

**Technical Report
Spring 2012 Test Administration**

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

January 11, 2013



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

Developed and published under contract with the 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 © 2012 by the District of Columbia Office of the State Superintendent of Education. 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.

Table of Contents

List of Tables.....4

Section 1. Overview5

Section 2. Item and Test Development6

Overview6

Content Standards and Item Development6

Test Development.....7

Test Design.....7

Section 3. Test Administration Guidelines and Requirements10

Overview10

Guidelines and Requirements for Administering DC CAS.....10

Materials Orders, Delivery, and Retrieval.....11

Secure Inventory.....11

Section 4. Student Participation.....13

Tests Administered.....13

Participation in DC CAS13

Definition of Valid Test Administration13

Participation Rates.....13

Special Accommodation.....14

Section 5. Methods.....17

Classical Item Level Analyses17

Item Bias Analyses.....17

Calibration and Equating.....18

Goodness of Fit18

Establishing Upper and Lower Bounds for the Grade Level Scales19

Reliability Coefficients.....20

Standard Errors of Measurement.....20

Section 6. Evidence for Reliability and Validity21

Reliability21

Validity.....21

Item Level Evidence.....21

Classical Item Statistics.....21

Differential Item Function.....22

Test and Strand Level Evidence22

Total Test Scores22

Strand Level Scores22

Standard Errors of Measurement.....22

References32

Appendix A: Checklist for DC Educator Review of DC CAS Items33

Appendix B: Health and PE Test Item Adjusted *P* Values35

List of Tables

Table 1. DC CAS 2012 Operational Test Form Blueprints: Health and PE8

Table 2. Number and Percent of Examinees with Valid Health and PE Test Administrations and Responding to Opt-Out Items.....15

Table 3. Number and Percent of Examinees with Valid Health and PE Test Administrations across Subgroups15

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

Table 5. DC CAS 2012 Classical Item Level Statistics23

Table 6. Numbers of Operational and Opt-Out Items Flagged for DIF Using the.....24
Mantel-Haenszel Procedure24

Table 7. Total Test Scale and Raw Score Means and Reliability Statistics25

Table 8. Coefficient Alpha Reliability for Strand Scores26

Table 9. DC CAS 2012 Strand-to-Strand Correlations28

Table 10. DC CAS 2012 Number Correct to Scale Score Conversions with Associated Standard Errors of Measurement (SEM) Operational31

Table B1. DC CAS 2012 Operational Form Item Adjusted *P* Values, Grade 535

Table B2. DC CAS 2012 Operational Form Item Adjusted *P* Values, Grade 836

Table B3. DC CAS 2012 Operational Form Item Adjusted *P* Values, High School37

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 2012 to assess students' skills in Grades 5, 8, and High School Health and Physical Education. Scores from these assessments were not reported at an individual student level in 2012. This technical report is written to document procedures and results from developing, analyzing, and validating the 2012 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 2012 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

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 the test's development, we establish evidence of validity for the DC CAS Health and PE assessment 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.

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 in April 2012. Each assessment contained 50 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 DC CAS Health and PE tests are designed as an operational/field test, since 2012 is the first year of administration and in which test results are required. In this way, the newly developed items were field tested and reviewed for statistical quality, and only those items that were of acceptable quality and that collectively met the blueprints contributed to student scores.

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 2012 Operational Test Form Blueprints: Health and PE

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

Table 1. DC CAS 2012 Operational Test Form Blueprints: Health and PE (continued)

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

Section 3. Test Administration Guidelines and Requirements

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*. 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 (e.g., reading only Mathematics test questions to examinees) and identifies accommodations that are considered modifications to the test, which will result in invalidated test scores (e.g., assisted reading of Reading passages). The *Test Chairperson's Manual* 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 (e.g., reading of [only] Mathematics test questions), and response accommodations (e.g., dictating responses). The *Test Chairperson's Manual* and *Test Directions* specify accommodations approved for English language learners; 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 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 for all of DC CAS content areas were accounted for; 212 secure test booklets were missing for the 2012 administration, compared with 103 test booklets missing in 2011.

