

**Technical Report
Spring 2013 Test Administration**

**Washington, D.C.
Comprehensive Assessment System
(DC CAS)
Health and Physical Education
Assessment
for Grades 5, 8, and High School**

September 11, 2013



**CTB/McGraw-Hill
Monterey, California 93940**

Developed and published under contract with District of Columbia Office of the State Superintendent of Education (OSSE) by CTB/McGraw-Hill LLC, 20 Ryan Ranch Road, Monterey, California 93940-5703. Copyright © 2013 by the District of Columbia Office of the State Superintendent of Education. Based on a template, copyright © CTB/McGraw-Hill LLC. All rights reserved. Only authorized customers may copy, download and/or print the document, located online at <http://osse.dc.gov>. Any other use or reproduction of this document, in whole or in part, requires written permission of the District of Columbia Office of the State Superintendent of Education and the publisher.

Table of Contents

List of Tables4

Section 1. Overview.....5

Section 2. Item and Test Development.....6

 Overview.....6

 Content Standards and Item Development6

 Test Development.....7

 Test Design7

Section 3. Test Administration Guidelines and Requirements.....10

 Overview.....10

 Guidelines and Requirements for Administering DC CAS10

 Materials Orders, Delivery, and Retrieval11

 Secure Inventory11

Section 4. Student Participation13

 Tests Administered13

 Participation in DC CAS.....13

 Definition of Valid Test Administration.....13

 Participation Rates13

 Special Accommodation14

Section 5. Methods17

 Classical Item Level Analyses17

 Item Bias Analyses17

 Calibration and Equating18

 Goodness of Fit.....18

 Year-to-Year Equating Procedures.....19

 Establishing Upper and Lower Bounds for the Grade Level Scales.....21

 Reliability Coefficients21

 Standard Errors of Measurement22

Section 6. Evidence for Reliability and Validity23

 Reliability.....23

 Validity23

 Item Level Evidence23

 Classical Item Statistics23

 Differential Item Function24

 Test and Strand Level Evidence24

 Total Test Scores.....24

 Strand Level Scores24

 Standard Errors of Measurement24

References.....36

Appendix A: Item Acceptability Checklist.....37

Appendix B: Health and PE Test Item Adjusted *P* Values.....38

List of Tables

Table 1. DC CAS 2013 Operational Test Form Blueprints: Health and PE.....8

Table 2. Number of Examinees with Valid Health and PE Test Administrations and Responding to Opt-Out Items, and Percent of Students Who Chose to “Opt Out”.....15

Table 3. Number and Percent of Examinees with Valid Health and PE Test Administrations across Subgroups*15

Table 4. Number and Percent of Students Receiving One or More Test Administration Accommodations16

Table 5. Correlations Between the Item Parameters for the Reference Form and 2013 DC CAS Operational Test Form22

Table 6. Scaling Constants Across Administrations.....22

Table 7. DC CAS 2013 Classical Item Level Statistics.....25

Table 8. Numbers of Operational and OP-Opt-Out Items Flagged for DIF Using the Mantel-Haenszel Procedure.....26

Table 9. Numbers of Field Test and FT-Opt-Out Items Flagged for DIF Using the Mantel-Haenszel Procedure.....27

Table 10. Total Test Scale and Raw Score Means and Reliability Statistics28

Table 11. Adjusted *P* Value Means and Standard Deviations, and Coefficient Alpha Reliability for Strand Scores.....29

Table 12. DC CAS 2013 Strand-to-Strand Correlations.....31

Table 13. DC CAS 2013 Number Correct to Scale Score Conversions with Associated Standard Errors of Measurement (SEM) Operational and OP-Opt-Out34

Table 14. DC CAS 2013 Number Correct to Scale Score Conversions with Associated Standard Errors of Measurement (SEM) Operational.....35

Table B1. DC CAS 2013 Operational and OP-Opt-Out Item Adjusted *P* Values, Grade 538

Table B2. DC CAS 2013 Operational and OP-Opt-Out Item Adjusted *P* Values, Grade 839

Table B3. DC CAS 2013 Operational and OP-Opt-Out Item Adjusted *P* Values, High School.....40

Table B4. DC CAS 2013 Field Test and FT-Opt-Out Item Adjusted *P* Values, Grade 541

Table B5. DC CAS 2013 Field Test and FT-Opt-Out Item Adjusted *P* Values, Grade 842

Table B6. DC CAS 2013 Field Test and FT-Opt-Out Item Adjusted *P* Values, High School.....43

Section 1. Overview

This technical report describes the Health and Physical Education (Health and PE) assessment, as required by Section 405 of the Healthy Schools Act of 2010. The Health and PE assessment is considered part of the operational District of Columbia Comprehensive Assessment System (DC CAS) and was administered to students in the spring of 2013 to assess students' skills in Grades 5, 8, and High School Health and PE. Scores from these assessments were not reported at an individual student level in 2012 and 2013. In 2013, the Office of the State Superintendent of Education (OSSE) will generate score summary reports at the school and district level to monitor school and district Health and PE student knowledge. This technical report is written to document procedures and results from developing, analyzing, and validating the 2013 DC CAS Health and PE assessment.

Technical reports provide information relevant to an evaluation of the validity of intended interpretations and uses of results from the 2013 DC CAS tests. According to the Standards for educational and psychological testing, the 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 for Health and PE, technical reports target audiences with some level of technical training and understanding.

Section 2. Item and Test Development

Overview

The procedures test developers used to develop the test's content and the alignment of items with the test blueprint and specifications are a key piece of validity evidence for the DC CAS Health and PE assessment. Therefore, the purpose of Section 2 is to provide a description of the events that took place in the development of the DC CAS Health and PE assessment.

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.

