.87	0.97	0.68
		Kappa	0.60	0.74	0.73	0.54
	Classification Accuracy	Accuracy	0.87	0.91	0.98	0.76
		False Positive Errors	0.06	0.03	0.01	0.09
		False Negative Errors	0.07	0.06	0.01	0.15
High School	Classification Consistency	Consistency	0.81	0.83	0.98	0.65
		Kappa	0.54	0.65	0.68	0.48
	Classification Accuracy	Accuracy	0.86	0.88	0.98	0.73
		False Positive Errors	0.06	0.04	0.00	0.10
		False Negative Errors	0.08	0.08	0.02	0.17

Table 48. Classification Consistency and Accuracy Rates by Grade and Cut Score: Composition

Grade	Composition Classification Consistency and Accuracy		Basic	Proficient	Advanced	All Cuts
4	Classification Consistency	Consistency	0.81	0.78	0.85	0.53
		Kappa	0.55	0.53	0.51	0.36
	Classification Accuracy	Accuracy	0.87	0.84	0.89	0.63
		False Positive Errors	0.09	0.07	0.05	0.19
		False Negative Errors	0.04	0.09	0.06	0.18
7	Classification Consistency	Consistency	0.86	0.81	0.84	0.58
		Kappa	0.57	0.62	0.56	0.42
	Classification Accuracy	Accuracy	0.91	0.87	0.89	0.68
		False Positive Errors	0.06	0.08	0.04	0.16
		False Negative Errors	0.03	0.05	0.08	0.15
10	Classification Consistency	Consistency	0.78	0.78	0.84	0.52
		Kappa	0.46	0.55	0.57	0.34
	Classification Accuracy	Accuracy	0.84	0.85	0.88	0.62
		False Positive Errors	0.09	0.07	0.04	0.18
		False Negative Errors	0.07	0.08	0.07	0.20

Table 49. Correlations Between Reading, Mathematics, Science/Biology, and Composition Total Test Raw Scores, by Grade

Grade	Mathematics	Science/Biology*	Composition
Reading			
Grade 2	0.72	--	--
Grade 3	0.78	--	--
Grade 4	0.80	--	0.57
Grade 5	0.76	0.78	--
Grade 6	0.78	--	--
Grade 7	0.78	--	0.64
Grade 8	0.77	0.74	--
Grade 9	--	0.56	--
Grade 10	0.74	0.65	0.58
Mathematics			
Grade 4	--	--	0.55
Grade 5	--	0.72	--
Grade 7	--	--	0.58
Grade 8	--	0.79	--
Grade 10	--	0.63	0.56
Science/Biology			
Grade 10	--	--	0.46

***Note*:** "--" = not applicable.

*In Biology all grades were used in the analyses but only Grades 9 and 10 can be used for the correlations since the other grades are not in common with other content areas.

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. (1995). PARDUX (Version 1.7) [Computer program]. Unpublished.

Burket, G. R. (2000). ITEMWIN [Computer program]. Unpublished.

CTB/McGraw-Hill. (2011). *District of Columbia Comprehensive Assessment System (DC CAS) grade 9 reading bookmark standard setting technical report 2011*. Monterey, CA: Author.

CTB/McGraw-Hill. (2011). *District of Columbia Public Schools (DCPS) grade 9 reading technical report 2011*. Monterey, CA: Author.

CTB/McGraw-Hill. (2011). *District of Columbia Comprehensive Assessment System (DC CAS) test chairperson's manual: Reading and mathematics, composition, science, and biology.* Monterey, CA: Author.

CTB/McGraw-Hill. (2011). *District of Columbia Comprehensive Assessment System (DC CAS) test directions: Reading and mathematics (grades 4–8 and 10), composition (grades 4, 7, and 10), science (grades 5 and 8), and biology.* Monterey, CA: Author.

CTB/McGraw-Hill. (2012). *District of Columbia Comprehensive Assessment System (DC CAS) Standard Setting Technical Report for Grades 3–10 Reading, Grade 2 Reading and Mathematics, and Grades 4, 7, and 10 Composition*. Monterey, CA: Author.

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.

Jaeger, R. M. (1995). Setting standards for complex performances: An iterative, judgmental policy-capturing strategy. *Educational Measurement: Issues and Practice, 14*(4): 16–20.

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.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32*, 179–197.

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.

No Child Left Behind Act of 2001, Pub. L. No. 107—110, 115 Stat.1425 (2002).

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 Assessments 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.

U.S. Department of Education. (2009, January). *Standards and assessments peer review guidance: Information and examples for meeting requirements of the No Child Left Behind Act of 2001.* Retrieved December 7, 2010, from **http://www.ed.gov/policy/elsec/guid/saaprguidance.pdf**.

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

Understanding Literary or Informational Text

Score	Description
4	The response demonstrates an understanding of the complexities of the text. • Fully addresses the demands of the question or prompt • Effectively uses explicitly stated text as well as inferences drawn from the text to support an answer or claim
3	The response demonstrates an understanding of the text. • Addresses the demands of the question or prompt • Uses some explicitly stated text and/or some inferences drawn from the text to support an answer or claim
2	The response is incomplete or oversimplified and demonstrates a partial or literal understanding of the text. • Attempts to answer the question or address the prompt • Uses explicitly stated text that demonstrates some understanding
1	The response shows evidence of a minimal understanding of the text. • Shows evidence that some meaning has been derived from the text to answer the question • Has minimal textual evidence

Note: The Composition prompt will also be aligned to a Common Core Reading standard. Responses will demonstrate degrees of mastery of that reading standard. Reading standards that the composition prompts will align to may include:

- Grade 4: CC.4.R.I.1, CC.4.R.L.2, and CC.4.R.L.4 (see Reading tested standards)
- Grade 7: CC.7.R.I.1, CC.7.R.I.8, CC.7.R.L.1, and CC.7.R.L.2 (see Reading tested standards)
- Grade 10: CC.9-10.R.I.1, CC.9-10.R.I.2, CC.10.R.I.3, CC.9-10.R.L.2, and CC.9-10.R.L.6 (see Reading tested standards)

Appendix C: Operational and Field Test Item Adjusted *P* Values

Table C1. DC CAS 2012 Operational Form Item Adjusted *P* Values: Reading

Reading Grade 2								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	4,457	1	0.93		25	4,320	1	0.50
2	4,393	1	0.88		26	4,411	1	0.90
3	4,416	1	0.85		27	4,403	1	0.86
4	4,399	1	0.61		28	3,812	1	0.71
5	4,448	1	0.71		29	4,257	1	0.78
6	4,432	1	0.74		30	4,252	1	0.29
7	4,397	1	0.53		31	4,246	1	0.53
8	4,319	3	0.31		32	4,207	1	0.34
9	4,438	1	0.40		33	4,272	1	0.72
10	4,419	1	0.44		34	4,285	1	0.78
11	4,412	1	0.51		35	4,280	1	0.79
12	4,404	1	0.75					
13	4,444	1	0.87					
14	4,418	1	0.66					
15	4,345	1	0.56					
16	4,080	1	0.68					
17	4,397	1	0.41					
18	4,430	1	0.50					
19	4,419	1	0.62					
20	4,373	1	0.64					
21	4,420	1	0.61					
22	4,373	1	0.71					
23	4,310	3	0.45					
24	4,387	1	0.82					

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

Table C1. DC CAS 2012 Operational Form Item Adjusted *P* Values: Reading *(continued)*

Reading Grade 3							
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**	**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	4,726	1	0.88	25	4,708	1	0.74
2	4,720	1	0.79	26	4,691	1	0.70
3	4,727	1	0.71	27	4,598	3	0.44
4	4,733	1	0.85	28	4,711	1	0.48
5	4,718	1	0.74	29	4,696	1	0.34
6	4,732	1	0.78	30	4,690	1	0.45
7	4,724	1	0.79	31	4,691	1	0.25
8	4,703	1	0.46	32	4,703	1	0.54
9	4,707	1	0.63	33	4,704	1	0.91
10	4,714	1	0.60	34	4,701	1	0.76
11	4,702	1	0.53	35	4,696	1	0.65
12	4,554	3	0.43	36	4,693	1	0.60
13	4,702	1	0.50	37	4,686	1	0.62
14	4,625	1	0.70	38	4,707	1	0.72
15	4,691	1	0.58	39	4,696	1	0.54
16	4,688	1	0.73	40	4,698	1	0.71
17	4,688	1	0.72	41	4,670	1	0.76
18	4,537	3	0.46	42	4,662	1	0.52
19	4,717	1	0.79	43	4,609	1	0.73
20	4,714	1	0.81	44	4,580	1	0.79
21	4,706	1	0.74	45	4,673	1	0.58
22	4,704	1	0.90	46	4,564	1	0.58
23	4,715	1	0.80	47	4,664	1	0.52
24	4,718	1	0.63	48	4,662	1	0.77

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

Table C1. DC CAS 2012 Operational Form Item Adjusted *P* Values: Reading *(continued)*

Reading Grade 4							
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**	**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	4,556	1	0.67	25	4,551	1	0.73
2	4,556	1	0.49	26	4,545	1	0.66
3	4,558	1	0.65	27	4,542	1	0.54
4	4,533	1	0.57	28	4,550	1	0.62
5	4,421	3	0.42	29	4,554	1	0.47
6	4,549	1	0.52	30	4,552	1	0.73
7	4,544	1	0.71	31	4,553	1	0.66
8	4,542	1	0.40	32	4,549	1	0.51
9	4,539	1	0.78	33	4,553	1	0.51
10	4,538	1	0.79	34	4,550	1	0.77
11	4,535	1	0.77	35	4,552	1	0.72
12	4,535	1	0.69	36	4,549	1	0.58
13	4,529	1	0.76	37	4,548	1	0.59
14	4,526	1	0.67	38	4,537	1	0.65
15	4,516	1	0.72	39	4,548	1	0.75
16	4,503	1	0.52	40	4,544	1	0.52
17	4,479	1	0.90	41	4,540	1	0.33
18	4,376	3	0.45	42	4,539	1	0.55
19	4,555	1	0.66	43	4,535	1	0.51
20	4,554	1	0.66	44	4,510	1	0.60
21	4,553	1	0.54	45	4,516	1	0.79
22	4,554	1	0.59	46	4,510	1	0.58
23	4,547	1	0.48	47	4,493	1	0.77
24	4,551	1	0.61	48	4,456	3	0.39

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

Table C1. DC CAS 2012 Operational Form Item Adjusted *P* Values: Reading *(continued)*