Section 4. Student Participation

Tests Administered

All public schools in the District of Columbia administered the DC CAS tests between April 17 and April 27, 2012.

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 Grade 5, 4,100 in Grade 8, and about 2,700 in High School. 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, where valid attempt on the test is defined as:

- At least one item marked with a correct response OR
- At least 5 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 9% (Grade 5), 2% (Grade 8), and 2% (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 and Percent of Examinees with Valid Health and PE Test Administrations and Responding to Opt-Out Items

| Grade | Students with Test Scores | Students Responding to “Opt-Out” Items | Percentage of Students Who Chose to “Opt Out” |
|-------------|---------------------------|--|---|
| 5 | 4,560 | 4,135 | 9% |
| 8 | 4,122 | 4,042 | 2% |
| High School | 2,704 | 2,640 | 2% |

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 | 4,560 | 2,300 | 50% | 2,230 | 49% | 81 | 2% | 3,501 | 77% | 582 | 13% | 356 | 8% |
| 8 | 4,122 | 2,030 | 49% | 2,051 | 50% | 56 | 1% | 3,293 | 80% | 479 | 12% | 227 | 6% |
| High School | 2,704 | 1,220 | 45% | 1,384 | 51% | 44 | 2% | 2,161 | 80% | 271 | 10% | 128 | 5% |

| Grade | Students with Test Scores | Special Education | | English Language Learner | | Section 504 | | Title I Targeted | | Home Schooling | |
|-------------|---------------------------|-------------------|-----|--------------------------|----|-------------|----|------------------|----|----------------|----|
| | | N | % | N | % | N | % | N | % | N | % |
| 5 | 4,560 | 554 | 12% | 190 | 4% | 33 | 1% | 203 | 4% | 0 | 0% |
| 8 | 4,122 | 481 | 12% | 222 | 5% | 24 | 1% | 138 | 3% | 1 | 0% |
| High School | 2,704 | 272 | 10% | 103 | 4% | 7 | 0% | 97 | 4% | 1 | 0% |

*Note that the percentages may not sum to 100% given 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 | % |
| 5 | 4,560 | 171 | 4% | 112 | 2% | 174 | 4% | 0 | 0% |
| 8 | 4,122 | 147 | 4% | 129 | 3% | 164 | 4% | 1 | 0% |
| High School | 2,704 | 91 | 3% | 72 | 3% | 93 | 3% | 0 | 0% |

| Grade | Students with Test Scores | Timing/Scheduling | | Setting | | Presentation | | Response | | Other | | Students with Special Education Code | |
|-------------|---------------------------|-------------------|-----|---------|-----|--------------|-----|----------|-----|-------|----|--------------------------------------|-----|
| | | N | % | N | % | N | % | N | % | N | % | N | % |
| 5 | 4,560 | 601 | 13% | 611 | 13% | 561 | 12% | 302 | 7% | 15 | 0% | 554 | 12% |
| 8 | 4,122 | 580 | 14% | 581 | 14% | 542 | 13% | 402 | 10% | 7 | 0% | 481 | 12% |
| High School | 2,704 | 241 | 9% | 245 | 9% | 185 | 7% | 130 | 5% | 4 | 0% | 272 | 10% |

| Grade | Students with Test Scores | Breaks | | Small Group and Individual Administrations | | Read or Translate Test Questions | | Responses Dictated | |
|-------------|---------------------------|--------|-----|--|-----|----------------------------------|-----|--------------------|----|
| | | N | % | N | % | N | % | N | % |
| 5 | 4,560 | 532 | 12% | 588 | 13% | 456 | 10% | 73 | 2% |
| 8 | 4,122 | 479 | 12% | 550 | 13% | 414 | 10% | 49 | 1% |
| High School | 2,704 | 211 | 8% | 226 | 8% | 109 | 4% | 19 | 1% |

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 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 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 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 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 ($Q1$) statistic that takes into account differing numbers of score levels as well as sample size:

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

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

Evidence of the validity of the scalings is provided by model fit. If the IRT model fits the empirical item response distributions for the population we want to generalize to (i.e., District of Columbia students), then the claim that the scores are valid indicators of an underlying proficiency is strengthened. Fit statistics 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 2 operational items were flagged for poor fit to the IRT model.

