Maryland School Assessment (MSA)
Science

Grades 5 and 8

Technical Report
2009 Operational Test

April, 2010

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 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. In 2006, 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 2008-2009 test administration cycle.

Purpose

The purpose of this MSA Technical Report is to provide objective information regarding technical aspects of the 2009 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, 1999).

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.

Each of these facets in the MSA Science test development and use cycle is critical to validity of inferences based upon and interpretation of results. This technical manual covers all of these topics for the 2008-2009 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 administration was conducted in the spring of 2007, followed by the first operational test in 2008. The MSA Science test is currently targeted at grade 5 and grade 8 students 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 to measures performance levels of students, schools and districts against the Science academic standards. Second, it provides information about what students learned in school to parents, teachers, and educators to inform the 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. According to MSDE's website, the MSC defines what students should know and be able to do and "is the document that aligns the Maryland Content Standards and the Maryland Assessment Program". 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 and begin to narrow the focus for teachers. 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 different 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 three test administrations-a census field test in 2007 and two operational tests (2008 and 2009). All administrations were conducted under the same testing conditions. Accordingly, the field test was designed so that it met the requirements of the operational administration test blueprint. The major difference is that there were fewer scored items on the operational form, but 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 (yielding a student score) items and 11 field test items for grade 5. 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 2009 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 2009 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.

MSA Science Operational 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 Table 1 and 2.

Table 1. Grade 5 MSA Science Blueprint

<table>
<tr><th colspan="2">Standard</th><th>Number of Selected Response Items (1 point)</th><th>Brief Constructed Response Items (0 - 3 Points)</th><th>Tot Items</th><th>Tot Points</th></tr>
<tr><td>1.0</td><td>Skills and Processes</td><td>9</td><td>1</td><td>10</td><td>12</td></tr>
<tr><td>2.0</td><td>Earth/Space Science</td><td rowspan="5">9 to 12 (where 3 standards have 12 SRs and 2 have 9 SRs)</td><td rowspan="5">2 - for standards with only 9 SR items</td><td>9 to 12</td><td>12</td></tr>
<tr><td>3.0</td><td>Life Science</td><td>9 to 12</td><td>12</td></tr>
<tr><td>4.0</td><td>Chemistry</td><td>9 to 12</td><td>12</td></tr>
<tr><td>5.0</td><td>Physics</td><td>9 to 12</td><td>12</td></tr>
<tr><td>6.0</td><td>Environmental</td><td>9 to 12</td><td>12</td></tr>
<tr><td colspan="2">Total</td><td>63</td><td>3</td><td>66</td><td>72</td></tr>
</table>

Note: All standards have 12 score points broken down as follows:
-- 3 of standards 2.0 to 6.0 will have 12 SR items, the rest will have 9.
-- 2 of standards 2.0 to 6.0 with only 9 SR items will have 1 BCR item.

Table 2. Grade 8 MSA Science Blueprint

<table>
<tr><th colspan="2">Standard</th><th>Number of Selected Response Items (1 point)</th><th>Brief Constructed Response Items (0 - 3 Points)</th><th>Tot Items</th><th>Tot Points</th></tr>
<tr><td>1.0</td><td>Skills and Processes</td><td>9</td><td>1</td><td>10</td><td>12</td></tr>
<tr><td>2.0</td><td>Earth/Space Science</td><td rowspan="5">9 to 12 (where 2 standards have 12 SRs and 2 have 9 SRs)</td><td rowspan="5">3 - for standards with only 9 SR items</td><td>9 to 12</td><td>12</td></tr>
<tr><td>3.0</td><td>Life Science</td><td>9 to 12</td><td>12</td></tr>
<tr><td>4.0</td><td>Chemistry</td><td>9 to 12</td><td>12</td></tr>
<tr><td>5.0</td><td>Physics</td><td>9 to 12</td><td>12</td></tr>
<tr><td>6.0</td><td>Environmental</td><td>9 to 12</td><td>12</td></tr>
<tr><td colspan="2">Total</td><td>60</td><td>4</td><td>64</td><td>72</td></tr>
</table>

Note: All standards have 12 score points broken down as follows:
-- 2 of standards 2.0 to 6.0 will have 12 SR items, the rest will have 9.
-- 3 of standards 2.0 to 6.0 with only 9 SR items will have 1 BCR item.

MSA Science 2009 Operational Test Construction

The 2009 operational tests were created according to the test blueprints (see Table 1 and 2) and reflective of the Voluntary State Curriculum (VSC) 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 2008 tests in terms of content, total number of items/score points, and item types. Additionally, each operational form was created with five unique sets of embedded field test items (see MSA Science 2009 Field Test Design). 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 2009 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 VSC 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. Items with reasonably strong point biserial correlations (>.30) and a spread of item difficulties in line with the 2008 forms were guiding principles. Items flagged for any reason based on the data review criteria (also including differential item functioning as described later), were discouraged from being used. 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 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 2009 Field Test Design

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. Field test item sets 1-5 were embedded in Form A and 6-10 in Form B. In other words, operational forms 1 through 5 share the same operational items and are differentiated by a unique field test item set within each form. Table 3 presents a graphical representation of this field test design. Items common to both forms are also depicted.

Table 3. 2009 MSA Science Test Form Design

Operational Items	Field test Item Sets									
	1	2	3	4	5	6	7	8	9	10
Form A	X									
		X								
			X							
				X						
Common					X					
Items						X				
Form B							X			
								X		
									X	
										X

MSDE and Pearson worked together to finalize the structure of the 2009 field test forms. At each grade, 10 field test forms were produced. The intent of the test build 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 equivalent to the operational base forms plus embedded field test in terms of total numbers of SR and BCR items. All 10 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 10 field test forms per grade. On a per form basis, initial item selections were performed by Pearson and then shared with MSDE for review and approval. Since Form 1 at each grade was the Braille/large print form, items were selected for Form 1 on the basis of feedback provided by the low-vision panel.

The 2009 forms (and all subsequent operational assessments) 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.

Item Development and Review

MSDE and Pearson worked together to define the development targets in support of the 2009 field test. Overall, development was structured to spread the items across the six standards specified within the Maryland (Voluntary) State Curriculum (VSC/MSC) and across the topics, indicators, objectives and assessment limits within each standard. Targets were developed at both grades 5 and 8; item development began once the development targets were finalized. The target number of items developed in 2008 for the 2009 administration was approximately 170 items for each grade: 150 SR and 20 BCR items.

During 2007 published technical passages to be approved for item development were selected and reviewed by Pearson content staff, MSDE content experts, and three separate Maryland content and bias committees. An item writer training was held in early December 2008. Current or former non-Maryland Science educators were recruited to write items and lab stimuli on behalf of the program. During the training, writers were introduced to a number of topics by both MSDE and Pearson staff. Topics for training included:

- an introduction to the VSC/MSC;
- the concept of assessment limits;
- the types of items on the MSA Science test;
- elements of universal design in assessment (see Thompson, Johnstone, & Thurlow, 2002 for an overview of universal design within large scale testing);
- how to develop items aligned to standards;
- identifying potential bias/sensitivity issues within the materials written, and;
- guidelines for writing SR and BCR items.

Following training, writers were given an opportunity to begin drafting items, which were then reviewed by Pearson content staff.

Once Pearson received items from writers, each item underwent an extensive internal review by Pearson content specialists for total item quality, including but not limited to:

- accurate Science content;
- appropriate and engaging context;
- effectiveness as a measurement of assessment limits within the VSC/MSC;
- age and grade-level appropriate language and vocabulary;
- adherence to established MSDE style guidelines.

Additionally, Pearson content specialists reviewed all items within each grade for the full range of item difficulty and consideration of a range of cognitive complexity. Cognitive complexity refers how items are solved. For example, complexity may range from items where students only need to rely on memory to answer a question versus having to evaluate and synthesize something to respond correctly. After this review, items went through an iterative development process between content specialist and copy editors, universal design specialists, and research librarians. In addition, all art and graphical supports for the items were produced. Finally, all BCR items were reviewed by Pearson Performance Scoring Center staff for scorability. Once Pearson completed the internal development, items were released to MSDE for review via Pearson's Item Tracker system. In May of 2008, Pearson and MSDE content experts met to review and discuss each new item and collaborate on revisions. Once revisions were made and reviewed again through the internal Pearson development team, the items were prepared for another series of content and bias reviews in Maryland.

Review panels of Maryland residents were convened in July 2008. Three different panels were convened to review items for each grade. Content review was conducted at each grade by Maryland educators within the appropriate grade range to further confirm content accuracy and grade-level appropriate vocabulary and language, and to identify and discuss potential improvements to the item stem or distractors. A separate bias/sensitivity panel at each grade was convened to examine the items for any possible socio-economic, geographical, cultural or gender biases. Finally, another committee of educators reviewed item text and graphics with particular focus on possible issues for blind or visually impaired students. Before reviewing materials, MSDE and Pearson provided an overview to the panelists on the purpose of each panel, the VSC/MSC, and the criteria by which they were asked to evaluate the items. Since the evaluation criteria were different, the content panelists and bias/sensitivity panelists were trained separately.

Content panelists were asked to evaluate the materials on the basis of the following criteria:

- alignment to the VSC/MSC;
- clarity and grade-appropriateness of text and graphic supports;
- accuracy of the underlying Science content.

Bias/sensitivity panelists were asked to evaluate the materials as an additional check on whether the materials:

- reflected favoritism towards a gender or ethnic group;
- were free of potentially offensive or inappropriate language;
- discriminated in any way against individuals who have special needs;
- contained any underlying assumptions not shared across ethnic, racial, and gender groups, socioeconomic levels, and geographic areas;
- contained language and/or dialect that is not commonly used across the state or has different connotations in different parts of the state;
- had graphic supports that were appropriate and accessible for all students.

In addition to the panels reviewing the items to be field tested in spring 2009, separate bias and content panels were convened for both grade 5 and grade 8 to read and evaluate the technical passages that were proposed to be used on the spring 2010 embedded field test. On the basis of input from these groups, MSDE and Pearson selected the passages for which items would be developed for the 2010 field test.

Following the panels, MSDE and Pearson met to reconcile the comments from the various groups. Each item and stimulus was reviewed along with the comments from the bias, content and low-vision panels. From this, a final decision was made by MSDE with respect to all edits and the disposition of the item.

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 Tables 4 and 5, respectively.

Table 4. Demographic Characteristics of Grade 5 and Grade 8 Sample for Overall, Online, and Paper

	Grade			
	5		8	
	N	%	N	%
Mode of Administration				
Online	38671	64.15	44945	71.74
Paper	21613	35.85	17707	28.26
Form				
1	4688	7.77	5808	9.27
2	5788	9.59	5919	9.45
3	7329	12.15	7286	11.63
4	5929	9.83	6013	9.60
5	5791	9.6	5906	9.43
6	5926	9.82	5891	9.40
7	7255	12.02	6066	9.68
8	5867	9.72	5975	9.54
9	5899	9.78	7611	12.15
10	5861	9.71	6177	9.86
Gender				
Female	30750	48.93	30525	48.72
Male	29521	50.97	32106	51.24
Unknown	62	0.10	21	0.03
Ethnicity				
Native American	224	0.37	237	0.38
Asian	3674	6.09	3562	5.69
African American	22867	37.90	23683	37.80
White	27964	46.35	29716	47.43
Hispanic	5542	9.19	5433	8.67
Unknown	62	0.10	21	0.03
All	**60333**	**100**	**62652**	**100**

* Differences in values reflect missing data

Distribution of Students across Forms

As described, MSA Science test forms are comprised 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 5 presents this distribution of tests across forms by mode of administration at each grade. Within-form overages (i.e. online Form 3) reflect the inclusion of additional forms for special accommodations (i.e. read-aloud, audio presentation, etc.).

Table 5. Distribution of Forms by Grade

		Form									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Grade 5	Online	2599	3671	5225	3785	3637	3755	4948	3686	3700	3665
	Paper	2084	2113	2099	2139	2149	2166	2302	2176	2194	2191
	Overall	4683	5784	7324	5924	5786	5921	7250	5862	5894	5856
Grade 8	Online	4117	4221	5581	4291	4183	4164	4322	4236	5373	4457
	Paper	1691	1698	1705	1722	1723	1727	1744	1739	2238	1720
	Overall	5808	5919	7286	6013	5906	5891	6066	5975	7611	6177

Key Check Analysis of Operational Test Data

Using preliminary data collected from the 2009 operational test (a minimum of 200 responses were required for each form by mode of administration), Pearson computed Classical Test Theory 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. All operational MSA Science items were confirmed as correctly keyed and functioning sufficiently within the statistical parameters (described below) to conduct the classic and IRT analysis described in the next sections.

The key check analysis included the following Classical Test Theory 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

Analysis

Following the complete processing of answer documents, student demographic and item response data were transmitted to Pearson's Psychometric and Research Services division. Pearson psychometric staff had primary responsibility for analyzing MSA Science data to ensure accuracy and validity of scoring. Most of the psychometric work was carried out using SAS Version 9.1 and MULTILOG 7.0, commercially available statistical analysis software. Traditional item analysis and data file QC analysis were conducted with SAS programs. Item response theory (IRT) analysis were conducted with the MUTLTILOG program (Thissen, Chen, & Bock, 2003). MULTILOG allows for estimation of IRT item parameters for dichotomously or polytomous scored items. It has been thoroughly tested and is currently utilized by several high-stakes testing programs administered by Pearson.

All technical support and analysis were carried out in accordance with both the *Standards* (AERA, APA, & NCME, 1999) and the Pearson Quality Assurance Program. Pearson staff verified the MSA Science data and analysis process at several steps in the procedure. This included verification of the SAS and MULTILOG programs prior to use on actual field data through review by a second member of the psychometric services staff and by using simulated data sets. Additionally, the output from the traditional and IRT item analysis programs were verified for out of range values and for consistent results across programs.

Classical Item Analysis

The following classical item statistics that 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 classical item analysis were banked for use during the construction of subsequent MSA Science tests. P-value and point-biserial statistics for the 2009 MSA operational items are reported in Appendix A.

IRT Calibration

Pearson used a concurrent calibration IRT estimation procedure for placing all Form A and Form B operational MSA Science items on a common theta scale that was then equated to the original 2007 base scale (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).

<u>Dichotomous Item Response Theory Model</u>

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:

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

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 $(0, 1, \ldots, m_j-1)$, 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 $k = 0, 1, 2, \ldots, m_j-1$ thresholds denoting the intersections of the respective m_j response functions.