According to the Healthy Schools Act of 2010 (D.C. Law 18-209) Report (2012), the Office of the State Superintendent of Education (OSSE) “convened a task force in summer of 2010, comprised of representatives from the Office of the State Superintendent of Education (OSSE), District of Columbia Public Schools (DCPS), Public Charter School Board, Friends of Choice in Urban Schools (FOCUS), Student Support Center, State Board of Education, DC Department of Health, DC Council Committee on Health, Friendship PCS, Metro Teen AIDS, George Washington University, and American University. The task force recommended the development of a standards-based Comprehensive Assessment System (DC CAS) for health and physical education. This assessment was developed and administered to 5th and 8th graders and high school students enrolled in health, as part of the DC CAS tests starting in April 2012. Each assessment contained 55 to 60 multiple choice items, covering topics such as nutrition, communication and emotional health, disease prevention, safety skills, and sexual health.”

Content Standards and Item Development

The Healthy Schools Act Report goes on to describe the standards to which the items were developed.

“The items on the assessment were derived from the Health Education Assessment Project (HEAP) of the Council of Chief State School Officers (CCSSO). The items were aligned to the OSSE health and physical education learning standards and edited to be unique to the standards and the District of Columbia.

Similar to the process of sexual health education, a passive consent form was sent home with students, and parents/guardians were able to “opt out” of the sexual health questions. Depending on grade level, these questions were the final three, four, or five test questions, and students either stopped the test prior to these questions or completed all 50 questions.

Physical education standards were also covered on the DC CAS for health and physical education; however, most physical education standards cannot be assessed with a multiple-choice test. Many schools use a tool to assess achievement in regards to the physical education standards; however, this tool varies by Local Education Agency.

DCPS uses the FitnessGram for students in grade four and above. Appendix G (of the Healthy Schools Act) has more information on this tool. This data is collected once per year and assess:

- Aerobic Capacity, as measured by a progressive aerobic cardiovascular endurance run (PACER)
- Body Composition, as measured by either a skin fold test or body mass index (BMI)
- Muscular Strength and Endurance, as measured by curl-ups and push-ups
- Flexibility, as measured by a back-saver sit and reach.”

The newly developed Health and PE items were exclusively of multiple choice (MC) type, and were examined through a rigorous content and psychometric review and approval process. CTB content and style editors, supervisors, and managers reviewed all items for content and grade appropriateness, and alignment to the content standards. 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, assembled test forms based on the Health and PE items designed to measure student performance. The total number of items and score points emphasized within each reporting category served as the test blueprint, details of which are provided in Table 1.

Test Design

The 2013 DC CAS Health and PE 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. All 2013 operational items are the same as 2012, and there is only one form per grade.

Unique to the Health and PE tests are items aligned to sexual health standards that students, prior to the start of testing, can be permitted to omit or “opt out” of responding. These items contain content to which parents may have requested limiting student exposure. In this report, we refer to those items as “opt-out” items.

Table 1. DC CAS 2013 Operational Test Form Blueprints: Health and PE

Grade	Content Standard		Operational Items	Opt-Out Items	Operational and Opt-Out Items		Field Test Items
			Number of Items	Number of Items	Total Number of Items	% of Total Points	Number of Items
5	1	Communication and Emotional Health	7	0	7	16%	1
	2	Safety Skills	5	0	5	12%	3
	3	Human Body and Personal Health	4	1	5	12%	2
	4	Disease Prevention	4	2	6	14%	2
	5	Nutrition	5	0	5	12%	1
	6	Alcohol, Tobacco and Other Drugs	4	0	4	9%	2
	7	Health Decision Making	6	0	6	14%	0
	8	Physical Education	5	0	5	12%	1
	Total		40	3	43	100%	12
8	1	Communication and Emotional Health	6	0	6	13%	1
	2	Safety Skills and Community Health	5	0	5	11%	2
	3	Human Development and Sexuality	0	5	5	11%	3
	4	Disease Prevention	7	0	7	16%	1
	5	Nutrition	6	0	6	13%	2
	6	Alcohol, Tobacco and Other Drugs	5	0	5	11%	2
	7	Health Information and Advocacy	5	0	5	11%	3
	8	Physical Education	6	0	6	13%	1
	Total		40	5	45	100%	15

Table 1. DC CAS 2013 Operational Test Form Blueprints: Health and PE (continued)

Grade	Content Standard		Operational Items	Opt-Out Items	Operational and Opt-Out Items		Field Test Items
			Number of Items	Number of Items	Total Number of Items	% of Total Points	Number of Items
High School	1	Human Growth and Development	4	0	4	9%	2
	2	Sexuality and Reproduction	0	5	5	11%	3
	3	Disease Prevention and Treatment	9	0	9	20%	1
	4	Nutrition	5	0	5	11%	3
	5	Alcohol, Tobacco and Other Drugs	4	0	4	9%	2
	6	Locate Health Information and Assistance	6	0	6	13%	1
	7	Safety Skills	6	0	6	13%	2
	8	Physical Education	6	0	6	13%	1
		Total	40	5	45	100%	15

Note: All Operational items are anchors.

Section 3. Test Administration Guidelines and Requirements

Overview

Administration of the DC CAS assessments each spring is managed by 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 2013 *Test Chairperson's Manual*. They, in turn, trained all Test Administrators and proctors. Test Administrators administered all DC CAS assessments according to requirements and steps in the *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* document covers similar topics and requirements. In addition, it provides instructions on scheduling test administrations, preparing students for the test administration, using standardized testing procedures, and verbatim instructions for administering each test to students. It also provides information on available accommodations for students with disabilities and English language learners.

Recall that students have the option to “opt out” of taking certain items aligned to sexual health standards. The *Test Chairperson's Manual* and *Test Directions* both cover the procedures to follow during testing to accommodate students that chose to opt out of taking these items.

Guidelines and Requirements for Administering DC CAS

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

The *Test Chairperson's Manual* and *Test Directions* provide information on available test administration accommodations for students with disabilities and English language learners. It specifies approved accommodations that maintain standard testing conditions and identifies accommodations that are considered modifications to the test, which will result in invalidated test. The *Test Chairperson's Manual* specifies how to indicate opt-out status on a student's answer booklet (“Special Use Only” bubbles are filled). The *Test Directions* provide verbatim directions for Test Administrators to collect test materials from students who have opted out prior to having other students complete the sexual health items, which are located at the end of the Health assessment.