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.

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 5 summarizes classical item level statistics (adjusted p values, point biserial correlations, omit rates, and rates of items not reached) for Health and PE operational, opt-out, and field test items. On average, the operational collection of items on the tests was above average (0.50 p value) at 0.64 for Grades 5 and 8, and 0.62 for High School. Opt-out items and field test items are, on average, slightly more difficult for Grades 5 and 8, though less difficult for 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 opt-out items (flagged with an asterisk).

The point biserial is one measure of the correlation between each item and the overall test. The correlations are all highest for the operational items and, in Grades 5 and 8, opt-out item correlations are slightly lower than operational. However, the correlations are slightly higher for High School opt-out items.

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 are low and less than the 5% criteria. However, a larger percentage of students in the group taking the opt-out items actually omitted those very 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. Table 6 summarizes the 2012 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 7, 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 0.84.

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 8, highlights the strands that tend to be the 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 4 items and as many as 9, range between 0.45 and 0.78. As an additional measure of internal consistency, correlations have been produced between strands within each grade. These are provided in Table 9. 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 10 lists the number correct to scale score values, along with their associated IRT SEM values. 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 5. DC CAS 2012 Classical Item Level Statistics

| Grade | Item Type | Number of Items | Mean | Mean | Mean Omit Rate | Mean Not Reached Rate |
|-------------|-------------|-----------------|-------------------------|------------------------|----------------|-----------------------|
| | | | <i>Adjusted p value</i> | Item-Total Correlation | | |
| 5 | Operational | 40 | 0.64 | 0.35 | 0.21 | 0.13 |
| | Opt-Out | 3 | 0.57 | 0.29 | 13.12 | 13.08 |
| | Field Test | 7 | 0.34 | 0.01 | 0.25 | 0.12 |
| 8 | Operational | 40 | 0.64 | 0.33 | 0.23 | 0.16 |
| | Opt-Out | 5 | 0.58 | 0.19 | 6.76 | 6.64 |
| | Field Test | 5 | 0.48 | 0.07 | 0.14 | 0.07 |
| High School | Operational | 40 | 0.62 | 0.32 | 0.50 | 0.41 |
| | Opt-Out | 5 | 0.75 | 0.39 | 11.79 | 11.75 |
| | Field Test | 5 | 0.67 | 0.17 | 0.44 | 0.36 |

Note: Omit and not reached rates are percentages.

Table 6. Numbers of Operational and 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 | 36 | 6 | 1 | 0 | 0 | 0 |
| African American | Asian | N/A | N/A | N/A | N/A | N/A | 43 |
| | Hispanic | 37 | 2 | 2 | 0 | 2 | 0 |
| | White | 26 | 4 | 2 | 9 | 1 | 1 |
| Grade 8 (total 45 items) | | | | | | | |
| Male | Female | 36 | 2 | 3 | 3 | 1 | 0 |
| African American | Asian | N/A | N/A | N/A | N/A | N/A | 45 |
| | Hispanic | 35 | 5 | 3 | 0 | 2 | 0 |
| | White | 29 | 5 | 0 | 9 | 2 | 0 |
| High School (total 45 items) | | | | | | | |
| Male | Female | 37 | 4 | 1 | 2 | 1 | 0 |
| African American | Asian | N/A | N/A | N/A | N/A | N/A | 45 |
| | Hispanic | 32 | 4 | 5 | 2 | 2 | 0 |
| | White | 27 | 2 | 1 | 9 | 2 | 4 |

