Maryland School Assessment (MSA)
Science

Grades 5 and 8

Technical Report
2016 Operational Test

August 2016

Pearson

Table of Contents

List of Tables

List of Figures

Table of Appendices

Test Overview and Design

Introduction

The Maryland School Assessment (MSA) tests are measures of students' knowledge relative to the Maryland State Curriculum at grades 5 and 8. The MSA Science test was originally added to established assessments in Reading and Mathematics to form part of the MSA program. Administered annually in the spring, the MSA program was established to meet the requirements of the No Child Left Behind Act (NCLB) of 2001. Though it should be noted that Maryland adopted the Next Generation Science Standards (NGSS) in 2013 and is currently within phase 3 of implementation (MSDE, 2016). In 2015, Pearson was contracted by Maryland State Department of Education (MSDE) to develop, administer, and maintain the MSA Science test. This report provides technical details of work accomplished during the 2015-2016 test administration cycle.

Purpose

The purpose of this MSA Technical Report is to provide objective information regarding technical aspects of the 2016 MSA Science operational test. This volume is intended to be one source of information to Maryland K-12 educational stakeholders (including testing coordinators, educators, parents, and other interested citizens) about the development, implementation, scoring, and technical attributes of the MSA Science tests. Other sources of information regarding the MSA Science test, provided in paper or online format, include the MSA Science administration manual, implementation materials, and training materials.

The information provided here fulfills professional and scientific guidelines for technical reports of large scale educational assessments and is intended for use by qualified users within schools who use and interpret the results of the MSA Science tests. Specifically, information was selected for inclusion in this report based on NCLB requirements and standards from the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014).

This manual provides information about the MSA Science test regarding:

1. Content of the tests;
2. Test form design;
3. Identification of ineffective items;
4. Reliability of the tests;
5. Difficulty of the test questions;
6. Equating of test forms;
7. Detection of item bias;
8. Scoring and reporting the results of the tests.

From test development to final reporting, each of these facets of the MSA Science test contributes to the validity of the inferences made about the test results. This technical manual addresses these topics for the 2015-2016 testing year.

Test Overview

In 2002, the Maryland State Department of Education adopted the testing program known as the Maryland School Assessment (MSA). The first two subjects to be established under this new testing program were Reading and Mathematics. The Science test was added and the first field

test administration was conducted in the spring of 2007, followed by the first operational test in 2008. The MSA Science test is currently given to grade 5 and grade 8 students in order to assess achievement in Science. Score reports are provided to parents and include total test scale score results and performance level classifications (described in more detail in following sections).

Purpose and Use

By assessing student achievement against the Science academic standards, the MSA Science test serves two important purposes. First, the MSA Science test provides an accountability tool that measures performance levels of students, schools, and districts against the Science academic standards. Second, it provides parents, teachers, and educators with critical information about what students have learned, which, if applied constructively, can foster improvement of instructional programs, classroom education, and school performance.

Test Content, Specifications and Design

The MSA Science test was designed to align to the Maryland State Curriculum (MSC) that specifies curricular indicators and objectives that contributed directly to measuring content standards. The MSC is formatted so that content standards delineate broad, measurable statements about what students should know and be able to do. Each standard has multiple indicator statements that provide the next level of specificity, thereby narrowing the focus for teachers further. Finally, objectives provide teachers with very clear information about what specific learning should occur. The MSC is widely disseminated to Maryland educational stakeholders, including teachers, central office staff, students, parents and other stakeholders.

In order to ensure that MSDE is in accordance with the federal law that requires states to align their tests to their content standards, the MSC serves as the guiding document for test development and design. Developing the items for testing was a collaborative effort between MSDE, educators, and Pearson. Teachers, administrators, and content specialists were recruited from all over Maryland for several test development committees. These committees reviewed items developed for MSA Science test.

The basic test specifications were established by MSDE and provided to Pearson to guide the test development and administration. Since the inception of the Science test, there have been nine test administrations—a census field test in 2007 and nine operational tests (2008 through 2016). All administrations were conducted under the same testing conditions. Accordingly, the field test was designed to match the requirements of the operational administration test blueprint, i.e., a student taking the census field test and the operational test would respond to the same number and type of items. However, because of embedding of field test items on the operational form, there were fewer scored items on the operational form, even with the same number of overall items. Beginning with the 2008 operational test, two base forms (i.e., two forms of scored operational items) were used. Each form had a total of 77 items on the grade 5 form and 75 items on the grade 8 form. Grade 5 tests had 66 operational items and 11 field test items. The grade 8 test had 64 operational items with 11 field test items. For both grade tests, only operational items contributed to student scores. The two base forms share a set of 20 common items. These common items are discrete (i.e., non-passage based, stand-alone) selected response (SR) items.

MSA Science Item Types

The 2016 operational MSA Science included two types of items: selected response (SR) and brief constructed response (BCR). SR items require students to select a correct answer from several alternatives. For the 2016 MSA Science tests, students selected an answer from four options. Each SR item was scored dichotomously (i.e., 0 or 1). BCR items require students to provide a short answer using words, numbers, and/or symbols. All BCR items are scored using a generic

rubric and scores range from 0-3 based on concordant scores from two independent raters. In cases where the scores differ by one point, the higher score is used. In cases where the rater scores differ by two or more points, a third expert rater's independent score is used as a resolution.

In addition to these formats, a new item type was administered at the end of the online operational tests. MSDE has been exploring the incorporation of technology enhanced (TE) items for a number of years as a means of potentially measuring more complex skills in line with steps towards NGSS Assessment. TE items make use of the interactive capacity of computers to allow for enhanced presentation and capture of stimuli and responses. They can range from the simple (i.e. drag-and-drop, hot spot, etc.) to fully interactive multi-step scenario based formats.

Given that MSA Science is currently administered both online and on paper it was important to ensure that inclusion of the TE items was handled in such a way that year-to-year score comparability was preserved. This was addressed by administering a single TE item at the end of the online forms. Additionally, the TE items used were comparable in terms of seat time to complete and complexity to existing SR items.

MSA Science Test Blueprints

There are two MSA Science test blueprints available, one for grade 5 and one for grade 8 and there are six standards assessed across each grade with 66 items in the grade 5 test and 64 items in the grade 8 test, as presented in Tables 1 and 2.

Table 1. Grade 5 MSA Science Standards Assessed

	Standard
1.0	Skills and Processes
2.0	Earth/Space Science
3.0	Life Science
4.0	Chemistry
5.0	Physics
6.0	Environmental
	Total Number of items: 66
	Total number of points:72

Table 2. Grade 8 MSA Science Standards Assessed

	Standard
1.0	Skills and Processes
2.0	Earth/Space Science
3.0	Life Science
4.0	Chemistry
5.0	Physics
6.0	Environmental
	Total Number of items: 64
	Total number of points: 72

MSA Science 2016 Operational Test Construction

The 2016 operational tests were created according to the test blueprints (see Table 1 and 2) and reflective of the Maryland State Curriculum for Science in the form of measureable Indicators

and Objectives. As such, each of the two operational forms yielding student scores has the same test composition as that of the 2008 tests in terms of content, total number of items/score points, and item types. Additionally, each operational form was created with embedded field test items (see MSA Science 2016 Field Test Design). As Maryland is currently transitioning to NGSS, 2016 marks the last administration of this MSA Science assessment targeted to the MSC. As such the field test items were not analyzed for future use and served as placeholders to help ensure year-to-year comparability.

As noted in the previous section, the two operational forms were created with a common set of 20 SR items. These items were chosen to reflect a miniature version of the overall operational tests and provide a mechanism for placing all operational items from both forms onto a common scale.

The process of selecting items for the two 2016 MSA Science operational test forms was an iterative process primarily involving Pearson content experts, MSDE, and Pearson psychometricians. Initial test forms were created to meet the respective blueprints, reflect the MSC measureable Indicators and Objectives, and align with statistical characteristics of the 2008 operational tests. Only items deemed eligible after being administered live (field tested) and reviewed by content experts based on statistical indicators (see Data Review of the Field Test Items) were used. Additional content-related characteristics that were part of the creation of the operational test forms had to do with ensuring there was no cuing from one item to the next. That is, items were scrutinized to make sure nothing in any one question or passage would provide information relevant to answering any other item correctly.

Classical item statistics were used in conjunction with item response theory (IRT) statistics to help target the overall test forms. The guiding principles were choosing items with reasonably strong point biserial correlations (ideally >.30) and matching a spread of item difficulties in line with the 2008 forms. Items flagged for any reason based on the data review criteria (also including differential item functioning, as described later) were identified as such, and content experts were discouraged from using them. Item level statistical targets based on overall test, by standard, and by item type were also used for guidance. IRT test characteristic curves (TCCs), test information functions (TIFs), and conditional standard error of measurement (CSEM) plots for each test form were also compared to the respective 2008 plots to help ensure the overall IRT measurement properties were captured across the scale (see Test Analysis, Operational Scaling and Scoring).

This process of content and psychometric review and modification of each operational test form proceeded iteratively, where each group would evaluate the most recent proposed forms and provide feedback. Once operational test forms were created that best met all content and statistical targets, the proposed forms were submitted to MSDE for review and/or modification.

MSA Science 2016 Field Test Design

The 2016 field test design is premised on the design implemented throughout the life of MSA Science. However, since this year's test is anticipated to be the last administration as Maryland transitions to NGSS, the 2016 field test items served only as placeholders. That is, they provide a comparable testing experience to test takers so that scores can be directly compared to previous performance.

Field test forms were composed of selected response (SR) items and brief constructed response (BCR). Items were either stand-alone (not linked to other items), linked to a lab set stimulus (e.g., technical graph or figure), or linked to a technical passage stimulus. One set of unique field test items was administered per core form.

MSDE and Pearson worked together to finalize the structure of the 2016 field test forms. At each grade, 2 field test forms were produced. The intent of the test building process was to have each form be parallel in terms of number of SR items, BCR items, and stimulus materials. In addition, the field test forms were designed to be proportionately reflective of the overall test blueprint in terms of content representation. Each of 2 forms per grade had the same number of SR and BCR items. In addition, a goal of item selection was to balance, to the extent possible, coverage of the standards across the 2 field test forms per grade. On a per form basis, initial item selections were conducted by Pearson and then shared with MSDE for review and approval.

The 2016 forms were spiraled at the student-level. Spiraling at the student-level supports the assumption that examinee groups responding to each test form are randomly equivalent; an assumption that will further strengthen the link across forms.

Operational Item Analysis and Equating

Testing Population

Maryland Students in grade 5 and 8 took the Science operational test as part of the MSA program. Mode of testing (whether a test is administered by paper or via online administration) was determined by each school. The number of students per form, including demographic breakdowns and accommodations for grade 5 and grade 8, appear in Table 3.