Calibration of the mixed test format (3PL/GPC model) items was conducted using MULTILOG 7.0 (Thissen, Chen, & Bock, 2003) and included only the students who:

- attempted at least one item on the test,
- attempted at least one BCR item, and
- the student's score was not invalidated.

MULTILOG estimates parameters simultaneously for dichotomous and polytomous items via marginal maximum likelihood procedures. 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 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 2009 MSA Science tests were concurrently calibrated, it was necessary to place each respective scale (Grade 5 and Grade 8) onto the originating 2007 base scale. This was carried out using what is referred to as a common item, non-equivalent groups design (CINEG; Kolen & Brennan, 2004). In this case, the common item sets from the operational forms were comprised of *all* operational SR items. That is, all operational items aside from BCRs served as linking items back to the base scale. For the item parameter estimates reflecting the base form, the most current parameter estimates were used, whether from the 2007 or 2008 field test calibration or from the 2008 operational administration.

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). It is customary to evaluate changes in item parameters, and evaluate how those changes affect the results of equating. Thus, it was necessary to examine the stability of the MSA Science anchor item parameters after equating. Specifically, Pearson evaluated stability in the operational linking item parameters by examining differences in the originating (base) and transformed item characteristic curves. All items used for linking the 2009 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 for in 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 classical item analysis, 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 6 items removed from the grade 5 common item set and 1 item removed from the grade 8 common item set. The final transformation constants for each grade following this procedure are listed in Table 6.

Table 6. Operational Transformation Constants

	Grade 5		**Grade 8**	
	Slope	**Intercept**	**Slope**	**Intercept**
Operational (09 OP items -> 07 base scale)	1.006562	0.12983	1.066698	0.132356

The transformation constants were applied to the 2009 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

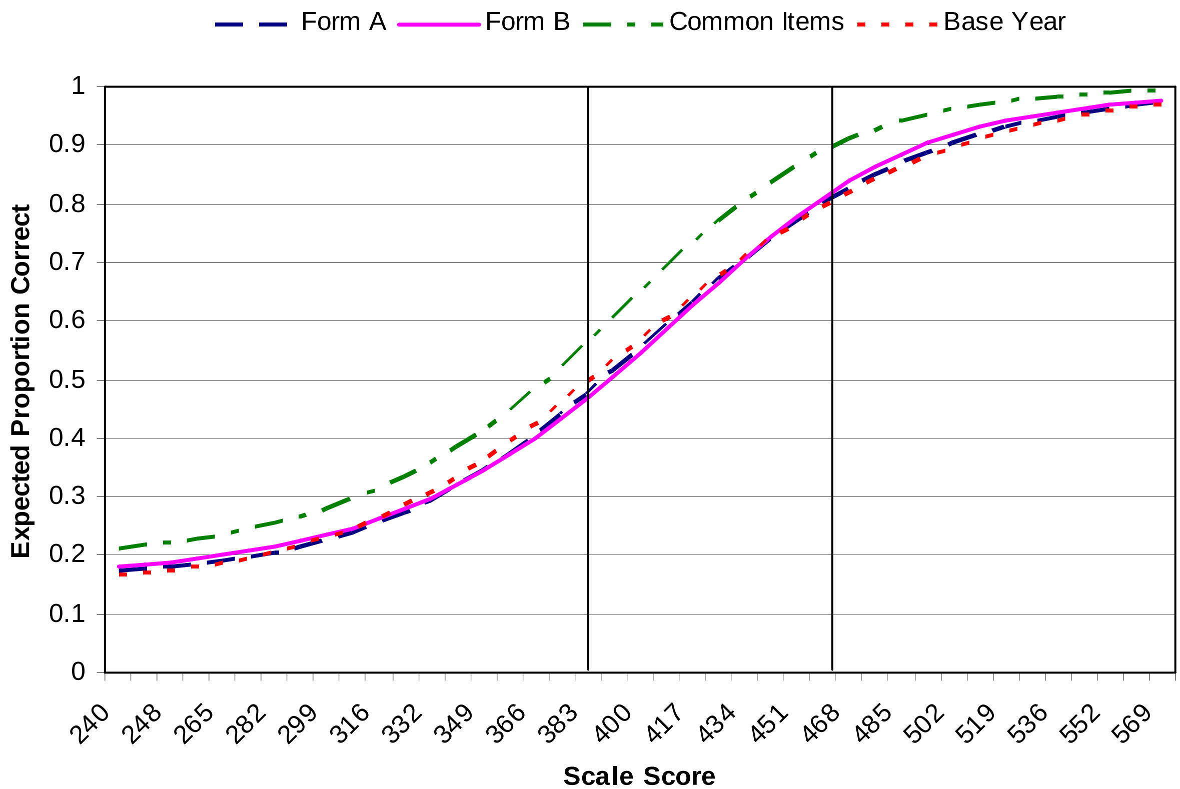
IRT item parameter estimates were used to generate test characteristic curves (TCCs), test information functions (TIFs), and conditional standard errors of measure (CSEM). These indices were computed for each of the current year operational forms (A and B), form-to-form linking items (common items), and the base-year operational item pool. In order to facilitate comparisons of these curves, the TCC, TIF, and SEM values were divided by the total number of score points for each form so that the curves can be plotted on the same scale.

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 Form A and B curves and the originating base scaling that was established in 2007. These are the primary comparisons because they reflect how well the 2009 forms reflect the original scale. It should be noted that this base scale was initially set on the entire 2007 item pool and this is reflected in the figures below (as Base Year curves).

Figure 1 shows the overlaid TCC plots for Form A, Form B, form-to-form linking items and base-year item pool for grade 5. These plots illustrate that the operational form A and B scales are very closely aligned to the 2007 base scale (and to each other). With respect to the Form A/B common item sets, recall that these are a common set of 20 items appearing on both forms A and B which allow for a concurrent calibration to be carried out while placing all 2010 items onto a common metric. This item set is ideally chosen to reflect a miniature version of the overall test in terms of content as well as statistical characteristics. In this instance, however, content considerations discussed between Pearson and MSDE content experts in arriving at the best *overall* test forms weighed more heavily into the final selection of the common item set and is likely the reason for the differences in the Common Item curve relative the other curves.

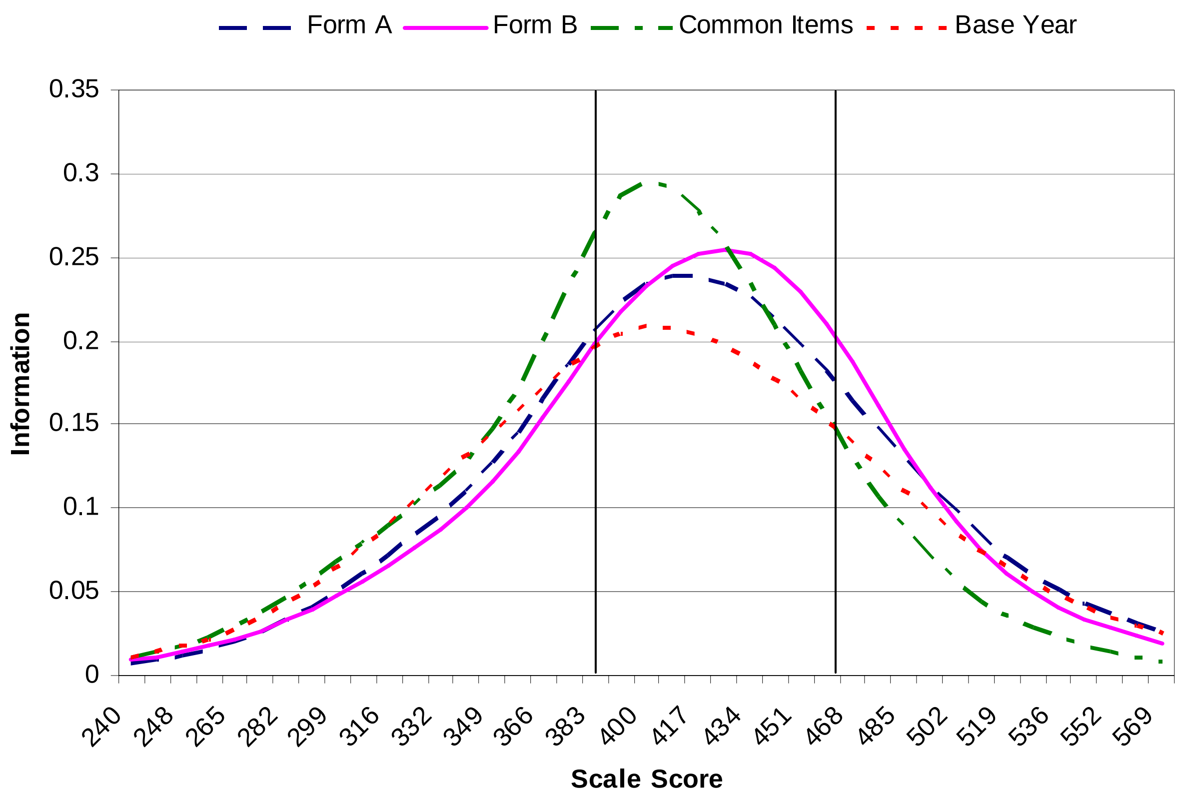
Figure 2 also displays test information curves for Form A, Form B, form-to-form linking items and the base-year item pool. Figure 3 illustrates the conditional standard error of measurements for the four item sets. The vertical lines in each figure represent the location of the Proficient and Advanced performance standards on the reportable 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 of the Grade 5 Science Test



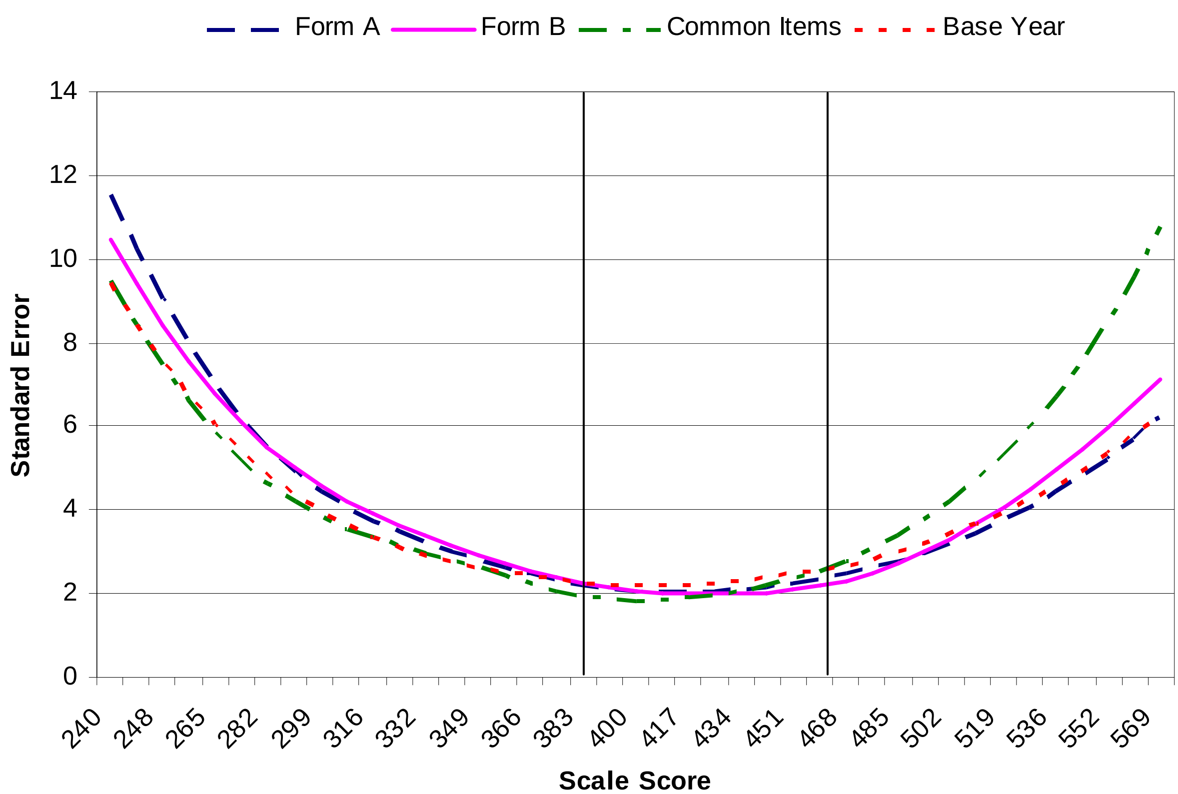
Note: The 2 vertical lines reflect the Proficient and Advanced cut scores which result in three performance levels: Basic, Proficient, and Advanced (Proficient Cut = 391, Advanced Cut = 467).

Figure 2. Test Information Function of the Grade 5 Science Test



Note: The 2 vertical lines reflect the Proficient and Advanced cut scores which result in three performance levels: Basic, Proficient, and Advanced (Proficient Cut = 391, Advanced Cut = 467).

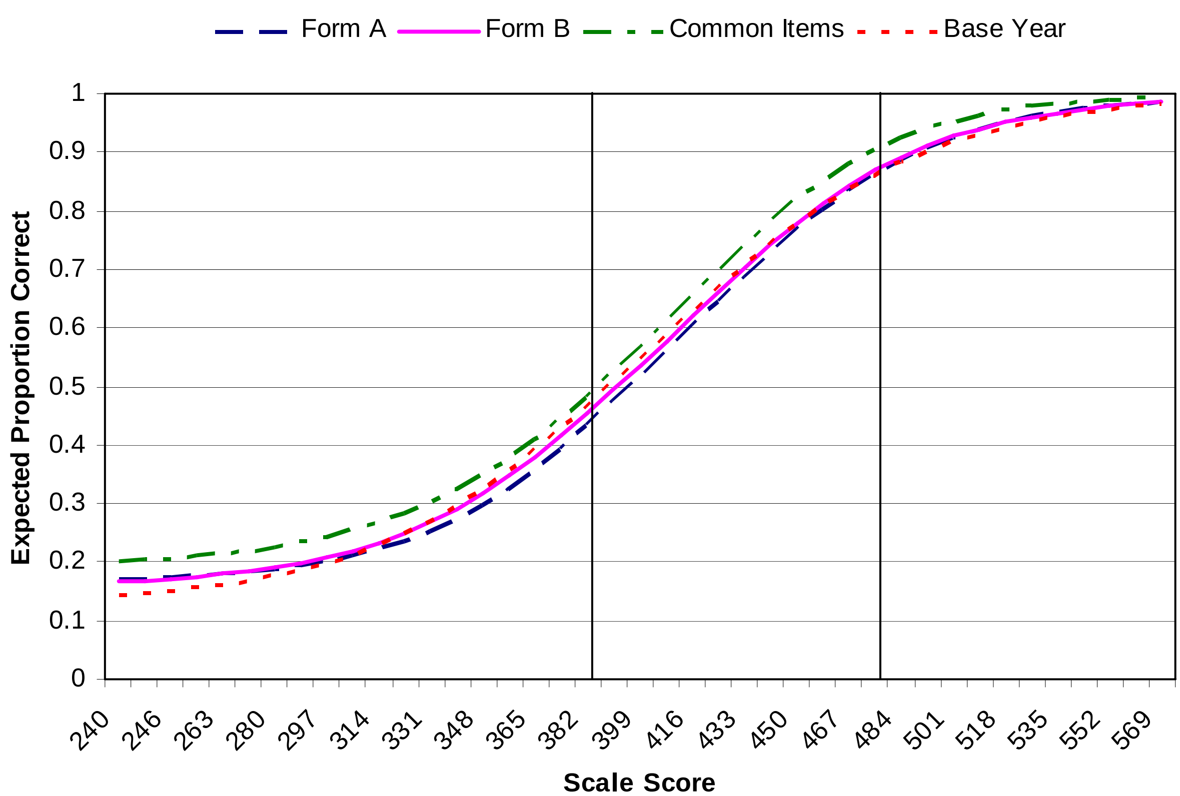
Figure 3. Conditional Standard Error of Measurement for the Grade 5 Science Test



Note: The 2 vertical lines reflect the Proficient and Advanced cut scores which result in three performance levels: Basic, Proficient, and Advanced (Proficient Cut = 391, Advanced Cut = 467).