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 7. Total Test Scale and Raw Score Means and Reliability Statistics

| Grade | Item Type | Students with Test Scores | Number of Items | Alpha | Feldt-Raju | Criterion Score | |
|-------------|-------------------------|---------------------------|-----------------|-------|------------|-----------------|------|
| | | | | | | Mean | SD |
| 5 | Operational | 4,560 | 40 | 0.86 | 0.86 | 25.52 | 6.89 |
| | Operational and Opt-Out | 4,135 | 43 | 0.86 | 0.87 | 26.90 | 7.36 |
| 8 | Operational | 4,119 | 40 | 0.85 | 0.85 | 25.62 | 6.72 |
| | Operational and Opt-Out | 4,039 | 45 | 0.86 | 0.86 | 28.29 | 7.28 |
| High School | Operational | 2,697 | 40 | 0.84 | 0.85 | 24.55 | 6.64 |
| | Operational and Opt-Out | 2,633 | 45 | 0.87 | 0.87 | 27.84 | 7.64 |

Table 8. Adjusted *P* Value Means and Standard Deviations, and Coefficient Alpha Reliability for Strand Scores

| Grade | Content Strand | | Operational | | | | Operational and 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 | 0.77 | 0.13 | 0.70 | 7 | 0.77 | 0.13 | 0.70 |
| | 2 | Safety Skills | 5 | 0.66 | 0.21 | 0.39 | 5 | 0.66 | 0.21 | 0.39 |
| | 3 | Human Body and Personal Health | 4 | 0.45 | 0.18 | 0.25 | 5 | 0.44 | 0.16 | 0.30 |
| | 4 | Disease Prevention | 4 | 0.67 | 0.25 | 0.37 | 6 | 0.66 | 0.22 | 0.48 |
| | 5 | Nutrition | 5 | 0.70 | 0.23 | 0.46 | 5 | 0.71 | 0.23 | 0.46 |
| | 6 | Alcohol, Tobacco and Other Drugs | 4 | 0.52 | 0.08 | 0.32 | 4 | 0.52 | 0.08 | 0.33 |
| | 7 | Health Decision Making | 6 | 0.59 | 0.20 | 0.47 | 6 | 0.60 | 0.20 | 0.48 |
| | 8 | Physical Education | 5 | 0.63 | 0.20 | 0.43 | 5 | 0.64 | 0.20 | 0.44 |
| 8 | 1 | Communication and Emotional Health | 6 | 0.76 | 0.11 | 0.54 | 6 | 0.76 | 0.11 | 0.54 |
| | 2 | Safety Skills and Community Health | 5 | 0.68 | 0.21 | 0.28 | 5 | 0.68 | 0.21 | 0.28 |
| | 3 | Human Development and Sexuality | 0 | — | — | — | 5 | 0.58 | 0.24 | 0.46 |
| | 4 | Disease Prevention | 7 | 0.70 | 0.20 | 0.55 | 7 | 0.71 | 0.20 | 0.55 |
| | 5 | Nutrition | 6 | 0.50 | 0.29 | 0.32 | 6 | 0.50 | 0.29 | 0.32 |
| | 6 | Alcohol, Tobacco and Other Drugs | 5 | 0.64 | 0.16 | 0.42 | 5 | 0.64 | 0.16 | 0.42 |
| | 7 | Health Information and Advocacy | 5 | 0.70 | 0.06 | 0.66 | 5 | 0.70 | 0.06 | 0.66 |
| | 8 | Physical Education | 6 | 0.51 | 0.20 | 0.41 | 6 | 0.51 | 0.20 | 0.41 |