Reading Grade 5								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	4,730	1	0.82		25	4,724	1	0.73
2	4,733	1	0.73		26	4,723	1	0.70
3	4,733	1	0.83		27	4,720	1	0.65
4	4,731	1	0.75		28	4,722	1	0.54
5	4,727	1	0.73		29	4,718	1	0.62
6	4,733	1	0.87		30	4,720	1	0.64
7	4,731	1	0.53		31	4,721	1	0.86
8	4,729	1	0.77		32	4,718	1	0.61
9	4,720	1	0.50		33	4,720	1	0.84
10	4,723	1	0.68		34	4,721	1	0.70
11	4,723	1	0.38		35	4,718	1	0.56
12	4,727	1	0.52		36	4,720	1	0.62
13	4,724	1	0.78		37	4,718	1	0.75
14	4,723	1	0.72		38	4,718	1	0.57
15	4,722	1	0.48		39	4,718	1	0.87
16	4,718	1	0.70		40	4,717	1	0.71
17	4,707	1	0.74		41	4,711	1	0.61
18	4,681	1	0.70		42	4,635	1	0.42
19	4,543	3	0.32		43	4,636	1	0.57
20	4,722	1	0.71		44	4,632	1	0.76
21	4,725	1	0.90		45	4,637	1	0.56
22	4,717	1	0.65		46	4,633	1	0.79
23	4,643	3	0.36		47	4,622	1	0.58
24	4,725	1	0.66		48	4,547	3	0.24

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

Table C1. DC CAS 2012 Operational Form Item Adjusted *P* Values: Reading *(continued)*

Reading Grade 6								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	4,532	1	0.57		25	4,529	1	0.76
2	4,539	1	0.69		26	4,523	1	0.52
3	4,535	1	0.54		27	4,490	1	0.48
4	4,535	1	0.70		28	4,449	3	0.46
5	4,529	1	0.69		29	4,520	1	0.93
6	4,538	1	0.76		30	4,517	1	0.90
7	4,531	1	0.52		31	4,517	1	0.60
8	4,532	1	0.57		32	4,519	1	0.57
9	4,529	1	0.39		33	4,513	1	0.39
10	4,532	1	0.77		34	4,516	1	0.61
11	4,533	1	0.70		35	4,520	1	0.87
12	4,530	1	0.56		36	4,519	1	0.81
13	4,510	1	0.58		37	4,518	1	0.41
14	4,452	3	0.37		38	4,517	1	0.86
15	4,520	1	0.67		39	4,517	1	0.76
16	4,518	1	0.44		40	4,511	1	0.62
17	4,517	1	0.46		41	4,515	1	0.66
18	4,514	1	0.52		42	4,509	1	0.67
19	4,510	1	0.75		43	4,495	1	0.63
20	4,533	1	0.79		44	4,494	1	0.80
21	4,533	1	0.69		45	4,492	1	0.52
22	4,529	1	0.65		46	4,486	1	0.81
23	4,531	1	0.73		47	4,414	3	0.48
24	4,531	1	0.69		48	4,382	1	0.85

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

Table C1. DC CAS 2012 Operational Form Item Adjusted *P* Values: Reading *(continued)*

Reading Grade 7								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	4,273	1	0.56		25	4,265	1	0.45
2	4,278	1	0.62		26	4,259	1	0.68
3	4,278	1	0.60		27	4,257	1	0.61
4	4,268	1	0.58		28	4,258	1	0.65
5	4,280	1	0.62		29	4,258	1	0.73
6	4,273	1	0.48		30	4,260	1	0.82
7	4,273	1	0.51		31	4,256	1	0.76
8	4,268	1	0.60		32	4,255	1	0.50
9	4,269	1	0.54		33	4,256	1	0.63
10	4,265	1	0.49		34	4,257	1	0.56
11	4,262	1	0.76		35	4,243	1	0.58
12	4,238	1	0.68		36	4,192	3	0.53
13	4,212	3	0.55		37	4,253	1	0.73
14	4,245	1	0.87		38	4,253	1	0.68
15	4,245	1	0.89		39	4,252	1	0.51
16	4,230	1	0.59		40	4,254	1	0.39
17	4,165	3	0.51		41	4,247	1	0.46
18	4,271	1	0.82		42	4,251	1	0.64
19	4,273	1	0.84		43	4,250	1	0.58
20	4,273	1	0.63		44	4,248	1	0.77
21	4,272	1	0.59		45	4,247	1	0.71
22	4,270	1	0.59		46	4,244	1	0.61
23	4,270	1	0.67		47	4,244	1	0.59
24	4,270	1	0.59		48	4,246	1	0.78

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

Table C1. DC CAS 2012 Operational Form Item Adjusted *P* Values: Reading *(continued)*

Reading Grade 8								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	4,326	1	0.60		25	4,317	1	0.53
2	4,335	1	0.90		26	4,319	1	0.59
3	4,336	1	0.93		27	4,304	1	0.56
4	4,333	1	0.49		28	4,213	3	0.42
5	4,336	1	0.88		29	4,316	1	0.57
6	4,326	1	0.86		30	4,314	1	0.73
7	4,330	1	0.72		31	4,314	1	0.51
8	4,331	1	0.63		32	4,309	1	0.51
9	4,329	1	0.65		33	4,311	1	0.39
10	4,330	1	0.77		34	4,313	1	0.63
11	4,323	1	0.68		35	4,312	1	0.60
12	4,319	1	0.38		36	4,305	1	0.27
13	4,320	1	0.49		37	4,308	1	0.30
14	4,313	1	0.63		38	4,310	1	0.65
15	4,313	1	0.46		39	4,309	1	0.25
16	4,299	1	0.78		40	4,314	1	0.76
17	4,174	3	0.34		41	4,306	1	0.58
18	4,325	1	0.66		42	4,306	1	0.44
19	4,320	1	0.48		43	4,275	1	0.67
20	4,321	1	0.70		44	4,277	1	0.66
21	4,324	1	0.66		45	4,271	1	0.49
22	4,323	1	0.57		46	4,275	1	0.56
23	4,324	1	0.71		47	4,268	1	0.64
24	4,320	1	0.75		48	4,161	3	0.33

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

Table C1. DC CAS 2012 Operational Form Item Adjusted *P* Values: Reading *(continued)*

Reading Grade 9							
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**	**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	3,532	1	0.77	25	3,480	1	0.48
2	3,522	1	0.59	26	3,489	1	0.78
3	3,526	1	0.71	27	3,488	1	0.60
4	3,523	1	0.79	28	3,384	1	0.43
5	3,525	1	0.58	29	3,384	1	0.65
6	3,525	1	0.83	30	3,385	1	0.63
7	3,512	1	0.78	31	3,378	1	0.32
8	3,512	1	0.62	32	3,378	1	0.43
9	3,324	2	0.42	33	3,380	1	0.47
10	3,510	1	0.43	34	3,381	1	0.64
11	3,511	1	0.77	35	3,376	1	0.49
12	3,508	1	0.83	36	3,380	1	0.68
13	3,481	1	0.46	37	3,376	1	0.72
14	3,485	1	0.66	38	3,376	1	0.70
15	3,480	1	0.39	39	3,360	1	0.42
16	3,476	1	0.54	40	3,368	1	0.54
17	3,453	1	0.50	41	3,371	1	0.37
18	3,027	3	0.22	42	3,369	1	0.48
19	3,495	1	0.67	43	3,359	1	0.47
20	3,494	1	0.66	44	2,942	3	0.30
21	3,493	1	0.59	45	3,348	1	0.64
22	3,494	1	0.66	46	3,349	1	0.69
23	3,492	1	0.65	47	3,350	1	0.65
24	3,490	1	0.40	48	3,341	1	0.64

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

Table C1. DC CAS 2012 Operational Form Item Adjusted *P* Values: Reading *(continued)*

Reading Grade 10								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	4,225	1	0.55		25	4,142	1	0.60
2	4,224	1	0.69		26	4,147	1	0.65
3	4,217	1	0.68		27	4,139	1	0.75
4	4,210	1	0.64		28	3,897	3	0.34
5	4,208	1	0.67		29	4,139	1	0.42
6	3,835	3	0.31		30	4,142	1	0.72
7	4,218	1	0.71		31	4,140	1	0.59
8	4,220	1	0.82		32	4,136	1	0.67
9	4,218	1	0.78		33	4,143	1	0.55
10	4,217	1	0.58		34	4,132	1	0.46
11	4,217	1	0.72		35	4,133	1	0.55
12	4,207	1	0.66		36	4,130	1	0.24
13	4,212	1	0.70		37	4,135	1	0.48
14	4,206	1	0.80		38	4,132	1	0.51
15	4,206	1	0.75		39	4,128	1	0.69
16	3,947	3	0.54		40	4,130	1	0.53
17	4,202	1	0.72		41	4,100	1	0.63
18	4,198	1	0.67		42	4,102	1	0.65
19	4,196	1	0.46		43	4,101	1	0.67
20	4,200	1	0.60		44	4,100	1	0.59
21	4,201	1	0.74		45	4,101	1	0.54
22	4,195	1	0.49		46	4,093	1	0.49
23	4,143	1	0.57		47	4,101	1	0.53
24	4,149	1	0.75		48	4,101	1	0.81

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

Table C2. DC CAS 2012 Operational Form Item Adjusted *P* Values: Mathematics

Mathematics Grade 2								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	suppressed*	1	n/a		28	4,470	1	0.64
2	4,485	1	0.88		29	4,463	1	0.80
3	4,490	1	0.83		30	4,455	1	0.76
4	4,483	1	0.85		31	4,467	1	0.80
5	4,473	1	0.92		32	4,436	1	0.53
6	4,476	2	0.38		33	4,473	1	0.90
7	4,470	1	0.76		34	suppressed	1	n/a
8	4,476	1	0.74		35	4,475	1	0.46
9	suppressed	1	n/a		36	4,461	1	0.87
10	4,467	1	0.74		37	4,474	1	0.91
11	4,474	1	0.50		38	4,447	1	0.68
12	suppressed	1	n/a		39	4,458	1	0.88
13	4,439	1	0.62		40	4,473	1	0.80
14	4,474	1	0.65					
15	4,464	1	0.66					
16	4,472	1	0.64					
17	suppressed	1	n/a					
18	4,472	1	0.72					
19	4,471	1	0.89					
20	suppressed	2	n/a					
21	4,470	1	0.74					
22	suppressed	1	n/a					
23	suppressed	1	n/a					
24	4,426	1	0.67					
25	4,387	1	0.84					
26	4,472	1	0.72					
27	4,461	1	0.77					

***Note*:** The adjusted *p* value for an item includes responses only for examinees with valid responses to that item.
*Items deemed statically unacceptable were suppressed.