Table 3. Demographic Characteristics of Grades 5 and 8 Sample for Overall, Online, and Paper

	Grade			
	5		8	
	N	%	N	%
Mode of Administration				
Online	60067	92.60	55152	87.97
Paper	4800	7.40	7540	12.03
Form				
1	31058	47.88	32356	51.61
2	33809	52.12	30336	48.39
Gender				
Female	31786	49.00	30634	48.86
Male	33080	51.00	32058	51.14
Unknown	1	0.00	0	0.00
Ethnicity				
Hispanic/Latino	10048	15.49	8943	14.26
Non-Hispanic/Latino	54810	84.50	53725	85.70
Unknown	9	0.01	24	0.04
Race				
American Indian	168	0.31	153	0.29
Asian/Pacific Islander	4323	7.89	3837	7.28
African American	21704	39.59	21214	40.23
Native Hawaiian	111	0.20	60	0.11
White	25659	46.81	24991	47.40
Two or more races	2845	5.19	2449	4.64
Unknown	9	0.02	24	0.05
All	64867	100.00	62692	100.00

Distribution of Students across Forms

As described, MSA Science test forms are composed of a set of operational items and field test items. Ideally, each respective test form will be administered to randomly equivalent groups of students. This helps ensure that any item and test level statistics are more directly comparable. The administration of multiple test forms is commonly referred to as "spiraling." The MSA Science test forms were spiraled at the student level and within mode of administration so that there would be an even distribution of tests across forms. Table 4 presents number of students taking each test form by mode of administration at a given grade. Within-form overages (i.e.

Grade 5 online Form 2) reflect the inclusion of additional forms for special accommodations (i.e. read-aloud, audio presentation, etc.).

Table 4. Distribution of Forms by Grade

		Form	
		1	**2**
Grade 5	**Online**	28678	31389
	Paper	2380	2420
	Overall	31058	33809
Grade 8	**Online**	28489	26663
	Paper	3867	3673
	Overall	32356	30336

Key Check Analysis of Operational Test Data

Using preliminary data collected from the 2016 operational test (a minimum of 200 responses were required for each form by mode of administration), Pearson computed Classical Test Theory (CTT) statistics on all multiple choice items in order to screen for items with characteristics that could be associated with an item being scored with a wrong correct-answer key (mis-keyed). Any items identified during this process were presented to Pearson content specialists for review to ensure that items were keyed properly. Findings of key check analysis suggested that all operational MSA Science items were correctly keyed and all the items had CTT values with the criteria described below.

The key check analysis included the following CTT statistics:

- **P-Value:** proportion of students who answered the item correctly. An item's p-value shows how difficult the item was for the students who took the test.

- **Point-Biserial Correlation (Pt Bis):** describes the relationship between a student's performance on the item (correct or incorrect) and the student's performance on the subject area test form as a whole (number of correct items on the test form).

- **P-Value by Response Option:** These data indicate the proportion of students who selected each response option.

The following criteria were used to designate items as potentially mis-keyed:

- P-value < 0.15
- Point-biserial < 0.20
- P-value for a single unkeyed response >= .40

Data Analysis

Each functional group within Pearson followed complete quality control and quality assurance (QC&QA) steps before the data of student demographic and item responses were delivered to Pearson's Psychometric Services (PS) division. Pearson PS staff had primary responsibility for analyzing MSA Science data to ensure accuracy and validity of scoring.

The data analyses for this report were generated using SAS software, of the Version 9.2 of the SAS System for Windows. Copyright © 2002-2008 SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA. In addition, IRTPRO version 2.1 (Cai, Thissen, and du Toit, 2011) was used for the original 2016 IRT analyses. IRTPRO allows for estimation of IRT item parameters for both dichotomously and polytomous scored items. It has been thoroughly tested and is currently utilized by several high-stakes testing programs administered by Pearson.

Upon receiving the student data file, Pearson psychometric staff conducted their internal quality checks by verifying the MSA Science data and analysis process at several steps in the procedure. These steps included verification of the SAS programs prior to use on actual field data through review by a second member of the psychometric services staff. Additionally, the output from the item analysis programs were verified for out-of-range values and for consistent results across programs. All technical support and analyses were carried out in accordance with both the *Standards* (AERA, APA, & NCME, 2014) and the Pearson's quality control steps.

Classical Item Analysis

The following classical item statistics were calculated:

- P-value of SR items
- Mean of BCR items
- Point-Biserial Correlation
- Item Option Point-Biserial for SR items
- P-value by Item Option for SR items
- Item Score Distribution for BCR items

The results of the CTT item analyses were stored in the item. P-value and point-biserial statistics for the 2016 MSA operational items are reported in Appendix A.

IRT Calibration

The IRT calibration, equating, and scaling work was conducted during the original 2016 operational administration and described herein. ISE pattern scoring for 2016 student responses (described later) was conducted using the original 2016 IRT parameters.

Pearson used a concurrent calibration IRT estimation procedure for placing all 2016 Form A and Form B operational MSA Science items on a common theta scale that was then equated to the original 2007 base calibration (as described in the next section). The 3 parameter logistic (3-PL) model was used for SR items and the generalized partial credit (GPC) model was used for BCR items because of the mixed format of the test (i.e., multiple-choice and constructed response or polytomous items).

Dichotomous Item Response Theory Model

For the SR items, or dichotomously scored items, calibration was done using Birnbaum's 3-PL item response theory (IRT) model (Lord & Novick, 1968). The formulation of the 3-PL model is presented below:

$$P_i(\theta) = c_i + (1 - c_i)\frac{1}{1 + e^{-Da_i(\theta - b_i)}}, \quad (1)$$

where θ (theta) is the student proficiency parameter, a_i is the item discrimination parameter, b_i is the item difficulty parameter, c_i is the lower asymptote parameter and D is a scaling constant. The scaling constant is traditionally 1.7. With multiple-choice items it is assumed that, due to guessing, examinees with minimal proficiency have a probability greater than zero of responding correctly to an item. This probability is represented in the 3-PL model by the c_i parameter.

Polytomous Item Response Theory Model

For the BCR items, or polytomously scored items, calibration was done using the GPC model (Muraki, 1992). For an item j with m_j possible scores (0, 1, . . . , m_j-1), the GPC model gives the probability of response r as a function of latent variable θ as

$$\Pr(X_j = r \mid \theta) = \frac{e^{z_{jr}}}{1+\sum_{k=0}^{m_j-1} e^{z_{jk}}}, \tag{2}$$

where

$$z_{ji} = \sum_{k=0}^{i} a_j(\theta - b_j + d_k), \tag{3}$$

X_j is a random variable representing a response to item j, a_j is item discrimination, b_j is the item location parameter, and d_k, is a threshold or "step" difficulty for k = 0, 1, 2, ... , m_j-1 thresholds denoting the intersections of the respective m_j response functions.

Calibration of the mixed test format (3PL/GPC model) items was conducted using IRTPRO (Cai, Thissen, and du Toit, 2011) and included only the students who:

- attempted at least one item on the test,
- had a student score that was not invalidated.

In the analyses, parameters of both dichotomous and polytomous items were estimated simultaneously using marginal maximum likelihood estimation procedure.

As mentioned in the test design section of this document, the MSA Science tests utilize two operational forms (Form A and Form B) per grade with a set of 20 items common to both forms. This set of 20 items was used to create an incomplete data matrix so that the unique items (or field test items) from each form could be calibrated concurrently, thus placing the parameters for all operational items administered at each grade on a common scale.

Equating

The purpose of equating is to maintain a common scale (theta) for expressing the item parameter estimates across versions (i.e., annual administrations) of a test. The theta distribution is commonly scaled to have the mean set to 0 and the standard deviation set to 1. Once the 2016 MSA Science tests were concurrently calibrated, it was necessary to place each respective scale (Grade 5 and Grade 8) onto the originating 2007 base calibration. The *common item non-equivalent groups* (CINEG) data collection design was used (Kolen & Brennan, 2004). In particular, the common item sets are *all* operational SR items. In other words, all operational items but BCRs served as anchor items to place their parameter estimates back to the base scale.

For the item parameter estimates reflecting the base form, the most current parameter estimates were used.

When conducting equating with nonequivalent groups, the parameters from different forms (Form X and Form Y) need to be placed on the same IRT scale. This can be accommodated under the IRT framework, because when the IRT model holds, the parameter estimates from different groups are on linearly related theta scales (Lord, 1980). Thus, a linear equation can be used to place IRT parameter estimates onto an existing (base) scale. A publicly available equating program, STUIRT (Kim & Kolen, 2004), was used to calculate transformation constants from the Stocking and Lord Procedure. In the Stocking and Lord approach (Stocking & Lord, 1983), the difference between two test characteristic curves is first squared for a fixed theta value:

$$\boldsymbol{SLdiff}(\theta_i) = \left[\sum_{j:V} \boldsymbol{P}_{ij}(\theta_{Yi}; \hat{\boldsymbol{a}}_{Yj}, \hat{\boldsymbol{b}}_{Yj}, \hat{\boldsymbol{c}}_{Yj}) - \sum_{j:V} \boldsymbol{P}_{ij}(\theta_{Yi}; \frac{\hat{\boldsymbol{a}}_{Xj}}{\boldsymbol{A}}, \boldsymbol{A}\hat{\boldsymbol{b}}_{Xj} + \boldsymbol{B}, \hat{\boldsymbol{c}}_{Xj}) \right]^2 .$$

The estimation proceeds by finding the combination of *A* and *B* minimizing the following criterion:

$$\boldsymbol{SLcrit} = \sum_{i} \boldsymbol{SLdiff}(\theta_i) ,$$

where the summation is over examinees. An iterative approach needs to be used to solve for *A* and *B* in the above equations.

Stability Check Procedure

Dramatic changes in item parameter values can result in systematic errors in equating results (Kolen & Brennan, 2004). Thus, it is expected to track changes in item parameters and to evaluate how those changes affect the results of equating. Pearson has conducted analyses to examine the stability of the MSA Science anchor item parameters after equating. Specifically, stability in the operational linking item parameters were evaluated by examining differences in the originating (base) and transformed item characteristic curves. All items used for linking the 2016 MSA Science tests to the base scales were included in this stability check.