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

Table 8. Adjusted *P* Value Means and Standard Deviations, and Coefficient Alpha Reliability for Strand Scores
(continued)

| Grade | Content Strand | | Operational | | | | Operational and 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 |
| High School | 1 | Human Growth and Development | 4 | 0.67 | 0.26 | 0.31 | 4 | 0.67 | 0.26 | 0.31 |
| | 2 | Sexuality and Reproduction | 0 | — | — | — | 5 | 0.75 | 0.18 | 0.79 |
| | 3 | Disease Prevention and Treatment | 9 | 0.60 | 0.18 | 0.58 | 9 | 0.60 | 0.18 | 0.58 |
| | 4 | Nutrition | 5 | 0.62 | 0.26 | 0.33 | 5 | 0.63 | 0.26 | 0.33 |
| | 5 | Alcohol, Tobacco and Other Drugs | 4 | 0.72 | 0.14 | 0.37 | 4 | 0.72 | 0.14 | 0.37 |
| | 6 | Locate Health Information and Assistance | 6 | 0.46 | 0.20 | 0.38 | 6 | 0.46 | 0.20 | 0.38 |
| | 7 | Safety Skills | 6 | 0.78 | 0.13 | 0.54 | 6 | 0.78 | 0.13 | 0.55 |
| | 8 | Physical Education | 6 | 0.52 | 0.18 | 0.45 | 6 | 0.51 | 0.18 | 0.46 |

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

Table 9. DC CAS 2012 Strand-to-Strand Correlations

| Grade | Content Strand | Operational and 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 | — | 0.53 | 0.42 | 0.53 | 0.56 | 0.42 | 0.58 | 0.49 |
| | Safety Skills | 0.53 | — | 0.33 | 0.40 | 0.42 | 0.31 | 0.42 | 0.36 |
| | Human Body and Personal Health | 0.42 | 0.33 | — | 0.41 | 0.37 | 0.33 | 0.37 | 0.35 |
| | Disease Prevention | 0.53 | 0.40 | 0.41 | — | 0.44 | 0.34 | 0.43 | 0.40 |
| | Nutrition | 0.56 | 0.42 | 0.37 | 0.44 | — | 0.31 | 0.45 | 0.42 |
| | Alcohol, Tobacco and Other Drugs | 0.42 | 0.31 | 0.33 | 0.34 | 0.31 | — | 0.42 | 0.35 |
| | Health Decision Making | 0.58 | 0.42 | 0.37 | 0.43 | 0.45 | 0.42 | — | 0.43 |
| | Physical Education | 0.49 | 0.36 | 0.35 | 0.40 | 0.42 | 0.35 | 0.43 | — |
| | Total Raw Score | 0.83 | 0.67 | 0.63 | 0.72 | 0.70 | 0.61 | 0.75 | 0.67 |

Table 9. DC CAS 2012 Strand-to-Strand Correlations (*continued*)

| 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 | — | 0.40 | 0.27 | 0.53 | 0.38 | 0.47 | 0.60 | 0.39 |
| | Safety Skills and Community Health | 0.40 | — | 0.22 | 0.41 | 0.29 | 0.35 | 0.41 | 0.29 |
| | Human Development and Sexuality | 0.27 | 0.22 | — | 0.30 | 0.23 | 0.26 | 0.29 | 0.23 |
| | Disease Prevention | 0.53 | 0.41 | 0.30 | — | 0.42 | 0.52 | 0.58 | 0.45 |
| | Nutrition | 0.38 | 0.29 | 0.23 | 0.42 | — | 0.36 | 0.42 | 0.34 |
| | Alcohol, Tobacco and Other Drugs | 0.47 | 0.35 | 0.26 | 0.52 | 0.36 | — | 0.53 | 0.41 |
| | Health Information and Advocacy | 0.60 | 0.41 | 0.29 | 0.58 | 0.42 | 0.53 | — | 0.45 |
| | Physical Education | 0.39 | 0.29 | 0.23 | 0.45 | 0.34 | 0.41 | 0.45 | — |
| | Total Raw Score | 0.75 | 0.60 | 0.51 | 0.79 | 0.62 | 0.71 | 0.80 | 0.66 |