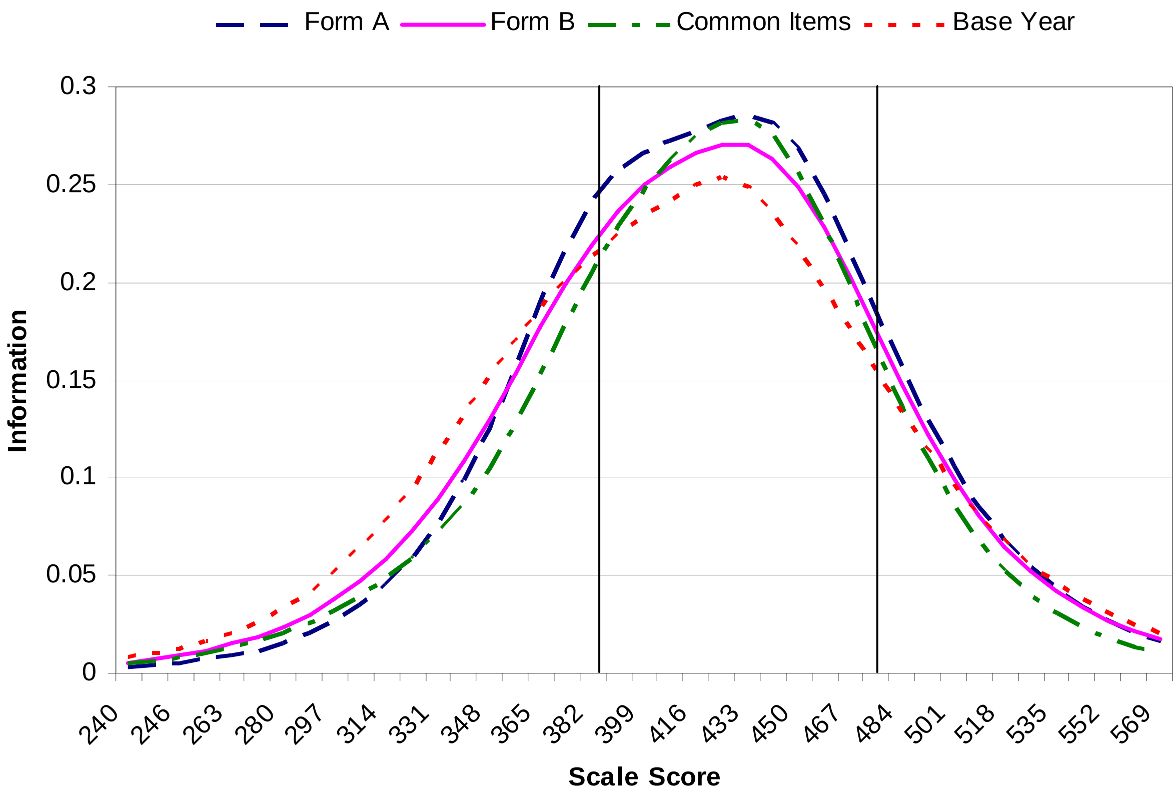
Similar to grade 5, IRT item parameter estimates were used to generate characteristic curves (TCCs), test information functions (TIFs), and conditional standard errors of measure (CSEM) were computed for each of the base forms, form-to-form linking items, and base-year operational test for grade 8. Figure 4 shows the overlaid TCC plots for Form A, B, linking item and base-year pools. The TCC and TIF values were divided by the total number of score points for each form so that the curves can be plotted on the same scale. Figure 5 displays test information curves for Form A, B, linking item and base-year pools. Figure 6 illustrates the conditional standard error of measurements for the four item sets. The vertical lines in each figure represent the location of the Proficient and Advanced performance standards on the reportable 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 of the Grade 8 Science Test



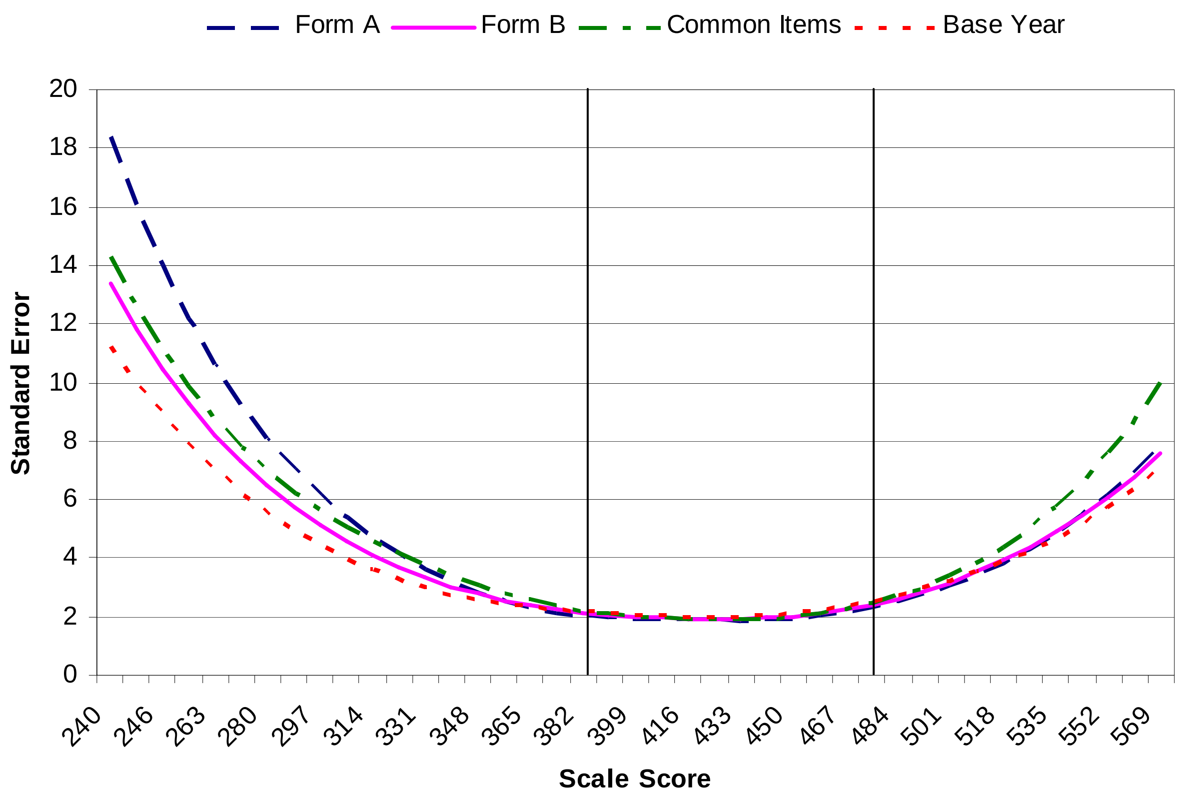
Note: The 2 vertical lines reflect the Proficient and Advanced cut scores which result in three performance levels: Basic, Proficient, and Advanced (Proficient Cut = 387, Advanced Cut = 478).

Figure 5. Test Information Function of the Grade 8 Science Test



Note: The 2 vertical lines reflect the Proficient and Advanced cut scores which result in three performance levels: Basic, Proficient, and Advanced (Proficient Cut = 387, Advanced Cut = 478).

Figure 6. Conditional Standard Error of Measurement for Grade 8 Science Test



Note: The 2 vertical lines reflect the Proficient and Advanced cut scores which result in three performance levels: Basic, Proficient, and Advanced (Proficient Cut = 387, Advanced Cut = 478).

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 transformed to a reporting scale which can be more meaningfully interpreted by students, teachers and other stakeholders. In order to facilitate the use and interpretation of the results of the 2009 MSA Science operational administration, scale scores were created through the application of scaling constants determined from the base 2007 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 7.

Table 7. 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

Pearson used an internally developed software program called IRT Score Estimation (ISE; Chien, Hsu, & Shin, 2007) to conduct pattern scoring for the spring 2009 administration of the MSA Science tests for grades 5 and 8. The program has been extensively tested and compared to commercially available software programs (e.g., MULTILOG, PARSCALE; Tong, Um, Turhan, Parker, Shin, Chien, & Hsu, 2007). The report concluded 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, technical manual and evaluation report and are available upon request.

The 2009 operational scores were estimated by the pattern scoring approach. The 2009 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. The theta estimates were transformed onto the MSA Science operational scale using the scaling constants described above.

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 (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 scaling 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. 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's denote 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:

$$Stratified\,\alpha = 1 - \frac{((\sigma_{SR}^2(1-\rho_{SR}) + (\sigma_{CR}^2(1-\rho_{CR}))}{\sigma_t^2}$$

where

σ_{SR}^2 = variance associated with SR items;

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

σ_t^2 = 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 8.

Table 8. Reliability Estimate by Grade, Form, Gender and Ethnicity

Group		**Grade 5**		**Grade 8**	
		Form A	**Form B**	**Form A**	**Form B**
Overall		0.92	0.92	0.94	0.94
Gender	Male	0.93	0.93	0.95	0.94
	Female	0.92	0.92	0.93	0.93
Ethnicity	Native American	0.90	0.91	0.93	0.94
	Asian	0.92	0.92	0.94	0.93
	Black	0.90	0.90	0.91	0.91
	White	0.90	0.90	0.93	0.92
	Hispanic	0.90	0.90	0.93	0.92

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 2009 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 2009 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 2 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 2009 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 9. 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 9. Scale score cut scores for grades 5 and 8 MSA Science.

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

Tables 10 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. 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	36	56	8	405	45.7	60592	33	58	8	408	44.0	38946	42	50	8	400	48.2	21646
Gender																		
Female	37	56	7	404	44.3	29662	34	59	7	407	42.9	19146	42	50	7	399	46.4	10516
Male	35	55	9	406	47.0	30914	32	58	9	409	45.0	19799	41	50	9	400	49.7	11115
Ethnicity																		
Native American	30	64	5	409	38.9	227	32	63	4	409	37.4	158	26	67	7	409	42.5	69
Asian	19	64	18	426	44.7	3683	18	68	15	425	42.5	2170	20	59	22	428	47.7	1513
Black	55	43	2	384	41.4	22999	54	45	2	386	40.0	12543	58	41	2	381	42.9	10456
White	20	67	13	423	40.8	28096	21	67	12	422	40.1	21334	18	65	17	427	42.6	6762
Hispanic	51	46	3	388	42.5	5571	50	47	3	389	43.1	2740	52	46	2	387	41.9	2831
Note: Performance Levels, B=Basic, P=Proficient, A=Advanced																		

Tables 11 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 5 as well as by administration mode.

Table 11. 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	60	5	403	49.8	62767	33	62	5	405	47.8	44994	38	56	5	398	54.1	17773
Gender																		
Female	34	61	4	402	46.3	30559	33	63	4	403	44.4	22036	38	58	5	399	50.7	8523
Male	35	59	6	403	52.8	32185	33	61	6	406	50.8	22958	39	55	6	398	57.0	9227
Ethnicity																		
Native American	38	59	3	397	49.6	240	40	56	4	398	47.2	182	33	67	<1	395	56.8	58
Asian	14	72	14	431	45.7	3564	15	73	13	430	45.6	2148	14	71	15	432	45.9	1416
Black	56	44	1	377	45.8	23751	53	46	1	381	43.7	16615	62	37	<1	368	49.2	7138
White	18	74	8	423	41.9	29742	18	75	8	423	41.3	22583	17	74	9	425	43.6	7159
Hispanic	50	49	1	383	47.4	5445	51	47	2	382	47.8	3466	48	51	1	385	46.7	1979
Note: Performance Levels, B=Basic, P=Proficient, A=Advanced																		

Field Test Item Analysis and Calibration

Key Check Analysis of Field Test Data

Using preliminary data collected from the 2009 administration (a minimum of 200 responses were required for each form by mode of administration), Pearson computed Classical Test Theory 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). These analyses were carried out in the same manner as those described for the operational key check analysis (see page 9). Any items identified during this process were presented to Pearson content specialists for review to ensure that items were keyed properly. No mis-keyed items were identified on either of the MSA Science tests.

Classical Item Analysis

The following classical item statistics that 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 classical item analysis were banked for use during the construction of subsequent MSA Science tests. P-value and point-biserial statistics for the 2009 MSA field test items are reported in Appendix A.

Field Test Calibration

Field test items are embedded within each session of the MSA Science tests with unique items appearing in the same positions across the field test forms. A total of ten field test forms were created by embedding unique field test items into each operational form. Table 3 provides a graphical depiction of the field test design. This design ensured that one of two sets of operational test items were common to each field test form. This allows all field test item parameters to be estimated concurrently, thus placing all items on a common scale as is done with the two operational forms during operational equating. During this concurrent calibration all items (operational and field test) are freely estimated. As a result the item parameter estimated obtained for the field test items are not on the base scale. In order to place these parameter estimates on the base scale so that they may be use to construct equivalent operational test forms for subsequent administrations the Stocking and Lord procedure is used to calculate transformation constants with the anchor set being formed from all of the operational items (comparing the operational item parameters obtained during field test calibration to those banked following post-equating). This process was used to place all 2009 field test items on the base scale. The transformation constants derived and applied at each grade during this are shown in Table 12. The IRT parameters for grade 5 and 8 field test items are presented in Appendix A.