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 sections of the tests), setting (e.g., individual and small group administrations), presentation, 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 linguistic support. Both manuals indicate that Test Administrators must record on the student's answer document all test administration accommodations that are provided.

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

Materials Orders, Delivery, and Retrieval

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

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

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

After the test administration was complete, the materials were packaged for retrieval and picked up according to a verified schedule. Daily tracking reports also served for OSSE to monitor retrievals. When the materials were back in CTB's custody, all books with security bar codes 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 bar coding 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 to 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 August 18, 2013, approximately 99.15% of secure materials for DC CAS Health and PE were accounted for; 141 secure test booklets were missing from the 16,674 test booklets that were shipped during the 2013 administration.

Section 4. Student Participation

Tests Administered

All public schools in the District of Columbia administered the DC CAS tests between April 22 and May 2, 2013.

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 4,300 students were assessed in Grade 5; 4,000 in Grade 8; and about 3,200 in High School. Only students with a valid test administration as required by the type of analysis, as defined below, are included in this report.

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 five.
- 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 2013 Business Requirements, are included in these analyses, where valid attempt on the test is defined as:

- At least one item marked with a correct response OR
- At least five items validly marked in the content area

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.

Participation Rates

The total number and percent of students with valid tests and those who participated in the opt-out items are provided in Table 2. As can be seen, the large majority of students responded to all items, including opt-out items. In each grade, the percentage of students who chose to "opt out" and not take the items was 4% in Grade 5, and less than 1% in both Grade 8 and High School.

The total number and percent of students and the number and percent of students in the subgroups of gender and race/ethnicity, as well as in special subgroups such as special education, 504 plans, and English language learners (ELLs), are provided in Table 3.

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 rates of the various accommodations documented are provided in Table 4. For more information on these accommodations, please refer to the DC CAS *Test Chairperson's Manual*.

Table 2. Number of Examinees with Valid Health and PE Test Administrations and Responding to Opt-Out Items, and Percent of Students Who Chose to “Opt Out”.

Grade	Students with Test Scores	Students Responding to “Opt-Out” Items	Percentage of Students Who Chose to “Opt Out”
5	4,331	4,145	4.29%
8	3,985	3,960	0.63%
High School	3,263	3,259	0.12%

Table 3. Number and Percent of Examinees with Valid Health and PE Test Administrations across Subgroups*

Grade	Students with Test Scores	Males		Females		Asian		African American		Hispanic		White	
		N	%	N	%	N	%	N	%	N	%	N	%
5	4,331	2,172	50%	2,135	49%	84	2%	3,131	72%	579	13%	450	10%
8	3,985	1,981	50%	1,993	50%	50	1%	3,084	77%	496	12%	271	7%
High School	3,263	1,529	47%	1,596	49%	59	2%	2,476	76%	395	12%	188	6%

Table 3. Number and Percent of Examinees with Valid Health and PE Test Administrations across Subgroups* (continued)

Grade	Students with Test Scores	Special Education		English Language Learner		Section 504		Title I Targeted		Home Schooling	
		N	%	N	%	N	%	N	%	N	%
5	4,331	455	11%	167	4%	43	1%	88	2%	0	0%
8	3,985	387	10%	158	4%	13	0%	94	2%	2	0%
High School	3,263	294	9%	33	1%	21	1%	5	0%	0	0%

*Note that the percentages may not sum to 100% given that not all students provided complete demographic information.

Table 4. Number and Percent of Students Receiving One or More Test Administration Accommodations

Grade	Students with Test Scores	Direct Linguistic Support—Oral		Direct Linguistic Support—Written		Indirect Linguistic Support		Other	
	N	N	%	N	%	N	%	N	%
5	4,331	126	3%	84	2%	130	3%	2	2%
8	3,985	190	5%	135	3%	199	5%	0	0%
High School	3,263	163	5%	102	3%	162	5%	0	0%

Table 4. Number and Percent of Students Receiving One or More Test Administration Accommodations (continued)

Grade	Students with Test Scores	Timing/Scheduling		Setting		Presentation		Response		Other		Students with Special Education Code	
	N	N	%	N	%	N	%	N	%	N	%	N	%
5	4,331	507	12%	519	12%	470	11%	316	7%	23	1%	455	11%
8	3,985	429	11%	465	12%	426	11%	307	8%	7	0%	387	10%
High School	3,263	319	10%	328	10%	259	8%	228	7%	6	0%	294	9%

Table 4. Number and Percent of Students Receiving One or More Test Administration Accommodations(continued)

Grade	Students with Test Scores	Breaks		Small Group and Individual Administrations		Read or Translate Test Questions		Responses Dictated	
	N	N	%	N	%	N	%	N	%
5	4,331	467	11%	500	12%	390	9%	98	2%
8	3,985	388	10%	434	11%	326	8%	44	1%
High School	3,263	278	9%	305	9%	135	4%	62	2%

Section 5. Methods

This section describes the methods used to analyze the item and test level data for the DC CAS Health and PE assessments. Results of the item and test level analyses described here are provided as evidence for reliability and validity in Section 6.

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 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 p values 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 values 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 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. When any point biserial on the distractor is a positive correlation or when the correlation is very low, then the item is reviewed for potentially having more than one correct response or having been miskeyed.

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 may be 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 2013 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 review all items that are flagged for DIF after each administration to identify whether content appears in the items that may favor or disadvantage examinee subgroups.

The statistic flags items for potential DIF using the following criteria:

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

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) IRT model for item calibration and scaling. These software programs were developed at CTB/McGraw-Hill to enable scaling and linking of complex assessment data.

In PARDUX (Burket, 1995), a marginal maximum likelihood procedure was used to simultaneously estimate the item parameters under the 3PL model (used for multiple choice items) (Bock & Aitkin, 1981; Thiessen, 1982). 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:

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

In the equation, 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.

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 10 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 model is 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 $DF = 10(2-1) - 3 = 7$. 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 $(N/1500 \times 4)$ 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 10 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. Only three operational items were flagged for poor fit to the IRT model in Grade 5, one item in Grade 8, and three items in Grade 10.