Pearson used an iterative anchor stability check approach that is analogous to examining differential item functioning. The steps of this process are as follows:

1) Place the current item parameters for all anchor items on the base-year scale by computing Stocking & Lord (SL) transformation constants using STUIRT (Kim & Kolen, 2004) and all anchor items.
2) For each linking item, calculate the weighted sum of the squared deviation (d^2) between the Item Characteristic Curves (ICC) using a theoretical weighted posterior theta distribution with 40 quadrature points:
   a) Apply the SL constants to the thetas associated with the standard normal theta distribution used to generate the SL constants.
   b) For each anchor item calculate a weighted sum of the squared deviation between the ICCs based on old (x) and new (y) parameters at each point in this theta distribution.

$$d_i^{\,2} = \sum^{k} \left[P_{ix}(\theta_k) - P_{iy}(\theta_k) \right]^2 \bullet g(\theta_k)$$

c) Compute the mean and standard deviation of the d^2 values, and flag any item with a d^2 more than two standard deviations above the mean.
d) Review and sort the items in a descending (largest to smallest) fashion according to the d^2 value.
e) Step 2d) results in an item with the largest area between pre- and post-equated ICCs at the top of the list of anchor items:
    i) Drop the largest d^2 item from the anchor set.
    ii) Repeat steps 1 through 2d – omitting 2c (use the original mean and standard deviation) until no more items are flagged or more than 20% of the operational items appearing across the two OP forms will be dropped.
f) Review all dropped items with a d^2 flag to determine at what point in the process no more items should be dropped. Items not flagged in this process should not be dropped, but a flag alone is not the sole criteria for removing an item from the linking set. In other words, the flag is a necessary, but not sufficient criterion for dropping an anchor item.

Flagged items were further reviewed through examination of the CTT and IRT estimates, item characteristic curves, fit statistics, item sequence change (change from location of the most recent administration), and impact on the test blueprint representation. Any item considered for removal was evaluated by a Pearson Content Specialist to determine of the content of the item or an event in the item's development history might explain the change in item performance. Decisions about whether to keep or remove an item were evaluated on a per item basis. When an item (note, only one item can be removed at a time) was removed from the anchor set, then this process (beginning with the computation of transformation constants) was repeated until there were no further items to be removed.

This process resulted in seven items removed from grade 5 and three items removed from the grade 8 common item sets. The final transformation constants for each grade following this procedure are listed in Table 5.

Table 5. Operational Transformation Constants

	Grade 5		**Grade 8**	
	Slope	**Intercept**	**Slope**	**Intercept**
Operational (14 OP items >> 07 base scale)	1.118251	0.02287	1.053803	0.070592

The transformation constants were applied to the 2016 item parameters so that all items in the MSA Science pool can be put onto the original base scales. The equated IRT parameters for grade 5 and 8 items are presented in Appendix A.

Test Analysis, Operational Scaling and Scoring

Test Analysis

IRT item parameter estimates were used to generate test characteristic curves (TCCs), test information functions (TIFs), and conditional standard errors of measure (CSEM). These indices were computed for each of the current year operational forms (A and B) and the base-year operational forms (A and B).

These graphs show how well a given test form compares to another in terms of the measurement (scale) characteristics across the scale range. Here the primary comparisons are between the 2016

Form A and B curves and curves reflective of operational items from the 2008 (first operational) administration.

Figure 1 shows the overlaid TCC plots for current year operational forms (A and B) and base-year operational forms (A and B) for grade 5. Figure 2 displays overlaid TIF plots for current year operational forms (A and B) and base-year operational forms (A and B) for grade 5. Figure 3 shows the overlaid CSEM plots for current year operational forms (A and B) and base-year operational forms (A and B) for grade 5.

The vertical lines in each figure represent the location of the Proficient and Advanced performance standards on the reporting scale metric (each performance level is denoted at the top of the plot: Basic, Proficient, and Advanced). It should also be noted that each curve is presented according to the MSA Science scale score metric, which is described in the Defining Scale Ranges section.

Figure 1. Test Characteristic Curves - Grade 5

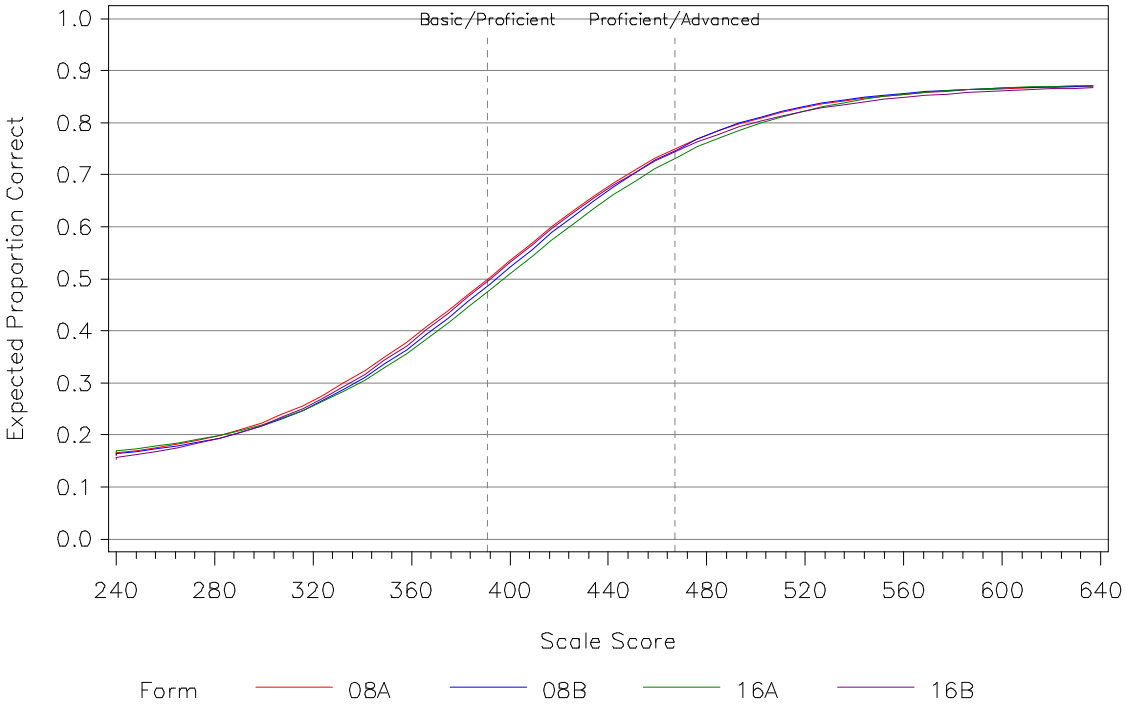


Figure 2. Test Information Function - Grade 5

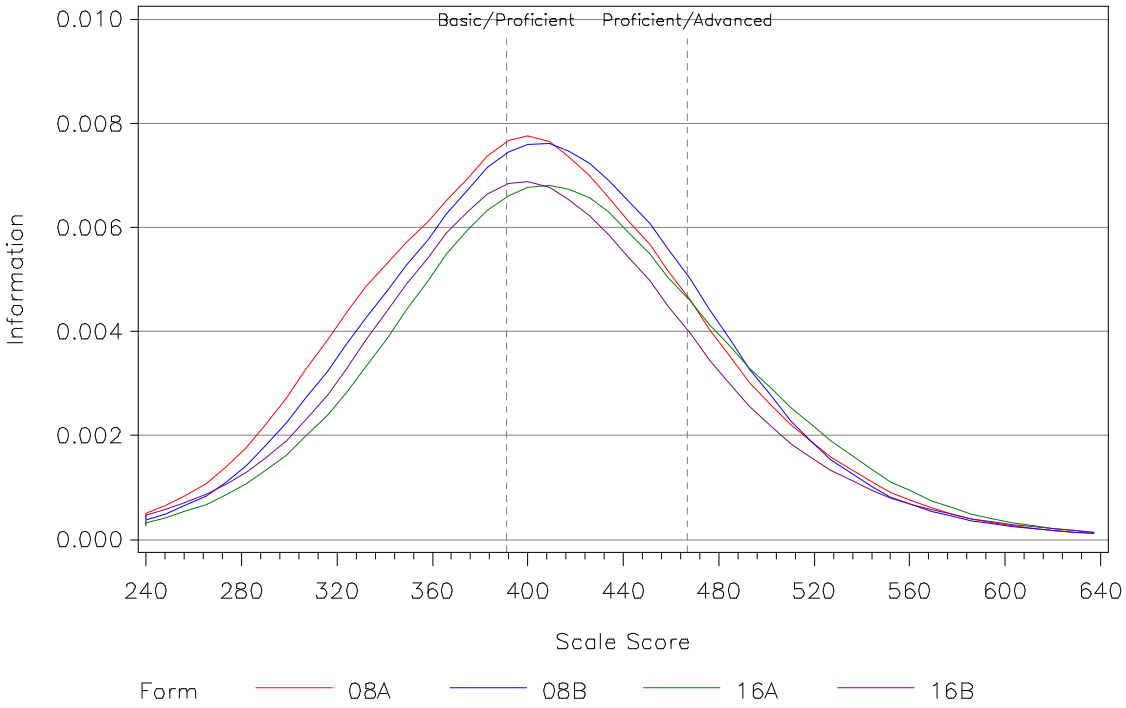
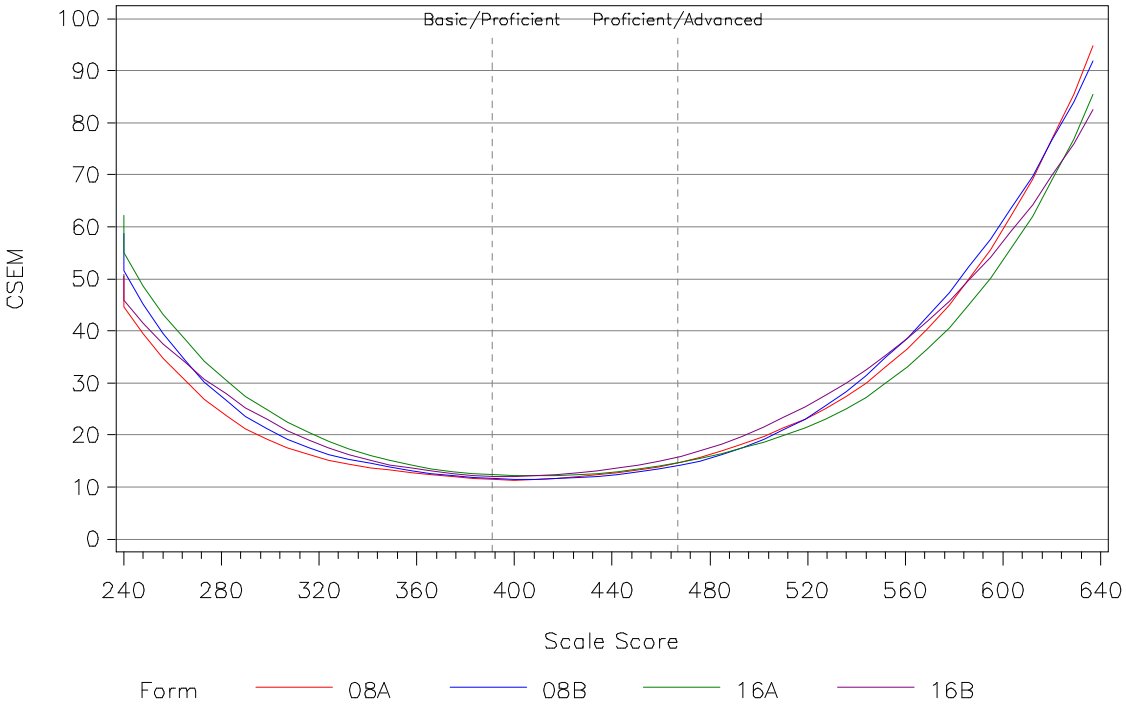


Figure 3. Conditional Standard Error of Measurement - Grade 5



As with grade 5, IRT item parameter estimates were used to generate test characteristic curves (TCCs), test information functions (TIFs), and conditional standard errors of measure (CSEM) were computed for each of the current year operational forms (A and B), and the base-year operational forms (A and B) for grade 8.