Table 12. Field Test Transformation Constants

	Grade 5		Grade 8	
	Slope	**Intercept**	**Slope**	**Intercept**
Field Test (09 FT items -> 09 OP items)	1.008967	0.121778	1.065291	0.111547

Differential Item Functioning (DIF) Analysis

One of the goals of the MSA Science test development is to assemble a set of items that provides a measure of a student's ability that is as fair and accurate as possible for all subgroups within the population. Differential item functioning (DIF) analysis refers to procedures that assess whether items are differentially difficult for different groups of examinees. DIF procedures typically control for overall between-group differences on a criterion, usually total test scores. Between-group performance on each item is then compared within sets of examinees having similar test scores. If the item is differentially more difficult for an identifiable subgroup when conditioned on ability, the item may be measuring something different from the intended construct. However, it is important to recognize that DIF-flagged items might be related to actual differences in relevant knowledge or skills or statistical Type 1 error. As a result, DIF statistics are used to identify potential sources of item bias. Subsequent review by content experts and bias committees are required to determine the source and meaning of performance differences. In the MSA Science DIF analysis, DIF statistics were estimated for all major subgroups of students with sufficient sample size: Black, Hispanic and Female[1]. Items with statistically significant differences in performance were flagged so that items could be carefully examined for possible biased or unfair content that was undetected in earlier fairness and bias content review meetings held prior to form construction.

Pearson used the Mantel-Haenszel (MH) chi-square approach to detect DIF in SR items. Pearson calculated the Mantel-Haenszel *delta* statistic (MH D-DIF, Holland & Thayer, 1988) to measure the degree and magnitude of DIF. The student group of interest is the *focal* group, and the group to which performance on the item is being compared is the *reference* group. The referent groups for this DIF analysis were White for ethnicity and male for gender. The focal groups were females and minority ethnicity groups.

Items were separated into one of three categories on the basis of DIF statistics (Holland & Thayer 1988; Dorans & Holland 1993): negligible DIF (category A), intermediate DIF (category B), and large DIF (category C). The items in category C, which exhibit significant DIF, are of primary concern.

Positive values of *delta* indicate that the item is easier for the *focal* group, suggesting that the item favors the *focal* group. A negative value of *delta* indicates that the item is more difficult for the *focal* group. The item classifications are based on the Mantel-Haenszel chi-square and the MH delta (Δ) value as follows:

- The item is classified as C category if the absolute value of the MH delta value (i.e., $|\Delta|$) is significantly greater than 1 and also greater than or equal to 1.5.
- The item is classified as B category if the MH delta value (Δ) is significantly different from 0 and either the absolute value of the MH delta ($|\Delta|$) is less than 1.5 or the absolute value of the MH delta ($|\Delta|$) is not significantly different from 1.

[1] DIF analysis on the Asian students was not conducted due to small sample size.

- The item is classified as A category if the delta value (Δ) is not significantly different from 0 or the absolute value of delta (|Δ|) is less than or equal to 1.

The effect size of the standardized mean difference (SMD) was used to flag DIF for the BCR items. The SMD reflects the size of the differences in performance on CR items between student groups matched on the total score. The following equation defines SMD:

$$SMD = \sum_k w_{Fk} m_{Fk} - \sum_k w_{Fk} m_{Rk}$$

where $w_{Fk} = n_{F+k} / n_{F++}$ is the proportion of focal group members who are at the kth stratification variable, $m_{Fk} = (1/n_{F+k})F_k$ is the mean item score for the focal group in the kth stratum, and $m_{Rk} = (1/n_{R+k})R_k$ is the analogous value for the reference group. In words, the SMD is the difference between the unweighted item mean of the focal group and the weighted item mean of the reference group. The weights applied to the reference group are applied so that the weighted number of reference group students is the same as in the focal group (within the same ability group). The SMD is divided by the total group item standard deviation to get a measure of the effect size for the SMD using the following equation:

$$\text{Effect Size} = \frac{SMD}{SD}$$

The SMD effect size allows each item to be placed into one of three categories: negligible DIF (AA), moderate DIF (BB), or large DIF (CC). The following rules are applied for the classification (Allen, Carlson & Zalanak, 1999). Only categories BB and CC were flagged in the results.

- The item is classified as CC category if the probability is <.05 and if |Effect Size| is >.25.
- The item is classified as BB category if the probability is < .05 and if .17<|Effect Size|≤.25.
- The item is classified as AA category if the probability is >.05 or |Effect Size| is ≤ .17.

Table 13 summarizes the results of the DIF analysis appearing in Appendix B for SR (B/C) and BCR (BB/CC) items. Items with a statistical indication of DIF were reviewed for bias by subject matter experts during data review.

Table 13. DIF Flag Summaries from all MSA Science Field Test Items

Grade	DIF Classification Level				Total
	B	BB	C	CC	
5	6	2	0	2	10
8	4	8	0	1	12

Data Review of the Field Test Items

Background

Data review represents a critical step in the test development cycle. Pearson psychometricians provided a list of flagged items for the 2009 MSA Science field test data review based on the following criteria:

SR items will be flagged if:

- P-value < .10 or P-value > 0.90
- Point biserial correlation < 0.30
- Item omission > 5%
- Incorrect distractor p-value > 0.40
- Incorrect distractor point biserial correlation > 0.05
- 100% non-response to any distractor
- IRT *a* parameter < 0.50
- IRT *b* parameter < -4.00, or IRT *b* parameter > 4.00
- IRT *c* parameter > 0.50
- C level DIF

BCR items will be flagged if:

- BCR mean < 0.30 or BCR mean > 2.70
- Point biserial correlation < 0.30
- Any score point where 0% of students earn that score
- IRT *a* parameter < 0.50
- IRT *b* parameter < -4.00, or IRT *b* parameter > 4.00
- IRT step values (*d*) < -4.00, or IRT step value > 4.00
- CC level DIF

The flagged items were reviewed by Pearson Content team and MSDE content experts. The final decision about the suppression of the flagged items was made in collaboration between MSDE and Pearson.

Results of Data Review

A total of 46 items in grade 5 and 43 items in grade 8 were inspected during data review as a result of the item not meeting the statistical flagging criteria. Five of the 46 total flagged were rejected from the grade 5 pool and nine of the 43 flagged items for grade 8 were rejected.

Validity

Pearson subscribes to the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999). The standards define validity as

> … the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing and evaluating tests. The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations.

Validity can be established through the collection of evidence to demonstrate the alignment of item content with the curriculum, compliance to the test specifications, test fairness, and valid uses and interpretations of test scores. This section describes various analysis to evaluate the validity and reliability evidence for the 2009 MSA Science test.

Content-related Validity

All MSA Science items were explicitly developed to measure the specific knowledge and skills described in the Voluntary State Curriculum (VSC). In addition, the alignment of the items to the standards was reviewed and verified independently by multiple content reviewers and Maryland educators. The MSA Science operational items were handed over to Pearson after the extensive reviews by the Maryland educators and external reviewers.

Construct-related Validity

Construct validity refers to what test scores mean and what kinds of inferences they support. Construct validity is the central concept underlying the MSA Science test validation process. Evidence for construct validity is comprehensive and integrates evidence from both content- and criterion-related validity.

Construct-related validity evidence (internal consistency validity evidence) can come from many sources. The American Psychological Association provides the following list of possible sources (AERA, APA & NCME, 1999):

- high inter-correlations among assessment items or tasks attest that the items are measuring the same trait, such as a content objective, sub-domain or construct;
- substantial relationships between the assessment results and other measures of the same defined construct;
- little or no relationship between the assessment results and other measures which are clearly not of the defined construct;
- substantial relationships between different methods of measurement regarding the same defined construct;
- relationships to non-assessment measures of the same defined construct.

The collection of construct-related evidence is a continuous process, and at present substantial evidence is available representing internal structure (the first of the five bullets above). This section describes four sources of internal structure-based construct validity evidence for the MSA Science test: item-total/point-biserial correlations, inter-correlation among standards/subscales, unidimensionality, and DIF analysis.

Item-total Correlation

Item-total correlations provide another measure of the congruence between the way an item functions and our expectations. Typically students with high ability (i.e., those who perform well on the MSA Science overall) answer items correctly, and students with low ability (i.e., those who perform poorly on the MSA Science overall) answer items incorrectly. If these expectations are met, the point-biserial (i.e., item-total) correlation between the item and the total test score will be high and positive, indicating that the item is a good discriminator between high ability and low ability students. A correlation value above 0.20 is considered acceptable; values closer to 1.00 indicate greater discrimination. A test comprised of maximally discriminating items will maximize internal consistency reliability.

Assuming that the total test score represents the extent to which a student possesses the construct being measured by the test, high point-biserial correlations indicate that the tasks on the test require this construct to be answered correctly. Table 14 reports the mean, minimum, and maximum point-biserial correlation values for the MSA Science tests. The adjusted point-biserial removes the item score from the total score so that the index can be an unbiased estimate of the item with the test. As can be observed from this table, the average adjusted point-biserial ranged from 0.32 to 0.42 across the MSA Science tests for grades 5 and 8. MSA Science operational items in general seem to perform well in terms of differentiating students with high ability from low-performing students and measuring a common underlying construct. A portion of the field test items were somewhat less effective, which is to be expected.

Table 14. Summary of Adjusted Point-Biserial Correlations

Subject	Grade	Status	Adjusted Point-biserial		
			Mean	**Minimum**	**Maximum**
SC	5	OP	0.38	0.16	0.61
SC	5	FT	0.32	-0.10	0.60
SC	8	OP	0.42	0.17	0.70
SC	8	FT	0.35	-0.09	0.71

Note: OP=operational, FT=field test

Inter-Correlations among Standards

There are six standards within the VSC frameworks for MSA Science. 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 provided on MSA Science.

Table 15 reports the correlations among the six standards based on scale scores. The standard-level (subtest) inter-correlations ranged from 0.55 to 0.67 with majority of correlations around 0.61. The standards are moderately highly related to one another and more strongly related to the total test scores. This suggests there is some uniqueness to items grouped by standard and that they are collectively measuring a dominant overall construct (science).

Table 15. Correlation among MSA Science content standards

Grade 5 Form A	**Mean**	**sd**		**Str1**	**Str2**	**Str3**	**Str4**	**Str5**	**Str6**	**Total**
	406.66	56.12	**Str1**	1						
	406.77	70.71	**Str2**	0.59487	1					
	406.52	63.92	**Str3**	0.65268	0.59238	1				
	403.27	75.76	**Str4**	0.61157	0.55575	0.60052	1			
	410.45	70.67	**Str5**	0.60663	0.55291	0.58018	0.56679	1		
	398.57	67.93	**Str6**	0.63652	0.5891	0.64126	0.58384	0.57181	1	
	403.83	45.38	**Total**	0.8404	0.77326	0.82693	0.77722	0.77717	0.81651	1
Grade 5 Form B				**Str1**	**Str2**	**Str3**	**Str4**	**Str5**	**Str6**	**Total**
	408.81	60.22	**Str1**	1						
	406.95	76.14	**Str2**	0.58663	1					
	412.42	65.06	**Str3**	0.62338	0.57984	1				
	405.41	62.64	**Str4**	0.62177	0.57457	0.59758	1			
	421.99	86.38	**Str5**	0.59923	0.55069	0.56322	0.56082	1		
	410.12	77.48	**Str6**	0.59868	0.57298	0.59474	0.58303	0.56181	1	
	407.20	45.53	**Total**	0.82801	0.77321	0.80656	0.80219	0.75998	0.78938	1
Grade 8 Form A				**Str1**	**Str2**	**Str3**	**Str4**	**Str5**	**Str6**	**Total**
	405.75	68.61	**Str1**	1						
	402.23	75.65	**Str2**	0.62142	1					
	411.14	83.78	**Str3**	0.63343	0.59636	1				
	406.83	71.27	**Str4**	0.65626	0.60298	0.62369	1			
	401.54	71.96	**Str5**	0.63738	0.60052	0.60181	0.63544	1		
	405.84	69.71	**Str6**	0.67142	0.61615	0.64115	0.65052	0.63142	1	
	403.74	50.12	**Total**	0.84013	0.77708	0.7885	0.82561	0.80459	0.83452	1
Grade 8 Form B				**Str1**	**Str2**	**Str3**	**Str4**	**Str5**	**Str6**	**Total**
	406.49	70.15	**Str1**	1						
	401.19	87.15	**Str2**	0.58071	1					
	404.40	70.96	**Str3**	0.64227	0.59004	1				
	403.35	65.82	**Str4**	0.65249	0.59776	0.66101	1			
	407.25	80.00	**Str5**	0.61758	0.58045	0.62026	0.63602	1		
	403.56	62.62	**Str6**	0.66161	0.59986	0.6486	0.66779	0.62826	1	
	402.91	48.90	**Total**	0.81885	0.75746	0.82564	0.8432	0.79474	0.84814	1

*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 2009 MSA Science tests to further examine construct validity by evaluating the relationship between the subtest scores. Subtest raw scores were used for this analysis. 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 which 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 ≤ .06 as the guideline for close fit. Table 16 summarizes fit indicators estimated from the confirmatory factor analysis for the 2009 MSA Science tests. The confirmatory factor analysis results provide additional evidence to support the validity of the MSA Science tests. For both grades, the lowest AGFI was 0.992, and the highest RMSEA was 0.032. The AGFI and RMSEA indicators supported the model fit.

Table 16. Fit indicators for confirmatory factor analysis on MSA Science

Grade/Form	AGFI	RMSEA
Grade 5 Form A	0.995	0.025
Grade 5 Form B	0.994	0.030
Grade 8 Form A	0.997	0.021
Grade 8 Form B	0.992	0.032

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

Validity 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 17). 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 directly comparable.

Table 17. Fit indicators for accommodations/non-accommodations based CFA

	Accommodations		No Accommodations	
Grade/Form	AGFI	RMSEA	AGFI	RMSEA
Grade 5 Form A	0.996	0.019	0.995	0.027
Grade 5 Form B	0.996	0.017	0.993	0.032
Grade 8 Form A	0.999	0.000	0.996	0.023
Grade 8 Form B	0.989	0.035	0.993	0.032

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

Validity Evidence for Different Populations

The primary evidence for the validity of the MSA Science lies in the content and construct being measured. The evidence of validity is sought from a statistical analysis to detect differential item functioning that could favor a particular sub-group over and beyond the difference in ability.