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. From 2012 to 2013, 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 TCC based on estimates from the previous calibration and $\hat{\psi}_j^*$ be the TCC 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 were all operational MC items. Anchor items were 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 Table 1 (all Operational items are anchors). The blueprint is fully represented in the anchor sets.

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 2012 and 2013 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. Because IRT guessing (c) parameters typically fluctuate considerably, there were held to their fixed reference values during calibration and were not considered in this evaluation. The correlations are shown in Table 5 for the discrimination (a) and difficulty (b) parameters and are high, ranging from 0.90 to 0.94 for a parameters and from 0.98 to 0.99 for b parameters. 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. Further, statistical differences between the reference and equated item parameters 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 2012–2013 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 2013 and the 2012 DC CAS administration are provided in Table 6.

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 5 are highly correlated at 1.00 for all grades. This is an indication that the anchor items performed similarly in the examinee populations in 2012 and 2013.

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. Specifically, the LOSS and HOSS for Grade 5 are 500 and 599, for Grade 8 are 800 and 899, and for High School are 900 and 999, respectively. These values remain constant from year to year.

Reliability Coefficients

Total test reliability statistics (alpha and CSEMs) measure the level of internal 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 total test reliabilities of the operational forms were evaluated first by Cronbach's α (1951) index of internal consistency. The specific calculation for Cronbach's α is calculated as

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

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.

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.

Table 5. Correlations Between the Item Parameters for the Reference Form and 2013 DC CAS Operational Test Form

Grade	Discrimination (a)	Difficulty (b)	<i>P</i> Value Correlation
5	0.94	0.98	1.00
8	0.94	0.99	1.00
High School	0.90	0.99	1.00

Table 6. Scaling Constants Across Administrations

Grade	2012		2013	
	Multiplicative	Additive	Multiplicative	Additive
5	10.00	550.00	10.47	549.45
8	10.70	853.00	11.20	851.99
High School	10.00	945.00	11.13	942.49

Section 6. Evidence for Reliability and Validity

Reliability

Reliability refers to the degree to which students' scores are free from measurement errors 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 evidence 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 established. 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 items are all reviewed for statistical accuracy and quality. Table 7 summarizes classical item level statistics (adjusted p values, point biserial correlations, omit rates, and rates of items not reached) for Health and PE operational, Op-Opt-Out, operational and OP-Opt-Out, field test, and FT-Opt-Out items. On average, the operational collection of items on the tests was above average (0.50 p value) at 0.65 for Grades 5 and 8, and 0.62 for High School. The mean p value of operational and Op-Opt-Out items was very close to that of the operational items at 0.64 for Grades 5 and 8, and 0.63 for High School. Op-Opt-Out items and field test items were, on average, slightly more difficult than operational items for Grades 5 and 8. For High School, field test items were more difficult and Op-Opt-Out were less difficult than operational items. There was no FT-Opt-Out item in Grade 5 and FT-Opt-Out items were less difficult than operational items for Grade 8 and High School. The tables in Appendix B display the item-specific difficulty for each item at each grade and include the operational items and the Op-Opt-Out items (flagged with an asterisk).

The point biserial (Item-Total Correlation) is one measure of the correlation between each item and the overall test. The correlations for the operational or operational and OP-Opt-Out items were higher than those for OP-Opt-Out and Field Test items in Grades 5 and 8. However, the correlations for High School Opt-Out items were slightly higher than operational or operational and OP-Opt-Out.

With respect to omit rates and number of items not reached, 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 and to ensure the administration conditions, such as testing time and accurate printing and scanning. Overall, the omit rates were low and less than the 5% criteria. However, a larger percentage of students in the group taking the opt-out items actually omitted those items. This is an indication that some students who were supposed to take all items—since their responses were not flagged during administration as opting out of the sex-ed items—actually did not respond. The rates were as high as 13% at Grade 5, about 7% at Grade 8, and 12% at High School.

Differential Item Function

Differential item function (DIF) analyses were conducted for all grades 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 8-9 summarize the 2013 DIF analysis results for Health and PE items. Modest numbers of items were flagged for DIF at levels B and C.

Test and Strand Level Evidence

Total Test Scores

Total test level raw score and scale score means and standard deviations are provided in Table 10, along with the test level reliability coefficients, including Cronbach alpha and stratified coefficient alpha. The scale score and raw score means and standard deviations are consistent across grades. The reliabilities all show high levels of internal consistency, with reliabilities all greater than or equal to 0.85.

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 it is 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, provided in Table 11, highlights the strands that tend to be more or less difficult for students.

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 for each strand. The internal reliability estimates for these strand scores, which include as few as four items and as many as nine, range between 0.24 and 0.80. As an additional measure of internal consistency, correlations have been produced between strands within each grade. These are provided in Table 12. A review of the correlations shows only moderate relationships amongst strands.

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. Table 13 and Table 14 list the number correct to scale score values along with their associated IRT SEM values for operational and op-opt-out items, and operational items respectively. The SEMs in the extreme

scores tend to be larger, as expected, and where the majority of students are likely to fall in their score performance, the SEMs are quite low.

Table 7. DC CAS 2013 Classical Item Level Statistics

Grade	Item Type	Number of Items	Mean	Mean	Mean Omit Rate	Mean Not Reached Rate
			Adjusted <i>p</i> value	Item-Total Correlation		
5	Operational	40	0.65	0.35	0.26	0.17
	OP-Opt-Out	3	0.54	0.29	12.85	12.80
	OP + OP Opt Out	43	0.64	0.35	1.14	1.05
	Field Test	12	0.56	0.28	0.41	0.25
	FT-Opt-Out	0	N/A	N/A	N/A	N/A
8	Operational	39	0.65	0.33	0.46	0.31
	OP-Opt-Out	5	0.59	0.21	6.57	6.38
	OP + OP Opt Out	44	0.64	0.31	1.16	1.00
	Field Test	13	0.58	0.29	0.62	0.40
	FT-Opt-Out	2	0.77	0.40	6.59	6.59
High School	Operational	39	0.62	0.34	0.64	0.38
	OP-Opt-Out	5	0.73	0.43	12.23	12.14
	OP + OP Opt Out	44	0.63	0.35	1.96	1.72
	Field Test	14	0.57	0.34	1.03	0.79
	FT-Opt-Out	1	0.83	0.46	12.44	12.44

Note: Omit and not reached rates are percentages.