Table C2. DC CAS 2012 Operational Form Item Adjusted *P* Values: Mathematics *(continued)*

Mathematics Grade 3							
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**	**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	4,753	1	0.85	28	4,755	1	0.54
2	4,753	1	0.58	29	4,747	1	0.78
3	4,751	1	0.51	30	4,714	1	0.85
4	4,712	1	0.72	31	4,668	1	0.84
5	4,755	1	0.31	32	4,740	1	0.80
6	4,706	3	0.21	33	4,745	1	0.66
7	4,744	1	0.57	34	4,737	1	0.59
8	4,746	1	0.95	35	4,729	1	0.62
9	4,738	1	0.73	36	4,680	1	0.55
10	4,725	1	0.51	37	4,713	1	0.51
11	4,740	1	0.80	38	4,735	1	0.55
12	4,753	1	0.61	39	4,746	1	0.67
13	4,724	1	0.52	40	4,747	1	0.83
14	4,703	1	0.79	41	4,746	1	0.87
15	4,759	1	0.53	42	4,748	1	0.33
16	4,758	1	0.63	43	4,743	1	0.88
17	suppressed*	1	n/a	44	4,737	1	0.70
18	4,754	1	0.78	45	4,730	1	0.69
19	4,748	1	0.65	46	4,724	1	0.42
20	4,745	1	0.83	47	4,728	1	0.73
21	4,574	3	0.80	48	4,711	3	0.32
22	4,726	1	0.75	49	4,727	1	0.86
23	4,739	1	0.87	50	4,716	1	0.52
24	4,748	1	0.77	51	4,522	1	0.67
25	4,648	1	0.33	52	4,698	1	0.64
26	4,755	1	0.84	53	4,726	1	0.61
27	4,726	1	0.55	54	4,736	1	0.83

***Note*:** The adjusted *p* value for an item includes responses only for examinees with valid responses to that item.
*Items deemed statically unacceptable were suppressed.

Table C2. DC CAS 2012 Operational Form Item Adjusted *P* Values: Mathematics *(continued)*

Mathematics Grade 4							
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**	**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	4,585	1	0.69	28	4,578	1	0.90
2	4,584	1	0.71	29	4,576	1	0.60
3	4,590	1	0.73	30	4,573	1	0.91
4	4,576	1	0.65	31	4,569	1	0.43
5	4,549	1	0.72	32	4,570	1	0.62
6	4,545	3	0.43	33	4,566	1	0.49
7	4,578	1	0.55	34	4,571	1	0.70
8	4,580	1	0.69	35	4,569	1	0.92
9	4,576	1	0.79	36	4,568	1	0.82
10	4,578	1	0.58	37	4,566	1	0.73
11	4,574	1	0.66	38	4,558	1	0.75
12	4,574	1	0.74	39	4,554	1	0.84
13	4,573	1	0.52	40	4,539	1	0.58
14	4,572	1	0.70	41	4,570	1	0.39
15	4,576	1	0.59	42	4,572	1	0.41
16	4,575	1	0.81	43	4,567	1	0.41
17	4,577	1	0.71	44	4,567	1	0.91
18	4,570	1	0.72	45	4,566	1	0.28
19	4,570	1	0.48	46	4,559	1	0.75
20	4,539	1	0.52	47	4,497	1	0.63
21	4,534	3	0.20	48	4,550	3	0.72
22	4,571	1	0.51	49	4,568	1	0.73
23	4,568	1	0.51	50	4,568	1	0.32
24	4,565	1	0.55	51	4,567	1	0.81
25	4,571	1	0.85	52	4,563	1	0.64
26	4,571	1	0.86	53	4,567	1	0.36
27	4,548	1	0.49	54	4,553	1	0.61

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

Table C2. DC CAS 2012 Operational Form Item Adjusted *P* Values: Mathematics *(continued)*

Mathematics Grade 5							
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**	**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	4,743	1	0.39	28	4,732	1	0.54
2	4,746	1	0.84	29	4,724	1	0.63
3	4,738	1	0.70	30	4,728	1	0.36
4	4,739	1	0.67	31	4,730	1	0.82
5	4,703	1	0.35	32	4,726	1	0.57
6	4,696	3	0.71	33	4,728	1	0.82
7	4,745	1	0.68	34	4,719	1	0.58
8	4,743	1	0.68	35	4,730	1	0.64
9	4,739	1	0.62	36	4,725	1	0.64
10	4,742	1	0.67	37	4,730	1	0.70
11	4,734	1	0.48	38	4,713	1	0.57
12	4,740	1	0.67	39	4,713	1	0.76
13	4,738	1	0.62	40	4,710	1	0.77
14	4,741	1	0.90	41	4,730	1	0.92
15	4,737	1	0.66	42	4,728	1	0.73
16	4,735	1	0.85	43	4,726	1	0.61
17	4,737	1	0.94	44	4,722	1	0.67
18	suppressed*	1	n/a	45	4,721	1	0.62
19	4,733	1	0.73	46	4,722	1	0.89
20	4,708	1	0.62	47	4,694	1	0.65
21	4,711	3	0.66	48	4,685	3	0.48
22	4,729	1	0.50	49	4,728	1	0.77
23	4,733	1	0.53	50	4,727	1	0.90
24	4,730	1	0.64	51	4,721	1	0.55
25	4,733	1	0.88	52	4,726	1	0.75
26	4,726	1	0.49	53	4,725	1	0.67
27	4,710	1	0.76	54	4,729	1	0.84

***Note*:** The adjusted *p* value for an item includes responses only for examinees with valid responses to that item.
*Items deemed statically unacceptable were suppressed.

Table C2. DC CAS 2012 Operational Form Item Adjusted *P* Values: Mathematics *(continued)*

Mathematics Grade 6							
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**	**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	4,545	1	0.56	28	4,535	1	0.64
2	4,547	1	0.44	29	4,534	1	0.53
3	4,546	1	0.80	30	4,541	1	0.63
4	4,541	1	0.66	31	4,535	1	0.70
5	4,533	1	0.73	32	4,532	1	0.51
6	4,495	3	0.46	33	4,530	1	0.36
7	4,546	1	0.71	34	4,530	1	0.55
8	4,545	1	0.81	35	4,533	1	0.63
9	4,541	1	0.86	36	4,536	1	0.56
10	4,536	1	0.37	37	4,531	1	0.66
11	4,542	1	0.62	38	4,533	1	0.74
12	4,536	1	0.36	39	4,536	1	0.59
13	4,538	1	0.36	40	4,526	1	0.52
14	4,537	1	0.52	41	4,535	1	0.62
15	4,517	1	0.53	42	4,534	1	0.71
16	4,529	1	0.59	43	4,537	1	0.61
17	4,538	1	0.71	44	4,527	1	0.54
18	4,525	1	0.64	45	4,529	1	0.33
19	4,521	1	0.60	46	4,527	1	0.66
20	4,479	1	0.66	47	4,499	1	0.54
21	4,493	3	0.33	48	4,456	3	0.21
22	4,540	1	0.66	49	4,533	1	0.28
23	4,535	1	0.55	50	4,534	1	0.72
24	4,531	1	0.47	51	4,534	1	0.74
25	4,536	1	0.70	52	4,534	1	0.66
26	4,528	1	0.74	53	4,530	1	0.45
27	4,529	1	0.47	54	4,528	1	0.62

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

Table C2. DC CAS 2012 Operational Form Item Adjusted *P* Values: Mathematics *(continued)*

Mathematics Grade 7							
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**	**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	4,274	1	0.54	28	4,273	1	0.64
2	4,288	1	0.63	29	4,269	1	0.71
3	4,294	1	0.86	30	4,270	1	0.64
4	4,232	1	0.43	31	4,278	1	0.82
5	4,286	1	0.48	32	4,270	1	0.56
6	4,190	3	0.20	33	4,275	1	0.69
7	4,287	1	0.23	34	4,279	1	0.71
8	4,287	1	0.51	35	4,274	1	0.55
9	4,291	1	0.48	36	4,267	1	0.46
10	4,284	1	0.58	37	4,277	1	0.65
11	4,290	1	0.51	38	4,273	1	0.63
12	4,288	1	0.73	39	4,264	1	0.61
13	suppressed*	1	.	40	4,248	1	0.58
14	4,289	1	0.35	41	4,267	1	0.57
15	4,277	1	0.52	42	4,256	1	0.43
16	4,273	1	0.82	43	4,261	1	0.65
17	4,281	1	0.79	44	4,265	1	0.67
18	4,280	1	0.77	45	4,268	1	0.69
19	4,270	1	0.61	46	4,261	1	0.62
20	4,231	1	0.49	47	4,250	1	0.36
21	4,171	3	0.50	48	.	3	.
22	4,271	1	0.50	49	4,268	1	0.41
23	4,266	1	0.49	50	4,260	1	0.60
24	4,269	1	0.79	51	4,264	1	0.62
25	4,258	1	0.45	52	4,267	1	0.70
26	4,260	1	0.76	53	4,260	1	0.61
27	4,243	1	0.65	54	4,259	1	0.52

***Note*:** The adjusted *p* value for an item includes responses only for examinees with valid responses to that item.
*Items deemed statically unacceptable were suppressed.