Figure 4 shows the overlaid TCC plots for current year operational forms (A and B) and base-year operational forms (A and B) for grade 8. Figure 5 displays overlaid TIF plots for current year operational forms (A and B) and base-year operational forms (A and B) for grade 8. Figure 6 shows the overlaid CSEM plots for current year operational forms (A and B) and base-year operational forms (A and B) for grade 8. The vertical lines in each figure represent the location of the Proficient and Advanced performance standards on the reporting scale metric. Note that each curve is presented relative to the scale score metric described in the Defining Scale Ranges section.

Figure 4. Test Characteristic Curves - Grade 8

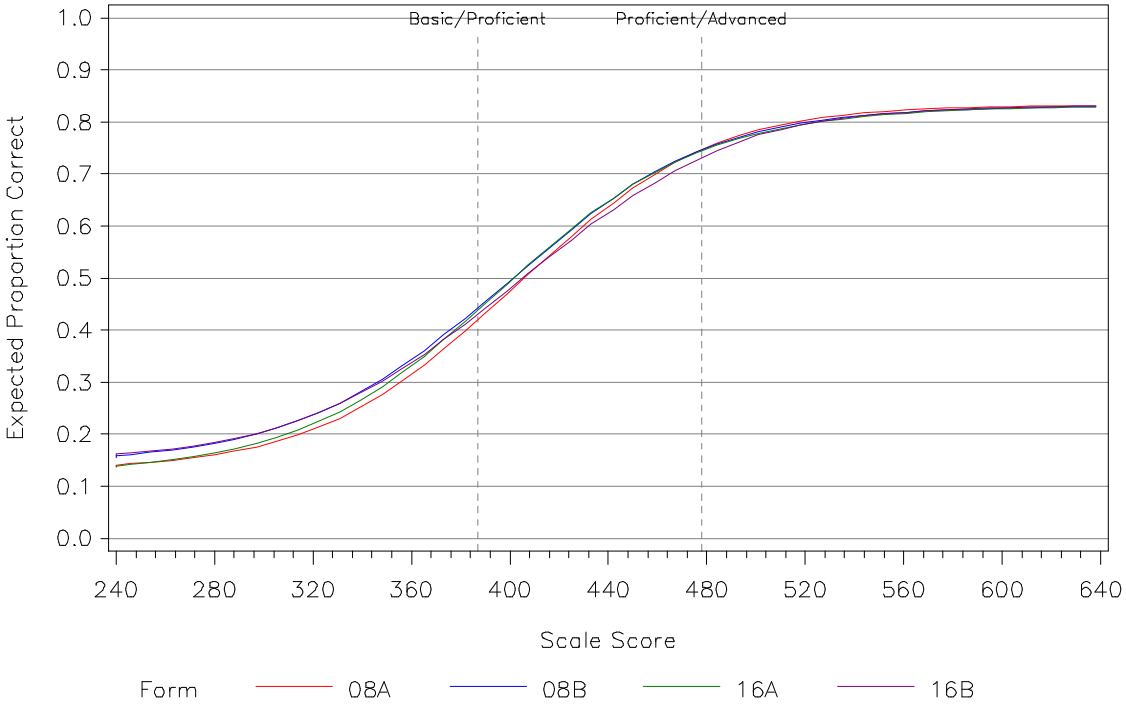


Figure 5. Test Information Function - Grade 8

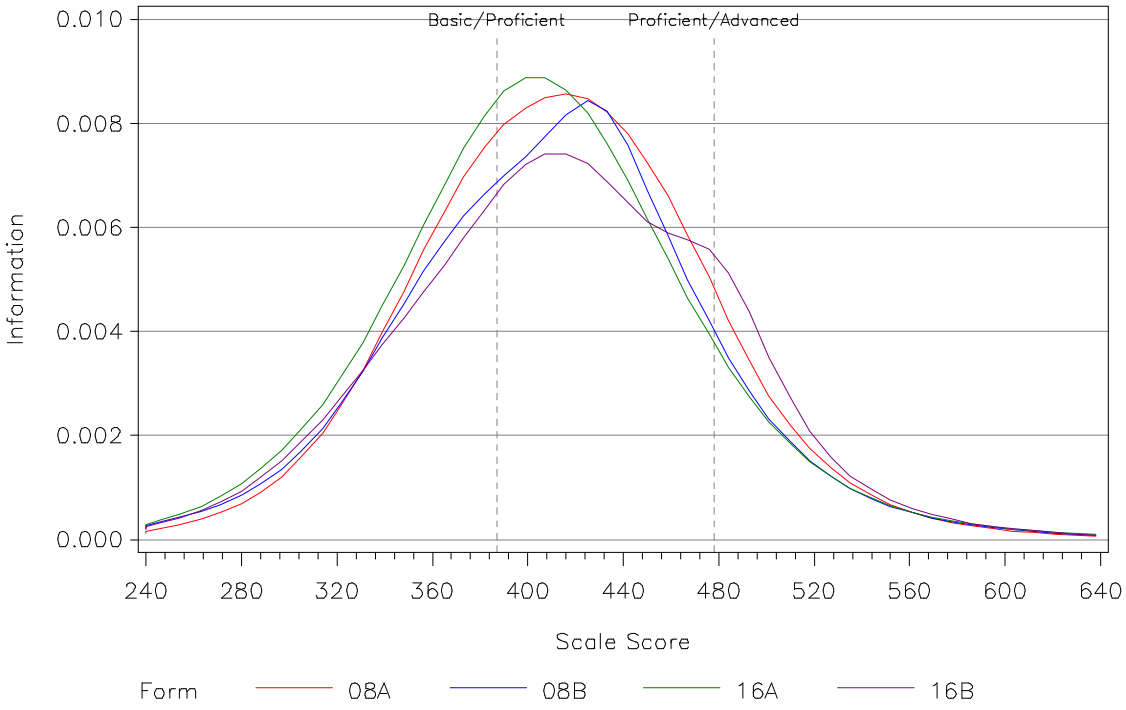
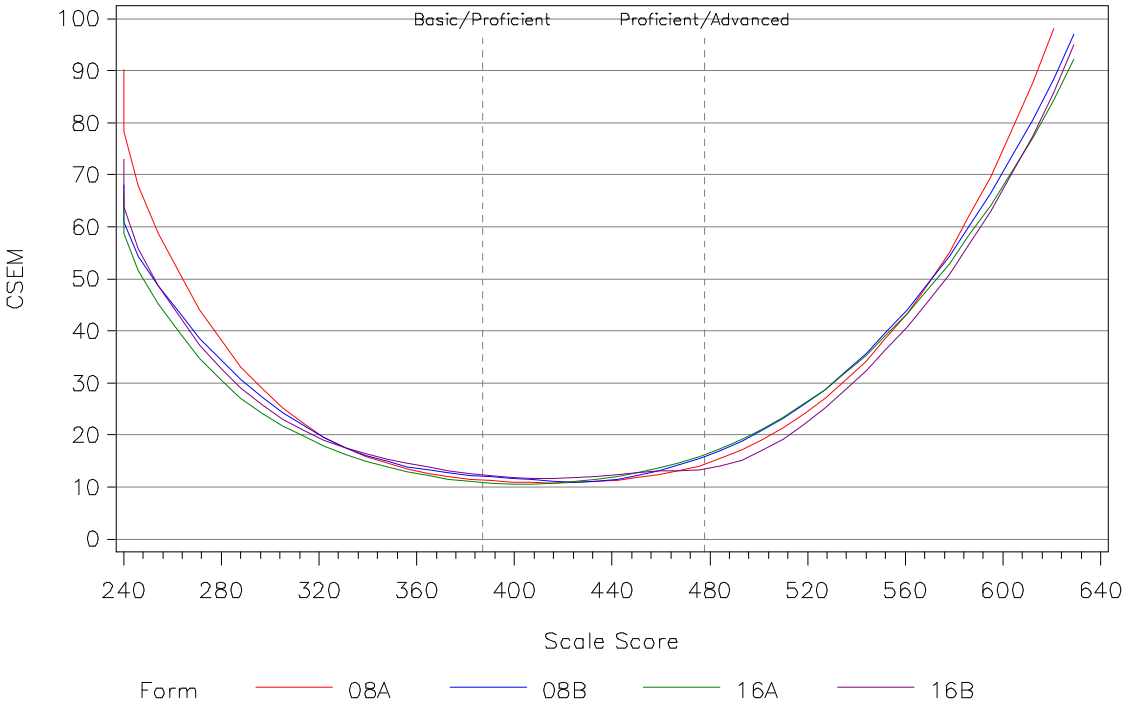


Figure 6. Conditional Standard Error of Measurement - Grade 8



Defining Scale Ranges

The theta scale is not often used for reporting because of interpretation issues arising from a scale with values typically ranging from -4.0 to +4.0. Therefore, following the calibration and equating phases, the resulting theta values are linearly transformed to a reporting scale that can be more meaningfully interpreted by students, teachers and other stakeholders. In order to facilitate the use and interpretation of the results of the 2016 MSA Science operational administration, scale scores were created through the application of scaling constants determined from the base 2007 field test administration. Scale scores were computed using the following simple linear transformation equation:

$$SS = M1(\theta) + M2$$

where, M1 is a multiplicative term, M2 is an additive term, and θ is an IRT based measure of student ability. These scaling constants (M1 and M2) were developed to meet MSDE requirements that the mean and standard deviation (sd) be established in the base year at mean scale score = 400 and sd = 40, while maintaining the lowest obtainable scale score (LOSS) at 240 and the highest obtainable scale score (HOSS) at 650. The LOSS and HOSS set the minimum and maximum values that are possible on the MSA Science test. These scaling constants as well as the LOSS and HOSS for each grade appear in Table 6.

Table 6. Target LOSS, HOSS, and Scaling Constants for Grades 5 and 8.

Grade	LOSS	HOSS	M1	M2
5	240	650	42.3077	400.1688
8	240	650	42.617	398.9311

ISE Pattern Scoring

The 2016 operational scores were estimated by the pattern scoring approach. As noted previously, the 2016 student responses were scored using the original 2016 item parameters. The 2016 operational item parameters were first equated to the base theta scale established in 2007. The equated item parameters were then used to estimate student ability (theta) using Pearson's ISE program which was described in the next paragraph. Final theta estimates from ISE were transformed onto the MSA Science operational scale using the scaling constants provided in Table 6.