Table 8. Numbers of Operational and OP-Opt-Out Items Flagged for DIF Using the Mantel-Haenszel Procedure

Reference Group	Focal Group	A	B	B-	C	C-	N/A
Grade 5 (total 43 items)							
Male	Female	37	5	1	0	0	0
African American	Asian	N/A	N/A	N/A	N/A	N/A	43
	Hispanic	37	3	1	0	2	0
	White	28	5	1	8	1	0
Grade 8 (total 44 items*)							
Male	Female	39	1	2	2	0	0
African American	Asian	N/A	N/A	N/A	N/A	N/A	44
	Hispanic	38	1	2	1	2	0
	White	31	5	0	7	1	0
High School (total 44 items*)							
Male	Female	37	5	0	1	1	0
African American	Asian	N/A	N/A	N/A	N/A	N/A	44
	Hispanic	35	4	2	2	1	0
	White	N/A	N/A	N/A	N/A	N/A	44

Note: Positive flags indicate DIF that favors the focal group. A = no DIF; B = moderate DIF; C = considerable DIF. N/A = not applicable because case count requirements for the reference (400) and focal (200) groups were not met.

See Table 3 for the numbers of examinees in each grade and subgroup.

*One item deemed statistically unacceptable is suppressed in Grade 8 and High School.

Table 9. Numbers of Field Test and FT-Opt-Out Items Flagged for DIF Using the Mantel-Haenszel Procedure

Reference Group	Focal Group	A	B	B-	C	C-	N/A
Grade 5 (total 12 items)							
Male	Female	12	0	0	0	0	0
African American	Asian	N/A	N/A	N/A	N/A	N/A	12
	Hispanic	8	1	1	0	2	0
	White	9	1	1	1	0	0
Grade 8 (total 15 items)							
Male	Female	12	1	0	2	0	0
African American	Asian	N/A	N/A	N/A	N/A	N/A	15
	Hispanic	14	0	1	0	0	0
	White	14	0	0	1	0	0
High School (total 15 items)							
Male	Female	11	3	0	1	0	0
African American	Asian	N/A	N/A	N/A	N/A	N/A	15
	Hispanic	13	1	1	0	0	0
	White	N/A	N/A	N/A	N/A	N/A	15

Note: Positive flags indicate DIF that favors the focal group. A = no DIF; B = moderate DIF; C = considerable DIF. N/A = not applicable because case count requirements for the reference (400) and focal (200) groups were not met.

See Table 3 for the numbers of examinees in each grade and subgroup.

Table 10. Total Test Scale and Raw Score Means and Reliability Statistics

Grade	Item Type	Students with Test Scores	Number of Items	Alpha	Feldt-Raju	Scale Score		Criterion Score	
						Mean	SD	Mean	SD
5	Operational	4,328	40	0.86	0.86	548.94	12.64	25.87	6.84
	Operational and OP-Opt-Out	4,142	43	0.86	0.87	548.58	12.51	27.16	7.30
8	Operational	3,979	39	0.85	0.85	851.07	13.69	25.26	6.59
	Operational and OP-Opt-Out	3,954	44	0.86	0.86	851.08	13.46	28.00	7.17
High School	Operational	3,252	39	0.86	0.86	942.70	13.56	24.07	6.87
	Operational and OP-Opt-Out	3,248	44	0.88	0.88	942.59	13.16	27.26	7.88

Table 11. Adjusted *P* Value Means and Standard Deviations, and Coefficient Alpha Reliability for Strand Scores

Grade	Content Strand		Operational				Operational and OP-Opt-Out			
			Number of Items	Mean Adj. <i>P</i> Value	Adj. <i>P</i> Value STD	Reliability	Number of Items	Mean Adj. <i>P</i> Value	Adj. <i>P</i> Value STD	Reliability
5	1	Communication and Emotional Health	7	0.78	0.13	0.70	7	0.78	0.13	0.70
	2	Safety Skills	5	0.66	0.23	0.37	5	0.66	0.23	0.37
	3	Human Body and Personal Health	4	0.46	0.19	0.34	5	0.45	0.17	0.37
	4	Disease Prevention	4	0.68	0.25	0.38	6	0.66	0.22	0.49
	5	Nutrition	5	0.72	0.23	0.47	5	0.72	0.23	0.47
	6	Alcohol, Tobacco and Other Drugs	4	0.52	0.09	0.29	4	0.52	0.09	0.29
	7	Health Decision Making	6	0.59	0.21	0.46	6	0.59	0.21	0.46
	8	Physical Education	5	0.65	0.20	0.43	5	0.65	0.20	0.43
8	1	Communication and Emotional Health	6	0.76	0.11	0.55	6	0.76	0.11	0.55
	2	Safety Skills and Community Health	5	0.66	0.22	0.24	5	0.66	0.22	0.24
	3	Human Development and Sexuality	0	—	—	—	5	0.59	0.23	0.48
	4	Disease Prevention	7	0.71	0.20	0.55	7	0.71	0.20	0.55
	5	Nutrition	6	0.50	0.29	0.26	6	0.50	0.29	0.26
	6	Alcohol, Tobacco and Other Drugs	5	0.64	0.15	0.44	5	0.64	0.15	0.44
	7	Health Information and Advocacy	5	0.71	0.06	0.65	5	0.71	0.06	0.65
	8	Physical Education	5	0.55	0.21	0.44	5	0.55	0.21	0.44

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

Table 11. Adjusted P Value Means and Standard Deviations, and Coefficient Alpha Reliability for Strand Scores
(continued)