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. This is done at multiple points in the item development process, and by multiple levels of reviews.

The DIF analysis was carried out on all MSA Science field test items. 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.

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) (1999). *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.

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.

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).

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.

Thissen, D., Chen, W-H., & Bock, R. D. (2003). *MUTILOG for Windows, Version 7* [Computer Software]. Lincolnwood, IL: Scientific Software International.

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

Appendix A
Item Statistics

Table A.1. Grade 5 item statistics

UIN	Status	Pvalue	Ptbis	*a*	*b*	*c*	d_1	d_2	d_3
50005	OP	0.82	0.46	1.11734	-0.81842	0.25904			
50016	OP	0.71	0.53	1.32229	-0.22105	0.25479			
50026	OP	0.80	0.31	0.51802	-1.24002	0.23931			
50028	OP	0.60	0.41	0.73274	0.18209	0.21592			
50033	OP	0.58	0.45	0.84184	0.20188	0.19847			
50034	OP	0.90	0.25	0.50764	-2.74708	0.04334			
50041	OP	0.60	0.37	0.48631	-0.37115	0.02669			
50052	OP	0.80	0.36	0.58714	-1.43702	0.08774			
50058	OP	0.77	0.43	0.75243	-1.00476	0.07442			
50059	OP	0.92	0.36	1.01880	-1.87655	0.04907			
50062	OP	0.34	0.28	0.67644	1.70351	0.18889			
50067	OP	0.33	0.51	0.54762	1.28808	0.00000	2.59137	-0.59373	-1.99764
50078	OP	0.44	0.62	0.71148	0.67735	0.00000	2.00219	0.46062	-2.46282
50079	OP	0.45	0.50	1.34947	0.69509	0.17484			
50083	OP	0.66	0.52	1.11995	-0.12212	0.19591			
50086	OP	0.56	0.38	0.80506	0.64141	0.28696			
50088	OP	0.67	0.47	0.97473	-0.10704	0.24069			
50107	OP	0.56	0.39	0.56320	0.06442	0.10709			
50109	OP	0.73	0.36	0.52847	-1.08247	0.05562			
50121	OP	0.58	0.40	0.65171	0.17538	0.19576			
50123	OP	0.58	0.47	0.92554	0.25013	0.19301			
50173	OP	0.74	0.43	0.85236	-0.38527	0.27116			
50194	OP	0.64	0.44	0.85117	-0.01316	0.24076			
50229	OP	0.49	0.36	0.50883	0.40350	0.08069			
50232	OP	0.28	0.32	0.79076	1.56002	0.12247			
50238	OP	0.59	0.31	1.00931	0.91644	0.42890			
50290	OP	0.58	0.49	0.99523	0.18620	0.18998			
50311	OP	0.92	0.37	1.03802	-1.85942	0.04565			
50329	OP	0.75	0.52	1.08490	-0.63060	0.13220			
50332	OP	0.77	0.43	0.85156	-0.62883	0.24967			
50335	OP	0.72	0.53	1.44031	-0.21008	0.29033			
50345	OP	0.64	0.45	0.98163	0.12764	0.27407			
50349	OP	0.84	0.32	0.56088	-1.90848	0.02867			
50364	OP	0.88	0.35	0.98519	-1.01235	0.47547			
50415	OP	0.45	0.34	0.81447	1.09693	0.24179			
50420	OP	0.57	0.34	0.84585	0.81806	0.35367			
50421	OP	0.57	0.43	0.81134	0.27658	0.20469			
50431	OP	0.61	0.40	0.71115	0.13139	0.23001			
50439	OP	0.60	0.42	0.71869	0.04629	0.17028			
50442	OP	0.47	0.45	0.93851	0.70117	0.17444			
50454	OP	0.60	0.41	0.68086	0.15667	0.19595			
50458	OP	0.48	0.43	0.91675	0.74227	0.20100			
50462	OP	0.42	0.42	1.07520	0.96707	0.19166			
50470	OP	0.57	0.38	0.80876	0.59523	0.28571			
50472	OP	0.44	0.47	1.16240	0.72871	0.17071			
50473	OP	0.46	0.35	0.54206	0.69579	0.12081			
50549	OP	0.73	0.33	0.46451	-1.06348	0.09547			
50550	OP	0.75	0.46	0.95135	-0.49308	0.23157			
50556	OP	0.52	0.31	0.46473	0.64058	0.18621			

UIN	Status	Pvalue	Ptbis	*a*	*b*	*c*	d_1	d_2	d_3
50566	OP	0.70	0.45	0.98063	-0.15974	0.29804			
50577	OP	0.47	0.52	0.57687	0.66448	0.00000	2.88936	0.43632	-3.32568
50578	OP	0.66	0.48	0.81426	-0.40956	0.09943			
50581	OP	0.48	0.54	1.09305	0.44943	0.10712			
50600	OP	0.75	0.41	0.73972	-0.71897	0.21096			
55149	OP	0.63	0.45	0.85731	0.08445	0.22995			
55174	OP	0.48	0.44	0.80574	0.60772	0.14556			
55198	OP	0.55	0.45	1.13296	0.46754	0.26092			
55202	OP	0.71	0.34	0.95773	0.43039	0.48336			
55206	OP	0.84	0.42	0.86529	-1.36754	0.03289			
55207	OP	0.90	0.34	0.78186	-1.91680	0.07238			
55208	OP	0.48	0.42	0.93708	0.74493	0.20746			
50056_01	OP	0.67	0.45	0.66495	-0.60604	0.03927			
50056_02	OP	0.67	0.48	0.98445	-0.19735	0.21933			
50056_03	OP	0.64	0.48	0.91248	-0.13914	0.17671			
50084_01	OP	0.60	0.44	0.90494	0.23534	0.25086			
50084_02	OP	0.40	0.32	0.85230	1.29145	0.23611			
50084_04	OP	0.42	0.41	0.75011	0.80519	0.13213			
50160_01	OP	0.75	0.44	0.84495	-0.58886	0.20980			
50160_06	OP	0.78	0.35	0.54353	-1.41746	0.04477			
50198_01	OP	0.58	0.29	0.32521	-0.47390	0.01579			
50198_06	OP	0.49	0.39	0.79043	0.70902	0.21643			
50198_07	OP	0.31	0.48	0.65254	1.45091	0.00000	2.95579	-0.93604	-2.01976
50199_02	OP	0.31	0.36	1.49572	1.36178	0.17441			
50199_06	OP	0.32	0.30	1.14213	1.55740	0.19699			
50199_08	OP	0.34	0.66	0.71532	0.97770	0.00000	1.31672	0.10619	-1.42291
50230_02	OP	0.81	0.46	1.16051	-0.73044	0.27636			
50230_04	OP	0.43	0.37	0.90994	1.05050	0.21817			
50230_05	OP	0.68	0.53	1.11114	-0.25054	0.16553			
50459_05	OP	0.55	0.39	1.00929	0.64393	0.30721			
50459_06	OP	0.39	0.56	0.59712	1.19671	0.00000	2.45536	0.50438	-2.95974
50486_01	OP	0.51	0.44	0.84527	0.52250	0.16373			
50486_05	OP	0.36	0.29	1.00927	1.54344	0.23367			
50486_06	OP	0.87	0.42	0.98556	-1.42358	0.05148			
50508_02	OP	0.27	0.23	1.13566	1.80429	0.18810			
50508_03	OP	0.75	0.41	0.77007	-0.58172	0.26453			
50508_04	OP	0.15	0.20	1.00263	2.29809	0.08863			
50510_02	OP	0.74	0.48	0.97369	-0.58369	0.18821			
50510_05	OP	0.56	0.48	1.14155	0.33935	0.23476			
50515_04	OP	0.53	0.45	1.02996	0.55244	0.22303			
50515_05	OP	0.82	0.32	0.55093	-1.34300	0.20856			
50516_01	OP	0.78	0.39	0.63612	-1.28060	0.02286			
50516_05	OP	0.37	0.40	0.96480	1.16055	0.16239			
50553_01	OP	0.75	0.47	0.99529	-0.53078	0.23989			
50553_04	OP	0.40	0.32	0.78173	1.28866	0.22426			
50553_05	OP	0.46	0.44	1.11264	0.74342	0.20146			
50558_01	OP	0.48	0.45	0.99802	0.68071	0.19053			
50558_02	OP	0.37	0.37	0.89052	1.21687	0.17053			
50587_05	OP	0.42	0.25	0.54019	1.66966	0.23887			
50587_06	OP	0.39	0.33	0.80291	1.33822	0.20599			
50588_02	OP	0.41	0.29	0.40094	1.15034	0.09796			

UIN	Status	Pvalue	Ptbis	*a*	*b*	*c*	d_1	d_2	d_3
50588_05	OP	0.53	0.40	0.94020	0.71296	0.27778			
50590_01	OP	0.71	0.38	0.66068	-0.40100	0.25238			
50590_02	OP	0.48	0.49	1.15143	0.56550	0.17221			
50590_04	OP	0.28	0.36	1.13146	1.39528	0.13390			
50592_02	OP	0.59	0.45	0.70798	-0.04374	0.09780			
50592_03	OP	0.64	0.45	0.72168	-0.29732	0.09678			
55011_01	OP	0.31	0.28	1.01800	1.59184	0.19034			
55011_02	OP	0.40	0.35	0.93895	1.17124	0.21800			
55011_06	OP	0.75	0.45	0.88112	-0.63403	0.21405			
55080_01	OP	0.52	0.40	0.64584	0.36642	0.13968			
55080_02	OP	0.82	0.45	0.92240	-1.21692	0.07132			
55080_05	OP	0.49	0.34	0.48343	0.46498	0.10160			
50019	FT	0.70	0.23	0.33555	-1.27919	0.06902			
50228	FT	0.77	0.32	0.72820	-0.23600	0.41904			
50336	FT	0.73	0.41	0.80933	-0.64389	0.17223			
50367	FT	0.29	0.50	0.62570	1.64549	0.00000	2.33257	-0.52985	-1.80273
50477	FT	0.80	0.42	0.81317	-1.12708	0.04971			
50529	FT	0.28	0.24	0.81595	1.84602	0.15481			
50552	FT	0.62	0.34	0.56693	-0.14752	0.15876			
50575	FT	0.86	0.34	0.73710	-1.45592	0.18218			
50656	FT	0.55	0.25	0.54744	0.79735	0.30443			
50658	FT	0.62	0.36	0.53403	-0.46333	0.09132			
50659	FT	0.85	0.39	0.94471	-1.10977	0.21678			
50661	FT	0.13	0.36	0.38775	2.82985	0.00000	0.39040	0.79867	-1.18906
50677	FT	0.53	0.40	0.85545	0.47134	0.19156			
50678	FT	0.72	0.36	0.59510	-0.83565	0.07969			
50679	FT	0.66	0.12	0.18136	-1.25711	0.13767			
50693	FT	0.65	0.46	0.86813	-0.37579	0.13987			
50694	FT	0.63	0.32	0.63833	0.11738	0.24979			
50696	FT	0.22	0.50	0.62700	1.90868	0.00000	1.73689	-0.35066	-1.38623
55110	FT	0.30	0.24	0.96309	1.65315	0.18235			
55167	FT	0.75	0.34	0.59627	-0.94578	0.10834			
55210	FT	0.36	0.56	0.58125	1.01351	0.00000	1.45907	0.21513	-1.67420
55230	FT	0.56	0.51	1.23542	0.29244	0.17483			
50149_01	FT	0.85	0.34	0.72315	-1.45217	0.12366			
50149_02	FT	0.60	0.40	0.76159	0.17753	0.19494			
50149_05	FT	0.58	0.29	0.45898	0.09526	0.13686			
50149_06	FT	0.41	0.08	1.32813	2.19760	0.37884			
50149_07	FT	0.13	0.10	0.65177	3.43717	0.09371			
50604_01	FT	0.72	0.32	0.56030	-0.73926	0.16936			
50604_02	FT	0.43	0.38	0.90352	0.82169	0.16210			
50604_03	FT	0.62	0.25	0.49866	0.35492	0.30641			
50604_05	FT	0.71	0.36	0.70420	-0.40782	0.23780			
50604_06	FT	0.31	0.48	0.50137	1.43108	0.00000	1.94172	-0.09817	-1.84355
50604_07	FT	0.38	0.50	0.70106	1.00623	0.00000	2.76626	-0.24520	-2.52106
50606_01	FT	0.74	0.34	0.65458	-0.56747	0.24769			
50606_02	FT	0.62	0.19	0.27307	-0.51859	0.10310			
50606_03	FT	0.78	0.42	0.93241	-0.74193	0.20830			
50606_04	FT	0.20	0.37	1.04267	1.62983	0.05131			
50606_05	FT	0.87	0.44	1.43778	-1.01226	0.20566			
50606_07	FT	0.28	0.46	0.57061	1.55513	0.00000	2.13857	-0.59585	-1.54272