Pearson used an internally developed software program called IRT Score Estimation (ISE; Chien, Hsu, & Shin, 2007) to conduct pattern scoring for the spring 2016 administration of the MSA Science tests for grades 5 and 8. ISE is a C++ computer program for estimating item response theory (IRT) pattern scores on various IRT models and the estimating methods include Brute-Force (BF) and Newton Raphson (NR) methods. ISE outputs IRT pattern scores and associated standard error (SE).The program has been extensively tested and compared to commercially available software programs such as MULTILOG, PARSCALE by (Tong, Um, Turhan, Parker, Shin, Chien, & Hsu, 2007). Tong et. al. (2007) found that that with normal cases the ISE program was able to replicate MULTILOG and PARSCALE theta estimates. However, "in problem cases, such as monotonically decreasing likelihood functions, in which MULTILOG and PARSCALE both produced theta estimates, ISE was able to produce the estimates that yielded the largest likelihood function, in alignment with the definition of the maximum likelihood algorithm" (p. 9). In addition, "with problem cases in which MULTILOG and PARSCALE failed to produce theta estimates, ISE was able to produce an estimate that yielded the largest likelihood from the likelihood function of a given response pattern" (p. 9). With regard to the CSEM, ISE produced

similar results to MULTILOG. More information about the ISE program can be found in the user manual, the technical manual, and the evaluation report, which are available upon request.

Conditional Standard Errors for LOSS and HOSS

Within ISE, student ability (theta) is determined via maximum likelihood estimation (MLE). One characteristic of MLE is that for students with scores of zero or perfect scores, abilities are not estimable (i.e., they effectively result in estimates of $\pm \infty$). Because of this it is typical to establish ability values or scale scores that are in line with the respective overall scale. For the MSA Science tests, the LOSS and HOSS values reflect the values associated with these extreme scores. Additionally, there are instances in which certain score patterns close to zero and perfect scores will provide ability estimates where the respective conditional standard errors of measurement (CSEM) are very large. These inflated CSEM estimates are problematic in that they are out of line with estimates from different score patterns but of the same ability. In addition to establishing reasonable scale scores for these points, it is also desirable to provide some reasonable associated standard error to promote appropriate score interpretation.

In order to provide students with appropriate score interpretations where ability estimates from the MSA Science tests are associated with the LOSS and HOSS scale scores (240 and 650), and Pearson recommended a maximum CSEM of 160 be used. This recommendation was based on multiple considerations.

First of all, consideration was given to the magnitude of standard errors relative to the overall scale score range. The current scale ranges from 240 to 650 (410 total points). When standard errors exceed 40% of a scale range, the utility of a test score interpretation is limited. With this in mind, the initial 2007 MSA Science base calibration was evaluated. The initial 2007 MSA Science administration involved the administration of ten field test forms per grade; each created in line with the MSA Science blueprints and served as the mechanism for establishing the base scales. For each form, ability estimates were generated and their associated standard errors were examined. Across grade 5 and 8 forms, the largest standard errors for the highest estimable abilities were roughly 155 scale score points and were within the 40% heuristic noted above.

In addition to evaluation of the base year calibrations, consideration was also given to standing practice for other Maryland assessments; specifically the Maryland High School Assessments (HSA). The 2004 HSA Technical Report describes principals adopted for the determination of optimal LOSS and HOSS values where associated standard errors are also described (Appendix 3.C). In determining a value for HOSS, it was recommended that the associated conditional standard error be lower than ten times the minimum conditional standard error on the overall test. For the LOSS, the recommendation was for the associated conditional standard error to be lower than fifteen times the minimum conditional standard error on the test. For the base year MSA Science administration, minimum CSEM values were roughly 11 scale score points.

Based on these considerations, a recommendation was made for the maximum CSEM be set to 160 for the LOSS and HOSS. This was in line with the observed standard errors from the base year calibrations for extreme scores and also in line with existing practice. Based upon state approval of the recommendation, the rule was implemented to report CSEM for all scores.

Test Score Reliability

The reliability of a test provides an estimate of the extent to which an assessment will yield the same results across subsequent administrations, provided the two administrations do not differ on relevant variables. Reliability coefficients are usually forms of correlation coefficients and must be interpreted within the context and design of the assessment and of the reliability study. The

forms of reliability below measure different dimensions of reliability and thus any or all might be used in assessing the reliability of MSA Science. The estimates of reliability reported here are measures of internal consistency and reflect the degree to which the components of a test are consistent with other components of the test. One of the most commonly used indices of internal consistency reliability is Cronbach's coefficient *alpha* (α; Cronbach, 1951). In this formula, the s_i^2 denotes the variances for the k individual items; s_{sum}^2 denotes the variance for the sum of all items.

$$\alpha = (k/(k-1)) * [1- \Sigma(s^2_i)/s^2_{sum}]$$

Because of the mixed item types on the MSA Science test (i.e., SR and BCR), a stratified alpha (Cronbach, Schönemann, & McKie, 1965) is more appropriate. Stratified alpha accounts for the fact that different groups of items ("strata") may have different variances. Since the Cronbach alpha relies on a single overall variance, it may not be the best estimate of "true" reliability. Because of this, stratified alpha reliability coefficients were computed for the MSA Science tests. The formula is:

$$Stratified\ \alpha = 1 - \frac{((\sigma^2_{SR}(1-\rho_{SR}) + (\sigma^2_{CR}(1-\rho_{CR}))}{\sigma^2_t}$$

where

σ^2_{SR} = variance associated with SR items;

σ^2_{CR} = variance associated with BCR items;

σ^2_t = variance of total score;

ρ_{SR} = reliability associated with the SR items; and

ρ_{CR} = reliability associated with BCR items.

These results are presented in Table 7.

Table 7. Reliability Estimate by Grade, Form, Gender and Ethnicity

Group		Grade 5		Grade 8	
		Form A	**Form B**	**Form A**	**Form B**
Overall		0.90	0.91	0.91	0.90
Gender	Female	0.90	0.90	0.91	0.90
	Male	0.91	0.91	0.92	0.91
Ethnicity	Non-Hispanic/Latino	0.91	0.91	0.91	0.90
	Hispanic/Latino	0.89	0.90	0.91	0.89
Race	American Indian	0.89	0.88	0.91	0.87
	Asian/Pacific Islander	0.89	0.91	0.91	0.89
	African American	0.88	0.88	0.89	0.87
	Native Hawaiian	0.88	0.89	0.94	0.89
	White	0.89	0.89	0.89	0.88

The coefficient alpha estimates for all forms meet conventional guidelines for applied test reliability (i.e., α > .85).

Student Performance

Score Interpretation

To help provide appropriate interpretation of the 2016 MSA Science operational test scores, two types of scores were created: scale scores and performance levels and descriptions.

Scale Scores

As explained in the proceeding section, the 2016 MSA Science tests yield scale scores that range between 240 and 650. As a result of calibration, equating, and scaling the scale scores from the two base forms are comparable within the same grade, but not across grade levels. The only inferences that can be appropriately drawn from scale scores are that higher scale scores represent higher performance on the MSA Science test. Thus, performance levels and descriptions can give a specific interpretation other than a simple interpretation because they were developed to bring meaning to the scale scores.

Performance Levels and Descriptions

Performance levels and descriptions provide specific information about students' performance levels and help interpret the 2016 MSA Science scale scores. They describe what students at a particular level generally know and are able to do and can be applicable to all students within a grade level.

Performance standards for the MSA Science tests were established in 2007. Details of the standard-setting process and outcomes are provided in MSA Science standard-setting technical report (Pearson, 2007). The Maryland State Board of Education reviewed the performance standards recommended by the standard-setting committee and made a modification in the recommendation. The performance standards approved by the State Board are listed in Table 8. Students whose scale scores are lower than the Proficient cut score are classified as "Basic." The highest performance group whose scale score is equal or higher than Advanced cut score belongs to the "Advanced" group. The middle group is called "Proficient."

Table 8. Scale score cut scores for grades 5 and 8 MSA Science.

Grade	Proficient Cut score	Advanced Cut score
5	391	467
8	387	478

Table 9 reports percentages of grade 5 students in three performance groups and the descriptive statistics for the selected subgroups (gender and ethnicity). The analysis was conducted for all students in grades 5 as well as by administration mode.

Table 10 reports percentages of grade 8 students in three performance groups and the descriptive statistics for the selected subgroups (gender and ethnicity). The analysis was conducted for all students in grades 8 as well as by administration mode.

Table 9. Grade 5 Performance Level Percentages and Summary Statistics

	Overall						Online Administration						Paper Administration					
	Performance Levels			Mean	SD	N	Performance Levels			Mean	SD	N	Performance Levels			Mean	SD	N
	B	P	A				B	P	A				B	P	A			
Subgroup																		
All Students																		
All	40	51	9	401	50.1	64867	40	51	9	400	50.3	60067	33	55	11	410	47.3	4800
Gender*																		
Female	40	52	8	401	48.3	31786	40	52	8	400	48.4	29385	33	56	11	409	46.7	2401
Male	40	51	10	401	51.7	33080	40	50	9	401	52.0	30681	33	55	12	410	47.9	2399
Ethnicity*																		
Hispanic/Latino	55	42	3	383	47.9	10048	55	42	3	383	48.1	9578	50	47	4	391	43.8	470
Non-Hispanic/Latino	37	53	10	405	49.8	54819	38	53	10	404	49.9	50489	31	56	12	412	47.2	4330
Race*																		
American Indian	49	48	2	392	46.4	168	49	49	3	392	47.2	154	57	43	0	389	38.5	14
Asian/Pacific Islander	19	62	19	427	47.0	4323	19	62	19	427	46.8	4043	24	55	21	423	50.3	280
African American	59	39	2	379	45.9	21704	59	39	2	379	46.1	20457	56	41	2	383	42.3	1247
Native Hawaiian	26	67	7	412	43.7	111	28	68	4	408	43.7	96	13	60	27	434	37.1	15
White	23	63	14	421	44.3	25659	23	63	14	421	44.4	23152	20	64	16	424	43.2	2507
Two or More Races	30	59	11	413	46.0	2854	30	59	11	413	46.1	2587	28	61	11	413	44.7	267

Note: Performance Levels, B=Basic, P=Proficient, A=Advanced

** 1 instance with missing Gender; 9 instances of missing Ethnicity/Race information*

Table 10. Grade 8 Performance Level Percentages and Summary Statistics

	Overall						Online Administration						Paper Administration					
	Performance Levels			Mean	SD	N	Performance Levels			Mean	SD	N	Performance Levels			Mean	SD	N
	B	P	A				B	P	A				B	P	A			
Subgroup																		
All Students																		
All	35	61	4	402	47.8	62692	36	60	4	400	48.0	55152	28	66	6	410	45.5	7540
Gender																		
Female	34	62	4	402	45.3	30634	35	61	3	401	45.5	26906	27	68	5	411	43.2	3728
Male	36	60	5	402	50.1	32058	37	59	5	400	50.3	28246	29	65	6	410	47.6	3812
Ethnicity*																		
Hispanic/Latino	50	49	1	382	48.9	8943	33	62	4	381	48.5	8176	41	56	3	390	51.7	767
Non-Hispanic/Latino	33	63	5	405	46.8	53749	51	48	1	404	47.1	46976	26	68	6	413	44.1	6773
Race*																		
American Indian	36	62	2	395	42.6	159	37	60	2	395	42.7	131	13	28	59	397	43.1	28
Asian/Pacific Islander	15	74	11	428	43.9	4062	15	75	11	429	43.9	3631	17	76	6	423	43.9	431
African American	55	45	1	379	42.9	21225	56	44	1	378	43.1	19068	46	53	1	390	39.4	2157
Native Hawaiian	31	64	5	397	49.9	78	31	63	6	398	50.4	67	1	2	97	394	48.8	11
White	17	75	7	422	40.7	25566	18	76	7	421	40.5	21831	18	82	0	425	41.5	3735
Two or More Races	27	69	4	411	43.3	2659	27	69	4	411	43.2	2248	29	71	0	411	43.9	411

Note: Performance Levels, B=Basic, P=Proficient, A=Advanced
** 24 instances with missing Ethnicity/Race information*

Validity

As noted in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999), "validity is the most important consideration in test evaluation."