Grade	Content Strand		Operational				Operational and OP-Opt-Out			
			Number of Items	Mean Adj. P Value	Adj. P Value STD	Reliability	Number of Items	Mean Adj. P Value	Adj. P Value STD	Reliability
High School	1	Human Growth and Development	4	0.67	0.25	0.34	4	0.67	0.25	0.34
	2	Sexuality and Reproduction	0	—	—	—	5	0.73	0.18	0.80
	3	Disease Prevention and Treatment	9	0.60	0.16	0.63	9	0.60	0.16	0.63
	4	Nutrition	5	0.61	0.25	0.36	5	0.61	0.25	0.36
	5	Alcohol, Tobacco and Other Drugs	4	0.72	0.14	0.39	4	0.72	0.14	0.39
	6	Locate Health Information and Assistance	5	0.49	0.20	0.39	5	0.49	0.20	0.39
	7	Safety Skills	6	0.76	0.13	0.58	6	0.76	0.13	0.58
	8	Physical Education	6	0.53	0.17	0.45	6	0.53	0.17	0.45

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

Table 12. DC CAS 2013 Strand-to-Strand Correlations

Grade	Content Strand	Operational and OP-Opt-Out							
		Communication and Emotional Health	Safety Skills	Human Body and Personal Health	Disease Prevention	Nutrition	Alcohol, Tobacco and Other Drugs	Health Decision Making	Physical Education
5	Communication and Emotional Health	—	0.50	0.42	0.53	0.57	0.43	0.58	0.49
	Safety Skills	0.50	—	0.31	0.40	0.44	0.29	0.40	0.38
	Human Body and Personal Health	0.42	0.31	—	0.41	0.37	0.35	0.40	0.36
	Disease Prevention	0.53	0.40	0.41	—	0.44	0.34	0.44	0.40
	Nutrition	0.57	0.44	0.37	0.44	—	0.33	0.45	0.41
	Alcohol, Tobacco and Other Drugs	0.43	0.29	0.35	0.34	0.33	—	0.39	0.33
	Health Decision Making	0.58	0.40	0.40	0.44	0.45	0.39	—	0.43
	Physical Education	0.49	0.38	0.36	0.40	0.41	0.33	0.43	—
	Total Raw Score	0.83	0.66	0.65	0.72	0.71	0.60	0.74	0.67

Table 12. DC CAS 2013 Strand-to-Strand Correlations (continued)

		Operational and OP-Opt-Out							
Grade	Content Strand	Communication and Emotional Health	Safety Skills and Community Health	Human Development and Sexuality	Disease Prevention	Nutrition	Alcohol, Tobacco and Other Drugs	Health Information and Advocacy	Physical Education
8	Communication and Emotional Health	—	0.37	0.28	0.53	0.36	0.47	0.58	0.42
	Safety Skills and Community Health	0.37	—	0.22	0.39	0.25	0.33	0.36	0.27
	Human Development and Sexuality	0.28	0.22	—	0.31	0.20	0.26	0.30	0.24
	Disease Prevention	0.53	0.39	0.31	—	0.41	0.52	0.57	0.49
	Nutrition	0.36	0.25	0.20	0.41	—	0.36	0.41	0.37
	Alcohol, Tobacco and Other Drugs	0.47	0.33	0.26	0.52	0.36	—	0.53	0.44
	Health Information and Advocacy	0.58	0.36	0.30	0.57	0.41	0.53	—	0.49
	Physical Education	0.42	0.27	0.24	0.49	0.37	0.44	0.49	—
	Total Raw Score	0.75	0.56	0.52	0.79	0.61	0.72	0.79	0.69

Table 12. DC CAS 2013 Strand-to-Strand Correlations (continued)

Operational and OP-Opt-Out									
Grade	Content Strand	Human Growth and Development	Sexuality and Reproduction	Disease Prevention and Treatment	Nutrition	Alcohol, Tobacco and Other Drugs	Locate Health Information and Assistance	Safety Skills	Physical Education
High School	Human Growth and Development	—	0.34	0.50	0.41	0.40	0.38	0.48	0.40
	Sexuality and Reproduction	0.34	—	0.41	0.35	0.37	0.33	0.40	0.33
	Disease Prevention and Treatment	0.50	0.41	—	0.48	0.48	0.48	0.53	0.54
	Nutrition	0.41	0.35	0.48	—	0.44	0.38	0.49	0.38
	Alcohol, Tobacco and Other Drugs	0.40	0.37	0.48	0.44	—	0.39	0.51	0.39
	Locate Health Information and Assistance	0.38	0.33	0.48	0.38	0.39	—	0.41	0.41
	Safety Skills	0.48	0.40	0.53	0.49	0.51	0.41	—	0.45
	Physical Education	0.40	0.33	0.54	0.38	0.39	0.41	0.45	—
	Total Raw Score	0.66	0.65	0.82	0.68	0.68	0.66	0.75	0.70

Table 13. DC CAS 2013 Number Correct to Scale Score Conversions with Associated Standard Errors of Measurement (SEM) Operational and OP-Opt-Out

Raw Score	Grade 5		Grade 8		High School	
	Scale Score	SEM	Scale Score	SEM	Scale Score	SEM
0	500	33	800	33	900	28
1	500	33	800	33	900	28
2	500	33	800	33	900	28
3	500	33	800	33	900	28
4	500	33	800	33	900	28
5	500	33	800	33	900	28
6	500	33	800	33	900	28
7	500	33	800	33	900	28
8	500	33	800	33	900	28
9	500	33	800	33	900	28
10	516	17	806	27	909	19
11	522	11	815	18	915	13
12	525	8	820	12	919	9
13	527	6	824	10	922	7
14	529	5	827	8	924	6
15	531	5	830	7	926	5
16	533	5	832	6	927	5
17	534	4	834	6	929	4
18	536	4	836	5	930	4
19	537	4	837	5	931	4
20	538	4	839	5	933	4
21	540	4	841	5	934	4
22	541	4	842	4	935	4
23	542	4	843	4	937	4
24	544	4	845	4	938	4
25	545	4	846	4	939	4
26	546	4	848	4	940	4
27	548	4	849	4	942	4
28	549	4	851	4	943	4
29	551	4	852	4	944	4
30	552	4	854	4	946	4
31	554	4	855	4	947	4
32	556	4	857	5	949	4
33	557	4	859	5	950	4
34	559	4	861	5	952	4
35	561	4	863	5	954	5
36	563	5	865	5	956	5
37	566	5	868	5	958	5
38	568	5	870	4	961	5
39	571	6	872	5	963	6
40	575	7	875	5	967	6
41	580	8	879	7	971	7
42	589	12	884	8	977	11
43	599	19	893	14	990	19
44	.	.	899	19	999	26