Table C2. DC CAS 2012 Operational Form Item Adjusted *P* Values: Mathematics *(continued)*

Mathematics Grade 8							
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**	**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	4,323	1	0.30	28	4,304	1	0.54
2	4,314	1	0.38	29	4,315	1	0.35
3	4,326	1	0.28	30	4,315	1	0.51
4	4,335	1	0.51	31	4,303	1	0.32
5	4,258	1	0.40	32	4,317	1	0.59
6	4,316	3	0.59	33	4,315	1	0.39
7	4,323	1	0.46	34	4,310	1	0.48
8	4,337	1	0.49	35	4,311	1	0.60
9	4,330	1	0.37	36	4,318	1	0.64
10	4,339	1	0.59	37	4,318	1	0.48
11	4,330	1	0.53	38	4,317	1	0.36
12	4,336	1	0.39	39	4,314	1	0.57
13	4,332	1	0.46	40	4,310	1	0.44
14	4,337	1	0.53	41	4,297	1	0.37
15	4,311	1	0.51	42	4,293	1	0.51
16	4,324	1	0.81	43	4,297	1	0.55
17	4,318	1	0.57	44	4,298	1	0.49
18	4,301	1	0.41	45	4,291	1	0.61
19	4,315	1	0.73	46	4,294	1	0.76
20	4,278	1	0.52	47	4,254	1	0.53
21	4,144	3	0.14	48	4,207	3	0.36
22	4,325	1	0.55	49	4,280	1	0.45
23	4,322	1	0.64	50	4,298	1	0.56
24	4,314	1	0.57	51	4,299	1	0.48
25	4,319	1	0.73	52	4,297	1	0.56
26	4,322	1	0.86	53	4,298	1	0.38
27	4,311	1	0.69	54	4,296	1	0.55

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

Table C2. DC CAS 2012 Operational Form Item Adjusted *P* Values: Mathematics *(continued)*

Mathematics Grade 10								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	3,436	1	0.39		28	3,370	1	0.44
2	3,461	1	0.62		29	3,385	1	0.33
3	3,454	1	0.64		30	3,407	1	0.51
4	3,431	1	0.48		31	3,405	1	0.58
5	3,418	1	0.38		32	3,400	1	0.40
6	3,136	3	0.52		33	3,402	1	0.57
7	3,431	1	0.40		34	3,403	1	0.38
8	3,445	1	0.40		35	3,365	1	0.36
9	3,457	1	0.83		36	3,401	1	0.58
10	3,440	1	0.49		37	3,400	1	0.47
11	3,442	1	0.51		38	3,395	1	0.46
12	3,404	1	0.24		39	3,397	1	0.78
13	3,455	1	0.72		40	3,396	1	0.53
14	3,447	1	0.77		41	3,363	1	0.53
15	3,437	1	0.72		42	3,378	1	0.28
16	3,435	1	0.63		43	3,395	1	0.68
17	3,425	1	0.47		44	3,374	1	0.48
18	3,423	1	0.49		45	3,387	1	0.40
19	3,419	1	0.42		46	3,364	1	0.35
20	3,383	1	0.44		47	3,350	1	0.46
21	3,218	3	0.20		48	3,040	3	0.20
22	3,428	1	0.53		49	3,369	1	0.30
23	3,431	1	0.45		50	3,391	1	0.43
24	3,424	1	0.63		51	3,393	1	0.56
25	3,424	1	0.45		52	3,380	1	0.33
26	3,418	1	0.38		53	3,381	1	0.56
27	3,418	1	0.62		54	3,382	1	0.30

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

Table C3. DC CAS 2012 Operational Form Item Adjusted *P* Values: 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	4,694	1	0.45		26	4,673	1	0.44
2	4,696	1	0.56		27	4,659	1	0.68
3	4,697	1	0.81		28	4,670	1	0.43
4	4,696	1	0.46		29	4,672	1	0.66
5	4,689	1	0.36		30	4,661	1	0.34
6	4,694	1	0.70		31	4,659	1	0.47
7	4,692	1	0.63		32	4,655	1	0.42
8	4,686	1	0.46		33	4,656	1	0.60
9	4,667	1	0.34		34	4,652	1	0.32
10	4,565	2	0.20		35	4,656	1	0.45
11	4,684	1	0.62		36	4,648	1	0.44
12	4,674	1	0.39		37	4,648	1	0.46
13	4,660	1	0.26		38	4,641	1	0.50
14	4,666	1	0.36		39	4,567	2	0.23
15	4,661	1	0.53		40	4,660	1	0.53
16	4,657	1	0.30		41	4,658	1	0.77
17	4,654	1	0.26		42	4,652	1	0.42
18	4,684	1	0.57		43	4,654	1	0.29
19	4,680	1	0.37		44	4,655	1	0.54
20	4,666	1	0.36		45	4,652	1	0.68
21	4,594	2	0.71		46	4,657	1	0.37
22	4,678	1	0.50		47	4,655	1	0.65
23	4,679	1	0.62		48	4,651	1	0.37
24	4,677	1	0.27		49	4,651	1	0.62
25	4,674	1	0.42		50	4,642	1	0.40

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

Table C3. DC CAS 2012 Operational Form Item Adjusted *P* Values: Science/Biology *(continued)*

Science Grade 8							
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**	**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	4,248	1	0.57	26	4,222	1	0.29
2	4,247	1	0.38	27	4,227	1	0.44
3	4,229	1	0.32	28	4,230	1	0.28
4	4,241	1	0.42	29	4,228	1	0.38
5	4,243	1	0.59	30	4,218	1	0.37
6	4,241	1	0.31	31	4,220	1	0.34
7	4,232	1	0.41	32	4,206	1	0.47
8	4,236	1	0.50	33	4,212	1	0.44
9	4,228	1	0.36	34	4,204	1	0.44
10	3,532	2	0.15	35	4,208	1	0.39
11	4,233	1	0.44	36	4,201	1	0.46
12	4,228	1	0.49	37	4,204	1	0.34
13	4,232	1	0.55	38	4,182	1	0.37
14	4,223	1	0.33	39	3,936	2	0.29
15	4,229	1	0.46	40	4,203	1	0.37
16	4,225	1	0.36	41	4,206	1	0.51
17	4,218	1	0.47	42	4,201	1	0.25
18	4,227	1	0.42	43	4,209	1	0.44
19	4,215	1	0.28	44	4,205	1	0.51
20	4,211	1	0.35	45	4,201	1	0.42
21	4,064	2	0.51	46	4,207	1	0.31
22	4,234	1	0.20	47	4,208	1	0.47
23	4,226	1	0.26	48	4,206	1	0.54
24	4,231	1	0.66	49	4,201	1	0.35
25	4,231	1	0.56	50	4,208	1	0.53

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

Table C3. DC CAS 2012 Operational Form Item Adjusted *P* Values: Science/Biology *(continued)*

High School Biology								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	3,670	1	0.23		26	3,660	1	0.33
2	3,684	1	0.37		27	3,648	1	0.39
3	3,689	1	0.35		28	3,661	1	0.62
4	3,689	1	0.49		29	3,658	1	0.34
5	3,680	1	0.30		30	3,640	1	0.44
6	3,689	1	0.52		31	3,637	1	0.39
7	3,686	1	0.41		32	3,634	1	0.43
8	3,682	1	0.57		33	3,633	1	0.56
9	3,663	1	0.28		34	3,608	1	0.48
10	3,480	2	0.74		35	3,608	1	0.47
11	3,676	1	0.25		36	3,611	1	0.36
12	3,672	1	0.27		37	3,602	1	0.28
13	3,672	1	0.52		38	3,570	1	0.41
14	3,657	1	0.70		39	3,279	2	0.28
15	3,658	1	0.57		40	3,602	1	0.27
16	3,659	1	0.27		41	3,613	1	0.29
17	3,653	1	0.60		42	3,614	1	0.45
18	3,657	1	0.34		43	3,612	1	0.38
19	3,662	1	0.26		44	3,605	1	0.39
20	3,666	1	0.42		45	3,609	1	0.54
21	3,334	2	0.27		46	3,605	1	0.33
22	3,666	1	0.38		47	3,605	1	0.39
23	3,668	1	0.38		48	3,608	1	0.40
24	3,667	1	0.49		49	3,608	1	0.48
25	3,662	1	0.51		50	3,607	1	0.33

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

Table C4. DC CAS 2012 Operational Form Item Adjusted *P* Values: Composition

Composition Grade 4								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	1,142	6	0.33		5	1,104	6	0.44
2	1,142	4	0.50		6	1,104	4	0.60
3	1,111	6	0.40		7	1,063	6	0.42
4	1,111	4	0.55		8	1,063	4	0.59
Composition Grade 7								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	1,037	6	0.46		5	1,030	6	0.48
2	1,037	4	0.64		6	1,030	4	0.64
3	1,036	6	0.51		7	991	6	0.42
4	1,036	4	0.68		8	991	4	0.61
Composition Grade 10								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	814	6	0.51		5	795	6	0.33
2	814	4	0.70		6	795	4	0.61
3	793	6	0.38		7	822	6	0.43
4	793	4	0.63		8	822	4	0.64

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

Table C5. DC CAS 2012 Field Test Form Item Adjusted *P* Values: Reading

Reading Grade 2								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	2,342	1	0.51		25	2,032	1	0.49
2	2,341	1	0.42		26	2,076	1	0.63
3	2,340	1	0.45		27	2,071	1	0.43
4	2,324	1	0.53		28	2,061	1	0.31
5	2,343	1	0.64		29	2,058	1	0.35
6	2,332	1	0.50		30	2,046	1	0.49
7	2,313	1	0.60		31	2,053	1	0.45
8	2,332	1	0.53		32	2,094	1	0.56
9	2,327	1	0.46		33	2,089	1	0.23
10	2,328	1	0.64		34	2,085	1	0.47
11	2,347	1	0.24		35	2,087	1	0.31
12	2,347	1	0.45		36	2,072	1	0.46
13	2,349	1	0.25		37	2,070	1	0.65
14	2,337	1	0.59		38	2,043	1	0.50
15	2,320	1	0.22		39	2,080	1	0.50
16	2,349	1	0.48		40	2,068	1	0.51
17	2,341	1	0.40		41	2,064	1	0.25
18	2,331	1	0.42		42	1,951	3	0.17
19	2,349	1	0.61					
20	2,339	1	0.49					
21	2,236	3	0.41					
22	2,077	1	0.67					
23	2,054	1	0.24					
24	2,062	1	0.32					

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

Table C5. DC CAS 2012 Field Test Form Item Adjusted *P* Values: Reading *(continued)*

Reading Grade 3								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	2,392	1	0.73		25	2,323	1	0.56
2	2,393	1	0.30		26	2,311	1	0.54
3	2,389	1	0.56		27	2,320	1	0.31
4	2,384	1	0.41		28	2,323	1	0.28
5	2,362	1	0.23		29	2,320	1	0.50
6	2,367	1	0.30		30	2,260	3	0.30
7	2,367	1	0.27		31	2,323	1	0.85
8	2,377	1	0.30		32	2,321	1	0.76
9	2,383	1	0.37		33	2,324	1	0.54
10	2,382	1	0.55		34	2,324	1	0.72
11	2,287	3	0.33		35	2,302	1	0.71
12	2,394	1	0.83		36	2,267	1	0.39
13	2,354	1	0.86		37	2,316	1	0.72
14	2,392	1	0.74		38	2,266	3	0.31
15	2,374	1	0.47					
16	2,380	1	0.32					
17	2,364	1	0.64					
18	2,372	1	0.70					
19	2,318	3	0.36					
20	2,326	1	0.19					
21	2,324	1	0.62					
22	2,320	1	0.69					
23	2,317	1	0.48					
24	2,313	1	0.40					