UIN	Status	Pvalue	Ptbis	*a*	*b*	*c*	d_1	d_2	d_3
50607_01	FT	0.56	0.36	1.07865	0.66281	0.31631			
50607_02	FT	0.41	0.36	0.76737	0.86249	0.16380			
50607_03	FT	0.60	0.20	0.32137	-0.01232	0.20850			
50607_04	FT	0.73	0.34	0.74015	-0.30135	0.30370			
50607_06	FT	0.30	0.42	0.58278	1.83329	0.00000	3.07241	-0.63236	-2.44005
50607_07	FT	0.29	0.60	0.75772	1.34049	0.00000	1.75938	0.13587	-1.89525
50608_01	FT	0.71	0.35	0.64004	-0.59830	0.21046			
50608_02	FT	0.56	0.38	0.95217	0.50973	0.29077			
50608_03	FT	0.55	0.29	0.40887	0.03193	0.08577			
50608_04	FT	0.61	0.25	0.39780	0.07362	0.19090			
50608_05	FT	0.36	0.24	1.31961	1.56345	0.26433			
50615_01	FT	0.21	0.17	1.00071	2.13812	0.14500			
50615_02	FT	0.79	0.32	0.67923	-0.66299	0.32276			
50615_03	FT	0.58	0.45	1.09159	0.36442	0.23153			
50615_04	FT	0.41	0.22	0.71256	1.54147	0.25602			
50616_01	FT	0.14	0.06	1.41327	2.40428	0.11472			
50616_02	FT	0.53	0.33	1.00305	0.81192	0.30255			
50616_03	FT	0.92	0.32	0.89725	-1.97057	0.10328			
50616_05	FT	0.47	0.48	1.35450	0.58069	0.18449			
50617_02	FT	0.36	0.10	0.63415	2.75463	0.30449			
50617_03	FT	0.81	0.44	1.06744	-1.02173	0.16986			
50617_04	FT	0.45	0.16	0.90344	1.81876	0.36613			
50617_05	FT	0.19	0.05	1.46596	2.50592	0.16891			
50618_01	FT	0.65	0.28	0.54293	0.08525	0.29270			
50618_03	FT	0.61	0.30	0.54114	0.16449	0.21884			
50618_04	FT	0.68	0.39	0.79408	-0.19743	0.23859			
50618_05	FT	0.73	0.50	1.10861	-0.46980	0.11196			
50619_01	FT	0.85	0.40	0.93900	-1.54694	0.06938			
50619_02	FT	0.22	0.39	1.16744	1.29983	0.06347			
50619_03	FT	0.14	-0.09	-0.56440	-3.42365	0.10033			
50619_04	FT	0.27	0.14	0.69538	2.47433	0.19857			
50620_01	FT	0.41	0.39	1.13912	1.02495	0.19116			
50620_02	FT	0.40	0.26	0.54563	1.34509	0.16276			
50620_03	FT	0.65	0.40	0.68478	-0.28322	0.11223			
50620_04	FT	0.73	0.44	0.89246	-0.45036	0.19743			
50622_01	FT	0.44	0.21	1.06955	1.51992	0.33154			
50622_02	FT	0.33	0.32	1.15525	1.35309	0.18129			
50622_04	FT	0.50	0.44	0.98913	0.51850	0.18325			
50622_05	FT	0.50	0.06	1.40083	2.24658	0.47879			
50624_01	FT	0.71	0.06	0.10636	-3.32944	0.17774			
50624_02	FT	0.41	0.39	1.38545	1.02650	0.21501			
50624_03	FT	0.42	0.40	1.03486	0.95440	0.18282			
50624_04	FT	0.45	0.17	0.26537	1.43289	0.12959			
50628_01	FT	0.91	0.34	0.92649	-2.02981	0.05570			
50628_02	FT	0.30	0.22	0.62472	1.87861	0.16160			
50628_03	FT	0.57	0.43	0.73963	0.05512	0.09870			
50628_05	FT	0.24	0.13	0.48145	3.03337	0.14831			
50629_01	FT	0.61	0.44	1.12042	0.32155	0.27922			
50629_02	FT	0.86	0.21	0.41995	-2.52843	0.06911			
50629_03	FT	0.80	0.37	0.77383	-0.86950	0.22132			
50629_05	FT	0.36	0.57	0.60529	1.10004	0.00000	1.51447	0.32376	-1.83823

UIN	Status	Pvalue	Ptbis	*a*	*b*	*c*	d_1	d_2	d_3
50630_01	FT	0.32	0.21	0.89235	1.72738	0.22418			
50630_02	FT	0.46	0.34	0.94296	0.86114	0.25865			
50630_03	FT	0.63	0.25	0.35136	-0.70107	0.04731			
50630_05	FT	0.82	0.40	0.91640	-0.93722	0.20557			
50632_01	FT	0.61	0.44	0.91242	0.06777	0.20303			
50632_02	FT	0.40	0.33	0.74920	1.07461	0.16158			
50632_03	FT	0.69	0.46	0.84847	-0.46369	0.11493			
50632_04	FT	0.71	0.50	1.26588	-0.33520	0.22551			
50633_01	FT	0.67	0.12	0.18880	-1.33400	0.13442			
50633_02	FT	0.43	0.39	0.94076	0.90944	0.17351			
50633_03	FT	0.38	0.36	0.76472	1.00462	0.11685			
50633_04	FT	0.28	-0.01	-0.03393	-32.73241	0.17287			
50634_01	FT	0.39	0.21	0.31786	1.42708	0.08247			
50634_02	FT	0.43	0.23	0.48097	1.40828	0.21548			
50634_03	FT	0.21	0.04	0.78598	3.53751	0.18740			
50634_04	FT	0.47	0.26	0.58782	1.18144	0.23686			
50635_01	FT	0.31	0.15	0.35941	2.73100	0.15971			
50635_03	FT	0.28	0.33	1.42043	1.36193	0.15433			
50635_04	FT	0.33	0.18	0.72848	2.09833	0.22540			
50635_05	FT	0.54	0.34	0.97309	0.81754	0.31012			

UIN=Unique Item Number; Status=Administration condition (OP = Operational item; FT = Field Test 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*	d_1	d_2	d_3
80013	OP	0.49	0.33	0.50994	0.85531	0.18553			
80027	OP	0.38	0.25	1.08273	1.66330	0.27617			
80032	OP	0.36	0.34	0.81023	1.41108	0.18172			
80046	OP	0.45	0.35	1.13014	1.17475	0.28419			
80048	OP	0.38	0.27	0.76647	1.66532	0.24305			
80052	OP	0.46	0.42	1.43421	0.95329	0.26446			
80061	OP	0.65	0.48	0.83423	-0.20186	0.16131			
80071	OP	0.53	0.44	0.84460	0.49786	0.20940			
80074	OP	0.57	0.35	0.58303	0.45582	0.24784			
80080	OP	0.74	0.24	0.70762	0.70711	0.58609			
80081	OP	0.69	0.43	0.65025	-0.61648	0.11660			
80087	OP	0.68	0.34	0.42731	-1.01764	0.01888			
80104	OP	0.68	0.45	0.69071	-0.52576	0.12062			
80112	OP	0.63	0.45	0.73990	-0.09344	0.17168			
80117	OP	0.43	0.47	1.21380	0.79763	0.17827			
80121	OP	0.77	0.33	0.60085	-0.49284	0.36885			
80131	OP	0.59	0.45	1.23505	0.49613	0.32134			
80132	OP	0.80	0.49	1.02244	-0.98888	0.15497			
80205	OP	0.60	0.55	1.14808	0.01683	0.15591			
80209	OP	0.49	0.35	0.53896	0.69139	0.15748			
80222	OP	0.77	0.40	0.71882	-0.69121	0.27156			
80229	OP	0.56	0.42	0.73148	0.34969	0.20616			
80257	OP	0.72	0.50	0.81882	-0.71969	0.05999			
80276	OP	0.63	0.52	1.03232	-0.04694	0.18022			
80277	OP	0.80	0.42	0.85795	-0.73738	0.32264			
80279	OP	0.43	0.48	1.02769	0.75676	0.14823			
80280	OP	0.74	0.51	0.87293	-0.90681	0.03020			
80284	OP	0.63	0.42	0.94803	0.31181	0.31994			
80296	OP	0.71	0.55	1.03633	-0.58182	0.11036			
80313	OP	0.61	0.53	0.92421	-0.11045	0.11183			
80315	OP	0.40	0.69	0.77876	0.62085	0.00000	1.47547	0.10196	-1.57744
80319	OP	0.36	0.42	1.09662	1.11807	0.15787			
80325	OP	0.76	0.43	0.81134	-0.62431	0.26037			
80330	OP	0.78	0.47	0.84487	-0.97683	0.08992			
80336	OP	0.67	0.45	0.88474	-0.09516	0.27370			
80337	OP	0.67	0.39	0.67049	-0.13606	0.25406			
80421	OP	0.41	0.48	0.74414	0.63438	0.05236			
80425	OP	0.47	0.32	0.66949	1.16197	0.25236			
80447	OP	0.81	0.49	0.97682	-1.06826	0.08645			
80460	OP	0.75	0.39	0.67572	-0.59653	0.27172			
80495	OP	0.63	0.39	0.49519	-0.54629	0.02164			
80501	OP	0.66	0.50	1.01385	-0.04627	0.24305			
80546	OP	0.32	0.68	0.78642	1.05697	0.00000	1.30937	-0.09706	-1.21231
80559	OP	0.75	0.40	0.57933	-1.20659	0.00854			
80567	OP	0.67	0.51	1.15269	-0.02515	0.26179			
80576	OP	0.73	0.36	0.51681	-1.02227	0.10101			
80579	OP	0.70	0.55	1.30042	-0.27056	0.20927			
80617	OP	0.39	0.74	0.86196	0.60884	0.00000	1.06908	-0.19545	-0.87363
80648	OP	0.66	0.53	1.06380	-0.12172	0.19050			

UIN	Status	Pvalue	Ptbis	*a*	*b*	*c*	d_1	d_2	d_3
85165	OP	0.32	0.37	0.71327	1.39253	0.10031			
85197	OP	0.74	0.46	0.86443	-0.60895	0.20885			
85201	OP	0.81	0.42	0.72233	-1.37799	0.06924			
80002_01	OP	0.72	0.38	0.81122	0.06350	0.40784			
80002_04	OP	0.45	0.49	1.13010	0.76098	0.17271			
80002_06	OP	0.41	0.52	1.07014	0.76559	0.10743			
80049_01	OP	0.40	0.31	1.12830	1.34743	0.26591			
80049_02	OP	0.46	0.43	0.96683	0.78393	0.20212			
80049_04	OP	0.57	0.48	0.96699	0.26656	0.20532			
80084_01	OP	0.71	0.58	1.23054	-0.44271	0.11227			
80084_05	OP	0.57	0.54	0.96973	0.10302	0.12145			
80084_06	OP	0.38	0.32	0.80214	1.44603	0.20542			
80139_01	OP	0.82	0.50	1.29995	-0.88058	0.23090			
80139_05	OP	0.35	0.41	1.33263	1.16512	0.17158			
80139_07	OP	0.48	0.63	0.59938	0.19877	0.00000	1.74731	-0.09968	-1.64763
80238_01	OP	0.41	0.42	1.46062	1.09761	0.22782			
80238_02	OP	0.45	0.41	0.62635	0.72743	0.11158			
80238_04	OP	0.47	0.21	0.42700	1.88593	0.29404			
80328_03	OP	0.32	0.50	1.59711	1.05775	0.10216			
80328_04	OP	0.61	0.38	0.89264	0.56842	0.35107			
80328_06	OP	0.83	0.42	0.77701	-1.40242	0.08554			
80338_03	OP	0.70	0.52	1.10329	-0.33562	0.21923			
80338_04	OP	0.62	0.46	0.86944	0.03563	0.21481			
80338_06	OP	0.69	0.51	1.05145	-0.27337	0.23198			
80452_02	OP	0.57	0.40	0.90052	0.53191	0.30070			
80452_04	OP	0.73	0.39	0.58800	-0.90760	0.13033			
80455_02	OP	0.37	0.46	0.97304	0.93976	0.12464			
80455_03	OP	0.75	0.47	0.77069	-1.05533	0.01899			
80455_05	OP	0.23	0.39	1.20276	1.48651	0.09434			
80497_03	OP	0.45	0.43	0.80661	0.75577	0.15652			
80497_06	OP	0.33	0.61	0.65764	1.17677	0.00000	2.04930	-0.15299	-1.89631
80507_02	OP	0.54	0.44	1.02652	0.62427	0.26080			
80507_04	OP	0.67	0.48	0.85815	-0.18161	0.20357			
80507_05	OP	0.71	0.40	0.54517	-0.96951	0.02558			
80528_01	OP	0.47	0.43	0.72256	0.58414	0.13388			
80528_04	OP	0.64	0.41	0.99657	0.34778	0.36147			
80528_05	OP	0.74	0.40	0.90694	-0.15359	0.40422			
80529_03	OP	0.68	0.46	0.85839	-0.15407	0.23813			
80529_05	OP	0.84	0.45	1.17531	-0.83352	0.34704			
80530_01	OP	0.77	0.59	1.56027	-0.61634	0.15087			
80530_03	OP	0.58	0.39	0.80495	0.51958	0.28524			
80530_04	OP	0.71	0.54	1.11715	-0.40475	0.17029			
80534_02	OP	0.68	0.53	1.05586	-0.22936	0.17774			
80534_03	OP	0.65	0.52	1.05699	-0.03580	0.20369			
80534_08	OP	0.26	0.58	0.52275	1.39500	0.00000	0.91312	-0.22438	-0.68874
80663_03	OP	0.79	0.45	0.77673	-1.17011	0.06750			
80663_04	OP	0.72	0.45	0.74773	-0.62548	0.15386			
80663_06	OP	0.49	0.40	0.87456	0.75412	0.23535			
80666_04	OP	0.69	0.38	0.62976	-0.20756	0.26346			
80666_06	OP	0.68	0.34	0.62067	0.08519	0.35121			
80667_01	OP	0.53	0.30	0.48524	0.72386	0.22477			