Table 9. DC CAS 2012 Strand-to-Strand Correlations (continued)

| 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 | — | 0.33 | 0.47 | 0.36 | 0.39 | 0.39 | 0.45 | 0.41 |
| | Sexuality and Reproduction | 0.33 | — | 0.39 | 0.34 | 0.36 | 0.32 | 0.40 | 0.31 |
| | Disease Prevention and Treatment | 0.47 | 0.39 | — | 0.44 | 0.45 | 0.49 | 0.52 | 0.53 |
| | Nutrition | 0.36 | 0.34 | 0.44 | — | 0.36 | 0.36 | 0.47 | 0.38 |
| | Alcohol, Tobacco and Other Drugs | 0.39 | 0.36 | 0.45 | 0.36 | — | 0.37 | 0.49 | 0.39 |
| | Locate Health Information and Assistance | 0.39 | 0.32 | 0.49 | 0.36 | 0.37 | — | 0.40 | 0.43 |
| | Safety Skills | 0.45 | 0.40 | 0.52 | 0.47 | 0.49 | 0.40 | — | 0.44 |
| | Physical Education | 0.41 | 0.31 | 0.53 | 0.38 | 0.39 | 0.43 | 0.44 | — |
| | Total Raw Score | 0.65 | 0.64 | 0.81 | 0.65 | 0.66 | 0.68 | 0.74 | 0.71 |

Table 10. DC CAS 2012 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 | 23 | 809 | 25 | 909 | 19 |
| 10 | 520 | 13 | 817 | 16 | 915 | 12 |
| 11 | 524 | 9 | 822 | 11 | 919 | 9 |
| 12 | 527 | 7 | 826 | 9 | 922 | 7 |
| 13 | 529 | 6 | 829 | 7 | 924 | 6 |
| 14 | 531 | 5 | 831 | 6 | 926 | 5 |
| 15 | 532 | 5 | 833 | 6 | 927 | 5 |
| 16 | 534 | 4 | 835 | 5 | 929 | 4 |
| 17 | 535 | 4 | 837 | 5 | 931 | 4 |
| 18 | 537 | 4 | 839 | 5 | 932 | 4 |
| 19 | 538 | 4 | 840 | 5 | 934 | 4 |
| 20 | 540 | 4 | 842 | 4 | 935 | 4 |
| 21 | 541 | 4 | 844 | 4 | 937 | 4 |
| 22 | 542 | 4 | 845 | 4 | 938 | 4 |
| 23 | 544 | 4 | 847 | 4 | 940 | 4 |
| 24 | 545 | 4 | 848 | 4 | 941 | 4 |
| 25 | 547 | 4 | 850 | 4 | 943 | 4 |
| 26 | 548 | 4 | 851 | 4 | 945 | 5 |
| 27 | 550 | 4 | 853 | 4 | 946 | 5 |
| 28 | 552 | 4 | 854 | 4 | 948 | 5 |
| 29 | 553 | 4 | 856 | 5 | 950 | 5 |
| 30 | 555 | 4 | 858 | 5 | 952 | 5 |
| 31 | 557 | 4 | 860 | 5 | 954 | 5 |
| 32 | 559 | 5 | 863 | 5 | 956 | 5 |
| 33 | 561 | 5 | 865 | 5 | 958 | 5 |
| 34 | 564 | 5 | 868 | 5 | 961 | 5 |
| 35 | 566 | 5 | 871 | 5 | 963 | 5 |
| 36 | 569 | 6 | 874 | 6 | 966 | 6 |
| 37 | 573 | 7 | 878 | 6 | 970 | 8 |
| 38 | 579 | 8 | 883 | 7 | 977 | 11 |
| 39 | 588 | 13 | 890 | 10 | 990 | 19 |
| 40 | 599 | 20 | 899 | 19 | 999 | 25 |