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

Table C5. DC CAS 2012 Field Test Form Item Adjusted *P* Values: Reading *(continued)*

Reading Grade 4							
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**	**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	2,306	1	0.39	25	2,244	1	0.60
2	2,303	1	0.23	26	2,243	1	0.34
3	2,306	1	0.83	27	2,238	1	0.69
4	2,302	1	0.42	28	2,234	1	0.47
5	2,300	1	0.34	29	2,220	1	0.49
6	2,304	1	0.41	30	2,165	3	0.34
7	2,302	1	0.59	31	2,240	1	0.19
8	2,298	1	0.74	32	2,240	1	0.48
9	2,299	1	0.46	33	2,240	1	0.36
10	2,282	1	0.44	34	2,236	1	0.55
11	2,231	3	0.40	35	2,237	1	0.48
12	2,309	1	0.66	36	2,232	1	0.58
13	2,308	1	0.41	37	2,221	1	0.44
14	2,306	1	0.43	38	2,191	3	0.37
15	2,304	1	0.61				
16	2,305	1	0.37				
17	2,299	1	0.25				
18	2,291	1	0.45				
19	2,258	3	0.34				
20	2,242	1	0.45				
21	2,244	1	0.49				
22	2,244	1	0.36				
23	2,244	1	0.42				
24	2,243	1	0.37				

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

Table C5. DC CAS 2012 Field Test Form Item Adjusted *P* Values: Reading *(continued)*

Reading Grade 5								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	2,383	1	0.46		25	2,333	1	0.49
2	2,382	1	0.58		26	2,321	1	0.32
3	2,382	1	0.42		27	2,202	3	0.12
4	2,383	1	0.45		28	2,334	1	0.67
5	2,381	1	0.49		29	2,335	1	0.64
6	2,378	1	0.46		30	2,334	1	0.70
7	2,370	1	0.55		31	2,334	1	0.56
8	2,278	3	0.21		32	2,333	1	0.46
9	2,388	1	0.66		33	2,333	1	0.56
10	2,388	1	0.50		34	2,335	1	0.62
11	2,388	1	0.43		35	2,332	1	0.58
12	2,385	1	0.45		36	2,329	1	0.49
13	2,388	1	0.48		37	2,324	1	0.78
14	2,388	1	0.64		38	2,243	3	0.25
15	2,386	1	0.65					
16	2,387	1	0.35					
17	2,383	1	0.52					
18	2,369	1	0.47					
19	2,324	3	0.36					
20	2,331	1	0.26					
21	2,335	1	0.32					
22	2,333	1	0.30					
23	2,332	1	0.37					
24	2,332	1	0.44					

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

Table C5. DC CAS 2012 Field Test Form Item Adjusted *P* Values: Reading *(continued)*

Reading Grade 6								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	2,261	1	0.67		25	2,257	1	0.57
2	2,261	1	0.83		26	2,247	1	0.67
3	2,258	1	0.60		27	2,195	3	0.37
4	2,259	1	0.60		28	2,254	1	0.50
5	2,261	1	0.74		29	2,257	1	0.63
6	2,258	1	0.24		30	2,256	1	0.52
7	2,248	1	0.78		31	2,255	1	0.74
8	2,191	3	0.22		32	2,258	1	0.76
9	2,260	1	0.44		33	2,254	1	0.56
10	2,258	1	0.43		34	2,254	1	0.64
11	2,260	1	0.54		35	2,255	1	0.39
12	2,258	1	0.47		36	2,250	1	0.43
13	2,259	1	0.48		37	2,251	1	0.51
14	2,255	1	0.33		38	2,218	3	0.15
15	2,257	1	0.47					
16	2,251	1	0.54					
17	2,249	1	0.39					
18	2,238	1	0.50					
19	2,180	3	0.18					
20	2,259	1	0.68					
21	2,257	1	0.51					
22	2,255	1	0.52					
23	2,255	1	0.57					
24	2,257	1	0.64					

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

Table C5. DC CAS 2012 Field Test Form Item Adjusted *P* Values: Reading *(continued)*

Reading Grade 7								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	2,169	1	0.54		25	2,098	1	0.46
2	2,171	1	0.67		26	2,092	1	0.49
3	2,166	1	0.46		27	2,015	3	0.23
4	2,167	1	0.53		28	2,098	1	0.61
5	2,166	1	0.69		29	2,100	1	0.50
6	2,166	1	0.43		30	2,096	1	0.50
7	2,156	1	0.36		31	2,099	1	0.79
8	2,119	3	0.31		32	2,095	1	0.50
9	2,158	1	0.49		33	2,096	1	0.47
10	2,158	1	0.51		34	2,097	1	0.46
11	2,154	1	0.33		35	2,099	1	0.41
12	2,157	1	0.36		36	2,097	1	0.56
13	2,159	1	0.38		37	2,085	1	0.67
14	2,158	1	0.57		38	2,068	3	0.41
15	2,159	1	0.32					
16	2,158	1	0.17					
17	2,158	1	0.32					
18	2,156	1	0.21					
19	2,071	3	0.14					
20	2,098	1	0.31					
21	2,099	1	0.26					
22	2,097	1	0.32					
23	2,100	1	0.39					
24	2,102	1	0.13					

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

Table C5. DC CAS 2012 Field Test Form Item Adjusted *P* Values: Reading *(continued)*

Reading Grade 8								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	2,178	1	0.50		25	2,106	1	0.40
2	2,180	1	0.58		26	2,096	1	0.49
3	2,177	1	0.46		27	2,108	1	0.31
4	2,174	1	0.58		28	2,105	1	0.50
5	2,178	1	0.42		29	2,099	1	0.65
6	2,177	1	0.37		30	2,050	3	0.37
7	2,176	1	0.35		31	2,117	1	0.80
8	2,175	1	0.45		32	2,117	1	0.62
9	2,170	1	0.58		33	2,112	1	0.47
10	2,093	3	0.39		34	2,115	1	0.67
11	2,178	1	0.35		35	2,116	1	0.61
12	2,175	1	0.30		36	2,116	1	0.73
13	2,175	1	0.58		37	2,109	1	0.53
14	2,174	1	0.66		38	2,115	1	0.68
15	2,177	1	0.52		39	2,107	1	0.51
16	2,174	1	0.54		40	2,075	3	0.17
17	2,176	1	0.24					
18	2,175	1	0.46					
19	2,162	1	0.48					
20	2,108	3	0.34					
21	2,115	1	0.29					
22	2,111	1	0.43					
23	2,114	1	0.40					
24	2,106	1	0.36					

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

Table C5. DC CAS 2012 Field Test Form Item Adjusted *P* Values: Reading *(continued)*

Reading Grade 9								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	1,748	1	0.67		25	1,717	1	0.46
2	1,762	1	0.54		26	1,717	1	0.46
3	1,761	1	0.53		27	1,719	1	0.58
4	1,763	1	0.41		28	1,716	1	0.49
5	1,761	1	0.46		29	1,715	1	0.55
6	1,761	1	0.28		30	1,709	1	0.49
7	1,760	1	0.35		31	1,664	1	0.34
8	1,760	1	0.43		32	1,666	1	0.81
9	1,756	1	0.60		33	1,662	1	0.58
10	1,586	3	0.23		34	1,662	1	0.69
11	1,714	1	0.73		35	1,661	1	0.60
12	1,712	1	0.48		36	1,661	1	0.31
13	1,711	1	0.41		37	1,662	1	0.28
14	1,713	1	0.65		38	1,660	1	0.62
15	1,712	1	0.67		39	1,642	1	0.18
16	1,706	1	0.36		40	1,489	3	0.20
17	1,713	1	0.63					
18	1,712	1	0.28					
19	1,712	1	0.58					
20	1,710	1	0.50					
21	1,703	1	0.50					
22	1,716	1	0.18					
23	1,716	1	0.42					
24	1,720	1	0.59					

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

Table C5. DC CAS 2012 Field Test Form Item Adjusted *P* Values: Reading *(continued)*

Reading Grade 10								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	2,114	1	0.31		25	2,084	1	0.68
2	2,109	1	0.54		26	2,084	1	0.64
3	2,112	1	0.60		27	2,080	1	0.56
4	2,109	1	0.49		28	2,082	1	0.49
5	2,108	1	0.57		29	2,078	1	0.27
6	2,106	1	0.57		30	1,941	3	0.47
7	2,106	1	0.31		31	2,049	1	0.65
8	2,105	1	0.57		32	2,044	1	0.36
9	2,099	1	0.63		33	2,047	1	0.52
10	1,927	3	0.31		34	2,040	1	0.71
11	2,072	1	0.63		35	2,047	1	0.46
12	2,069	1	0.41		36	2,042	1	0.56
13	2,072	1	0.47		37	2,047	1	0.42
14	2,068	1	0.21		38	2,045	1	0.44
15	2,073	1	0.65		39	2,036	1	0.55
16	2,063	1	0.46		40	1,893	3	0.42
17	2,063	1	0.58					
18	2,062	1	0.55					
19	2,057	1	0.42					
20	1,857	3	0.33					
21	2,082	1	0.60					
22	2,081	1	0.70					
23	2,080	1	0.44					
24	2,079	1	0.58					

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

Table C6. DC CAS 2012 Field Test Form Item Adjusted *P* Values: Mathematics

Mathematics Grade 2								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	2,364	1	0.73		19	2,071	1	0.77
2	2,366	1	0.62		20	2,101	1	0.87
3	2,361	1	0.86		21	2,109	1	0.80
4	2,370	1	0.57		22	2,109	1	0.69
5	2,354	1	0.78		23	2,106	1	0.84
6	2,367	1	0.97		24	2,100	1	0.63
7	2,365	1	0.78		25	2,096	1	0.75
8	2,368	1	0.55		26	2,093	1	0.68
9	2,370	1	0.82		27	2,093	1	0.63
10	2,362	1	0.52		28	2,099	1	0.84
11	2,360	1	0.53		29	2,095	1	0.67
12	2,367	1	0.69		30	2,093	1	0.55
13	2,351	3	0.43		31	2,089	3	0.60
14	2,360	1	0.81		32	2,104	1	0.45
15	2,369	1	0.91		33	2,100	1	0.93
16	2,371	1	0.66		34	2,097	1	0.48
17	2,353	1	0.50		35	2,105	1	0.52
18	2,361	1	0.60		36	2,101	1	0.79

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

Table C6. DC CAS 2012 Field Test Form Item Adjusted *P* Values: Mathematics *(continued)*