Messick (1989) defined validity as follows:

> Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment. (p.5)

This definition implies that test validation is the process of accumulating evidence to support intended use of test scores. Consequently, test validation is a series of ongoing and independent processes that are essential investigations of the appropriate use or interpretation of test scores from a particular measurement procedure (Suen, 1990).

In addition, test validation embraces all of the experimental, statistical, and philosophical means by which hypotheses and scientific theories can be evaluated. This is the reason that validity is now recognized as a unitary concept (Messick, 1989).

To investigate the validity evidence of the 2016 MSA-Science tests, content-related evidence, differential item functioning (DIF) analysis on gender and ethnicity, and evidence based on internal structure were collected.

Content-related Evidence

Content related validity is frequently defined in terms of the sampling adequacy of test items. That is, content validity is the extent to which the items in a test adequately represent the domain of items or the construct of interest (Suen, 1990). Consequently, content validity provides judgmental evidence in support of the domain relevance and representativeness of the content in the test (Messick, 1989).

As described in the Item Development and Review section, all MSA Science items were explicitly developed to measure the specific knowledge and skills described in the Maryland State Curriculum. As noted, the alignment of the items to the six Science standards was reviewed and verified independently by multiple content experts to include Pearson staff, MSDE staff, and Maryland educators.

The Test Overview and Design section details the connection between the MSA Science blueprint and the MSC. The 2016 MSA Science tests were constructed exclusively using items that met not only the statistical criteria described in this report, but also verified as aligning to the MSC by Maryland science content experts. As described, tests were constructed according to the test blueprints and as such, scores provided are reflective of overall Science ability as defined within the state standards.

Differential Item Functioning (DIF)

Since the test assesses the statewide content standards, which are required to be taught to all students, the test should not be more or less valid for use with one subpopulation of students relative to another. Great care has been taken to ensure that the MSA Science items are fair for students of various backgrounds. During the item development and review processes, efforts were made to avoid the use of language or context that might offer an advantage or disadvantage to particular subpopulations within Maryland. Besides these content-based efforts that are put forth in the test development process, data-driven statistical procedures are also employed to identify items that behave differently for different populations. Statistical indices of Differential Item

Functioning (DIF) are only a quantitative marker; bias is a qualitative condition that can only be determined by an examination of the content of the item. The MSA Science test development approaches incorporate both perspectives when reviewing test questions with respect to fairness. Bias and sensitivity committee review of all field tested items occurs each year as described in the Item Development and Review section.

DIF analyses are carried out on all MSA Science field test items according to the procedures in the Differential Item Functioning Analysis section. DIF statistics are used to identify items on which members of a focal group have different probability of getting the items correct from members of a reference group after members of both groups have been matched by the students' ability level on the test. In the DIF analysis, the total raw score on the operational items is used as the ability-matching variable. Any items displaying DIF that are also judged to contain language or context favoring or disadvantaging a given subpopulation are removed from the pool of eligible items during data review. Because of this ongoing and thorough approach, the majority of items on the MSA Science operational tests exhibit no DIF or weak DIF, and no items judged to show bias are selected for operational use.

Inter-Correlations among Standards

There are six standards within the MSC frameworks for MSA Science that together contribute to the overall reported Science test score. Items are written to capture performance that not only reflects the overall construct of science as defined within the frameworks, but to capture content and skills by standard. To assess the extent to which items aligned with the standards are offering some unique characteristics based on each respective standard, while more strongly capturing an overall "science" construct, a correlation matrix was computed among the total scores of competencies. It should be noted that only overall scale scores and performance levels are reported for MSA Science.

Table 11 reports the correlations among the six standards based on scale scores. The standard-level (subtest) inter-correlations ranged from 0.546 to 0.850. The standard sub-scores are moderately highly related to one another and more strongly related to the total test score. This suggests there is some uniqueness to items grouped by standard but that they are collectively measuring a dominant overall construct (science).

Table 11. Correlation among MSA Science content standards

Grade 5 Form A	**Mean**	**SD**		**Str1**	**Str2**	**Str3**	**Str4**	**Str5**	**Str6**	**Total**
	405.92	66.38	**Str1**	1.000						
	405.71	68.03	**Str2**	0.588	1.000					
	403.57	66.34	**Str3**	0.605	0.592	1.000				
	415.32	84.15	**Str4**	0.563	0.564	0.555	1.000			
	405.59	67.33	**Str5**	0.596	0.584	0.603	0.546	1.000		
	401.68	64.26	**Str6**	0.650	0.619	0.627	0.594	0.609	1.000	
	403.60	48.65	**Total**	0.814	0.800	0.810	0.760	0.792	0.850	1.000
Grade 5 Form B				**Str1**	**Str2**	**Str3**	**Str4**	**Str5**	**Str6**	**Total**
	401.17	63.61	**Str1**	1.000						
	399.91	65.15	**Str2**	0.652	1.000					
	400.43	66.40	**Str3**	0.660	0.646	1.000				
	401.03	76.14	**Str4**	0.598	0.617	0.590	1.000			
	408.88	81.87	**Str5**	0.603	0.599	0.606	0.573	1.000		
	401.20	78.60	**Str6**	0.626	0.627	0.623	0.594	0.587	1.000	
	398.95	51.31	**Total**	0.841	0.836	0.830	0.784	0.783	0.810	1.000
Grade 8 Form A				**Str1**	**Str2**	**Str3**	**Str4**	**Str5**	**Str6**	**Total**
	398.70	63.43	**Str1**	1.000						
	399.05	62.69	**Str2**	0.668	1.000					
	396.98	68.09	**Str3**	0.677	0.673	1.000				
	399.91	62.16	**Str4**	0.675	0.662	0.665	1.000			
	399.78	70.76	**Str5**	0.620	0.615	0.618	0.614	1.000		
	408.26	74.21	**Str6**	0.628	0.630	0.624	0.622	0.586	1.000	
	399.30	48.19	**Total**	0.845	0.848	0.847	0.843	0.777	0.798	1.000
Grade 8 Form B				**Str1**	**Str2**	**Str3**	**Str4**	**Str5**	**Str6**	**Total**
	404.45	69.91	**Str1**	1.000						
	405.94	67.23	**Str2**	0.588	1.000					
	413.04	75.07	**Str3**	0.625	0.589	1.000				
	399.39	70.15	**Str4**	0.583	0.574	0.582	1.000			
	396.87	76.94	**Str5**	0.562	0.561	0.555	0.553	1.000		
	405.53	60.05	**Str6**	0.644	0.624	0.642	0.607	0.582	1.000	
	404.22	47.27	**Total**	0.806	0.796	0.805	0.779	0.742	0.847	1.000

*Str1=Skills and Processes; Str2=Earth/Space Science; Str3=Life Science; Str4=Chemistry; Str5=Physics; Str6=Environmental

Confirmatory Factor Analysis

A confirmatory factor analysis (CFA) was conducted for the 2016 MSA Science administration to examine the relationship between the subtest scores relative the total test score. CFA used SAS Proc Calis and the maximum likelihood estimation (MLE; Anderson & Gerbing, 1988) procedure. The model hypothesized that the subtest scores belong to a single latent trait. Model fit was tested through indices including adjusted goodness of fit (AGFI), and Root Mean Square Error of Approximation (RMSEA). Values of the AGFI statistic that indicate good fit are higher than 0.90 (Tabachnick & Fidell, 2001). The RMSEA is a function of the estimated discrepancy between the population covariance matrix and the model-implied covariance matrix, with a value of less than or equal to .05 indicating close fit and a value between .05 and .08 indicating a "reasonable error of approximation" (Browne & Cudeck, 1993, p. 144). Hu and Bentler (1999) propose an RMSEA $\leq$ .06 as the guideline for close fit. Table 12 summarizes fit indicators estimated from the confirmatory factor analysis for the 2014 MSA Science tests. The confirmatory factor analysis results provide additional evidence to support the conclusion that scores from the MSA Science tests reflect a single latent trait (Science). For both grades, the lowest AGFI was 0.9961, and the highest RMSEA was 0.0228. The AGFI and RMSEA indicators supported the model fit.

Table 12. Fit indicators for confirmatory factor analysis on MSA Science

Grade/Form	AGFI	RMSEA
Grade 5 Form A	0.9961	0.0228
Grade 5 Form B	0.9961	0.0227
Grade 8 Form A	0.9989	0.0109
Grade 8 Form B	0.9965	0.0214

*AGFI: Adjusted Goodness of Fit; RMSEA: Root Mean Square Error of Approximation

Evidence for Scores from Accommodated Testing

Accommodations are offered to students with disabilities that preclude them from being fairly assessed by the tests as they are written (e.g., visually impaired students). In order to examine whether or not these accommodations are effective (i.e., result in valid test scores) the CFA conducted to examine the relationship between standards was repeated using only students testing with accommodations and then again using only students testing without accommodations. The results of this analysis showed comparable levels of model fit based on the two groups (see Table 13). This suggests that the accommodations offered to disabled students are effective at preserving the underlying latent structure of the MSA Science tests in comparison to that standard (non-accommodated) administration. By extension, MSA Science scores for accommodated and non-accommodated students are comparable.