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) test chairperson's manual: Reading and mathematics, composition, science, biology, and health and physical education*. Monterey, CA: Author.
- CTB/McGraw-Hill. (2012). *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 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.
- 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: 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
- 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

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: Health and PE Test Item Adjusted *P* Values

Table B1. DC CAS 2012 Operational Form 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 | 4,135 | 1 | 0.79 | 25 | 4,115 | 1 | 0.38 |
| 2 | 4,135 | 1 | 0.77 | 26 | 4,121 | 1 | 0.46 |
| 3 | 4,134 | 1 | 0.48 | 27 | 4,120 | 1 | 0.31 |
| 4 | 4,134 | 1 | 0.43 | 28 | 4,122 | 1 | 0.53 |
| 5 | 4,133 | 1 | 0.88 | 29 | 4,119 | 1 | 0.52 |
| 6 | 4,133 | 1 | 0.41 | 30 | 4,122 | 1 | 0.49 |
| 7 | 4,134 | 1 | 0.91 | 31 | 4,123 | 1 | 0.48 |
| 8 | 4,132 | 1 | 0.59 | 32 | 4,122 | 1 | 0.81 |
| 9 | 4,131 | 1 | 0.52 | 33 | 4,119 | 1 | 0.82 |
| 10 | 4,135 | 1 | 0.75 | 34 | 4,121 | 1 | 0.45 |
| 11 | 4,134 | 1 | 0.79 | 35 | 4,117 | 1 | 0.68 |
| 12 | 4,135 | 1 | 0.85 | 36 | 4,122 | 1 | 0.83 |
| 13 | 4,130 | 1 | 0.90 | 37 | 4,118 | 1 | 0.74 |
| 14 | 4,133 | 1 | 0.81 | 38 | 4,117 | 1 | 0.65 |
| 15 | 4,132 | 1 | 0.91 | 39 | 4,116 | 1 | 0.57 |
| 16 | 4,129 | 1 | 0.67 | 40 | 4,107 | 1 | 0.30 |
| 17 | 4,131 | 1 | 0.92 | 41* | 3,595 | 1 | 0.49 |
| 18 | 4,128 | 1 | 0.36 | 42* | 3,594 | 1 | 0.78 |
| 19 | 4,129 | 1 | 0.81 | 43* | 3,586 | 1 | 0.43 |
| 20 | 4,127 | 1 | 0.31 | | | | |
| 21 | 4,126 | 1 | 0.38 | | | | |
| 22 | 4,124 | 1 | 0.69 | | | | |
| 23 | 4,123 | 1 | 0.87 | | | | |
| 24 | 4,123 | 1 | 0.61 | | | | |

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 2012 Operational Form Item Adjusted P Values, Grade 8