Mathematics Grade 3								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	2,407	1	0.73		17	2,339	1	0.44
2	2,406	1	0.61		18	2,281	1	0.95
3	2,398	1	0.78		19	2,333	1	0.77
4	2,398	1	0.60		20	2,320	1	0.55
5	2,335	3	0.82		21	2,321	3	0.37
6	2,403	1	0.42		22	2,340	1	0.62
7	2,385	1	0.58		23	2,321	1	0.64
8	2,401	1	0.87		24	2,325	1	0.82
9	2,343	3	0.58		25	2,276	3	0.42
10	2,395	1	0.50		26	2,340	1	0.71
11	2,401	1	0.55		27	2,333	1	0.83
12	2,390	1	0.73		28	2,329	1	0.79
13	2,390	1	0.20		29	2,315	1	0.56
14	2,395	1	0.58		30	2,334	1	0.42
15	2,394	1	0.68		31	2,341	1	0.96
16	2,366	1	0.59		32	2,324	1	0.35

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

Table C6. DC CAS 2012 Field Test Form Item Adjusted *P* Values: Mathematics *(continued)*

Mathematics Grade 4								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	2,327	1	0.62		17	2,237	1	0.34
2	2,327	1	0.74		18	2,239	1	0.52
3	2,321	1	0.73		19	2,232	1	0.46
4	2,308	1	0.74		20	2,220	1	0.47
5	2,290	3	0.33		21	2,177	3	0.46
6	2,295	1	0.32		22	2,209	1	0.44
7	2,294	1	0.45		23	2,211	1	0.79
8	2,277	1	0.45		24	2,206	1	0.25
9	2,287	3	0.32		25	2,205	3	0.30
10	2,318	1	0.45		26	2,232	1	0.79
11	2,315	1	0.68		27	2,231	1	0.45
12	2,302	1	0.44		28	2,218	1	0.62
13	2,319	1	0.69		29	2,246	1	0.66
14	2,317	1	0.47		30	2,240	1	0.76
15	2,319	1	0.68		31	2,240	1	0.86
16	2,302	1	0.47		32	2,217	1	0.28

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

Table C6. DC CAS 2012 Field Test Form Item Adjusted *P* Values: Mathematics *(continued)*

Mathematics Grade 5								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	2,390	1	0.42		17	2,345	1	0.27
2	2,395	1	0.35		18	2,346	1	0.27
3	2,393	1	0.46		19	2,345	1	0.72
4	2,374	1	0.33		20	2,334	1	0.31
5	2,304	3	0.40		21	2,302	3	0.68
6	2,372	1	0.49		22	2,328	1	0.31
7	2,369	1	0.22		23	2,319	1	0.32
8	2,362	1	0.43		24	2,310	1	0.69
9	2,351	3	0.10		25	2,307	3	0.22
10	2,379	1	0.38		26	2,332	1	0.76
11	2,382	1	0.55		27	2,329	1	0.56
12	2,372	1	0.80		28	2,319	1	0.46
13	2,387	1	0.50		29	2,333	1	0.61
14	2,388	1	0.21		30	2,335	1	0.68
15	2,384	1	0.34		31	2,332	1	0.35
16	2,380	1	0.47		32	2,325	1	0.45

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

Table C6. DC CAS 2012 Field Test Form Item Adjusted *P* Values: Mathematics *(continued)*

Mathematics Grade 6								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	2,282	1	0.80		17	2,256	1	0.68
2	2,282	1	0.41		18	2,254	1	0.49
3	2,279	1	0.30		19	2,253	1	0.40
4	2,280	1	0.81		20	2,251	1	0.74
5	2,206	3	0.13		21	2,211	3	0.23
6	2,276	1	0.43		22	2,251	1	0.79
7	2,273	1	0.39		23	2,249	1	0.34
8	2,267	1	0.47		24	2,251	1	0.70
9	2,242	3	0.36		25	2,228	3	0.15
10	2,276	1	0.61		26	2,243	1	0.40
11	2,271	1	0.53		27	2,248	1	0.72
12	2,265	1	0.59		28	2,243	1	0.42
13	2,276	1	0.44		29	2,254	1	0.43
14	2,276	1	0.20		30	2,254	1	0.45
15	2,278	1	0.19		31	2,256	1	0.30
16	2,274	1	0.17		32	2,251	1	0.25

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

Table C6. DC CAS 2012 Field Test Form Item Adjusted *P* Values: Mathematics *(continued)*

Mathematics Grade 7								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	2,182	1	0.39		17	2,106	1	0.43
2	2,179	1	0.36		18	2,107	1	0.36
3	2,174	1	0.33		19	2,102	1	0.35
4	2,171	1	0.41		20	2,105	1	0.32
5	2,096	3	0.14		21	2,053	3	0.28
6	2,151	1	0.57		22	2,064	1	0.54
7	2,154	1	0.59		23	2,064	1	0.62
8	2,152	1	0.50		24	2,062	1	0.56
9	2,136	3	0.32		25	2,046	3	0.37
10	2,170	1	0.51		26	2,073	1	0.27
11	2,169	1	0.42		27	2,071	1	0.44
12	2,171	1	0.47		28	2,073	1	0.47
13	2,165	1	0.20		29	2,097	1	0.34
14	2,166	1	0.46		30	2,096	1	0.52
15	2,165	1	0.33		31	2,096	1	0.23
16	2,165	1	0.49		32	2,091	1	0.52

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

Table C6. DC CAS 2012 Field Test Form Item Adjusted *P* Values: Mathematics *(continued)*

Mathematics Grade 8								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	2,196	1	0.51		17	2,134	1	0.71
2	2,196	1	0.37		18	2,133	1	0.53
3	2,199	1	0.17		19	2,127	1	0.29
4	2,199	1	0.41		20	2,131	1	0.62
5	2,072	3	0.25		21	1,935	3	0.10
6	2,183	1	0.38		22	2,121	1	0.50
7	2,186	1	0.66		23	2,115	1	0.41
8	2,188	1	0.37		24	2,120	1	0.35
9	2,077	3	0.10		25	2,061	3	0.08
10	2,188	1	0.61		26	2,123	1	0.34
11	2,191	1	0.44		27	2,122	1	0.33
12	2,188	1	0.45		28	2,122	1	0.32
13	2,180	1	0.46		29	2,116	1	0.36
14	2,178	1	0.27		30	2,117	1	0.30
15	2,182	1	0.34		31	2,116	1	0.28
16	2,173	1	0.37		32	2,115	1	0.46

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

Table C6. DC CAS 2012 Field Test Form Item Adjusted *P* Values: Mathematics *(continued)*

Mathematics Grade 10								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	1,741	1	0.58		17	1,702	1	0.37
2	1,739	1	0.28		18	1,701	1	0.38
3	1,733	1	0.31		19	1,697	1	0.42
4	1,737	1	0.35		20	1,704	1	0.50
5	1,317	3	0.20		21	1,554	3	0.12
6	1,725	1	0.30		22	1,689	1	0.21
7	1,715	1	0.30		23	1,689	1	0.14
8	1,724	1	0.66		24	1,691	1	0.37
9	1,485	3	0.07		25	1,432	3	0.18
10	1,713	1	0.23		26	1,676	1	0.49
11	1,712	1	0.55		27	1,680	1	0.51
12	1,711	1	0.49		28	1,680	1	0.41
13	1,703	1	0.41		29	1,679	1	0.50
14	1,706	1	0.31		30	1,670	1	0.35
15	1,706	1	0.57		31	1,677	1	0.40
16	1,707	1	0.48		32	1,675	1	0.33

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

Table C7. DC CAS 2012 Field Test Form Item Adjusted *P* Values: 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	2,395	1	0.71		15	2,301	1	0.59
2	2,394	1	0.42		16	2,300	1	0.43
3	2,393	1	0.48		17	2,299	1	0.54
4	2,297	2	0.39		18	2,207	2	0.45
5	2,380	1	0.24		19	2,292	1	0.43
6	2,379	1	0.41		20	2,291	1	0.48
7	2,383	1	0.23		21	2,295	1	0.51
8	2,382	1	0.31		22	2,295	1	0.62
9	2,376	1	0.45		23	2,292	1	0.40
10	2,376	1	0.61		24	2,287	1	0.36
11	2,357	1	0.54		25	2,280	1	0.70
12	2,288	2	0.18		26	2,224	2	0.34
13	2,365	1	0.69		27	2,280	1	0.48
14	2,371	1	0.67		28	2,284	1	0.40

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

Table C7. DC CAS 2012 Field Test Form Item Adjusted *P* Values: Science/Biology *(continued)*

Science Grade 8								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	2,148	1	0.31		15	2,092	1	0.57
2	2,153	1	0.43		16	2,095	1	0.50
3	2,152	1	0.43		17	2,093	1	0.54
4	1,800	2	0.12		18	1,926	2	0.03
5	2,143	1	0.38		19	2,085	1	0.33
6	2,145	1	0.41		20	2,090	1	0.30
7	2,142	1	0.55		21	2,091	1	0.50
8	2,141	1	0.28		22	2,092	1	0.43
9	2,142	1	0.27		23	2,090	1	0.33
10	2,134	1	0.27		24	2,086	1	0.33
11	2,120	1	0.13		25	2,077	1	0.31
12	1,965	2	0.35		26	1,874	2	0.12
13	2,124	1	0.39		27	2,081	1	0.38
14	2,127	1	0.34		28	2,078	1	0.34

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

Table C7. DC CAS 2012 Field Test Form Item Adjusted *P* Values: Science/Biology *(continued)*

High School Biology								
Operational Item Sequence Number	**N**	**Max Points**	**Adjusted *P* Value**		**Operational Item Sequence Number**	**N**	**Max Points**	**Adjusted *P* Value**
1	1,860	1	0.32		15	1,818	1	0.54
2	1,864	1	0.35		16	1,817	1	0.30
3	1,867	1	0.50		17	1,818	1	0.48
4	1,497	2	0.08		18	1,653	2	0.20
5	1,859	1	0.42		19	1,809	1	0.40
6	1,856	1	0.36		20	1,803	1	0.26
7	1,857	1	0.30		21	1,807	1	0.32
8	1,854	1	0.28		22	1,804	1	0.34
9	1,857	1	0.37		23	1,808	1	0.44
10	1,851	1	0.45		24	1,805	1	0.57
11	1,848	1	0.29		25	1,799	1	0.41
12	1,724	2	0.49		26	1,546	2	0.34
13	1,822	1	0.35		27	1,775	1	0.24
14	1,823	1	0.17		28	1,781	1	0.15