Table 13. Fit indicators for accommodations/non-accommodations based CFA

	Accommodations		No Accommodations	
Grade/Form	**AGFI**	**RMSEA**	**AGFI**	**RMSEA**
Grade 5 Form A	0.9966	0.0128	0.9960	0.0231
Grade 5 Form B	0.9942	0.0254	0.9964	0.0219
Grade 8 Form A	0.9966	0.0156	0.9989	0.0108
Grade 8 Form B	0.9967	0.0075	0.9958	0.0233

*AGFI: Adjusted Goodness of Fit; RMSEA: Root Mean Square Error of Approximation

References

Allen, N.L., Carlson, J.E., & Zalanak, C.A. (1999). *The NAEP 1996 technical report.* Washington, DC: National Center for Education Statistics.

American Educational Research Association (AERA), American Psychological Association (APA), and the National Council on Measurement in Education (NCME) (2014). *Standards for educational and psychological testing.* Washington, DC: AERA.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.) *Testing structural equation models*. Pp. 136-162. Beverly Hills, CA: Sage.

Cai, L., Thissen, D.J., & du Toit, S. (2011). *IRTPRO* (Version 2.1) [Computer sotware]. Skokie, IL: Scientific Software International, Inc.

Chien,Y, Hsu, Y. & Shin, D. (2007). *ISE (IRT Score Estimation) program (version 1.0)*. [Computer sotware]. Iowa City, IA: Pearson.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16, 292-334.*

Cronbach, L. J., Schönemann, P., & McKie, D. (1965). Alpha coefficients for stratified parallel tests. *Educational and Psychological Measurement, 25,* 291-312.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In *Differential item functioning,* edited by Paul W. Holland & Howard Wainer. Hillsdale, NJ: Lawrence Erlbaum.

Feldt, L. S., & Brennen, R. L. (1989) *Reliability*. In Linn, R. L. (ed.), *Educational measurement.* New York:Macmillan.

Holland, P. W., & Thayer, D. T. (1988). "Differential Item Performance and the Mantel-Haenszel Procedure." In *Test Validity,* edited by Howard Wainer and Henry I. Braun. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria in fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1-55.

Kim, S., & Kolen, M. J. (2004). *STUIRT: A computer program for scale transformation under unidimensional item response theory models*. Iowa City, IA: Iowa Testing Programs, The University of Iowa.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading Massachusetts: Addison-Wesley Publishing Company.

Maryland State Department of Education (2016). Maryland Next Generation Science Standards Implementation and Planning Document; Retrieved from http://mdk12.msde.maryland.gov/share/VSC/UpdatedNGSS_IT.pdf

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18,* 5-11.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological measurement, 16,* 159–176.

No Child Left Behind Act of 2001, 20 U.S.C. 6301 et seq (2001) (PL 107-110).

SAS Institute Inc. 2004 SAS/*STAT*® *9.1 User's Guide.* Cary, NC: SAS Institute Inc

Spray, J. A., Ackerman, T. A., Reckase, M. D., & Carlson, J. E. (1989). Effect of the medium of item presentation on examinee performance and item characteristics. *Journal of Educational measurement, 26,* 261-271.

Suen, H. K. (1990). *Principles of test theories.* Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved [today's date], from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/Synthesis44.html

Tong, Y., Um, K., Turhan, A., Parker, B., Shin, D., Chien,Y., & Hsu, Y. (2007). IRT score estimation: Evaluation document. Iowa City, IA: Pearson.

Appendix A
Item Statistics

Table A.1. Grade 5 item statistics

UIN	Status	Pvalue	Ptbis	a	b	c	d1	d2	d3
50015	OP	0.65	0.31	0.38720	-1.00856	0.03335			
50033	OP	0.57	0.42	0.84372	0.23331	0.22785			
50041	OP	0.62	0.43	0.60964	-0.49726	0.01752			
50059	OP	0.85	0.49	1.18027	-1.38637	0.13423			
50083	OP	0.57	0.54	1.03062	0.04402	0.14664			
50110	OP	0.27	0.14	0.65845	2.63226	0.19598			
50117	OP	0.54	0.41	0.73789	0.42894	0.21087			
50174	OP	0.68	0.42	0.64177	-0.70482	0.12022			
50216	OP	0.71	0.46	0.79292	-0.71478	0.08846			
50219	OP	0.88	0.40	0.81000	-1.98454	0.02169			
50221	OP	0.65	0.41	0.62473	-0.24945	0.20306			
50227	OP	0.68	0.49	0.88344	-0.52319	0.14147			
50229	OP	0.56	0.39	0.55319	0.10039	0.14713			
50277	OP	0.37	0.20	0.89942	1.79126	0.27033			
50289	OP	0.69	0.33	0.53458	-0.50112	0.24576			
50319	OP	0.57	0.36	0.83343	0.47520	0.30334			
50335	OP	0.69	0.56	1.49335	-0.30685	0.23389			
50336	OP	0.73	0.41	0.64593	-1.06102	0.02431			
50348	OP	0.46	0.30	0.67080	1.02564	0.23423			
50416	OP	0.48	0.39	0.63953	0.63085	0.16142			
50421	OP	0.54	0.42	0.81520	0.32106	0.19620			
50423	OP	0.27	0.08	0.95911	2.67610	0.23732			
50433	OP	0.41	0.19	0.72506	1.78157	0.30365			
50461	OP	0.56	0.29	0.41484	0.10427	0.12351			
50463	OP	0.65	0.38	0.80806	0.11230	0.31411			
50466	OP	0.63	0.44	0.65433	-0.46150	0.02686			
50475	OP	0.50	0.44	1.31372	0.63010	0.23492			
50477	OP	0.74	0.47	0.79632	-0.93863	0.02230			
50546	OP	0.76	0.42	0.68868	-1.09908	0.14184			
50552	OP	0.58	0.40	0.72418	0.30172	0.25000			
50563	OP	0.40	0.25	0.44609	1.36951	0.16203			
50568	OP	0.78	0.43	0.68545	-1.25159	0.10017			
50598	OP	0.35	0.39	0.41967	1.59020	0	3.76120	-0.34033	-3.42088
50607_02	OP	0.30	0.25	0.94198	1.72394	0.18299			
50607_03	OP	0.50	0.33	0.56342	0.55176	0.15394			
50607_06	OP	0.19	0.43	0.52962	2.97442	0	2.94983	-0.40977	-2.54006
50620_02	OP	0.38	0.30	0.56873	1.27614	0.13306			
50620_04	OP	0.70	0.44	0.76149	-0.58194	0.13635			
50659	OP	0.81	0.48	0.96029	-1.08910	0.17751			
50694	OP	0.60	0.42	0.53932	-0.39910	0.05147			
50900	OP	0.52	0.36	0.59887	0.34832	0.13986			
50908	OP	0.41	0.21	0.59014	1.75365	0.26155			
50923_01	OP	0.87	0.38	0.82586	-1.60748	0.14227			
50923_04	OP	0.42	0.29	0.87732	1.25744	0.25280			
50929_04	OP	0.76	0.51	1.06049	-0.81169	0.17601			
50929_05	OP	0.68	0.43	0.61007	-0.89439	0.03070			

UIN	Status	Pvalue	Ptbis	a	b	c	d1	d2	d3
51004	OP	0.76	0.30	0.43110	-1.56377	0.10785			
51006	OP	0.52	0.42	0.65920	0.40840	0.16570			
51009	OP	0.67	0.43	1.16328	0.11082	0.33679			
51028	OP	0.39	0.50	0.59387	0.71668	0	3.09659	-0.50233	-2.59425
51041_02	OP	0.60	0.32	0.39369	-0.66772	0.02317			
51041_05	OP	0.56	0.25	0.28265	-0.48325	0.01536			
51057_01	OP	0.74	0.39	0.72807	-0.59281	0.24293			
51057_04	OP	0.73	0.40	0.65801	-0.85732	0.12872			
51127_01	OP	0.75	0.34	0.72233	-0.31797	0.37899			
51127_03	OP	0.53	0.23	0.35530	0.64463	0.19365			
51134_02	OP	0.42	0.20	0.32451	1.61054	0.17075			
51134_05	OP	0.45	0.33	0.53140	0.71637	0.13374			
51157_02	OP	0.48	0.26	0.37746	0.71735	0.13526			
51157_04	OP	0.55	0.33	0.49669	0.15622	0.15214			
51157_06	OP	0.40	0.17	0.21405	1.50170	0.04253			
51164	OP	0.67	0.34	0.42846	-0.75014	0.12439			
51183	OP	0.68	0.37	0.46043	-0.94121	0.05176			
51188	OP	0.64	0.55	1.13792	-0.28219	0.09710			
51207_02	OP	0.30	0.11	0.69105	2.73826	0.24649			
51207_04	OP	0.53	0.34	1.03096	0.76672	0.32142			
51212_04	OP	0.56	0.47	0.74350	-0.13559	0.07470			
51212_05	OP	0.66	0.31	0.39357	-1.07394	0.01386			
51218_01	OP	0.70	0.45	0.70954	-0.80648	0.08648			
51218_04	OP	0.80	0.39	0.62301	-1.65539	0.01948			
51261	OP	0.86	0.40	0.84934	-1.67618	0.02904			
51265	OP	0.43	0.32	0.61168	0.96801	0.17988			
51266	OP	0.55	0.44	0.83796	0.19191	0.18504			
51271	OP	0.48	0.28	0.74648	1.12402	0.29616			
51272	OP	0.74	0.45	0.85637	-0.74797	0.14967			
51273	OP	0.82	0.31	0.64599	-0.87206	0.39601			
51274	OP	0.49	0.13	0.56558	2.20758	0.40305			
51281	OP	0.45	0.42	0.73493	0.70073	0.14562			
51290	OP	0.25	0.54	0.69169	1.91800	0	2.52244	-0.39127	-2.13116
51296	OP	0.54	0.42	0.64909	0.21823	0.14730			
51301_02	OP	0.39	0.29	0.74522	1.36447	0.20769			
51301_03	OP	0.69	0.33	0.70232	0.03731	0.36802			
51301_08	OP	0.20	0.53	0.72799	2.27306	0	2.24782	-0.34154	-1.90628
51302_01	OP	0.35	0.26	0.52228	1.58065	0.14557			
51302_02	OP	0.66	0.48	0.70258	-0.72338	0.01316			
51302_08	OP	0.31	0.46	0.58156	1.76461	0	3.40484	-0.57782	-2.82702
51313_01	OP	0.68	0.33	0.42211	-1.14438	0.03027			
51313_03	OP	0.70	0.35	0.46758	-1.24023	0.01318			
51317_02	OP	0.64	0.29	0.89203	0.70326	0.45510			
51317_03	OP	0.44	0.36	0.66057	0.84013	0.14337			
51318_01	OP	0.37	0.28	1.15802	1.30798	0.23787			
51318_02	OP	0.53	0.45	1.11191	0.38536	0.21887			
51323_01	OP	0.49	0.32	1.02366	0.97338	0.29783			