UIN	Status	Pvalue	Ptbis	*a*	*b*	*c*	d_1	d_2	d_3
80667_02	OP	0.27	0.24	0.49323	2.30939	0.12108			
80667_05	OP	0.42	0.46	1.18129	0.84857	0.17777			
85057_01	OP	0.55	0.52	0.98820	0.20570	0.14056			
85057_05	OP	0.48	0.41	0.66265	0.58716	0.13928			
85057_08	OP	0.38	0.64	0.81603	0.61989	0.00000	2.11999	-0.79932	-1.32067
85080_01	OP	0.78	0.55	1.26308	-0.74580	0.14535			
85080_03	OP	0.32	0.39	0.92473	1.31359	0.13006			
85080_07	OP	0.31	0.65	0.78750	1.05060	0.00000	1.65874	-0.61343	-1.04530
80066	FT	0.67	0.54	1.15771	-0.25035	0.14105			
80347	FT	0.53	0.26	0.62827	1.05884	0.31112			
80575	FT	0.34	0.29	0.64131	1.57737	0.14036			
80608	FT	0.32	0.67	0.87395	1.04188	0.00000	1.37621	-0.09153	-1.28468
80609	FT	0.29	0.20	1.32412	1.72709	0.21527			
80610	FT	0.75	0.41	0.84984	-0.45560	0.28039			
80624	FT	0.32	0.37	1.26726	1.23639	0.14908			
80625	FT	0.55	0.49	0.92386	0.18015	0.11909			
80632	FT	0.81	0.45	0.96370	-0.95970	0.12879			
80642	FT	0.60	0.41	0.60460	-0.37179	0.08485			
80649	FT	0.42	0.71	0.89324	0.49323	0.00000	1.26222	-0.11238	-1.14985
80660	FT	0.31	0.10	0.87575	2.54058	0.26439			
80661	FT	0.61	0.41	0.55407	-0.51151	0.02865			
80743	FT	0.20	0.59	0.71149	1.61516	0.00000	0.41085	0.48274	-0.89358
80745	FT	0.38	0.57	0.73594	0.88234	0.00000	2.28473	-0.50817	-1.77656
80747	FT	0.34	0.52	0.52979	1.23190	0.00000	1.99422	0.03326	-2.02748
80748	FT	0.32	0.42	0.86079	1.09393	0.08197			
80749	FT	0.47	0.36	0.71089	0.85242	0.18773			
80765	FT	0.62	0.37	0.70375	0.26968	0.25248			
80775	FT	0.55	0.55	1.22682	0.24709	0.14470			
80154_01	FT	0.71	0.41	0.68324	-0.64712	0.08994			
80154_03	FT	0.76	0.40	0.69541	-0.97144	0.07893			
80154_04	FT	0.47	0.19	0.24434	0.75732	0.06007			
80154_05	FT	0.38	0.25	0.92841	1.54962	0.24002			
80154_06	FT	0.51	0.31	0.75695	0.91720	0.27371			
80154_08	FT	0.41	0.64	0.66391	0.85587	0.00000	0.90389	0.94185	-1.84574
80671_01	FT	0.69	0.52	1.03535	-0.29904	0.13295			
80671_02	FT	0.50	0.33	0.63013	0.78926	0.20114			
80671_03	FT	0.21	0.23	0.90511	2.06762	0.11564			
80671_04	FT	0.30	0.24	0.58699	1.92533	0.14352			
80671_05	FT	0.46	0.42	0.93006	0.77624	0.16725			
80671_06	FT	0.20	0.60	0.81877	1.70233	0.00000	1.22474	-0.36503	-0.85972
80672_01	FT	0.58	0.28	0.39785	-0.19414	0.07461			
80672_02	FT	0.42	0.04	0.05815	7.73094	0.14128			
80672_03	FT	0.65	0.45	0.84371	-0.14457	0.17774			
80672_04	FT	0.27	0.14	1.38296	2.21428	0.23340			
80672_05	FT	0.72	0.36	0.71129	-0.31837	0.29907			
80672_06	FT	0.36	0.60	0.60693	0.86375	0.00000	1.09902	0.00062	-1.09963
80674_01	FT	0.54	0.27	0.92602	1.14322	0.37727			
80674_02	FT	0.39	0.42	0.78660	0.85233	0.10623			
80674_03	FT	0.73	0.46	1.11727	-0.18013	0.30983			
80674_04	FT	0.53	0.30	0.88785	0.99490	0.35123			
80674_05	FT	0.26	0.07	2.20399	2.04858	0.22180			

UIN	Status	Pvalue	Ptbis	*a*	*b*	*c*	d_1	d_2	d_3
80674_06	FT	0.35	0.66	0.89121	0.85561	0.00000	1.91826	-0.38357	-1.53468
80675_01	FT	0.55	0.31	1.45850	0.84718	0.39938			
80675_02	FT	0.46	0.30	0.76243	1.07463	0.25153			
80675_03	FT	0.36	0.23	1.87579	1.47229	0.27369			
80675_04	FT	0.55	0.45	0.97551	0.36360	0.20670			
80675_05	FT	0.66	0.46	0.97879	-0.22826	0.25338			
80675_07	FT	0.19	0.55	0.64304	1.82941	0.00000	1.36039	-0.32815	-1.03224
80690_01	FT	0.33	0.17	1.11096	2.07472	0.26701			
80690_02	FT	0.47	0.30	0.82874	1.16396	0.27363			
80690_03	FT	0.42	0.40	0.67368	0.79517	0.08824			
80690_04	FT	0.42	0.42	0.96731	0.91614	0.14681			
80691_02	FT	0.59	0.47	1.10532	0.30672	0.22327			
80691_03	FT	0.28	-0.08	-0.16035	-6.05559	0.13827			
80691_04	FT	0.56	0.41	0.81049	0.37496	0.23910			
80691_05	FT	0.25	0.25	1.37598	1.79222	0.16676			
80692_01	FT	0.54	0.26	0.50077	0.85138	0.26745			
80692_02	FT	0.41	0.34	0.73241	1.11218	0.16679			
80692_03	FT	0.51	0.27	0.63016	1.09888	0.28106			
80692_04	FT	0.54	0.35	0.93633	0.84202	0.30157			
80694_02	FT	0.38	0.31	0.60184	1.22061	0.15036			
80694_03	FT	0.42	0.21	0.54770	1.75821	0.26821			
80694_04	FT	0.38	0.03	0.04869	12.99072	0.16183			
80694_05	FT	0.35	0.18	0.41151	2.27208	0.17675			
80696_01	FT	0.56	0.32	0.56590	0.48037	0.26062			
80696_02	FT	0.56	0.33	0.49986	0.22788	0.12649			
80696_03	FT	0.19	0.02	1.24212	2.96374	0.18199			
80696_04	FT	0.55	0.30	0.44555	0.20678	0.10833			
80697_01	FT	0.58	0.41	0.78914	0.10019	0.22167			
80697_02	FT	0.61	0.44	1.11993	0.18639	0.31005			
80697_04	FT	0.81	0.36	0.63856	-1.45401	0.08863			
80697_05	FT	0.71	0.30	0.70581	0.16697	0.42341			
80698_02	FT	0.71	0.40	0.73657	-0.41667	0.20786			
80698_03	FT	0.51	0.34	0.92613	0.90935	0.29406			
80698_04	FT	0.54	0.33	0.64416	0.56991	0.22002			
80698_05	FT	0.52	0.24	0.51498	1.03730	0.27907			
80701_01	FT	0.24	0.23	1.06369	1.97359	0.15409			
80701_02	FT	0.63	0.41	0.67810	-0.10431	0.13419			
80701_03	FT	0.62	0.28	0.38589	-0.36297	0.10246			
80701_05	FT	0.23	0.25	0.76062	2.06013	0.11202			
80702_01	FT	0.42	0.30	1.00003	1.24590	0.25003			
80702_02	FT	0.78	0.25	0.38757	-1.85676	0.07695			
80702_03	FT	0.61	0.36	1.00459	0.53865	0.34875			
80702_04	FT	0.35	0.36	0.92883	1.20964	0.15016			
80704_01	FT	0.79	0.51	1.19949	-0.78858	0.13338			
80704_02	FT	0.64	0.37	0.90812	0.34572	0.34125			
80704_05	FT	0.63	0.29	0.48530	0.02985	0.23540			
80704_06	FT	0.62	0.45	0.97009	0.09977	0.21591			
80711_01	FT	0.47	0.43	0.76209	0.46302	0.13472			
80711_02	FT	0.41	0.18	0.64528	1.93098	0.29632			
80711_04	FT	0.30	0.19	1.29350	1.75040	0.22890			
80711_05	FT	0.44	0.34	0.48230	0.61723	0.04504			

UIN	Status	Pvalue	Ptbis	*a*	*b*	*c*	d_1	d_2	d_3
80715_01	FT	0.28	0.34	0.80606	1.40381	0.10902			
80715_02	FT	0.44	0.15	2.32918	1.48169	0.36681			
80715_03	FT	0.72	0.39	0.67700	-0.75938	0.20268			
80715_04	FT	0.57	0.35	0.49405	-0.14406	0.05741			
80716_02	FT	0.62	0.26	0.36653	-0.42955	0.08722			
80716_03	FT	0.56	0.38	0.71507	0.39717	0.20404			
80716_04	FT	0.82	0.37	0.67614	-1.38272	0.11178			
80716_05	FT	0.66	0.40	0.88085	0.06765	0.28797			
80721_01	FT	0.54	0.21	0.31347	0.60763	0.16170			
80721_02	FT	0.74	0.47	0.95175	-0.56001	0.17188			
80721_04	FT	0.21	-0.09	-0.25882	-4.83239	0.11508			
80721_05	FT	0.58	0.39	0.74174	0.29258	0.20272			
80731_01	FT	0.55	0.50	1.10869	0.33174	0.16873			
80731_02	FT	0.62	0.47	0.98237	0.11588	0.20110			
80731_03	FT	0.52	0.40	0.80209	0.58136	0.19013			
80731_04	FT	0.75	0.40	0.79137	-0.53738	0.24019			

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

Appendix B

DIF Analysis

Table B.1 Grade 5 DIF results

	Black/White				Hispanic/White				Male/Female			
UIN	**Delta**	**SMD**	**Sig**	**Favor**	**Delta**	**SMD**	**Sig**	**Favor**	**Delta**	**SMD**	**Sig**	**Favor**
50019	-0.01	0.0007			0.29	0.0286			-0.30	-0.0259		
50228	-0.71	-0.0489			0.65	0.0411			-0.19	-0.0109		
50336	0.60	0.0484			-0.06	0.0024			-0.11	-0.0062		
50367	N/A	-0.0222			N/A	-0.0382			N/A	0.0924		
50477	-0.45	-0.0198			0.00	-0.0041			0.37	0.0195		
50529	0.17	0.0156			0.09	0.0150			-0.57	-0.0414		
50552	-0.48	-0.0395			0.20	0.0219			0.13	0.0125		
50575	-0.05	-0.0033			0.52	0.0349			0.56	0.0264		
50656	0.24	0.0420			-0.08	-0.0060			0.33	0.0322		
50658	-0.53	-0.0460			-0.59	-0.0519			-0.10	-0.0075		
50659	-0.14	-0.0188			0.12	0.0034			0.15	0.0079		
50661	N/A	0.0563			N/A	0.0896			N/A	0.0377		
50677	-0.25	-0.0154			-0.89	-0.0760			0.01	0.0002		
50678	-1.03	-0.0672	B	W	-0.69	-0.0401			-0.03	-0.0004		
50679	-0.01	-0.0094			-0.07	-0.0133			0.45	0.0453		
50693	-0.18	-0.0182			0.13	0.0096			0.12	0.0108		
50694	0.21	0.0295			-0.02	-0.0044			0.56	0.0513		
50696	N/A	0.0414			N/A	-0.0123			N/A	0.0114		
55110	0.11	0.0202			0.22	0.0248			-0.38	-0.0296		
55167	0.07	0.0185			-0.02	0.0070			-0.10	-0.0063		
55210	N/A	-0.0133			N/A	0.0039			N/A	0.1016		
55230	0.07	-0.0035			0.18	0.0130			-0.19	-0.0124		
50149_01	-0.13	-0.0001			0.49	0.0356			0.22	0.0121		
50149_02	0.51	0.0534			0.61	0.0574			0.45	0.0395		
50149_05	0.35	0.0391			0.94	0.0813			0.33	0.0330		
50149_06	0.91	0.0840			-0.18	-0.0101			0.04	0.0045		
50149_07	0.06	0.0007			0.23	0.0042			-0.13	-0.0053		
50604_01	0.79	0.0769			0.85	0.0883			0.29	0.0234		
50604_02	0.23	0.0200			0.66	0.0522			-0.36	-0.0301		
50604_03	0.30	0.0290			-0.21	-0.0099			0.14	0.0144		
50604_05	0.21	0.0062			-0.27	-0.0179			0.51	0.0383		
50604_06	N/A	0.0749			N/A	0.0388			N/A	0.2030	CC	F
50604_07	N/A	-0.0123			N/A	0.0336			N/A	0.1604	CC	F
50606_01	-0.16	-0.0172			-0.39	-0.0360			-0.17	-0.0121		
50606_02	0.24	0.0135			0.44	0.0442			0.06	0.0049		
50606_03	0.22	0.0170			0.38	0.0359			0.14	0.0080		
50606_04	0.33	0.0139			0.10	0.0039			0.35	0.0157		
50606_05	-0.74	-0.0341			-1.06	-0.0438	B	W	0.10	0.0040		
50606_07	N/A	0.0054			N/A	-0.0164			N/A	0.1352	BB	F
50607_01	-0.31	-0.0242			-0.05	0.0048			-0.37	-0.0341		
50607_02	0.31	0.0210			0.10	0.0076			-0.15	-0.0113		
50607_03	0.29	0.0265			0.57	0.0566			-0.14	-0.0134		
50607_04	0.46	0.0443			0.61	0.0540			0.37	0.0284		
50607_06	N/A	0.0846			N/A	0.0443			N/A	0.1090	BB	F
50607_07	N/A	-0.0386			N/A	-0.0359			N/A	0.0578		
50608_01	0.60	0.0453			0.41	0.0369			0.08	0.0071		
50608_02	-0.14	-0.0089			-0.26	-0.0241			0.00	0.0013		
50608_03	0.55	0.0520			0.26	0.0259			-0.05	-0.0044		

	Black/White				Hispanic/White				Male/Female			
UIN	Delta	SMD	Sig	Favor	Delta	SMD	Sig	Favor	Delta	SMD	Sig	Favor
50608_04	-0.34	-0.0265			-0.69	-0.0650			0.30	0.0283		
50608_05	0.28	0.0340			0.66	0.0574			0.01	0.0000		
50615_01	-0.10	0.0141			0.34	0.0293			-0.46	-0.0314		
50615_02	-0.37	-0.0179			0.56	0.0479			-0.28	-0.0183		
50615_03	-0.17	0.0072			0.88	0.0787			-1.03	-0.0823	B	M
50615_04	-0.64	-0.0450			-0.40	-0.0357			-0.42	-0.0376		
50616_01	0.59	0.0306			-0.07	-0.0006			-0.08	-0.0046		
50616_02	0.42	0.0458			0.24	0.0268			0.44	0.0415		
50616_03	-0.27	-0.0116			0.22	0.0074			-0.03	-0.0005		
50616_05	0.56	0.0315			-0.21	-0.0191			0.04	0.0049		
50617_02	0.23	0.0253			0.06	0.0045			-0.11	-0.0108		
50617_03	0.15	0.0040			0.05	0.0015			0.72	0.0348		
50617_04	-0.17	-0.0086			0.20	0.0265			0.06	0.0057		
50617_05	0.25	0.0101			-0.01	0.0031			-0.19	-0.0115		
50618_01	-0.20	-0.0032			0.96	0.0910			-0.05	-0.0053		
50618_03	0.13	0.0235			0.62	0.0657			-0.27	-0.0234		
50618_04	-0.38	-0.0168			0.50	0.0474			-0.23	-0.0163		
50618_05	-0.02	0.0045			0.65	0.0520			0.13	0.0095		
50619_01	-0.29	-0.0113			-0.53	-0.0274			0.06	0.0042		
50619_02	0.04	0.0094			0.72	0.0301			-0.14	-0.0085		
50619_03	0.07	-0.0006			-0.11	-0.0129			-0.05	-0.0017		
50619_04	0.19	0.0076			0.26	0.0249			0.38	0.0317		
50620_01	-0.17	-0.0120			-0.99	-0.0763			0.10	0.0091		
50620_02	0.56	0.0394			0.68	0.0587			-0.15	-0.0158		
50620_03	0.07	0.0079			0.18	0.0233			0.48	0.0397		
50620_04	-0.17	-0.0085			-0.87	-0.0672			0.21	0.0155		
50622_01	-0.38	-0.0461			-0.19	-0.0161			-0.25	-0.0241		
50622_02	-0.11	-0.0182			-0.30	-0.0148			-0.19	-0.0156		
50622_04	-0.51	-0.0282			-0.78	-0.0606			-0.89	-0.0709		
50622_05	0.48	0.0637			0.06	0.0148			-0.12	-0.0120		
50624_01	0.80	0.0632			0.70	0.0669			-0.09	-0.0080		
50624_02	-0.02	0.0097			0.35	0.0266			-0.23	-0.0189		
50624_03	0.32	0.0299			0.60	0.0578			-0.09	-0.0047		
50624_04	0.25	0.0089			0.17	0.0132			0.18	0.0188		
50628_01	0.02	0.0013			1.04	0.0291	B	H	-0.41	-0.0111		
50628_02	-0.01	0.0003			-0.08	-0.0050			0.10	0.0094		
50628_03	-0.25	-0.0137			0.50	0.0507			0.19	0.0125		
50628_05	-0.12	-0.0088			-0.01	-0.0059			-0.10	-0.0093		
50629_01	0.02	-0.0083			0.16	0.0074			-0.14	-0.0105		
50629_02	0.79	0.0362			-0.05	-0.0043			-0.30	-0.0125		
50629_03	-0.30	-0.0077			-0.57	-0.0329			0.47	0.0282		
50629_05	N/A	0.0062			N/A	-0.0207			N/A	0.0927		
50630_01	-0.10	-0.0133			0.28	0.0148			0.14	0.0103		
50630_02	0.41	0.0361			0.40	0.0418			-0.32	-0.0293		
50630_03	0.50	0.0593			-0.27	-0.0155			-0.36	-0.0335		
50630_05	0.90	0.0522			0.26	0.0095			0.04	0.0025		
50632_01	-0.83	-0.0313			-1.06	-0.0808	B	W	-0.24	-0.0176		
50632_02	-0.84	-0.0665			-0.23	-0.0205			-0.14	-0.0117		
50632_03	-0.41	-0.0161			-0.75	-0.0550			0.20	0.0131		
50632_04	-0.34	-0.0061			0.54	0.0455			0.54	0.0323		