Table 14. DC CAS 2013 Number Correct to Scale Score Conversions with Associated Standard Errors of Measurement (SEM) Operational

Raw Score	Grade 5		Grade 8		High School	
	Scale Score	SEM	Scale Score	SEM	Scale Score	SEM
0	500	33	800	34	900	28
1	500	33	800	34	900	28
2	500	33	800	34	900	28
3	500	33	800	34	900	28
4	500	33	800	34	900	28
5	500	33	800	34	900	28
6	500	33	800	34	900	28
7	500	33	800	34	900	28
8	500	33	800	34	900	28
9	511	22	809	25	911	17
10	520	13	817	16	917	11
11	524	9	822	11	920	8
12	527	7	826	9	923	6
13	529	6	829	8	925	5
14	531	5	831	7	926	5
15	532	5	834	6	928	5
16	534	4	836	5	930	4
17	535	4	837	5	931	4
18	537	4	839	5	932	4
19	538	4	841	5	934	4
20	540	4	842	4	935	4
21	541	4	844	4	937	4
22	542	4	846	4	938	4
23	544	4	847	4	940	4
24	545	4	849	4	942	4
25	547	4	850	4	943	5
26	548	4	852	4	945	5
27	550	4	853	4	947	5
28	552	4	855	4	949	5
29	553	4	857	5	951	5
30	555	4	859	5	953	5
31	557	4	861	5	955	5
32	559	5	864	5	958	5
33	561	5	867	5	960	5
34	564	5	869	5	963	6
35	566	5	872	5	966	6
36	570	6	876	6	971	7
37	573	7	882	8	977	11
38	579	8	890	11	990	19
39	588	12	899	20	999	26
40	599	20

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.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334.
- CTB/McGraw-Hill. (2012). *District of Columbia Comprehensive Assessment System (DC CAS) health and physical education assessment for grades 5, 8, and high school*. Monterey, CA: Author.
- CTB/McGraw-Hill. (2013). *District of Columbia Comprehensive Assessment System (DC CAS) test chairperson's manual: Reading and mathematics, composition, science, biology, and health and physical education*. Monterey, CA: Author.
- CTB/McGraw-Hill. (2013). *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, 8, and biology), and health (grades 5, 8, and high school)*. Monterey, CA: Author.
- OSSE, (2012). *Healthy Schools Act of 2010* (D.C. Law 18-209) Report. Retrieved from http://osse.dc.gov/sites/default/files/dc/sites/osse/publication/attachments/HSA%20Council%20Report%20FY12_health_physed.pdf
- 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.
- Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*, 201–210.
- Thiessen, D. (1982). Marginal maximum-likelihood estimation for the one-parameter logistic model. *Psychometrika*, *47*, 175–186.
- Yen, W.M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, *5*, 245–262.
- Zwick, R., Donoghue, J.R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*, *30*, 233–251.

Appendix A: Item Acceptability Checklist

Item Acceptability Checklist (all seven criteria must be met)		
	<u>Criteria of an Acceptable Item</u>	<u>Criteria for Automatic Item Rejection</u>
1	Item is clearly aligned to the intended standard/objective/node/subskill, does not require prior knowledge (unless required by target skill/spec), AND is grade-appropriate for content, skill, concept, and vocabulary/language.	The item does not align to the intended standard/objective/node/subskill OR the information needed to respond to item is not present in item or stimulus (e.g., item requires prior knowledge) OR item is not grade-appropriate in content or readability.
2	Multiple-choice items contain one and only one clear correct response, with strong, plausible distractors and clearly written and valid answer choice rationales for every answer choice.	There is no single correct response (no correct or multiple correct responses in multiple-choice items) OR the answer choice rationales are missing or contain significant errors.
3	Items/stimulus free of factual inaccuracies.	The item or stimulus contains <u>significant</u> factual errors.
4	Use of varied, creative concepts/ideas in items across the assignment, and Items does not clue other items in a set.	The concept/idea is overly measured in items submitted (e.g., same idea is present in other items/batch from vendor) OR item clues other items in a set.
5	Items use current, realistic contexts that engage the test taker and are free of content that could be considered sensitive or biased.	There is non-realistic context that could affect response to item OR there are bias and/or sensitivity issues inherent in the item.
6	Stimulus and item are well aligned, cohesive and relevant to each other, AND when stimulus is present and required by target skill, it must be referenced in order for item to be answered.	There is a <u>significant</u> mismatch between stimulus and item element, OR when stimulus is present and required by target skill, <u>but item can be answered without referencing the stimulus.</u>
7	Items that do not infringe on content copyrights, or plagiarize content, or rely on trademarks or pop- culture references, AND that include complete and accurate source documentation when required by the content of the item.	There is a copyright infringement, permissions violation, or plagiarism, OR item relies on trademarks or pop-culture references, OR the required source documentation is significantly incomplete, missing or deemed unreliable.

Appendix B: Health and PE Test Item Adjusted *P* Values

Table B1. DC CAS 2013 Operational and OP-Opt-Out Item Adjusted *P* Values, Grade 5

Operational Item Sequence Number	N	Max Points	Adjusted <i>p</i> value	Operational Item Sequence Number	N	Max Points	Adjusted <i>p</i> value
1	4,138	1	0.80	25	4,122	1	0.34
2	4,140	1	0.76	26	4,130	1	0.44
3	4,139	1	0.48	27	4,125	1	0.30
4	4,138	1	0.43	28	4,129	1	0.57
5	4,138	1	0.89	29	4,126	1	0.51
6	4,139	1	0.41	30	4,122	1	0.51
7	4,138	1	0.92	31	4,123	1	0.53
8	4,138	1	0.57	32	4,120	1	0.85
9	4,136	1	0.54	33	4,122	1	0.83
10	4,141	1	0.79	34	4,119	1	0.44
11	4,140	1	0.81	35	4,118	1	0.70
12	4,138	1	0.86	36	4,118	1	0.85
13	4,139	1	0.91	37	4,117	1	0.75
14	4,139	1	0.82	38	4,110	1	0.66
15	4,141	1	0.91	39	4,109	1	0.59
16	4,132	1	0.66	40	4,112	1	0.31
17	4,139	1	0.93	41*	3,614	1	0.47
18	4,138	1	0.40	42*	3,611	1	0.74
19	4,141	1	0.81	43*	3,602	1	0.41
20	4,134	1	0.31				
21	4,134	1	0.40				
22	4,131	1	0.72				
23	4,133	1	0.89				
24	4,129	1	0.63				