| Operational Item Sequence Number | N | Max Points | Adjusted p value | | Operational Item Sequence Number | N | Max Points | Adjusted p value |
|----------------------------------|-------|------------|--------------------|--|----------------------------------|-------|------------|--------------------|
| 1 | 4,027 | 1 | 0.12 | | 25 | 4,022 | 1 | 0.27 |
| 2 | 4,037 | 1 | 0.68 | | 26 | 4,023 | 1 | 0.46 |
| 3 | 4,039 | 1 | 0.90 | | 27 | 4,024 | 1 | 0.79 |
| 4 | 4,038 | 1 | 0.76 | | 28 | 4,024 | 1 | 0.43 |
| 5 | 4,037 | 1 | 0.68 | | 29 | 4,026 | 1 | 0.63 |
| 6 | 4,037 | 1 | 0.85 | | 30 | 4,024 | 1 | 0.50 |
| 7 | 4,038 | 1 | 0.81 | | 31 | 4,023 | 1 | 0.91 |
| 8 | 4,036 | 1 | 0.92 | | 32 | 4,017 | 1 | 0.35 |
| 9 | 4,036 | 1 | 0.33 | | 33 | 4,024 | 1 | 0.87 |
| 10 | 4,037 | 1 | 0.84 | | 34 | 4,022 | 1 | 0.61 |
| 11 | 4,032 | 1 | 0.73 | | 35 | 4,014 | 1 | 0.63 |
| 12 | 4,037 | 1 | 0.72 | | 36 | 4,021 | 1 | 0.59 |
| 13 | 4,037 | 1 | 0.77 | | 37 | 4,021 | 1 | 0.65 |
| 14 | 4,038 | 1 | 0.88 | | 38 | 4,021 | 1 | 0.47 |
| 15 | 4,037 | 1 | 0.80 | | 39 | 4,020 | 1 | 0.57 |
| 16 | 4,038 | 1 | 0.85 | | 40 | 4,020 | 1 | 0.62 |
| 17 | 4,038 | 1 | 0.77 | | *41 | 3,770 | 1 | 0.70 |
| 18 | 4,037 | 1 | 0.64 | | *42 | 3,763 | 1 | 0.21 |
| 19 | 4,035 | 1 | 0.67 | | *43 | 3,761 | 1 | 0.66 |
| 20 | 4,036 | 1 | 0.44 | | *44 | 3,767 | 1 | 0.84 |
| 21 | 4,034 | 1 | 0.49 | | *45 | 3,767 | 1 | 0.50 |
| 22 | 4,015 | 1 | 0.24 | | | | | |
| 23 | 4,023 | 1 | 0.73 | | | | | |
| 24 | 4,025 | 1 | 0.64 | | | | | |

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 2012 Operational Form 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 | 2,633 | 1 | 0.74 | | 25 | 2,606 | 1 | 0.30 |
| 2 | 2,629 | 1 | 0.42 | | 26 | 2,608 | 1 | 0.89 |
| 3 | 2,627 | 1 | 0.28 | | 27 | 2,606 | 1 | 0.32 |
| 4 | 2,626 | 1 | 0.42 | | 28 | 2,609 | 1 | 0.88 |
| 5 | 2,630 | 1 | 0.72 | | 29 | 2,606 | 1 | 0.53 |
| 6 | 2,629 | 1 | 0.43 | | 30 | 2,609 | 1 | 0.41 |
| 7 | 2,631 | 1 | 0.78 | | 31 | 2,608 | 1 | 0.93 |
| 8 | 2,627 | 1 | 0.28 | | 32 | 2,602 | 1 | 0.76 |
| 9 | 2,633 | 1 | 0.78 | | 33 | 2,608 | 1 | 0.93 |
| 10 | 2,633 | 1 | 0.81 | | 34 | 2,608 | 1 | 0.82 |
| 11 | 2,631 | 1 | 0.68 | | 35 | 2,607 | 1 | 0.55 |
| 12 | 2,631 | 1 | 0.77 | | 36 | 2,607 | 1 | 0.90 |
| 13 | 2,628 | 1 | 0.57 | | 37 | 2,608 | 1 | 0.67 |
| 14 | 2,631 | 1 | 0.61 | | 38 | 2,607 | 1 | 0.68 |
| 15 | 2,630 | 1 | 0.45 | | 39 | 2,608 | 1 | 0.38 |
| 16 | 2,627 | 1 | 0.47 | | 40 | 2,607 | 1 | 0.85 |
| 17 | 2,629 | 1 | 0.88 | | *41 | 2,327 | 1 | 0.90 |
| 18 | 2,629 | 1 | 0.44 | | *42 | 2,325 | 1 | 0.94 |
| 19 | 2,628 | 1 | 0.58 | | *43 | 2,320 | 1 | 0.69 |
| 20 | 2,628 | 1 | 0.66 | | *44 | 2,318 | 1 | 0.72 |
| 21 | 2,627 | 1 | 0.29 | | *45 | 2,319 | 1 | 0.49 |
| 22 | 2,627 | 1 | 0.38 | | | | | |
| 23 | 2,609 | 1 | 0.66 | | | | | |
| 24 | 2,611 | 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