	Black/White				Hispanic/White				Male/Female			
UIN	**Delta**	**SMD**	**Sig**	**Favor**	**Delta**	**SMD**	**Sig**	**Favor**	**Delta**	**SMD**	**Sig**	**Favor**
50633_01	0.24	0.0214			0.59	0.0429			0.04	0.0042		
50633_02	0.56	0.0466			1.16	0.0949	B	H	-0.32	-0.0270		
50633_03	0.24	0.0212			0.24	0.0226			0.29	0.0265		
50633_04	0.31	0.0255			0.19	0.0084			0.15	0.0139		
50634_01	-0.01	-0.0026			0.06	0.0017			-0.26	-0.0239		
50634_02	0.05	0.0072			0.44	0.0374			-0.21	-0.0219		
50634_03	-0.56	-0.0325			-0.28	-0.0188			0.06	0.0042		
50634_04	-0.26	-0.0218			0.08	0.0151			0.29	0.0251		
50635_01	-0.49	-0.0328			-0.02	0.0003			0.12	0.0109		
50635_03	-0.13	-0.0142			0.12	0.0008			0.07	0.0060		
50635_04	0.33	0.0099			0.27	0.0192			0.21	0.0167		
50635_05	0.61	0.0375			0.43	0.0364			0.30	0.0264		

UIN=Unique Item Number; Delta= Mantel-Haenszel *delta* statistic; SMD=Standardized Mean Difference statistic; Sig=denotes whether the Delta value is significantly different across compared groups and by what degree (B/BB denotes intermediate DIF, C/CC denotes large DIF); Favor=which subgroup the DIF favors (B=black, W=white, H=Hispanic, M=male, F=female)

Table B.2 Grade 8 DIF results

	Black/White				Hispanic/White				Male/Female			
UIN	**Delta**	**SMD**	**Sig**	**Favor**	**Delta**	**SMD**	**Sig**	**Favor**	**Delta**	**SMD**	**Sig**	**Favor**
80066	0.19	-0.0046			0.10	0.0008			-0.39	-0.0254		
80347	-0.43	-0.0473			-0.13	-0.0139			0.51	0.0466		
80575	-0.04	-0.0107			0.33	0.0264			0.17	0.0174		
80608	N/A	-0.1511	BB	W	N/A	-0.2009	BB	W	N/A	-0.0015		
80609	-0.32	-0.0138			-0.02	0.0115			0.44	0.0345		
80610	-0.23	-0.0091			0.24	0.0147			0.11	0.0067		
80624	-0.33	-0.0104			-0.53	-0.0329			0.11	0.0081		
80625	0.00	-0.0096			-0.06	-0.0073			0.95	0.0745		
80632	-0.04	-0.0101			0.56	0.0395			0.51	0.0267		
80642	0.22	0.0241			0.69	0.0641			0.41	0.0333		
80649	N/A	-0.1764	BB	W	N/A	-0.0821			N/A	-0.0006		
80660	0.07	0.0256			-0.05	0.0013			0.28	0.0270		
80661	0.23	0.0210			0.12	0.0129			0.50	0.0422		
80743	N/A	-0.0517			N/A	-0.0306			N/A	0.2233	CC	F
80745	N/A	-0.0679			N/A	-0.1346			N/A	0.0435		
80747	N/A	0.0957			N/A	0.0035			N/A	0.1319		
80748	-0.15	-0.0162			0.19	0.0068			-0.87	-0.0624		
80749	-0.65	-0.0443			-0.58	-0.0525			0.10	0.0097		
80765	-0.30	-0.0226			-0.03	0.0148			-0.19	-0.0140		
80775	0.93	0.0598			0.92	0.0652			0.52	0.0386		
80154_01	0.82	0.0687			0.33	0.0356			0.67	0.0504		
80154_03	0.64	0.0416			0.72	0.0523			0.26	0.0180		
80154_04	0.49	0.0327			0.60	0.0571			-0.35	-0.0332		
80154_05	-0.34	-0.0399			-0.66	-0.0670			0.39	0.0285		
80154_06	0.46	0.0488			0.58	0.0592			-0.19	-0.0179		
80154_08	N/A	-0.1801	BB	W	N/A	-0.0995			N/A	0.1174		
80671_01	0.22	0.0078			0.23	0.0158			0.71	0.0465		
80671_02	-0.26	-0.0245			-0.30	-0.0161			-0.18	-0.0186		
80671_03	-0.40	-0.0267			-0.42	-0.0172			-0.12	-0.0075		
80671_04	-0.39	-0.0222			0.04	0.0072			0.30	0.0289		
80671_05	-0.79	-0.0515			-0.82	-0.0630			-0.09	-0.0072		
80671_06	N/A	-0.0866			N/A	-0.1046			N/A	0.1370	BB	F
80672_01	0.26	0.0190			0.26	0.0177			0.51	0.0472		
80672_02	0.00	0.0125			-0.17	-0.0115			0.12	0.0156		
80672_03	0.18	0.0272			0.46	0.0442			0.02	0.0023		
80672_04	0.49	0.0138			0.31	0.0156			0.15	0.0168		
80672_05	0.11	-0.0045			0.29	0.0277			0.13	0.0089		
80672_06	N/A	0.0066			N/A	0.0351			N/A	0.2352	BB	F
80674_01	-0.09	-0.0087			0.04	-0.0075			-0.39	-0.0365		
80674_02	-0.33	-0.0313			-0.48	-0.0408			-0.63	-0.0516		
80674_03	0.59	0.0389			0.67	0.0530			-0.30	-0.0182		
80674_04	0.33	0.0513			0.50	0.0555			-0.12	-0.0118		
80674_05	0.15	0.0181			-0.21	-0.0062			-0.74	-0.0563		
80674_06	N/A	-0.0126			N/A	0.0746			N/A	0.1579	BB	F
80675_01	0.24	0.0179			0.01	0.0038			-0.14	-0.0134		
80675_02	0.28	0.0296			0.19	0.0225			-0.31	-0.0281		
80675_03	0.56	0.0369			0.50	0.0280			0.04	0.0019		
80675_04	0.07	0.0063			0.14	0.0070			0.05	0.0050		

	Black/White				Hispanic/White				Male/Female			
UIN	Delta	SMD	Sig	Favor	Delta	SMD	Sig	Favor	Delta	SMD	Sig	Favor
80675_05	-0.07	-0.0122			-0.29	-0.0304			-0.42	-0.0313		
80675_07	N/A	-0.0035			N/A	0.0119			N/A	0.1780	BB	F
80690_01	-0.03	-0.0176			-0.06	-0.0190			-0.05	-0.0052		
80690_02	-0.65	-0.0541			0.15	0.0152			0.66	0.0673		
80690_03	-0.49	-0.0314			0.26	0.0273			0.13	0.0133		
80690_04	0.09	0.0001			0.27	0.0262			-0.34	-0.0266		
80691_02	-0.51	-0.0290			-0.66	-0.0470			-0.18	-0.0133		
80691_03	0.13	-0.0078			0.65	0.0566			-0.24	-0.0193		
80691_04	0.49	0.0509			0.58	0.0609			-0.49	-0.0442		
80691_05	-0.32	-0.0017			0.26	0.0334			-0.15	-0.0092		
80692_01	0.22	0.0333			0.12	0.0201			0.10	0.0071		
80692_02	0.34	0.0234			-0.40	-0.0303			0.17	0.0147		
80692_03	0.46	0.0451			0.35	0.0325			0.33	0.0340		
80692_04	0.32	0.0188			-0.07	-0.0087			-0.74	-0.0669		
80694_02	-0.22	-0.0239			-0.77	-0.0561			0.06	0.0077		
80694_03	0.21	0.0341			0.31	0.0439			0.26	0.0276		
80694_04	-0.64	-0.0650			-0.40	-0.0463			0.05	0.0047		
80694_05	0.37	0.0234			0.55	0.0452			-0.13	-0.0145		
80696_01	0.78	0.0764			0.12	0.0157			0.55	0.0506		
80696_02	0.59	0.0471			-0.19	-0.0211			0.41	0.0377		
80696_03	0.20	0.0125			0.25	0.0177			-0.45	-0.0307		
80696_04	0.80	0.0812			1.28	0.1176	B	H	0.49	0.0444		
80697_01	-0.07	-0.0162			-0.28	-0.0227			0.59	0.0471		
80697_02	0.18	0.0100			-0.07	0.0005			-0.03	-0.0027		
80697_04	0.27	0.0186			0.32	0.0203			-0.02	-0.0022		
80697_05	0.55	0.0519			0.04	0.0075			0.09	0.0086		
80698_02	0.01	0.0080			0.72	0.0575			-0.24	-0.0151		
80698_03	0.48	0.0363			0.75	0.0654			-0.54	-0.0513		
80698_04	-0.03	-0.0079			-0.24	-0.0259			-0.24	-0.0237		
80698_05	0.12	0.0184			0.67	0.0591			-0.75	-0.0730		
80701_01	-0.34	-0.0090			-0.04	0.0063			-0.23	-0.0147		
80701_02	0.25	0.0312			0.43	0.0487			-0.17	-0.0165		
80701_03	0.27	0.0523			0.58	0.0647			0.33	0.0278		
80701_05	0.54	0.0127			-0.35	-0.0274			-0.32	-0.0193		
80702_01	0.31	0.0314			0.70	0.0568			-0.59	-0.0490		
80702_02	0.15	0.0102			1.12	0.0784	B	H	-0.15	-0.0080		
80702_03	-0.02	0.0122			0.24	0.0224			-0.96	-0.0834		
80702_04	-0.23	-0.0107			-0.69	-0.0478			-0.03	0.0017		
80704_01	-0.34	-0.0296			-0.53	-0.0304			-0.07	-0.0055		
80704_02	0.03	0.0015			-0.04	0.0029			-0.74	-0.0642		
80704_05	0.58	0.0534			-0.07	-0.0046			0.10	0.0105		
80704_06	0.03	0.0157			0.39	0.0436			0.97	0.0768		
80711_01	0.50	0.0344			0.14	0.0154			-0.04	-0.0007		
80711_02	0.03	-0.0026			0.21	0.0195			0.35	0.0340		
80711_04	0.20	0.0000			0.00	-0.0109			-0.13	-0.0101		
80711_05	0.87	0.0609			0.30	0.0229			0.29	0.0274		
80715_01	-0.01	-0.0061			-0.29	-0.0238			-0.83	-0.0589		
80715_02	-0.20	-0.0135			-0.84	-0.0816			-0.27	-0.0239		
80715_03	0.62	0.0552			1.07	0.0974	B	H	-0.27	-0.0187		
80715_04	0.33	0.0296			-0.23	-0.0261			0.17	0.0138		

	Black/White				Hispanic/White				Male/Female			
UIN	**Delta**	**SMD**	**Sig**	**UIN**	**Delta**	**SMD**	**Sig**	**UIN**	**Delta**	**SMD**	**Sig**	**UIN**
80716_02	0.30	0.0448			0.67	0.0614			0.46	0.0442		
80716_03	0.60	0.0541			0.54	0.0545			0.21	0.0192		
80716_04	0.84	0.0483			0.84	0.0488			0.12	0.0047		
80716_05	0.22	0.0060			0.75	0.0632			-0.34	-0.0279		
80721_01	0.62	0.0624			0.15	0.0262			0.56	0.0564		
80721_02	-0.61	-0.0670			0.09	-0.0009			-0.64	-0.0401		
80721_04	0.14	0.0284			-0.75	-0.0410			-0.14	-0.0069		
80721_05	0.01	0.0299			0.10	0.0160			-0.08	-0.0066		
80731_01	-0.95	-0.0780			-1.12	-0.0927	B	W	0.35	0.0253		
80731_02	-0.37	-0.0397			-0.77	-0.0780			-0.22	-0.0184		
80731_03	0.05	0.0163			-0.09	-0.0119			0.02	0.0010		
80731_04	0.23	0.0181			-0.07	0.0020			0.19	0.0110		

UIN=Unique Item Number; Delta= Mantel-Haenszel *delta* statistic; SMD=Standardized Mean Difference statistic; Sig=denotes whether the Delta value is significantly different across compared groups and by what degree (B/BB denotes intermediate DIF, C/CC denotes large DIF); Favor=which subgroup the DIF favors (B=black, W=white, H=Hispanic, M=male, F=female)