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

Appendix D: Internal Consistency Reliability Coefficients for Examinee Subgroups

(See Section 8. Evidence for Reliability and Validity, *Internal Consistency Reliability*, Table 31)

Table D1. Internal Consistency Reliability Coefficients for Examinee Subgroups: Reading

Grade	Subgroup	N	Alpha	Stratified Alpha	Feldt-Raju	Mean	SD
2	All Examinees	4,469	0.88	0.88	0.88	241.97	15.78
	Male	2,260	0.88	0.89	0.89	240.05	16.11
	Female	2,186	0.87	0.87	0.87	244.00	15.15
	Asian	95	0.85	0.86	0.86	252.17	13.76
	African	3,195	0.86	0.87	0.87	239.24	14.70
	Hispanic	625	0.86	0.87	0.87	240.84	14.47
	White	521	0.80	0.81	0.82	258.07	13.44
3	All Examinees	4,737	0.93	0.94	0.94	348.65	15.37
	Male	2,390	0.93	0.94	0.94	346.78	15.61
	Female	2,329	0.93	0.93	0.93	350.58	14.87
	Asian	94	0.90	0.90	0.91	359.27	11.91
	African	3,459	0.92	0.93	0.93	346.05	14.62
	Hispanic	664	0.92	0.92	0.92	349.03	14.04
	White	479	0.90	0.90	0.90	364.76	11.98
4	All Examinees	4,559	0.92	0.92	0.92	452.42	15.09
	Male	2,299	0.92	0.93	0.93	450.83	15.85
	Female	2,241	0.91	0.91	0.91	454.13	14.03
	Asian	102	0.92	0.93	0.93	460.84	16.16
	African	3,330	0.91	0.91	0.91	449.79	14.12
	Hispanic	629	0.90	0.90	0.90	452.58	13.59
	White	461	0.87	0.87	0.87	469.20	12.02
5	All Examinees	4,734	0.92	0.92	0.92	553.75	15.09
	Male	2,395	0.92	0.92	0.92	551.45	15.65
	Female	2,324	0.91	0.91	0.91	556.22	13.99
	Asian	78	0.88	0.89	0.89	565.37	12.41
	African	3,686	0.91	0.91	0.91	551.66	14.44
	Hispanic	578	0.91	0.91	0.91	554.59	14.31
	White	365	0.81	0.82	0.82	571.01	9.96

Table D1. Internal Consistency Reliability Coefficients for Examinee Subgroups: Reading *(continued)*

Grade	Subgroup	N	Alpha	Stratified Alpha	Feldt-Raju	Mean	SD
6	All Examinees	4,539	0.91	0.92	0.92	650.16	14.20
	Male	2,293	0.92	0.92	0.92	648.28	14.91
	Female	2,220	0.91	0.91	0.91	652.15	13.08
	Asian	68	0.91	0.92	0.92	659.51	13.95
	African American	3,590	0.91	0.91	0.91	648.70	13.46
	Hispanic	566	0.91	0.91	0.91	650.62	14.06
	White	268	0.91	0.92	0.92	666.40	13.03
7	All Examinees	4,283	0.90	0.91	0.90	754.13	14.25
	Male	2,150	0.91	0.91	0.91	751.95	14.93
	Female	2,118	0.89	0.89	0.89	756.37	13.10
	Asian	55	0.88	0.89	0.89	762.96	11.12
	African American	3,442	0.89	0.90	0.90	752.69	13.97
	Hispanic	507	0.89	0.90	0.90	755.59	12.67
	White	240	0.90	0.91	0.91	768.72	12.97
8	All Examinees	4,337	0.90	0.91	0.91	853.86	14.32
	Male	2,154	0.90	0.91	0.91	851.40	14.96
	Female	2,159	0.89	0.90	0.90	856.42	13.12
	Asian	54	0.90	0.91	0.91	862.48	12.81
	African American	3,526	0.89	0.90	0.90	852.64	13.82
	Hispanic	475	0.89	0.89	0.89	854.04	13.65
	White	235	0.89	0.90	0.90	869.98	13.03
9	All Examinees	3,534	0.92	0.93	0.93	947.17	16.94
	Male	1,701	0.92	0.92	0.92	944.95	16.72
	Female	1,778	0.93	0.93	0.93	949.64	16.72
	Asian	28	0.94	0.94	0.95	958.18	16.48
	African American	2,940	0.92	0.92	0.92	946.91	16.14
	Hispanic	388	0.94	0.94	0.94	944.50	19.69
	White	96	0.92	0.92	0.93	969.07	11.92
10	All Examinees	4,230	0.92	0.92	0.92	951.32	15.45
	Male	2,014	0.92	0.92	0.92	949.08	15.95
	Female	2,170	0.91	0.92	0.92	953.56	14.57
	Asian	64	0.91	0.91	0.91	961.02	12.91
	African American	3,518	0.91	0.92	0.92	950.26	15.12
	Hispanic	444	0.91	0.91	0.91	952.63	14.11
	White	153	0.92	0.93	0.93	969.75	13.18

Table D2. Internal Consistency Reliability Coefficients for Examinee Subgroups: Mathematics

Grade	Subgroup	N	Alpha	Stratified Alpha	Feldt-Raju	Mean	SD
2	All Examinees	4,499	0.89	0.89	0.90	253.94	15.27
	Male	2,276	0.90	0.90	0.90	253.85	15.82
	Female	2,198	0.88	0.88	0.89	254.05	14.67
	Asian	100	0.83	0.82	0.85	265.78	15.62
	African	3,209	0.88	0.88	0.88	251.10	13.73
	Hispanic	632	0.88	0.88	0.89	253.41	14.10
	White	523	0.80	0.81	0.81	269.50	14.85
3	All Examinees	4,771	0.93	0.94	0.94	352.29	17.70
	Male	2,413	0.93	0.94	0.94	351.76	17.78
	Female	2,339	0.93	0.94	0.94	352.92	17.55
	Asian	97	0.89	0.89	0.90	368.14	13.88
	African	3,477	0.93	0.93	0.93	348.98	16.98
	Hispanic	674	0.92	0.92	0.92	354.33	15.13
	White	482	0.89	0.90	0.90	370.36	13.38
4	All Examinees	4,590	0.93	0.93	0.93	456.65	15.75
	Male	2,314	0.93	0.93	0.94	455.55	16.66
	Female	2,258	0.92	0.93	0.93	457.87	14.62
	Asian	103	0.92	0.92	0.93	469.13	13.11
	African	3,349	0.92	0.92	0.92	453.72	15.04
	Hispanic	638	0.92	0.92	0.92	458.47	14.00
	White	463	0.89	0.89	0.89	472.59	11.67
5	All Examinees	4,747	0.93	0.93	0.93	557.66	16.67
	Male	2,409	0.93	0.93	0.93	556.10	16.94
	Female	2,322	0.93	0.93	0.93	559.41	16.12
	Asian	81	0.93	0.93	0.93	572.89	15.89
	African	3,681	0.92	0.93	0.93	555.72	16.35
	Hispanic	586	0.92	0.93	0.93	558.93	15.55
	White	368	0.87	0.87	0.87	572.14	11.77

Table D2. Internal Consistency Reliability Coefficients for Examinee Subgroups: Mathematics *(continued)*

Grade	Subgroup	N	Alpha	Stratified Alpha	Feldt-Raju	Mean	SD
6	All Examinees	4,551	0.93	0.94	0.94	651.21	17.11
	Male	2,295	0.93	0.94	0.94	650.10	17.44
	Female	2,229	0.93	0.93	0.93	652.43	16.63
	Asian	69	0.94	0.95	0.95	669.06	17.84
	African American	3,594	0.92	0.93	0.93	649.04	16.26
	Hispanic	572	0.93	0.93	0.93	654.13	15.41
	White	269	0.93	0.94	0.94	669.81	16.14
7	All Examinees	4,297	0.92	0.92	0.93	753.33	17.49
	Male	2,150	0.93	0.93	0.93	751.91	18.36
	Female	2,131	0.92	0.92	0.92	754.82	16.37
	Asian	55	0.93	0.93	0.93	771.02	16.60
	African American	3,435	0.91	0.91	0.92	751.38	16.56
	Hispanic	527	0.92	0.92	0.92	754.97	16.70
	White	240	0.93	0.94	0.94	773.05	18.03
8	All Examinees	4,341	0.92	0.92	0.92	850.23	16.59
	Male	2,161	0.91	0.91	0.92	848.52	17.18
	Female	2,156	0.92	0.92	0.92	852.05	15.75
	Asian	57	0.94	0.94	0.95	865.98	14.79
	African American	3,508	0.90	0.90	0.91	848.57	16.17
	Hispanic	493	0.90	0.91	0.91	851.80	14.59
	White	234	0.93	0.93	0.93	868.26	14.43
10	All Examinees	3,466	0.91	0.92	0.92	946.80	18.80
	Male	1,678	0.92	0.92	0.92	945.67	19.59
	Female	1,748	0.91	0.91	0.91	948.10	17.84
	Asian	60	0.89	0.89	0.89	964.82	12.73
	African American	2,819	0.90	0.91	0.91	945.27	18.18
	Hispanic	409	0.90	0.91	0.91	948.37	17.38
	White	134	0.94	0.94	0.95	968.23	19.64

Table D3. Internal Consistency Reliability Coefficients for Examinee Subgroups: Science/Biology

Grade	Subgroup	N	Alpha	Stratified Alpha	Feldt-Raju	Mean	SD
5	All Examinees	4,697	0.89	0.89	0.89	548.40	13.28
	Male	2,374	0.89	0.89	0.90	547.45	14.22
	Female	2,296	0.89	0.89	0.89	549.44	12.13
	Asian	79	0.90	0.91	0.91	557.41	9.53
	African American	3,632	0.86	0.86	0.86	546.51	13.14
	Hispanic	588	0.86	0.86	0.86	550.11	11.52
	White	365	0.86	0.86	0.86	562.42	6.68
8	All Examinees	4,253	0.88	0.88	0.88	848.66	17.76
	Male	2,088	0.89	0.89	0.89	847.90	18.47
	Female	2,120	0.87	0.87	0.87	849.69	16.79
	Asian	57	0.90	0.90	0.90	861.11	8.39
	African American	3,416	0.85	0.85	0.86	847.23	17.82
	Hispanic	493	0.84	0.85	0.85	850.23	15.10
	White	233	0.91	0.91	0.91	865.06	11.40
High School	All Examinees	3,693	0.85	0.86	0.86	947.91	14.76
	Male	1,731	0.87	0.87	0.87	947.22	15.63
	Female	1,882	0.84	0.84	0.84	948.92	13.45
	Asian	69	0.88	0.88	0.88	955.90	9.04
	African American	2,947	0.82	0.82	0.82	946.83	14.59
	Hispanic	396	0.84	0.85	0.85	948.80	13.74
	White	197	0.89	0.89	0.89	962.38	7.86