UIN	Status	Pvalue	Ptbis	a	b	c	d1	d2	d3
51323_02	OP	0.78	0.47	1.13482	-0.64217	0.27678			
51324_01	OP	0.63	0.37	0.50961	-0.58422	0.02475			
51324_03	OP	0.27	0.20	0.72892	2.09558	0.16911			
51329_01	OP	0.52	0.35	0.66297	0.53369	0.19950			
51329_03	OP	0.68	0.34	0.48336	-0.89305	0.04292			
51359	OP	0.40	0.43	0.88172	0.87316	0.14576			
51362	OP	0.65	0.43	0.66665	-0.31208	0.16868			
51363	OP	0.76	0.47	0.95437	-0.72331	0.25343			
51365	OP	0.68	0.49	0.82026	-0.46669	0.16281			
51367	OP	0.78	0.40	0.57891	-1.52832	0.01092			
51368	OP	0.47	0.43	0.78509	0.50398	0.10644			
51383	OP	0.49	0.19	0.69299	1.67294	0.36633			
51391	OP	0.67	0.37	0.53490	-0.72874	0.12292			
51395	OP	0.64	0.51	0.93412	-0.25015	0.15943			
51399	OP	0.60	0.40	0.48727	-0.42905	0.04803			
55102	OP	0.83	0.32	0.53315	-1.94256	0.01719			
55169	OP	0.74	0.33	0.49391	-1.24881	0.05880			
55172	OP	0.67	0.35	0.46019	-0.99114	0.03165			
55241	OP	0.87	0.42	0.97507	-1.48979	0.23049			

UIN=Unique Item Number; Status=Administration condition (OP = Operational item); Pvalue=Item p-value; Ptbis=Item Point Biserial; IRT 3PL and GPC model item parameters (a, b, c, d_k)

Table A.2. Grade 8 item statistics

UIN	Status	Pvalue	Ptbis	a	b	c	d1	d2	d3
80026	OP	0.79	0.39	0.81822	-0.71950	0.28488			
80072	OP	0.49	0.34	0.79950	0.72988	0.23517			
80083	OP	0.36	0.34	0.60253	1.27450	0.12065			
80124	OP	0.94	0.32	0.98072	-2.29859	0.10047			
80178	OP	0.45	0.41	0.94903	0.78173	0.17162			
80187	OP	0.50	0.29	0.39105	0.22604	0.05350			
80202	OP	0.72	0.45	0.83205	-0.45041	0.23993			
80209	OP	0.49	0.31	0.62484	0.84127	0.21201			
80213	OP	0.65	0.47	0.96977	-0.02773	0.25715			
80228	OP	0.51	0.37	0.82009	0.78087	0.26596			
80257	OP	0.75	0.47	0.86655	-0.90669	0.01782			
80272	OP	0.67	0.37	0.87691	0.25733	0.39444			
80330	OP	0.80	0.41	0.72906	-1.35965	0.08077			
80344	OP	0.81	0.53	1.39066	-0.95545	0.17199			
80425	OP	0.42	0.25	0.92470	1.35144	0.28301			
80447	OP	0.83	0.46	0.90994	-1.40193	0.02332			
80529_01	OP	0.69	0.44	1.08441	-0.05311	0.30300			
80529_03	OP	0.72	0.41	0.80322	-0.42601	0.22207			
80529_06	OP	0.43	0.69	0.93560	0.45400	0	1.35175	-0.01222	-1.33952
80552	OP	0.58	0.51	1.14012	0.07746	0.18004			
80574	OP	0.38	0.42	1.22176	0.86601	0.14970			
80585	OP	0.79	0.28	0.44959	-1.79228	0.02141			
80610	OP	0.77	0.44	0.73122	-0.98243	0.12988			
80628	OP	0.40	0.43	0.78986	0.80213	0.10772			
80634	OP	0.19	0.56	0.79613	1.60916	0	1.27188	-0.41304	-0.85884
80642	OP	0.74	0.32	0.79348	-0.01800	0.43631			
80666_04	OP	0.62	0.31	0.86311	0.59398	0.38898			
80666_06	OP	0.69	0.30	0.75593	0.33896	0.43039			
80674_01	OP	0.48	0.28	1.10439	1.18739	0.33233			
80674_05	OP	0.27	0.17	1.89193	1.77830	0.20702			
80674_06	OP	0.39	0.58	0.80661	0.84709	0	2.40256	-0.16452	-2.23804
80748	OP	0.35	0.48	0.88074	0.83362	0.05017			
80912	OP	0.63	0.34	0.45377	-0.76563	0.01013			
80924_01	OP	0.67	0.50	0.88070	-0.54099	0.07662			
80924_03	OP	0.35	0.37	0.89373	1.04066	0.13184			
80924_04	OP	0.58	0.49	1.11329	0.11959	0.20156			
80926	OP	0.35	0.36	0.70766	1.20115	0.12281			
80929	OP	0.50	0.31	0.35472	0.15145	0.01269			
80932_01	OP	0.63	0.29	0.60076	0.37600	0.33364			
80932_04	OP	0.65	0.30	0.88129	0.58446	0.43324			
80934_01	OP	0.54	0.31	0.45567	0.09802	0.07804			
80934_02	OP	0.70	0.35	0.56959	-0.60588	0.15463			
80960	OP	0.90	0.38	1.02001	-1.54809	0.19078			
81009	OP	0.81	0.41	0.77118	-1.27615	0.02648			
81015	OP	0.62	0.44	0.87978	0.07687	0.20127			
81019	OP	0.66	0.41	0.63657	-0.32195	0.16664			

UIN	Status	Pvalue	Ptbis	a	b	c	d1	d2	d3
81023	OP	0.67	0.48	0.76719	-0.48494	0.08891			
81045_03	OP	0.37	0.30	0.54227	1.29923	0.10940			
81045_04	OP	0.33	0.20	0.63841	1.98445	0.20563			
81053_01	OP	0.73	0.50	0.89763	-0.83703	0.05119			
81053_03	OP	0.71	0.34	0.49583	-1.21365	0.01435			
81121_02	OP	0.55	0.47	0.77389	-0.03353	0.06111			
81121_04	OP	0.59	0.57	1.16062	-0.11547	0.09082			
81121_05	OP	0.58	0.51	0.90746	-0.12555	0.08830			
81138_01	OP	0.68	0.48	1.01656	-0.24070	0.18494			
81138_02	OP	0.84	0.40	0.82187	-1.45640	0.01755			
81138_03	OP	0.69	0.42	0.65583	-0.71785	0.01867			
81154_01	OP	0.25	0.12	1.77336	1.99193	0.20925			
81154_02	OP	0.87	0.39	0.86225	-1.63798	0.02237			
81154_04	OP	0.55	0.25	0.32475	-0.21592	0.01466			
81183	OP	0.41	0.29	0.59104	1.34229	0.20872			
81197	OP	0.62	0.52	1.44166	0.09648	0.21764			
81198	OP	0.35	0.36	0.71077	1.26056	0.12761			
81204_03	OP	0.54	0.37	0.76275	0.44107	0.23542			
81204_04	OP	0.73	0.39	0.79286	-0.42677	0.31019			
81206_01	OP	0.45	0.26	0.34062	0.49442	0.02002			
81206_05	OP	0.69	0.50	1.27717	-0.19896	0.27742			
81212_01	OP	0.56	0.40	0.66403	0.02993	0.12681			
81212_04	OP	0.54	0.47	1.24568	0.37386	0.22538			
81212_05	OP	0.57	0.33	0.43454	-0.34040	0.02216			
81216_01	OP	0.59	0.28	0.48872	0.33220	0.23845			
81216_04	OP	0.45	0.40	0.76741	0.68353	0.12662			
81261	OP	0.72	0.28	0.38402	-1.55002	0.01579			
81269	OP	0.58	0.42	0.65469	-0.11232	0.04883			
81287	OP	0.49	0.49	1.19984	0.49763	0.14809			
81290	OP	0.16	0.45	0.53555	2.11261	0	1.10673	-0.23400	-0.87273
81291	OP	0.38	0.46	0.53526	0.71257	0	2.87436	-0.81876	-2.05560
81295	OP	0.76	0.35	0.56040	-1.27320	0.03318			
81298	OP	0.18	0.51	0.54965	1.54830	0	-0.27151	0.25082	0.02070
81300_02	OP	0.34	0.21	0.49196	1.83072	0.16507			
81300_05	OP	0.71	0.53	1.40980	-0.31009	0.25190			
81300_07	OP	0.33	0.58	0.80906	1.44705	0	2.64197	0.09893	-2.74090
81304_04	OP	0.55	0.43	1.16925	0.41397	0.25832			
81304_05	OP	0.66	0.25	0.33307	-1.16694	0.02376			
81304_08	OP	0.20	0.44	0.55214	2.53002	0	2.66936	-0.71006	-1.95930
81310_02	OP	0.77	0.45	0.85156	-0.95533	0.15096			
81310_05	OP	0.66	0.39	0.56172	-0.78157	0.00914			
81311_01	OP	0.57	0.19	0.62266	1.24521	0.42376			
81311_02	OP	0.50	0.41	0.68550	0.28835	0.10193			
81311_04	OP	0.59	0.45	0.88470	0.04921	0.17682			
81320_04	OP	0.45	0.37	0.71565	0.76025	0.14277			
81320_05	OP	0.45	0.36	0.69803	0.78662	0.15346			
81361	OP	0.77	0.40	0.68785	-1.14872	0.04744			

UIN	Status	Pvalue	Ptbis	a	b	c	d1	d2	d3
81364	OP	0.49	0.43	0.57147	0.15831	0.01223			
81365	OP	0.81	0.34	0.71954	-0.85031	0.33485			
81366	OP	0.38	0.23	1.07194	1.49423	0.26832			
81367	OP	0.50	0.30	0.64697	0.91555	0.25202			
81370	OP	0.78	0.49	1.07920	-0.70064	0.25003			
81374	OP	0.70	0.29	0.48305	-0.52284	0.22440			
81375	OP	0.67	0.44	0.68375	-0.64322	0.06351			
81376	OP	0.78	0.52	0.97942	-1.01938	0.03242			
81378	OP	0.63	0.51	0.83358	-0.28705	0.08258			
81380	OP	0.35	0.24	0.97815	1.66388	0.24101			
81381	OP	0.42	0.24	0.38163	1.09828	0.10910			
81383	OP	0.49	0.41	0.94915	0.71292	0.23647			
85142	OP	0.62	0.50	1.09214	0.06945	0.23535			
85201	OP	0.76	0.42	0.88518	-0.61919	0.28694			
85229	OP	0.85	0.49	1.11528	-1.41153	0.02093			

UIN=Unique Item Number; Status=Administration condition (OP = Operational item); Pvalue=Item p-value; Ptbis=Item Point Biserial; IRT 3PL and GPC model item parameters (a, b, c, d_k)