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

* Opt-Out/Sex-Ed items

Table B2. DC CAS 2013 Operational and OP-Opt-Out Item Adjusted *P* Values, Grade 8

Operational Item Sequence Number	N	Max Points	Adjusted <i>p</i> value		Operational Item Sequence Number	N	Max Points	Adjusted <i>p</i> value
1	3,938	1	0.12		25	3,916	1	0.26
2	3,950	1	0.69		26	3,923	1	0.45
3	3,954	1	0.90		27	3,923	1	0.82
4	3,953	1	0.76		28	3,922	1	0.42
5	3,950	1	0.67		29	3,924	1	0.62
6	3,951	1	0.86		30	3,921	1	0.48
7	3,951	1	0.79		31	3,923	1	0.91
8	3,947	1	0.92		32	3,909	1	0.34
9	.	1	.		33	3,924	1	0.87
10	3,949	1	0.85		34	3,923	1	0.62
11	3,947	1	0.73		35	3,916	1	0.67
12	3,950	1	0.73		36	3,922	1	0.60
13	3,945	1	0.76		37	3,916	1	0.65
14	3,947	1	0.87		38	3,914	1	0.45
15	3,948	1	0.82		39	3,913	1	0.57
16	3,948	1	0.85		40	3,917	1	0.64
17	3,948	1	0.76		*41	3,702	1	0.68
18	3,947	1	0.60		*42	3,692	1	0.24
19	3,939	1	0.67		*43	3,685	1	0.64
20	3,946	1	0.46		*44	3,696	1	0.85
21	3,942	1	0.50		*45	3,695	1	0.52
22	3,941	1	0.24					
23	3,943	1	0.76					
24	3,944	1	0.67					

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

* Opt-Out/Sex-Ed items

Table B3. DC CAS 2013 Operational and OP-Opt-Out Item Adjusted *P* Values, High School

Operational Item Sequence Number	N	Max Points	Adjusted <i>p</i> value		Operational Item Sequence Number	N	Max Points	Adjusted <i>p</i> value
1	3,244	1	0.76		25	.	1	.
2	3,240	1	0.41		26	3,219	1	0.89
3	3,231	1	0.28		27	3,211	1	0.37
4	3,243	1	0.44		28	3,216	1	0.86
5	3,243	1	0.70		29	3,194	1	0.51
6	3,239	1	0.42		30	3,208	1	0.43
7	3,245	1	0.77		31	3,216	1	0.92
8	3,237	1	0.29		32	3,206	1	0.75
9	3,241	1	0.75		33	3,215	1	0.92
10	3,243	1	0.81		34	3,212	1	0.80
11	3,241	1	0.66		35	3,211	1	0.53
12	3,240	1	0.78		36	3,213	1	0.88
13	3,235	1	0.57		37	3,210	1	0.70
14	3,239	1	0.59		38	3,206	1	0.66
15	3,236	1	0.46		39	3,207	1	0.41
16	3,237	1	0.52		40	3,208	1	0.83
17	3,239	1	0.85		*41	2,854	1	0.88
18	3,233	1	0.45		*42	2,853	1	0.93
19	3,236	1	0.59		*43	2,851	1	0.65
20	3,234	1	0.66		*44	2,848	1	0.67
21	3,233	1	0.30		*45	2,847	1	0.50
22	3,235	1	0.36					
23	3,229	1	0.65					
24	3,229	1	0.70					

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

* Opt-Out/Sex-Ed items

Table B4. DC CAS 2013 Field Test and FT-Opt-Out Item Adjusted *P* Values, Grade 5

Field Test Item Sequence Number	N	Max Points	Adjusted <i>p</i> value
1	4,135	1	0.27
2	4,137	1	0.35
3	4,138	1	0.34
4	4,138	1	0.85
5	4,132	1	0.33
6	4,132	1	0.63
7	4,130	1	0.88
8	4,124	1	0.91
9	4,110	1	0.60
10	4,110	1	0.91
11	4,104	1	0.25
12	4,101	1	0.43

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

* There are no Opt-Out/Sex-Ed items in Grade 5.

Table B5. DC CAS 2013 Field Test and FT-Opt-Out Item Adjusted P Values, Grade 8

Field Test Item Sequence Number	N	Max Points	Adjusted p value
1	3,942	1	0.52
2	3,948	1	0.94
3	3,947	1	0.86
4	3,943	1	0.72
5	3,945	1	0.69
6	3,946	1	0.18
7	3,924	1	0.85
8	3,919	1	0.51
9	3,918	1	0.19
10	3,914	1	0.63
11	3,915	1	0.34
12	3,912	1	0.26
13	3,909	1	0.87
*14	3,696	1	0.79
*15	3,690	1	0.75

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

* Opt-Out/Sex-Ed items

Table B6. DC CAS 2013 Field Test and FT-Opt-Out Item Adjusted *P* Values, High School

Field Test Item Sequence Number	N	Max Points	Adjusted <i>p</i> value
1	3,234	1	0.72
2	3,244	1	0.47
3	3,226	1	0.48
4	3,233	1	0.41
5	3,208	1	0.53
6	3,208	1	0.83
7	3,205	1	0.86
8	3,208	1	0.77
9	3,207	1	0.58
10	3,206	1	0.14
11	3,209	1	0.43
12	3,203	1	0.71
13	3,206	1	0.58
14	3,203	1	0.53
*15	2,844	1	0.83

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

* Opt-Out/Sex-Ed items