Table D4. Internal Consistency Reliability Coefficients for Examinee Subgroups: Composition

Grade	Subgroup	N	Alpha	Stratified Alpha	Feldt-Raju	Mean	SD
4	All Examinees	4,508	0.92	0.92	0.93	451.73	18.87
	Male	2,284	0.93	0.92	0.93	448.48	19.16
	Female	2,215	0.92	0.91	0.92	455.14	17.93
	Asian	104	0.92	0.92	0.93	459.50	18.29
	African American	3,293	0.91	0.91	0.91	448.95	18.15
	Hispanic	623	0.90	0.90	0.91	453.93	17.44
	White	458	0.85	0.85	0.87	466.71	17.74
7	All Examinees	4,176	0.91	0.90	0.92	754.33	15.76
	Male	2,086	0.91	0.91	0.92	751.07	15.91
	Female	2,083	0.90	0.89	0.91	757.60	14.90
	Asian	55	0.89	0.88	0.90	765.44	14.56
	African American	3,360	0.90	0.90	0.91	752.37	15.12
	Hispanic	498	0.90	0.90	0.91	758.36	14.80
	White	228	0.90	0.91	0.91	770.39	15.35
10	All Examinees	3,429	0.92	0.92	0.93	952.18	20.11
	Male	1,616	0.93	0.93	0.93	948.47	19.80
	Female	1,801	0.92	0.92	0.93	955.53	19.85
	Asian	43	0.91	0.91	0.93	965.28	15.21
	African American	2,879	0.92	0.92	0.93	950.67	20.24
	Hispanic	364	0.91	0.91	0.92	956.28	15.71
	White	123	0.91	0.91	0.92	970.93	18.80

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 2					
Males	0.70	0.58	0.79	0.10	0.11
Females	0.69	0.56	0.77	0.11	0.12
Asian	0.69	0.53	0.78	0.12	0.10
African American	0.70	0.56	0.78	0.10	0.12
Hispanic	0.69	0.56	0.78	0.11	0.12
White	0.70	0.52	0.78	0.12	0.10
Grade 3					
Males	0.79	0.70	0.85	0.07	0.08
Females	0.78	0.68	0.84	0.07	0.09
Asian	0.73	0.54	0.80	0.10	0.11
African American	0.79	0.69	0.85	0.06	0.08
Hispanic	0.78	0.67	0.85	0.07	0.08
White	0.77	0.56	0.84	0.07	0.09
Grade 4					
Males	0.78	0.67	0.84	0.07	0.09
Females	0.77	0.65	0.84	0.07	0.09
Asian	0.78	0.66	0.84	0.07	0.08
African American	0.77	0.66	0.84	0.07	0.09
Hispanic	0.77	0.65	0.84	0.07	0.09
White	0.80	0.61	0.72	0.02	0.26
Grade 5					
Males	0.81	0.74	0.74	0.08	0.18
Females	0.76	0.64	0.83	0.09	0.09
Asian	0.76	0.63	0.83	0.10	0.07
African American	0.76	0.64	0.83	0.08	0.09
Hispanic	0.76	0.63	0.83	0.08	0.09
White	0.74	0.54	0.81	0.10	0.10
Grade 6					
Males	0.78	0.67	0.85	0.07	0.08
Females	0.76	0.63	0.83	0.08	0.09
Asian	0.75	0.61	0.82	0.09	0.09
African American	0.77	0.65	0.84	0.08	0.09
Hispanic	0.77	0.64	0.83	0.09	0.08
White	0.75	0.57	0.82	0.09	0.09

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

Grade 7					
Males	0.72	0.59	0.80	0.10	0.10
Females	0.71	0.58	0.79	0.10	0.11
Asian	0.73	0.59	0.80	0.09	0.10
African American	0.71	0.58	0.79	0.10	0.11
Hispanic	0.70	0.56	0.78	0.11	0.11
White	0.79	0.62	0.85	0.09	0.06
Grade 8					
Males	0.72	0.60	0.81	0.09	0.10
Females	0.73	0.62	0.81	0.09	0.10
Asian	0.73	0.60	0.81	0.11	0.08
African American	0.72	0.59	0.80	0.09	0.10
Hispanic	0.72	0.58	0.80	0.10	0.10
White	0.77	0.60	0.84	0.09	0.07
Grade 9					
Males	0.74	0.62	0.82	0.09	0.10
Females	0.74	0.64	0.82	0.09	0.09
Asian	0.77	0.65	0.85	0.06	0.09
African American	0.74	0.63	0.81	0.09	0.09
Hispanic	0.74	0.64	0.82	0.09	0.09
White	0.80	0.69	0.86	0.06	0.08
Grade 10					
Males	0.75	0.64	0.83	0.07	0.10
Females	0.74	0.62	0.82	0.09	0.09
Asian	0.74	0.63	0.82	0.09	0.09
African American	0.75	0.63	0.82	0.08	0.10
Hispanic	0.72	0.60	0.81	0.09	0.10
White	0.76	0.57	0.83	0.10	0.08

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 2					
Males	0.72	0.61	0.79	0.11	0.11
Females	0.72	0.60	0.79	0.10	0.11
Asian	0.73	0.58	0.78	0.10	0.12
African American	0.72	0.60	0.79	0.10	0.11
Hispanic	0.71	0.59	0.78	0.10	0.11
White	0.74	0.55	0.76	0.14	0.10
Grade 3					
Males	0.78	0.69	0.85	0.07	0.08
Females	0.78	0.69	0.85	0.07	0.08
Asian	0.76	0.63	0.83	0.08	0.09
African American	0.84	0.78	0.79	0.04	0.17
Hispanic	0.76	0.65	0.83	0.08	0.10
White	0.74	0.59	0.82	0.09	0.09
Grade 4					
Males	0.77	0.67	0.83	0.08	0.08
Females	0.77	0.67	0.83	0.08	0.08
Asian	0.82	0.72	0.87	0.06	0.07
African American	0.76	0.66	0.83	0.08	0.08
Hispanic	0.77	0.66	0.83	0.08	0.09
White	0.79	0.64	0.85	0.08	0.07
Grade 5					
Males	0.77	0.67	0.83	0.08	0.09
Females	0.76	0.66	0.83	0.09	0.09
Asian	0.78	0.66	0.83	0.08	0.09
African American	0.76	0.67	0.83	0.08	0.09
Hispanic	0.76	0.66	0.83	0.08	0.09
White	0.76	0.61	0.82	0.09	0.09
Grade 6					
Males	0.76	0.67	0.83	0.08	0.08
Females	0.76	0.66	0.83	0.09	0.08
Asian	0.94	0.88	0.81	0.03	0.16
African American	0.76	0.65	0.83	0.08	0.09
Hispanic	0.76	0.66	0.83	0.09	0.07
White	0.82	0.67	0.87	0.06	0.07
Grade 7					
Males	0.75	0.64	0.82	0.08	0.09
Females	0.75	0.63	0.82	0.09	0.09
Asian	0.80	0.65	0.85	0.08	0.06
African American	0.74	0.62	0.81	0.09	0.10
Hispanic	0.76	0.64	0.83	0.09	0.08
White	0.84	0.70	0.89	0.06	0.06

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

Grade 8					
Males	0.72	0.60	0.80	0.09	0.11
Females	0.73	0.60	0.81	0.10	0.09
Asian	0.79	0.67	0.85	0.09	0.06
African American	0.72	0.58	0.80	0.10	0.10
Hispanic	0.73	0.59	0.81	0.10	0.09
White	0.83	0.70	0.88	0.06	0.06
Grade 10					
Males	0.73	0.62	0.80	0.09	0.10
Females	0.73	0.60	0.80	0.10	0.10
Asian	0.80	0.63	0.86	0.06	0.08
African American	0.72	0.60	0.80	0.10	0.10
Hispanic	0.72	0.59	0.80	0.10	0.11
White	0.80	0.69	0.86	0.08	0.06

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	0.72	0.60	0.80	0.10	0.10
Females	0.71	0.58	0.80	0.10	0.11
Asian	0.76	0.63	0.82	0.08	0.10
African American	0.71	0.57	0.80	0.10	0.11
Hispanic	0.70	0.55	0.79	0.10	0.11
White	0.77	0.61	0.83	0.08	0.09
Grade 8					
Males	0.68	0.54	0.76	0.10	0.13
Females	0.68	0.54	0.76	0.11	0.13
Asian	0.74	0.59	0.82	0.09	0.09
African American	0.67	0.52	0.76	0.10	0.14
Hispanic	0.66	0.51	0.75	0.12	0.14
White	0.79	0.65	0.85	0.08	0.07
High School					
Males	0.67	0.51	0.75	0.11	0.14
Females	0.65	0.48	0.74	0.13	0.13
Asian	0.74	0.55	0.81	0.09	0.10
African American	0.65	0.48	0.74	0.13	0.14
Hispanic	0.66	0.49	0.75	0.12	0.13
White	0.78	0.62	0.84	0.08	0.08

Table E4. Classification Consistency and Accuracy Rates for All Cut Scores and Examinee Subgroups: Composition

Grade/Subgroup	Classification Consistency		Classification Accuracy		
	Consistency	Kappa	Accuracy	False Positive Errors	False Negative Errors
Grade 4					
Males	0.55	0.37	0.65	0.17	0.18
Females	0.52	0.34	0.62	0.20	0.18
Asian	0.54	0.35	0.64	0.18	0.18
African American	0.53	0.35	0.63	0.18	0.19
Hispanic	0.52	0.33	0.62	0.20	0.18
White	0.56	0.34	0.65	0.22	0.13
Grade 7					
Males	0.59	0.43	0.69	0.15	0.15
Females	0.57	0.40	0.68	0.18	0.15
Asian	0.60	0.39	0.69	0.19	0.12
African American	0.57	0.40	0.68	0.17	0.16
Hispanic	0.59	0.43	0.70	0.16	0.14
White	0.70	0.44	0.78	0.14	0.08
Grade 10					
Males	0.52	0.34	0.62	0.16	0.21
Females	0.52	0.34	0.61	0.19	0.19
Asian	0.55	0.34	0.64	0.22	0.15
African American	0.51	0.34	0.61	0.18	0.21
Hispanic	0.51	0.32	0.60	0.19	0.21
White	0.66	0.42	0.73	0.16	0.10